

LONDON
SCHOOL *of*
HYGIENE
& TROPICAL
MEDICINE



Strategies for imputing missing covariate values in observational data

Author:

Orlagh CARROLL

Supervisors:

Prof. Ruth KEOGH

Dr. Tim MORRIS

Thesis submitted in accordance with the requirements for the degree of Doctor of
Philosophy of the University of London

January 2022

Department of Medical Statistics
Faculty of Epidemiology and Population Health
London School of Hygiene and Tropical Medicine

Funded by The Economic and Social Research Council

Declaration of Authorship

I, Orlagh Ursula Carroll, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Signed: Orlagh U. Carroll

Date: 22.01.2022

Abstract

Multivariable model-building is an important aspect of statistical analyses and should be given careful consideration. A common issue when conducting an analysis is the presence of partially-observed covariates. Missing data in covariates are known to result in biased estimates of associations with the outcome and loss of power to detect associations. The impact of missing data in the prediction context has been less studied. When using a dataset to train a model for prediction it is essential to evaluate its performance. Two popular internal validation methods for evaluating a prediction model are K -fold cross-validation and using the bootstrap algorithm to correct for optimism. Methods for handling missing data in this process are not well established and will be the primary focus of this thesis.

Multiple imputation is a method commonly used to handle missing data involving replacing a missing value with a plausible value across multiple copies of the original dataset and will be used here to handle the various challenges that missing data pose. This thesis will assess how to combine multiple imputation with internal validation techniques in an ‘ideal’ and ‘pragmatic’ setting. The use of two imputation models is proposed, one to impute the dataset to estimate the coefficients of the prediction model and the other to evaluate the prediction model. Consideration is given to data leakage which can occur during the imputation process. The presence of missing data further presents challenges when selecting covariates and flexibly modelling covariates. An extension to the internal validation methods will include covariate selection and assessment of the functional form of continuous covariates using fractional polynomials. Finally, methods will be demonstrated using the Rotterdam breast cancer study data which is a publicly available dataset.

The final part of the thesis turns to the handling of missing data in studies of associations. While methods for handling missing data in this context are well established for simple settings, extensions to deal with considerations such as functional forms, covariate selection and time-varying effects are more challenging, and it is not clear to what extent they have been used in practice. This thesis presents findings from a systematic review investigating how researchers commonly handle missing data in observational time-to-event studies. A particular focus is given to the methods researchers used to deal with unobserved values, assess the functional forms of continuous covariates and select covariates for the model of interest. Recommendations for dealing with missing values in practice while handling these complicated aspects of an analysis are given.

Acknowledgements

A huge thank you to my supervisors Ruth Keogh and Tim Morris for their guidance, support and advice throughout my PhD. I have been so unbelievably lucky to have had two fantastic supervisors, who have made time for me and also kept me motivated when I needed it. I really enjoyed our weekly meetings and the discussions we had in them (both academic and those that went way off-tangent). Your mentorship has been much appreciated and will not be forgotten.

A big thank you also to Ian White for sitting on my advisory committee, providing advice and asking excellent, insightful questions. I would also like to acknowledge and thank Angela Wood who posed the question that sent me down the path of missing data and prediction modelling. In addition, I would like to thank the members of the informal missing data group, MiDIA, who have presented interesting work in missing data over the years which has helped to maintain my interest in methodological work.

Thank you to the Economic and Social Research Council and the UBEL Doctoral Training Programme who funded this research.

A huge thank you to my parents who have supported and motivated me through every stage of my education, and without them I would not have gotten to where I am today. I would also like to thank Pdraig for being with me through every step of my PhD and for providing constant love, support and motivation. Finally, to all of my friends, thank you for the fun tea breaks and chats through the years.

Contents

List of Figures23

List of Tables25

List of Abbreviations26

List of Notation27

1 Introduction28

| | | |
|--------|---|-----|
| 1.1 | Background. | .28 |
| 1.2 | Motivation. | .30 |
| 1.3 | Datasets to be used in the thesis. | .30 |
| 1.4 | The aim of this chapter. | .31 |
| 1.5 | An introduction to missing data. | .32 |
| 1.5.1 | Missing data mechanisms. | .32 |
| 1.6 | Methods to handle missing data. | .33 |
| 1.6.1 | Complete-case (CC) analysis. | .33 |
| 1.6.2 | Multiple Imputation (MI). | .34 |
| 1.6.3 | Congeniality in MI. | .35 |
| 1.6.4 | Imputation of covariates in Cox regression. | .36 |
| 1.7 | Flexible transformation of continuous covariates. | .37 |
| 1.7.1 | Fractional Polynomials. | .37 |
| 1.8 | Prediction Models. | .40 |
| 1.9 | Introduction to internal validation of prediction models. | .41 |
| 1.9.1 | Apparent performance. | .41 |
| 1.9.2 | Split the dataset into a training and test set. | .41 |
| 1.9.3 | Cross-validation (CV) methods. | .42 |
| 1.9.4 | Bootstrap (BS) methods. | .43 |
| 1.10 | Performance measures. | .46 |
| 1.10.1 | Continuous outcome: Mean-squared error (MSE). | .46 |
| 1.10.2 | Binary outcome: Brier Score. | .47 |
| 1.10.3 | Binary outcome: Area under the curve (AUC). | .47 |
| 1.10.4 | Binary outcome: Calibration. | .48 |
| 1.11 | Data leakage in prediction models. | .49 |
| 1.12 | Outline of thesis. | .50 |

2 Internal validation when missing data are present in covariates52

| | | |
|-------|---|-----|
| 2.1 | Introduction. | .52 |
| 2.2 | Pragmatic and Ideal performance. | .52 |
| 2.2.1 | Imputation models to assess ideal or pragmatic performance. | .53 |

| | | |
|----------|---|-----------|
| 2.3 | A summary of the current published literature. | .54 |
| 2.4 | A short note on pooling prediction models versus keeping prediction models unpooled when internally validating. | .61 |
| 2.5 | Using separate imputation models to impute the training and test sets. | .61 |
| 2.5.1 | A simple scenario with a single training and test split. | .62 |
| 2.5.2 | Relating the training and test imputation models to ideal and pragmatic performance. | .63 |
| 2.6 | Proposed methods for Cross-validation. | .66 |
| 2.6.1 | An additional consideration for MI. | .70 |
| 2.7 | Proposed methods for the bootstrap algorithms. | .71 |
| 2.7.1 | BS-then-MI. | .72 |
| 2.7.2 | MI-then-BS. | .74 |
| 2.7.3 | Other considerations. | .76 |
| 2.8 | Data Leakage. | .77 |
| 2.8.1 | Data Leakage in cross-validation. | .79 |
| 2.8.2 | Data Leakage in the bootstrap algorithms. | .79 |
| 2.9 | Conclusion. | .83 |
| 3 | Designing a simulation study to evaluate methods for combining MI and internal validation techniques | 84 |
| 3.1 | Aim. | .84 |
| 3.2 | Data-generating mechanisms (DGM). | .84 |
| 3.2.1 | Continuous outcome. | .84 |
| 3.2.2 | Binary outcome. | .85 |
| 3.2.3 | Introducing missingness. | .85 |
| 3.2.4 | Factors to vary in the simulation. | .86 |
| 3.3 | Estimands. | .87 |
| 3.4 | Methods. | .88 |
| 3.5 | Performance Measures. | .88 |
| 3.5.1 | Continuous outcome. | .88 |
| 3.5.2 | Binary outcome. | .88 |
| 3.6 | Finding the ‘Target’ performance measure. | .89 |
| 3.6.1 | Generating very large datasets to estimate the target MSE. | .89 |
| 3.6.2 | Using the fully-observed data. | .90 |
| 3.6.3 | Simulating AUC target performance for the binary outcome. | .90 |
| 3.6.4 | Generating a test set for each repetition. | .90 |
| 3.7 | Conclusion. | .92 |
| 4 | Simulation study results for cross-validation: continuous outcome | 93 |
| 4.1 | Introduction. | .93 |

| | | |
|----------|--|------------|
| 4.2 | Summary of the fully-observed data. | .93 |
| 4.3 | A brief summary of the cross-validation methods. | .94 |
| 4.4 | A brief overview of results for cross-validation. | .95 |
| 4.5 | Detailed results for cross-validation. | .97 |
| 4.5.1 | Comparing results to the MSE estimate when data are fully-observed | 97 |
| 4.5.2 | Increasing the number of imputed datasets from 5 to 25. | .101 |
| 4.5.3 | Increasing the percentage of missingness to 40%. | .102 |
| 4.5.4 | Comparing to the target performance. | .104 |
| 4.6 | Is data leakage an issue within the imputation process?. | .108 |
| 4.7 | Discussion of results for continuous outcome. | .113 |
| 5 | Simulation study results for cross-validation: binary outcome | 115 |
| 5.1 | Introduction. | .115 |
| 5.2 | Summary of the simulated fully-observed data. | .115 |
| 5.3 | Detailed results: Area under the ROC curve. | .116 |
| 5.3.1 | Comparing the methods' AUC to the estimate of the AUC when data are fully-observed. | .116 |
| 5.3.2 | Increasing the number of imputed datasets from 5 to 25. | .120 |
| 5.3.3 | Increasing the percentage of missingness to 40%. | .122 |
| 5.3.4 | Comparing each method's AUC to the target estimate of the AUC from a larger validation set. | .124 |
| 5.4 | Detailed results: Brier score. | .128 |
| 5.4.1 | Comparing each method's Brier score to the estimate of the Brier score when data are fully-observed. | .128 |
| 5.4.2 | Increasing the number of imputed datasets from 5 to 25. | .132 |
| 5.4.3 | Increasing the percentage of missingness to 40%. | .134 |
| 5.4.4 | Comparing each method's Brier score to the target estimate of the Brier score from a larger validation set. | .136 |
| 5.5 | Detailed results: Calibration intercept. | .140 |
| 5.5.1 | Comparing each method's Calibration intercept to the estimate of the Calibration intercept when data are fully-observed. | .140 |
| 5.5.2 | Increasing the number of imputed datasets from 5 to 25. | .144 |
| 5.5.3 | Increasing the percentage of missingness to 40%. | .146 |
| 5.5.4 | Comparing each method's calibration intercept to the target esti- mate of the calibration intercept from a larger validation set. | .148 |
| 5.6 | Detailed results: Calibration slope. | .152 |
| 5.6.1 | Comparing each method's Calibration slope to the estimate of the Calibration slope when data are fully-observed. | .152 |
| 5.6.2 | Increasing the number of imputed datasets from 5 to 25. | .156 |
| 5.6.3 | Increasing the percentage of missingness to 40%. | .158 |

| | | |
|----------|---|------------|
| 5.6.4 | Comparing each method’s calibration slope to the target estimate of the calibration slope from a larger validation set. | .160 |
| 5.7 | Is data leakage an issue within the imputation process?. | .164 |
| 5.8 | Discussion of results for the binary outcome. | .172 |
| 5.9 | Conclusions. | .173 |
| 6 | Simulation study results for the bootstrap: continuous outcome | 174 |
| 6.1 | Introduction. | .174 |
| 6.2 | A brief overview of the <i>BS-then-MI</i> and <i>MI-then-BS</i> methods for the <i>standard</i> bootstrap algorithm. | .178 |
| 6.3 | A comparison of reusing versus re-imputing test datasets: <i>standard</i> bootstrap algorithm. | .180 |
| 6.4 | Detailed results for the <i>standard</i> bootstrap algorithm. | .182 |
| 6.4.1 | Comparing results to the MSE estimate when data are fully-observed | 182 |
| 6.4.2 | Increasing the number of imputed datasets from 5 to 25. | .187 |
| 6.4.3 | Increasing the percentage of missingness to 40%. | .189 |
| 6.4.4 | Comparing results to the target performance. | .191 |
| 6.5 | Detailed results for the 0.632 bootstrap algorithm. | .195 |
| 6.5.1 | Comparing results to the MSE estimate when data are fully-observed | 195 |
| 6.5.2 | Increasing the number of imputed datasets from 5 to 25. | .199 |
| 6.5.3 | Increasing the percentage of missingness to 40%. | .201 |
| 6.5.4 | Comparing results to the target performance. | .203 |
| 6.6 | Is data leakage an issue within the imputation process for the <i>standard</i> and 0.632 bootstrap algorithms?. | .206 |
| 6.7 | Comparing internal validation algorithms. | .208 |
| 6.8 | Discussion of the results when the outcome is continuous. | .211 |
| 6.9 | Summary and discussion of results when the outcome is binary. | .212 |
| 6.10 | Conclusions. | .213 |
| 7 | Multiple imputation and internal validation with fractional polynomial terms in the prediction model | 215 |
| 7.1 | Introduction. | .215 |
| 7.2 | Background. | .215 |
| 7.2.1 | Approximate Bayesian bootstrapping. | .216 |
| 7.2.2 | Selecting an exponent in order to impute missing values. | .217 |
| 7.3 | Adapting the exponent selection and imputation process to handle prediction models. | .217 |
| 7.4 | <i>MI-then-validate</i> methods. | .218 |
| 7.4.1 | <i>MI-then-CV</i> | .219 |
| 7.4.2 | <i>MI-then-BS</i> | .220 |

| | | |
|----------|--|------------|
| 7.5 | <i>Validate-then-MI</i> methods. | .221 |
| 7.5.1 | <i>CV-then-MI</i> | .221 |
| 7.5.2 | <i>BS-then-MI</i> | .223 |
| 7.6 | Additional considerations when applying the proposed methods. | .224 |
| 7.6.1 | Application of an origin-shift transformation before applying the FPS or MFP procedure. | .224 |
| 7.6.2 | Choice of α -levels in the FPS and MFP procedure. | .224 |
| 7.7 | Conclusion. | .224 |
| 8 | Designing a simulation study to evaluate methods for combining MI and internal validation techniques while incorporating fractional polynomials and covariate selection | 225 |
| 8.1 | Introduction. | .225 |
| 8.2 | Aim. | .225 |
| 8.3 | Data-generating mechanisms (DGM). | .226 |
| 8.3.1 | Introducing missingness. | .227 |
| 8.3.2 | Factors to vary in the simulation. | .228 |
| 8.4 | Estimands. | .229 |
| 8.5 | Methods. | .229 |
| 8.6 | Performance Measures. | .230 |
| 8.6.1 | Assessment of the predictions. | .231 |
| 8.6.2 | Assessment of the exponent selection for imputation or the MFP procedure. | .231 |
| 8.6.3 | Assessment of covariate selection in the MFP procedure. | .231 |
| 8.7 | The ‘Target’ performance measure. | .232 |
| 8.8 | Conclusion. | .232 |
| 9 | Simulation study results for combining multiple imputation, internal validation and fractional polynomials | 233 |
| 9.1 | Introduction. | .233 |
| 9.2 | Summary of the simulated fully-observed data. | .235 |
| 9.3 | Comparing results from the FPS procedure to the MSE estimate when data are fully-observed. | .237 |
| 9.3.1 | The impact of α_E when an origin-shift transformation is not used. | .237 |
| 9.3.2 | An explanation for the large MSE estimates when an origin-shift transformation is not used. | .239 |
| 9.3.3 | The impact of α_E when an origin-shift transformation is used. | .241 |
| 9.3.4 | The impact of the origin-shift transformation on the estimated MSE when using fractional polynomials. | .245 |

| | | |
|-----------|---|------------|
| 9.3.5 | An explanation for the large MSE estimates when an origin-shift transformation is used. | .247 |
| 9.3.6 | Overall summary of the estimated MSE results. | .248 |
| 9.4 | Selection of exponents for imputation. | .250 |
| 9.5 | Selection of exponents from the fractional polynomial selection algorithm. | 252 |
| 9.5.1 | Selection of exponents when the best-fitting fractional polynomial is selected ($\alpha_E = 1$). | .252 |
| 9.5.2 | Selection of exponents when the best-fitting fractional polynomial is compared against the default inclusion as a linear covariate ($\alpha_E = 0.05$). | .254 |
| 9.6 | Results for the Multivariable fractional polynomial (MFP) algorithm. . . | .256 |
| 9.6.1 | Covariate selection of X_1 when using the MFP algorithm. | .256 |
| 9.6.2 | Covariate selection of X_2 when using the MFP algorithm. | .259 |
| 9.7 | Discussion. | .261 |
| 10 | Rotterdam breast cancer data | 264 |
| 10.1 | Introduction. | .264 |
| 10.2 | Background to the Rotterdam data. | .264 |
| 10.3 | Developing a prediction model in the Rotterdam dataset. | .265 |
| 10.4 | Splitting data into a training and external validation dataset. | .266 |
| 10.4.1 | Sample size calculation for building a prediction model. | .266 |
| 10.4.2 | Creating a training and holdout set. | .266 |
| 10.5 | Introducing missing data to covariates. | .267 |
| 10.6 | The imputation model used when multiply imputing. | .267 |
| 10.7 | Evaluating performance when data are fully-observed. | .269 |
| 10.7.1 | Internal validation. | .269 |
| 10.7.2 | External validation. | .269 |
| 10.8 | Evaluating performance when data are partially-observed. | .270 |
| 10.8.1 | Internal validation. | .270 |
| 10.8.2 | External validation. | .270 |
| 10.9 | Results. | .271 |
| 10.9.1 | Baseline characteristics of the training and holdout datasets. . . . | .271 |
| 10.9.2 | Training a prediction model when the data are fully-observed. . . | .273 |
| 10.9.3 | Applying the proposed methods to the partially-observed Rotterdam dataset. | .273 |
| 10.10 | Discussion. | .275 |
| 11 | How are missing data in covariates handled in observational time-to-event studies in oncology? A systematic review | 277 |
| 11.1 | Relevant material not included in the finalised journal article. | .312 |

| | | |
|-----------|---|------------|
| 11.1.1 | Prediction modelling and Inference modelling. | .312 |
| 11.1.2 | Validation methods used in prediction modelling. | .313 |
| 11.1.3 | Discussion of the additional information. | .313 |
| 12 | Discussion | 314 |
| 12.1 | Summary of the key results. | .314 |
| 12.1.1 | Combining internal validation with multiple imputation. | .314 |
| 12.1.2 | How are missing data handled in practice?. | .316 |
| 12.1.3 | Training and validating a prediction model in practice. | .317 |
| 12.1.4 | What does this thesis contribute to the field?. | .320 |
| 12.2 | Limitations. | .320 |
| 12.3 | Future extensions. | .322 |
| 12.4 | Conclusions. | .324 |
| | Bibliography | 325 |
| | Appendices | 335 |
| A | Chapter1: The Introduction | 335 |
| A.1 | Multivariable FPs (MFP): Selecting a FP for multiple covariates. | .335 |
| B | Chapter3: The simulation set-up | 337 |
| C | Chapter6: Write-up of the simulations results when the outcome is binary | 338 |
| C.1 | Introduction. | .338 |
| C.2 | Detailed results for the AUC performance of the 0.632 algorithm. | .340 |
| C.2.1 | Comparing results to the AUC estimate when data are fully-observed | 340 |
| C.2.2 | Increasing the number of imputed datasets from 5 to 25. | .344 |
| C.2.3 | Increasing the percentage of missingness to 40%. | .346 |
| C.2.4 | Comparing results to the target performance. | .348 |
| C.3 | Detailed results for the Brier score performance of the 0.632 algorithm. | .352 |
| C.3.1 | Comparing results to the Brier score estimate when data are fully-observed. | .352 |
| C.3.2 | Increasing the number of imputed datasets from 5 to 25. | .356 |
| C.3.3 | Increasing the percentage of missingness to 40%. | .358 |
| C.3.4 | Comparing results to the target performance. | .360 |
| C.4 | Detailed results for the calibration intercept performance of the 0.632 algorithm. | .364 |
| C.4.1 | Comparing results to the calibration intercept estimate when data are fully-observed. | .364 |
| C.4.2 | Increasing the number of imputed datasets from 5 to 25. | .368 |

| | | |
|----------|--|------|
| C.4.3 | Increasing the percentage of missingness to 40%. | .370 |
| C.4.4 | Comparing results to the target performance. | .372 |
| C.5 | Detailed results for the calibration slope performance of the 0.632 algorithm | 376 |
| C.5.1 | Comparing results to the calibration slope estimate when data are fully-observed. | .376 |
| C.5.2 | Increasing the number of imputed datasets from 5 to 25. | .380 |
| C.5.3 | Increasing the percentage of missingness to 40%. | .382 |
| C.5.4 | Comparing results to the target performance. | .384 |
| C.6 | Comparing reusing and re-imputing test imputed datasets of the original dataset. | .388 |
| C.7 | Overview of results for the standard bootstrap algorithm. | .390 |
| C.7.1 | Area under the ROC curve. | .390 |
| C.7.2 | Brier Score. | .402 |
| C.7.3 | Calibration intercept. | .414 |
| C.7.4 | Calibration slope. | .426 |
| C.8 | Is data leakage an issue for the <i>standard</i> and 0.632 bootstrap algorithms? | .438 |
| C.9 | Comparing internal validation algorithms. | .439 |
| | | |
| D | Chapter11: Systematic review - Supplementary file 1442 | |
| D.1 | Search Terms. | .442 |
| D.2 | Data extraction checklist. | .445 |
| D.3 | Papers included in the review. | .450 |

Supplementary Plots471

S1 Chapter4: Cross-validation and MI for the continuous outcome471

S1.1 Continuous outcome. 471

 S1.1.1 MSE from methods compared to the fully-observed MSE ($MSE_{imp} - MSE_{obs}$). 473

 S1.1.2 Comparing data leakage ($MSE_{imp} - MSE_{obs}$). 479

 S1.1.3 The proportion of missingness is 40% ($MSE_{imp} - MSE_{obs}$). 486

 S1.1.4 Comparing M=5 versus M=25 ($MSE_{imp} - MSE_{obs}$). 498

 S1.1.5 MSE from imputation methods compared to the target MSE (MSE_{target}) using a larger validation set. 511

 S1.1.6 Comparing data leakage ($MSE_{imp} - MSE_{target}$). 517

S2 Chapter4: Cross-validation and MI for the binary outcome524

S2.1 AUC. 524

 S2.1.1 AUC from imputation methods compared to the fully-observed AUC ($AUC_{imp} - AUC_{obs}$). 524

 S2.1.2 The proportion of missingness is 40% ($AUC_{imp} - AUC_{obs}$). 527

 S2.1.3 Comparing M=5 versus M=25 ($AUC_{imp} - AUC_{obs}$). 531

 S2.1.4 AUC from imputation methods compared to the target AUC (AUC_{target}) using a larger validation set. 536

S2.2 Brier Score. 539

 S2.2.1 Brier score from imputation methods compared to the fully-observed Brier score ($Brier_{imp} - Brier_{obs}$). 539

 S2.2.2 The proportion of missingness is 40% ($Brier_{imp} - Brier_{obs}$). 542

 S2.2.3 Comparing M=5 versus M=25 ($Brier_{imp} - Brier_{obs}$). 547

 S2.2.4 Brier score from imputation methods compared to the target Brier score ($Brier_{target}$) using a larger validation set. 552

S2.3 Calibration intercept. 555

 S2.3.1 Calibration intercept from imputation methods compared to the fully-observed calibration intercept ($intercept_{imp} - intercept_{obs}$). 555

 S2.3.2 The proportion of missingness is 40% ($intercept_{imp} - intercept_{obs}$). 558

 S2.3.3 Comparing M=5 versus M=25 ($intercept_{imp} - intercept_{obs}$). 563

 S2.3.4 Calibration intercept from imputation methods compared to the target calibration intercept ($intercept_{target}$) using a larger validation set 568

S2.4 Calibration slope. 571

 S2.4.1 Calibration slope from imputation methods compared to the fully-observed calibration slope ($slope_{imp} - slope_{obs}$). 571

 S2.4.2 The proportion of missingness is 40% ($slope_{imp} - slope_{obs}$). 574

 S2.4.3 Comparing M=5 versus M=25 ($slope_{imp} - slope_{obs}$). 579

| | | |
|-----------|---|------------|
| S2.4.4 | Calibration slope from imputation methods compared to the target calibration slope (slope_{target}) using a larger validation set. | .584 |
| S2.5 | Is data leakage an issue?. | .587 |
| S2.5.1 | AUC. | .587 |
| S2.5.2 | Brier Score. | .593 |
| S2.5.3 | Calibration intercept. | .599 |
| S2.5.4 | Calibration slope. | .605 |
| | | |
| S3 | Chapter6: Bootstrap and MI (continuous outcome) | 612 |
| S3.1 | Reusing versus re-imputing for test performance of the standard algorithm. | 612 |
| S3.2 | The standard bootstrap. | .626 |
| S3.2.1 | MSE from imputation methods compared to the fully-observed MSE ($\text{MSE}_{imp}-\text{MSE}_{obs}$). | .626 |
| S3.2.2 | The proportion of missingness is 40% ($\text{MSE}_{imp}-\text{MSE}_{obs}$). | .633 |
| S3.2.3 | Comparing M=5 versus M=25 ($\text{MSE}_{imp}-\text{MSE}_{obs}$). | .643 |
| S3.2.4 | MSE from imputation methods compared to the target MSE (MSE_{target}) using a larger validation set. | .649 |
| S3.3 | The 0.632 bootstrap. | .656 |
| S3.3.1 | MSE from imputation methods compared to the fully-observed MSE ($\text{MSE}_{imp}-\text{MSE}_{obs}$). | .656 |
| S3.3.2 | The proportion of missingness is 40% ($\text{MSE}_{imp}-\text{MSE}_{obs}$). | .663 |
| S3.3.3 | Comparing M=5 versus M=25 ($\text{MSE}_{imp}-\text{MSE}_{obs}$). | .675 |
| S3.3.4 | MSE from imputation methods compared to the target MSE (MSE_{target}) using a larger validation set. | .681 |
| S3.4 | Comparing internal validation methods. | .687 |
| | | |
| S4 | ChapterC: Bootstrap and MI (binary outcome) | 694 |
| S4.1 | The <i>standard</i> bootstrap: AUC. | .694 |
| S4.1.1 | Reusing versus re-imputing for test performance of the <i>standard</i> algorithm. | .694 |
| S4.1.2 | AUC from imputation methods compared to the fully-observed AUC ($\text{AUC}_{imp}-\text{AUC}_{obs}$). | .698 |
| S4.1.3 | The proportion of missingness is 40% ($\text{AUC}_{imp}-\text{AUC}_{obs}$). | .700 |
| S4.1.4 | Comparing M=5 versus M=25 ($\text{AUC}_{imp}-\text{AUC}_{obs}$). | .704 |
| S4.1.5 | AUC from imputation methods compared to the target AUC (AUC_{target}) using a larger validation set. | .706 |
| S4.2 | The standard bootstrap: Brier Score. | .709 |
| S4.2.1 | Reusing versus re-imputing for test performance of the standard algorithm. | .709 |

| | | |
|--------|---|------|
| S4.2.2 | Brier Score from imputation methods compared to the fully-observed Brier Score ($\text{Brier}_{imp}-\text{Brier}_{obs}$) | .713 |
| S4.2.3 | The proportion of missingness is 40% ($\text{Brier}_{imp}-\text{Brier}_{obs}$) | .716 |
| S4.2.4 | Comparing M=5 versus M=25 ($\text{Brier}_{imp}-\text{Brier}_{obs}$) | .720 |
| S4.2.5 | Brier Score from imputation methods compared to the target Brier Score (Brier_{target}) using a larger validation set. | .722 |
| S4.3 | The standard bootstrap: Calibration intercept and slope. | .724 |
| S4.3.1 | Reusing versus re-imputing for test performance of the standard algorithm. | .724 |
| S4.3.2 | Calibration intercept and slope from imputation methods compared to the fully-observed Calibration intercept and slope ($\text{Cal}_{imp}-\text{Cal}_{obs}$) | .734 |
| S4.3.3 | The proportion of missingness is 40% ($\text{Cal}_{imp}-\text{Cal}_{obs}$) | .739 |
| S4.3.4 | Comparing M=5 versus M=25 ($\text{Cal}_{imp}-\text{Cal}_{obs}$) | .749 |
| S4.3.5 | Calibration intercept and slope from imputation methods compared to the target Calibration intercept and slope (Cal_{target}) using a larger validation set. | .754 |
| S4.4 | The 0.632 bootstrap: AUC. | .760 |
| S4.4.1 | AUC from imputation methods compared to the fully-observed AUC ($\text{AUC}_{imp}-\text{AUC}_{obs}$) | .760 |
| S4.4.2 | The proportion of missingness is 40% ($\text{AUC}_{imp}-\text{AUC}_{obs}$) | .762 |
| S4.4.3 | Comparing M=5 versus M=25 ($\text{AUC}_{imp}-\text{AUC}_{obs}$) | .766 |
| S4.4.4 | AUC from imputation methods compared to the target AUC (AUC_{target}) using a larger validation set. | .768 |
| S4.5 | The 0.632 bootstrap: Brier Score. | .771 |
| S4.5.1 | Brier Score from imputation methods compared to the fully-observed Brier Score ($\text{Brier}_{imp}-\text{Brier}_{obs}$) | .771 |
| S4.5.2 | The proportion of missingness is 40% ($\text{Brier}_{imp}-\text{Brier}_{obs}$) | .774 |
| S4.5.3 | Comparing M=5 versus M=25 ($\text{Brier}_{imp}-\text{Brier}_{obs}$) | .778 |
| S4.5.4 | Brier Score from imputation methods compared to the target Brier Score (Brier_{target}) using a larger validation set. | .780 |
| S4.6 | The 0.632 bootstrap: Calibration intercept and slope. | .782 |
| S4.6.1 | Calibration intercept and slope from imputation methods compared to the fully-observed Calibration intercept and slope ($\text{Cal}_{imp}-\text{Cal}_{obs}$) | .782 |
| S4.6.2 | The proportion of missingness is 40% ($\text{Cal}_{imp}-\text{Cal}_{obs}$) | .787 |
| S4.6.3 | Comparing M=5 versus M=25 ($\text{Cal}_{imp}-\text{Cal}_{obs}$) | .797 |
| S4.6.4 | Calibration intercept and slope from imputation methods compared to the target Calibration intercept and slope (Cal_{target}) using a larger validation set. | .802 |
| S4.7 | Comparing internal validation methods. | .808 |
| S4.7.1 | AUC. | .808 |

| | | |
|-----------|--|------------|
| S4.7.2 | Brier score. | .810 |
| S4.7.3 | Calibration intercept. | .812 |
| S4.7.4 | Calibration slope. | .814 |
| | | |
| S5 | Chapter9: Simulation study results for FPS, comparison of MSE (Section9.3) | 816 |
| S5.1 | Cross-validation. | .816 |
| S5.1.1 | No origin shift transformation applied. | .816 |
| S5.1.2 | An origin shift transformation applied. | .829 |
| S5.2 | The 0.632 bootstrap. | .842 |
| S5.2.1 | No origin shift transformation applied. | .842 |
| S5.2.2 | An origin shift transformation applied. | .855 |
| | | |
| S6 | Chapter9: Simulation study results for FPS, exponent selection (Section 9.3) | 868 |
| S6.1 | ABB exponent selection. | .868 |
| S6.1.1 | Cross-validation. | .868 |
| S6.1.2 | The 0.632 bootstrap. | .883 |
| S6.2 | Exponent selection from the FPS procedure. | .898 |
| S6.2.1 | Cross-validation, $\alpha_E = 1$ and no origin-shift. | .898 |
| S6.2.2 | Cross-validation, $\alpha_E = 0.05$ and no origin-shift. | .913 |
| S6.2.3 | Cross-validation, $\alpha_E = 1$ and an origin-shift has been applied. | .928 |
| S6.2.4 | Cross-validation, $\alpha_E = 0.05$ and an origin-shift has been applied. | .943 |
| S6.2.5 | The 0.632 bootstrap, exponents selected in the bootstrap samples: $\alpha_E = 1$ and no origin-shift. | .958 |
| S6.2.6 | The 0.632 bootstrap, exponents selected in the bootstrap samples: $\alpha_E = 0.05$ and no origin-shift. | .973 |
| S6.2.7 | The 0.632 bootstrap, exponents selected in the bootstrap samples: $\alpha_E = 1$ and an origin-shift has been applied. | .988 |
| S6.2.8 | The 0.632 bootstrap, exponents selected in the bootstrap samples: $\alpha_E = 0.05$ and an origin-shift has been applied. | .1003 |
| S6.2.9 | The 0.632 bootstrap, exponents selected using all the data: $\alpha_E = 1$ and no origin-shift. | .1018 |
| S6.2.10 | The 0.632 bootstrap, exponents selected using all the data: $\alpha_E =$ 0.05 and no origin-shift. | .1033 |
| S6.2.11 | The 0.632 bootstrap, exponents selected using all the data: $\alpha_E = 1$ and an origin-shift has been applied. | .1048 |
| S6.2.12 | The 0.632 bootstrap, exponents selected using all the data: $\alpha_E =$ 0.05 and an origin-shift has been applied. | .1063 |

S7 Chapter9: Simulation study results for MFP, comparison of MSE (Section9.6)1078

S7.1 Cross-validation.1078
 S7.1.1 $\beta_2 = 1$ and an origin shift transformation has not been applied. . .1078
 S7.1.2 $\beta_2 = 1$ and an origin shift transformation has been applied. . . .1091
 S7.1.3 $\beta_2 = 0$ and an origin shift transformation has not been applied. . .1104
 S7.1.4 $\beta_2 = 0$ and an origin shift transformation has been applied. . . .1117
 S7.2 The 0.632 bootstrap.1130
 S7.2.1 $\beta_2 = 1$ and an origin shift transformation has not been applied. . .1130
 S7.2.2 $\beta_2 = 1$ and an origin shift transformation has been applied. . . .1143
 S7.2.3 $\beta_2 = 0$ and an origin shift transformation has not been applied. . .1156
 S7.2.4 $\beta_2 = 0$ and an origin shift transformation has been applied. . . .1169

S8 Chapter9: Simulation study results for MFP, exponent selection (Section9.6)1182

S8.1 ABB exponent selection.1182
 S8.1.1 Cross-validation, $\beta_2 = 1$1182
 S8.1.2 Cross-validation, $\beta_2 = 0$1197
 S8.1.3 The 0.632 bootstrap, $\beta_2 = 1$1212
 S8.1.4 The 0.632 bootstrap, $\beta_2 = 0$1227
 S8.2 Exponent selection from the FPS procedure.1242
 S8.2.1 Cross-validation, $\beta_2 = 1$, $\alpha_E = 1$ and no origin-shift.1242
 S8.2.2 Cross-validation, $\beta_2 = 1$, $\alpha_E = 0.05$ and no origin-shift.1257
 S8.2.3 Cross-validation, $\beta_2 = 1$, $\alpha_E = 1$ and an origin-shift has been used.1272
 S8.2.4 Cross-validation, $\beta_2 = 1$, $\alpha_E = 0.05$ and an origin-shift has been used1287
 S8.2.5 Cross-validation, $\beta_2 = 0$, $\alpha_E = 1$ and no origin-shift.1302
 S8.2.6 Cross-validation, $\beta_2 = 0$, $\alpha_E = 0.05$ and no origin-shift.1317
 S8.2.7 Cross-validation, $\beta_2 = 0$, $\alpha_E = 1$ and an origin-shift has been used.1332
 S8.2.8 Cross-validation, $\beta_2 = 0$, $\alpha_E = 0.05$ and an origin-shift has been used1347
 S8.2.9 The 0.632 bootstrap, exponents selected in the bootstrap samples:
 $\beta_2 = 1$, $\alpha_E = 1$ and no origin-shift.1362
 S8.2.10The 0.632 bootstrap, exponents selected in the bootstrap samples:
 $\beta_2 = 1$, $\alpha_E = 0.05$ and no origin-shift.1377
 S8.2.11The 0.632 bootstrap, exponents selected in the bootstrap samples:
 $\beta_2 = 1$, $\alpha_E = 1$ and an origin-shift has been applied.1392
 S8.2.12The 0.632 bootstrap, exponents selected in the bootstrap samples:
 $\beta_2 = 1$, $\alpha_E = 0.05$ and an origin-shift has been applied.1407
 S8.2.13The 0.632 bootstrap, exponents selected in the bootstrap samples:
 $\beta_2 = 0$, $\alpha_E = 1$ and no origin-shift.1422

| | | |
|---------|---|-------|
| S8.2.14 | The 0.632 bootstrap, exponents selected in the bootstrap samples: $\beta_2 = 0, \alpha_E = 0.05$ and no origin-shift. | .1437 |
| S8.2.15 | The 0.632 bootstrap, exponents selected in the bootstrap samples: $\beta_2 = 0, \alpha_E = 1$ and an origin-shift has been applied. | .1452 |
| S8.2.16 | The 0.632 bootstrap, exponents selected in the bootstrap samples: $\beta_2 = 0, \alpha_E = 0.05$ and an origin-shift has been applied. | .1467 |
| S8.2.17 | The 0.632 bootstrap, exponents selected using all the data: $\beta_2 = 1,$ $\alpha_E = 1$ and no origin-shift. | .1482 |
| S8.2.18 | The 0.632 bootstrap, exponents selected using all the data: $\beta_2 = 1,$ $\alpha_E = 0.05$ and no origin-shift. | .1497 |
| S8.2.19 | The 0.632 bootstrap, exponents selected using all the data: $\beta_2 = 1,$ $\alpha_E = 1$ and an origin-shift has been applied. | .1512 |
| S8.2.20 | The 0.632 bootstrap, exponents selected using all the data: $\beta_2 = 1,$ $\alpha_E = 0.05$ and an origin-shift has been applied. | .1527 |
| S8.2.21 | The 0.632 bootstrap, exponents selected using all the data: $\beta_2 = 0,$ $\alpha_E = 1$ and no origin-shift. | .1542 |
| S8.2.22 | The 0.632 bootstrap, exponents selected using all the data: $\beta_2 = 0,$ $\alpha_E = 0.05$ and no origin-shift. | .1557 |
| S8.2.23 | The 0.632 bootstrap, exponents selected using all the data: $\beta_2 = 0,$ $\alpha_E = 1$ and an origin-shift has been applied. | .1572 |
| S8.2.24 | The 0.632 bootstrap, exponents selected using all the data: $\beta_2 = 0,$ $\alpha_E = 0.05$ and an origin-shift has been applied. | .1587 |

S9 Chapter9: Simulation study results for MFP, covariate selection (Section9.6)1603

| | | |
|--------|--|-------|
| S9.1 | Cross-validation. | .1603 |
| S9.1.1 | Covariate selection of X_2 : $\beta_2 = 1, \alpha_E = 1,$ no origin-shift. | .1603 |
| S9.1.2 | Covariate selection of X_2 : $\beta_2 = 1, \alpha_E = 0.05,$ no origin-shift. | .1619 |
| S9.1.3 | Covariate selection of X_2 : $\beta_2 = 1, \alpha_E = 1$ and an origin-shift has been applied. | .1635 |
| S9.1.4 | Covariate selection of X_2 : $\beta_2 = 1, \alpha_E = 0.05$ and an origin-shift has been applied. | .1651 |
| S9.1.5 | Covariate selection of X_1 : $\beta_2 = 1, \alpha_E = 1,$ no origin-shift. | .1667 |
| S9.1.6 | Covariate selection of X_1 : $\beta_2 = 1, \alpha_E = 0.05,$ no origin-shift. | .1683 |
| S9.1.7 | Covariate selection of X_1 : $\beta_2 = 1, \alpha_E = 1$ and an origin-shift has been applied. | .1699 |
| S9.1.8 | Covariate selection of X_1 : $\beta_2 = 1, \alpha_E = 0.05$ and an origin-shift has been applied. | .1715 |
| S9.2 | The 0.632 bootstrap. | .1731 |
| S9.2.1 | Covariate selection of X_2 using all data: $\beta_2 = 1, \alpha_E = 1,$ no origin-shift | 1731 |

| | |
|--|-------|
| S9.2.2 Covariate selection of X_2 using all data: $\beta_2 = 1$, $\alpha_E = 0.05$, no origin-shift. | .1747 |
| S9.2.3 Covariate selection of X_2 using all data: $\beta_2 = 1$, $\alpha_E = 1$ and an origin-shift has been applied. | .1763 |
| S9.2.4 Covariate selection of X_2 using all data: $\beta_2 = 1$, $\alpha_E = 0.05$ and an origin-shift has been applied. | .1779 |
| S9.2.5 Covariate selection of X_1 using all data: $\beta_2 = 1$, $\alpha_E = 1$, no origin-shift | 1795 |
| S9.2.6 Covariate selection of X_1 using all data: $\beta_2 = 1$, $\alpha_E = 0.05$, no origin-shift. | .1811 |
| S9.2.7 Covariate selection of X_1 using all data: $\beta_2 = 1$, $\alpha_E = 1$ and an origin-shift has been applied. | .1827 |
| S9.2.8 Covariate selection of X_1 using all data: $\beta_2 = 1$, $\alpha_E = 0.05$ and an origin-shift has been applied. | .1843 |
| S9.2.9 Covariate selection of X_2 in the bootstrap samples: $\beta_2 = 1$, $\alpha_E = 1$, no origin-shift. | .1859 |
| S9.2.10 Covariate selection of X_2 in the bootstrap samples: $\beta_2 = 1$, $\alpha_E = 0.05$, no origin-shift. | .1875 |
| S9.2.11 Covariate selection of X_2 in the bootstrap samples: $\beta_2 = 1$, $\alpha_E = 1$ and an origin-shift has been applied. | .1891 |
| S9.2.12 Covariate selection of X_2 in the bootstrap samples: $\beta_2 = 1$, $\alpha_E = 0.05$ and an origin-shift has been applied. | .1907 |
| S9.2.13 Covariate selection of X_1 in the bootstrap samples: $\beta_2 = 1$, $\alpha_E = 1$, no origin-shift. | .1923 |
| S9.2.14 Covariate selection of X_1 in the bootstrap samples: $\beta_2 = 1$, $\alpha_E = 0.05$, no origin-shift. | .1939 |
| S9.2.15 Covariate selection of X_1 in the bootstrap samples: $\beta_2 = 1$, $\alpha_E = 1$ and an origin-shift has been applied. | .1953 |
| S9.2.16 Covariate selection of X_1 in the bootstrap samples: $\beta_2 = 1$, $\alpha_E = 0.05$ and an origin-shift has been applied. | .1968 |

List of Figures

| | | |
|-----|---|------|
| 1.1 | An example of complete-case analysis. | .33 |
| 1.2 | Diagram depicting the multiple imputation procedure. | .35 |
| 1.3 | An example of splitting data into a training and test set.. . . . | .42 |
| 1.4 | An example of cross-validation for $K = 3$ | .42 |
| 1.5 | The difference between the <i>standard</i> and 0.632 algorithms for one bootstrap sample b | .45 |
| 1.6 | An example of a ROC curve (black line).. | .47 |
| 1.7 | Simple example of data leakage: k -means. | .50 |
| 2.1 | An example of splitting data into a training set and a test set when using MI.63 | |
| 2.2 | An example of splitting data into a training set (grey) and a test set (purple) before or after imputing the original data.. . . . | .78 |
| 2.3 | Data leakage flow in the <i>standard</i> algorithm for combining multiple imputation and the bootstrap. | .80 |
| 2.4 | Data leakage in the 0.632 algorithm for combining multiple imputation (MI) and the bootstrap (BS). | .82 |
| 4.1 | The difference between $MSE_{imp} - MSE_{obs}$ when $R^2 = 0.1$ for the cross-validation methods.. . . . | .96 |
| 4.2 | The difference $MSE_{imp} - MSE_{obs}$ when data are weakly covariate-dependent MAR for $M = 5$ when 25% of values are missing in X_1 | .98 |
| 4.3 | The difference $MSE_{imp} - MSE_{obs}$ when data are weakly outcome- and covariate-dependent MAR for $M = 5$ when 25% of values are missing in X_1 | .100 |
| 4.4 | The difference $MSE_{imp} - MSE_{obs}$ when data are weakly outcome- and covariate-dependent MAR for cross-validation when $M = 25$ | .101 |
| 4.5 | Comparing the impact of increasing the percentage of missingness on the difference $MSE_{imp} - MSE_{obs}$ when $M = 5$ and data are weakly outcome- and covariate-dependent MAR.. . . . | .103 |
| 4.6 | The difference $MSE_{imp} - MSE_{target}$ when data are weakly covariate-dependent MAR for cross-validation.. . . . | .105 |
| 4.7 | The difference $MSE_{imp} - MSE_{target}$ when data are weakly outcome- and covariate-dependent MAR for cross-validation.. . . . | .107 |
| 4.8 | Assessing data leakage within the imputation process for cross-validation. The difference $MSE_{imp} - MSE_{obs}$ is compared when data are weak outcome- and strong covariate-dependent MAR.. . . . | .109 |
| 4.9 | Assessing data leakage within the imputation process for cross-validation. The difference $MSE_{imp} - MSE_{target}$ is compared when when $R^2 = 0.1, 0.3$ and data are weak outcome- and strong covariate-dependent MAR.. . . . | .112 |
| 5.1 | The difference $AUC_{imp} - AUC_{obs}$ when data are MCAR or covariate-dependent MAR for cross-validation.. . . . | .117 |

| | | |
|------|--|------|
| 5.2 | The difference $AUC_{imp} - AUC_{obs}$ when data are outcome-dependent or outcome- and covariate-dependent MAR for cross-validation.. | .119 |
| 5.3 | The difference $AUC_{imp} - AUC_{obs}$ when data are outcome-dependent or outcome- and covariate-dependent MAR for cross-validation when $M = 25$. | 121 |
| 5.4 | Comparing the impact of increasing the percentage of missingness on the difference $AUC_{imp} - AUC_{obs}$ when data are outcome- and covariate-dependent MAR for cross-validation when $M = 5$ | .123 |
| 5.5 | The difference $AUC_{imp} - AUC_{target}$ when data are MCAR or covariate-dependent MAR for cross-validation when $M = 5$ | .125 |
| 5.6 | The difference $AUC_{imp} - AUC_{target}$ when data are outcome-dependent or outcome- and covariate-dependent MAR for cross-validation when $M = 5$. | 127 |
| 5.7 | The difference $Brier_{imp} - Brier_{obs}$ when data are MCAR or covariate-dependent MAR for cross-validation.. | .129 |
| 5.8 | The difference $Brier_{imp} - Brier_{obs}$ when data are outcome-dependent or outcome- and covariate-dependent MAR for cross-validation.. | .131 |
| 5.9 | The difference $Brier_{imp} - Brier_{obs}$ when data are outcome-dependent or outcome- and covariate-dependent MAR for cross-validation when $M = 25$. | 133 |
| 5.10 | Comparing the impact of increasing the percentage of missingness on the difference $Brier_{imp} - Brier_{obs}$ when data are outcome- and covariate-dependent MAR for cross-validation when $M = 5$ | .135 |
| 5.11 | The difference $Brier_{imp} - Brier_{target}$ when data are MCAR or covariate-dependent MAR for cross-validation when $M = 5$ | .137 |
| 5.12 | The difference $Brier_{imp} - Brier_{target}$ when data are outcome-dependent or outcome- and covariate-dependent MAR for cross-validation when $M = 5$. | 139 |
| 5.13 | The difference $Intercept_{imp} - Intercept_{obs}$ when data are MCAR or covariate-dependent MAR for cross-validation.. | .141 |
| 5.14 | The difference $Intercept_{imp} - Intercept_{obs}$ when data are outcome-dependent or outcome- and covariate-dependent MAR for cross-validation.. | .143 |
| 5.15 | The difference $Intercept_{imp} - Intercept_{obs}$ when data are outcome-dependent or outcome- and covariate-dependent MAR for cross-validation when $M = 25$. | 145 |
| 5.16 | Comparing the impact of increasing the percentage of missingness on the difference $Intercept_{imp} - Intercept_{obs}$ when data are outcome- and covariate-dependent MAR for cross-validation when $M = 5$ | .147 |
| 5.17 | The difference $Intercept_{imp} - Intercept_{target}$ when data are MCAR or covariate-dependent MAR for cross-validation when $M = 5$ | .149 |
| 5.18 | The difference $Intercept_{imp} - Intercept_{target}$ when data are outcome-dependent or outcome- and covariate-dependent MAR for cross-validation when $M = 5$. | 151 |
| 5.19 | The difference $Slope_{imp} - Slope_{obs}$ when data are MCAR or covariate-dependent MAR for cross-validation.. | .153 |

| | | |
|------|--|-----|
| 5.20 | The difference $\text{Slope}_{imp} - \text{Slope}_{obs}$ when data are outcome-dependent or outcome- and covariate-dependent MAR for cross-validation.. | 155 |
| 5.21 | The difference $\text{Slope}_{imp} - \text{Slope}_{obs}$ when data are outcome-dependent or outcome- and covariate-dependent MAR for cross-validation when $M = 25$. | 157 |
| 5.22 | Comparing the impact of increasing the percentage of missingness on the difference $\text{Slope}_{imp} - \text{Slope}_{obs}$ when data are outcome- and covariate-dependent MAR for cross-validation when $M = 5$ | 159 |
| 5.23 | The difference $\text{Slope}_{imp} - \text{Slope}_{target}$ when data are MCAR or covariate-dependent MAR for cross-validation when $M = 5$ | 161 |
| 5.24 | The difference $\text{Slope}_{imp} - \text{Slope}_{target}$ when data are outcome-dependent or outcome- and covariate-dependent MAR for cross-validation when $M = 5$. | 163 |
| 5.25 | Assessing data leakage within the imputation process for cross-validation for the AUC. | 166 |
| 5.26 | Assessing data leakage within the imputation process for cross-validation for the Brier score. | 168 |
| 5.27 | Assessing data leakage within the imputation process for cross-validation for the calibration intercept. | 170 |
| 5.28 | Assessing data leakage within the imputation process for cross-validation for the calibration slope. | 171 |
| 6.1 | The difference between $\text{MSE}_{imp} - \text{MSE}_{obs}$ when $R^2 = 0.1$ for the <i>standard</i> bootstrap algorithm.. | 179 |
| 6.2 | A comparison of reusing versus re-imputing test datasets for the <i>standard</i> bootstrap algorithm. | 181 |
| 6.3 | The difference $\text{MSE}_{imp} - \text{MSE}_{obs}$ when data are weakly covariate-dependent MAR for the <i>standard</i> bootstrap algorithm.. | 184 |
| 6.4 | The difference $\text{MSE}_{imp} - \text{MSE}_{obs}$ when data are weakly outcome- and covariate-dependent MAR for the <i>standard</i> bootstrap algorithm.. | 186 |
| 6.5 | The difference $\text{MSE}_{imp} - \text{MSE}_{obs}$ when data are weakly outcome- and covariate-dependent MAR for the <i>standard</i> bootstrap algorithm when $M = 25$ | 188 |
| 6.6 | Comparing the impact of increasing the percentage of missingness on the difference $\text{MSE}_{imp} - \text{MSE}_{obs}$ when data are weakly outcome- and covariate-dependent MAR for the <i>standard</i> bootstrap algorithm when $M = 5$ | 190 |
| 6.7 | The difference $\text{MSE}_{imp} - \text{MSE}_{target}$ when data are weakly covariate-dependent MAR for the <i>standard</i> bootstrap algorithm.. | 192 |
| 6.8 | The difference $\text{MSE}_{imp} - \text{MSE}_{target}$ when data are weakly outcome- and covariate-dependent MAR for the <i>standard</i> bootstrap algorithm.. | 194 |
| 6.9 | The difference $\text{MSE}_{imp} - \text{MSE}_{obs}$ when data are weakly covariate-dependent MAR for the 0.632 bootstrap algorithm.. | 196 |
| 6.10 | The difference $\text{MSE}_{imp} - \text{MSE}_{obs}$ when data are weakly outcome- and covariate-dependent MAR for the 0.632 bootstrap algorithm.. | 198 |

| | | |
|------|--|------|
| 6.11 | The difference $MSE_{imp} - MSE_{obs}$ when data are weakly covariate-dependent MAR for the 0.632 bootstrap algorithm when $M = 25$ | .200 |
| 6.12 | Comparing the impact of increasing the percentage of missingness on the difference $MSE_{imp} - MSE_{obs}$ when data are weakly outcome- and covariate-dependent MAR for the 0.632 bootstrap algorithm when $M = 5$ | .202 |
| 6.13 | The difference $MSE_{imp} - MSE_{target}$ when data are weakly covariate-dependent MAR for the 0.632 bootstrap algorithm.. | .204 |
| 6.14 | Comparing cross-validation, and the 0.632 and <i>standard</i> Bootstrap using the target MSE.. | .210 |
| 9.1 | The difference $MSE_{imp} - MSE_{obs}$ for cross-validation when the true exponent is -2, data are MCAR or covariate-dependent MAR and $R^2 = 0.1$. An origin-shift transformation has been applied.. | .243 |
| 9.2 | The proportion of times an exponent was selected to impute missing values using the ABB exponent selection. | .251 |
| 9.3 | The proportion of times an exponent was selected via FPS post-imputing when $\alpha_E = 1$ and an origin-shift transformation was used. | .253 |
| 9.4 | The proportion of times an exponent was selected via FPS post-imputing when $\alpha_E = 0.05$ and an origin-shift transformation was used. | .255 |
| 9.5 | The proportion of times covariate X_1 is selected for inclusion to the prediction model when using the MFP algorithm. | .258 |
| 9.6 | The proportion of times covariate X_2 is selected for inclusion to the prediction model when using the MFP algorithm. | .260 |

List of Tables

| | | |
|-----|--|-----|
| 1.1 | The degrees of freedom for comparing the difference in deviances in fractional polynomial (FP) models. | .38 |
| 2.1 | Proposed methods by Wood, Royston and White for handling missing data in training and test sets.. | .56 |
| 2.2 | Imputing a covariate with missing data, X_1 , in the test set under an ideal or pragmatic performance estimand. X_2 and Y are a fully-observed additional covariate and outcome, respectively.. | .64 |
| 2.3 | Methods for combining cross-validation and MI for a pragmatic scenario when a covariate is partially-observed. Cross-validation will be applied first, followed by MI.. | .67 |
| 2.4 | Methods for combining cross-validation and pragmatic imputation for an incomplete covariate when imputing first. | .71 |
| 3.1 | Specification of parameter values ψ_0, ψ_2, ψ_3 to ensure MCAR, weak MAR and strong MAR with approximately 25% ($\psi_{0,25}$) or 40% ($\psi_{0,40}$) of observations induced to be missing.. | .86 |

| | | |
|-----|---|-----|
| 3.2 | Specification of parameter values ψ_0, ψ_2, ψ_3 to ensure MCAR, weak and strong MAR for missingness dependent and not dependent on the outcome. | 86 |
| 3.3 | Factors which will be varied for the continuous outcome simulations. | 87 |
| 4.1 | The mean and variance of the outcome Y across the 2000 simulated datasets. The min and max values of Y are the minimum and maximum across all repetitions.. | 93 |
| 4.2 | Summary of the MSE estimates when data are fully-observed. This is summarised from the 2000 simulated repetitions.. | 94 |
| 4.3 | Brief summary of methods A-K for combining multiple imputation and cross-validation. | 94 |
| 4.4 | Brief summary of methods B-E and J-K for combining multiple imputation and cross-validation. | 108 |
| 5.1 | Summarising performance when data are fully-observed for the 2000 simulated datasets. | 115 |
| 6.1 | A brief summary of the various methods under consideration to combine the bootstrap (BS) algorithms with multiple imputation (MI).. | 175 |
| 6.2 | Summary of MSE_{obs} for cross-validation (CV) and the <i>standard</i> (Std) and 0.632 bootstrap algorithms.. | 208 |
| 8.1 | Variance of the outcome (σ^2) used in the simulation study for each choice of the exponent and level of R^2 | 226 |
| 8.2 | Specification of parameter values ψ_0, ψ_2, ψ_3 to ensure MCAR, weak MAR and strong MAR with approximately 25% ($\psi_{0,25}$) of observations induced to be missing.. | 228 |
| 8.3 | Factors which will be varied for the continuous outcome simulations. | 229 |
| 8.4 | Factors which will be varied in the MFP analysis. | 230 |
| 9.1 | Parameters in the analysis procedure which will be assessed.. | 234 |
| 9.2 | The mean and variance of the outcome Y across the 2000 simulated datasets. The min and max values of Y are the minimum and maximum across all repetitions.. | 235 |
| 9.3 | The MSE estimates when data are fully-observed (MSE_{obs}) for cross-validation and the 0.632 bootstrap. | 236 |
| 9.4 | An example of the impact of low imputed values on the performance of a prediction model. | 239 |
| 9.5 | Applying an origin-shift transformation to X_1 to improve the performance of a prediction model. | 240 |
| 9.6 | The estimated MSE and Monte Carlo 95% confidence interval (CI) when the exponent is -2 and an origin-shift has been applied.. | 244 |
| 9.7 | Estimated MSE results before and after applying an origin-shift transformation when $E = 0$ and $\alpha_E = 0.05$ | 246 |

| | | |
|------|--|------|
| 9.8 | An example demonstrating how large imputed values of X_1 can lead to poor predictions. | .247 |
| 10.1 | Specification of the ψ parameter values to ensure covariate-dependent MAR (scenario 1) and outcome- and covariate-dependent MAR.. . . . | .267 |
| 10.2 | Covariates which will be included in the imputation models used to impute each of the partially-observed covariates.. . . . | .268 |
| 10.3 | Baseline Characteristics stratified by those who had the event or were censored within 231 months of follow-up. | .272 |
| 10.4 | Results from the MFP algorithm on the Rotterdam dataset when data are fully-observed ($N = 2,684$). | .273 |
| 10.5 | The estimated C -statistic when using the proposed methods on the Rotterdam data.. . . . | .274 |
| 11.1 | Comparing covariate selection for predictive and inference modelling. . . | .312 |

List of Abbreviations

| | |
|-------------|--|
| ABB | Approximate Bayesian bootstrap |
| AUC | Area under the curve |
| BS | Bootstrap |
| CC | Complete-case |
| CI | Confidence interval |
| CV | Cross-validation |
| DF | Degrees of Freedom |
| DGM | Data-generating mechanism |
| FCS | Full conditional specification |
| FP | Fractional polynomial |
| FPD | Fractional polynomial of degree D |
| FPS | Fractional polynomial selection |
| MAR | Missing at random |
| MCAR | Missing completely at random |
| MFP | Multivariable fractional polynomial |
| MFPT | Multivariable fractional polynomial time |
| MI | Multiple imputation |
| MICE | Multivariate imputation by chained equations |
| MNAR | Missing not at random |
| MSE | Mean-squared error |
| OCP | Optimism-corrected performance |
| OOB | Out-of-bag |
| PMM | Predictive mean matching |
| ROC | Receiver operating characteristic (curve) |
| SD | Standard deviation |

List of Notation

| <i>Data</i> | |
|-------------------------------|---|
| Y | An outcome. |
| X_1, X_2 | Covariates related to Y . X_1 is typically partially-observed and X_2 is fully-observed. |
| N_{obs} | The number of observations in a dataset. |
| P | The number of covariates considered for inclusion into a model. |
| <i>Fractional Polynomials</i> | |
| E | A fractional polynomial exponent which will be applied to a continuous covariate X_1^E . |
| α_E | Significance level for the exponent selection process. Default value of 0.05. |
| α_β | Significance level for covariate selection. Default value of 0.05. |
| I_δ | Indicator for the use of an origin-shift transformation on a continuous covariate. |
| <i>Models</i> | |
| $g(Y X_1, X_2, \dots; \beta)$ | informal notation of an analysis model regressing an outcome Y on covariates, parameterised by β . |
| $f(X_1 X_2, Y; \psi)$ | informal notation for an imputation model to impute X_1 , parameterised by ψ . |
| <i>Performance</i> | |
| $\text{Perf}(P, D)$ | A value of performance from applying prediction model P to dataset D . |
| Perf_{obs} | Performance can represent the MSE, AUC, Brier score or calibration intercept or slope which is estimated and averaged across 2000 simulated datasets. |
| Perf_{imp} | Performance can represent the MSE, AUC, Brier score or calibration intercept or slope which is estimated in 2000 simulated datasets using a proposed method imp . The 2000 estimates are then averaged. |
| Perf_{target} | Performance can represent the MSE, AUC, Brier score or calibration intercept or slope which is estimated in a validation dataset |

Note to the reader

Internal hyperlinks have been used throughout this thesis. When reading the thesis on a computer, the reader may click on a hyperlink to refer back to a relevant section. Once finished, the reader can click Alt and \leftarrow (at the same time) to return to the original location.

1 Introduction

1.1 Background

Missing data are a common problem in observational data, occurring due to a failure to observe a value for a covariate. This may be due to any number of reasons such as failure to respond to questions in a questionnaire, data entry errors or patient loss to follow-up. The presence of missing data can cause several issues for researchers when conducting a statistical analysis. It can lead to a loss of power to detect associations between covariates and the outcome of interest, as well as introduce bias into the estimates for these associations [1]. In addition, it can cause difficulties when making decisions on common analysis issues. Two examples of this are the selection of covariates into a statistical model or allowing for the flexible transformation of continuous covariates in the presence of missing data.

Multivariable model-building is an important aspect of many statistical analysis, being commonly used in many types of studies. There are typically three classified aims of quantitative research which are exploratory, causal or predictive in nature. An exploratory model is used for descriptive purposes, it can suggest that an exposure or treatment is associated with an outcome but cannot help with drawing firm causal conclusions. A causal model can be used to try and establish evidence for a causal relationship between a treatment and an outcome. For an exploratory or causal model, the presence of missing data can be problematic in inference as regression parameters can potentially be biased while other covariates may incorrectly be noted to not be associated with an outcome. There are now many recommendations in place concerning the handling of missing data in inference modelling (I have summarised recommendations in Table 4 of Chapter 11). However, another multivariable model-building setting involves prediction modelling which aims to determine a patient's risk of having or developing a health outcome. To date, much of the published literature has focused on the effects of missing data in an inference setting while the effects of missing data in a prediction modelling setting have been less formally investigated. While much of the published literature and recommendations, which focus on an inference setting, may transcribe to a prediction setting, it is important to note some key differences.

Missing data in an inference setting is often concerned with bias which may be introduced to regression parameters. However, bias in a prediction setting is only problematic if it causes a model to produce worse predicted values. An additional difference, is using information from the outcome when imputing missing values. In inference, maintaining associations between imputed values and the outcome is essential. However, in a prediction setting the question arises as to whether the very thing that is about to be predicted should be used to impute missing values. Another consideration is that with inference modelling a 'final' inference model is desired from which to draw associations between

exposures and an outcome. Using missing data methods such as multiple imputation (this will be introduced in Section 1.6.2) will involve combining several analysis models together to get one overall model. This may not be necessary in a prediction setting, which could avoid difficulties associated with getting an overall model from multiple models which each include different covariates or transformations of continuous covariates. In this thesis, I will primarily focus on the handling of missing data in a prediction setting.

Covariate selection and the transformation of continuous covariates are two common decisions made, in addition to missing data, during the development of multivariable prediction models. Several systematic reviews have assessed the reporting of prediction models for various health areas. Collins et al. (2011) [2] found that 41% of studies developing prediction models for type II diabetes did not consider missing data. More recently, Navarro et al. (2021) [3] noted that 41% of studies had handled missing data inappropriately when developing prediction models using supervised machine-learning approaches, either omitting records with missing data or using a ‘flawed’ imputation approach. Limited detail was available from the review on why a complete-case analysis was considered flawed (there are some circumstances for which a complete-case analysis can be acceptable to use [4]). Similarly, Tsvetanova et al. (2021) [5] found a lack of reported detail on how missing data are handled during the development, validation and implementation stages of a prediction model. The handling of missing data is not solely a problem in a traditional regression-based prediction model but also in machine-learning [6].

The TRIPOD statement [7] is a set of recommendations focusing on the analysis and reporting of prediction models. These recommendations range from detailing the study objectives clearly to explaining how the sample size of the study was decided and stating what type of prediction model was used. Specifically in relation to missing data, the TRIPOD statement gives three recommendations. These are (i) stating the proportion of missingness in each covariate, (ii) stating what method was used to handle missing data and (iii) discussing any limitations that missing data has caused. As stated in a systematic review on the reporting and handling of missing data in predictive research by Masconi et al. (2015) [8], there is little consideration given to the effect of missing data in risk prediction. Masconi concludes that formal guidelines may improve the reporting and handling of missing data for future studies.

There are several published articles on handling missing data in prediction modelling but advice can often conflict which can make it difficult to implement guidelines and recommendations in practice. A specific issue concerns the handling of internal validation methods when data are partially-observed. A thorough study into combining missing data methods with internal validation techniques is required and will be investigated in this thesis.

In the introduction chapter of this thesis I will detail the motivation for the work conducted during the PhD. This will include the impact of missing data on observational studies and how this thesis is a response to the practical question of how to combine internal validation methods with multiple imputation. I will give an introduction to missing data, prediction models and internal validation in this chapter.

1.2 Motivation

The tentative aim of my PhD was to develop strategies for handling missing data in time-to-event analyses. This would have involved incorporating covariate selection, selecting the functional form of continuous covariates (i.e. covariate transformation) and the handling of time-varying effects when using multiple imputation (a method used to impute missing values in datasets). In the first year of my PhD I focused on reviewing the current literature on multiple imputation and conducting a systematic review concerning how missing data are handled in practice. The systematic review covered different study types, including studies of associations and prediction studies. This resulted in the work presented in Chapter 11 and a corresponding paper was published [9]. I also investigated fractional polynomials to handle covariate selection, covariate transformation and time-varying effects with the ultimate aim being to combine an algorithm called ‘multivariable fractional polynomial time’ with multiple imputation.

In November 2018 in the second year of the PhD I attended an informal missing data discussion group where the question was raised by Professor Angela Wood regarding the best way to combine multiple imputation and cross-validation when developing and validating a prediction model. As Professor Wood had no time to investigate this and it was a problem which my supervisors and I found to be highly interesting, it was decided that I would undertake a detailed investigation into the validation of prediction models in the presence of missing covariate values as a primary aim of my PhD. I have since focused on the handling of missing data when using cross-validation or the bootstrap optimism-corrected algorithms.

1.3 Datasets to be used in the thesis

The majority of the data that will be used in this thesis to assess methods for handling missing data in the development of prediction models will be simulated datasets. This will allow for the evaluation of the methods under controlled circumstances where the underlying data-generating processes are known [10].

In addition, the methods that perform well (based on findings from the simulation studies)

will be applied to a real dataset. The *Rotterdam breast cancer dataset* is a publicly available dataset, available for [download](#) from the Institute of Medical Biometry and Statistics. It has 2,982 fully-observed records and is used throughout the ‘Multivariable Model-Building’ book by Royston and Sauerbrei [11] for example analyses. In this thesis, I will use the dataset to illustrate the final methods selected from the simulation studies.

1.4 The aim of this chapter

There are two key areas of statistical research which must be introduced in this chapter. The first is missing data. I will briefly describe the underlying missing data mechanisms which can cause missing values to arise, followed by discussing two common methods (complete-case analysis and multiple imputation) which are used to handle missing data in practice.

The second area is that of prediction modelling. I will briefly describe the uses of prediction models before providing an overview of methods used for the evaluation of prediction models, focusing on internal validation. This will detail two common internal validation approaches (cross-validation and the optimism-corrected bootstrap) and explain several performance measures that are used when evaluating model performance. Finally I will introduce the concept of data leakage which will play an important role in the methods I will propose in later chapters.

I will finish by outlining the remainder of the thesis at the end of this Chapter, giving a brief summary of each subsequent chapter.

1.5 An introduction to missing data

In this section I will give a brief overview of the area of missing data. As missing data is a very common problem which arises in healthcare analyses, there are various texts available giving a detailed overview of the area [12,13,14]. Here, I shall give a brief introduction to the missing data concepts which are relevant to the thesis. This will include a description of the underlying missing data mechanisms and a description of two methods, complete-case analysis and multiple imputation, which are most commonly used in practice to handle partially-observed data. In this section and throughout the thesis, I will be focusing on the handling of missing data in covariates rather than outcomes.

1.5.1 Missing data mechanisms

Three ‘missing data mechanisms’ were defined by Rubin and Little (2002) [13, p. 12] to explain the potential relationship between missing values and the rest of a dataset. Let R_i be an indicator variable specifying whether a value is missing (1) or not (0). Let a dataset \mathbf{D} contain an outcome Y and a matrix of covariates \mathbf{X} . The subset of observed and missing covariates for patient i in \mathbf{D} are denoted $\mathbf{D}_{i,Obs}$ and $\mathbf{D}_{i,Miss}$ i.e. $\mathbf{D}_i = \{\mathbf{D}_{i,Obs}, \mathbf{D}_{i,Miss}\}$.

Missing completely at random (MCAR) implies that the probability of missingness is not conditional on whether data are observed or missing.

$$\Pr(R_i = 1 \mid \mathbf{D}_{i,Obs}, \mathbf{D}_{i,Miss}) = \Pr(R_i = 1)$$

Missing at random (MAR) implies that the probability of data being missing is conditionally independent of the missing data given the observed data.

$$\Pr(R_i = 1 \mid \mathbf{D}_{i,Obs}, \mathbf{D}_{i,Miss}) = \Pr(R_i = 1 \mid \mathbf{D}_{i,Obs})$$

Missing Not at random (MNAR) implies that the probability of data being missing depends on both the observed and unobserved data.

$$\Pr(R_i = 1 \mid \mathbf{D}_{i,Obs}, \mathbf{D}_{i,Miss}) \neq \Pr(R_i = 1 \mid \mathbf{D}_{i,Obs})$$

It is not possible to determine whether data are actually MAR or MNAR, though it is possible to test MCAR against MAR (if we are willing to rule out MNAR). Instead, it is determined by the plausibility of the missing mechanism within the context of the data. In this thesis, I shall mainly discuss the performance of methods when data are MCAR or MAR.

1.6 Methods to handle missing data

A number of statistical methods have been developed for how to deal with missing data, including weighting approaches like inverse probability weighting, looking only at the observed values or imputing missing values with the mean or mode value for the covariate being assessed. Here, I shall detail two methods: complete-case analysis and multiple imputation. I previously stated in Section 1.1 that much of the published literature regarding missing data focused on the inference setting. Altering the missing data methods discussed here to handle a prognostic setting, instead of the inference unbiased parameter estimation setting, will be discussed in Chapter 2.

1.6.1 Complete-case (CC) analysis

This is a common method used by researchers to deal with missing observations and is often the default method for dealing with missing data in statistical software such as R or Stata. It involves restricting the analysis of interest to the dataset of those who have fully-observed data as seen in Figure 1.1.

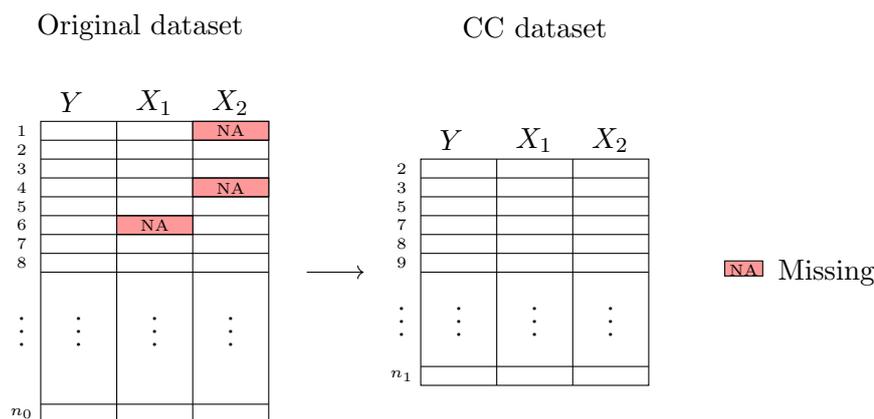


Figure 1.1: An example of complete-case analysis. The total sample size for the data before and after CC analysis is n_0 and n_1 , respectively, where $n_1 < n_0$. The CC dataset is then used for the analysis.

If data are MCAR then the results from complete-case analysis will lead to unbiased estimates, as the records are still simply a random sample of the population. However, this method is inefficient due to the discarding of information i.e. a smaller sample size is used, as seen in Figure 1.1. The complete-case analysis could also provide valid inference in certain MAR scenarios such as regression analysis when the missing mechanism does not depend on the outcome. For example, if there are missing values in either the outcome, the covariates or both, then as long as the probability of being fully-observed is independent of the outcome when conditioned on the covariates, a complete-case analysis will be unbiased [12, p.24-25, 34-35].

1.6.2 Multiple Imputation (MI)

Originally proposed by Rubin [15], the idea behind MI is to create M copies of the original dataset and replace the missing data in each imputed dataset with plausible values using a model. Using several imputed datasets helps to account for the uncertainty in the estimates of interest due to the imputation process. By accounting for this uncertainty in imputed values, multiple imputation is efficient and can produce unbiased estimates of regression parameters and standard errors under the MAR assumption. It is also flexible and can handle various covariate types (continuous, binary etc.) or different datasets such as longitudinal or multi-level data. A final result is obtained by summarising across these imputed datasets, each of which has different imputed values.

The steps in the MI process are visualised in Figure 1.2 and described below:

1. Create M copies of the original dataset, \mathbf{D} .
2. Replace missing data in each copy with plausible values drawn from the posterior predictive distribution of the missing data conditional on the observed.
 - This involves first forming an imputation model with parameters ψ , $f(\mathbf{D}_{Miss} | \mathbf{D}_{Obs}; \psi)$, under an assumption about the missingness mechanism.
 - Initial values for ψ are estimated on the complete-cases. Given the initial values, a draw of ψ can be taken from its posterior distribution. This can then be used to impute the missing values.
 - Taking draws from the model and the posterior distribution of ψ is repeated M times.
3. Apply the analysis procedure (e.g. fit the analysis model of interest) to each imputed dataset and get estimates of the parameters of interest, $\hat{\beta}_m$ for $m = 1, \dots, M$.
4. Combine or ‘pool’ these estimates using Rubin’s rules. An overall point estimate is obtained using Rubin’s first rule $\hat{\beta} = \sum_{m=1}^M \hat{\beta}_m$. Rubin’s second rule estimates the total variance of $\hat{\beta}$ [16]:

$$\text{Var}(\hat{\beta}) = \frac{1}{M} \sum_{m=1}^M W_m + \left(1 + \frac{1}{M}\right) * \frac{1}{M-1} \sum_{m=1}^M (\hat{\beta}_m - \hat{\beta})^2$$

where W_m is the estimated variance of $\hat{\beta}_m$.

MI is typically conducted using the MCAR or MAR assumption, which I will focus on within the PhD, although it can be extended to be implemented with MNAR [17].

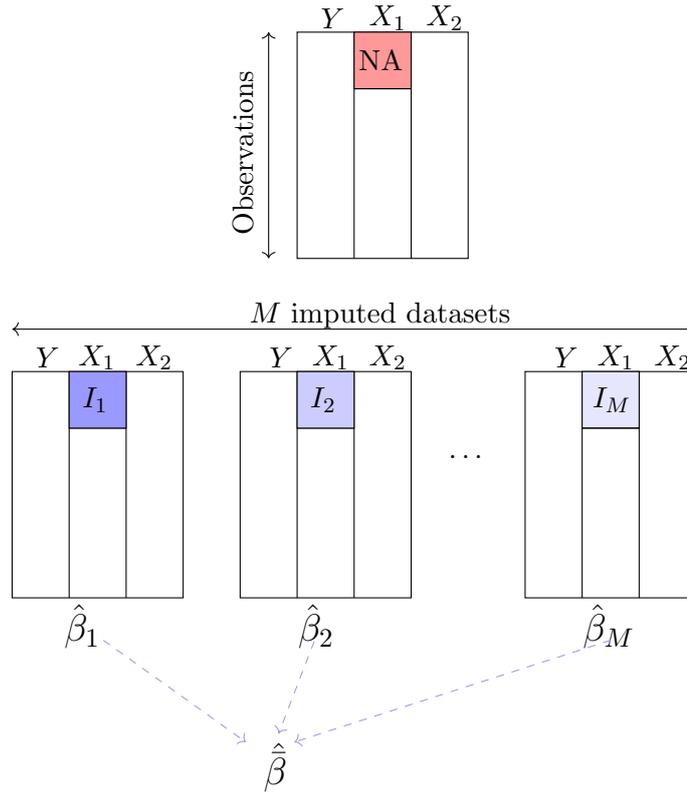


Figure 1.2: Diagram explaining the multiple imputation procedure. NA denotes missing values and I represents the imputed values which replace NA.

A popular MI method is joint modelling which involves drawing values for missing data simultaneously from a multivariate distribution, typically multivariate normal. Another MI method is *full conditional specification* (FCS), which is also known as *multivariate imputation by chained equations* (MICE). MICE imputes missing data by cycling through a series of univariate conditional models for each covariate with missing values conditional on other covariates and the outcome. When cycling through these conditional models, the most up-to-date values for the missing covariates not currently being imputed are used (this algorithm is clearly detailed in [18, Section 4.1]). While joint modelling assumes a multivariate distribution for all covariates, MICE allows each covariate to have its own individual distribution for imputing which is beneficial when dealing with missing values in continuous, binary and categorical covariates.

1.6.3 Congeniality in MI

Xie and Meng (2017) [19] discusses congeniality which is an important concept for the validity of MI. When applying MI, there are two models which need to be specified:

1. The analysis model:

This is the model of interest (also known as the substantive model). An example would be a generalized linear model regressing an outcome on several covariates, with parameters β : $g(Y | X_1, X_2, X_3, \dots; \beta)$

2. The imputation model:

This is the model used to impute the missing values in a covariate. If one covariate X_1 is partially-observed, this regresses the covariate with the missing values, X_1 , against all other relevant covariates in the dataset. For example: $f(X_1 | Y, X_2, X_3, \dots; \psi)$

For Rubin’s rules to hold these models must be congenial which implies there must exist a joint model which has conditional models which corresponds to the analysis and imputation models. If this is not the case the analysis model may lead to biased parameter estimates. Auxiliary variables not included in the analysis model could be included in the imputation model to improve efficiency of MI but would cause an uncongenial “richer” model [12, p.64] and lead to bias in the Rubin’s rules variance estimator. Alternatively, removing variables from the imputation model which are in the analysis model leads to a “poorer” model which invalidates both Rubin’s rules parameter and variance estimates.

I will now give a brief example of uncongeniality. As discussed by Bartlett et al. (2015) [18], uncongeniality can be seen in incorrect modelling of the functional form of a continuous covariate. A covariate containing missing values, X_1 , has a quadratic association with the outcome, Y , such that $Y | X_1, X_1^2$ is normal with mean a function of X_1 and X_1^2 . An imputation model could introduce bias if it assumes that X_1 is conditionally normal given Y with mean a linear function of Y . This is because the two models cannot simultaneously hold, i.e. they are uncongenial. The imputed data will only reflect a linear relationship with the outcome whereas the observed are associated quadratically.

1.6.4 Imputation of covariates in Cox regression

In the description of MI above, I have focused on a generic outcome Y which could be continuous or binary. There are some special considerations needed for a time-to-event outcome. As the Rotterdam breast cancer dataset will be used to illustrate any future methods, the imputation of covariates in Cox regression will be briefly discussed.

Bartlett et al. [18] and White and Royston (2009) [20] have proposed methods for handling missing covariate values in the case of the analysis model being a Cox model. White and Royston suggested an approximately valid imputation model which should contain the event indicator, other covariates X and the Nelson-Aalen estimate of the cumulative hazard, $\hat{H}_0(t)$, at the person’s observed event or censoring time. This is an uncongenial approach unless parameters β from the analysis model and the imputation parameters are equal to zero.

In comparison, Bartlett et al. use a modified MICE approach which accounts for the analysis model in the imputation process in order to make the analysis and imputation model congenial. This approach can be used in Cox, linear and logistic regression and can

accommodate transformations of continuous covariates in the analysis model.

While the approach from Bartlett et al. is considered to be the gold standard MI approach (because full Bayes or likelihood are superior but harder to actually do) [9], it is not, as yet, able to handle the selection of fractional polynomials (described in Section 1.7), which are relevant to later sections of the thesis. The ‘simpler’ approaches involving multiple imputation and MICE are able to handle fractional polynomials [21]. As such, when imputing the Rotterdam data in Chapter 10, the method from White and Royston will be used.

1.7 Flexible transformation of continuous covariates

There are many ways to handle continuous covariates in an analysis, some of which are not recommended. One example is to categorise a continuous covariate using ‘cutpoints’, for example those with a covariate value between 0 and 10 are group 1, those with values between 10 and 20 are group 2 etc. However, categorising covariates is generally not recommended due to a loss of information, the analysis results can change depending on the cutpoints used [22] and can cause ‘jumps’ in predicted values which are ‘unnatural’ [23, p.178-180].

A covariate is commonly included into a model as a linear term [24]. However, transforming covariates using non-linear functions can improve the fit of a prediction model [23, p.180-184]. Splines can be used to flexibly model covariates by increasing the degree of the piecewise polynomials, the level of flexibility can be controlled by either the number of knots (used to section off the data) or by changing the allocated number of degrees of freedom. An alternative to splines for flexible covariate transformations are fractional polynomials (FPs) developed by Royston and Sauerbrei [11] which I will focus on in this thesis.

1.7.1 Fractional Polynomials

FPs are a method for flexibly modelling nonlinear effects of a continuous variable X_1 where $X_1 > 0$ for all observations. FPs of degree 1 (FP1) are of the form X_1^E for $E \in S$. S is a set of powers such that $S = (-2, -1, -\frac{1}{2}, 0, \frac{1}{2}, 1, 2, 3)$ where 0 represents a log transformation. While this is the common set of fractional polynomials in the literature, other values can also be considered. Due to set S including a logarithmic transformation and negative exponents, there is a requirement that the values of a covariate X_1 should be greater than zero ($X_1 \in \mathbb{R}_{>0}^+$). A FP of degree 2 (FP2) applied to a covariate, X_1 , is of the form

$$X_1^{\mathbf{E}} = X_1^{(E_1, E_2)} = \begin{cases} (X_1^{E_1}, X_1^{E_2}), & E_1 \neq E_2 \\ (X_1^{E_1}, X_1^{E_1} \log(X_1)), & E_1 = E_2 \end{cases} \quad (1.1)$$

A FP of degree D in linear regression for patient i is of the form:

$$\mathbb{E}[Y_i | \mathbf{X}_i] = \beta_0 + \sum_{p=1}^P \beta_p X_{i,p}^{\mathbf{E}} \quad (1.2)$$

Logistic regression uses a logit function to link the linear predictor of covariates to the probability of having an outcome. The logit link function of a value a is $\log\left(\frac{a}{1-a}\right)$. A FP of degree D in logistic regression for patient i is of the form:

$$\text{logit}(P(Y_i = 1 | \mathbf{X}_i)) = \beta_0 + \sum_{p=1}^P \beta_p X_{i,p}^{\mathbf{E}} \quad (1.3)$$

A FP of degree D in a Cox model for patient i is of the form

$$h(t_i | \mathbf{X}_i) = h_0(t) \exp\left(\sum_{p=1}^P \beta_p X_{i,p}^{\mathbf{E}}\right) \quad (1.4)$$

where exponent \mathbf{E} has dimension D and P is the number of covariates included in the model.

One way to select among FP models (e.g. FP1 versus FP2) is by comparing the difference in deviances to a chi-squared distribution. The appropriate degrees of freedom (DF) for comparing models can be found in Table 1.1.

Table 1.1: The degrees of freedom (DF) for comparing the difference in deviances in fractional polynomial (FP) models. The parameters highlighted in blue indicate those related to the degrees of freedom.

| Model A | Model B | DF | DF explanation |
|---------|---------|----|--|
| FP2 | Null | 4 | Null: α_0 FP2: $\beta_0 + \beta_1 X_1^{E_1} + \beta_2 X_1^{E_2}$ |
| FP2 | Linear | 3 | Lin: $\alpha_0 + \alpha_1 X_1$ FP2: $\beta_0 + \beta_1 X_1^{E_1} + \beta_2 X_1^{E_2}$ |
| FP2 | FP1 | 2 | FP1: $\alpha_0 + \alpha_1 X_1^{Q_1}$ FP2: $\beta_0 + \beta_1 X_1^{E_1} + \beta_2 X_1^{E_2}$ |
| FP1 | Null | 2 | Null: α_0 FP1: $\beta_0 + \beta_1 X_1^{E_1}$ |

FP algorithm: Selecting a FP for one covariate

The FP selection (FPS) algorithm below states the selection procedure for choosing an appropriate FP for one covariate. The default function is linear.

1. Assume a linear function for covariate X_1 in the regression model for Y .
2. Choose appropriate level of Type I error (α_E) and appropriate maximum degree D , typically $D = 2$.
3. Test the best FP2 model for X_1 at the α_E level against the null model. If the test yields a non-significant p-value (at the chosen α_E level), stop, otherwise continue. This is equivalent to testing for an association with X_1 .
4. Test the best FP2 model against a linear model. If the test yields a non-significant p-value, stop, otherwise continue. This is equivalent to testing for non-linearity.
5. Test the best FP2 for X_1 against the best FP1. If the test yields a non-significant p-value, the final model is FP1, otherwise FP2.

This algorithm can be extended to handle multiple covariates in the multivariable FP (MFP) algorithm [11, p.117-118]. This iteratively cycles through all covariates to be considered for selection into the model and also determines whether any continuous covariates should be transformed. The algorithm is available in Appendix A.

Combining FPs and MI

Typically MI assumes that the analysis model is fixed and already known but use of FPs involves model selection via the MFP algorithm. Moreover, FPs require that any imputed values be positive. Morris et al. (2015) [21] propose imputation methods for use when covariates are transformed using FPs and the MFP model selection procedure is to be applied. This involves either approximate Bayesian bootstrap or a rejection sampling approach. Note that Morris et al. focused on combining MI with fractional polynomials of degree 1, there is no satisfactory method for a degree greater than 1. Combining FPs with MI will be covered in more detail in Chapter 7 where I shall use the work conducted by Morris et al. to extend my proposed methods (which combine MI and internal validation) to handle covariate selection and assessment of the functional form of continuous covariates.

1.8 Prediction Models

A prediction model aims to use various characteristics about a patient in order to predict whether they have (or will have) an outcome of interest. A model can be written as some function of covariates for patient i as previously seen in equations 1.2-1.4.

In addition to missing data, there are various issues to consider when developing a prediction model [7]. These include, but are not limited to, ensuring the sample size is sufficient [25,26,27,28], selecting covariates into the final model and deciding on the functional form of continuous covariates [24,29].

There are various ways to include covariates into a model. Covariate selection could be conducted *a priori* by adding a pre-identified set of covariates into a prediction model. There are other more traditional methods such as stepwise regression which iteratively evaluates the contribution of a covariate to the model of interest. Forward selection starts with an empty model and then includes the covariate which is most significant (based on a predefined significance level). It proceeds to add in the most significant covariates until either all covariates are included or the inclusion of any of the remaining non-selected covariates to the model does not improve the ‘model fit’. Backwards selection is similar but reversed. The model starts with all potential covariates included and then evaluates whether the removal of the least significant covariate badly impacts the model. Stepwise methods can be straightforward to use but the selection process can have several disadvantages [23, p.213] such as covariate selection instability and it can also lead to worse internally and externally validated performance than if a full model including all covariates had been used.

Consideration of these various issues can help to prevent the overfitting of a prediction model to the data it has been trained in. A model which suffers from overfitting lacks generalisability and will not perform well when predicting outcomes for previously unseen data. Shrinkage methods such as Lasso regression are intended to help to address overfitting [30] and can be used when developing a prediction model.

The development of a final prediction model is not the focus of this thesis. However, the procedure used to develop a prediction model must also be accounted for in the validation process. Therefore, some of the issues that surround model-building, such as covariate selection and the transformation of continuous covariates, will feature in Chapters 7-10.

1.9 Introduction to internal validation of prediction models

When developing a prediction model it is essential to know how well the model will perform when used to predict the outcome for a new individual. This could be an assessment of how well the model performs in the data it was fitted to (internal validation). Alternatively, validation could include assessing how well the model performs in a population which is slightly different from that used to fit the prediction model or from a different time period (external validation).

Internal validation involves evaluating a prediction model using the same data used to fit it. Internal validation is typically conducted at the initial development stage to assess the validity of the model in the setting it was trained in. External validation is usually conducted afterwards to assess how generalisable a prediction model is. There are several ways to internally validate a model. In this thesis, I shall briefly overview apparent performance and splitting the data into a training and test set. I shall then detail the cross-validation and optimism-corrected bootstrap algorithms which will be used frequently throughout Chapters 2-10. These methods are available in more detail in [23, Chapter 17]. All methods detailed in this section will be for the scenario where data are fully-observed.

1.9.1 Apparent performance

A simple approach to model validation is one in which the prediction model is evaluated using the same data which was used for model development. This is the simplest form of model validation but will lead to a model performance estimate which is over-optimistic, particularly in small samples. This is due to all of the data having been used when fitting the model, it is therefore trained to give good predictions specifically for the dataset. The performance estimated when evaluating a prediction model using all of the data which trained it is known as the *apparent performance*. This will be relevant in Section 1.9.4 when discussing the optimism-corrected bootstrap algorithms.

1.9.2 Split the dataset into a training and test set

The split sample approach involves randomly splitting a dataset into two sub-datasets. An example is demonstrated in Figure 1.3 where a two third training set versus one third test set split has been used. The observations in the training set are used to fit a prediction model. The observations in the test set are used to evaluate how well the prediction model (fitted to the training set) will perform on ‘unseen’ data.

The advantage of splitting the data into a training and test set is that it is very easy to do. However, there are several disadvantages. By splitting the data into two, the overall amount of data available to train a prediction model or to evaluate it is reduced. The

trained prediction model is losing out on valuable information and may be at increased risk of overfitting to the data due to a smaller sample size i.e. the model will be less generalisable or robust to new data. In addition, the trained prediction model is entirely dependent on the way the data has been split into the training and test sets. A different choice of split could produce a different and either better or worse performing prediction model i.e. the results could be unstable and highly variable.

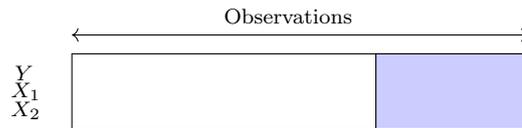


Figure 1.3: An example of splitting data into a training set (white) and a test set (purple).

As the training and test split is easy to visualise and understand, it will be used in Chapter 2 for illustrative purposes when discussing the imputation of data for the training and evaluation of prediction models. It also serves as a good introduction step to cross-validation which repeatedly splits the data into training and test sets.

1.9.3 Cross-validation (CV) methods

Cross-validation can be used to internally validate the predictive performance of a clinical prediction model. It can be thought of as repetitive splits of the data into training and test sets. The dataset is split into K folds. Figure 1.4 demonstrates cross-validation in the case where three folds are used ($K = 3$).

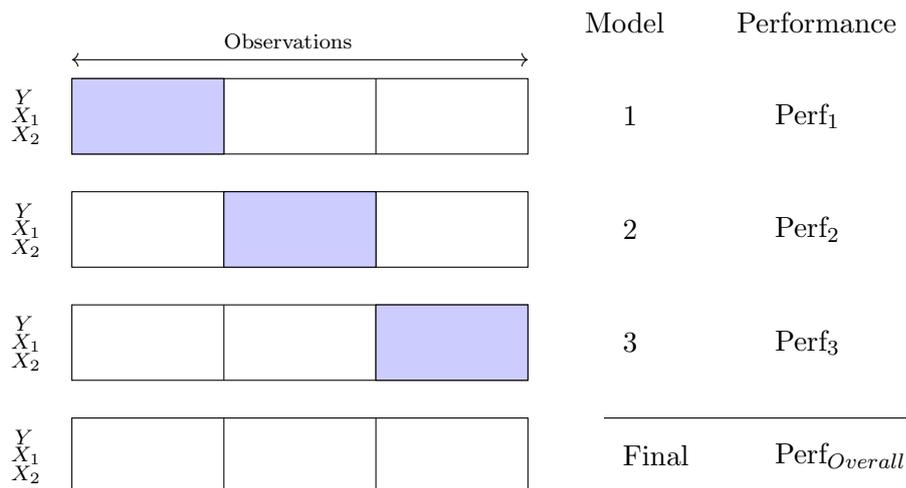


Figure 1.4: An example of cross-validation for $K = 3$. Each fold is iteratively used as a test set (purple) to evaluate each model formed in the training set (white). The three estimates of performance Perf _{k} will be averaged to get an overall estimate of performance for the final prediction model trained using all the data.

A training model is fit to the data which excludes the k^{th} fold and the model's performance will be tested in the excluded fold. This process is repeated while iterating through

the k folds for $k = 1, \dots, K$. In this way, there will be K estimates of the predictive performance (Perf_k) of the K fitted models. An example of a performance measure is the mean-squared error which will be introduced in Section 1.10. These K estimates of performance are combined by taking the mean to get an overall estimate of performance $\text{Perf}_{Overall} = \sum_{k=1}^K \frac{\text{Perf}_k}{K}$.

10-fold cross-validation ($K = 10$) is an improvement on validation using one training and test split as all the data are used to develop a final prediction model and it has lower Monte Carlo error. Then, when applying cross-validation the process used to develop the prediction model is repeated iteratively in 90% of the data (ensuring a more stable prediction model for evaluation) and tested on the remaining 10%. This produces less biased results while ensuring their variability is both reduced and less dependent on the choice of split. The method is more computationally intensive than a simple training and test split as K prediction models must be fitted for validation, in addition to the development of a final prediction model for use. However, repeating the cross-validation process (repeated cross-validation) may be necessary to obtain ‘stable’ estimates of performance [23, p. 333].

1.9.4 Bootstrap (BS) methods

The bootstrap is another method that is commonly used to internally validate the predictive performance of a clinical prediction model. A bootstrap sample involves randomly sampling observations (with replacement) from a dataset. The number of observations in each bootstrap sample $b = 1, \dots, B$ is the same as in the dataset, but the bootstrap sample may have repeated observations. The basic idea is to sample with replacement from the dataset to train and evaluate a model. In doing so, it is expected that 63.2% of observations are represented at least once in a bootstrap sample [31, p.253]. Here, we will describe several versions of using the bootstrap for internal validation when there are no missing data.

The out-of-bag (OOB) bootstrap

A simple method of using the bootstrap to validate a prediction model is to use the OOB method. This involves taking a bootstrap sample and fitting a prediction model to it. This ‘bootstrap prediction model’ is then evaluated in the observations which were not sampled in the bootstrap sample. The procedure is as follows:

1. Take a bootstrap sample, b , with replacement from the original data
2. Train a prediction model in bootstrap sample b using the same analysis procedure as intended on the original data
3. Evaluate the prediction model in those individuals who were not sampled for the bootstrap sample (out-of-bag)

4. Repeat steps 1-3 B times and average the estimates of predictive performance to get an overall performance measure.

This is also known as bootstrap cross-validation and provides a lower bound for the true predictive performance [32]. Less information is available in the bootstrap sample due to, on average, 63.2% of observations being sampled in a bootstrap sample. This means that the method underestimates performance compared to all observations being used in apparent performance [32].

The *standard* and 0.632 optimism-corrected bootstrap algorithms

There are several variations of using the bootstrap for correcting the optimism in the apparent performance estimate [33]. Here I detail two: the default method, which will be known as the *standard* method; and the *0.632* variation.

Evaluating a prediction model P in a dataset \mathbf{D} for a particular performance measure is noted as $\text{perf}(P, \mathbf{D})$. For example, $\text{perf}(P, \mathbf{D})$ could be the estimated MSE value from evaluating prediction model P in dataset \mathbf{D} . In the following algorithm a dataset is denoted as \mathbf{D}_d where d represents either the full dataset (o) or the bootstrap sample (b) and a prediction model developed in either the full dataset or bootstrap sample will be labelled P_d for $d = o, b$. The *standard* bootstrap method uses the following steps:

1. Train a prediction model P_o on the full dataset. Evaluate the performance of P_o in the original data to estimate the *apparent performance*, $\text{perf}(P_o, \mathbf{D}_o)$.
2. Take a bootstrap sample b from the original data. Train a prediction model P_b in the bootstrap sample b .
3. Evaluate the performance of P_b in bootstrap sample b to estimate the *bootstrap performance* ($\text{perf}(P_b, \mathbf{D}_b)$).
4. Evaluate the performance of P_b in the original data to estimate the *test performance* ($\text{perf}(P_b, \mathbf{D}_o)$).
5. The *bootstrap performance* (of the prediction model fitted to bootstrap sample b) is then compared to the *test performance*. This produces an estimate of optimism for the *bootstrap performance*. This is estimated as:

$$\text{Optimism}_b = \text{bootstrap performance}_b - \text{test performance}_b$$

6. Steps (2)-(5) are repeated for $b = 1, \dots, B$.

7. The optimism-corrected performance (OCP) is then calculated as

$$\text{OCP} = \text{Apparent performance} - \frac{1}{B} \sum_{b=1}^B \text{Optimism}_b$$

Step 1 results in an over-optimistic performance estimate as a prediction model is trained and evaluated in the same dataset. To adjust for this, the *standard* bootstrap algorithm is utilised to estimate how optimistic the apparent performance is.

The bootstrap performance can be thought of as the apparent performance in the bootstrap sample as it too is trained and evaluated in the same data (in this step, the data is the bootstrap sample). To gauge how optimistic the bootstrap apparent performance is, the model is then evaluated in the original dataset, which contains observations not sampled in the bootstrap sample. The performance of the bootstrap model in this larger dataset with new observations can be compared to the bootstrap performance to estimate optimism. This attempts to replicate the scenario of evaluating P_o in unobserved data, if that option was available.

A variation of this validation method is the ‘0.632’ algorithm which is similar to the OOB bootstrap. Figure 1.5 displays the key differences between the standard and 0.632 algorithms. The 0.632 algorithm differs from the *standard* bootstrap by evaluating the bootstrap prediction model P_b only in those observations which were not selected in the bootstrap sample in order to estimate the test performance. This means that Step 3 is ignored and Step 4 is modified to only evaluate P_b in those not bootstrap sampled. The OCP of the 0.632 method is $(0.368 \times \text{Apparent performance}) + (0.632 \times \text{mean}(\text{Test performance}))$ [31, p.253].

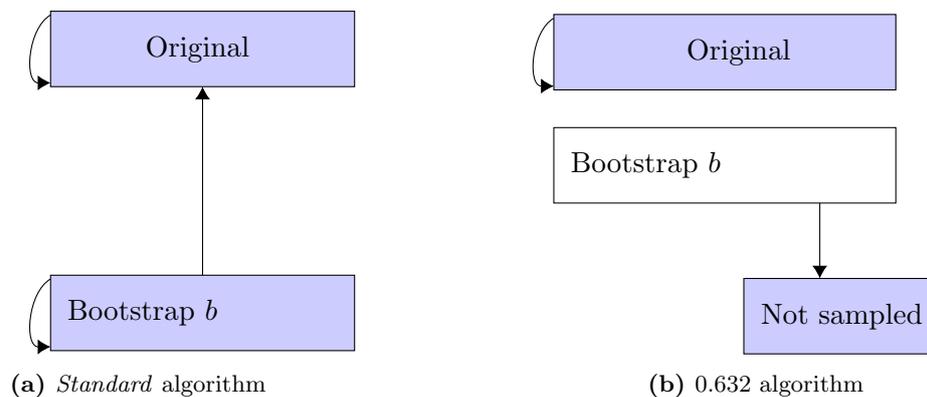


Figure 1.5: The difference between the *standard* and 0.632 algorithms for one bootstrap sample b . The solid lined arrow starts where the prediction model was developed and points to the data where it is evaluated. Datasets coloured in purple are used for the evaluation of a prediction model.

The 0.632 method is similar to the OOB bootstrap as both methods involve evaluating the ‘bootstrap prediction model’ in those observations which were not selected into the bootstrap sample. However, the 0.632 avoids the drawbacks of the OOB (underestimating the true performance) as it uses a weighted average of the apparent performance (which uses all observations) and the test performance.

1.10 Performance measures

When evaluating how well a model performs using validation methods, we require some measure by which to evaluate the model's performance. We can typically segregate performance measures into three categories: 'overall measures' (such as the mean-squared error or Brier score [23]), discriminative ability and calibration.

An overall measure of performance uses the distance between two points to determine how well a model performs, for example how close the predicted outcome is to the observed outcome. It is a measure of performance regardless of whether the outcome is continuous or binary. Examples of overall performance measures include the mean-squared error when the outcome is continuous and the Brier score when the outcome is binary. These are discussed in more detail below.

The discriminatory ability of a model measures whether a prediction model can differentiate between the different levels of an outcome, for example can the prediction model differentiate between high and low risk patients for a disease. An example of a discriminatory performance measure is the area under the receiver operating characteristic curve, which is equivalent to the c -statistic when the outcome is binary.

Finally, calibration assesses the agreement between predicted values and the outcomes which were observed. There are several levels of calibration but in this thesis, I will focus solely on weak calibration. The reasoning for this will be explained in Section 1.10.4.

In this section, I will give a brief summary of the four performance measures which will be used in the thesis to assess model performance. These measures are well documented and commonly used in practice, a more detailed description can be found in [23, Chapter 15].

1.10.1 Continuous outcome: Mean-squared error (MSE)

The MSE or Mean-squared prediction error (MSPE) can be used to summarise the predictive ability of a model when the outcome is continuous. It is an overall measure of model performance. The MSE measures how closely a model's predicted values are to the observed outcome, as seen in equation (1.5), and its values are non-negative ($\text{MSE} \in \mathbb{R}^+$). A lower value of the MSE indicates that the predicted values (\hat{y}_i) are close to the values that were observed (y_i) i.e. the model has good predictive ability. The MSE is therefore defined as:

$$\text{MSE} = \frac{1}{N_{obs}} \sum_{i=1}^{N_{obs}} (\hat{y}_i - y_i)^2 \quad (1.5)$$

N_{obs} is the number of records used to evaluate the prediction model. In the case of estimating apparent performance, N_{obs} would be the total number of records in the dataset on whom \hat{y}_i can be produced. If using a training and test split, N_{obs} would be the number of records in the test set used to evaluate the prediction model.

1.10.2 Binary outcome: Brier Score

The Brier score is parallel to the MSE when the outcome is binary. It compares the predicted probability of the outcome happening, \hat{p}_i , to what was observed, y_i using a quadratic loss function:

$$\text{Brier} = \frac{1}{N_{obs}} \sum_{i=1}^{N_{obs}} (\hat{p}_i - y_i)^2 \quad (1.6)$$

The value of the Brier score estimate can range between 0 and 1. For example, if an outcome occurs for a single observation ($y_i = 1$) and the model is poor giving a predicted probability of 0.2, the Brier score estimate would be $(0.2 - 1)^2 = 0.64$, whereas a better model may have a predicted probability of 0.8 which gives a Brier score estimate of $(0.8 - 1)^2 = 0.04$. A lower Brier score value indicates a better performing model.

1.10.3 Binary outcome: Area under the curve (AUC)

The AUC estimate assesses a model's discriminative ability i.e. how well it can differentiate between those who have or do not have the outcome. The AUC refers to the area under a receiver operating characteristic (ROC) curve. An ROC curve plots the probability of being predicted to have the outcome given you have the outcome (sensitivity/true positive rate) against the probability of being predicted to have the outcome given you do not have the outcome (1-specificity or the false positive rate), an example can be seen in Figure 1.6.

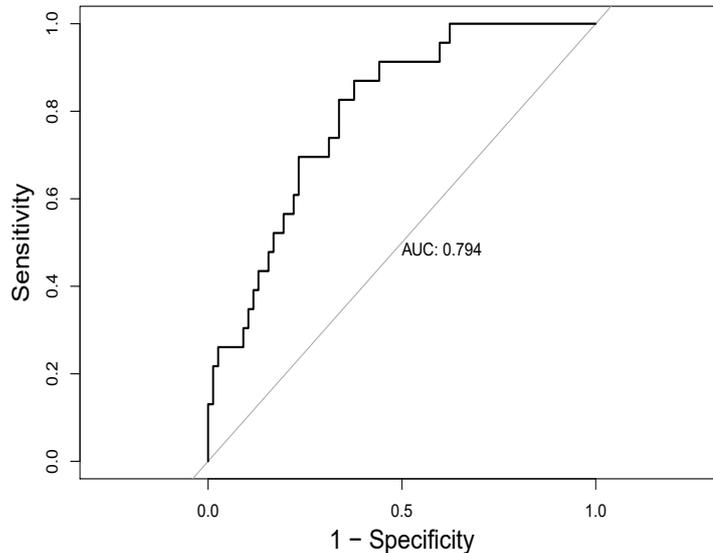


Figure 1.6: An example of a ROC curve (black line).

A well-performing model will have a curve tending to the top left corner of the ROC curve i.e. it has high sensitivity and high specificity. This equates to a higher value of the area under the curve which can range between 0 and 1. A value of 0.5 typically indicates a model which does not perform any better than chance, while a higher value indicates the model is doing better than a random guess. When the outcome is binary, the AUC is equivalent to the concordance statistic.

1.10.4 Binary outcome: Calibration

A well calibrated model should provide reliable predicted risks which correspond to the observed proportions of the event [34] i.e. if the predicted risk of an outcome is $r\%$, then we expect $r\%$ of observations to have the outcome at a group-level. There are several levels of calibration: mean, weak, moderate and strong. The moderate and strong levels assess calibration using plots. This would make assessment in multiply imputed datasets, within cross-validation folds or bootstrap samples within multiple repetitions of simulated data difficult. Here, I focus on weak calibration as it evaluates deviations in the slope and intercept i.e. quantifiable estimates.

Weak calibration requires that there is no systematic under- or overfitting or under- or overestimation of the risks [23]. This is evaluated by assessing deviation of the slope ζ away from 1. The logistic regression prediction model is applied to the observations which will be used to evaluate the prediction model (such as those in a test set). Each observation i will have a predicted probability \hat{p}_i of having the outcome. A logistic regression model is then used to compare the predicted probabilities with the outcome which was observed:

$$\text{logit}(y=1) = \gamma + \zeta * \hat{p} \quad (1.7)$$

A value of $\zeta < 1$ indicates overfitting, which means that the linear predictor has a tendency to give extreme values i.e. high risks are overestimated while low risks are underestimated [34].

Deviations in the intercept α are compared to zero by constraining $\zeta = 1$ i.e. the intercept is estimated using an offset [23, p.300]:

$$\text{logit}(y=1) = \alpha + \text{offset}(\hat{p}) \quad (1.8)$$

A value of $\alpha < 0$ implies that the predicted risks from a model are on average overestimated and a value of α greater than zero implies that the predicted risks are underestimated.

While calibration could also be assessed when the outcome is continuous, it is not a very popular measure for linear regression and was therefore not considered for a continuous outcome in this thesis.

1.11 Data leakage in prediction models

There are many considerations which must be taken into account when developing a prediction model. As outlined in Section 1.8 some of these include considerations such as sample size, covariate selection or how to handle missing data.

Another key consideration for the development of prediction models is that of data leakage. This concerns any prediction model which has an unfair advantage due to having access to the unseen data it will be evaluated in. To explain this concept further, I consider when it could arise in the situation of developing and evaluating a prediction model using the split sample method (Section 1.9.2). If the researcher trains any parameters or hyperparameters using just the training set, then no leakage has occurred. However, if the researcher had used all of the data to tune a hyperparameter, and then used this pre-estimated hyperparameter value when fitting a prediction model in the training set, this would cause data leakage. In terms of performance, data leakage can result in optimistic estimates of performance. Even optimism correction algorithms, such as the *standard* or 0.632 bootstrap (Section 1.9.4), are at risk of having optimistic estimates of optimism.

A very simple example of data leakage with a training and test split can be described using the k -means clustering algorithm applied to a simple simulated dataset. The k -means clustering method classifies a new observation to a cluster based on its proximity to the centre of a cluster and the number of clusters k must be estimated. Looking at all available data in Figure 1.7 one might state that $k = 3$. However, if the number of clusters was estimated based on the training data $k = 2$ seems plausible. By using all of the data to estimate the number of clusters k , information about the observations which were not sampled for the training data has been leaked. Therefore a k -means model with three clusters will do far better than the model with 2 clusters as would have been selected if only the training data had been looked at.

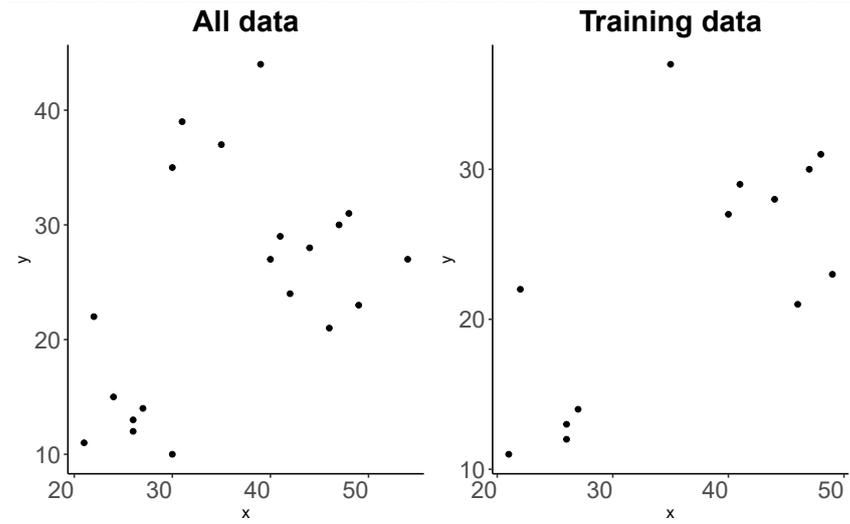


Figure 1.7: Simple example of data leakage: k -means

Another example could be using all available data to estimate the penalisation parameter, λ , in lasso regression [35] and using this estimate for the lasso regression model whose covariate coefficients were estimated using the training data only.

Data leakage will be further discussed in Section 2.8 by introducing the concept of data leakage through the imputation of missing data.

1.12 Outline of thesis

An outline of the remainder of the thesis is as follows:

- Chapter 2 reviews the current literature on combining MI with cross-validation and introduces key concepts that will be used throughout the thesis. I will propose methods for combining MI with cross-validation or with the bootstrap algorithms and discuss the concept of data leakage through the imputation process.
- Chapter 3 details the set-up of an extensive simulation study for data with a continuous or binary outcome. The aim of the simulation study will be to assess the proposed methods from Chapter 2 which combine MI with either cross-validation or the bootstrap optimism-corrected algorithm. The simulated data will explore multiple factors such as different missing data mechanisms, the influence of sample size and the effects of increasing the proportion of missingness. As the results are extensive, they are presented in the three following chapters.
- Chapters 4 and 5 present results from the simulation study described in Chapter 3. The results in these chapters focus on the proposed methods which combine MI with cross-validation.

- Chapter6 present results from the simulation study described in Chapter3. The results in these chapters focus on the proposed methods which combine MI with the bootstrap optimism-corrected algorithms.
- Chapter7 introduces the use of fractional polynomials for covariate selection and the choice of functional form when combining MI and internal validation algorithms.
- Chapter8 details the set-up of a simulation study for data with a continuous outcome. The aim of the study is to evaluate the methods proposed in Chapter7. These methods combine MI and internal validation, while allowing for covariate selection and the flexible transformation of continuous covariates using fractional polynomials.
- Chapter9 presents the simulated data results for combining internal validation, fractional polynomials and MI.
- Chapter10 demonstrates the methods for combining MI and internal validation which are considered to have the best properties (based on the simulation investigations) in the Rotterdam breast cancer dataset.
- Chapter11 details a systematic review which investigates the handling of missing data in observational time-to-event analyses. This review was published in the BMC Medical Research Methodology journal.
- Chapter12 contains a discussion of the methods and results presented in the thesis and proposes potential extensions to the work that has been conducted to-date.

Appendices and a supplementary file containing plots are available at the end of the thesis.

2 Internal validation when missing data are present in covariates

In this chapter I will discuss the problem that missing data present when developing and evaluating prediction models. Current literature in the area will be discussed, as will the intricacies of combining MI with validation techniques. Finally, I will propose several methods for combining MI with either cross-validation or the bootstrap optimism-corrected algorithms.

2.1 Introduction

Missing data complicate the development of clinical prediction models, determining their performance, and their use to obtain predictions for new patients. There are three stages to consider:

Stage 1: Handling missing data when training a clinical prediction model

Stage 2: Handling missing data while validating the prediction model

Stage 3: Handling missing data in new patients when applying the prediction model

MI is a common method used to handle missing data and may be a viable approach to handling missing data in the prediction setting. Stage 1 has been well-researched, including how to incorporate variable selection, functional forms of continuous covariates and other considerations of model development when multivariable model-building. A more exhaustive list of recommendations can be found in Table 4 of Chapter 11 but some examples include: Wood et al. (2008) [36] who investigated combining MI with variable selection, Morris et al. (2015) [21] who incorporated MI with fractional polynomials to handle variable selection and choice of functional forms of continuous covariates or Seaman et al. (2012) [37] who investigated covariate transformations when using MI.

The bootstrap and cross-validation algorithms are commonly used to internally validate the predictive performance of a clinical prediction model. When data are fully-observed the algorithms are implemented, as outlined in section 1.9. In this chapter, I will focus on the second stage and start by introducing two estimand-like measures which are important and used regularly throughout this thesis.

2.2 Pragmatic and Ideal performance

Wood, Royston and White (2015) [38] investigated the use of MI when using a prediction model to estimate predicted values and evaluate the model's performance. Two estimand-like measures were detailed in this paper. These measures are defined in terms of the practical context in which the predictions would be used for future individuals:

- **Ideal model performance** focuses on an ideal clinical setting where all individuals have fully-observed predictors
- **Pragmatic model performance** is based in a real-world clinical setting where some individuals have missing predictor values

Essentially, missing data for ideal model performance can be thought of as a feature that happens to occur in the data the prediction model will be developed in, but would not occur in future use i.e. in the data on which new predictions are to be made. When data is observed for a new patient it will be fully-observed and the outcome can be predicted with no need for MI.

Pragmatic performance is relevant when missing data are present in both the development data and the data to which the model is to be applied (i.e. it is expected that any future patient data will also have unobserved values in predictors). The handling of missing data in new individuals for whom predictions are to be obtained must be considered in addition to missing data in the dataset used for model development and evaluation. An example could involve using a complete-case analysis as the method to handle missing data. A prediction model is coded into a software programme and will throw up an error if an ‘input’ value for a patient is missing. In this case, predictions could only be made for those who are fully-observed (complete-case).

The phrasing of *ideal* and *pragmatic* is arguably not helpful and gives a negative connotation towards the scenario where future patients are not fully-observed. One scenario is not ‘lesser’ simply because it is more inconvenient. However, this is the terminology that has been chosen by Wood et al. and I will use it throughout this thesis.

2.2.1 Imputation models to assess ideal or pragmatic performance

Wood et al. focused on estimating the apparent performance of a prediction model. When assessing ideal performance, Wood et al. [38] recommended imputing the entire dataset M times using an imputation model which includes the outcome. This maintains the association between the imputed values and the outcome of interest. These M imputed datasets can each be used to train a prediction model and obtain predicted values which can be used to estimate apparent performance (Stage 1 and Stage 2). Apparent performance was previously introduced in Section 1.9.

For assessing pragmatic performance, Wood et al. state that basing predictions on imputed data derived using the observed outcome may bias the predictions for a new individual, because in practice the outcome would not be available for the MI process (as it would not yet be observed for new patients i.e. Stage 3). As such, they recommend imputing the

entire dataset of interest using two imputation models. The first model will include the outcome Y and will allow for the association between the outcome and missing covariate to be maintained for model development (Stage 1). The second imputation model will exclude the outcome and the imputed datasets using this model will be used to evaluate the trained prediction model to estimate the apparent performance (Stage 2).

2.3 A summary of the current published literature

A description of two potential ways to combine imputation methods with validation algorithms

Before commencing a summary of the published literature, I will first define two ways which can be considered to combine imputation methods such as MI with validation methods. The first is *MI-then-Validate*. This involves multiply imputing the entire dataset first and then applying the validation method to each imputed dataset. The second is *Validate-then-MI*, which involves multiply imputing within the validation algorithm. For example, *CV-then-MI* would involve first splitting the entire dataset, which is partially-observed, into K folds. Within one iteration using fold k as the test fold and the other $k - 1$ folds as the training set, multiple imputation would be applied within the training and test folds. That is, within the test fold (k), an imputation model would be fitted and used to obtain M imputed versions of the test fold. This is repeated within the training set. For a bootstrap method (*BS-then-MI*), it would involve taking a bootstrap sample of the partially-observed data and then applying MI to the bootstrap sample.

Literature outlining ideal and pragmatic performance

Wood et al. (2015) [38] was discussed in Section 2.2. They investigated using MI for estimating predictions and model evaluation. They defined ideal and pragmatic model performance and focused on how best to evaluate a prediction model when using apparent performance (Section 1.9) as a validation method. For ideal performance, they recommended multiply imputing the dataset (including an outcome in the imputation model) using one set of M imputed datasets to train and evaluate a prediction model. For pragmatic performance, they recommended using two sets of imputed datasets. The first set would be imputed including the outcome in the imputation model. These M imputed datasets would be used to fit M prediction models. The entire dataset would be multiply imputed a second time, this time with the outcome excluded from the imputation model, to produce M imputed datasets in which to evaluate the prediction models.

Literature recommending pooling performance measures instead of predicted values

Wood et al. primarily focused on when to apply Rubin's rules when evaluating a prediction model using multiple imputed datasets. Option 1 involved using the M prediction models to get M estimates of predicted values for each individual in the dataset. These M

predicted values could then be pooled using Rubin’s first rule to get a single predicted value for each individual. These predicted values could then be used to estimate a measure of performance. Option 2 involved using the prediction model fitted in the m^{th} imputed data set to get predicted values for the individuals in the entire dataset (as they were using apparent performance). These predicted values could then be used to estimate the performance measure of interest. This was then repeated across the M imputed datasets, for $m = 1, \dots, M$, resulting in a total of M estimates of performance which could then be pooled using Rubin’s first rule to get an overall estimate of performance. Wood et al. showed that option 2 was preferred as option 1 tended to over-estimate the performance of the MSE.

Literature recommending using one pooled prediction model for future use

In addition to options 1 and 2, Wood et al. made a comparison between using Rubin’s rules to collapse the M prediction models into one overall prediction model versus keeping the M prediction models separate. However, they do not state how the predicted values should be used if the prediction models are kept separate i.e. do they recommend pooling the predicted values from the M prediction models for a future individual or keeping the M predictions separate. Differences between predicted values were observed based on whether an overall model or the M models were used and Wood et al. stated that an overall model could give imprecise estimates of model performance. They state that pooling the prediction model can be used in practice as it is unlikely that researchers will keep prediction models unpooled. Similar work conducted by Miles (2015) [39] compared whether to use one overall model to produce predicted values or to keep the M prediction models separate and apply Rubin’s rules to the predicted values instead. Miles concluded that both methods perform similarly but that using one final overall model is faster and easier to implement.

Vergouwe et al. (2010) [40] illustrated how to develop and evaluate a prediction model using a temporal external validation dataset when missing data are present in both the development data and the external validation data. An overall prediction model was estimated by applying Rubin’s first rule to the M prediction models fitted to the M imputed datasets of the ‘training’ data. This overall model was then applied to the M imputed datasets of the external dataset. The performance of the overall model was estimated in each imputed external dataset and Rubin’s rules were then applied to get an averaged estimate of performance, as recommended by Wood et al. Vergouwe et al. gave no justification for (i) estimating an overall prediction model instead of keeping the M prediction models separate or (ii) pooling the performance measure estimates estimated in the external dataset instead of pooling the predicted values.

Literature recommending MI-then-Validate

While Wood et al. (2015) did not describe it as such, their main focus was on the apparent performance setting. They discussed several potential methods for handling training and test sets which I have summarised in Table 2.1. They stated that their method 1 (which I term the *MI-then-Validate* approach) may be the most appropriate when using internal validation algorithms, such as cross-validation, but also mentioned that their method 2 (a *Validate-then-MI* approach) could be used.

Table 2.1: Proposed methods by Wood, Royston and White for handling missing data in training and test sets. For methods 1 and 2, the outcome will be included in the imputation model if ideal performance is of interest.

| Proposed Method | Estimand |
|--|--------------------|
| 1 Impute the dataset and then split into a training and test set. | Ideal or Pragmatic |
| 2 Take the imputation model used to impute the training set and apply it to the test set | Ideal or Pragmatic |
| 3 Impute the training and test sets separately including outcome | Ideal |
| 4 Impute the training and test sets separately - training set can include outcome, test set will exclude | Pragmatic |

Steyerberg and Vergouwe (2014) [41] outline steps for the development and validation of prediction models. Their first step for developing a prediction model involves inspecting the available data. If missing data are present they recommend imputing missing values in this step 1 and validating the developed prediction model, using cross-validation or bootstrap resampling, in a later step. No justification for this approach is given in the paper. Their approach corresponds to *MI-then-Validate*. Similar recommendations are given by Steyerberg (2019) [23, p.335-336]. There, Steyerberg also considered the use of bootstrap in combination with MI. His focus is on the ideal performance setting and, citing Wood et al., his recommendation is to first apply MI to the complete cohort and then obtain bootstrap samples from these imputed datasets (i.e. *MI-then-Validate*).

Jaeger et al. (2020) [42] focused on the use of cross-validation in unsupervised learning. Their recommendations are to impute first before applying cross-validation as they state that this reduces variance in model error estimates. This method is discussed and critiqued in section 2.8. While they remark on data leakage (this concept was introduced in Section 1.11) being a concern while modelling they do not appear to have connected this with the use of imputation methods. As such, their recommended method used all available data to impute missing values. They used an unsupervised imputation approach called *k*-nearest-neighbour imputation (details of this method are available in the Jaeger et al.

paper). They compared imputing first, followed by cross-validation, with implementing cross-validation first, followed by imputing. Their version of *CV-then-impute* involved imputing the $k - 1$ training folds together. The same model used to impute the $k - 1$ folds is then used to impute the k^{th} test fold. This is then repeated K times, with each fold iteratively being used as the test fold. Jaeger et al. evaluated their methods using simulated data. Their outcome was continuous, the performance measure of interest was the root-MSE ($\text{MSE}^{0.5}$ i.e. this is a study-level performance estimate rather than a simulation performance measure) and they generated datasets of sample size 100, 500, 1,000 and 5,000.

Literature recommending Validate-then-MI

Musoro et al. (2014) [43] focused specifically on combining MI with, what I will term, the *standard* optimism-corrected bootstrap algorithm (Section 1.9.4). They stated that all available data, including the outcome, were used when multiply imputing (i.e. fitting the imputation model and drawing imputed values). I therefore suggest that they were implicitly estimating ideal model performance. Both *Validate-then-MI* and *MI-then-Validate* were considered and the *apparent performance* was estimated in the same manner for each combination, as follows. The entire dataset was imputed M times and a prediction model fitted to each imputed dataset. One overall prediction model was then estimated using Rubin’s rules. This overall prediction model was then evaluated in the same M imputed datasets. The M estimates of performance were then averaged to estimate the apparent performance. Three versions of *MI-then-Validate* were considered, each version involved applying the bootstrap algorithm to M imputed datasets. Version (i) involved applying the bootstrap sampling procedure each time to the M imputed datasets i.e. obtaining bootstrap samples separately in each of the M imputed datasets, so that the samples in each imputed dataset are different. Version (ii) involved using the same set of B bootstrap samples in each imputed dataset i.e. the bootstrap samples are fixed. Version (iii) applied the bootstrap sampling procedure to one imputed dataset (i.e. $M = 1$). In order to multiply impute within the bootstrap procedure (*Validate-then-MI*), a bootstrap sample is multiply imputed M times and a prediction model is fitted to each imputed bootstrap sample. These M prediction models are pooled to get one overall prediction model for the bootstrap sample (i.e. one overall prediction model per bootstrap sample). This overall model is then used to estimate the bootstrap performance in the M imputed datasets of the bootstrap sample. Rubin’s rules are applied to get an average estimate, and this is referred to as the *bootstrap performance*. The overall model is next applied to each of the M imputed datasets (containing all observations) to get M estimates of the *test performance*, and these are then averaged using Rubin’s rules.

Musoro et al. evaluated their proposed methods using a simulation study with a continuous outcome. One thousand simulated datasets were generated for a number of data-

generating scenarios. Methods were assessed when the sample size was 500 or 1000 and the MSE (Section 1.10) was used as the performance measure of interest. Performance estimates from the *Validate-then-MI* and *MI-then-Validate* methods were then compared to estimates from a simulated ‘external validation set’ with no missing data. Version (i) of *MI-then-Validate* (allowing the bootstrap samples to vary across the imputed datasets) was found to underestimate the estimate of optimism i.e. *bootstrap performance-test performance* was smaller than expected. Musoro et al. conclude that multiply imputing within the *standard* algorithm is recommended (*Validate-then-MI*).

Wahl et al. (2016) [44] investigated *Validate-then-MI* and *MI-then-Validate* methods to combine multiple imputation with several internal validation algorithms. These internal validation methods were bootstrapping (*standard*, 0.632, 0.632₊), *K*-fold cross-validation and subsampling. For *MI-then-Validate* they imputed the entire dataset, using one set of *M* imputed datasets, and then applied the validation algorithm. *MI-then-Validate* was considered for both ideal and pragmatic performance due to the inclusion and exclusion of the outcome in the imputation model. For *Validate-then-MI*, the outcome was included in the imputation models fitted to the ‘training’ and ‘test’ sets (i.e. this is an assessment of ideal performance). Specific details on how each internal validation was combined with MI for the *MI-then-Validate* methods were difficult to ascertain as the methods were presented in a diagram. I will summarise my interpretation of their *Validate-then-MI* method using the 0.632 bootstrap algorithm as an example. A bootstrap sample is taken from the entire partially-observed dataset. This bootstrap sample is then multiply imputed *M* times (including the outcome in the imputation model fitted to the bootstrap sample). The observations which were not selected into the bootstrap sample are also imputed *M* times using a separately fitted imputation model (the outcome is included in the imputation model fitted to the not-selected observations). A prediction model is fitted to the *m*th imputed dataset within a bootstrap sample. This prediction model is then evaluated in the *m*th not-selected imputed dataset. This is repeated *M* times to get *M* estimates of performance for the bootstrap sample. These *M* estimates are then averaged using Rubin’s first rule to get an overall estimate of performance. This is then repeated for the *B* bootstrap samples.

Wahl et al. used simulation studies to evaluate the various methods. These focused on datasets with a binary outcome and multiple factors were varied, such as sample size (100, 200, 500 and 1000) and the number of covariates included in the prediction model (1, 5, 10, 20). Two hundred and fifty simulated datasets were generated to assess performance in each data-generating scenario. The AUC, Brier score and calibration intercept and slope were used as performance measures. Wahl et al. concluded that the *Validate-then-MI* methods are preferred as they typically provide ‘unbiased estimates’. In addition, Wahl et al. found that increasing the number of imputed datasets beyond 5 had little effect on

the performance estimates but increasing the number of bootstrap samples from 10 to 50 improved accuracy of the estimates.

Mertens, Banzato and de Wreede (2020) [45] focused on combining cross-validation with MI using a *Validate-then-MI* method. Three variations of *Validate-then-MI* were considered. For version (i), they propose first assigning observations randomly to K folds. Within one iteration of cross-validation, fold k is used as a test fold and the remaining $k - 1$ folds are used as a training set. The observed outcomes in the test fold are temporarily set as missing. All K folds are then imputed together using multiple imputation (the imputation model includes the outcome) with $M = 1$ (i.e. one imputed dataset is used). I therefore suggest that they were implicitly estimating pragmatic model performance. The prediction model fitted to the $k - 1$ imputed training folds is then used to get predicted values for the imputed k^{th} fold. This is repeated for $k = 1, \dots, K$ until every observation has a predicted value. This entire process is then repeated, with the observations assigned to K different folds and imputed in a similar manner. Overall, each observation will have K predicted values and these are then averaged. These averaged values for each observation in the dataset can then be used to estimate performance.

Version (ii) is similar to version (i) except $M > 1$ and each observation is only assigned to a fold once (i.e. there is no repeating of the entire process as there is in version (i)). Observations are randomly assigned to K folds. For one iteration of cross-validation where the k^{th} fold is used as the test set, again the outcomes in the test fold are temporarily set as missing and the entire dataset is imputed together (with the outcome included in the imputation model). The M imputed datasets are then split into M imputed $k - 1$ training folds and M imputed test folds. Prediction models are fitted in the M training folds and Rubin's rules are used to estimate one overall prediction model. This overall prediction model is used to get predicted values in the M imputed test folds. This is then repeated for $k = 1, \dots, K$ so that each observation will have M predicted values which are then averaged. The averaged predicted values for each observation in the dataset can then be used to get a performance measure of interest. Version (iii) is similar to version (ii) as the M imputed test folds are collapsed into one imputed test fold by averaging the M imputed values for each observation which had missing values.

Mertens et al. evaluated their three methods on two 'real' data sets of sample size 524 (153 deaths and 38 censored records) and 694 (184 deaths and 46 censored records). They focused on a time-to-event outcome and the Brier score was used as a measure of accuracy. An additional simulation study was used to simulate a binary outcome. The simulation study used a sample size of 1,000 and considered an increasing number of imputed datasets $M = 1, 10, 100$. Overall, Mertens et al. found version (i) to be the preferred approach, recommending that averaging the M predicted values is preferred instead of using an

overall prediction model (as used in versions (ii) and (iii)).

Literature concerning stage 3: handling a new/out-of-sample individual

Both Fletcher and Blume (2018) [46] and Hoogland et al. (2020) [47] focus on the third stage of applying the prediction model in a missing data context i.e. how to obtain a predicted value for a new patient with missing observations. This stage is not the main focus of this thesis. Fletcher and Blume use a specific submodel for each missing data pattern. They note that their pattern submodel approach performed well and is easy to use in practice. When data are MCAR or MAR, they state that submodels and MI will have similar predictive accuracy. Hoogland et al. compares submodels based on observed data, marginalization over the missing variables and MI (using MICE). They found that using submodels or MI by fixed chained equations performed well, based on the C-statistic performance measure, when obtaining predictions for individual new patients.

Summary of this literature review

There is a wide array of literature in this field which often provides conflicting advice. It is possible to find literature stating that either *MI-then-Validate* or *Validate-then-MI* is the preferred way to combine MI and internal validation algorithms. In addition, much of the published literature has focused on the ideal performance estimand. Recommendations for the pragmatic performance of *Validate-then-MI* involve the removal of the outcome from the list of variables which should be included in the imputation model.

In addition, there are many ways to combine cross-validation with MI or the bootstrap methods with MI. One can average M prediction models (fitted within imputed training datasets) to get one overall prediction model, which can estimate predicted values in M test sets. Alternatively one can evaluate a prediction model fitted in the m^{th} imputed training set and evaluate it in the m^{th} imputed test set, repeating this for imputed training and test datasets $m = 1, \dots, M$. Recommendations from Wood et al. (2015) state that it is preferable to pool the performance estimates, rather than predictions. Whereas, when combining MI and cross-validation Mertens et al. (2020) recommend pooling predicted values and then using those to estimate a performance measure estimate.

Miles (2015) recommended pooling M prediction models to get one overall prediction model which can be used to estimate predicted values. Wood et al. (2015) note that using a pooled prediction model may give imprecise estimates of model performance, especially if the regression parameters are very different across the M imputed datasets but argue that, it is unlikely unpooled prediction models would be used in practice.

The aim of this chapter is not only to explore pragmatic performance methods, but also to explore combining cross-validation or the bootstrap with MI in more detail than has been

considered in the current literature, while taking into account the recommendation from Wood et al. to pool performance estimates rather than predicted values. In addition, I provide principled justification to determine the best way to incorporate these methods when the ideal or pragmatic setting is of interest.

2.4 A short note on pooling prediction models versus keeping prediction models unpooled when internally validating

Within this chapter I will propose methods for combining internal validation methods (cross-validation and the optimism-corrected bootstrap algorithms) with MI. Due to the nature of MI, there will be M prediction models used to estimate predicted values when applying internal validation. There are three potential ways to use these M prediction models. Option (i) involves pooling the M prediction models to get one ‘overall’ prediction model. This prediction model can then be used to get one predicted value for each new individual. Options (ii) and (iii) involves keeping the M prediction models unpooled and estimating M predicted values for each individual used to evaluate the prediction models. Option (ii) will pool these predicted values to get one overall predicted value for each patient. Option (iii) involves keeping these predicted values unpooled, estimating the performance for each prediction model and then pooling the performance measure.

For all of the methods proposed in this section I will proceed with option (iii). This involves keeping the prediction models unpooled for one bootstrap sample or within one iteration of cross-validation (for example, M predicted values per individual per fold). As stated by Wood et al. [38], pooling predicted values (option (ii)) can over-estimate performance. Wood et al. also state that a pooled prediction model (option (i)) can give imprecise estimates of model performance if the regression parameters are very different across imputed datasets. In addition, all methods proposed in this chapter will be extended to handle covariate selection and transformations of continuous covariates when fitting a prediction model in Chapter 7. Not only could regression parameters potentially be very different across imputed datasets, there is also the possibility that included covariates or the functional form of continuous covariates could vary across imputed datasets. Pooling prediction models that have selected different covariates for inclusion or which have different transformations for continuous covariates is possible [21,36] but an unnecessary layer of added difficulty for a researcher in practice when prediction models can just as easily be kept unpooled.

2.5 Using separate imputation models to impute the training and test sets

In this section, I will propose the use of two imputation models regardless of whether ideal or pragmatic performance is of interest. For explanatory purposes, I will focus on a simple

setting where a dataset will be split into a training set (to fit a prediction model) and a test set (to evaluate the prediction model). The dataset will contain an outcome, Y , a partially-observed covariate X_1 and a fully-observed covariate X_2 .

A lot of the literature to date has focused on using one set of imputed datasets to train and evaluate prediction models. For example, Wahl et al. and many others who assessed the *MI-then-validate* approach ([44,43,23]) produced one set of imputed datasets for which to train and evaluate a prediction model.

I argue that two separate imputation models should be fitted, which would in turn produce two sets of imputed datasets, regardless of whether your target is the ideal or pragmatic estimand. One of these imputation models (the *training imputation model*) and the imputed datasets it produces should be used for training a prediction model. The other imputation model (the *test imputation model*) and the imputed datasets it produces can be used to evaluate the prediction models. Two imputation models and two sets of imputed datasets should be used even if both imputation models contains the same covariates. If using one set of imputed datasets which are then split into a training and test set, the imputed training and test sets are correlated due to the same imputation model parameters having been used for the MI process, leading to optimism that would not be detected. More specifically when using one set of imputed datasets for ideal performance, the imputed test sets are correlated with the observed training set records due to the inclusion of the training set's outcome in the MI process. Similarly, the imputed training sets are correlated with the observed test set records. This correlation will be further discussed in section 2.8.

Whether focusing on ideal or pragmatic performance I suggest the use of two separate sets of imputed datasets, one set specifically for estimating the coefficients of the prediction model and the other set used for evaluating models.

2.5.1 A simple scenario with a single training and test split

To set up ideas, I will work through a very simple hypothetical description in the following sections. In the simplest validation procedure the original data $\{Y, X_1, X_2\}$ can be split randomly into a training and test set as seen in Figure 2.1. Y and X_2 are both fully-observed while X_1 contains missing values. The training and test set will be imputed separately using a training and test imputation model. There will be M_{train} imputed datasets of the training set and in each one, a prediction model will be fitted. The test set will be imputed M_{test} times and used to evaluate each of the M_{train} prediction models.

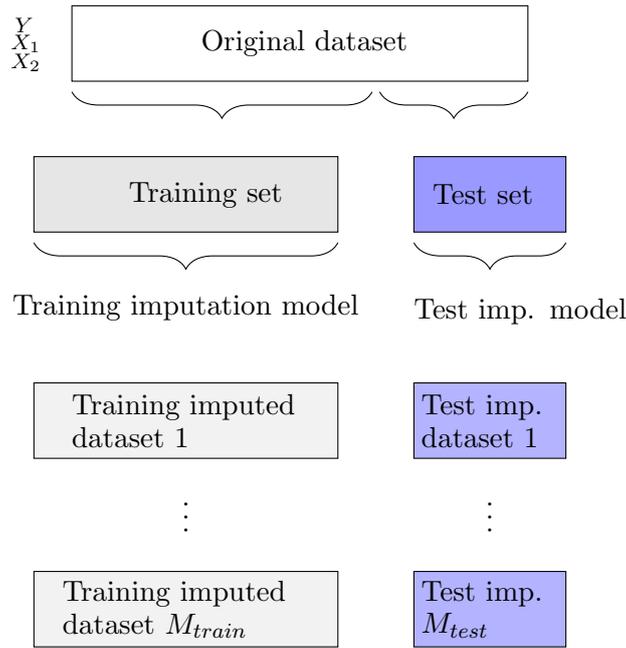


Figure 2.1: An example of splitting data into a training set (white) and a test set (purple). The training set is multiply imputed M_{train} times, fitting the imputation model only to the observations in the training set. This is repeated for the test set.

2.5.2 Relating the training and test imputation models to ideal and pragmatic performance

In Figure 2.1, a training and test imputation model was used to impute the training and test sets. In practice, when implementing either the ideal or pragmatic setting, the training imputation model should include the outcome and any relevant covariates which will improve the quality of the imputed values. Inclusion of the outcome will maintain the association between it and the values of the variable being imputed in the training set. However, multiple imputation of the test set will vary depending on whether ideal or pragmatic performance is of interest. For ideal and pragmatic performance, the training imputation model will both include the same relevant covariates (X_2) and the outcome (Y). It is the test imputation model which is different depending on the estimand. Table 2.2 describes test imputation models for ideal or pragmatic performance in a training and test split scenario (Figure 2.1), where X_1 is a covariate with missing values, X_2 is a fully-observed covariate and Y is the outcome.

To estimate ideal performance, the outcome Y can be included in the test imputation model for imputing the test set in order to maintain the association between the imputed values and the outcome. This replicates the scenario of a new patient being fully-observed i.e. there will be no missing values in X_1 and therefore the underlying true data-generating process relating X_1 to Y will be maintained, under the assumption of missing-at-random.

Table 2.2: Imputing a covariate with missing data, X_1 , in the test set under an ideal or pragmatic performance estimand. X_2 and Y are a fully-observed additional covariate and outcome, respectively.

| Estimand | Method | Test imputation model | Comments |
|-----------|--------|--|---|
| Ideal | | $X_1 = f(Y + X_2; \psi)$ | Using observed values of Y when fitting imputation model. |
| Pragmatic | A | $Y = f(X_1 + X_2; \psi_Y)$ $X_1 = f(Y + X_2; \psi)$ | Values of Y in test set are set as missing. Both Y and X_1 are imputed. Discard the imputed values of Y . |
| | B | $X_1 = f(X_2; \psi)$ | Y removed completely from test imputation model. |

$f(\mathbf{A}; \psi)$ denotes the imputation model; \mathbf{A} represents the covariates to be included in the imputation model and model parameters ψ

When estimating pragmatic performance, future patients are expected to have missing values in X_1 under the same mechanism as in the current data used to train the model. While it is possible to include the outcome Y in the test imputation model when estimating pragmatic performance (Pragmatic A) the actual observed values of Y in the test set should not be included, since these would not be available to impute X_1 in a practical context. In these ‘mock future patients’, the outcome Y will not be known in practice and therefore cannot be included to impute the missing values of X_1 . Instead, Y can be treated as if it were missing and imputed in the MI process. These imputed Y values can then be used to impute X_1 . The imputed values of Y will then be discarded. The imputed values of X_1 , alongside the fully-observed X_2 , can then be used to predict the outcome Y . The intuitive justification for Pragmatic A is that the association between the imputed values of X_1 and Y is maintained while also ensuring that X_1 is not using the observed values of Y as these would not be available for future patients. However, this approach may be difficult to implement in some situations. For example, in Figure 2.1 if we wished to multiply impute the test set independently of the training set using Pragmatic A, all observed values for Y would be temporarily ‘set’ to be missing and would have to be imputed. However, now that the entire outcome is ‘missing’ in the test set the question then arises on how Y can be imputed when there are no observed Y for the MI process.

One possible solution would be to take all of the original data, set the Y values for the patients selected to be in the test set as missing and then impute Y and X_1 as outlined in Table 2.2 method Pragmatic A. After imputing, the imputed dataset is restricted to those patients which are in the test set. However, in implementing this solution it is ensured that the test imputed datasets are correlated with the training set as the available outcome values from the training set are used.

Alternatively, Pragmatic B ensures complete independence of the training and test datasets. The outcome Y can be removed completely from the test imputation model and the other relevant covariates can be used to impute the missing values to estimate pragmatic performance (Pragmatic B). In the example in Figure 2.1, X_2 can be used to impute X_1 . While this model does not maintain the association between the missing values in the covariate and the outcome in the test set, it may be the only feasible model available for certain validation procedures. However, the training imputation model when estimating pragmatic performance should include the outcome, even if the outcome will be excluded in the test imputation model, in order to avoid bias to the parameters of the prediction model.

While one could argue that in a prediction setting we are less interested in the bias of model parameters, these parameters are used to estimate predicted values for new observations. Therefore, bias in the parameters will influence the predicted values. An alternative approximately unbiased single imputation method for the pragmatic setting could be to impute a missing value using the mean of X_1 conditioned on other covariates (i.e. a complete-case analysis estimate of $\mathbb{E}[X_1 | X_2; \psi]$) [48]. However, this would only produce unbiased estimates of β under MCAR.

In this section I have detailed the use of imputation models in relation to ideal or pragmatic performance. For the description, I have used a simpler validation technique which involves splitting a dataset into a training and test set (Figure 2.1). This serves as a good introduction to the application of imputation models to cross-validation which allocates data to K folds and then splits the observations into a training set (folds 1 to $k - 1$) and a test set (fold k), iteratively using each fold as a test set.

In the following sections I will detail methods which will either:

- multiply impute the data first, followed by applying the validation algorithm to each of the multiply imputed datasets. This will be known as *MI-then-Validate* where validate could be cross-validation or bootstrapping
- apply the validation algorithm first, followed by MI. For example, split the data into K folds and then split the data into a $k - 1$ training set and k^{th} fold test set, or take a bootstrap sample of the dataset. Then the $k - 1$ training set, the k fold test set or the bootstrap sample can be multiply imputed. This will be known as *Validate-then-MI*

2.6 Proposed methods for Cross-validation

When combining cross-validation with MI the order in which each algorithm should be implemented must be considered. This accounts for whether to multiply impute first and apply cross-validation to each imputed dataset or whether to cross-validate first (apply the K -fold splitting first) and then multiply impute the folds. Table 2.3 details several possible methods for combining the two when cross-validating first followed by MI in a pragmatic setting. While I use the MSE as the performance measure of interest to describe the methods, the methods are the same for other performance measures, for example: the AUC or Brier score.

The methods in Table 2.3 can be classified according to how to impute the training folds within one iteration of cross-validation. Impute training folds separately (Method A), impute the $k - 1$ training folds together (Methods B, C, F, G, I) or impute all folds together and then exclude fold k (Methods D, E, H).

Table 2.3: Methods for combining cross-validation and MI for a pragmatic scenario when a covariate is partially-observed. Cross-validation will be applied first (i.e. observations will be randomly assigned to K folds), followed by MI.

| Pragmatic performance | | Method |
|---|--|--------|
| Training | Test | |
| <p>1. Separately in each fold $k = 1, \dots, K$, fit a training imputation model which includes Y and produces M_{train} imputed datasets for each fold i.e. impute each fold separately.</p> <p>2. For the folds to be used for training (folds_{-k}) combine together their imputed datasets to make a training set of $K - 1$ folds for each $m_{train} = 1, \dots, M_{train}$ - this will produce M_{train} training imputed datasets.</p> <p>3. Fit a prediction model to each of the M_{train} training imputed datasets, to get models $P_{m_{train}}$. Keep the models unpooled i.e. do not use Rubin's rules to get a final model averaged over the imputed datasets.</p> | <p>1. In the k^{th} fold fit a test imputation model excluding Y but still including other covariates. This will produce M_{test} imputed datasets for the k^{th} fold.</p> <p>2. For $m_{test} = 1, \dots, M_{test}$, evaluate prediction model $P_{m_{train}}$ in the M_{test} test imputed datasets of fold k to get M_{test} estimates of the MSE. Use Rubin's first rule to average the M_{test} estimates of the MSE. This will produce an overall estimate of MSE for $P_{m_{train}}$.</p> <p>3. Repeat Step2 for each $P_{m_{train}}$ for $m_{train} = 1, \dots, M_{train}$.</p> <p>4. Take the M_{train} overall MSE estimates from each prediction model from Step3 and use Rubin's first rule to get a final estimate of MSE across the M_{train} prediction models.</p> | A |

folds_{- k} denotes the $k - 1$ training folds which exclude the k^{th} test fold

Table 2.3: Methods for combining cross-validation and MI for a pragmatic scenario when cross-validating first (continued)

| Pragmatic performance | | Method |
|--|---|--------|
| Training | Test | |
| <p>1. Combine the folds to be used for training (folds_{-k}). Fit the training imputation model including Y and relevant covariates to folds_{-k} and produce M_{train} imputed training datasets</p> <p>Then apply Training step 3 from method A to produce M_{train} prediction models: $P_{m_{train}}$ for $m_{train} = 1, \dots, M_{train}$</p> | <p>Refer to Test steps for Method A</p> | B |
| | <p>1. In k^{th} fold fit a test imputation model excluding Y and using the other covariates from all K folds. This will produce M_{test} imputed datasets which should be restricted to the observations included in the k^{th} fold.</p> <p>Then apply Test steps 2-4 from method A</p> | C |
| <p>1. Fit the imputation model including Y to all folds and produce M_{train} training imputed datasets. Restrict the imputed datasets to the observations included in folds_{-k}</p> <p>Then apply Training step 3 from method A to produce M_{train} prediction models</p> | <p>Refer to Test steps for Method A</p> | D |
| | <p>Refer to Test steps for Method C</p> | E |
| <p>Refer to Training steps for method B</p> | <p>1. In each training imputed dataset, m_{train}, impute the k^{th} fold using relevant covariates from all folds (the imputed values in folds_{-k} taken as ‘true’ values) and excluding Y. Restrict test imputed datasets to observations in the k^{th} fold.</p> <p>Then apply Test Steps 2-4 from method A</p> | F |

folds_{-k} denotes the $k - 1$ training folds which exclude the k^{th} test fold

Table 2.3: Methods for combining cross-validation and MI for a pragmatic scenario when cross-validating first (continued)

| Pragmatic performance | | Method |
|--------------------------------------|---|--------|
| Training | Test | |
| Refer to Training steps for method B | <p>1. In the k^{th} fold, set Y as missing.</p> <p>2. Using all folds fit a test imputation model including the outcome and relevant covariates to impute the incomplete covariate and outcome Y. This will output M_{test} test imputed datasets - restrict these to the observations that should be in the k^{th} fold and discard the imputed Y values.</p> <p>Then implement Test steps 2-4 from Method A to get an overall summary performance measure</p> | G |
| Refer to Training steps for method D | Refer to Test steps for Method G | H |
| Refer to Training steps for method B | <p>1. In the k^{th} fold, set Y as missing.</p> <p>2. To impute the k^{th} fold, use the same imputation model as in Training steps to impute Y and X - discard imputed values of Y.</p> <p>Then implement Test steps 2-4 from Method A to get an overall summary performance measure</p> | I |

While the above methods were described for a pragmatic performance scenario they can be used to estimate ideal performance by including the outcome Y in the test imputation model. All methods can be adapted for ideal performance by including the outcome in the test imputation model, except for Methods G-I (due to the observed outcome for fold k being set to missing). To adapt method C and F for ideal performance, it is important to set Y to missing in the $k - 1$ training folds when fitting the test imputation model to the test fold and producing the test imputed datasets.

All methods in Table 2.3 will be evaluated in the pragmatic and, where relevant, ideal

scenario. Method I would be difficult to implement in practice as it involves using the training imputation model to impute the test set. Currently, it is not a feature of the `mice` package in R [49] to conduct out-of-sample imputation i.e. new data/the test set are unable to be imputed based on the imputation model from the training data. The MI imputation model parameters can be extracted in Stata for the multivariate normal command (command: `mi impute mvn`) but not for chained equations (command: `mi impute chained`). As noted by [47], the only package in R which outputs imputation model parameters is the Amelia package ([50]) which, similar to Stata (command: `mi impute mvn`), assumes multivariate normality. This may dissuade analysts from using Method I in practice and it was therefore not examined. To implement Method I when software does allow for the extraction of the imputation model in R it is important to include fully missing records to the $k - 1$ training folds so that the imputation model will be able to impute the outcome, as well as the missing covariate, in the test set.

Table 2.4 details two methods (J and K) when imputing first, followed by applying cross-validation to the imputed datasets (*MI-then-CV*). Method J involves imputing the dataset once overall, either with or without the outcome depending on whether the ideal or pragmatic setting is of interest as used in [43,44]. Method K involves fitting two imputation models to the dataset. The first model (a *training imputation model*) will include the outcome Y in order to fit training models in the $k - 1$ training folds. The second imputation model (a *test imputation model*) can either include or exclude the outcome for the imputed test datasets to evaluate each of the training models.

2.6.1 An additional consideration for MI

The amount of overall missingness in the dataset and reflecting this within fold assignment should be considered. For the methods described above, I chose to randomly sample individuals in a stratified manner to ensure that each fold had the same proportion of observed and missing values as in the original dataset. This maintains the same distribution of missingness in each fold and also has the added benefit of preventing a test fold from having a large proportion of (or entirely consisting of) missing data. This is particularly relevant when fitting an imputation model to a test fold which has a small number of observations. For example, the test fold contains 20 observations, of which 17 randomly contain missing values in one or several covariates.

Table 2.4: Methods for combining cross-validation and pragmatic imputation for an incomplete covariate when imputing first

| MI procedure | Method |
|--|--------|
| <ol style="list-style-type: none"> 1. Impute the original dataset, including relevant covariates in the imputation model for the pragmatic scenario or relevant covariates and the outcome for the ideal scenario. This will produce M imputed datasets. 2. In each imputed dataset, apply the cross-validation procedure to get an overall estimate of performance 3. Apply Rubin's rules to the M estimates from Step 2 to get an overall estimate of performance across the M imputed datasets | J |
| <ol style="list-style-type: none"> 1. Impute the original dataset, including relevant covariates and the outcome in the imputation model. This will produce M_{train} training imputed datasets. 2. Impute the original dataset, including relevant covariates (and the outcome if focusing on an ideal setting) in the imputation model. This will produce M_{test} test imputed datasets. 3. For training imputed dataset m_{train} train a model $P_{m_{train}}$ on the $k - 1$ folds (with fold k to be used as the test set). This training model is then evaluated using the k^{th} fold in each of the M_{test} test imputed datasets. Rubin's rules are applied to get an overall estimate of performance for model $P_{m_{train}}$ when fold k is used as the test set. 4. Repeat Step 3 for $k = 1, \dots, K$. Use cross-validation averaging rules to get an overall estimate of performance for training imputed dataset m_{train}. 5. Repeat steps 3 and 4 for imputed dataset $m_{train} = 1, \dots, M_{train}$ and apply Rubin's rules to get an overall estimate of performance. | K |

2.7 Proposed methods for the bootstrap algorithms

The Ideal and Pragmatic B methods from Table 2.2 will be examined for the bootstrap algorithm. Recall that Pragmatic A allowed for the inclusion of Y in the test imputation model while Pragmatic B excluded the outcome from the test imputation model. It would be possible to use method Pragmatic A for the *standard* algorithm to estimate the test performance as it is evaluated in both those who were or were not sampled for the bootstrap. Those who were not sampled could have their outcome set as missing. However,

it would not be possible to use Pragmatic A in the 0.632 alternative as the prediction model fitted to the bootstrap sample is evaluated in those observations not included in the bootstrap sample. Therefore, when setting Y to missing in this ‘test set’ Y would be fully unobserved and it would not be possible to impute it. Therefore, inclusion of Y in the test imputation model for pragmatic performance is not examined here.

In the following sections, I will detail how to combine MI and the bootstrap algorithms. The performance measure estimate used to evaluate a prediction model P in a dataset \mathbf{D} is denoted $\text{perf}(p, \mathbf{D})$.

2.7.1 BS-then-MI

The *standard* algorithm

For the *standard* bootstrap algorithm there are three measures of interest: Apparent, Bootstrap (Apparent performance in the Bootstrap sample) and Test. Recall, a training imputation model will include any relevant covariates and the outcome to impute missing values in a covariate. The test imputation model can include or exclude the outcome, depending on whether ideal or pragmatic performance is of interest. The algorithm is as follows:

1. The original sample (o) contains missing data. Impute it using both training and test imputation models to get the *Apparent* performance:
 - (a) Use the training imputation model to get M_{train} imputed datasets
 - (b) In each of these imputed datasets m_{train} , train the prediction model $P_{m_{train}}$.
 - (c) Impute the original sample again but this time using the test imputation model to get M_{test} imputed datasets ($t = 1, \dots, M_{test}$).
 - (d) For each $m_{train} = 1, \dots, M_{train}$ calculate the prediction model’s performance across the test imputed datasets, t , and use Rubin’s first rule to get an overall estimate of performance for each model:

$$\text{perf}(P_{m_{train}}, o) = \sum_{t=1}^{M_{test}} \frac{\text{perf}(P_{m_{train}}, t)}{M_{test}}$$

- (e) Use Rubin’s first rule to get an overall estimate of apparent performance

$$Apparent = \sum_{m=1}^{M_{train}} \frac{\text{perf}(P_{m_{train}}, o)}{M_{train}}$$

2. Sample from the original data with replacement to get a bootstrap sample b . Impute it using both training and test imputation models to get the *Bootstrap* performance:

- (a) Use the training imputation model to get M_{train} imputed datasets of bootstrap sample b .
- (b) For each imputed dataset, $m_{train}^* = 1, \dots, M_{train}$ of bootstrap sample b , train the prediction model $P_{m_{train}^*}$
- (c) Impute bootstrap sample b using the test imputation model to get M_{test} imputed datasets
- (d) For each $m_{train}^* = 1, \dots, M_{train}$ calculate the prediction model's performance across the bootstrap test imputed datasets, t^* , and use Rubin's first rule to get an overall estimate of the bootstrap performance for each model:

$$perf(P_{m_{train}^*}, b) = \sum_{t^*=1}^{M_{test}} \frac{perf(P_{m_{train}^*}, t^*)}{M_{test}}$$

- (e) Use Rubin's first rule to get an overall estimate of the bootstrap performance for bootstrap sample b

$$Bootstrap_b = \sum_{m_{train}^*=1}^{M_{train}} \frac{perf(P_{m_{train}^*}, b)}{M_{train}}$$

3. Finally, to get the *Test* performance:

- (a) Impute the original sample using the test imputation model to get a different set of M_{test} imputed datasets ($t' = 1, \dots, M_{test}$) to those in step (1c)
- (b) For each bootstrap prediction model from step (2b) $P_{m_{train}^*}$ for $m_{train}^* = 1, \dots, M_{train}$, calculate the performance across the test imputed datasets $t' = 1, \dots, M_{test}$ from step (3a) and use Rubin's first rule to get an overall estimate of the test performance for each bootstrap model:

$$perf(P_{m_{train}^*}, o) = \sum_{t'=1}^{M_{test}} \frac{perf(P_{m_{train}^*}, t')}{M_{test}}$$

- (c) Use Rubin's first rule to get an overall estimate of the test performance across the M_{train} prediction models

$$Test_b = \sum_{m_{train}^*=1}^{M_{train}} \frac{perf(P_{m_{train}^*}, o)}{M_{train}}$$

4. The optimism is then estimated as the difference between the bootstrap and test performance:

$$Optimism_b = Bootstrap_b - Test_b$$

5. Repeat steps (2) - (4) B times

6. To estimate the optimism-corrected performance,

$$OCP = Apparent - \frac{1}{B} \sum_{b=1}^B Optimism_b$$

The 0.632 algorithm

In this variation of the bootstrap algorithm only the apparent and test performance are of interest.

The 0.632 algorithm when bootstrapping first, followed by MI, has similar steps to the *standard* algorithm above but with some minor changes as noted below:

- The bootstrap performance is not calculated (because the 0.632 version is using a more classic sample splitting approach) so steps (2c) - (2e) are not necessary
- For step (3a) take those observations not sampled in the bootstrap sample from the original sample. Impute this subsample of not selected observations using the test imputation model to get $t' = 1, \dots, M_{test}$ imputed datasets.
- For step (6) the optimism-corrected performance becomes a weighted average (recall that on average, approximately 63.2% of observations in a bootstrap sample are unique) of the apparent performance and B test performances

$$OCP = (0.368 \times \textit{Apparent}) + \left(0.632 \times \frac{1}{B} \sum_{b=1}^B \textit{Test}_b \right)$$

2.7.2 MI-then-BS

The *standard* algorithm

1. Step 1 to get the *Apparent* performance is the same as for *BS-then-MI* in section 2.7.1. The original sample (o) contains missing data. Impute it using both training and test imputation models to get the *Apparent* performance:

- (a) Use the training imputation model to get M_{train} imputed datasets
- (b) In each of these imputed datasets m_{train} , train the prediction model $P_{m_{train}}$.
- (c) Impute the original sample again but this time using the test imputation model to get M_{test} imputed datasets ($t = 1, \dots, M_{test}$).
- (d) For each $m_{train} = 1, \dots, M_{train}$ calculate the prediction model's performance across the test imputed datasets, t , and use Rubin's first rule to get an overall estimate of performance for each model:

$$\text{perf}(P_{m_{train}}, o) = \sum_{t=1}^{M_{test}} \frac{\text{perf}(P_{m_{train}}, t)}{M_{test}}$$

- (e) Use Rubin's first rule to get an overall estimate of apparent performance

$$\textit{Apparent} = \sum_{m=1}^{M_{train}} \frac{\text{perf}(P_{m_{train}}, o)}{M_{train}}$$

2. Sample with replacement from training imputed dataset m_{train} with replacement to get a bootstrap sample b . In order to estimate the *bootstrap* performance for bootstrap sample b :

- (a) Train a prediction model $P_{train,b}^*$ in bootstrap sample b
- (b) Set the imputed missing values in bootstrap sample b back to missing. Impute these missing values using the test imputation mode to get M_{test} imputed datasets
- (c) Calculate the performance across the test imputed datasets $t^* = 1, \dots, M_{test}$

$$Bootstrap_{m,b} = \sum_{t^*=1}^{M_{test}} \frac{\text{perf}(P_{train,b}^*, t^*)}{M_{test}}$$

3. In order to get the *test* performance from the prediction model trained on bootstrap sample b :

- (a) Impute the original sample using the test imputation model to get a different set of M_{test} imputed datasets ($t' = 1, \dots, M_{test}$) as in step (3a) of *BS-then-MI*
- (b) For prediction model $P_{train,b}^*$ calculate the performance across the test imputed datasets $t' = 1, \dots, M_{test}$ from step (3a) and use Rubin's first rule to get an overall estimate of the test performance for each bootstrap model:

$$Test_{m,b} = \sum_{t'=1}^{M_{test}} \frac{\text{perf}(P_{train,b}^*, t')}{M_{test}}$$

4. Calculate the optimism for bootstrap b of imputed dataset m :

$$Optimism_{m,b} = Bootstrap_{m,b} - Test_{m,b}$$

5. Repeat steps (2)-(4) B times to get B estimates of optimism. To get the optimism-corrected performance for imputed dataset m :

$$OCP_m = Apparent - \frac{1}{B} \sum_{b=1}^B Optimism_{m,b}$$

6. Repeat steps (2) - (5) for $m_{train} = 1, \dots, M_{train}$ to get M_{train} estimates of the OCP.

7. Using Rubin's first rule, take the mean of the M_{train} estimates of the OCP to get an overall estimate.

The 0.632 algorithm

When there are no missing data present two measures of performance are calculated: Apparent and Test.

The 0.632 algorithm when imputing first, followed by bootstrapping follows a similar algorithm as the *standard* algorithm above but with some minor changes as noted below:

- The bootstrap performance is not calculated so steps (2b) - (2c) are not necessary
- For step (3a) take those observations not sampled in the bootstrap sample from the original sample. Impute this subsample of not selected observations using the test imputation model to get $t' = 1, \dots, M_{test}$ imputed datasets.
- Step (4) is no longer performed
- For step (5) the optimism-corrected performance becomes a weighted average of the apparent performance and B test performances

$$OCP = (0.368 \times \text{Apparent}) + \left(0.632 \times \frac{1}{B} \sum_{b=1}^B \text{Test}_b \right)$$

2.7.3 Other considerations

For both *MI-then-BS* and *BS-then-MI*, an additional consideration will be whether the training and test imputed datasets used to calculate the apparent performance can be reused to calculate the bootstrap or test performance. As combining bootstrapping with MI can be a computationally intensive procedure, especially for *BS-then-MI*, which multiply imputes $B \times M$ times, reusing imputed datasets will reduce the number of times it is necessary to impute and help with computational efficiency.

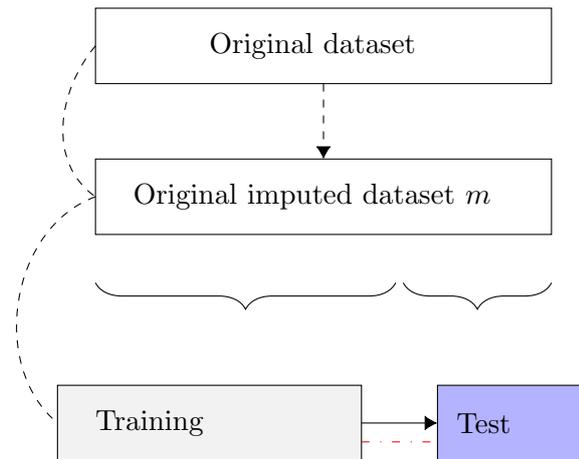
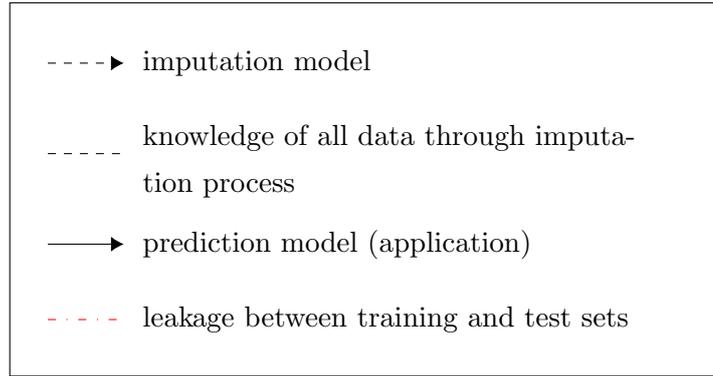
When conducting the bootstrap, to date, none of the current literature has made clear how they bootstrapped with respect to the missing data. I will use the stratified bootstrap method to ensure the same proportion of observed and missing is present in each bootstrap sample. The stratified bootstrap was implemented in order to avoid the possibility of sampling observations which all contained missing values in the covariates.

2.8 Data Leakage

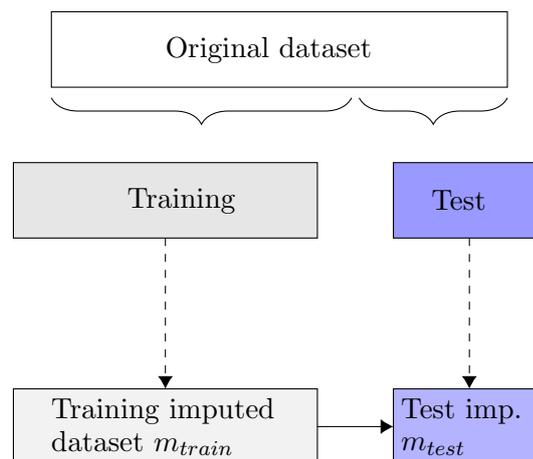
Data leakage is a prediction concept that needs to be carefully considered when estimating parameters of the prediction model. It can be a cause for a prediction model appearing to predict well during internal validation but when deployed for use in reality the model instead performs poorly. Thinking again in terms of a simple training and test split scenario, data leakage could occur due to information from the test set being ‘leaked’ to the prediction model being fitted in the training set. Therefore, the prediction model will perform well on the test set (of which it had prior knowledge of the observations) but when it encounters completely new observations (of which it has no prior knowledge on) it may perform badly. Data leakage has previously been detailed in Section 1.11 which also gives an intuitive example for how leakage can affect an analysis, demonstrated by the k -means classification method.

Without careful consideration, data leakage may also arise when imputing missing data as seen in Figure 2.2. When imputing first, all of the dataset is used to impute the missing values and as such these imputed values now contain information that came from fitting a model to the entire dataset. This association between training and test sets was previously discussed in section 2.5. When splitting the imputed dataset into a training and test set, observations in the test set are no longer completely independent of the data used to estimate the prediction model coefficients in the training set, and data leakage occurs.

This leakage does not occur in Figure 2.2(b) as the MI procedure takes place after splitting the data into a training and test set. The training imputed datasets only have access to the data in the training set during the MI process. Similarly, the test set does not gain any leakage from the training set as it is also independently imputed without including any observations from the training set.



(a) Imputing missing data followed by splitting the dataset into training and test sets



(b) Splitting data into a training and test set and then imputing each separately

Figure 2.2: An example of splitting data into a training set (grey) and a test set (purple) before or after imputing the original data.

2.8.1 Data Leakage in cross-validation

For cross-validation, data leakage can be considered in the same way as in the simple training and test split example. In one iteration of K -fold cross-validation, the dataset is split into $k - 1$ folds which make up the ‘training set’. The observations in the k^{th} fold are used to evaluate the prediction model fitted in the training set i.e. the k^{th} fold is a ‘test set’.

Similarly to the training and test split example in Figure 2.2(a), data leakage occurs when multiply imputing first and then applying cross-validation to each imputed dataset (Methods J and K). However, for *CV-then-MI* methods it is necessary to pay close attention as to which parts of the dataset are included when multiply imputing the training and test folds. When cross-validating first the data are split into the $k - 1$ training folds and the k^{th} test fold, followed by imputing the training and test folds separately which avoids data leakage. *CV-then-MI* methods with no data leakage are methods A and B. While Method C imputes the $k - 1$ training folds independently of the holdout fold k , fold k is imputed using the relevant covariates from the $k - 1$ training folds. However, as Y from the training folds is excluded from the MI process, method C should not be correlated with the training set. Methods D, E, F and H all include the holdout fold, including the outcome, when imputing the training set imputed datasets. Methods G and H include the outcome of the training folds when imputing the test fold.

It will be possible to examine the impact of data leakage in the MI step by comparing certain methods together.

| Methods | Comparing |
|------------------------|---|
| B vs. D, or C vs. E | the impact of including the test fold observations when drawing imputed values for the training set |
| B vs. C, or D vs. E | the inclusion of the training folds observations when drawing imputed values for the test set |

2.8.2 Data Leakage in the bootstrap algorithms

Figure 2.3 depicts data leakage which is present for both *BS-then-MI* and *MI-then-BS* in the *standard* algorithm. Due to imputing the entire original dataset first in *MI-then-BS*, the subsequent prediction models fitted to the bootstrap samples now have an association with all observations in the original dataset and not just those observations selected for the bootstrap sample. Therefore, when evaluating a prediction model (fitted to the bootstrap sample) in those who are not sampled, the model now has an unfair advantage through the MI process. It is also of note, that the *standard* algorithm calculates both the apparent and bootstrap performance where a model is trained and tested in the same sample (original and bootstrap, respectively). While this is data leakage, it is an inherent

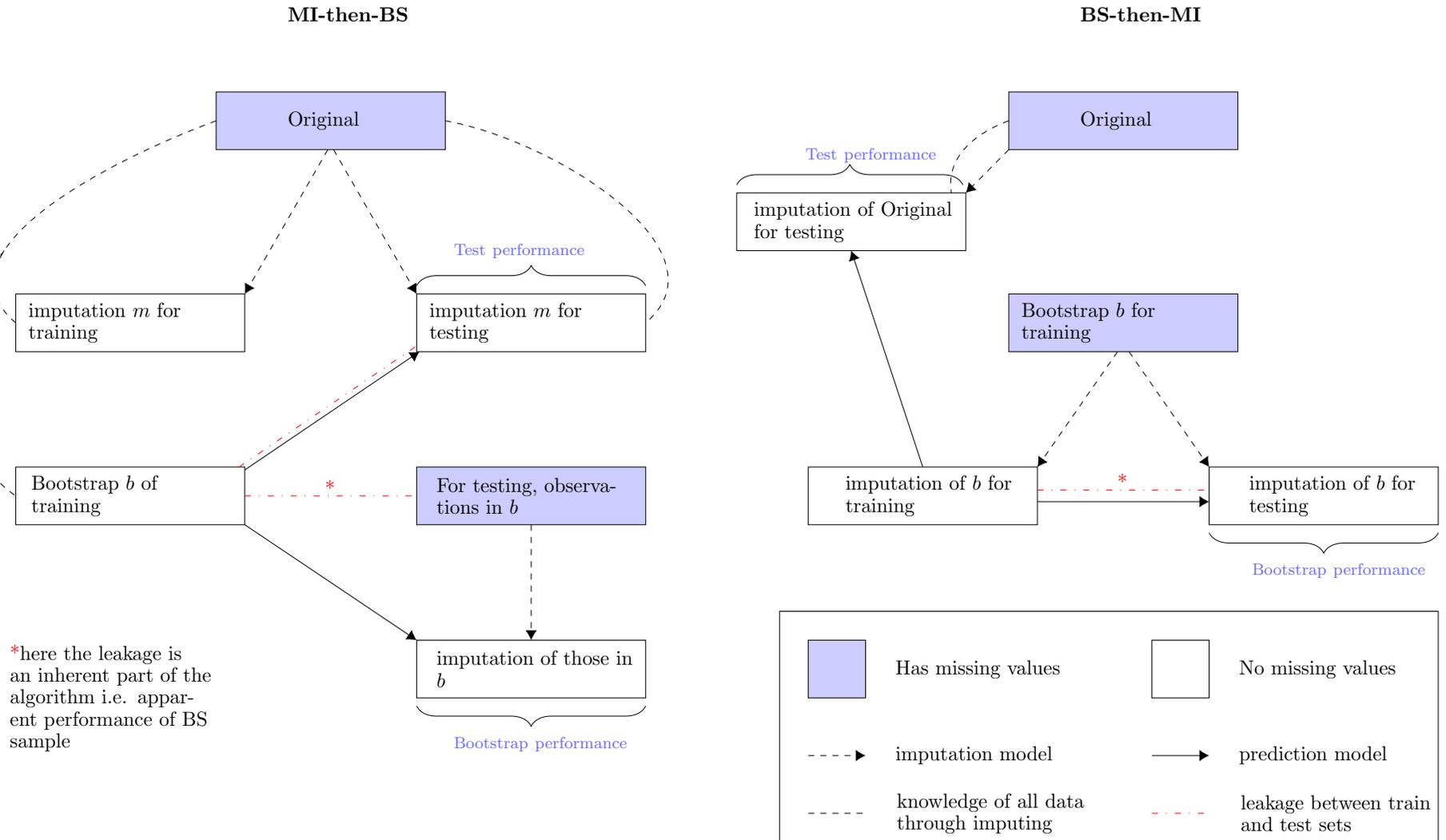


Figure 2.3: Data leakage flow in the standard algorithm for combining multiple imputation (MI) and the bootstrap (BS). This demonstrates leakage for one bootstrap sample b and one imputed dataset m .

part of the *standard* algorithm.

For *BS-then-MI* when using the *standard* algorithm the leakage for the apparent and bootstrap performance is present. However, when using *MI-then-BS* the bootstrap prediction models have an association with all observations in the original dataset. This is not an issue for *BS-then-MI* as MI occurs after the bootstrap sample.

Similarly for the 0.632 algorithm in Figure 2.4 when imputing the entire original dataset first in *MI-then-BS*, the subsequent bootstrap samples now have an association with all observations. When testing a bootstrap model in those who weren't sampled the model now has an unfair advantage through the MI process. This leakage is not present in *BS-then-MI*.

Similarly to using all folds to impute the training set or test set in cross-validation, reuse of the imputed datasets used to calculate apparent performance is expected to have the same type of leakage. The apparent performance imputed datasets use data from the entire dataset, and thus reusing these imputed datasets to sample observations for the bootstrap sample is also expected to cause data leakage.

Whether data leakage is an issue in the MI process is currently unknown and will be investigated in subsequent chapters using simulation studies.

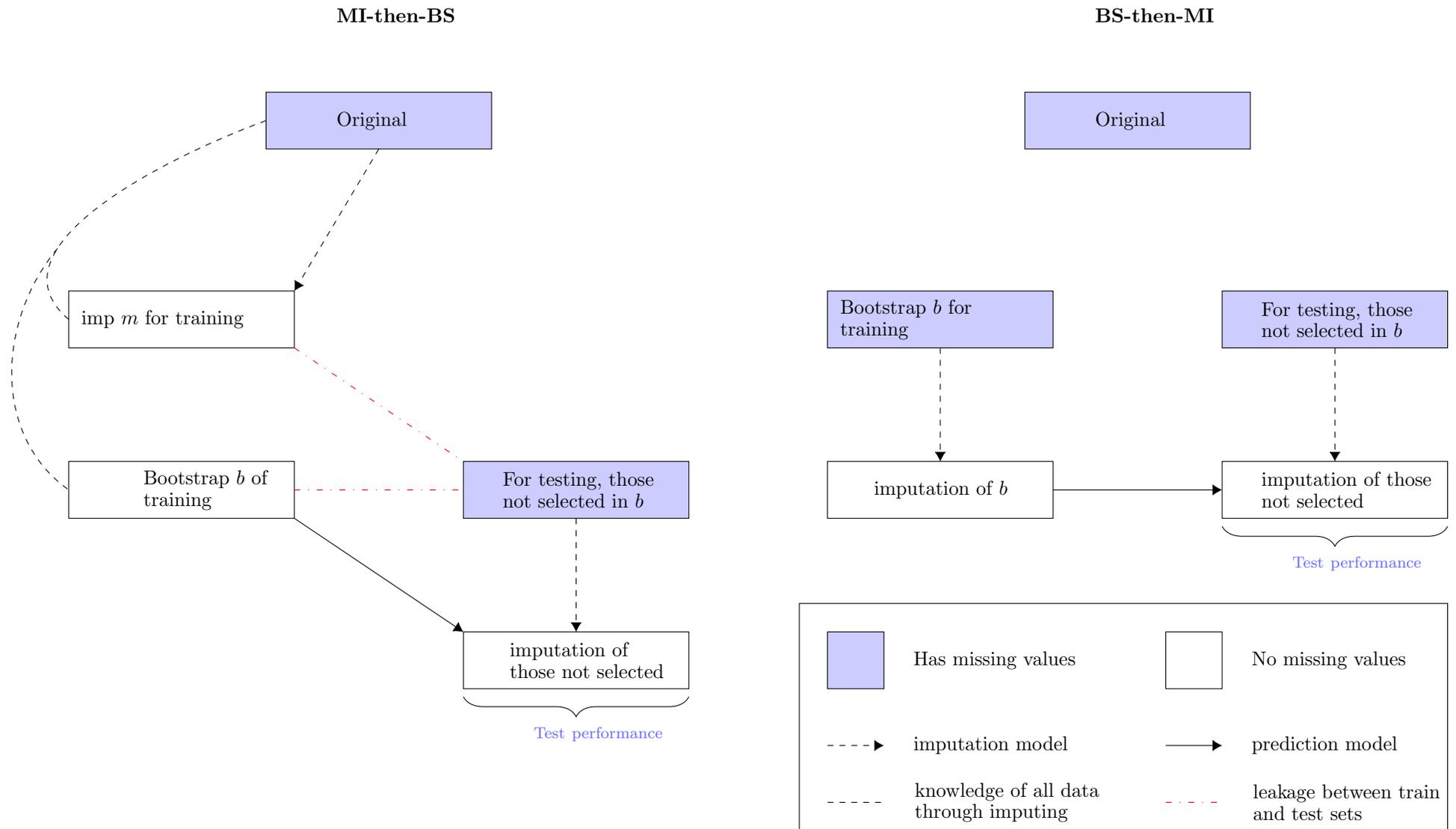


Figure 2.4: Data leakage in the 0.632 algorithm for combining multiple imputation (MI) and the bootstrap (BS). This demonstrates leakage for one bootstrap sample b and one imputed dataset m .

2.9 Conclusion

In this chapter I have summarised relevant current literature for combining internal validation with MI. Methods have been proposed to combine MI and two internal validation strategies. Despite the many methods under consideration here, the list of potential ways to combine is by no means exhaustive.

Much of the current literature has used the results of simulation studies to classify whether MI or validation should take place first. Wood [38] proposed several ways to combine, suggesting that imputing first may be most appropriate. Jaegar et al. [42] and Steyerberg [23] have suggested imputing first as a viable method. Alternatively, Musoro[43], Wahl [44] and Mertens [45] recommended or used a validation first approach. In this chapter, I have discussed the potential impact of data leakage through the MI process and this may be a potential justification for determining the best way to incorporate internal validation and MI. I have also explored and explicitly explained how to combine the two and which methods are prone to data leakage. This is in addition to evaluating the proposed methods via a simulation study presented in later chapters.

For ideal and pragmatic performance, I have proposed using two imputation models. The training imputation model will maintain the association between missing covariates and the outcome. For pragmatic performance, the test imputation model will exclude the observed outcome as in reality, this would not yet be known. For ideal performance, the test imputation model will include the observed outcome, as in practice we expect future values to be fully-observed.

In the next chapter, I will outline a simulation study which will be used to evaluate the methods I have proposed in this chapter. The simulation study will evaluate the methods for both a continuous and binary outcome.

3 Designing a simulation study to evaluate methods for combining MI and internal validation techniques

This chapter describes the design of simulation studies that aim to evaluate methods for combining internal validation techniques with MI. The simulation design is more complex than most simulation studies and therefore, I spend this chapter on the set-up and results which will be described in subsequent chapters. The outline of this chapter follows the ADEMP structure recommended by Morris et al. [10] for clear reporting of simulation studies. I will assess the proposed methods for both continuous and binary outcomes.

3.1 Aim

Chapter 2 outlined proposed methods for combining MI with cross-validation (Section 2.6) and the bootstrap algorithms (Section 2.7). The aim of the following simulation studies is to identify which of the proposed methods performs well across a range of different settings, including different amounts of missing data and multiple missing data mechanisms. The simulation studies will be used to assess the proposed methods for both cross-validation and the bootstrap optimism-corrected algorithms.

3.2 Data-generating mechanisms (DGM)

For the continuous outcome a linear model will be used as the prediction model, and logistic regression will be used for the binary outcome. For both scenarios, the linear predictor will have two correlated Normally distributed covariates, X_1 and X_2 . They will be generated with correlation $\rho = 0.5$.

$$\begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \sim \mathcal{N} \left(\begin{bmatrix} 25 \\ 2 \end{bmatrix}, \begin{bmatrix} 25 & \rho * 5 * 10 \\ \rho * 5 * 10 & 100 \end{bmatrix} \right)$$

3.2.1 Continuous outcome

The continuous outcome, Y , was generated using $Y \sim \mathcal{N}(\mu, \sigma^2)$, where

- $\mu = \beta_0 + \beta_1 X_1 + \beta_2 X_2$
- $\sigma^2 = (\beta_1^2 \text{Var}(X_1) + \beta_2^2 \text{Var}(X_2) + 2\beta_1 \beta_2 \text{Cov}(X_1, X_2)) \times \frac{1-R^2}{R^2}$

The model is written like this in order to allow the variance of the outcome (σ^2) to be adjusted for varying levels of R-squared (R^2) while keeping the values of β constant. Three values were considered for R^2 : 0.01, 0.1 and 0.3. The derivation for the adjustment of σ^2 to allow for different values of R^2 is available in Appendix B. In all simulation scenarios the values of the β_1 and β_2 parameters were one with the slope through the origin ($\beta_0 = 0$), allowing for assessment of the proposed methods in the simple scenario of a prediction model with no complexity.

3.2.2 Binary outcome

The binary outcome, Y , was generated using a Bernoulli distribution with the probability of the outcome for patient j ($j = 1, \dots, n_{obs}$):

$$p_j = \frac{\exp(\beta_0 + \beta_1 X_{j,1} + \beta_2 X_{j,2})}{1 + \exp(\beta_0 + \beta_1 X_{j,1} + \beta_2 X_{j,2})}$$

In all scenarios the values chosen for the log odds ratio parameters were $\beta_1 = \log(1.1)$ and $\beta_2 = \log(1.1)$ corresponding to odd ratios of 1.1. $\beta_0 = -3.676$ was selected so that approximately 30% of individuals have $Y = 1$. This value of β_0 was found by iterating through a range of values from -25 to 0 with 10,000 simulated datasets, each with a sample size of 1,000.

3.2.3 Introducing missingness

Missingness was induced in one covariate, X_1 , for the continuous and binary outcome DGMs. Scenarios in which the missingness in X_1 does and does not depend on X_2 or on the outcome Y are considered. For patient j the probability of X_1 being missing is:

$$\pi_{X_1,j} = \frac{\exp(\psi_0 + \psi_2 X_{2,j} + \psi_3 Y_j)}{1 + \exp(\psi_0 + \psi_2 X_{2,j} + \psi_3 Y_j)} \quad (3.1)$$

Using equation (3.1), three missing data scenarios were considered:

1. MCAR ($\psi_2 = 0, \psi_3 = 0$)
2. Covariate-dependent MAR ($\psi_2 \neq 0, \psi_3 = 0$)
3. Covariate- and outcome-dependent MAR ($\psi_2 \neq 0, \psi_3 \neq 0$)

For the two MAR mechanisms non-zero values of ψ_2 and ψ_3 were selected to produce weak and strong MAR. This strength was calibrated based on the area under a ROC curve (AUC) from regressing the missing indicator on the covariates related to missingness. Values for ψ_0 were then selected such that approximately 25% or 40% of observations in X_1 were set as missing.

Table 3.1 shows the finalised ψ parameter values and the AUC of missingness when the outcome is continuous. When missingness is MCAR or covariate-dependent MAR, missingness does not depend on the outcome and therefore the values of ψ are unaffected by the R^2 values. For covariate- and outcome-dependent MAR, the values of ψ are selected to maintain a similar missingness AUC for the three R^2 values.

Table 3.1: Specification of parameter values ψ_0, ψ_2, ψ_3 to ensure MCAR, weak MAR and strong MAR with approximately 25% ($\psi_{0,25}$) or 40% ($\psi_{0,40}$) of observations induced to be missing.

| Mechanism | R^2 | ψ_3 | ψ_2 | $\psi_{0,25}$ | $\psi_{0,40}$ | AUC |
|--|-----------|----------|----------|---------------|---------------|-------|
| MCAR | All R^2 | 0 | 0 | -1.1 | -0.41 | 0.500 |
| weak covariate-dependent MAR | All R^2 | 0 | 0.05 | -1.25 | -0.53 | 0.634 |
| strong covariate-dependent MAR | All R^2 | 0 | 0.1 | -1.52 | -0.69 | 0.743 |
| weak outcome-dependent MAR | 0.01 | 0.003 | 0 | -1.22 | -0.5 | 0.609 |
| | 0.1 | 0.009 | 0 | -1.375 | -0.665 | 0.604 |
| | 0.3 | 0.016 | 0 | -1.56 | -0.85 | 0.607 |
| weak outcome- and covariate-dependent MAR | 0.01 | 0.003 | 0.05 | -1.38 | -0.63 | 0.675 |
| | 0.1 | 0.009 | 0.05 | -1.76 | -0.79 | 0.685 |
| | 0.3 | 0.016 | 0.05 | -1.76 | -0.99 | 0.699 |
| weak outcome- and strong covariate-dependent MAR | 0.01 | 0.003 | 0.1 | -1.63 | -0.78 | 0.762 |
| | 0.1 | 0.009 | 0.1 | -2.05 | -0.95 | 0.772 |
| | 0.3 | 0.016 | 0.1 | -2.05 | -1.16 | 0.783 |

Table 3.2 displays the finalised parameter values for inducing missingness in X_1 when the outcome is binary.

Table 3.2: Specification of parameter values ψ_0, ψ_2, ψ_3 to ensure MCAR, weak and strong MAR for missingness dependent and not dependent on the outcome.

| Mechanism | ψ_2 | ψ_3 | $\psi_{0,25}$ | $\psi_{0,40}$ | AUC |
|--|----------|----------|---------------|---------------|-------|
| MCAR | 0 | 0 | -1.1 | -0.4 | 0.501 |
| weak covariate-dependent MAR | 0.05 | 0 | -1.25 | -0.54 | 0.635 |
| strong covariate-dependent MAR | 0.1 | 0 | -1.52 | -0.67 | 0.743 |
| weak outcome-dependent MAR | 0 | 0.9 | -1.4 | -0.7 | 0.600 |
| weak outcome- and covariate-dependent MAR | 0.05 | 0.9 | -1.6 | -0.83 | 0.707 |
| weak outcome- and strong covariate-dependent MAR | 0.1 | 0.9 | -1.88 | -0.97 | 0.791 |

3.2.4 Factors to vary in the simulation

Above I specified that different simulation scenarios will be considered for three values of R^2 for the continuous outcome, three missing data mechanisms, and two levels of missingness. Other factors that were varied included the sample size, and the number of imputations (M) used when performing MI - though this is a feature of the analysis rather than the DGM. The proposed methods will be assessed across 108 different simulated scenarios for the continuous outcome and 36 for the binary outcome. Each scenario was initially assessed with 1000 repetitions, however this was increased to 2000 in order to minimise Monte Carlo error. The factors varied (factorially) across scenarios and their values are

found in Table 3.3.

Some values in Table 3.3 such as a sample size of 100 patients or R-squared value of 0.01 were used to assess how the methods may perform in extreme scenarios. Increasing dependence of missingness on another covariate and also on the outcome is examined.

Table 3.3: Factors which will be varied for the continuous outcome simulations

| Factors | Notation | Values |
|------------------------------------|------------|--------------------------|
| <i>All scenarios</i> | | |
| Number of individuals | n_{obs} | {100, 300, 1000} |
| Number of repetitions used | n_{sim} | {2000} |
| Proportion of missingness | p_{miss} | {25%, 40%} |
| Dependence of missingness on X_2 | ψ_2 | {0, 0.05, 0.1} |
| <i>Continuous outcome only</i> | | |
| Level of R-squared | R^2 | {0.01, 0.1, 0.3} |
| Dependence of missingness on Y | ψ_3 | {0, 0.003, 0.009, 0.016} |
| <i>Binary outcome only</i> | | |
| Dependence of missingness on Y | ψ_3 | {0, 0.9} |

3.3 Estimands

In each simulation scenario and using each analysis method (see below) I assess the ideal and pragmatic estimates of performance measures. However, we lack a clear 'benchmark' for the performance of a method. The ideal and pragmatic performance measure estimates for each repetition will be compared to the performance measure estimated from the same repetition but with fully observed X_1 . We expect pragmatic estimates to underestimate those of ideal performance [38]. Similarly to the comparison with the fully-observed data, both ideal and pragmatic performance will be compared to a target ideal and pragmatic estimate using a larger simulated dataset generated using the same DGMs. This is discussed in more detail in section 3.6.

For all methods in the ideal and pragmatic setting an overall performance measure ($\widehat{\text{Perf}}_{imp}$) will be estimated. This will be compared to both:

- the performance measure calculated in the fully-observed case (Perf_{obs})

$$\widehat{\text{Perf}}_{imp} - \text{Perf}_{obs}$$

- a larger validation set to estimate the target performance (Perf_{target})

$$\widehat{\text{Perf}}_{imp} - \text{Perf}_{target}$$

3.4 Methods

The proposed methods for combining MI with cross-validation and bootstrapping described in Sections 2.6 and 2.7, respectively, will be assessed. They will be compared with methods already proposed in the literature such as using one set of imputations to train and evaluate prediction models. Particularly for the bootstrap, methods will also look at the reuse of imputed datasets for calculating the bootstrap or test performance to improve computational efficiency. MICE will be used [49], Bayesian linear regression [14, p.67-74] will be used when the outcome is continuous and predictive mean matching [14, p.77-84] when the outcome is binary.

3.5 Performance Measures

The choice of performance measures for the continuous and binary outcome will now be outlined.

3.5.1 Continuous outcome

The performance measure for the prediction models when the outcome is continuous is the MSE. For method *imp* and simulated repetition, $r = 1, \dots, n_{sim}$, the overall MSE for each DGM is:

$$\widehat{\text{MSE}}_{imp} = \frac{1}{n_{sim}} \sum_{r=1}^{n_{sim}} \widehat{\text{MSE}}_{r,imp}$$

The fully-observed MSE and the target MSE from a larger validation set will also be estimated:

$$\text{MSE}_{obs} = \frac{1}{n_{sim}} \sum_{r=1}^{n_{sim}} \widehat{\text{MSE}}_{r,obs}$$
$$\text{MSE}_{target} = \frac{1}{n_{sim}} \sum_{r=1}^{n_{sim}} \widehat{\text{MSE}}_{r,target}$$

As outlined in Section 3.3, $\widehat{\text{MSE}}_{imp}$ will be compared with the averaged MSE when data are fully-observed ($\widehat{\text{MSE}}_{imp} - \text{MSE}_{obs}$) and with the target performance ($\widehat{\text{MSE}}_{imp} - \text{MSE}_{target}$). These are equivalent to the Perf_{obs} , Perf_{target} and Perf_{imp} notation outlined in Section 3.3.

3.5.2 Binary outcome

Initially, the Brier score was considered as this is the binary outcome equivalent of the MSE. The AUC is another popular metric used and reported with logistic regression and was also considered.

In addition, the calibration intercept and slopes will be calculated. While calibration could also be assessed within the continuous outcome scenario, it is not a very popular measure

for linear regression and was not considered here.

Each of these performance measures will be averaged across the 2000 simulated repetitions and compared to their fully-observed and target performance, as detailed in the continuous outcome case above.

3.6 Finding the ‘Target’ performance measure

Comparing the averaged estimate of performance to a target value of the performance measure estimated using a larger validation set was discussed in section 3.3. In this section, I will detail the different steps that were taken to estimate Perf_{target} and explain why these methods were discarded. Finally, I will present the method used to generate Perf_{target} .

3.6.1 Generating very large datasets to estimate the target MSE

The first simulation study undertaken was when the outcome was continuous and it was of interest to compare $\widehat{\text{MSE}}_{imp}$ to a target MSE. Initially, 100 large external datasets of size 100,000 were simulated using the same DGMs (Section 3.2).

To estimate the target ideal performance, the prediction model (with the true value of parameters for X_1 and X_2) was applied to each large simulated dataset, with X_1 fully-observed, and the MSE was obtained. The resulting MSEs were then averaged over the 100 datasets to get an overall performance estimate for the ideal target MSE.

For the pragmatic setting, each large simulated dataset, with X_1 partially-observed, was imputed ($M = 25$) using a test imputation model i.e. the imputation model only included the other covariate X_2 . The prediction model (using the true parameter values for X_1 and X_2) was applied to each imputed dataset and Rubin’s first rule was used to get an overall MSE estimate. This was repeated across the 100 large datasets and the resulting MSEs were then averaged to get an overall performance estimate for the pragmatic target MSE.

When compared with the output from the proposed methods there was a tendency with both increasing sample size and increasing R-squared for the difference between the imputation methods and the target MSE ($\text{MSE}_{imp} - \text{MSE}_{target}$) to be overoptimistic. This implied that performance from the imputation methods were performing better than the target MSE. This trend was also present when using validation methods on the fully observed replications ($\text{MSE}_{obs} - \text{MSE}_{target}$).

This method was evaluating the performance of the true prediction model rather than the model of interest which is the one that is fit to the data. Therefore, this method to find

the target MSE of the performance measure was discarded and was not attempted for the binary outcome case.

3.6.2 Using the fully-observed data

To get around the fact that there was no target value for comparing methods, the fully-observed data was used. Comparing ideal performance to the performance in fully-observed data ($MSE_{imp,ideal} - MSE_{obs}$) was considered to be the gold standard approach for comparing methods by Wood et al. [38]. They also expect that pragmatic performance should underestimate the ideal performance i.e. $|MSE_{imp,prag} - MSE_{obs}| \geq |MSE_{imp,ideal} - MSE_{obs}|$.

Based on Wood et al. [38], ideal performance of methods could be compared to the fully-observed data. Pragmatic performance of methods would be compared using the fully-observed data and also by looking at how the methods performed in the ideal scenario. In this way, I could determine whether the pragmatic performance of a method was performing well ($MSE_{imp,prag} - MSE_{obs} \rightarrow 0$) or whether it actually had a tendency to under or overestimate the difference (underestimate: $MSE_{imp,prag} - MSE_{obs} < 0$; overestimate: $MSE_{imp,prag} - MSE_{obs} > 0$).

3.6.3 Simulating AUC target performance for the binary outcome

Another attempt to find a target value of performance included attempting to simulate data with a pre-specified value of the AUC. This can be simulated easily for one covariate, X_1 , by converting the AUC into a Cohen's d value [51] and sampling two vectors which are d standard normal distributions apart, these are then combined to create covariate X_1 . Therefore, a dataset can be simulated with outcome Y and covariate X_1 with the underlying true value set as the AUC of interest. This was considered alongside the method described below in section 3.6.4, which is easier to implement and therefore this was not considered any further.

3.6.4 Generating a test set for each repetition

The final way considered to generate a true value for the performance measures of the simulation study was to generate a sufficiently large dataset to test models in, using the same DGM detailed in section 3.2 for both the continuous and binary outcomes. The test dataset generated for each DGM contained 100,000 observations, missingness was induced under the same mechanisms as before.

Ideal performance estimates are compared with an estimate from the fully-observed generated test dataset. For pragmatic performance estimates the test dataset is imputed

($M = 5$) by fitting a test imputation model to the data (excluding the outcome). Pragmatic performance estimates are compared to the Rubin’s rule averaged performance estimate from the M pragmatically imputed test datasets.

There are two methods to estimate target performance from the large generated dataset. Both theoretically will produce true values of a performance measure and are described below.

Method 1: Test each internal validation prediction model

Sections 2.6 and 2.7 outlined the proposed methods for combining multiple imputation with cross-validation and bootstrapping, respectively.

For all proposed cross-validation methods, a prediction model is fitted to $k - 1$ folds. This model could then be applied to the larger test dataset to get an estimate of performance for this prediction model. This would be repeated for $k = 1, \dots, K$ and the resulting estimates would be averaged to get a target estimate. Similarly, when using the 0.632 bootstrap optimism-correction method, each bootstrap-trained prediction model can be evaluated in the large test set to get a test performance estimate.

However, this means that we would be finding a target value which is specific to cross-validation. This would be different to the target value for the bootstrapping algorithms.

Method 2: Use as an external validation style dataset

In general practice when using internal validation, a model is trained using the entire dataset available. It is then validated using the same dataset to get a performance estimate of the modelling procedure. Internal validation is an option when there is a lack of availability of a similar external dataset to evaluate the model.

Instead of using the large test dataset to evaluate each prediction model from the validation process (as described in Method 1 above), it could instead be used to evaluate the final model trained from the dataset. For example, in simulated repetition r the dataset is imputed using the training imputation model and an overall model is determined using Rubin’s rules. This model can be evaluated in the large test dataset (either the fully-observed version for ideal performance or its imputed version for pragmatic performance). The internal validation methods are then applied to repetition r and the estimated performance can be compared to the large test dataset performance.

Comparing the performance of this model evaluated in a larger similar dataset with the internal validation estimate has an advantage compared to testing each internal validation prediction model. Method 2 estimates one overall target value for each repetition which

allows cross-validation and the bootstrap optimism-corrected methods to be compared. This method was used to approximate the target value of each performance measure.

3.7 Conclusion

In this chapter I have discussed the set-up of the simulation study to be used for both cross-validation and the bootstrap internal validation algorithms. I have outlined two ways to compare the performance estimates from the proposed methods in the simulation study - the first comparing with the estimate from the fully-observed data and the second using a large test set to evaluate the final developed model fitted to imputed datasets.

4 Simulation study results for cross-validation: continuous outcome

In Chapter 3 I described the design for a simulation study to investigate the performance of methods when combining MI with internal validation.

4.1 Introduction

In this chapter I present the results from combining MI with cross-validation. The impact of data leakage, which was introduced in Chapter 2, on the methods to impute the missing data will also be assessed. The output from the simulation study for the continuous outcome will be presented here, the results for the binary outcome are available in Chapter 5. A small selection of graphs have been made available in this chapter, selected for either having important results or being representative of results across various DGMs. All graphical output from the simulation study is available in the supplementary plot chapter (Section S1).

Several factors were varied for the continuous setting as detailed in Table 3.3. Results will be detailed below for these factors which included sample size, value of R-squared and dependence of missingness on other covariates. In section 3.5 notation was presented for the averaged estimate of MSE in the fully-observed data (MSE_{obs}) and the larger validation set (MSE_{target}). In addition, MSE_{prag} will represent the pragmatic performance of an imputation method and MSE_{ideal} will represent the ideal performance.

4.2 Summary of the fully-observed data

I begin by summarising the fully-observed data, which is the simulated data before missingness is introduced in the covariate X_1 . With increasing R^2 and sample size the variation of the outcome Y decreases (Table 4.1). Similarly, the MSE decreases slightly with increased sample size (Table 4.2). Increasing R^2 causes the MSE to decrease from 17,388 for $N = 1000$ when $R^2 = 0.01$ to 410 when $R^2 = 0.3$.

Table 4.1: The mean and variance of the outcome Y across the 2000 simulated datasets. The min and max values of Y are the minimum and maximum across all repetitions.

| N_{obs} | Summary statistics | $R^2 = 0.01$ | $R^2 = 0.1$ | $R^2 = 0.3$ |
|-----------|--------------------|-----------------|-----------------|-----------------|
| 100 | Mean (var) | 26.91 (17502) | 26.97 (1748) | 26.98 (582) |
| | Min, Max | -611.05, 690.76 | -177.57, 228.47 | -85.72, 140.80 |
| 300 | Mean (var) | 27.10 (17555) | 27.02 (1757) | 27.00 (586) |
| | Min, Max | -583.63, 642.53 | -167.02, 224.88 | -83.72, 141.48 |
| 1000 | Mean (var) | 27.02 (17473) | 27.01 (1746) | 27.01 (582) |
| | Min, Max | -671.94, 680.66 | -190.08, 227.56 | -101.81, 140.20 |

Table 4.2: Summary of the MSE estimates when data are fully-observed. This is summarised from the 2000 simulated repetitions.

| R^2 | N=100 | | N=300 | | N=1000 | |
|-------|--------|------------|--------|-----------|--------|-----------|
| | Mean | Variance | Mean | Variance | Mean | Variance |
| 0.01 | 17,831 | 13,431,859 | 17,546 | 4,087,025 | 17,388 | 1,238,694 |
| 0.10 | 1,633 | 114,500 | 1,598 | 33,237 | 1,578 | 10,028 |
| 0.30 | 422 | 7,457 | 413 | 2,183 | 410 | 675 |

4.3 A brief summary of the cross-validation methods

The methods presented in this chapter are summarised in full in Table 2.3 but are briefly resummarised below in Table 4.3.

Table 4.3: Brief summary of methods A-K for combining multiple imputation and cross-validation

| Method | Training set | Test set |
|-------------------|---|--|
| <i>CV-then-MI</i> | | |
| A | Each fold imputed separately including Y and X_2 in the imputation model | k^{th} fold imputed by itself using X_2 and possibly Y |
| B | Y , X_2 used to impute $k - 1$ training folds by themselves | Same as A |
| C | Same as B | k^{th} fold imputed using all K folds and including X_2 and possibly Y |
| D | Y , X_2 used to impute $k - 1$ training folds using all K folds and restricting to $k - 1$ folds after imputation process | Same as A |
| E | Same as D | Same as C |
| F | Same as B | Take the imputed and observed values from the $k - 1$ training folds and use to impute the unobserved values in the k^{th} fold. |
| G | Same as B | Set Y missing in k^{th} fold and impute X_1 and Y using data from all K folds before restricting back to the k^{th} fold |
| H | Same as D | Same as G |
| <i>MI-then-CV</i> | | |
| J | Impute the dataset first using one set of imputed datasets | |
| K | Impute the dataset first using two sets of imputed datasets - one for training the model on $k - 1$ folds and the other for evaluating the model in the k^{th} fold | |

4.4 A brief overview of results for cross-validation

Due to the large number of results from the simulation studies presented in this chapter which assess the various methods under multiple DGMs, I will first present results for two methods when $R^2 = 0.1$. The aim is to introduce the reader to how the results are being displayed and interpreted as well as introducing the impact that data leakage can have on the results.

I will briefly compare method B, which applies cross-validation first and then imputes, to method J which imputes the data first before applying cross-validation. Method B has no data leakage issues while method J is considered to be the method with the highest risk of leakage. The MSE results from each method are compared to the estimates of the MSE from applying cross-validation to the fully-observed data (MSE_{obs}) i.e. $MSE_{imp} - MSE_{obs}$.

For all sample sizes and missing data scenarios, the estimated pragmatic performance of both method B and method J overestimates MSE_{obs} i.e. $MSE_{prag,imp} - MSE_{obs} > 0$. Method B tends to overestimate MSE_{obs} to a greater degree ($|MSE_{prag,B} - MSE_{obs}|$) than method J for all sample sizes. However, with increasing sample size the magnitude of the difference ($|MSE_{prag,imp} - MSE_{obs}|$) for both methods decreases and the difference becomes more similar between the two methods. This can be seen across all missing data scenarios for $R^2 = 0.1$.

The estimated ideal performance of method B tends to overestimate MSE_{obs} for all sample sizes. However, method J underestimates MSE_{obs} for all sample sizes. This means that the results from method J are over-optimistic for ideal performance i.e. the method gives better performance post-imputation than what would have been observed if missing data were not present. The magnitudes of under- or overestimation of the two methods are similar across all missing data scenarios and for sample sizes greater than 100.

In the following section, I will present a summary of results for all methods in a similar manner as above. Recall that a ‘good’ prediction model would have a lower MSE score. Therefore, over-estimation of MSE_{obs} implies worse performance after handling missing data than if the data had all been observed to begin with. Under-estimation of MSE_{obs} suggests that the method is over-optimistic; that is, it is performing better than if we had observed the data. In the following results, it will be shown that many of the methods which are subject to data leakage tend to have over-optimistic ideal performance.

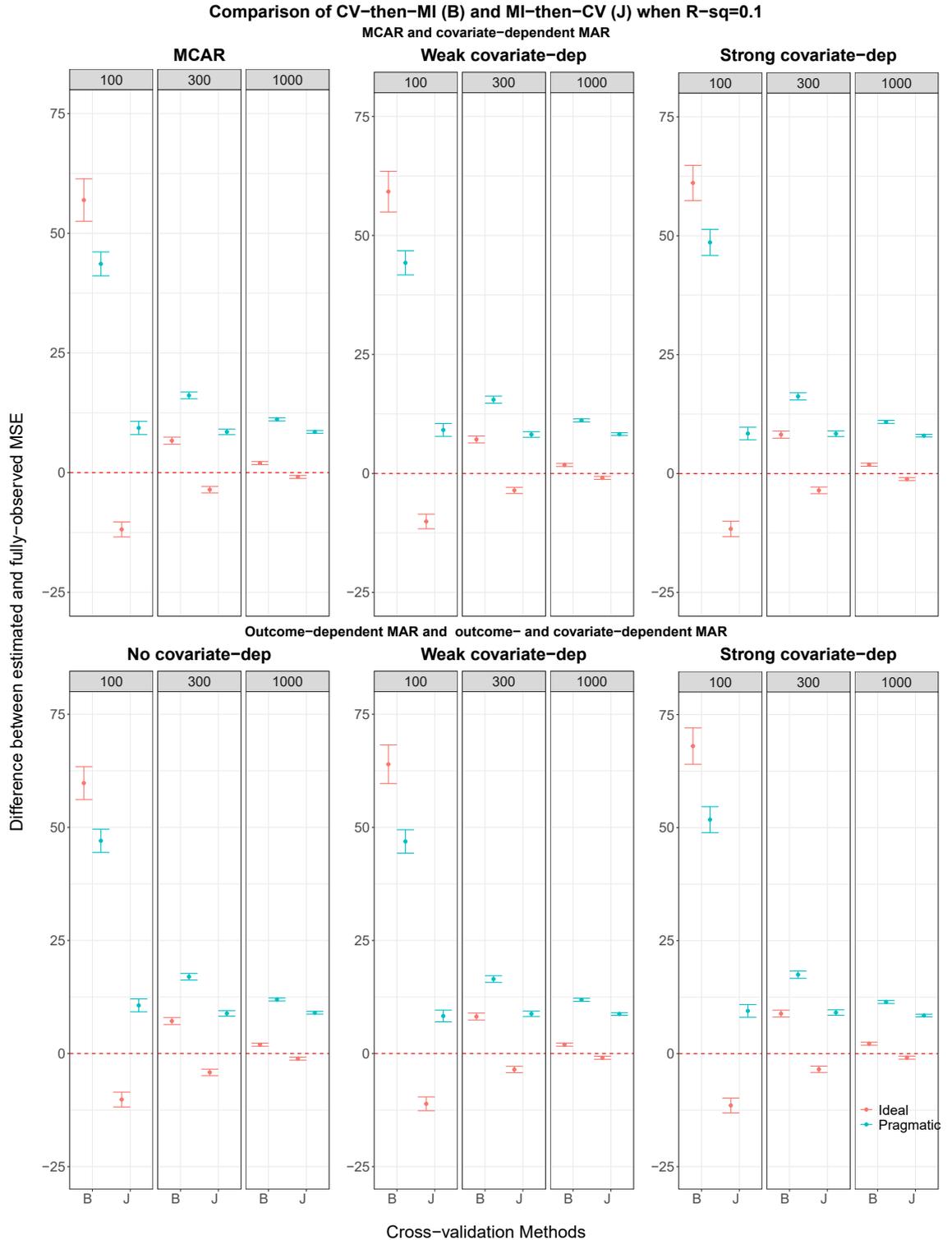


Figure 4.1: The difference $MSE_{imp} - MSE_{obs}$ when $R^2 = 0.1$ for $M = 5$ when 25% of values are missing in X_1 . Each sub-graph displays results for a sample size of 100, 300 and 1000. Row 1 presents results when data are MCAR or covariate-dependent MAR. Row 2 presents results when data are outcome-dependent MAR or outcome- and covariate-dependent MAR. Ideal performance is in red and pragmatic performance is in blue. The error bars summarise results from the 2000 repetitions and the limits represent the Monte Carlo 95% confidence interval of $MSE_{imp} - MSE_{obs}$. CC (complete-case); methods A-K are described in Table 2.3 and summarised in Table 4.4.

4.5 Detailed results for cross-validation

In this section I will summarise the results from the simulation study when the outcome is continuous. The results for 25% missingness and $M = 5$ will initially be presented before discussing increasing the number of imputed datasets, increasing the percentage of missingness or comparing the results to a target MSE estimate.

4.5.1 Comparing results to the MSE estimate when data are fully-observed

MCAR and covariate-dependent MAR

Figure 4.2 displays the estimates of various methods when compared to the MSE estimates when data are fully-observed. The plot shows results when data are weak covariate-dependent MAR but is representative of the MCAR or strong covariate-dependent MAR scenarios (additional figures in supplementary plots S1.1.1).

When data are MCAR or covariate-dependent MAR and for small values of R-squared the complete-case analysis tends to overestimate MSE_{obs} ($MSE_{CC} - MSE_{obs} > 0$) and is more variable than the MI methods. With increasing sample size, Monte Carlo standard error is reduced for the complete-case analysis and with increased R^2 the complete-case outperforms the other methods.

For a sample size of 100 and $R^2 = 0.01$, the pragmatic performance of method A (impute each fold separately) outperforms all other methods which exclude the holdout fold k from imputing the training $k - 1$ folds (B, C, F, G). When $R^2 = 0.1$ it performs similarly to methods C, F and G but is out-performed by methods C, F and G for $R^2 = 0.3$. Overall, method J has the best performance with the lowest difference ($MSE_{J,prag} - MSE_{obs}$) overall for increasing R^2 and sample size. With increasing sample size to 300 and 1000 the pragmatic performance of all methods is similar.

For low R-squared and small sample size, the ideal performance of method F has the smallest difference ($MSE_{F,ideal} - MSE_{obs}$) for all methods which exclude the holdout fold k from imputing the training $k - 1$ folds. With higher R^2 and increased sample size, the methods tend to have similar performance for the ideal setting. Method A outperforms method B for small and moderate sample size while they perform similarly for a sample size of 1000. Method K has the smallest difference overall across all methods but tends to be over-optimistic ($MSE_{K,ideal} - MSE_{obs} < 0$), as is method J. The ideal performance for methods C and E largely overestimate MSE_{obs} and tends to be more variable than the other methods.

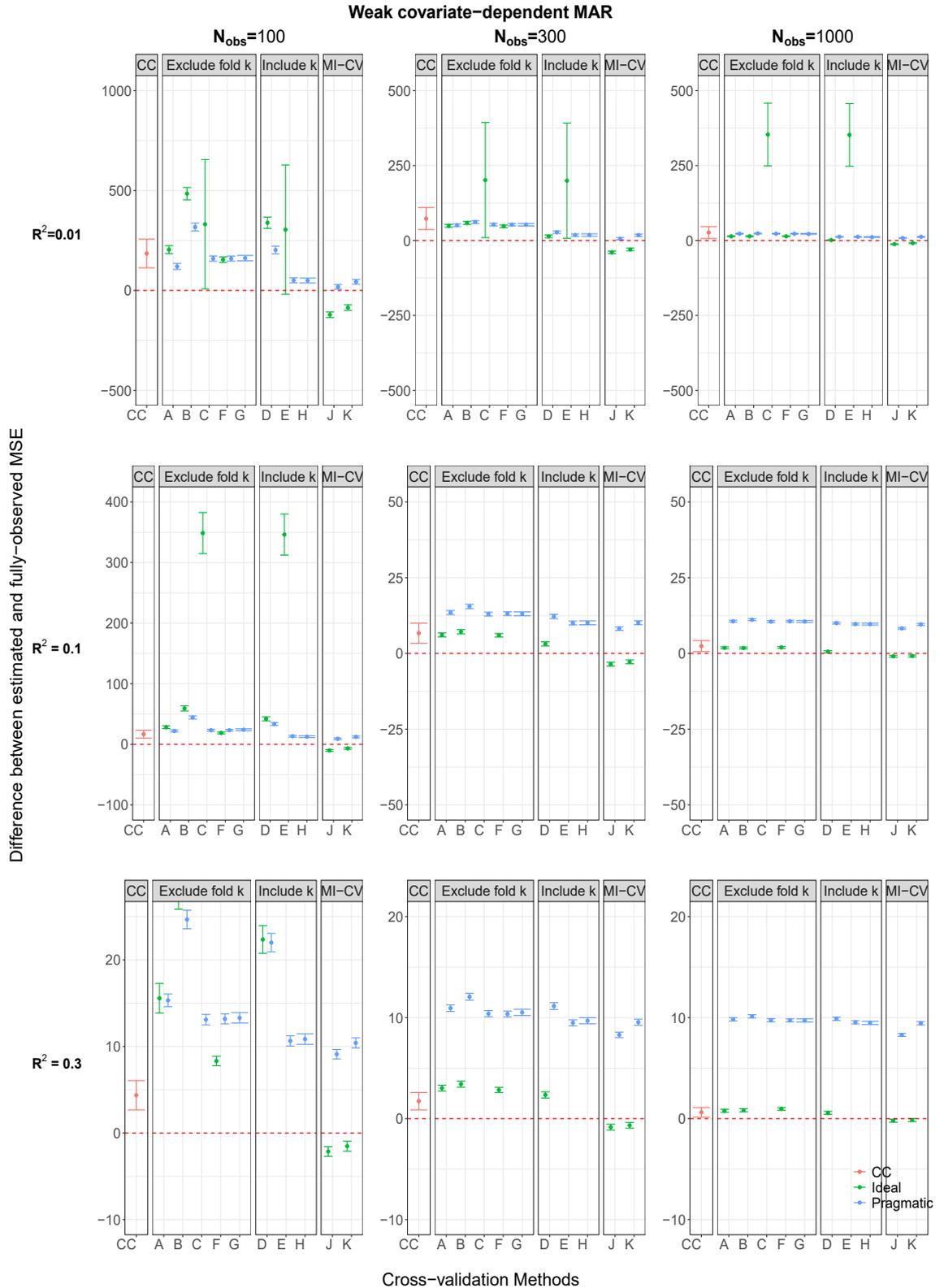


Figure 4.2: The difference $MSE_{imp} - MSE_{obs}$ when data are weakly covariate-dependent MAR for $M = 5$ when 25% of values are missing in X_1 . The error bars summarise results from the 2000 repetitions and the limits represent the Monte Carlo 95% confidence interval of $MSE_{imp} - MSE_{obs}$. CC (complete-case); methods A-K are described in Table 2.3 and summarised in Table 4.4.

Outcome-dependent MAR

Figure 4.3 displays the estimates of various methods when compared to MSE_{obs} when data are weak outcome- and covariate-dependent MAR. This graph is representative of all outcome-dependent MAR scenarios (additional graphs available in supplementary plots).

When data are outcome-dependent MAR and for all sample sizes and levels of R-squared complete-case analysis underestimates MSE_{obs} ($MSE_{CC} - MSE_{obs}$) and is more variable than the imputation methods.

For all sample sizes and levels of R-squared, the ideal performance of *MI-then-CV* methods J and K tend to underestimate the MSE_{obs} ($MSE_{imp,ideal} - MSE_{obs} < 0$, $imp = J, K$). With increasing sample size, the difference between the estimated and fully-observed MSE for all methods tends to zero and ideal performance tends to outperform pragmatic performance. Two exceptions are methods C and E whose ideal performance greatly overestimates MSE_{obs} . These methods have an average difference greater than 300 for R^2 of 0.1 and 0.3, and therefore are not visible in the figure for these values of R^2 due to the scale of the vertical axis. Across all scenarios the ideal performance for methods C and E is poor and overestimates MSE_{obs} . As can be seen in the first row of Figure 4.3, the ideal performance estimates for C and E are highly variable compared to the ideal performance of other methods. Across all scenarios, ideal performance of *MI-then-CV* methods J and K tends to underestimate the MSE whereas *CV-then-MI* methods A-H tend to overestimate the MSE.

For a sample size of 100, ideal performance for methods A, B and D tends to overestimate MSE_{obs} more so than pragmatic performance. With increasing R-squared ideal performance is better than pragmatic performance for method A but not for methods B or D. However, increasing the sample size to 300 results in ideal performance being better than pragmatic performance.

For pragmatic performance and sample sizes of 100 or 300, method A has a smaller difference than method B ($MSE_{A,prag} - MSE_{obs} < MSE_{B,prag} - MSE_{obs}$). When excluding the k^{th} fold from imputing the $k - 1$ training folds for small sample sizes, method F tends to have the smallest difference between its MSE estimate and MSE_{obs} . *MI-then-CV* method J tends to have the smallest difference ($MSE_{J,prag} - MSE_{obs}$ overall for all scenarios). With increasing sample size, the pragmatic performance of the imputation methods tends to perform similarly.

Overall (excluding ideal performance for methods C and D), methods A-K perform similarly with increasing levels of R-squared and increasing sample size for ideal performance when compared with MSE_{obs} . Similarly the pragmatic performance of the methods

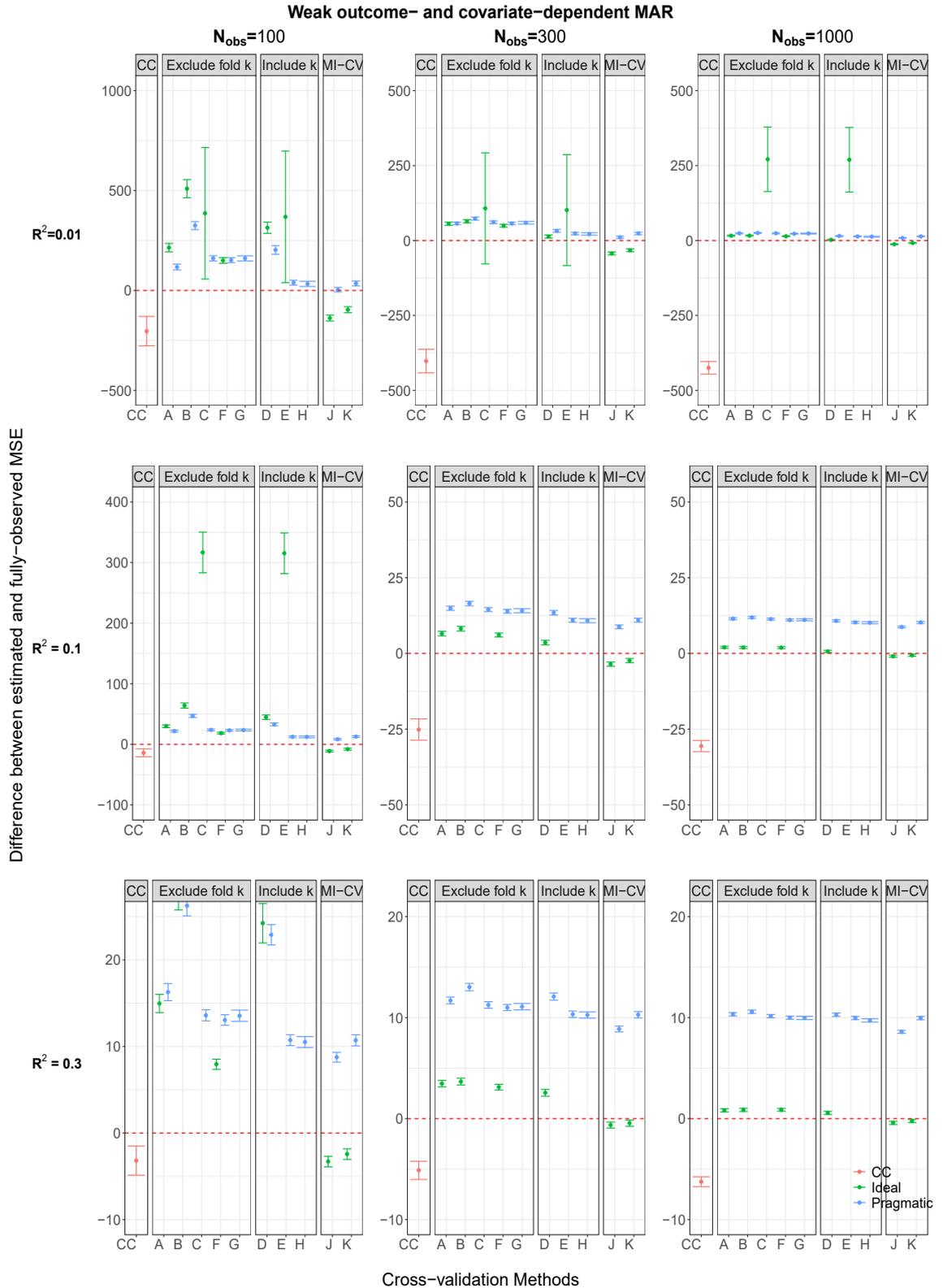


Figure 4.3: The difference $MSE_{imp} - MSE_{obs}$ when data are weakly outcome- and covariate-dependent MAR for $M = 5$ when 25% of values are missing in X_1 . The error bars summarise results from the 2000 repetitions and the limits represent the Monte Carlo 95% confidence interval of $MSE_{imp} - MSE_{obs}$. CC (complete-case); methods A-K are described in Table 2.3 and summarised in Table 4.4.

performs similarly with increasing sample size. The ideal performance of methods J and K tends to underestimate MSE_{obs} while the ideal or pragmatic performance of all other methods overestimate MSE_{obs} .

4.5.2 Increasing the number of imputed datasets from 5 to 25

Figure 4.4 shows results for comparing 5 versus 25 imputed datasets when estimating pragmatic performance and comparing it to MSE_{obs} ($MSE_{imp} - MSE_{obs}$). The results in the graph are for the scenario when data are weak outcome-dependent MAR and $R^2 = 0.01$ but are reflective of all scenarios for both pragmatic and ideal performance (additional plots in Supplementary plots S1.1.4).

Increasing the number of imputed datasets has little effect on the various methods' MSE estimates when comparing them to MSE_{obs} . For all imputation methods, using 25 imputed datasets results in similar estimates to using 5 imputed datasets, with similar Monte Carlo variability across the 2000 repetitions.

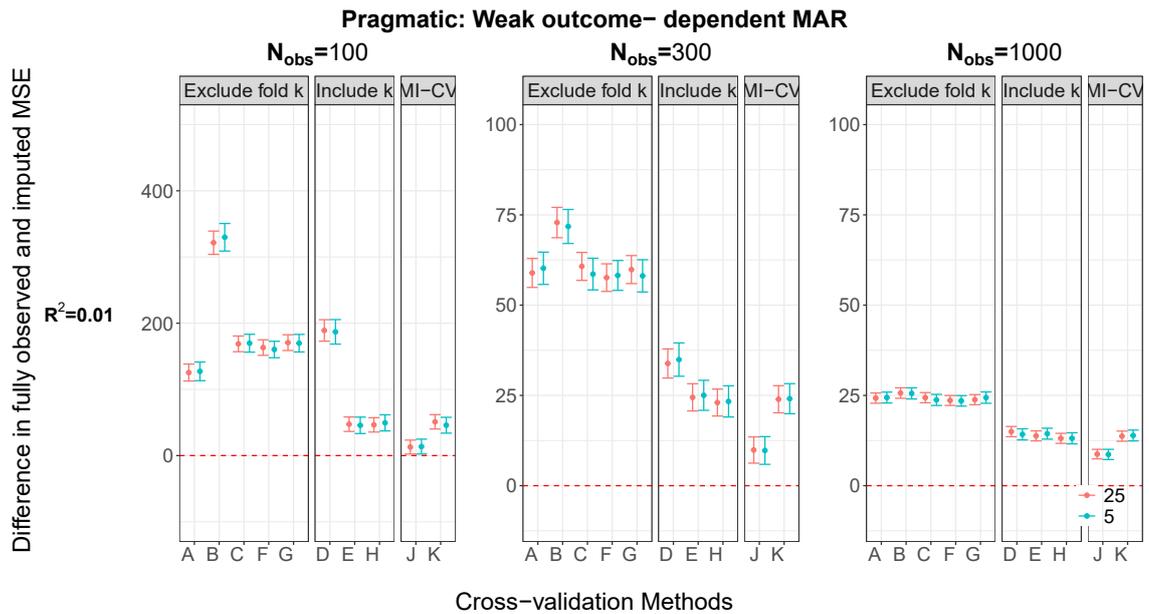


Figure 4.4: The difference $MSE_{imp} - MSE_{obs}$ when data are weakly outcome-dependent MAR for $M = 25$ versus $M = 5$ when 25% of values are missing in X_1 . The error bars summarise results from the 2000 repetitions and the limits represent the Monte Carlo 95% confidence interval of $MSE_{imp} - MSE_{obs}$. CC (complete-case); methods A-K are described in Table 2.3 and summarised in Table 4.4.

4.5.3 Increasing the percentage of missingness to 40%

Figure 4.5 displays the results for comparing missing data methods to the fully-observed MSE when 25% versus 40% of values in X_1 are missing. The graph presents results when data are weak outcome- and covariate-dependent MAR for $R^2 = 0.1$, results are similar for ideal and pragmatic performance in all other scenarios (additional plots in Supplementary plots S1.1.3).

When data are MCAR or covariate-dependent MAR, the complete-case analysis when 40% of X_1 values are missing performs similarly to when 25% of data are missing but has increased variability. When missingness is dependent on the outcome and potentially on covariate X_2 , as seen in Figure 4.5, complete-case analysis has an increased magnitude $|\text{MSE}_{CC} - \text{MSE}_{obs}|$ and variability when 40% of the values for X_1 are missing, compared to 25%.

For pragmatic performance, the MSE estimates when 40% of X_1 values are missing tend to overestimate MSE_{obs} compared to when 25% of values are missing ($\text{MSE}_{imp,40} - \text{MSE}_{obs} > \text{MSE}_{imp,25} - \text{MSE}_{obs}$). In some instances, the magnitude of the difference when 40% of the data are missing may be slightly smaller than when data are 25% missing, such as method J for a sample size of 300 in Figure 4.5, but this changes back to being bigger with increased sample size. The variability of the MSE estimates across repetitions is greatly increased when compared to 25% of values being missing.

For ideal performance, the difference between the imputation methods' MSE estimates and MSE_{obs} when 40% of X_1 values are missing tends to be similar or greater than when 25% of values are missing. Similarly to complete-case analysis and pragmatic performance, the variability of the ideal performance estimates of MSE ($\text{MSE}_{imp,40} - \text{MSE}_{obs}$) have greatly increased for 40% of values missing compared to 25%.

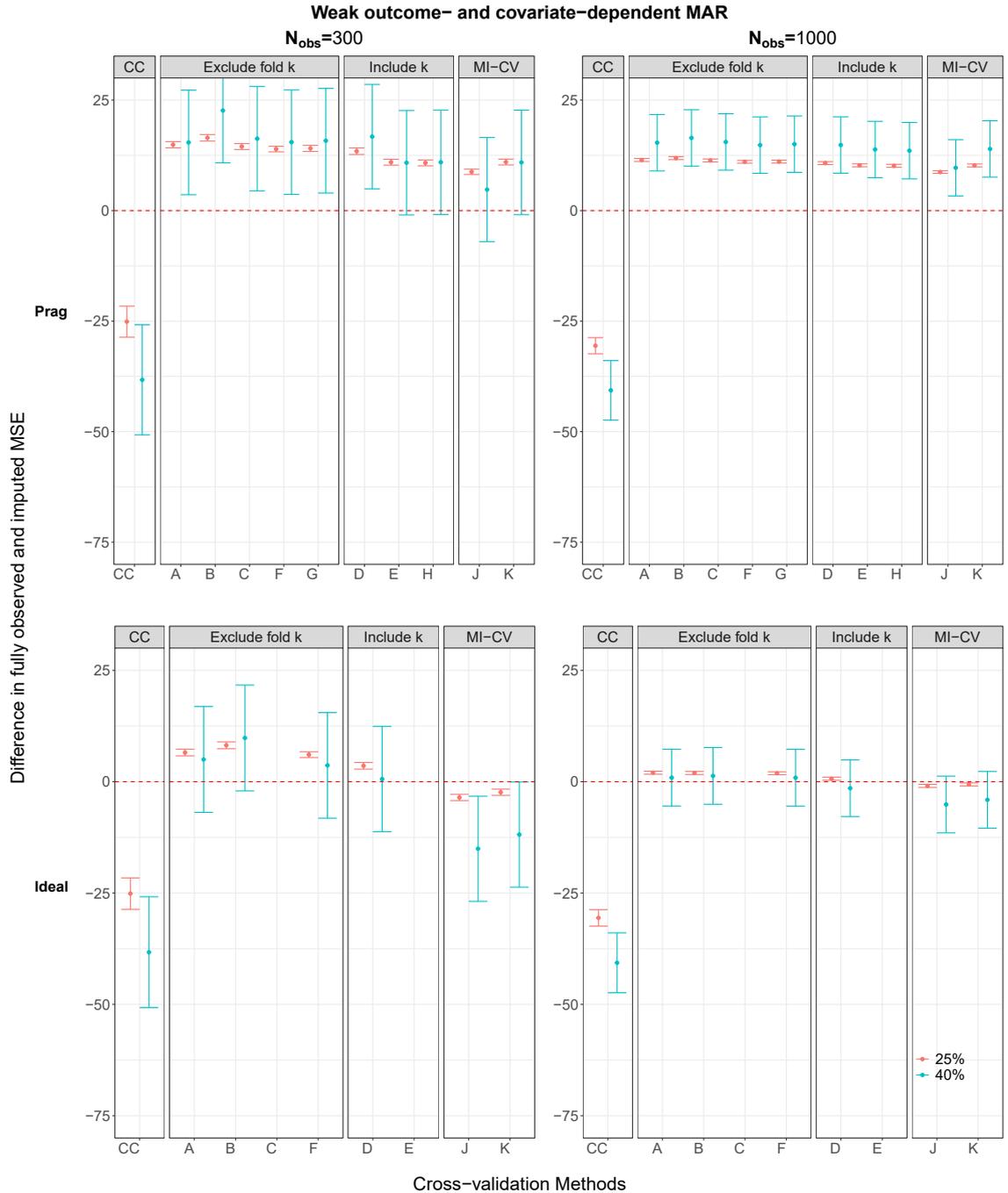


Figure 4.5: Comparing the impact of increasing the percentage of missingness on the difference $MSE_{imp} - MSE_{obs}$ when $M = 5$, $R^2 = 0.1$ and data are weakly outcome- and covariate-dependent MAR. The error bars summarise results from the 2000 repetitions and the limits represent the Monte Carlo 95% confidence interval of $MSE_{imp} - MSE_{obs}$. Red denotes $MSE_{imp} - MSE_{obs}$ when 25% of X_1 values are missing and blue denotes $MSE_{imp} - MSE_{obs}$ when 40% of X_1 values are missing. The top row presents the results for pragmatic performance and the bottom row presents results for ideal performance. CC (complete-case); methods A-K are described in Table 2.3 and summarised in Table 4.4.

4.5.4 Comparing to the target performance

Briefly as a reminder for the target MSE, the ideal performance of the proposed methods and MSE_{obs} were compared to the ideal target MSE estimate. This is estimated by applying a prediction model, developed using all data, to the fully-observed data in the larger test set to get an MSE estimate, $MSE_{target,obs}$ (Section 3.6). The pragmatic performance of the proposed methods is compared to applying a prediction model, developed using all data, to the imputed datasets of the larger test set ($MSE_{target,imputed}$). The complete-case estimate of the MSE is obtained from applying a prediction model to the observed cases of the larger test set ($MSE_{target,CC}$). Figure 6.13 displays results for comparing the various methods MSE estimate with their respective ideal, pragmatic or CC target MSE. Graphs from all scenarios are available in the supplementary plot section S1.1.5.

MCAR and covariate-dependent MAR

Figure 4.6 presents results for comparing MSE estimates to the target MSE ($MSE_{imp} - MSE_{target}$) when data are weakly covariate-dependent MAR. The results are similar for MCAR and strong covariate-dependent MAR.

For all scenarios when $R^2 = 0.01$ or for $R^2 = 0.3$ when the sample size is 100 or 300, MSE_{obs} tends to approximate the MSE performance in the fully-observed larger test set. In all other scenarios, MSE_{obs} tends to under- or overestimate $MSE_{target,obs}$. For low or high values of R-squared ($R^2 = 0.01, 0.3$) and for a sample size of 100 (or 300 when $R^2 = 0.3$), the complete-case analysis estimate tends to approximate the complete-case target estimate ($MSE_{CC} - MSE_{target,CC}$). For all other scenarios, the complete-case method tends to either over- or underestimate $MSE_{target,CC}$.

For $R^2 = 0.01$ and sample size of 100, the ideal performance MSE estimate of methods A, B, D and F tends to overestimate $MSE_{target,obs}$ ($MSE_{imp,ideal} - MSE_{target,obs}$ for $imp = A, B, D, F$). All other imputation methods tend to overestimate the target MSE but their 95% confidence intervals overlap with zero. With increasing sample size, all of the proposed methods have similar ideal performance when compared to $MSE_{target,obs}$, except methods C and D which continue to overestimate. For low R-squared and small sample size, the pragmatic performance MSE estimate of method B tends to overestimate $MSE_{target,imputed}$. With increasing sample size, all methods have similar pragmatic performance when compared to $MSE_{target,imputed}$ ($MSE_{imp,prag} - MSE_{target,imputed}$).

For $R^2 = 0.1$ and sample size is 300 or 1000, all ideal performance estimates underestimate $MSE_{target,obs}$ and the pragmatic performance of the proposed methods underestimates $MSE_{target,imputed}$. For a sample size of 100, the ideal performance estimate of methods A and F closely approximates $MSE_{target,obs}$. The ideal performance of method B, C, D and

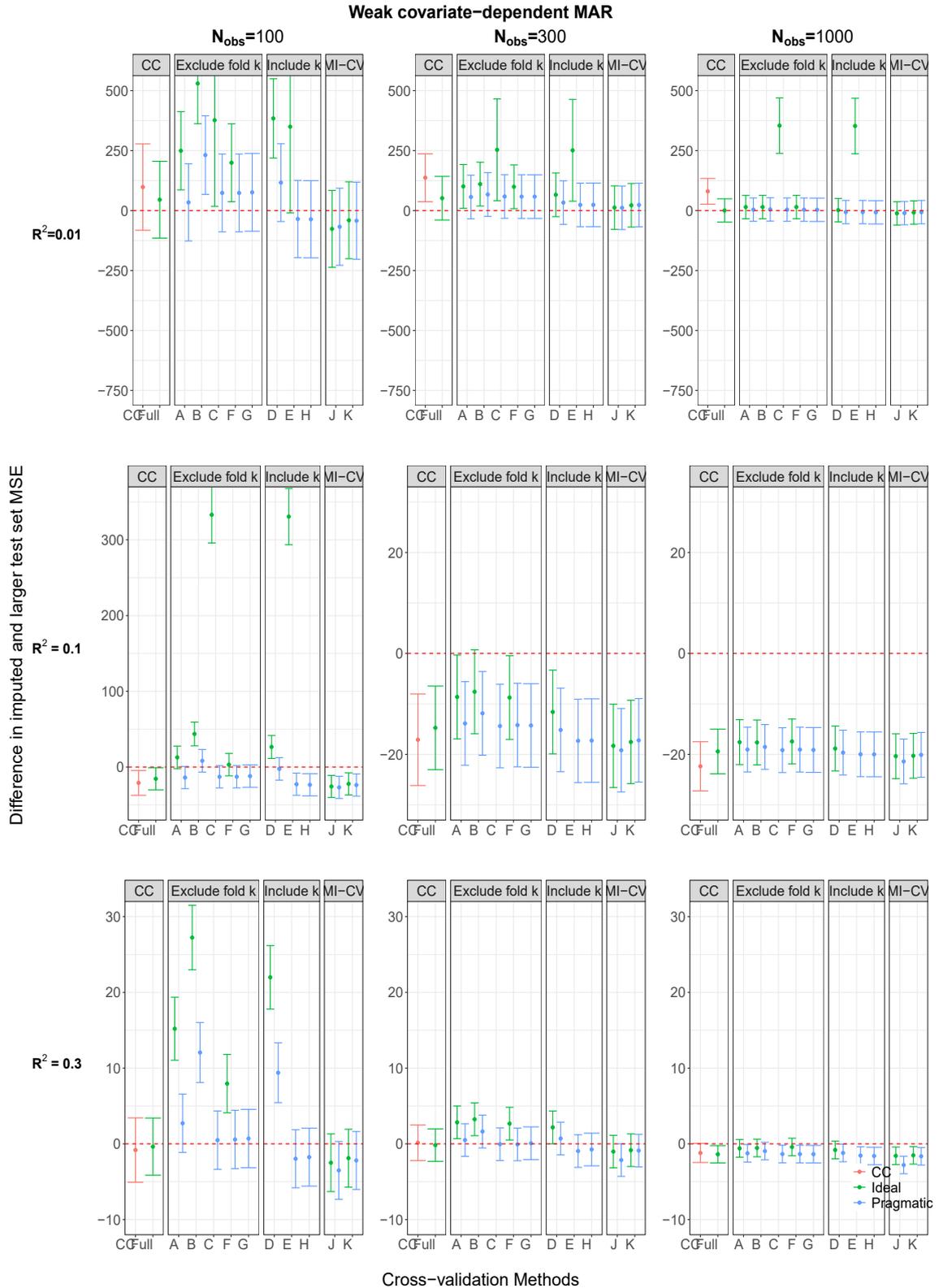


Figure 4.6: The difference $MSE_{imp} - MSE_{target}$ when data are weakly covariate-dependent MAR for $M = 5$ when 25% of values are missing in X_1 . The error bars summarise results from the 2000 repetitions and the limits represent the Monte Carlo 95% confidence interval of $MSE_{imp} - MSE_{target}$. CC (complete-case); methods A-K are described in Table 2.3 and summarised in Table 4.4.

E tend to overestimate $MSE_{target,obs}$ while all other methods tend to underestimate their respective target MSE. For a sample size of 100, the pragmatic performance of methods A, B, C, D, F and G approximates $MSE_{target,imputed}$. The pragmatic performance of method E, H, J and K tend to underestimate the target MSE in the fully-observed large test set while all other methods tend to underestimate $MSE_{target,imputed}$. With increasing sample size, all methods tend to perform similarly when compared to the various target estimates.

For $R^2 = 0.3$ and sample size of 100, the ideal performance for methods A-F tend to overestimate $MSE_{target,obs}$. The ideal performance of methods J and K tends to underestimate $MSE_{target,obs}$ but the 95% confidence intervals overlap with zero. The pragmatic performance of method B and D tend to overestimate $MSE_{target,imputed}$. All other methods confidence intervals overlap with zero with the mean point estimate of the pragmatic performance of methods A,C,F and G overestimating $MSE_{target,imputed}$ and the point estimate of E,H,J and K underestimating. With increasing sample size the performance of all methods tends to perform similarly when compared to the various target estimates.

Outcome-dependent MAR

Figure 4.7 presents results for comparing MSE estimates to the target MSE ($MSE_{imp} - MSE_{target}$) when data are weakly outcome- and covariate-dependent MAR. When data are outcome-dependent MAR and $R^2 = 0.1$ all methods tend to overestimate their respective target MSE. The MSE estimate when data are fully-observed or for the various proposed methods tends to overestimate $MSE_{target,obs}$ for the various scenarios.

When sample size is 300 or 1000, all methods tend to perform similarly when compared to the various target MSE estimates across the various scenarios, excluding the ideal performance of methods C and E which overestimates $MSE_{target,obs}$. When sample size is 100, the pragmatic performance of methods A, C, F and G are similar when compared to $MSE_{target,imputed}$, while methods B and D tend to have a larger magnitude of the difference. The ideal performance of methods A and F tend to be similar when compared to $MSE_{target,obs}$, while methods B and D have a larger magnitude of the difference. The ideal and pragmatic performance of methods J and K tend to approximate $MSE_{target,obs}$ or $MSE_{target,imputed}$ well.

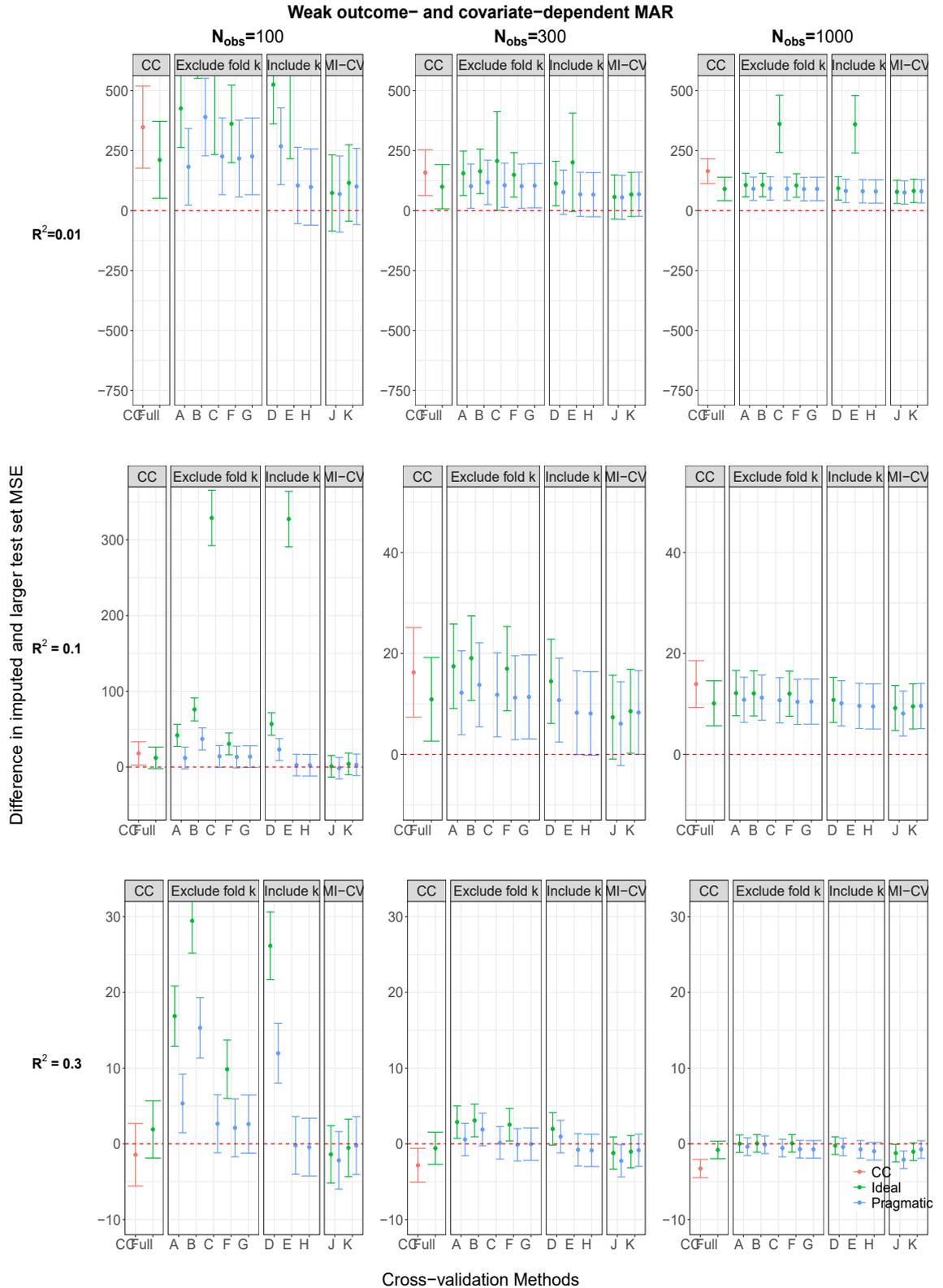


Figure 4.7: The difference $MSE_{imp} - MSE_{target}$ when data are weakly outcome- and covariate-dependent MAR for $M = 5$ when 25% of values are missing in X_1 . The error bars summarise results from the 2000 repetitions and the limits represent the Monte Carlo 95% confidence interval of $MSE_{imp} - MSE_{target}$. CC (complete-case); methods A-K are described in Table 2.3 and summarised in Table 4.4.

4.6 Is data leakage an issue within the imputation process?

In section 2.8I discussed the issue of data leakage in the imputation process and how we could investigate the impact of this leakage by comparing several methods, which are briefly re-summarised in Table 4.4. The methods to compare data leakage range from having no leakage (method B) to those with the highest amount of leakage (method J). Methods A (which has no leakage) and F-H have no similar methods from which to compare the inclusion and exclusion of folds to assess the impact of data leakage and, therefore, will not be discussed here.

Table 4.4: Brief summary of methods B-E and J-K for combining multiple imputation and cross-validation

| Method | Train imputations | Test imputations |
|--------|---|--|
| B | Y, X_2 used to impute $k - 1$ training folds by themselves | k^{th} fold imputed by itself using X_2 and possibly Y |
| C | Same as B | k^{th} fold imputed using all K folds and including X_2 and possibly Y |
| D | Y, X_2 used to impute $k - 1$ training folds using all K folds and restricting to $k - 1$ folds after imputation process | Same as B |
| E | Same as D | Same as C |
| J | Impute the dataset first using one set of imputations before applying cross-validation to each imputation set | |
| K | Impute the dataset first using two set of imputations - one for training the model on $k - 1$ folds and the other for evaluating the model in the k^{th} fold | |

Figure 4.8 compares the methods summarised above when data are weakly outcome- and covariate dependent MAR with high R^2 . Similar methods, with one method having a higher amount of data leakage, are compared side-by-side.

For sample size of 100 and $R^2 = 0.01$, the difference between the ideal and pragmatic performance of method B and MSE_{obs} is large and does not fit onto the scale of Figure 4.8 (refer to Figures 4.2 and 4.3). When sample size is 100 and for various R^2 values, the pragmatic performance of method B ($MSE_{B,prag} - MSE_{obs}$) is twice the difference of method C ($MSE_{C,prag} - MSE_{obs}$). This is similarly seen when comparing method D with E suggesting that using all the covariate data from all folds to impute the missing data in the holdout k^{th} fold has a strong impact on evaluating model performance. This difference can be seen for a moderate sample size of 300 but for a sample size of 1000 methods B and C perform similarly, as do D and E. For a sample size of 1000 each holdout fold contains 100 observations, of which approximately 75% are fully-observed. Any influence from

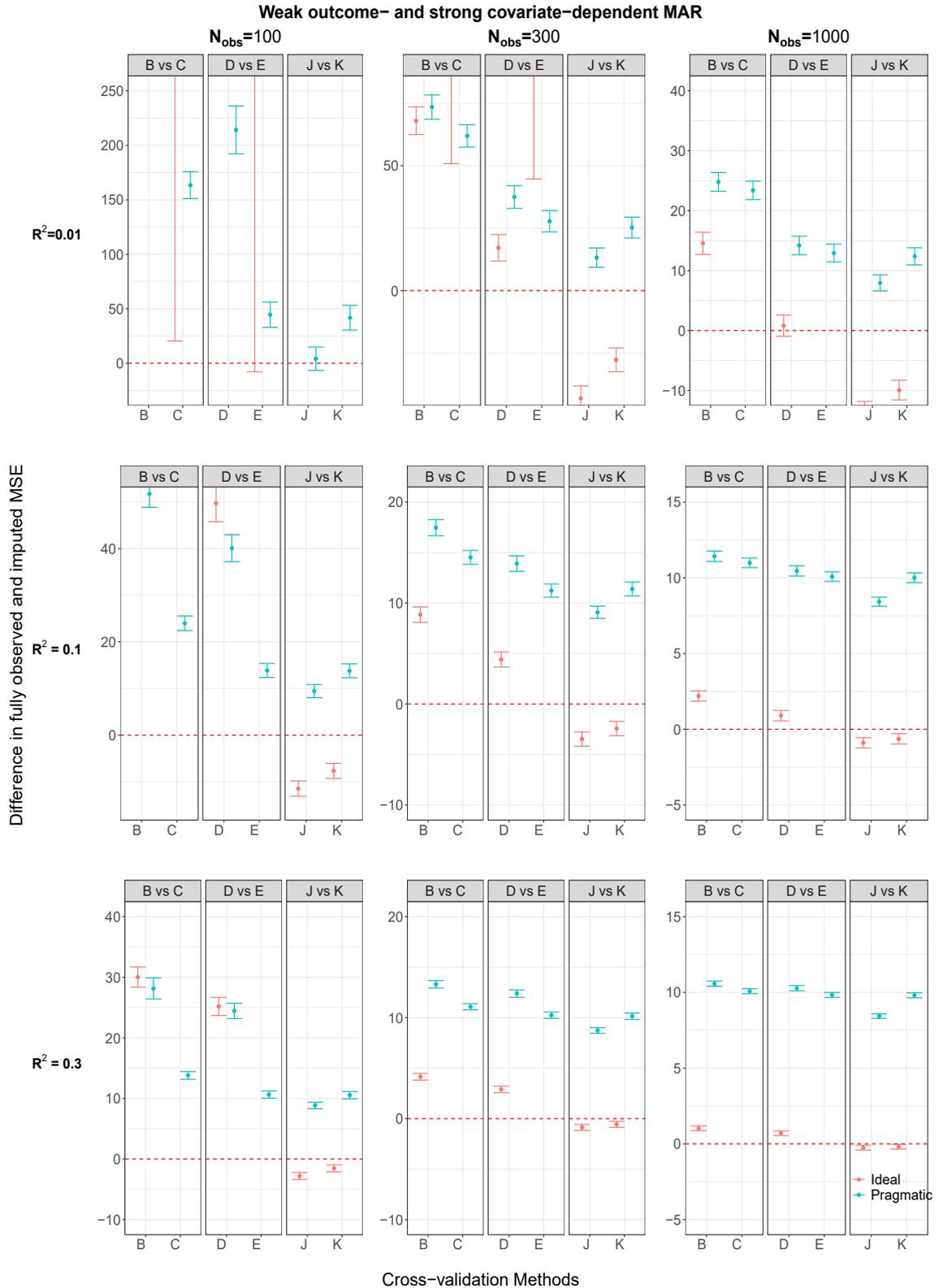


Figure 4.8: Assessing data leakage within the imputation process for cross-validation. The difference $MSE_{imp} - MSE_{obs}$ is compared when data are weak outcome- and strong covariate-dependent MAR. For $R^2 = 0.01, 0.1$ and 0.3 , the average MSE when data are fully-observed is approximately 17,800, 1600 and 400, respectively. CC (complete-case); methods A-K are described in Table 2.3 and summarised in Table 4.4.

the leakage of the additional 90% of the data appears to be minimal.

By comparing B with D it is possible to assess the leakage relevant to including observations from the test fold when drawing imputed values for the $k - 1$ training folds. For all sample sizes, we can see that D has a smaller difference ($\text{MSE}_{D,prag} - \text{MSE}_{obs}$) than method B ($\text{MSE}_{B,prag} - \text{MSE}_{obs}$). Although, the magnitude of this is not as large as when comparing B with C when $n_{obs} = 100$, with increasing sample size C and D perform similarly.

This leakage can also be seen for ideal performance when comparing methods B and D. However, this is not the case when comparing B with C for ideal performance. The difference between C's MSE estimate for ideal performance and the fully-observed MSE ($\text{MSE}_{C,ideal} - \text{MSE}_{obs}$) was approximately 350 across the three sample sizes. Recall that for imputing the holdout fold for method C the outcome was set to missing in the other $k - 1$ folds and was therefore imputed alongside X_1 in order to avoid any leakage by associating the outcome in the training folds with the test fold. By doing this, only Y in the holdout fold is available and the outcome Y in the $k - 1$ folds needs to be imputed, meaning that Y is missing for 90% of observations when performing imputation. This was found to introduce a large amount of bias and uncertainty into the ideal performance MSE estimates compared to their pragmatic version and is therefore not a recommended way to approach ideal performance in cross-validation.

Method E imputes the training folds using data from all K folds before restricting to the $k - 1$ folds to fit the prediction model. Similarly, all K folds are used to impute the dataset before restricting to the observations in the k^{th} test fold so, in total, two imputation sets are used. However, if ideal performance is of interest, the outcome in the $k - 1$ folds is excluded from the imputation of the k^{th} test fold. This process is repeated for $k = 1, \dots, K$. Method K is essentially the same as method E but all Y observations are used for imputing the test set. Method E and K perform similarly for pragmatic performance. However, removing values of Y in the training folds for ideal performance has caused method E to perform poorly by overestimating the MSE, similarly to method C. By using all available covariate and outcome data the ideal performance for method K has performed similarly to MSE_{obs} .

For the ideal scenario, methods J and K have a tendency to underestimate the MSE for all sample sizes across all values of R-squared. Method J has a tendency to have more optimistic performance than method K but with increasing sample size and R-squared they tend to perform comparably. This is also seen within the pragmatic version of the methods. While both methods J and K are impacted by data leakage due to the imputation step being performed first in these methods, method J is more prone to it than method K. For

ideal performance, the imputation model for both method J and K involves covariate X_2 and the outcome. However, method K involves using both a training imputation model to fit models and a test imputation model to evaluate the fitted prediction model. Whereas method J uses one set of imputations to both train and evaluate prediction models. We can see that by using two imputation models, method K tends to be less optimistic than method J.

Figure 4.9 compares methods in terms of whether they are subject to data leakage when data are outcome- and covariate- dependent MAR when compared with a larger test set ($\text{MSE}_{imp} - \text{MSE}_{target}$). With increasing sample size, all methods tend to perform similar when compared to the target MSE estimate. When sample size is 100 or 300, the methods with the most amount of data leakage tend to have a difference closer to zero ($\text{MSE}_{imp} - \text{MSE}_{target} \rightarrow 0$). In other missing data and R^2 scenarios, the methods with the most amount of data leakage tend to have the highest magnitude of the difference with MSE_{target} .

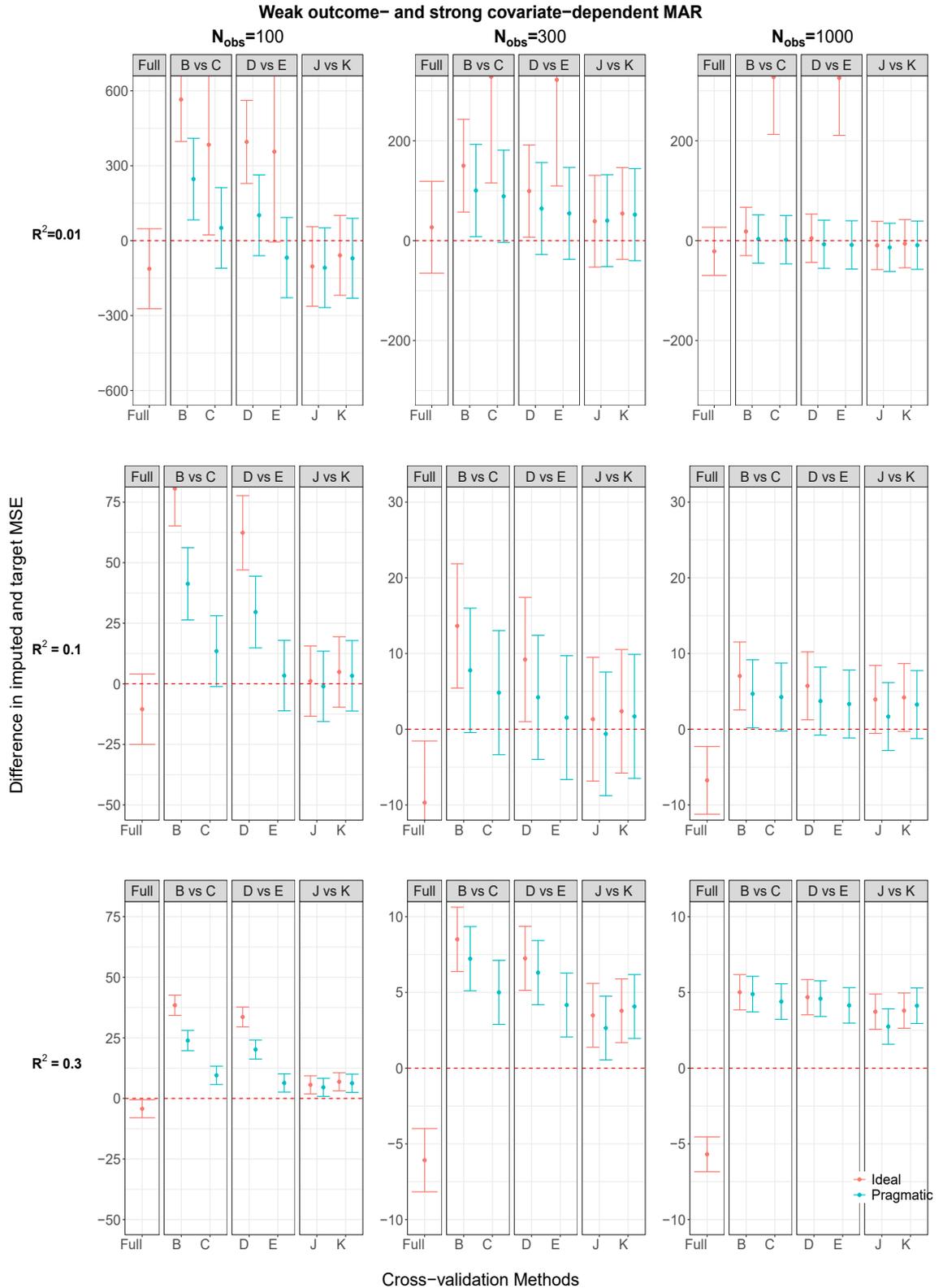


Figure 4.9: Assessing data leakage within the imputation process for cross-validation. The difference $MSE_{imp} - MSE_{target}$ is compared when when $R^2 = 0.1, 0.3$ and data are weak outcome- and strong covariate-dependent MAR. Method results are compared to estimates from a larger test set. CC (complete-case); methods A-K are described in Table 2.3 and summarised in Table 4.4.

4.7 Discussion of results for continuous outcome

The aim of this simulation study was to identify the most appropriate way to combine MI and cross-validation by using a simulation study covering a range of scenarios. In general, as the amount of data leakage increased across the methods the smaller the difference between imputed and fully-observed MSE tended to become. In certain scenarios there was a tendency for the methods with the most data leakage (*MI-then-CV* methods J and K) to underestimate the MSE.

Methods A and B are methods with no data leakage present and, with increasing sample size, performed similarly to those which had the advantage of data leakage. Method A imputes each fold separately using a train and test imputation model which can lead to a total of $2K^2M$ imputed datasets. Method B imputes the $k - 1$ training folds and k^{th} test fold separately which produces $2KM$ imputation datasets. While method A is more computationally intensive than B, this only added on an extra one to two hours of computational time in general across all scenarios on a high performance cluster. Imputing each fold separately appears to produce better results than imputing $k - 1$ folds together for small sample sizes of 100. Each fold to be imputed contains 10 observations, of which two or three are missing values. It is possible that the imputed values based on seven or eight fully-observed rows will be more variable than method B which has, on average, 68 fully-observed rows out of 90 observations. The variability in imputations for method A may lead to a more robust training model which is better at predicting observations in the test set than method B. The test set has also been imputed based on approximately 7 fully-observed rows out of 10 and therefore the imputed values will be more variable than the imputed values in the training set of method B.

I found that methods that are subject to more data leakage (methods C-E, J-K) tended to result in a smaller difference between the estimated MSE and MSE_{obs} , compared with methods that are not subject to leakage (methods A and B). An exception to this was the ideal performance of methods C and E. With both increasing sample size and R^2 , the difference in MSE_{imp} and MSE_{obs} decreased for their pragmatic versions but remained high for ideal performance. For both methods, in an effort to avoid data leakage through including the outcome from the training folds, Y in the $k - 1$ folds was set as missing. Therefore, for ideal performance both X_1 in all K folds and Y in the $k - 1$ folds needed to be imputed. However, imputing Y when only 10% of the values of Y have been observed is typically not advisable as seen by the large amount of over-estimation of the MSE compared to other methods.

As all methods will be further explored for a binary outcome scenario in the next chapter, I will not yet make any recommendations as more exploration is needed. Chapter 5 will

assess the simulation study results for the various cross-validation methods combined with MI when the outcome is binary.

5 Simulation study results for cross-validation: binary outcome

5.1 Introduction

Several performance measures were evaluated in the binary outcome setting: AUC, Brier score and ‘weak calibration’ which were originally described in Section 1.10. Similarly to the continuous outcome scenario, various sample sizes and levels of missingness were examined for the binary outcome case. For a quick reminder on how results will be compared, a brief overview of the analysis comparing methods B and J is available in Section 4.3 when the performance measure of interest is the MSE. All methods were previously described in Section 2.6 and summarised in Table 4.3.

For each performance measure, results for the complete-case analysis when sample size is 100 are available for at least 1920 repetitions. Thirty percent of the observations have the outcome and approximately 25% of the data have a missing value. A complete-case analysis resulted in records with the outcome present being removed from a fold which lead to difficulties in obtaining results.

5.2 Summary of the simulated fully-observed data

Table 5.1 presents a summary of the AUC, Brier score and calibration in the fully-observed data, before missingness is introduced. For the AUC, a value approaching 1 indicates good performance whereas for Brier score, a smaller value is preferred. For all sample sizes the mean AUC is approximately 0.78 and the mean Brier score is 0.17. Calibration is assessed using the calibration intercept and slope (Section 1.10). Large deviations from zero or one for the intercept and slope, respectively, can indicate poor calibration. For $N = 100$ the intercept and slope vary massively, with the slope still having some variation for $N = 300$. These unstable results will be discussed in Section 5.8. When sample size equals 1000, the calibration intercept and slope tend towards zero and one.

Table 5.1: Summarising performance when data are fully-observed for the 2000 simulated datasets

| | AUC | Brier | Intercept | Slope |
|------------|----------------|----------------|----------------------|-------------------|
| $N = 100$ | | | | |
| Mean (var) | 0.79 (0.003) | 0.17 (< 0.001) | -1.35e+11 (3.65e+25) | 80.94 (8.93e+05) |
| (Min, Max) | 0.63, 0.95 | 0.10, 0.25 | -2.70e+14, 0.13 | -556.91, 39710.70 |
| $N = 300$ | | | | |
| Mean (var) | 0.78 (< 0.001) | 0.17 (< 0.001) | -0.03 (< 0.001) | 1.80 (60.31) |
| (Min, Max) | 0.66, 0.88 | 0.12, 0.20 | -0.17, 0.02 | 0.91, 196.09 |
| $N = 1000$ | | | | |
| Mean (var) | 0.78 (< 0.001) | 0.16 (< 0.001) | -0.01 (< 0.001) | 1.04 (0.00) |
| (Min, Max) | 0.73, 0.84 | 0.14, 0.18 | -0.03, 0.01 | 0.99, 1.15 |

5.3 Detailed results: Area under the ROC curve

A higher AUC estimate generally suggests the model is performing well. Therefore, if a method overestimates the AUC estimated when data are fully-observed, AUC_{obs} , the method is considered to over-optimistic i.e. the model performs better when data have been imputed than if the data had not been missing to begin with.

5.3.1 Comparing the methods' AUC to the estimate of the AUC when data are fully-observed

MCAR and covariate-dependent MAR

Figure 5.1 displays results for the various cross-validation methods' (imp) estimates of the AUC which are compared to AUC_{obs} ($AUC_{imp} - AUC_{obs}$) when data are MCAR or covariate-dependent MAR.

When sample size is 100 and data are MCAR or covariate-dependent MAR, the complete-case analysis tends to overestimate the AUC_{obs} ($AUC_{CC} - AUC_{obs} > 0.01$). For the MCAR scenario, with increasing sample size the difference decreases to zero ($AUC_{CC} - AUC_{obs} \xrightarrow{n_{obs} \rightarrow \infty} 0$). For the covariate-dependent MAR scenarios when the sample size is 300 or 1000, the complete-case analysis estimate of the AUC tends to underestimate AUC_{obs} ($AUC_{CC} - AUC_{obs} < 0$).

The pragmatic performance of all methods underestimates AUC_{obs} ($AUC_{imp,prag} - AUC_{obs} < 0$). Similarly to the continuous outcome case, method A (impute each fold separately) has a smaller difference than method B (impute the k^{th} test fold separately to the $k - 1$ training folds) i.e. $|AUC_{A,prag} - AUC_{obs}| < |AUC_{B,prag} - AUC_{obs}|$. This can be observed for all sample sizes when data are MCAR or covariate-dependent MAR. This was also noted for the MSE when the outcome is continuous. Method B has the largest magnitude $|AUC_{B,prag} - AUC_{obs}|$ across all imputation methods while method J (impute all K folds together using one set of imputed datasets) tends to have the smallest magnitude of the difference. The pragmatic performance of methods C, F and G, is similar in relation to AUC_{obs} . Their magnitude ($|AUC_{imp,prag} - AUC_{obs}|$, $imp = C, F, G$) is smaller than the magnitude of method B but larger than the magnitude of method A when the sample size is 100 or 300. All methods tend to perform similarly when the sample size is 1000.

The ideal performance of *CV-then-MI* methods A-H underestimates AUC_{obs} for all sample sizes when data are MCAR or covariate-dependent MAR ($AUC_{imp,ideal} - AUC_{obs} < 0$, $imp = A-H$). Whereas the ideal performance of methods J and K (*MI-then-CV*) tends to overestimate AUC_{obs} ($AUC_{imp,ideal} - AUC_{obs} > 0$, $imp = J, K$).

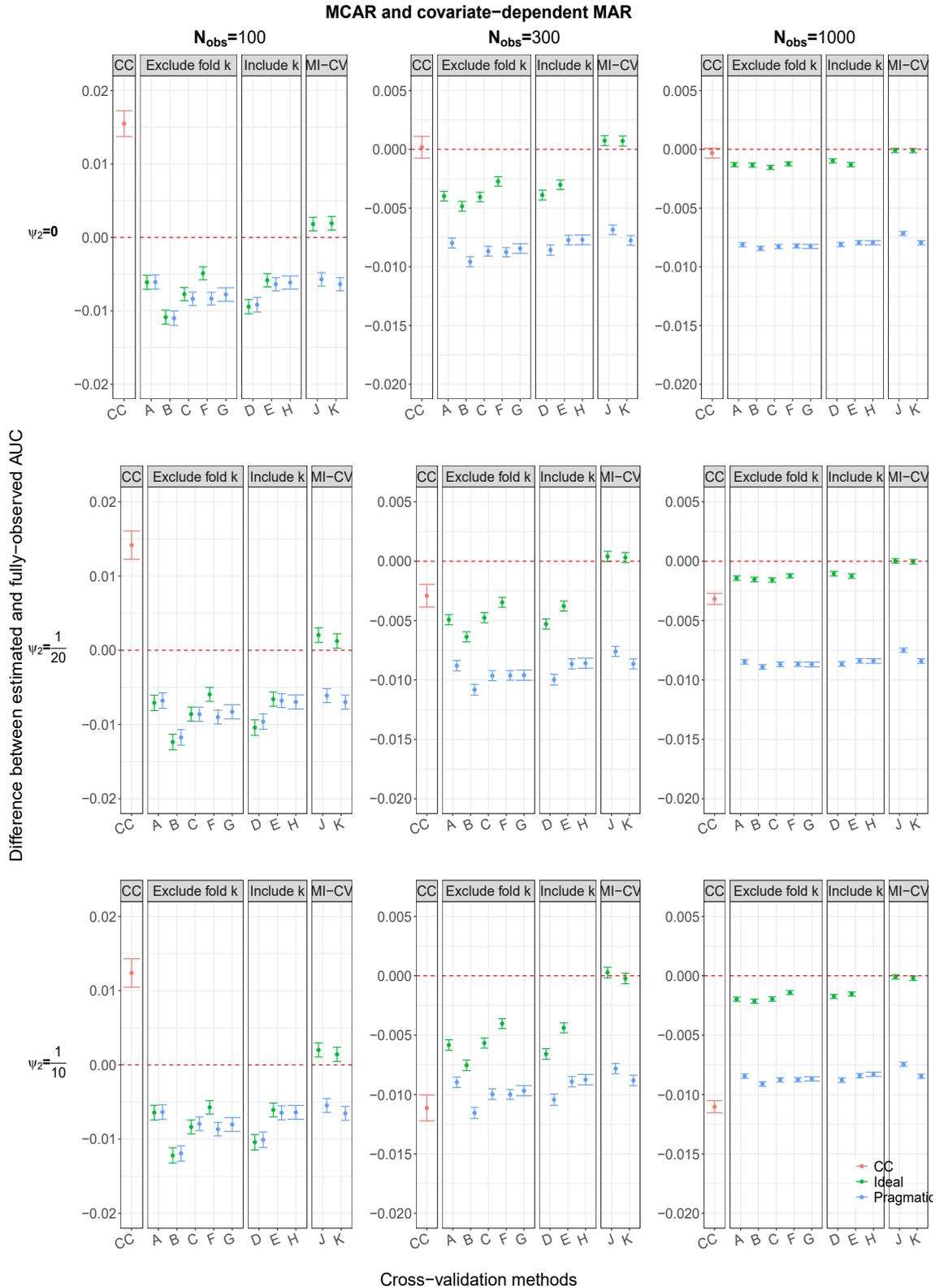


Figure 5.1: The difference $AUC_{imp} - AUC_{obs}$ when data are MCAR or covariate-dependent MAR for $M = 5$ when 25% of values are missing in X_1 . The error bars summarise results from the 2000 repetitions and the limits represent the Monte Carlo 95% confidence interval of $AUC_{imp} - AUC_{obs}$. The average AUC when data are fully-observed is 0.78. CC (complete-case); methods A-K are described in Table 2.3 and summarised in Table 4.4.

Across all sample sizes, the magnitude of the difference ($|AUC_{imp,ideal} - AUC_{obs}|$) for methods J and K tends to be the smallest while method B tends to have the largest. For a sample size of 100, the ideal performance of methods A, F and E perform similarly and have the smallest magnitude of difference for all *CV-then-MI* methods. With an increase in sample size to 300 methods E and F have the smallest magnitude of difference ($|AUC_{imp,ideal} - AUC_{obs}|$, $imp = E, F$) while method A performs similarly to methods C and E. With an increase in sample size to 1000 all *CV-then-MI* methods (methods A-H) tend to perform similarly with a magnitude less than 0.0025 ($|AUC_{imp,ideal} - AUC_{obs}| < 0.0025$).

Outcome-dependent MAR

Figure 5.2 displays results for the various cross-validation methods' (imp) estimates of the AUC which are compared to AUC_{obs} ($AUC_{imp} - AUC_{obs}$) when data are outcome-dependent or outcome- and covariate-dependent MAR.

Similarly to the MCAR and covariate-dependent MAR scenarios, the complete-case analysis tends to overestimate AUC_{obs} for a sample size of 100 ($AUC_{CC} - AUC_{obs} > 0$). When data are outcome-dependent MAR and sample size is 300 or 1000 the complete-case analysis estimates AUC_{obs} well. When data are outcome- and covariate-dependent MAR and sample size is 300 or 1000, the complete-case analysis underestimates the AUC value ($AUC_{CC} - AUC_{obs} < -0.01$).

The pragmatic performance of all methods underestimates AUC_{obs} ($AUC_{imp,prag} - AUC_{obs} < 0$, $imp = A-H, J, K$). For all sample sizes and missing data scenarios, the pragmatic performance of method B has the largest magnitude of the difference ($|AUC_{B,prag} - AUC_{obs}|$), with method D having the second largest magnitude. Method B (impute the training folds using the $k-1$ folds only) has a larger magnitude than method D (impute the training folds using all K folds before restricting to the $k-1$ folds to fit the prediction model). Method B imputes the test fold using only data available in the k^{th} fold and is outperformed by method D which uses all K folds to impute the test fold before restricting to the data in the k^{th} fold to evaluate the prediction model. Method A performs similarly to methods C, F and G while method J tends to have the smallest magnitude of all the methods ($|AUC_{J,prag} - AUC_{obs}|$) for all sample sizes. With increasing sample size the magnitude of the difference decreases for all sample sizes and the methods tend to perform similarly in relation to AUC_{obs} when the sample size is 1000.

The ideal performance of methods A-H underestimate AUC_{obs} ($|AUC_{imp,ideal} - AUC_{obs}| < 0$, $imp = A, \dots, H$) while methods J and K tend to overestimate AUC_{obs} when sample size is 100 or 300 ($|AUC_{imp,ideal} - AUC_{obs}| > 0$, $imp = J, K$) and underestimate AUC_{obs} for a sample size of 1000. Again, method B has the largest magnitude of difference for all sample sizes when data are outcome-dependent or outcome- and covariate-dependent

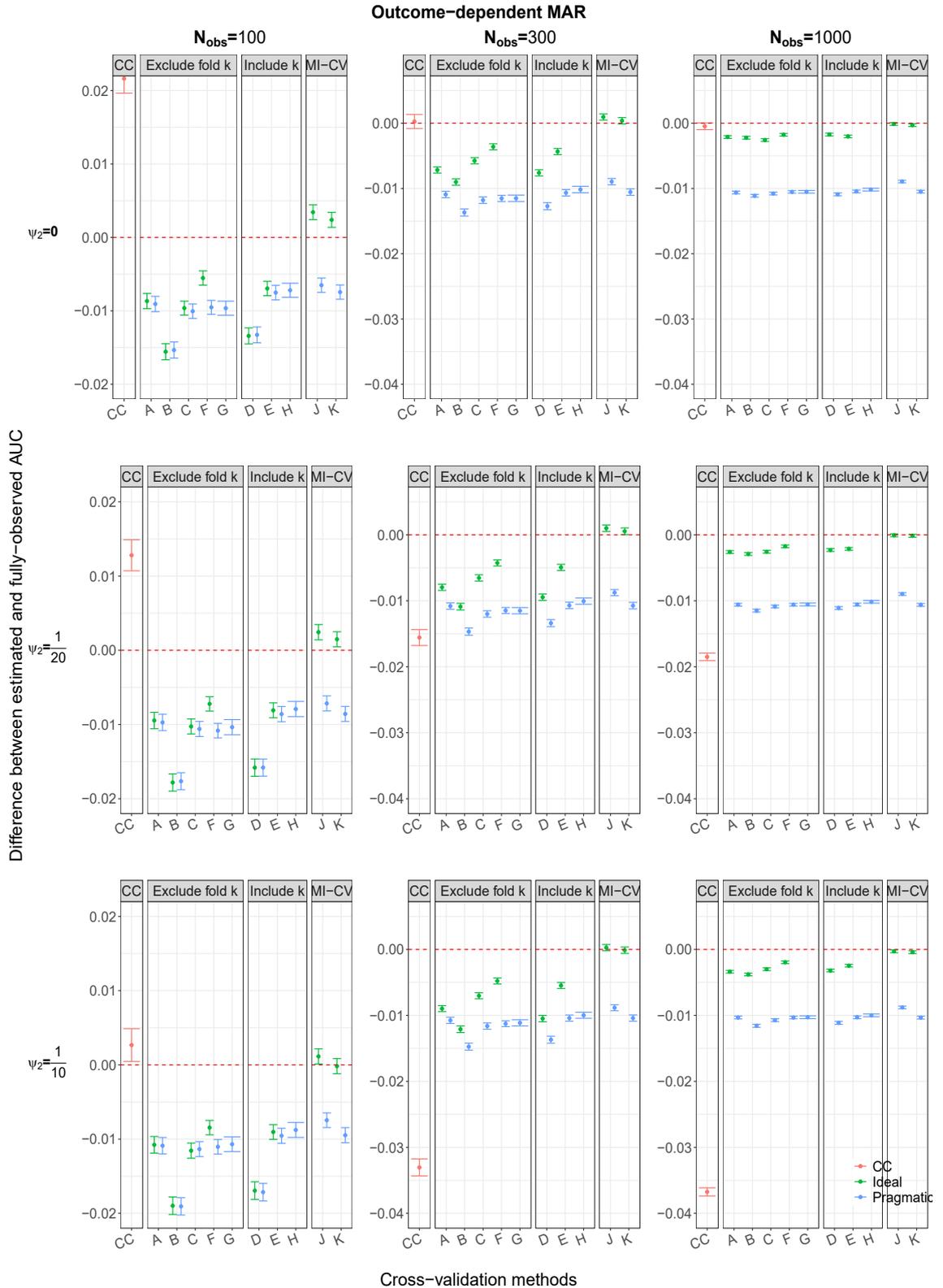


Figure 5.2: The difference $AUC_{imp} - AUC_{obs}$ when data are outcome-dependent or outcome- and covariate-dependent MAR for $M = 5$ when 25% of values are missing in X_1 . The error bars summarise results from the 2000 repetitions and the limits represent the Monte Carlo 95% confidence interval of $AUC_{imp} - AUC_{obs}$. The average AUC when data are fully-observed is 0.78. CC (complete-case); methods A-K are described in Table 2.3 and summarised in Table 4.4.

MAR. For all sample sizes and missing data scenarios, methods J and K have the smallest magnitudes of difference which are less than 0.005 ($|\text{AUC}_{imp,ideal} - \text{AUC}_{obs}| < 0.005$, $imp = J, K$).

The ideal performance of method D (observations from all folds are used to impute the $k - 1$ training folds) has a slightly smaller magnitude of the difference than method B (only observations from the $k - 1$ folds are used when imputing the training folds) for sample sizes of 100 and 300. The difference in magnitude between method D and B is due to the use of all folds in the training folds imputation process. This comparison can also be made for method E which has a smaller magnitude of the difference than method C.

Method C (observations from all folds used to impute the k^{th} test fold) has a much smaller magnitude than method B (only observations from the k^{th} test fold used to impute the test fold) for small and moderate sample sizes. This difference in magnitude due to the use of all folds in the test fold imputation process can also be seen when comparing method E (which uses all folds) to method D (which uses only the test fold).

For large sample sizes methods A-H (*CV-then-MI*) tend to perform similarly when compared to AUC_{obs} while methods J and K (*MI-then-CV*) have the smallest magnitude. When the sample size is 1000, the magnitude of the difference is less than 0.005 for all methods ($|\text{AUC}_{imp,ideal} - \text{AUC}_{obs}| < 0.005$).

5.3.2 Increasing the number of imputed datasets from 5 to 25

Figure 5.3 displays results comparing the use of 5 versus 25 imputed datasets (M) when data are outcome-dependent or outcome- and covariate-dependent MAR ($\text{AUC}_{imp,M} - \text{AUC}_{obs}$). The results are for the pragmatic performance but are generalisable also to the ideal performance in all missing data scenarios. All graphs comparing 5 versus 25 imputed datasets for the ideal and pragmatic performance are available in the Supplementary plots section S2.1.3.

Increasing the number of imputed datasets from 5 to 25 has had little impact on the methods' AUC estimates when compared to AUC_{obs} ($\text{AUC}_{imp,M=5} - \text{AUC}_{obs} \approx \text{AUC}_{imp,M=25} - \text{AUC}_{obs}$). This can be seen for the various cross-validation methods for all missing data and sample size scenarios.

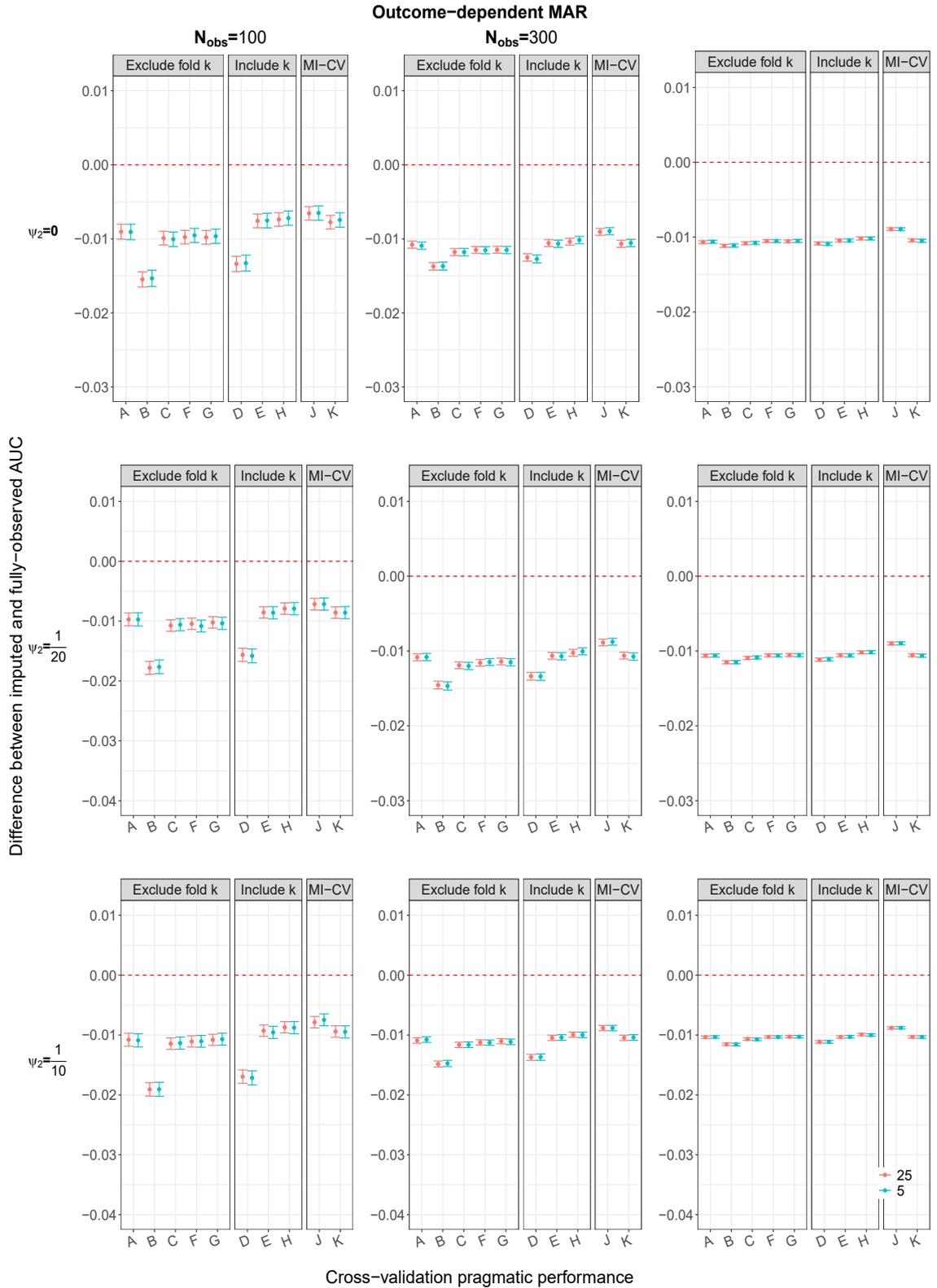


Figure 5.3: The difference $AUC_{imp} - AUC_{obs}$ when data are outcome-dependent or outcome- and covariate-dependent MAR for $M = 25$ versus $M = 5$ when 25% of values are missing in X_1 . The error bars summarise results from the 2000 repetitions for pragmatic performance and the limits represent the Monte Carlo 95% confidence interval of $AUC_{imp} - AUC_{obs}$. The average AUC when data are fully-observed is 0.78. CC (complete-case); methods A-K are described in Table 2.3 and summarised in Table 4.4.

5.3.3 Increasing the percentage of missingness to 40%

Figure 5.4 displays results demonstrating the impact that an increased percentage of missingness can have on the various cross-validation methods when data are weakly outcome- and covariate-dependent MAR. The figure presents the AUC estimates when 25% or 40% of X_1 values are missing compared to AUC_{obs} ($AUC_{imp,\%} - AUC_{obs}$). The results are generally representative of the comparison between 25% and 40% missingness for ideal and pragmatic performance for all missing data scenarios and sample sizes. All plots are available in Section S2.1.2 of the Supplementary Plots.

The complete-case analysis tends to perform similarly or have overlapping confidence intervals when 25% or 40% of the X_1 values are missing when data are MCAR or covariate-dependent MAR. When data are outcome-dependent MAR and sample size is 300 the larger percentage of missingness results in a larger magnitude than when 25% of values are missing ($|AUC_{CC,25\%} - AUC_{obs}| < |AUC_{CC,40\%} - AUC_{obs}|$). When the sample size is 1000 they both perform similarly when compared to AUC_{obs} . For weak outcome-dependent and weak or strong covariate-dependent MAR, the magnitude of the difference tends to be smaller for the higher percentage of missingness when sample size is 300 ($|AUC_{CC,40\%} - AUC_{obs}| < |AUC_{CC,25\%} - AUC_{obs}|$) but with increased sample size to 1000 this reverts to the higher percentage of missingness having a larger magnitude ($|AUC_{CC,25\%} - AUC_{obs}| < |AUC_{CC,40\%} - AUC_{obs}|$).

The pragmatic performance of methods A-F and H tend to have a larger magnitude of the difference between the methods' AUC estimates and AUC_{obs} for a sample size of 300 or 1000 for all missing data scenarios. Method G, J and K tend to have similar ($|AUC_{imp,25\%} - AUC_{obs}| \approx |AUC_{imp,40\%} - AUC_{obs}|$ where $imp = G, J$ or K).

The ideal performance of methods A-F (methods G and H do not have an ideal performance estimate) tend to have a larger magnitude when the percentage of missingness is 40% compared to a percentage of 25% ($|AUC_{imp,25\%} - AUC_{obs}| < |AUC_{imp,40\%} - AUC_{obs}|$ where $imp = A, \dots, F$) for all sample sizes and missing data scenarios. Similarly to the pragmatic performance, the ideal performance estimate of the AUC for methods J and K performs similarly regardless of the percentage of missing data present in variable X_1 .

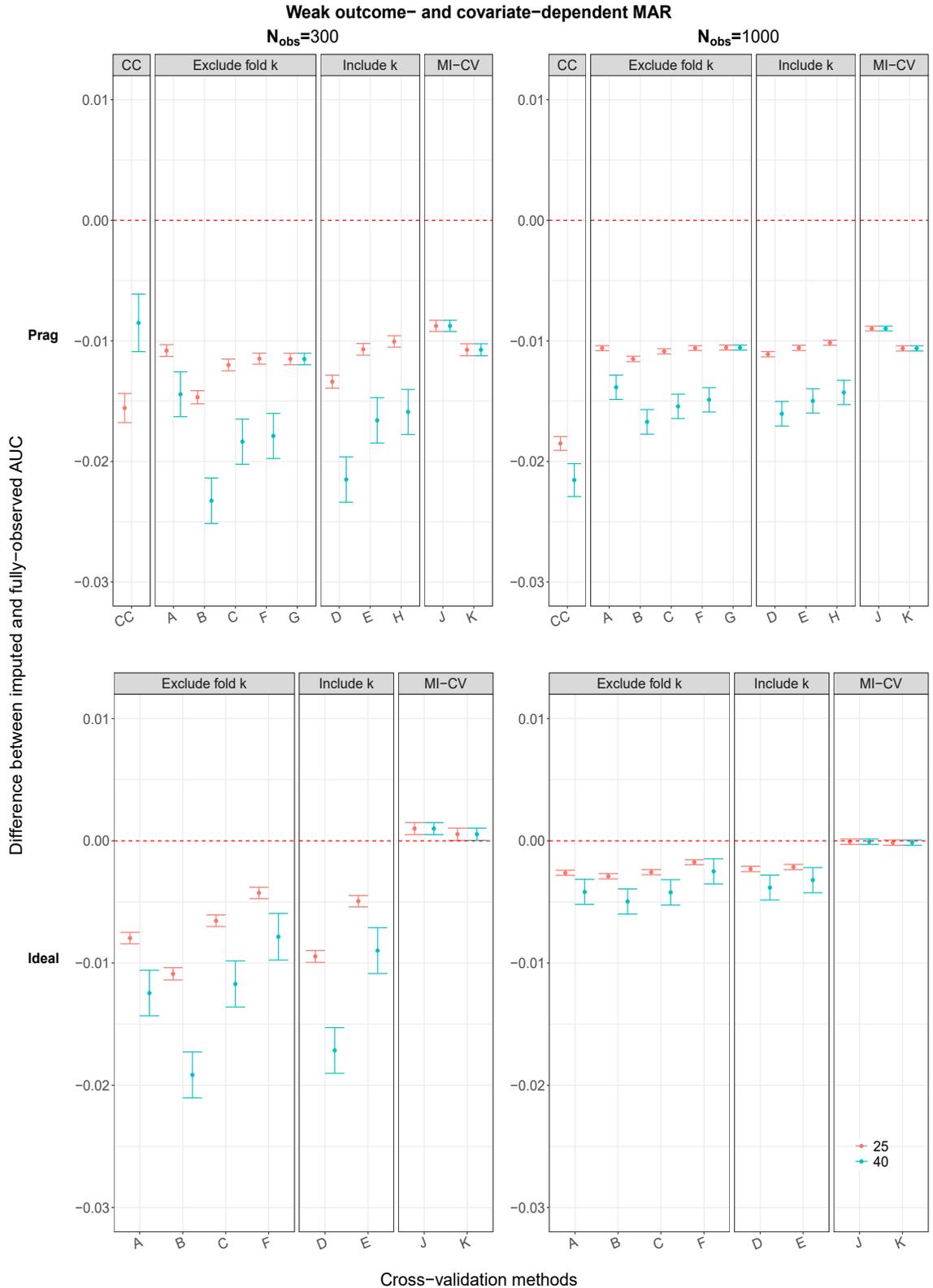


Figure 5.4: Comparing the impact of increasing the percentage of missingness on the difference $AUC_{imp} - AUC_{obs}$ when data are outcome- and covariate-dependent MAR when $M = 5$. The error bars summarise results from the 2000 repetitions and the limits represent the Monte Carlo 95% confidence interval of $AUC_{imp} - AUC_{obs}$. Red denotes $AUC_{imp} - AUC_{obs}$ when 25% of X_1 values are missing and blue denotes $AUC_{imp} - AUC_{obs}$ when 40% of X_1 values are missing. The top row presents the results for pragmatic performance and the bottom row presents results for ideal performance. The average AUC when data are fully-observed is 0.78. CC (complete-case); methods A-K are described in Table 2.3.

5.3.4 Comparing each method's AUC to the target estimate of the AUC from a larger validation set

Similarly to the continuous outcome scenario in Section 4.5, the ideal performance of the proposed methods and AUC_{obs} were compared to the ideal target AUC estimate, $AUC_{target,obs}$. This is estimated by applying a prediction model, based on all data in a repetition, to the fully-observed data in the larger test set. The pragmatic performance of the methods is compared to applying a repetition's prediction model to the imputed datasets of the larger test set ($AUC_{target,imputed}$). The complete-case estimate of the AUC is compared to applying a repetition's prediction model to the observed cases of the larger test set ($AUC_{target,CC}$).

MCAR and covariate-dependent MAR

Figure 5.5 presents the ideal and pragmatic performance estimates of the various methods when compared to their respective target AUC estimate, when data are MCAR or covariate-dependent MAR.

For a sample size of 100, the complete-case estimate tends to overestimate the complete-case target estimate of the AUC ($AUC_{CC} - AUC_{target,CC} > 0$) and does not fit onto the scale of Figure 5.5. With increasing sample size the magnitude of the difference between the complete-case analysis and $AUC_{target,CC}$ tends to be less than 0.005 ($AUC_{CC} - AUC_{target,CC} < 0.005$) and either under- or overestimates the target estimate.

When sample size is 100, the pragmatic performance of all methods tends to overestimate $AUC_{target,imputed}$ by approximately 0.02 ($AUC_{imp,prag} - AUC_{target,imputed} \approx 0.02$). When sample size is 100, method B, which tended to have the largest magnitude when compared to AUC_{obs} , now tends to have the smallest magnitude ($|AUC_{B,prag} - AUC_{target,imputed}|$). Methods J and K tend to have the largest magnitude when sample size is small, closely followed by methods A, E and H. With increasing sample size, the pragmatic performance of the methods tends to perform similarly with the magnitude of the pragmatic performance estimates tending to be less than 0.005 ($|AUC_{imp,prag} - AUC_{target,imputed}| < 0.005$) when sample size is 300 and less than 0.0025 when sample size is 1000. When data are MCAR or strong covariate-dependent MAR, the pragmatic performance of the methods tends to underestimate the pragmatic target AUC estimate, while the methods tend to overestimate the AUC target estimate when data are weak covariate-dependent MAR.

The ideal performance of all methods when sample size is 100 overestimates $AUC_{target,obs}$ ($AUC_{imp,ideal} - AUC_{target,obs} > 0$). When sample size is 100, method B tends to have the smallest magnitude of the difference while methods J and K have the largest. For a sample size of 300 there is a downwards pull for all methods, the *CV-then-MI* methods

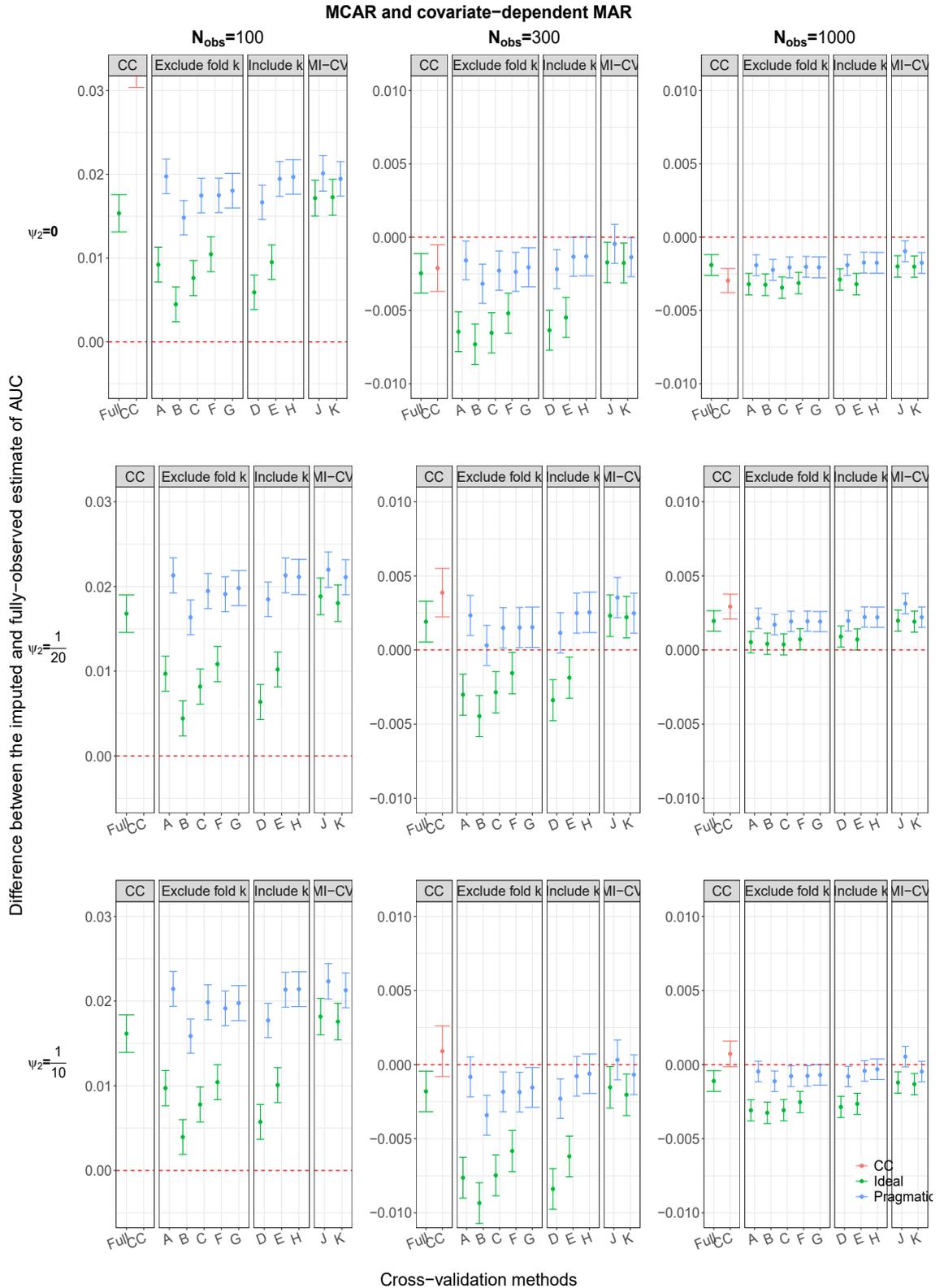


Figure 5.5: The difference $AUC_{imp} - AUC_{target}$ when data are MCAR or covariate-dependent MAR for $M = 5$ when 25% of values are missing in X_1 . The error bars summarise results from the 2000 repetitions and the limits represent the Monte Carlo 95% confidence interval of $AUC_{imp} - AUC_{target}$. The average AUC when data are fully-observed is 0.78. CC (complete-case); methods A-K are described in Table 2.3.

A-F now tend to underestimate the target estimate while the *MI-then-CV* methods tend to overestimate or approximate the target estimate well. With increasing sample size to 1000 all methods tend to perform similarly and the magnitude of the difference tends to be less than 0.005 for all methods ($|AUC_{imp,ideal} - AUC_{target,obs}| < 0.005$).

Outcome-dependent MAR

Figure 5.6 presents the ideal and pragmatic performance estimates of the various methods when compared to their respective target AUC estimate, when data are outcome-dependent or outcome- and covariate-dependent MAR.

The complete-case analysis estimate of the AUC tends to overestimate $AUC_{target,CC}$ for all sample sizes and missing data scenarios, at times not fitting onto the scale of the graph in Figure 5.6 when sample size is 100. With increasing sample size, the magnitude of the difference between the complete-case analysis estimate and the complete-case target estimate tends to decrease.

Similarly to the MCAR and covariate-dependent MAR scenario, the pragmatic performance of all methods tends to overestimate $AUC_{target,imputed}$ when sample size is 100 ($0.015 < AUC_{imp,prag} - AUC_{target,imputed} < 0.03$). Method B tends to have the smallest magnitude ($|AUC_{B,prag} - AUC_{target,imputed}|$) while methods J and K tend to have the largest. Increasing the sample size of the methods to 300, the magnitude of the difference for all methods tends to be less than 0.005 ($|AUC_{imp,prag} - AUC_{target,imputed}| < 0.005$). Methods A, C, E-H tend to perform well when sample size is 300, while methods B and D tend to have a slightly larger magnitude. For a sample size of 1000 all methods tend to perform similarly.

When sample size is 100, the ideal performance of all methods tends to overestimate $AUC_{target,obs}$. Method B tends to have the smallest magnitude ($|AUC_{B,ideal} - AUC_{target,obs}|$) while the ideal performance of the *CV-then-MI* methods A-F tend to have a smaller magnitude than the *MI-then-CV* methods, which tend to have a magnitude around 0.02. For a sample size of 300 the *CV-then-MI* methods' ideal performance underestimates $AUC_{target,obs}$ while the *MI-then-CV* methods overestimate $AUC_{target,obs}$ (i.e. they're over-optimistic). Methods B and D tend to have the largest magnitude while the other methods have a magnitude less than 0.0075. With increased sample size to 1000, all methods tend to perform similarly with a magnitude less than 0.005.

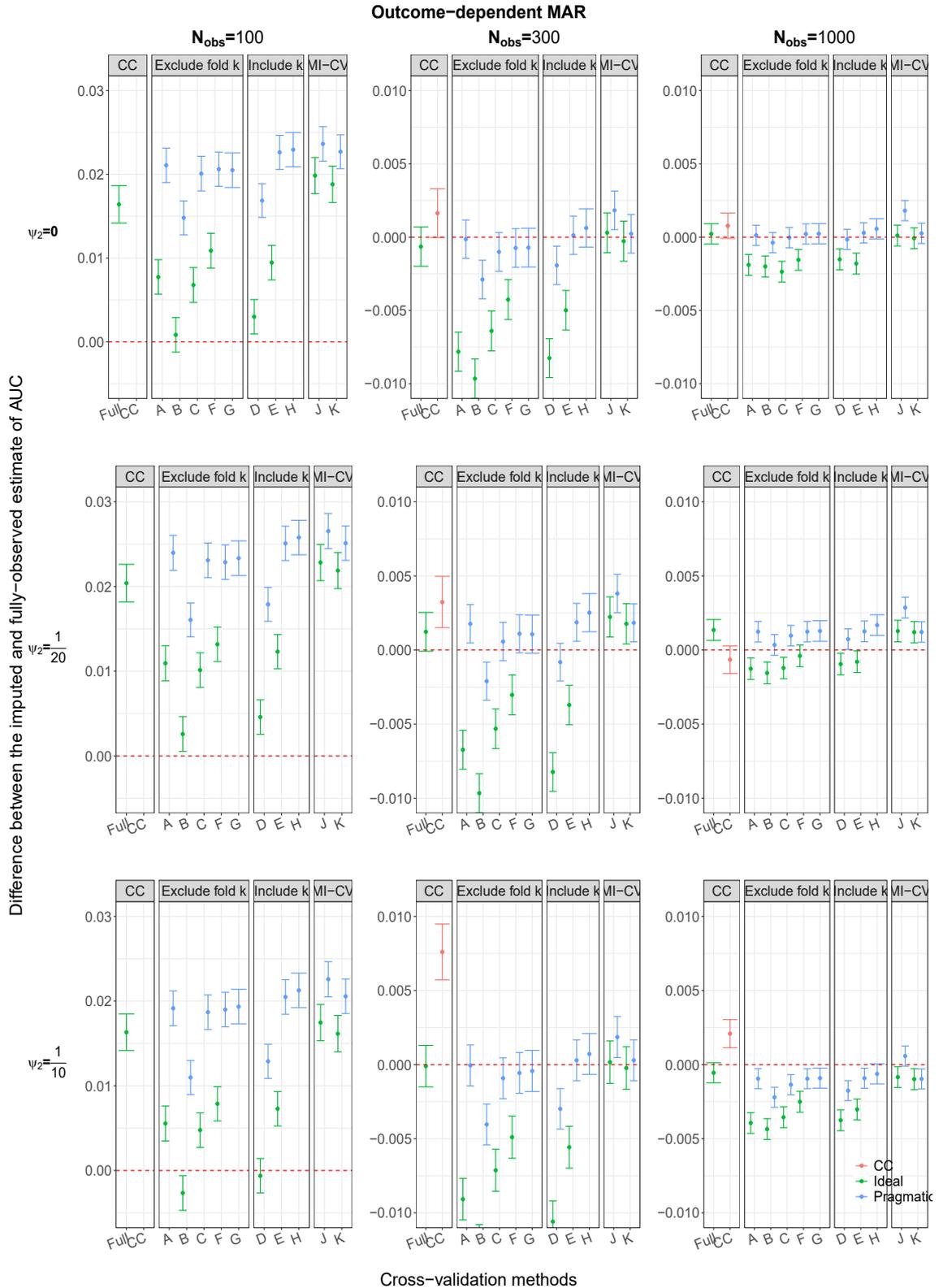


Figure 5.6: The difference $AUC_{imp} - AUC_{target}$ when data are outcome-dependent or outcome- and covariate-dependent MAR for $M = 5$ when 25% of values are missing in X_1 . The error bars summarise results from the 2000 repetitions and the limits represent the Monte Carlo 95% confidence interval of $AUC_{imp} - AUC_{target}$. The average AUC when data are fully-observed is 0.78. CC (complete-case); methods A-K are described in Table 2.3.

5.4 Detailed results: Brier score

A lower Brier score estimate generally suggests the model is performing well. Therefore, if a method underestimates the Brier score estimated when data are fully-observed the method is considered to be over-optimistic i.e. the model performs better when data have been imputed than if the data had not been missing to begin with.

5.4.1 Comparing each method's Brier score to the estimate of the Brier score when data are fully-observed

MCAR and covariate-dependent MAR

Figure 5.7 displays results for the various cross-validation methods' (*imp*) estimates of the Brier score which are compared to the Brier score estimate when data are fully-observed ($\text{Brier}_{imp} - \text{Brier}_{obs}$) when data are MCAR or covariate-dependent MAR.

When data are MCAR, the complete-case analysis estimate tends to overestimate Brier_{obs} . Increasing the sample size from 100 to 1000 causes the magnitude of this difference to decrease from 0.00122 to 0.00029. When data are covariate-dependent MAR, the complete-case analysis estimate underestimates Brier_{obs} with a magnitude of at least 0.005.

For all sample sizes when data are MCAR or covariate-dependent MAR, the pragmatic performance of all methods overestimates Brier_{obs} . For a sample size of 100 and 300 across all missing data scenarios, methods J and K have the smallest magnitude ($|\text{Brier}_{J,prag} - \text{Brier}_{obs}| \approx 0.0025$), followed by methods A and E. Method B has the largest magnitude (greater than 0.005 when sample size is 100 and approximately 0.003 for a sample size of 300). With increased sample size to 1000 all methods perform similarly when data are MCAR or covariate-dependent MAR.

For all sample sizes when data are MCAR or covariate-dependent MAR, the ideal performance of all *CV-then-MI* methods (methods A-F only as methods G and H do not have an ideal performance estimate) overestimates Brier_{obs} . Methods J and K (*MI-then-CV*) underestimate Brier_{obs} when sample size is 100 or 300 (over-optimistic) but tends to perform similarly to Brier_{obs} when sample size is 1000 across all missing data scenarios. When the sample size is 100, method A tends to have the smallest magnitude of the difference across all *CV-then-MI* methods. Method B tends to have the largest magnitude of difference ($|\text{Brier}_{B,ideal} - \text{Brier}_{obs}|$) while the methods similar to B but which either include the test fold when imputing the training fold (method D) or include the training folds when imputing the test fold (method C) tend to have a smaller magnitude than B. With increasing sample size, the methods tend to perform similarly when compared to Brier_{obs} .

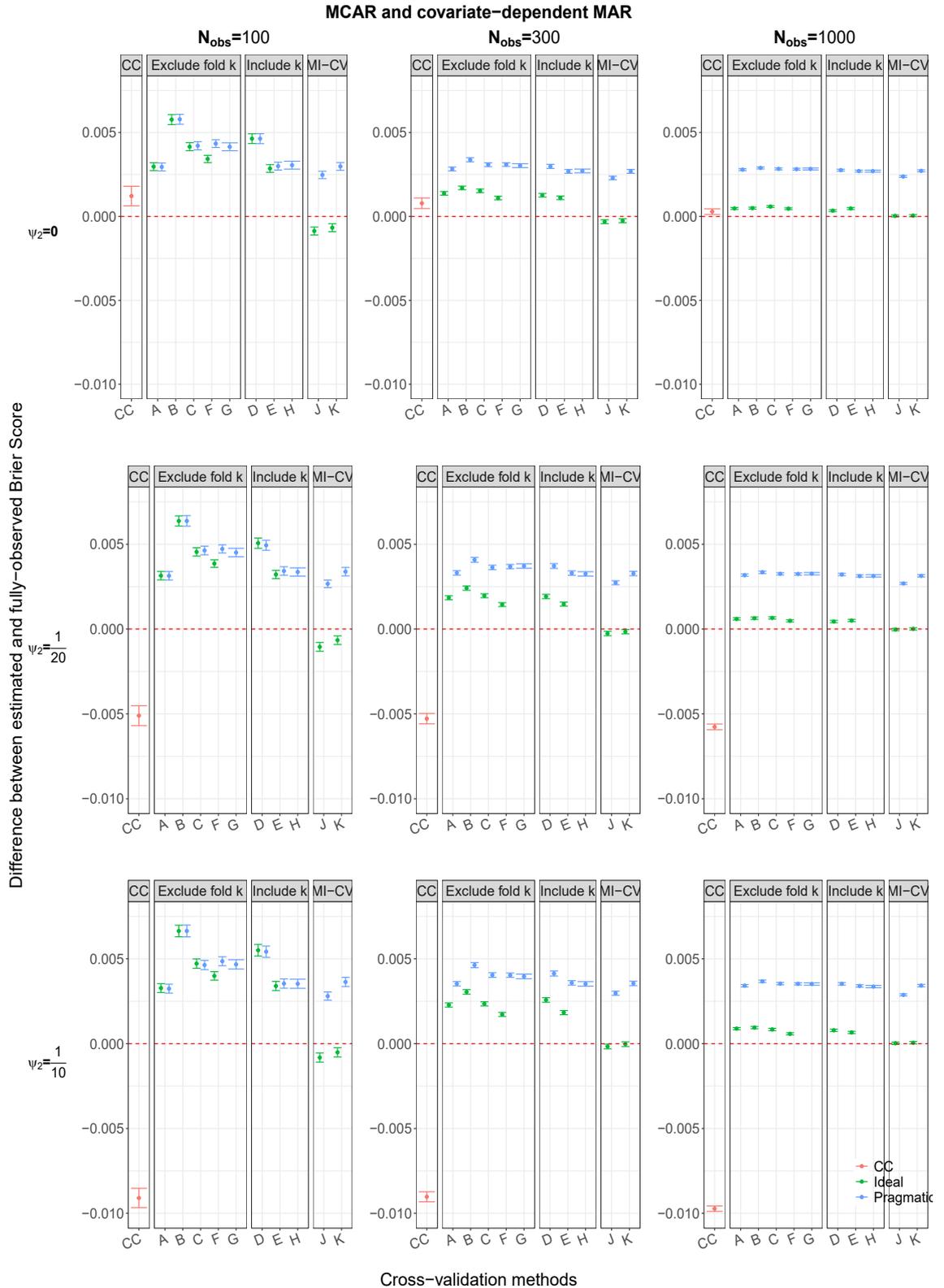


Figure 5.7: The difference $\text{Brier}_{imp} - \text{Brier}_{obs}$ when data are MCAR or covariate-dependent MAR for $M = 5$ when 25% of values are missing in X_1 . The error bars summarise results from the 2000 repetitions and the limits represent the Monte Carlo 95% confidence interval of $\text{Brier}_{imp} - \text{Brier}_{obs}$. The average Brier score when data are fully-observed is 0.17. CC (complete-case); methods A-K are described in Table 2.3 and summarised in Table 4.4.

Outcome-dependent MAR

Figure 5.8 displays results for the various cross-validation methods' (*imp*) estimates of the Brier score which are compared to Brier_{obs} ($\text{Brier}_{imp} - \text{Brier}_{obs}$) when data are outcome-dependent or outcome- and covariate-dependent MAR.

For all sample sizes and missing data scenarios, the complete-case analysis underestimates Brier_{obs} with a magnitude greater than 0.01 ($|\text{Brier}_{CC} - \text{Brier}_{obs}| > 0.01$). For the stronger MAR scenarios (rows two and three of Figure 5.8) the complete-case analysis estimate does not fit onto the scale of the graph.

For all sample sizes and missing data scenarios, the pragmatic performance of all methods overestimates Brier_{obs} . For a sample size of 100, method B tends to have the largest magnitude of the difference which is greater than 0.008 ($|\text{Brier}_{B,prag} - \text{Brier}_{obs}| > 0.008$) while method J tends to have the smallest magnitude of approximately 0.003. For *CV-then-MI*, methods A, E and H tend to have the smallest magnitudes which are approximately 0.004. With increasing sample size the methods tend to perform similarly, with method J having the smallest magnitude of the difference while method B has the largest.

For all sample sizes and missing data scenarios, the ideal performance of all *CV-then-MI* methods A-F overestimates Brier_{obs} . Methods J and K (*MI-then-CV*) underestimate Brier_{obs} when sample size is 100 or 300 (i.e. over-optimistic) but tends to perform similarly to Brier_{obs} when sample size is 1000 for all missing data scenarios. For a sample size of 100, methods A, E and F tend to have the smallest magnitudes of difference with Brier_{obs} across the *CV-then-MI* methods while methods J and K have the smallest magnitude overall. With increased sample size to 300, the magnitude of all methods is less than 0.005 ($|\text{Brier}_{imp,ideal} - \text{Brier}_{obs}| < 0.005$). For a sample size of 1000, methods J and K approximate Brier_{obs} well, while methods E and F have the smallest magnitude across the *CV-then-MI* methods while method B still has the largest magnitude.

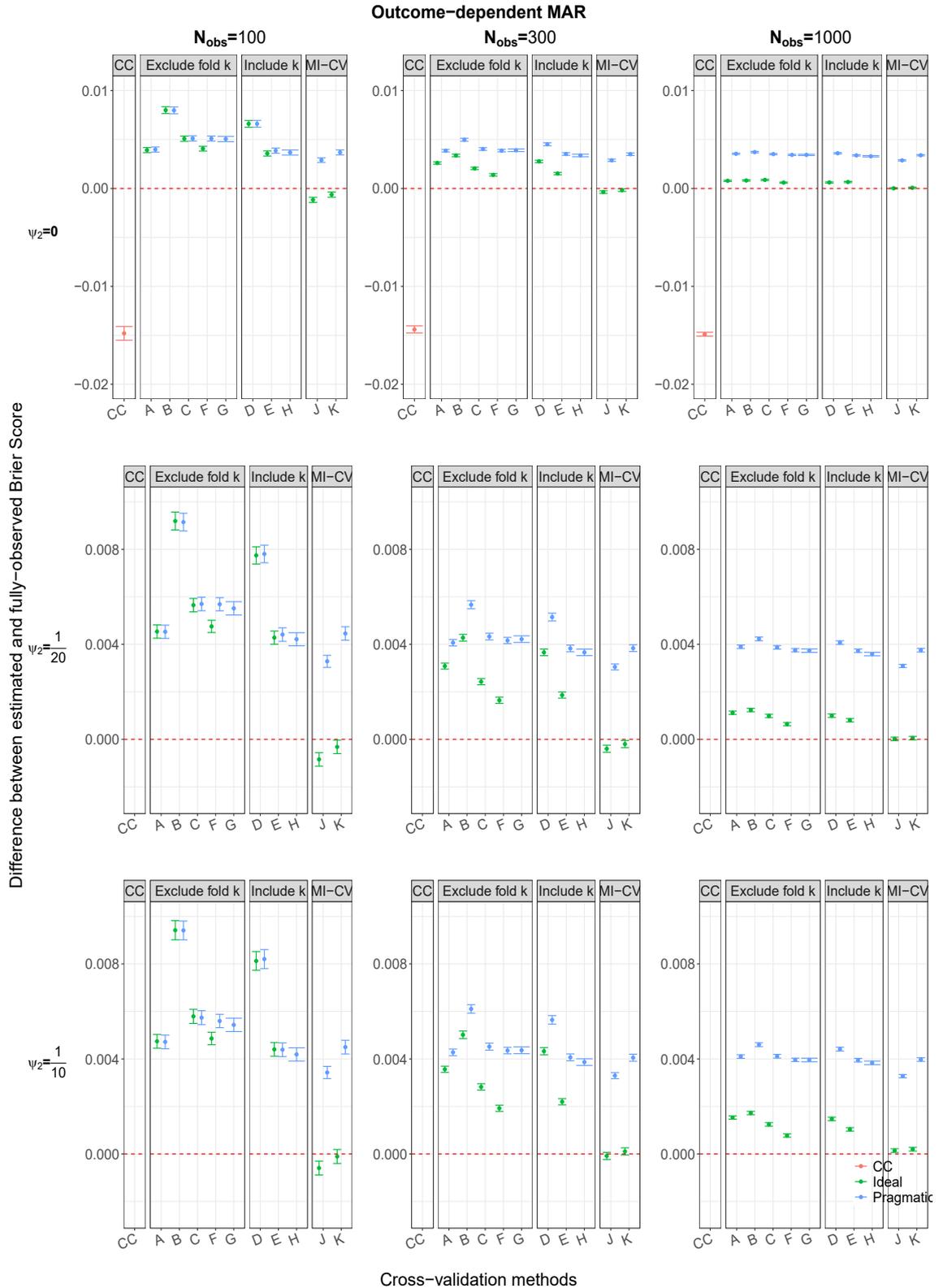


Figure 5.8: The difference $\text{Brier}_{imp} - \text{Brier}_{obs}$ when data are outcome-dependent or outcome- and covariate-dependent MAR for $M = 5$ when 25% of values are missing in X_1 . The error bars summarise results from the 2000 repetitions and the limits represent the Monte Carlo 95% confidence interval of $\text{Brier}_{imp} - \text{Brier}_{obs}$. The average Brier score when data are fully-observed is 0.17. CC (complete-case); methods A-K are described in Table 2.3 and summarised in Table 4.4.

5.4.2 Increasing the number of imputed datasets from 5 to 25

Figure 5.9 displays results comparing the use of 5 versus 25 imputed datasets when data are outcome-dependent or outcome- and covariate-dependent MAR ($\text{Brier}_{imp,M} - \text{Brier}_{obs}$). The results are for the pragmatic performance but are generalisable also to the ideal performance in all missing data scenarios. All graphs for the ideal and pragmatic performance are available in the Supplementary plots section S2.2.3.

As seen in Figure 5.9, the performance of all methods is unaffected by an increased number of imputed datasets when estimating the Brier score performance. This holds for all sample sizes and missing data scenarios for pragmatic and ideal performance.

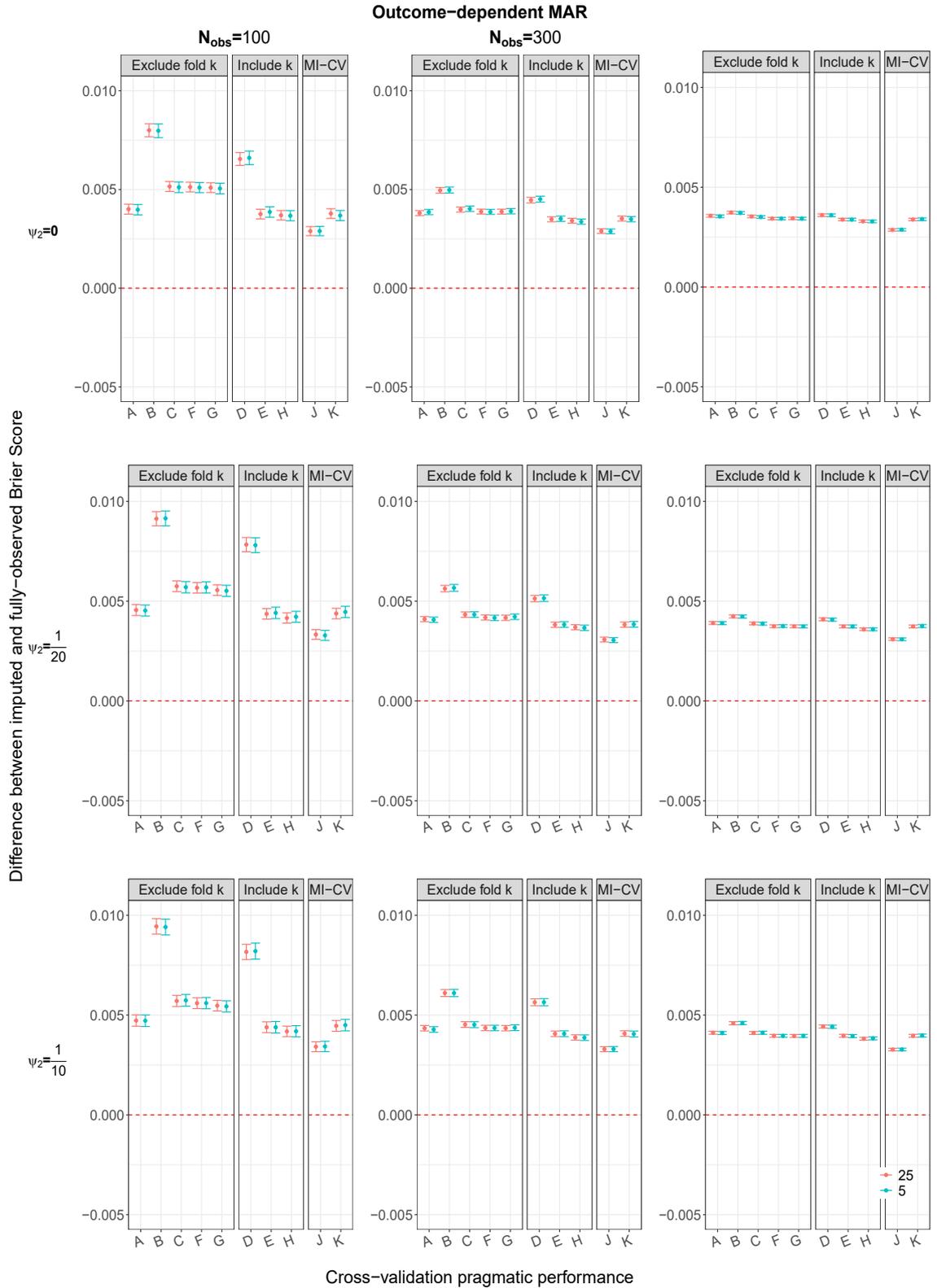


Figure 5.9: The difference $\text{Brier}_{imp} - \text{Brier}_{obs}$ when data are outcome-dependent or outcome- and covariate-dependent MAR for $M = 25$ versus $M = 5$ when 25% of values are missing in X_1 . The error bars summarise results from the 2000 repetitions for pragmatic performance and the limits represent the Monte Carlo 95% confidence interval of $\text{Brier}_{imp} - \text{Brier}_{obs}$. The average Brier score when data are fully-observed is 0.17. CC (complete-case); methods A-K are described in Table 2.3 and summarised in Table 4.4.

5.4.3 Increasing the percentage of missingness to 40%

Figure 5.10 displays results demonstrating the impact that an increased percentage of missingness can have on the various cross-validation methods when data are weakly outcome- and covariate-dependent MAR. The figure presents the Brier score estimates when 25% or 40% of X_1 values are missing compared to Brier_{obs} ($\text{Brier}_{imp,\%} - \text{Brier}_{obs}$). The results are generally representative of the comparison between 25% and 40% missingness for ideal and pragmatic performance for all missing data scenarios and sample sizes. All plots are available in Section S2.2.2 of the Supplementary Plots.

When data are MCAR, the complete-case analysis performs similarly regardless of whether 25% or 40% of X_1 values are missing. For all MAR scenarios, an increased percentage of missingness results in a larger magnitude of the complete-case analysis estimate when compared to Brier_{obs} ($|\text{Brier}_{CC,25\%} - \text{Brier}_{obs}| < |\text{Brier}_{CC,40\%} - \text{Brier}_{obs}|$).

For all sample sizes and missing data scenarios the pragmatic performance of methods G, J and K perform similarly regardless of the percentage of missing data in X_1 . For all other methods, an increased percentage of missingness causes an increase in the magnitude of the difference between their estimate of the Brier score and Brier_{obs} .

For all sample sizes and missing data scenarios the ideal performance of the *MI-then-CV* methods J and K perform similarly regardless of the percentage of missing data in X_1 . For all other methods (methods A-F), an increased percentage of missingness causes an increase in the magnitude ($|\text{Brier}_{imp,25\%} - \text{Brier}_{obs}| < |\text{Brier}_{imp,40\%} - \text{Brier}_{obs}|$ for $imp = A, \dots, F$).

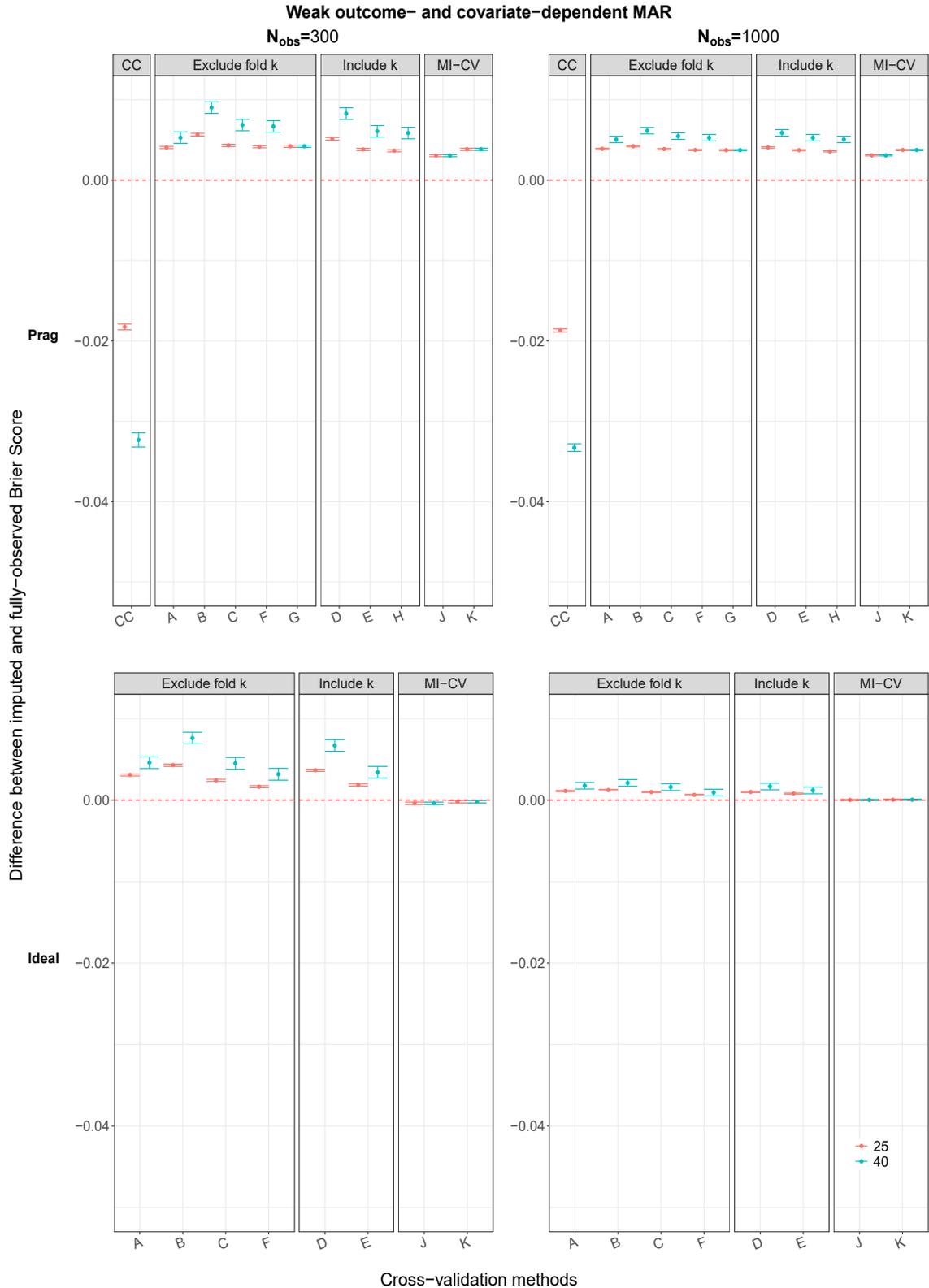


Figure 5.10: Comparing the impact of increasing the percentage of missingness on the difference $Brier_{imp} - Brier_{obs}$ when data are outcome- and covariate-dependent MAR when $M = 5$. The error bars summarise results from the 2000 repetitions and the limits represent the Monte Carlo 95% confidence interval of $Brier_{imp} - Brier_{obs}$. Red denotes $Brier_{imp} - Brier_{obs}$ when 25% of X_1 values are missing and blue denotes $Brier_{imp} - Brier_{obs}$ when 40% of X_1 values are missing. The top row presents the results for pragmatic performance and the bottom row presents results for ideal performance. The average Brier score when data are fully-observed is 0.17. CC (complete-case); methods A-K are described in Table 2.3.

5.4.4 Comparing each method's Brier score to the target estimate of the Brier score from a larger validation set

As previously discussed for the AUC results, the ideal performance of the proposed methods and $Brier_{obs}$ were compared to the ideal target Brier score estimate ($Brier_{target,obs}$). This is estimated by applying a prediction model, based on all data in a repetition, to the fully-observed data in the larger test set. The pragmatic performance of the imputation methods is compared to applying a repetition's prediction model to the imputed datasets of the larger test set ($Brier_{target,imputed}$). The complete-case estimate of the Brier score is compared to applying a repetition's prediction model to the observed cases of the larger test set ($Brier_{target,CC}$).

MCAR and covariate-dependent MAR

Figure 5.11 presents the ideal and pragmatic performance estimates of the various methods when compared to their respective target Brier score estimate, when data are MCAR or covariate-dependent MAR. The magnitude of the difference between the methods' Brier score estimate and the target estimate is less than 0.0075 when sample size is 100, less than 0.005 for a sample size of 300 and less than 0.0025 for a sample size of 1000.

When data are MCAR or strong covariate-dependent MAR, the complete-case analysis estimate overestimates $Brier_{target,CC}$ ($Brier_{CC} - Brier_{target,CC} > 0$). When data are weak covariate-dependent MAR and sample size is 100, the complete-case analysis estimate approximates the target estimate well but with increasing sample size the complete-case analysis tends to underestimate $Brier_{target,CC}$.

When data are MCAR or strong covariate-dependent MAR and sample size is 100, the pragmatic performance of methods A, E, H, J and K tend to approximate $Brier_{target,imputed}$ well ($Brier_{imp,prag} - Brier_{target,imputed} \approx 0$ for $imp = A, E, H, J, K$). The other methods tend to overestimate $Brier_{target,imputed}$ with method B having the largest magnitude of the difference ($|Brier_{B,prag} - Brier_{target,imputed}|$). With increasing sample size all methods tend to perform similarly and overestimate $Brier_{target,imputed}$. When data are weak covariate-dependent MAR and sample size is 100, all methods tend to approximate $Brier_{target,imputed}$ well (i.e. their confidence intervals overlap with zero) except for methods B and J who tend to over- and underestimate $Brier_{target,imputed}$, respectively. For a sample size of 300 or 1000, all methods tend to underestimate $Brier_{target,imputed}$. Method B has the smallest magnitude while method J has the largest. For a sample size of 1000 all methods perform similarly.

When data are MCAR or strong covariate-dependent MAR and sample size is 100, the

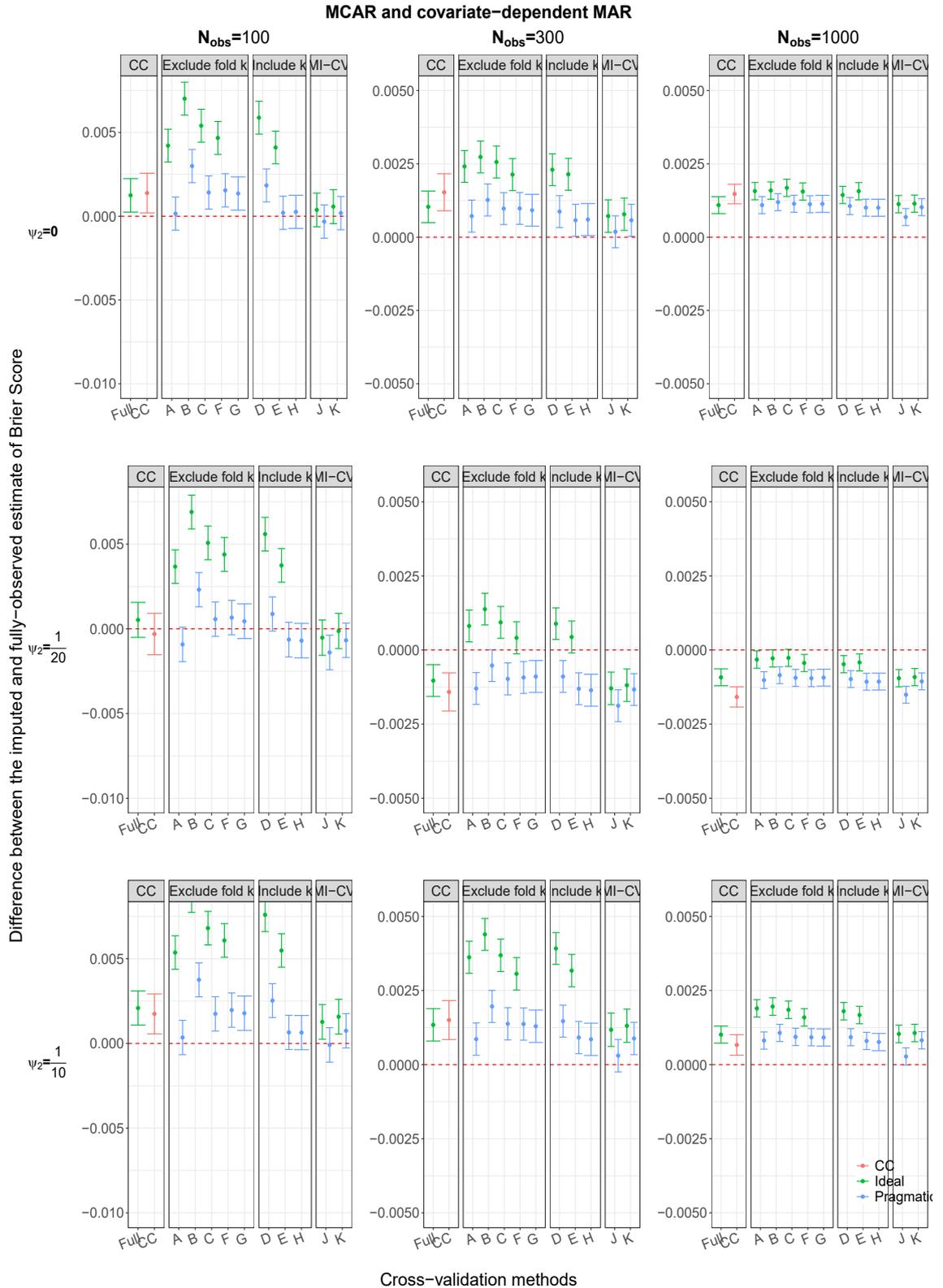


Figure 5.11: The difference $Brier_{imp} - Brier_{target}$ when data are MCAR or covariate-dependent MAR for $M = 5$ when 25% of values are missing in X_1 . The error bars summarise results from the 2000 repetitions and the limits represent the Monte Carlo 95% confidence interval of $Brier_{imp} - Brier_{target}$. The average Brier when data are fully-observed is 0.17. CC (complete-case); methods A-K are described in Table 2.3.

ideal performance of all methods overestimates $\text{Brier}_{target,obs}$. For *CV-then-MI*, methods A and E have the lowest magnitude ($|\text{Brier}_{imp,ideal} - \text{Brier}_{target,obs}|$) while methods J and K tend to have the lowest magnitude overall. With increasing sample size to 300 the magnitude of the difference for all methods is less than 0.0025 when data are MCAR or weak covariate-dependent MAR and less than 0.005 for strong covariate-dependent MAR. When data are MCAR or strong covariate-dependent MAR, all methods tend to slightly overestimate $\text{Brier}_{target,obs}$. When data are weak covariate-dependent MAR, methods A-F overestimate $\text{Brier}_{target,obs}$ and methods J and K underestimate $\text{Brier}_{target,obs}$ (i.e. over-optimistic). With increased sample size to 1000, all methods underestimate $\text{Brier}_{target,obs}$ and perform similarly.

Outcome-dependent MAR

Figure 5.12 presents the ideal and pragmatic performance estimates of the various methods when compared to their respective target Brier score estimate, when data are outcome-dependent or outcome- and covariate-dependent MAR. The magnitude of the difference between the methods' Brier score estimate and the target estimate is less than 0.01 when sample size is 100, less than 0.005 for a sample size of 300 and less than 0.0025 for a sample size of 1000. The complete-case analysis estimate underestimates $\text{Brier}_{target,CC}$ for all sample sizes and missing data scenarios ($\text{Brier}_{CC} - \text{Brier}_{target,CC} < 0$). With increasing strength of missingness, the magnitude of the difference ($|\text{Brier}_{CC} - \text{Brier}_{target,CC}|$) increases.

When the sample size is 100 or 300, the pragmatic performance of methods A, C, E-H, J and K tend to perform well, either approximating $\text{Brier}_{target,imputed}$ well or having very small magnitudes ($|\text{Brier}_{imp,prag} - \text{Brier}_{target,imputed}|$). Methods B and D tend to overestimate $\text{Brier}_{target,imputed}$ for all sample sizes and missing data scenarios. With increasing sample size to 1000 all methods perform similarly.

The ideal performance of the *CV-then-MI* methods A-F overestimates $\text{Brier}_{target,obs}$ for sample sizes of 100 or 300 for all missing data scenarios. The performance of methods J and K (*MI-then-CV*) tends to underestimate $\text{Brier}_{target,obs}$ when data are weak outcome- and covariate-dependent MAR and overestimates $\text{Brier}_{target,obs}$ for the other missing scenarios. For a sample size of 100 or 300, methods B and D tend to have the largest magnitude ($|\text{Brier}_{imp,ideal} - \text{Brier}_{target,obs}|$, $imp = B, D$) while methods J and K tends to have the smallest magnitude. With increasing sample size to 1000, the performance of all methods is similar except for the weak outcome- and strong covariate-dependent MAR scenario where methods J and K have the smallest magnitudes.

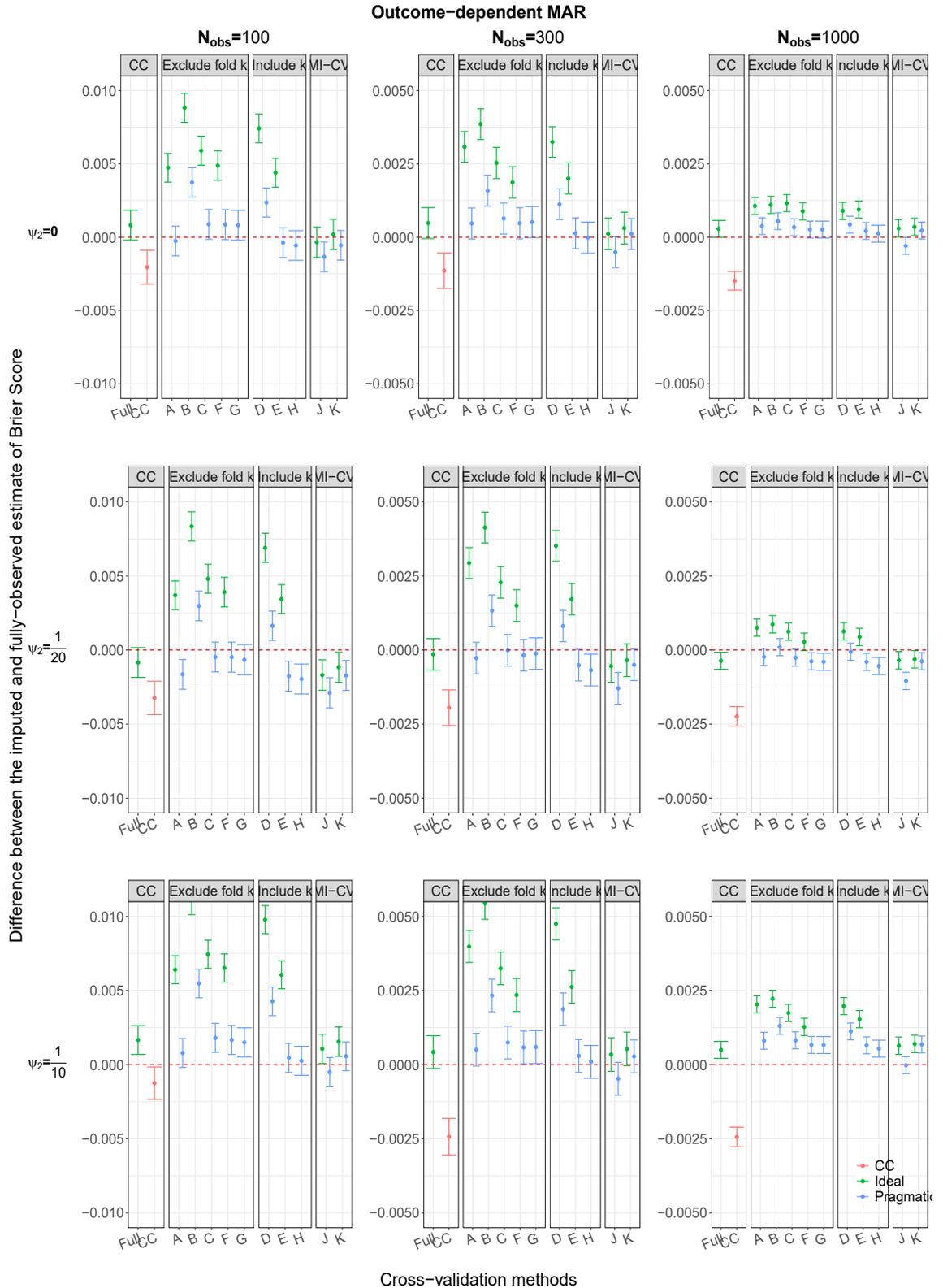


Figure 5.12: The difference $\text{Brier}_{\text{imp}} - \text{Brier}_{\text{target}}$ when data are outcome-dependent or outcome- and covariate-dependent MAR for $M = 5$ when 25% of values are missing in X_1 . The error bars summarise results from the 2000 repetitions and the limits represent the Monte Carlo 95% confidence interval of $\text{Brier}_{\text{imp}} - \text{Brier}_{\text{target}}$. The average Brier when data are fully-observed is 0.17. CC (complete-case); methods A-K are described in Table 2.3.

5.5 Detailed results: Calibration intercept

For a sample size of 100 when data are MAR, the various performance estimates of the calibration intercept estimate are very unstable when compared to the intercept estimated when data are fully-observed. This can also be seen for the bootstrap calibration results in Appendix C. The estimates of the calibration intercept for a sample size of 100 when data are fully-observed were previously noted to vary widely in Section 5.2 (Table 5.1). Here, we will focus on results for a sample size of 300 and 1000.

5.5.1 Comparing each method's Calibration intercept to the estimate of the Calibration intercept when data are fully-observed

MCAR and covariate-dependent MAR

Figure 5.13 displays results for the proposed methods' (imp) estimates of the calibration intercept which are compared to $Intercept_{obs}$ ($Intercept_{imp} - Intercept_{obs}$) when data are MCAR or covariate-dependent MAR.

The complete-case analysis estimate underestimates $Intercept_{obs}$ ($Intercept_{CC} - Intercept_{obs} < 0$). For covariate-dependent MAR when sample size is 300, the magnitude of the underestimation is greater than 0.015 and it does not fit onto the scale of Figure 5.13. However, with increasing sample size the magnitude of the difference ($|Intercept_{CC} - Intercept_{obs}|$) decreases when data are MCAR or covariate-dependent MAR.

For all sample sizes when data are MCAR or covariate-dependent MAR, the pragmatic performance of methods B and D overestimate $Intercept_{obs}$ ($Intercept_{imp,prag} - Intercept_{obs} > 0$ for $imp = B, D$). For a sample size of 300 when data are covariate-dependent MAR, they do not fit onto the scale of the graph. With increasing sample size, the magnitude of the difference decreases ($|Intercept_{imp,prag} - Intercept_{obs}| \rightarrow 0$, $imp = B, D$). The pragmatic performance of the other methods tends to underestimate $Intercept_{obs}$ when data are MCAR. When data are weak or strong covariate-dependent MAR, methods E and K tend to overestimate $Intercept_{obs}$ while the remaining methods either underestimate or approximate $Intercept_{obs}$ well.

Similarly for the ideal performance, methods B and D overestimate $Intercept_{obs}$ ($Intercept_{imp,ideal} - Intercept_{obs} > 0$ for $imp = B, D$) with a magnitude greater than 0.1. The ideal performance of the remaining methods either over- or underestimate $Intercept_{obs}$ but with similar magnitudes.

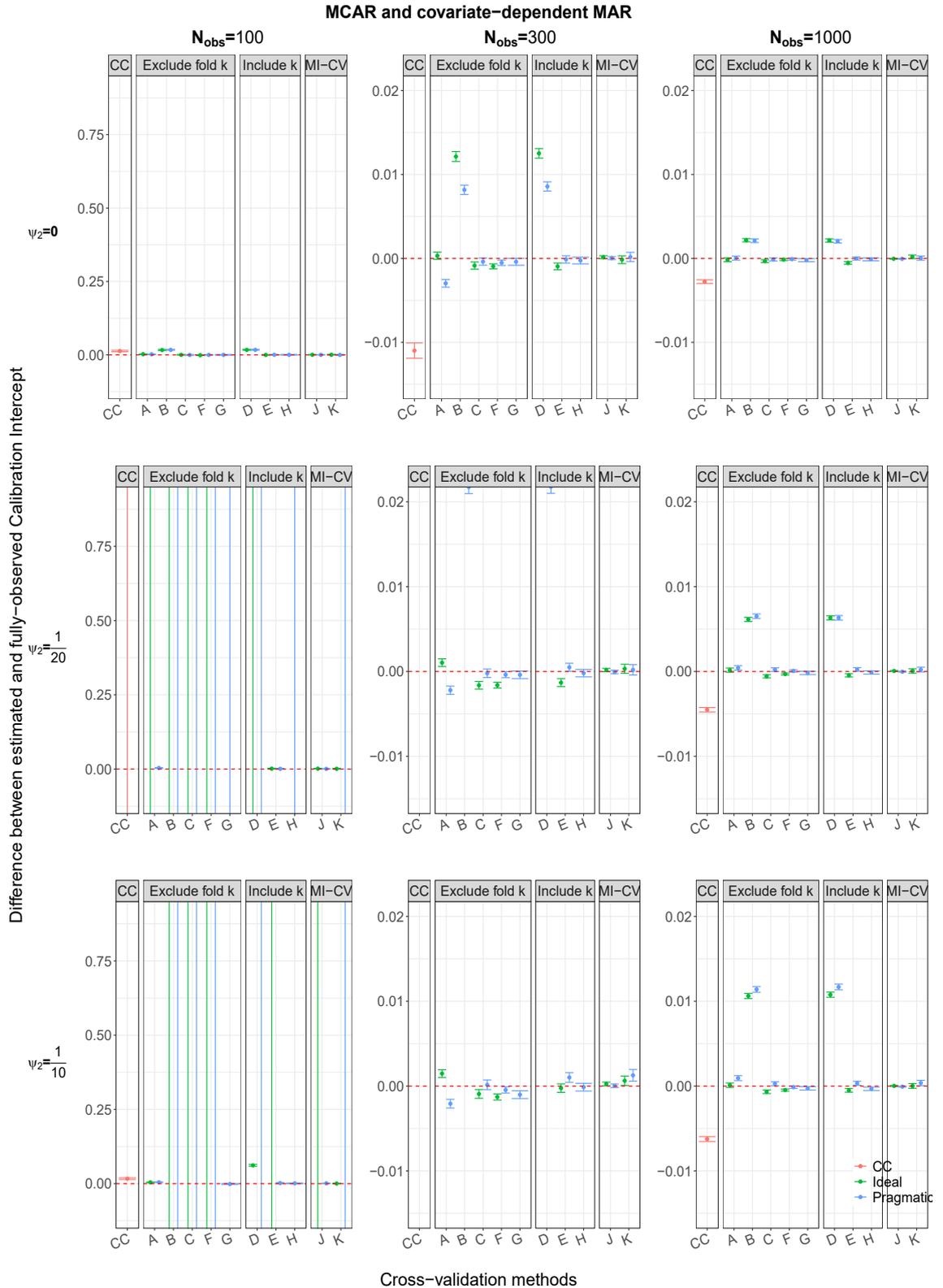


Figure 5.13: The difference $\text{Intercept}_{imp} - \text{Intercept}_{obs}$ when data are MCAR or covariate-dependent MAR for $M = 5$ when 25% of values are missing in X_1 . The error bars summarise results from the 2000 repetitions and the limits represent the Monte Carlo 95% confidence interval of $\text{Intercept}_{imp} - \text{Intercept}_{obs}$. The average Calibration intercept when data are fully-observed is 0.02 for larger sample sizes. CC (complete-case); methods A-K are described in Table 2.3 and summarised in Table 4.4.

Outcome-dependent MAR

Figure 5.14 displays results when data are outcome-dependent or outcome- and covariate-dependent MAR. The graph presents the comparison of the various methods' (imp) estimates of the calibration intercept to the intercept estimate when data are fully-observed ($\text{Intercept}_{imp} - \text{Intercept}_{obs}$).

The complete-case analysis estimate of the calibration intercept underestimates Intercept_{obs} for a sample size of 300 and 1000, at times not fitting onto the scale of the graph when sample size is 300. Increasing the sample size to 1000 decreases the magnitude of the underestimation ($|\text{Intercept}_{imp} - \text{Intercept}_{obs}| \rightarrow 0$).

The pragmatic performance of all methods overestimates Intercept_{obs} for a sample size of 300 or 1000 across all missing data scenarios. When sample size is 300, methods A and J have the smallest magnitudes ($|\text{Intercept}_{imp,prag} - \text{Intercept}_{obs}| < 0.0025$ for $imp = A, J$) while methods B and D have the largest. When sample size is 1000, method J has the smallest magnitude of the difference across all methods while methods A and F-H tend to perform similarly with the lowest magnitude for *CV-then-MI* methods.

The ideal performance of all methods also overestimates Intercept_{obs} for all sample sizes and missing data scenarios. For the *MI-then-CV* methods when sample size is 300 or 1000, method J (impute once) has a smaller magnitude than method K (impute using a training and test imputation model). For the *CV-then-MI* methods, methods A and F tend to have the smallest magnitude ($|\text{Intercept}_{imp,ideal} - \text{Intercept}_{obs}| < 0.0025$ for $imp = A, F$) while methods B and D have the largest ($|\text{Intercept}_{imp,ideal} - \text{Intercept}_{obs}| \geq 0.015$ for $imp = B, D$).

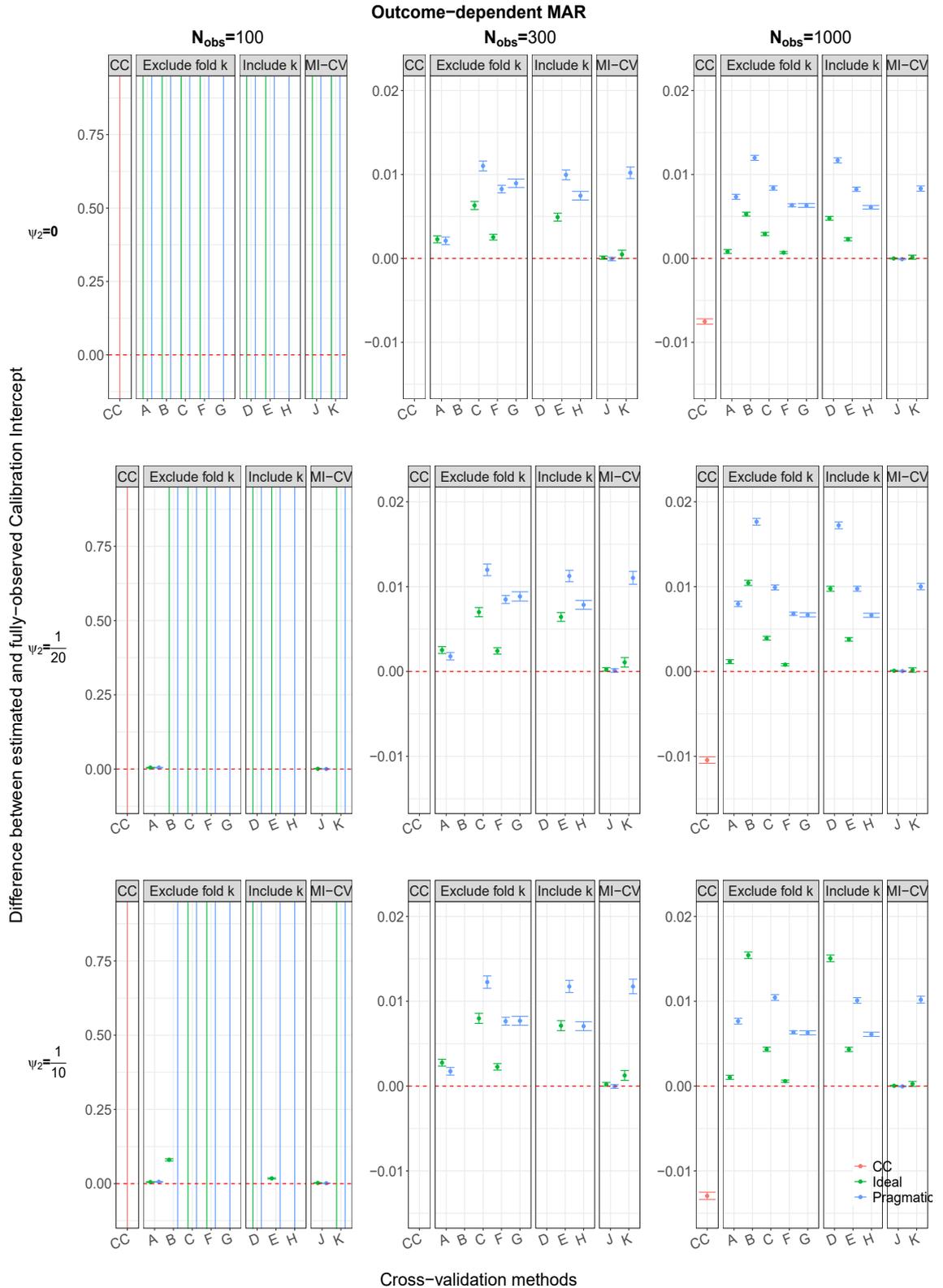


Figure 5.14: The difference $\text{Intercept}_{imp} - \text{Intercept}_{obs}$ when data are outcome-dependent or outcome- and covariate-dependent MAR for $M = 5$ when 25% of values are missing in X_1 . The error bars summarise results from the 2000 repetitions and the limits represent the Monte Carlo 95% confidence interval of $\text{Intercept}_{imp} - \text{Intercept}_{obs}$. The average Calibration intercept when data are fully-observed is 0.02 for larger sample sizes. CC (complete-case); methods A-K are described in Table 2.3 and summarised in Table 4.4.

5.5.2 Increasing the number of imputed datasets from 5 to 25

Figure 5.15 displays results comparing the use of 5 versus 25 imputed datasets when data are outcome-dependent or outcome- and covariate-dependent MAR ($\text{Intercept}_{imp,M} - \text{Intercept}_{obs}$). The results are for the pragmatic performance but are generalisable also to the ideal performance in all missing data scenarios. All graphs comparing 5 versus 25 imputed datasets for the ideal and pragmatic performance are available in the Supplementary plots section S2.3.3.

The use of 25 imputed datasets to estimate the calibration intercept has little effect when compared to 5 imputed datasets ($\text{Intercept}_{imp,M=5} - \text{Intercept}_{obs} \approx \text{Intercept}_{imp,M=25} - \text{Intercept}_{obs}$). This can be seen for all methods across all data-generating scenarios for sample sizes of 300 or 1000.

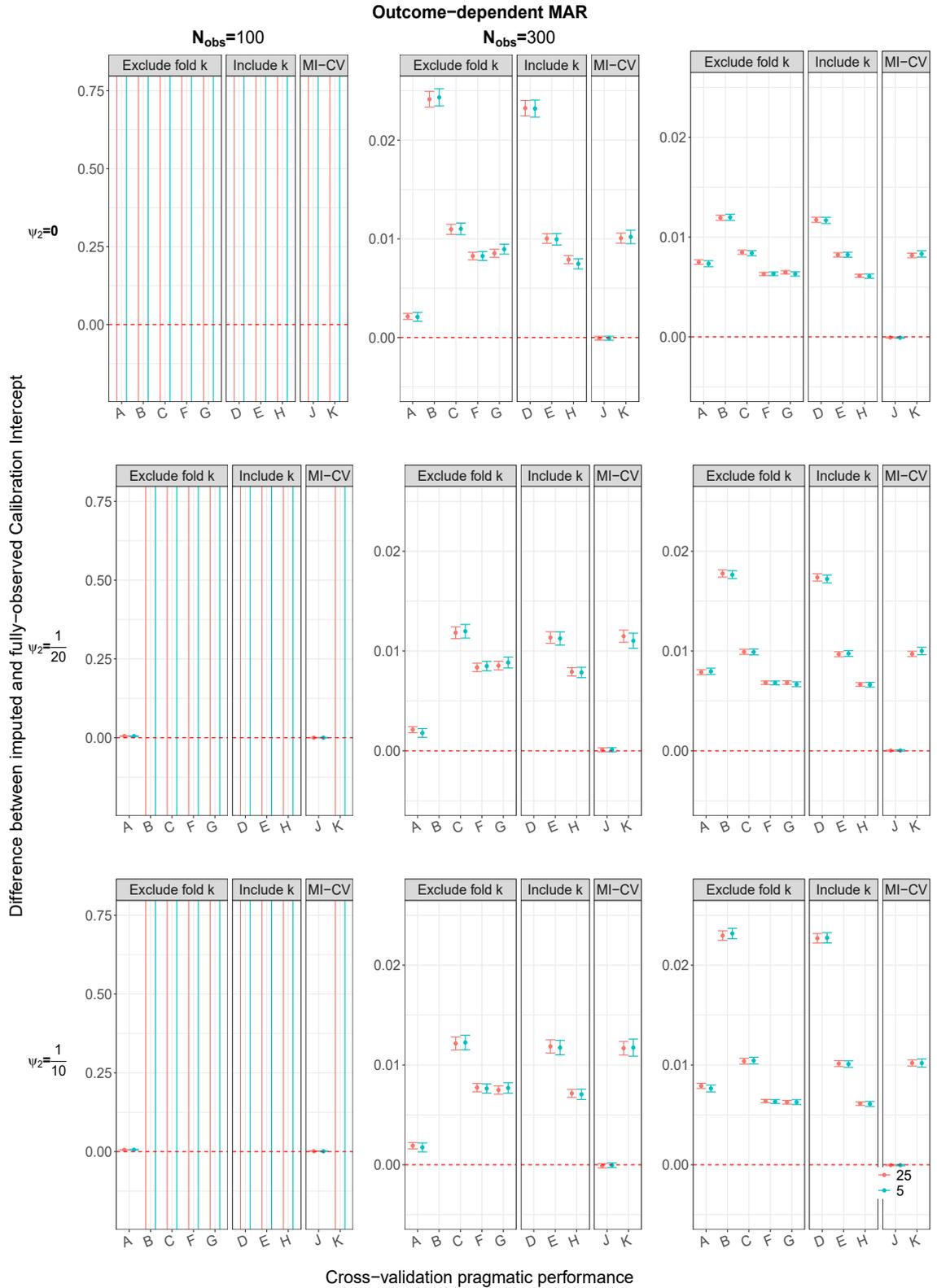


Figure 5.15: The difference $\text{Intercept}_{imp} - \text{Intercept}_{obs}$ when data are outcome-dependent or outcome- and covariate-dependent MAR for $M = 25$ versus $M = 5$ when 25% of values are missing in X_1 . The error bars summarise results from the 2000 repetitions for pragmatic performance and the limits represent the Monte Carlo 95% confidence interval of $\text{Intercept}_{imp} - \text{Intercept}_{obs}$. The average Calibration intercept when data are fully-observed is 0.02 for larger sample sizes. CC (complete-case); methods A-K are described in Table 2.3 and summarised in Table 4.4.

5.5.3 Increasing the percentage of missingness to 40%

Figure 5.16 displays results demonstrating the impact that an increased percentage of missingness can have on the various methods when data are weakly outcome- and covariate-dependent MAR. The figure presents the calibration intercept estimates when 25% or 40% of X_1 values are missing compared to Intercept_{obs} ($\text{Intercept}_{imp,\%} - \text{Intercept}_{obs}$). The results are generally representative of the comparison between 25% and 40% missingness for ideal and pragmatic performance for all missing data scenarios and sample sizes. All plots are available in Section S2.3.2 of the Supplementary Plots.

The complete-analysis estimate when 25% of X_1 values are missing tends to have a smaller magnitude than when 40% of values are missing ($|\text{Intercept}_{CC,25\%} - \text{Intercept}_{obs}| < |\text{Intercept}_{CC,40\%} - \text{Intercept}_{obs}|$) for all sample sizes and missing data scenarios. The complete-case analysis estimates do not fit onto the scale of Figure 5.16.

The pragmatic performance of methods A, B and D has a larger magnitude when the percentage of missingness is increased when data are MCAR or covariate-dependent MAR. The other methods either perform similarly or have overlapping confidence intervals when comparing the percentage of missingness. When data are outcome-dependent or outcome- and covariate-dependent, methods G, J and K tend to perform similarly regardless of the percentage of missing values. The magnitude for all other methods increases with an increased percentage ($|\text{Intercept}_{imp,25\%} - \text{Intercept}_{obs}| < |\text{Intercept}_{imp,40\%} - \text{Intercept}_{obs}|$ for $imp = A-F, H, K$).

When data are MCAR or covariate-dependent MAR, the ideal performance of methods B and D has a larger magnitude when the percentage of missing values increases to 40% compared to when 25% of X_1 values are missing. For all other methods, the confidence intervals for the intercept estimate when 40% of values are missing either overlap or encompass the point estimate and confidence intervals with a smaller percentage of missingness. When data are outcome-dependent or outcome- and covariate-dependent, methods A, J and K perform similarly or have overlapping confidence intervals regardless of the percentage of missing values while for all other methods $|\text{Intercept}_{imp,25\%} - \text{Intercept}_{obs}| < |\text{Intercept}_{imp,40\%} - \text{Intercept}_{obs}|$ for $imp = B-F, G$ and H.

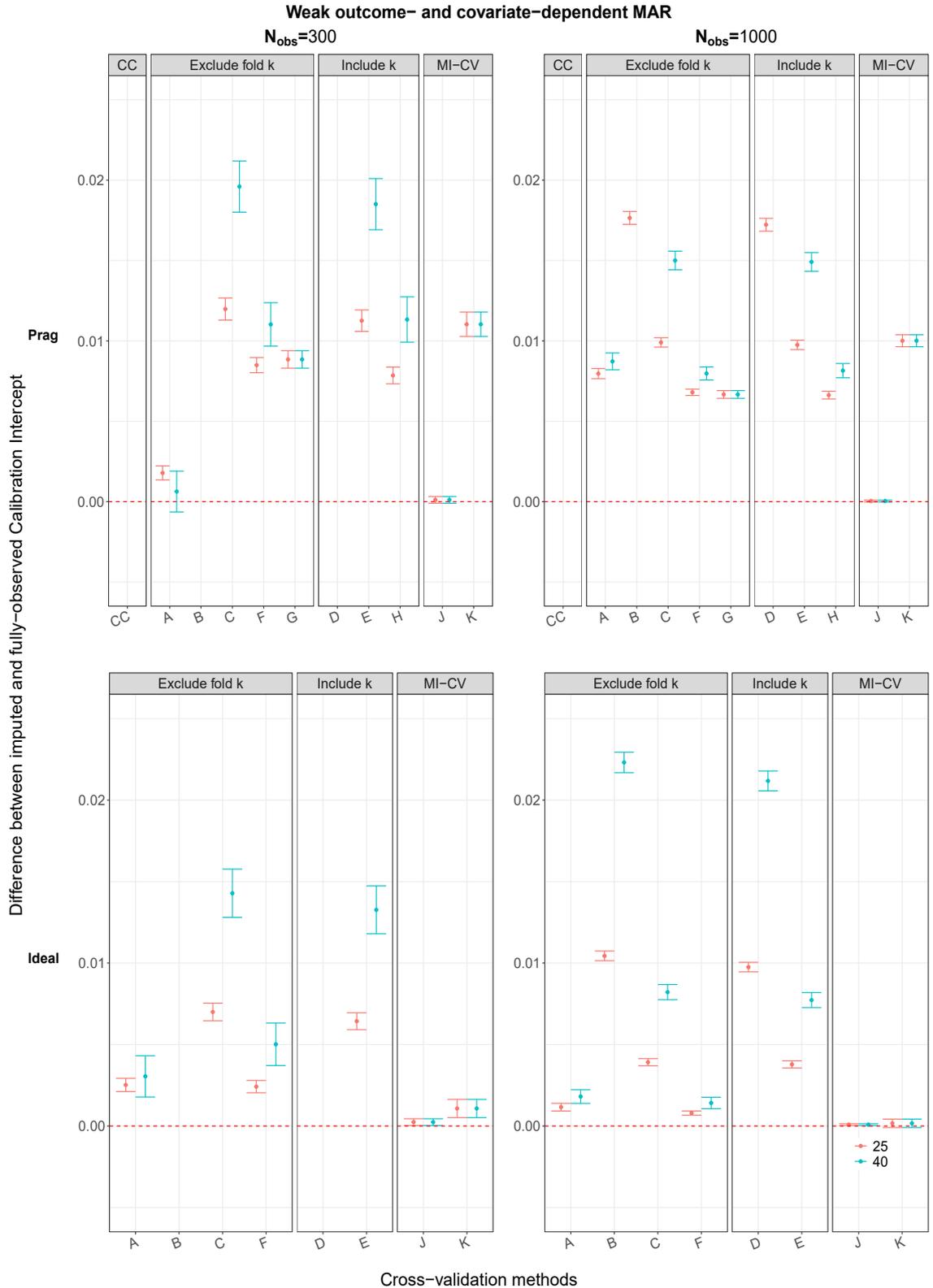


Figure 5.16: Comparing the impact of increasing the percentage of missingness on the difference $\text{Intercept}_{imp} - \text{Intercept}_{obs}$ when data are outcome- and covariate-dependent MAR when $M = 5$. The error bars summarise results from the 2000 repetitions and the limits represent the Monte Carlo 95% confidence interval of $\text{Intercept}_{imp} - \text{Intercept}_{obs}$. Red denotes $\text{Intercept}_{imp} - \text{Intercept}_{obs}$ when 25% of X_1 values are missing and blue denotes $\text{Intercept}_{imp} - \text{Intercept}_{obs}$ when 40% of X_1 values are missing. The top row presents the results for pragmatic performance and the bottom row presents results for ideal performance. The average Calibration intercept when data are fully-observed is 0.02 for larger sample sizes. CC (complete-case); methods A-K are described in Table 2.3.

5.5.4 Comparing each method’s calibration intercept to the target estimate of the calibration intercept from a larger validation set

As previously discussed for the continuous outcome scenario for the AUC and Brier score results, the ideal performance of the methods and Intercept_{obs} are compared to the ideal target intercept estimate ($\text{Intercept}_{target,obs}$). This is estimated by applying a prediction model, based on all data in a repetition, to the fully-observed data in the larger test set. The pragmatic performance of the imputation methods is compared to applying a repetition’s prediction model to the imputed datasets of the larger test set ($\text{Intercept}_{target,imputed}$). The complete-case estimate of the intercept is compared to applying a repetition’s prediction model to the observed cases of the larger test set ($\text{Intercept}_{target,CC}$).

MCAR and covariate-dependent MAR

Figure 5.17 presents the ideal and pragmatic performance estimates of the various methods when compared to their respective target calibration intercept estimate, when data are MCAR or covariate-dependent MAR. For all sample sizes, the complete-case analysis underestimates $\text{Intercept}_{target,CC}$ with a magnitude of approximately 0.5 ($|\text{Intercept}_{obs} - \text{Intercept}_{target,CC}| \approx 0.5$). As a result, the complete-case estimate can not be seen in Figure 5.17 due to the scale of the graph.

For a sample size of 300 or 1000, the pragmatic performance of all methods overestimates $\text{Intercept}_{target,imputed}$ ($\text{Intercept}_{imp,prag} - \text{Intercept}_{target,imputed} > 0$). Methods B and D tend to have a slightly larger magnitude than the other methods which all perform similarly with a magnitude less than 0.05.

When data are MCAR or weak covariate-dependent MAR and sample size is 300, the ideal performance of all methods tends to underestimate Intercept_{obs} ($\text{Intercept}_{obs,ideal} - \text{Intercept}_{target,obs} < 0$). Methods B and D tend to have a slightly smaller magnitude than the other methods which perform similarly to each other. When sample size is 300 and data are strong covariate-dependent MAR, methods B and D overestimate Intercept_{obs} while the other methods still underestimate Intercept_{obs} . When the sample size is increased to 1000, all methods overestimate Intercept_{obs} and methods B and D have the largest magnitude of the difference.

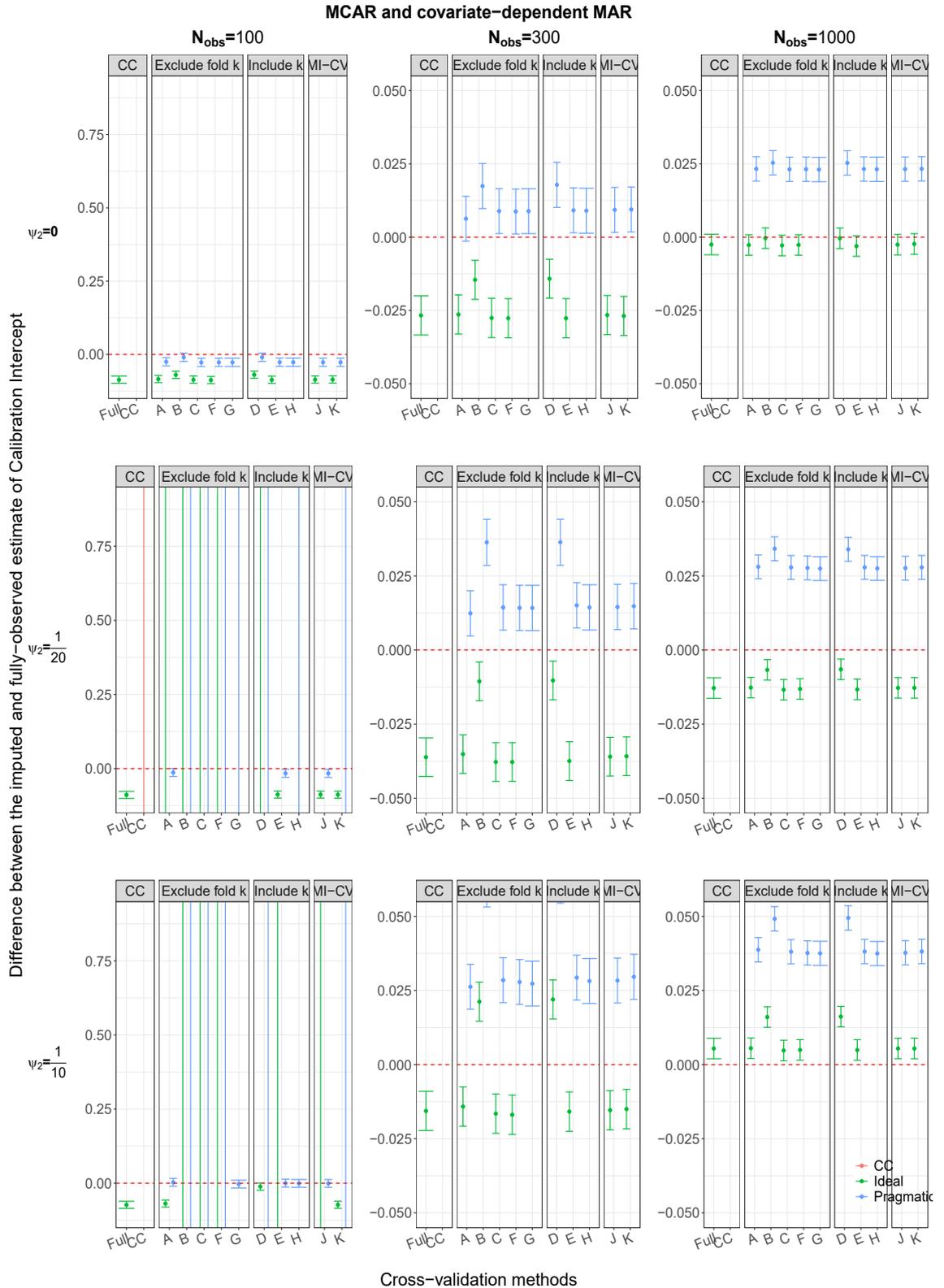


Figure 5.17: The difference $\text{Intercept}_{imp} - \text{Intercept}_{target}$ when data are MCAR or covariate-dependent MAR for $M = 5$ when 25% of values are missing in X_1 . The error bars summarise results from the 2000 repetitions and the limits represent the Monte Carlo 95% confidence interval of $\text{Intercept}_{imp} - \text{Intercept}_{target}$. The average Calibration intercept when data are fully-observed is 0.02 for larger sample sizes. CC (complete-case); methods A-K are described in Table 2.3.

Outcome-dependent MAR

Figure 5.18 presents the ideal and pragmatic performance estimates of the various methods when compared to their respective target calibration intercept estimate, when data are outcome-dependent or outcome- and covariate-dependent MAR.

For sample sizes of 300 or 1000, the complete-case analysis estimate of the calibration intercept underestimates $\text{Intercept}_{target,CC}$ with a magnitude of approximately 0.5 for all missing data scenarios. As such, the complete-case estimate does not fit onto the scale of Figure 5.18.

For sample sizes of 300 or 1000 and all missing data scenarios, the pragmatic performance of all methods overestimates $\text{Intercept}_{target,imputed}$ ($\text{Intercept}_{imp,prag} - \text{Intercept}_{target,imputed} > 0$). When the sample size is 300, methods A and J tend to have the smallest magnitude ($|\text{Intercept}_{imp,prag} - \text{Intercept}_{target,imputed}|$, $imp = A, J$), methods B and D have the largest magnitude, and the remaining methods (methods C, E-H and K) perform similarly to each other. When sample size is 1000, method J has the smallest magnitude and methods B and D have the largest magnitude. The remaining methods perform similarly when compared to the target pragmatic estimate.

For sample sizes of 300 or 1000 and all missing data scenarios, the ideal performance of methods B and D tend to approximate Intercept_{obs} well or slightly overestimate it ($\text{Intercept}_{imp,ideal} - \text{Intercept}_{target,imputed} \geq 0$, $imp = B, D$). All other methods (methods A, C, E, F, J and K) underestimate Intercept_{obs} for all sample sizes and missing data scenarios and perform similarly to each other.

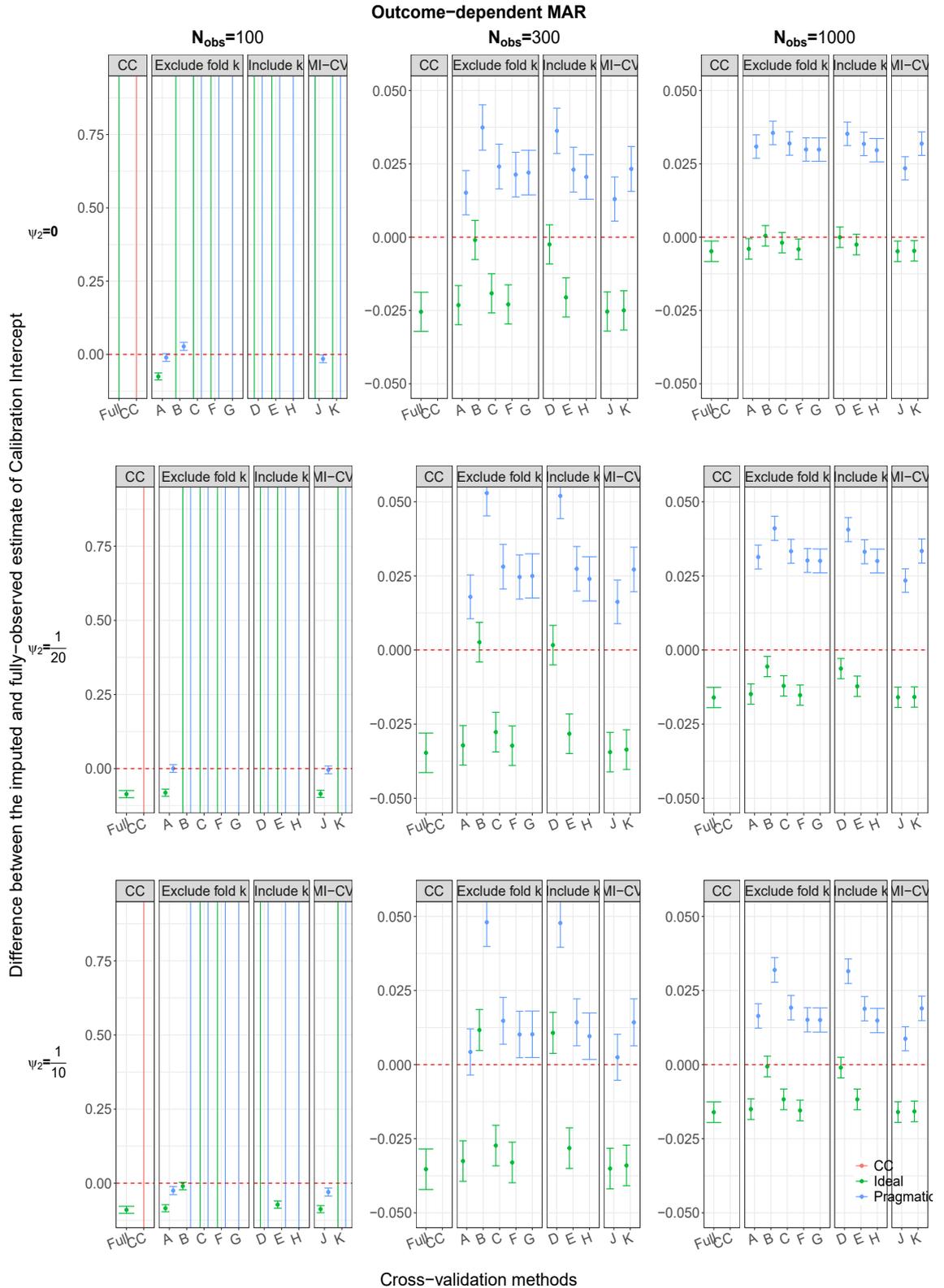


Figure 5.18: The difference $\text{Intercept}_{imp} - \text{Intercept}_{target}$ when data are outcome-dependent or outcome- and covariate-dependent MAR for $M = 5$ when 25% of values are missing in X_1 . The error bars summarise results from the 2000 repetitions and the limits represent the Monte Carlo 95% confidence interval of $\text{Intercept}_{imp} - \text{Intercept}_{target}$. The average Calibration intercept when data are fully-observed is 0.02 for larger sample sizes. CC (complete-case); methods A-K are described in Table 2.3.

5.6 Detailed results: Calibration slope

Similarly to the calibration intercept, for a sample size of 100 the various performance estimates of the calibration slope estimate are very unstable when compared to the slope estimated when data are fully-observed. For a sample size of 300, the results are slightly improved but still have large variation. Results will be discussed for a sample size of 1000.

5.6.1 Comparing each method's Calibration slope to the estimate of the Calibration slope when data are fully-observed

MCAR and covariate-dependent MAR

Figure 5.19 displays results for the various methods' (imp) estimates of the calibration slope which are compared to the slope estimate when data are fully-observed ($Slope_{imp} - Slope_{obs}$) when data are MCAR or covariate-dependent MAR.

The complete-case analysis overestimates $Slope_{obs}$ for a sample size of 1000 when data are MCAR or covariate-dependent MAR. The magnitude ($|Slope_{CC} - Slope_{obs}|$) is approximately 0.02.

The pragmatic performance of all methods underestimates $Slope_{obs}$ with a magnitude between 0.03 and 0.06, except for method J which has the smallest magnitude across all methods >0.001 . Methods A, B and D perform similarly with the smallest magnitudes across the *CV-then-MI* methods while methods C, E-H perform similarly with larger magnitudes. Method K performs similarly to methods C, E-H.

The ideal performance of methods A, B and D overestimate $Slope_{obs}$ ($Slope_{imp,ideal} - Slope_{obs}$, $imp = A, B, D$) while the other methods underestimate $Slope_{obs}$. Method A tends to have the smallest magnitude of the difference amongst the *CV-then-MI* methods while method D tends to have the largest. Overall, methods J and K have a smaller magnitude than the *CV-then-MI* methods while all methods have a magnitude less than 0.01.

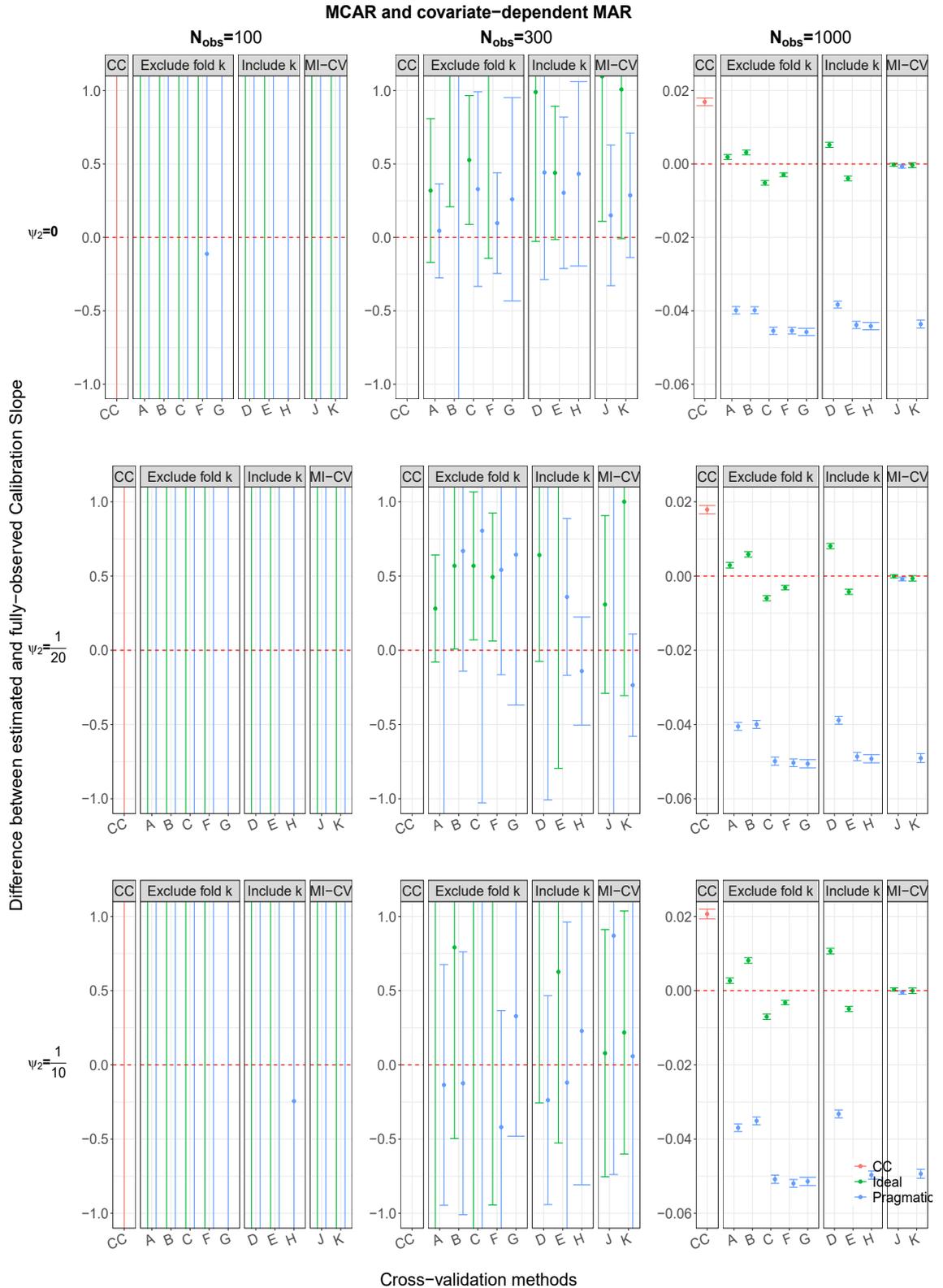


Figure 5.19: The difference $\text{Slope}_{imp} - \text{Slope}_{obs}$ when data are MCAR or covariate-dependent MAR for $M = 5$ when 25% of values are missing in X_1 . The error bars summarise results from the 2000 repetitions and the limits represent the Monte Carlo 95% confidence interval of $\text{Slope}_{imp} - \text{Slope}_{obs}$. The average Calibration slope when data are fully-observed is 1.04 for larger sample sizes. CC (complete-case); methods A-K are described in Table 2.3 and summarised in Table 4.4.

Outcome-dependent MAR

Figure 5.20 displays results for the various cross-validation methods' (*imp*) estimates of the calibration slope which are compared to the slope estimate when data are fully-observed ($\text{Slope}_{imp} - \text{Slope}_{obs}$) when data are outcome-dependent or outcome- and covariate-dependent MAR.

The complete-case analysis overestimates Slope_{obs} when sample size is 1000 for all missing data scenarios. The magnitude is approximately 0.02.

The pragmatic performance of all methods underestimates Slope_{obs} for a sample size of 1000 when data are outcome-dependent or outcome- and covariate-dependent MAR. Method A tends to have the smallest magnitude ($|\text{Slope}_{imp,prag} - \text{Slope}_{obs}|$), closely followed by methods B and D across the *CV-then-MI* methods while method J has the smallest magnitude overall. Methods C, E-H and K tend to perform similarly and have the largest magnitude across all methods.

The ideal performance of methods A, B and D overestimates Slope_{obs} with a smaller magnitude than methods C, E and F which underestimate Slope_{obs} . Across all *CV-then-MI* methods, method A tends to have the smallest magnitude while the ideal performance of methods J and K has the smallest magnitude of the difference overall.

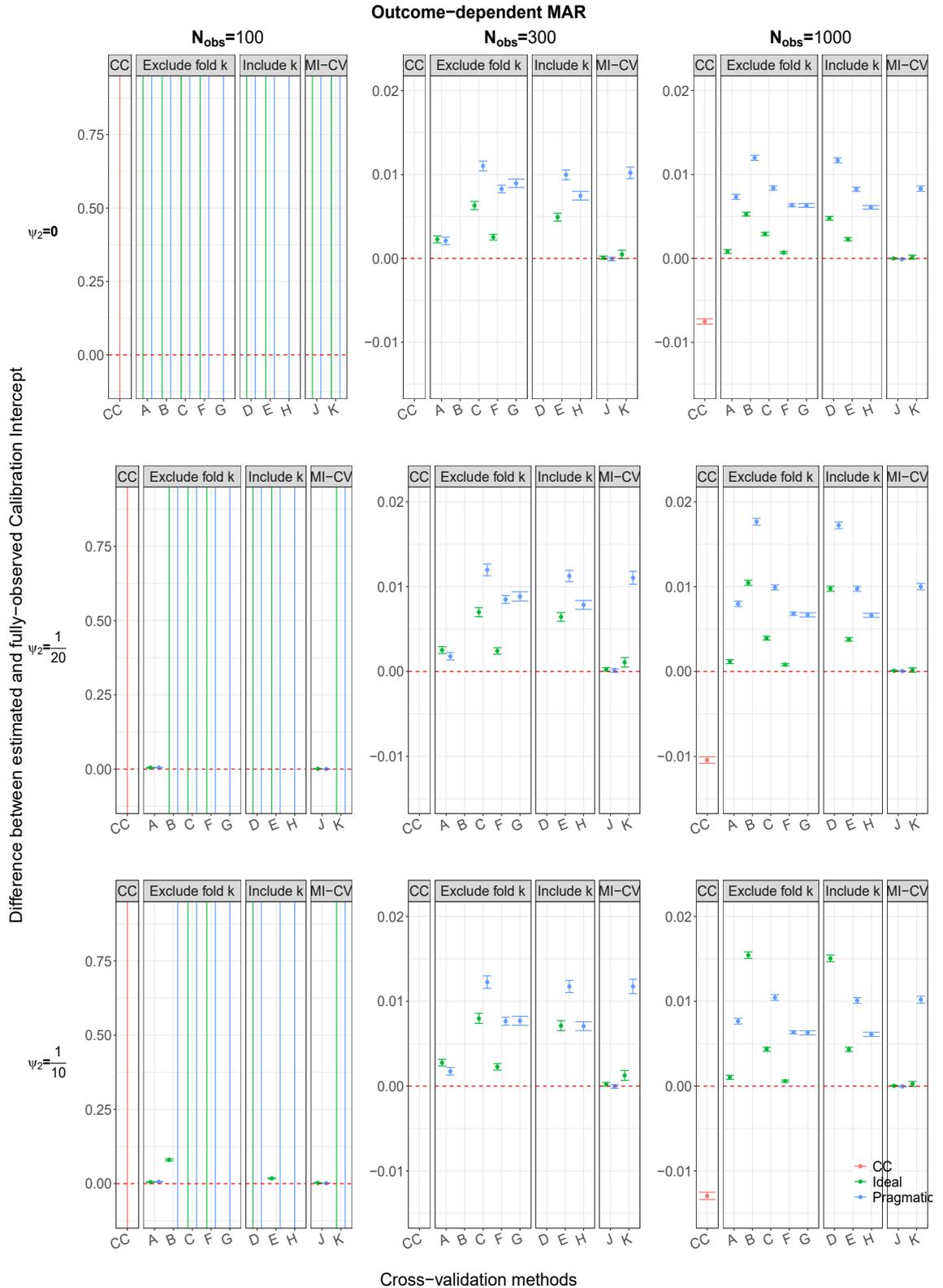


Figure 5.20: The difference $\text{Slope}_{imp} - \text{Slope}_{obs}$ when data are outcome-dependent or outcome- and covariate-dependent MAR for $M = 5$ when 25% of values are missing in X_1 . The error bars summarise results from the 2000 repetitions and the limits represent the Monte Carlo 95% confidence interval of $\text{Slope}_{imp} - \text{Slope}_{obs}$. The average Calibration slope when data are fully-observed is 1.04 for larger sample sizes. CC (complete-case); methods A-K are described in Table 2.3 and summarised in Table 4.4.

5.6.2 Increasing the number of imputed datasets from 5 to 25

Figure 5.21 displays results comparing the use of 5 versus 25 imputed datasets when data are outcome-dependent or outcome- and covariate-dependent MAR ($\text{Slope}_{imp,M} - \text{Slope}_{obs}$). The results are for the pragmatic performance but are generalisable also to the ideal performance in all missing data scenarios. All graphs comparing 5 versus 25 imputed datasets for the ideal and pragmatic performance are available in the Supplementary plots section S2.4.3.

The use of 25 imputed datasets to estimate the calibration intercept has little effect when compared to 5 imputed datasets ($\text{Slope}_{imp,M=5} - \text{Slope}_{obs} \approx \text{Slope}_{imp,M=25} - \text{Slope}_{obs}$). This can be seen for all methods across all data-generating scenarios for a sample size of 1000.

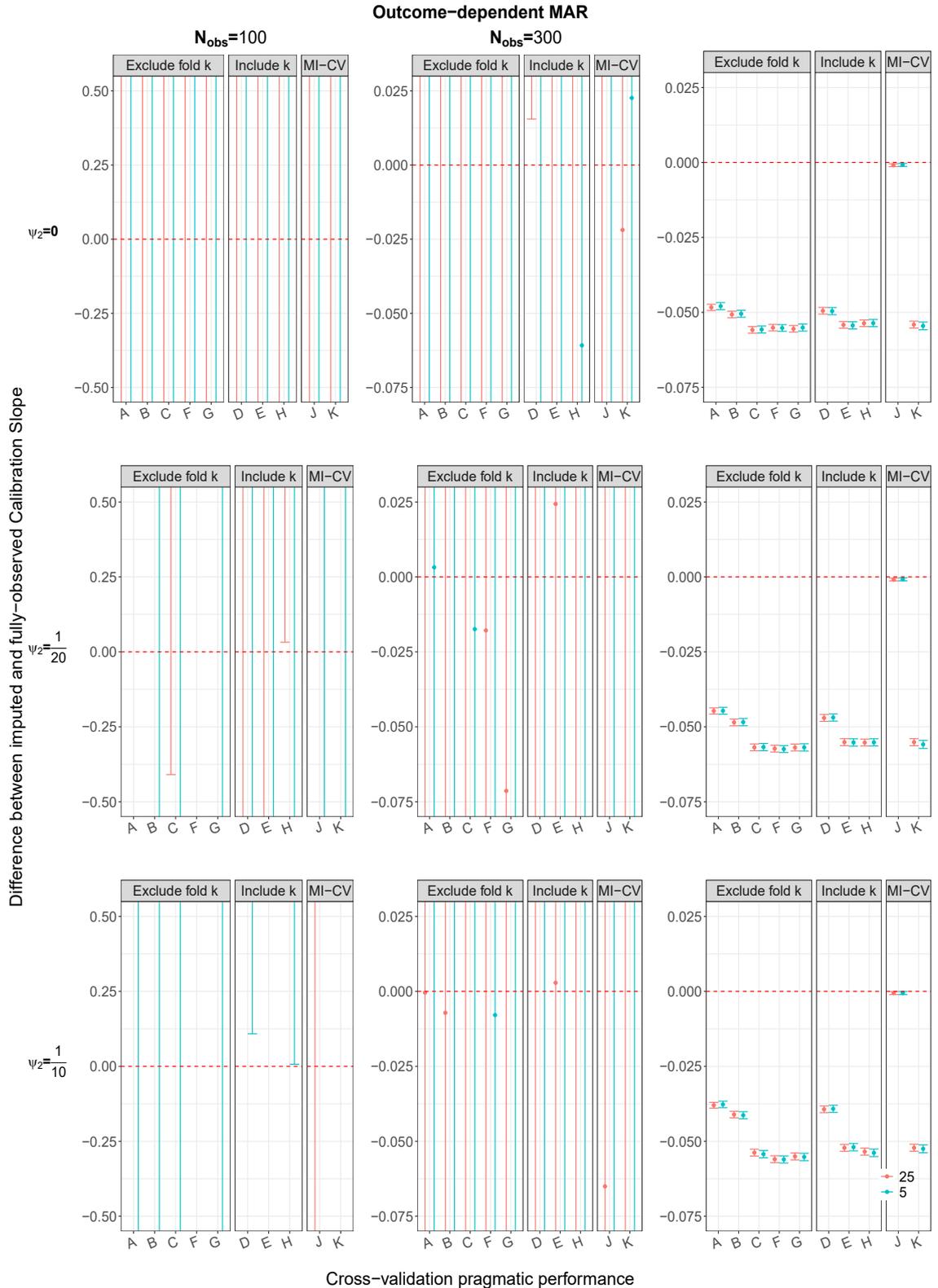


Figure 5.21: The difference $\text{Slope}_{imp} - \text{Slope}_{obs}$ when data are outcome-dependent or outcome- and covariate-dependent MAR for $M = 25$ versus $M = 5$ when 25% of values are missing in X_1 . The error bars summarise results from the 2000 repetitions for pragmatic performance and the limits represent the Monte Carlo 95% confidence interval of $\text{Slope}_{imp} - \text{Slope}_{obs}$. The average Calibration slope when data are fully-observed is 1.04 for larger sample sizes. CC (complete-case); methods A-K are described in Table 2.3 and summarised in Table 4.4.

5.6.3 Increasing the percentage of missingness to 40%

Figure 5.22 displays results demonstrating the impact that an increased percentage of missingness can have on the various methods when data are weakly outcome- and covariate-dependent MAR. The figure presents the calibration slope estimates when 25% or 40% of X_1 values are missing compared to Slope_{obs} ($\text{Slope}_{imp,\%} - \text{Slope}_{obs}$). The results are generally representative of the comparison between 25% and 40% missingness for ideal and pragmatic performance for all missing data scenarios and sample sizes. All plots are available in Section S2.4.2 of the Supplementary Plots.

For the complete-case analysis and pragmatic performance of all *CV-then-MI* methods (except method G), an increased percentage of missing values results in an increased magnitude of the difference between the estimated slope and Slope_{obs} ($|\text{Slope}_{imp,25\%} - \text{Slope}_{obs}| < |\text{Slope}_{imp,40\%} - \text{Slope}_{obs}|$). This holds across all missing data scenarios. Methods G, J and K tend to have similar performance in relation to Slope_{obs} , regardless of the percentage of missing values.

An increased percentage of missingness tends to result in an increased magnitude of the ideal performance for the majority of the *CV-then-MI* methods when data are MCAR or MAR. Methods A, B and D tend to have similar or small increases in magnitude with increased missingness ($|\text{Slope}_{imp,40\%} - \text{Slope}_{obs}|$) while methods C, E and F tend to have larger differences in magnitude. The *MI-then-CV* methods J and K perform similarly regardless of the percentage of missingness across all missing data scenarios.

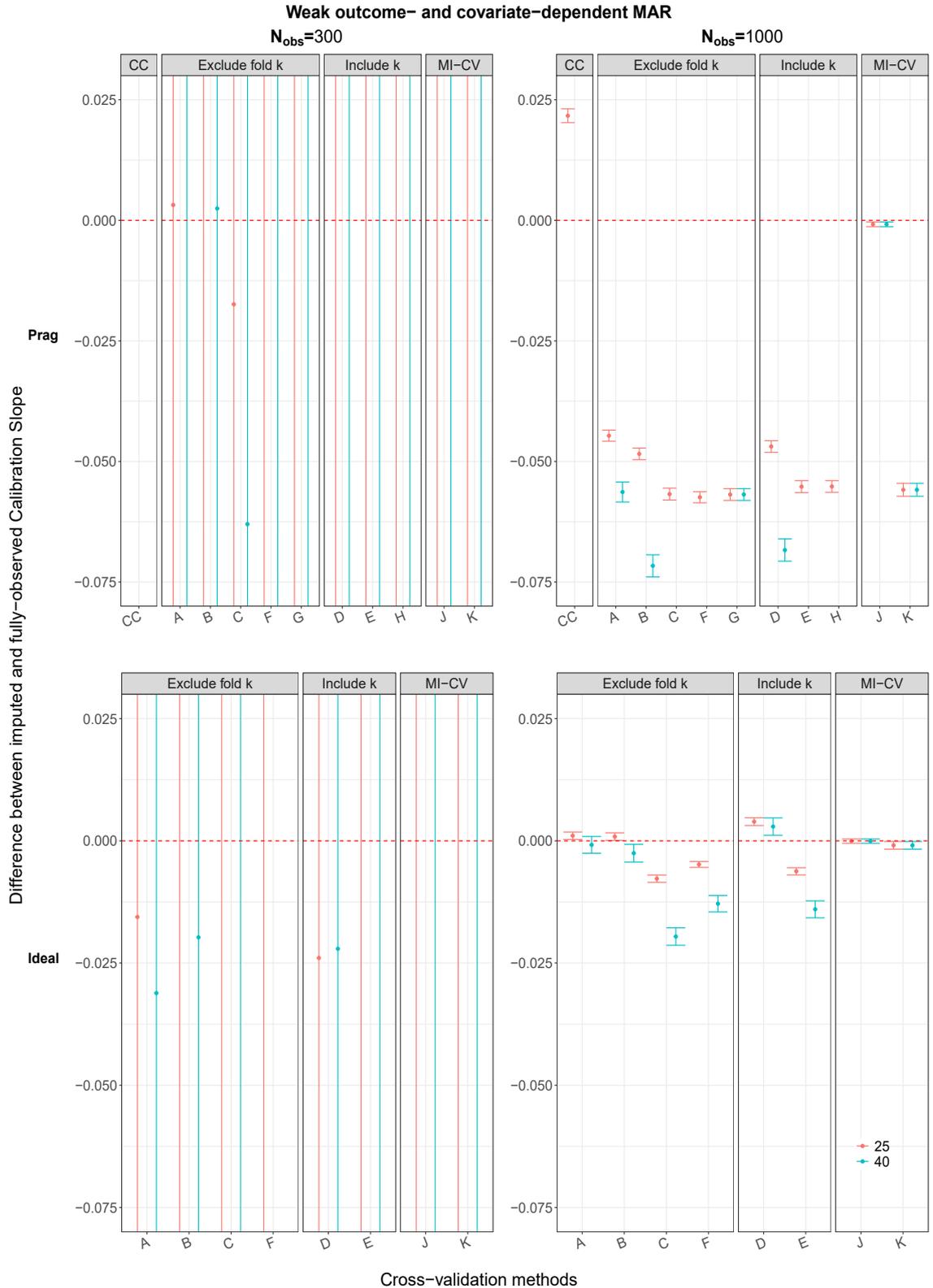


Figure 5.22: Comparing the impact of increasing the percentage of missingness on the difference $Slope_{imp} - Slope_{obs}$ when data are outcome- and covariate-dependent MAR when $M = 5$. The error bars summarise results from the 2000 repetitions and the limits represent the Monte Carlo 95% confidence interval of $Slope_{imp} - Slope_{obs}$. Red denotes $Slope_{imp} - Slope_{obs}$ when 25% of X_1 values are missing and blue denotes $Slope_{imp} - Slope_{obs}$ when 40% of X_1 values are missing. The top row presents the results for pragmatic performance and the bottom row presents results for ideal performance. The average Calibration slope when data are fully-observed is 1.04 for larger sample sizes. CC (complete-case); methods A-K are described in Table 2.3.

5.6.4 Comparing each method's calibration slope to the target estimate of the calibration slope from a larger validation set

As previously discussed for the AUC, Brier score and calibration intercept results, the ideal performance of the cross-validation imputation methods and Slope_{obs} were compared to the ideal target slope estimate ($\text{Slope}_{target,obs}$). This is estimated by applying a prediction model, based on all data in a repetition to the fully-observed data in the larger test set. The pragmatic performance of the imputation methods is compared to applying a repetition's prediction model to the imputed datasets of the larger test set ($\text{Slope}_{target,imputed}$). The complete-case estimate of the slope is compared to applying a repetition's prediction model to the observed cases of the larger test set ($\text{Slope}_{target,CC}$).

Similarly to the comparison of the methods with the Slope_{obs} , the slope estimates are unstable for small and moderate sample sizes. As such, the results will be discussed for sample size of 1000.

MCAR and covariate-dependent MAR

Figure 5.23 presents the ideal and pragmatic performance estimates of the various methods when compared to their respective target calibration slope estimate, when data are MCAR or covariate-dependent MAR.

The complete-case analysis overestimates $\text{Slope}_{target,CC}$ ($\text{Slope}_{CC} - \text{Slope}_{target,CC} > 0$) with a magnitude of approximately 0.9. The complete-case analysis estimate does not fit onto the scale of Figure 5.23 for the covariate-dependent MAR scenarios.

The pragmatic performance of all methods tends to overestimate $\text{Slope}_{target,imputed}$ for all methods. Method J tends to have the largest magnitude of overestimation ($|\text{Slope}_{J,prag} - \text{Slope}_{target,imputed}|$) of approximately 0.1 while the other methods tend to perform similarly with an overestimation of approximately 0.05.

The ideal performance of all methods also overestimates the target ideal estimate for all methods ($\text{Slope}_{imp,ideal} - \text{Slope}_{target,obs} > 0$). Again, the methods all tend to perform similarly in relation to $\text{Slope}_{target,obs}$, with magnitudes of approximately 0.05.

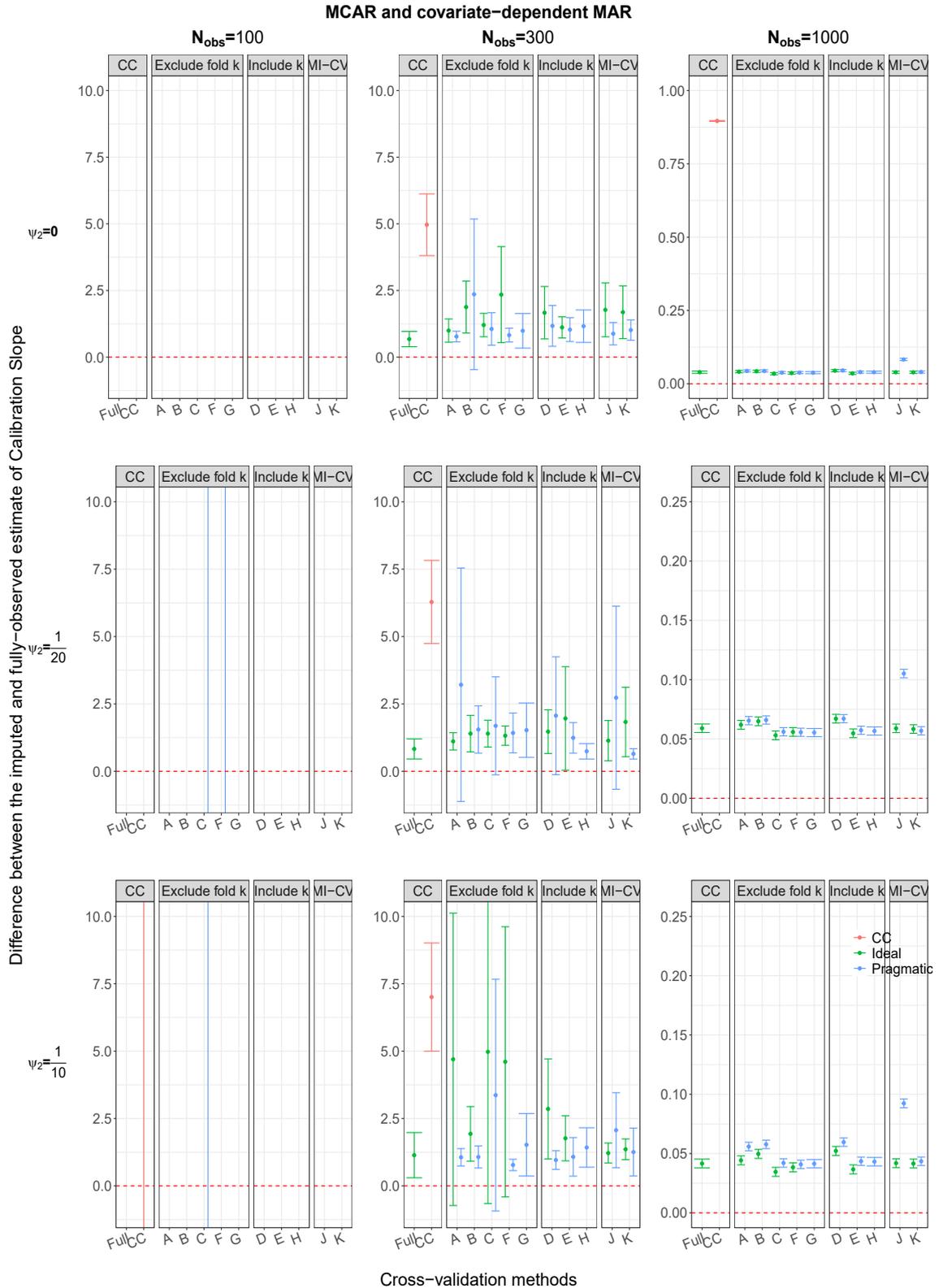


Figure 5.23: The difference $\text{Slope}_{imp} - \text{Slope}_{target}$ when data are MCAR or covariate-dependent MAR for $M = 5$ when 25% of values are missing in X_1 . The error bars summarise results from the 2000 repetitions and the limits represent the Monte Carlo 95% confidence interval of $\text{Slope}_{imp} - \text{Slope}_{target}$. The average Calibration slope when data are fully-observed is 1.8 for a sample size of 300 and 1.04 for a sample size 1000. CC (complete-case); methods A-K are described in Table 2.3.

Outcome-dependent MAR

Figure 5.24 presents the ideal and pragmatic performance estimates of the various methods when compared to their respective target calibration slope estimate, when data are outcome-dependent or outcome- and covariate-dependent MAR.

Similarly to the MCAR and covariate-dependent MAR scenarios, the complete-case analysis overestimates $\text{Slope}_{target,CC}$ ($\text{Slope}_{CC} - \text{Slope}_{target,CC} > 0$) and does not fit onto the scale of Figure 5.24 for the outcome- and weak/strong covariate-dependent MAR scenarios.

The pragmatic performance of all methods overestimates $\text{Slope}_{target,imputed}$ for all methods ($\text{Slope}_{imp,prag} - \text{Slope}_{target,imputed} > 0$). Method J has the largest magnitude of this difference. The other methods tend to perform similarly with methods A, B and D having a slightly larger magnitude than methods C, E, F-H and K but all approximately have a magnitude of 0.05. The ideal performance also overestimates $\text{Slope}_{target,obs}$ for all methods. The methods all perform similarly to each other when compared to $\text{Slope}_{target,obs}$.

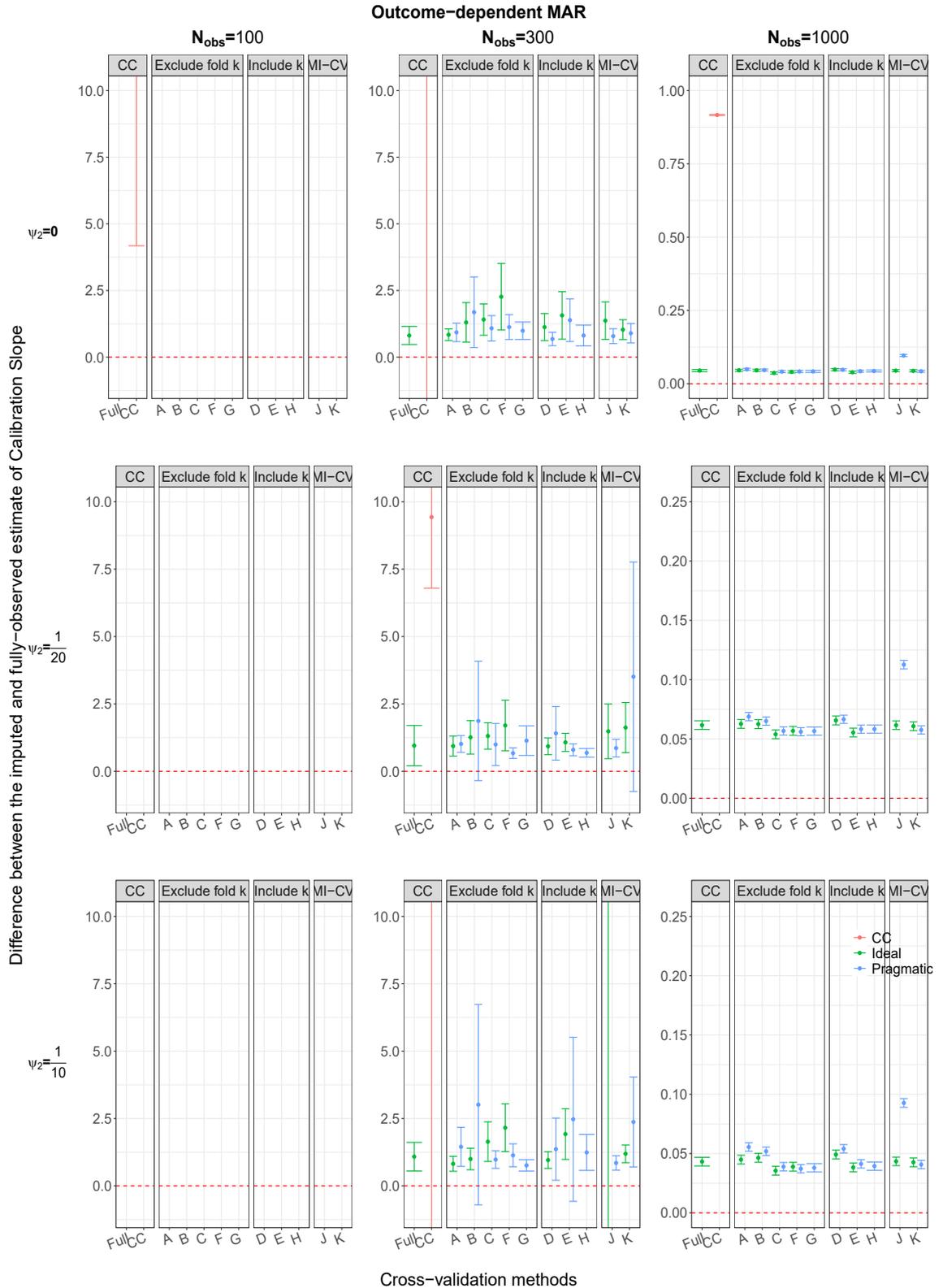


Figure 5.24: The difference $\text{Slope}_{\text{imp}} - \text{Slope}_{\text{target}}$ when data are outcome-dependent or outcome- and covariate-dependent MAR for $M = 5$ when 25% of values are missing in X_1 . The error bars summarise results from the 2000 repetitions and the limits represent the Monte Carlo 95% confidence interval of $\text{Slope}_{\text{imp}} - \text{Slope}_{\text{target}}$. The average Calibration slope when data are fully-observed is 1.8 for a sample size of 300 and 1.04 for a sample size 1000. CC (complete-case); methods A-K are described in Table 2.3.

5.7 Is data leakage an issue within the imputation process?

In section 2.8 I discussed the issue of data leakage in the imputation process and how we could investigate the impact of this leakage by comparing several methods, which were previously re-summarised in Table 4.4. The impact of data leakage was previously discussed for a continuous outcome scenario in Section 4.6.

The methods to compare data leakage range from having no leakage (method B) to those with the highest amount of leakage (method J). Methods A (which has no leakage) and F-H have no similar methods from which to compare the inclusion or exclusion of folds to assess the impact of data leakage when training or evaluating a prediction model and, therefore, will not be discussed here.

In the following analysis, method B will be compared with method C, and method D with method E. This comparison allows us to assess the impact of using only the observations available in the k^{th} test fold to impute the test fold (method B or D) versus using observations from all K folds to impute the k^{th} test fold (method C or E). Comparing method B with method D (or method C with E) will assess using the observations in the $k - 1$ training folds to impute the training set (method B or C) versus using all K folds to impute the $k - 1$ training folds before restricting to the $k - 1$ training folds to fit the prediction model. In other words, method B versus method C compares the use of all data on the imputation of the test fold while comparing method B versus D compares the inclusion of the test fold when imputing the training folds. Method E can be compared with method K to understand the impact of using two sets of imputed datasets for training and testing models but excluding values of the outcome (from the $k - 1$ training folds) to impute the test set to prevent data leakage. Method K can be compared with method J to assess the influence of using two sets of imputed datasets (one for training and the other for testing, method K) compared to using one set of imputed datasets (for both training and testing the prediction models, method J).

The comparisons of all methods (for example, method B versus method C) can be seen across all data-generating scenarios in Figures 5.1 and 5.2 for the AUC, in Figures 5.7 and 5.8 for the Brier score, in Figures 5.13 and 5.14 for the calibration intercept and finally, in Figures 5.19 and 5.20 for the calibration slope when comparing the performance measure of interest to the estimate when data are fully-observed. For the comparison of the methods' estimates with a target estimate in a larger validation set, please see Figures 5.5 and 5.6 for the AUC, Figures 5.11 and 5.12 for the Brier score, Figures 5.17 and 5.18 for the calibration intercept and Figures 5.23 and 5.24 for the calibration slope. For comparison purposes, I will focus on the weak outcome- and covariate-dependent scenario to investigate data leakage but the trends discussed will be generalisable across the majority of the missing

scenarios. All data leakage comparison graphs for the AUC, Brier score and calibration intercept and slope are available in Section S2.5 of the Supplementary Plots.

AUC

Figure 5.25 presents results assessing the impact of data leakage on the AUC by comparing methods and their inclusion or exclusion of training or test folds when data are weak outcome- and covariate-dependent MAR. The top row of the Figure compares the AUC with the AUC estimated when data are fully-observed ($AUC_{imp} - AUC_{obs}$) while the bottom row compares the AUC with the target ideal or pragmatic estimate of the AUC ($AUC_{imp} - AUC_{target}$).

Comparing $AUC_{imp} - AUC_{obs}$, the magnitude of this difference is two times larger for method B than method C for both ideal and pragmatic performance when sample size is 100 or 300. However, the absolute differences are small (approximately 0.0075 and 0.005 for sample sizes of 100 and 300, respectively). This can similarly be seen when comparing method D with E suggesting that using all the covariate data from all folds to impute the missing data in the k^{th} test fold has a strong impact on model performance. Increasing the sample size to 1000, this difference in magnitudes between methods B and C (or methods D versus E) is not as severe.

Methods B versus D and C versus E are compared to understand the impact of including the test fold when imputing the training folds. For all scenarios, method D has a smaller magnitude ($|AUC_D - AUC_{obs}|$) than B, similarly method E has a smaller magnitude than C. However, the differences between their magnitudes are all approximately 0.0025 when sample size is 100 or 300 and less than 0.001 for a sample size of 1000.

Method E versus K can be used to understand the impact of using two sets of imputed datasets for training and testing models but excluding values of the outcome (from the $k - 1$ training folds) to impute the test set to prevent data leakage. Methods E and K both have similar pragmatic performance across all sample sizes. However for the ideal performance, by removing values of Y from the training folds so that they are not used in the imputation of the test fold, method E performs poorly compared to method K which uses all of the outcome (from all K folds) to impute.

Comparing method K with J to assess the influence of using two sets of imputed datasets (one for training and the other for testing) compared to using one set of imputed datasets (for both training and testing the prediction models). Both methods perform similarly for ideal imputation, with method J having a smaller magnitude than K ($|AUC_{J,ideal} - AUC_{obs}| < |AUC_{K,ideal} - AUC_{obs}|$). For pragmatic performance, method J has a smaller

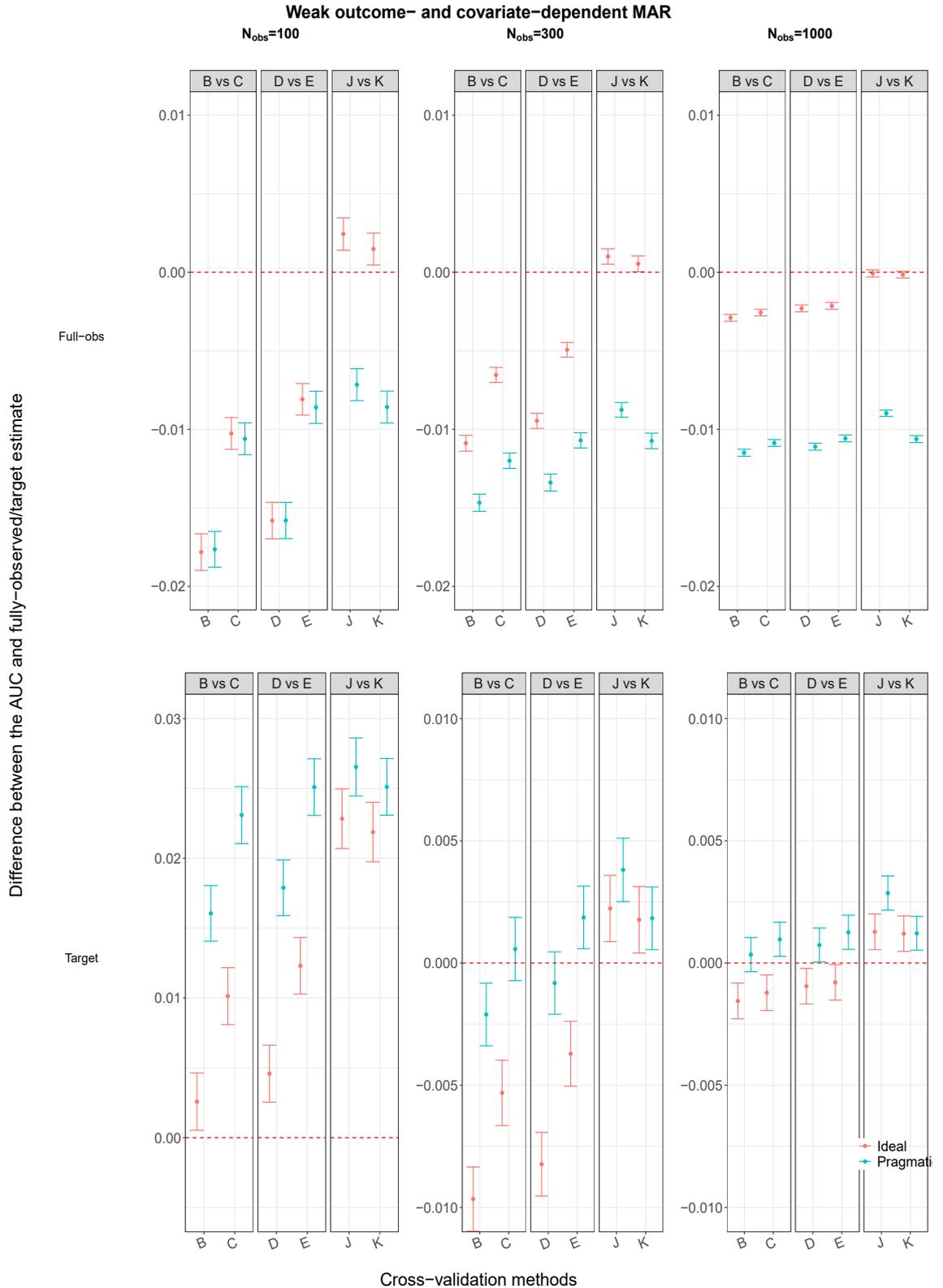


Figure 5.25: Assessing data leakage within the imputation process for cross-validation. The differences $AUC_{imp} - AUC_{obs}$ and $AUC_{imp} - AUC_{target}$ are compared when data are weak outcome- and strong covariate-dependent MAR. Methods are compared to both the AUC estimate when data are fully-observed (Full-obs, row 1) and the target estimate (Target, row 2) from a larger validation set. CC (complete-case); methods A-K are described in Table 2.3 and summarised in Table 4.4.

magnitude method K for the majority of scenarios. This is perhaps due to the increased correlation among imputed values used for training and testing in method J from using one set of imputed datasets. By contrast, method K uses two different imputation models and two sets of imputed datasets which are less correlated to each other.

When comparing the methods for handling missing data in a larger validation set and for a small sample size of 100, method B (no leakage) has a smaller magnitude $|AUC_B - AUC_{target}|$ than C (leakage when imputing the test set) and D (leakage when imputing the training folds) for both ideal and pragmatic imputation. Method K tends to have a smaller magnitude than method J for both ideal and pragmatic performance. Increasing the sample size to 300, the magnitude of the pragmatic performance has decreased to less than 0.005 for all methods. The ideal performance of method B is larger than method C and method D when sample size is 300. When the sample size is 1000 the various *CV-then-MI* methods have similar ideal and pragmatic performance. The pragmatic performance of methods J and K are larger than all other methods and method J has the largest magnitude of the difference for ideal performance. All methods for the ideal and pragmatic performance have a magnitude less than 0.005 when sample size is 1000.

Brier Score

The results for the Brier score are similar to those for the AUC when comparing data leakage using the various comparative methods. Figure 5.26 displays results comparing the Brier score to the Brier score when data are fully-observed and the target estimate from a validation set when data are weak outcome- and covariate-dependent MAR. The comparisons made are generally the same across DGMs.

For both ideal and pragmatic performance, methods C and D both have a smaller magnitude of the difference than method B when compared to $Brier_{obs}$. The ideal and pragmatic performance of method E has a smaller magnitude of the difference than method B. For pragmatic performance, both methods J and K overestimate $Brier_{obs}$. Method J (uses one set of imputed datasets to train and evaluate models) has a smaller magnitude of the difference than method K (uses two imputed datasets) for all sample sizes. However, the ideal performance of methods J and K underestimates $Brier_{obs}$ (i.e. they are over-optimistic) but method K has a smaller magnitude of the difference than method J.

For both ideal and pragmatic performance, methods C and D both have a smaller magnitude of the difference than method B when compared to $Brier_{target}$. The ideal performance of methods B, C and D either overestimate or approximate $Brier_{target,obs}$ while the ideal performance of method E tends to underestimate the target Brier score estimate (i.e. is over-optimistic). The ideal and pragmatic performance of methods J and K underestimates $Brier_{target}$ (i.e. they are over-optimistic). The ideal and pragmatic performance

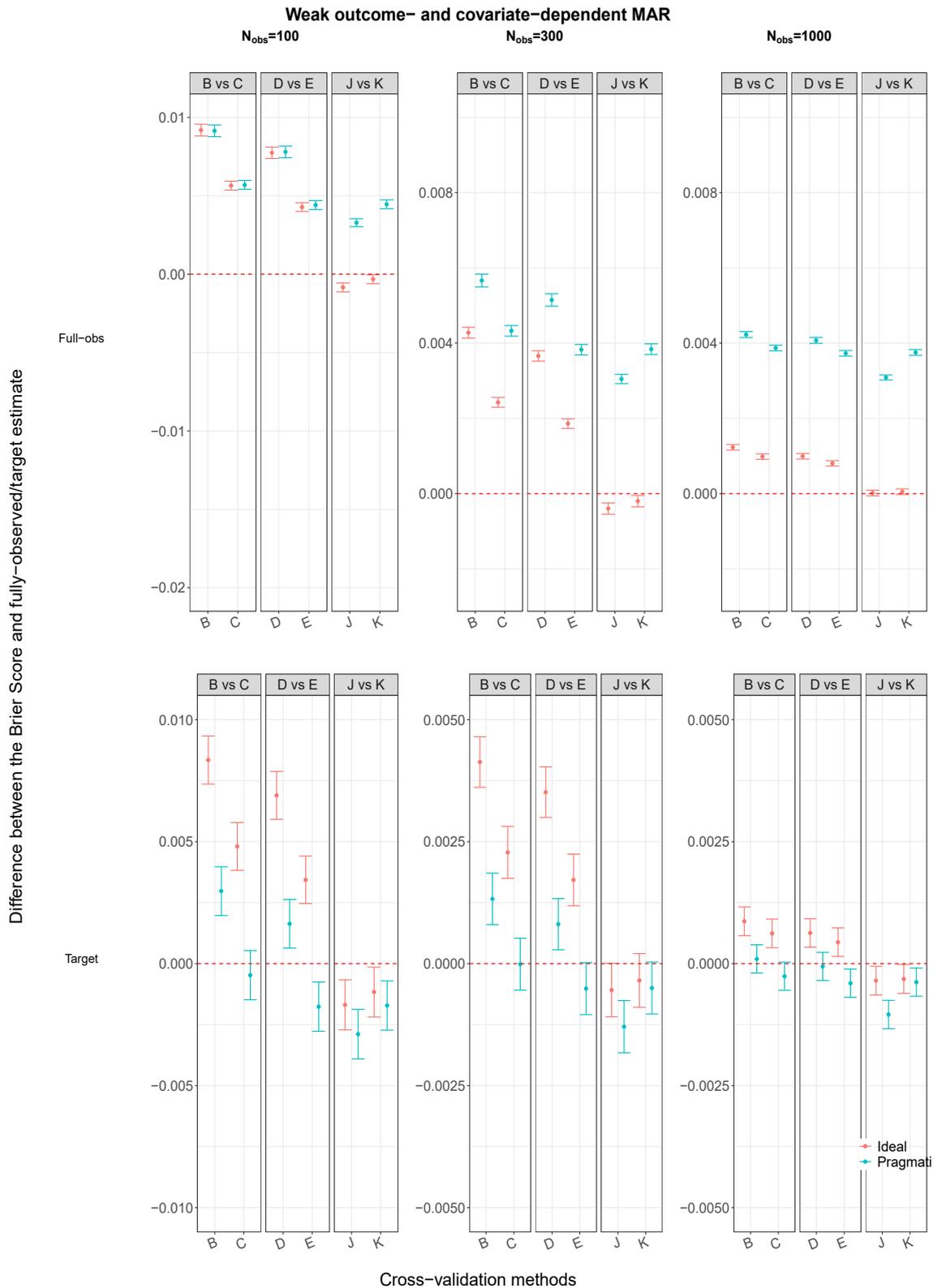


Figure 5.26: Assessing data leakage within the imputation process for cross-validation. The differences $Brier_{imp} - Brier_{obs}$ and $Brier_{imp} - Brier_{target}$ are compared when data are weak outcome- and strong covariate-dependent MAR. Methods are compared to both the Brier score estimate when data are fully-observed (Full-obs, row 1) and the target estimate (Target, row 2) from a larger validation set. CC (complete-case); methods A-K are described in Table 2.3 and summarised in Table 4.4.

of method J tends to be larger than the ideal and pragmatic performance of method K, respectively, when compared to Brier_{target} i.e. method J is more optimistic than method K.

For the majority of the scenarios the difference between the estimated Brier score and either Brier_{obs} or Brier_{target} is less than 0.01. The magnitude of the differences across methods between the estimated Brier score and either Brier_{obs} or $\text{Brier}_{target,obs}/\text{Brier}_{target,imputed}$ become more similar with increasing sample size.

Calibration intercept and slope

Figures 5.27 and 5.28 display data leakage comparisons for the calibration intercept and slope, respectively.

For the calibration intercept, both methods C and D had lower magnitudes of the difference than method B when comparing the estimated intercepts to Intercept_{obs} . With increasing sample size the difference between methods (when compared to Intercept_{obs}) decreases. The ideal and pragmatic performance of the *CV-then-MI* methods B, C, D, and E tend to overestimate the calibration intercept when compared to the Intercept_{obs} . The ideal performance of methods J and K is similar when compared to Intercept_{obs} . However, the pragmatic performance of method J has a smaller magnitude of the difference with Intercept_{obs} than method K. The pragmatic performance of all methods are somewhat similar when compared to $\text{Intercept}_{target,imputed}$. The ideal performance of methods B and D tend to overestimate $\text{Intercept}_{target,obs}$ while the other methods underestimate $\text{Intercept}_{target,obs}$ (with methods J and K having the largest magnitudes of the difference). With increasing sample size, the ideal performance of all methods approximates $\text{Intercept}_{target,obs}$.

The calibration slope had very large differences and variation in their estimates for both fully-observed and larger validation set comparisons when the sample size was small or moderate. For a sample size of 1000, method B has a smaller magnitude of the difference than method C when compared to Slope_{obs} for both ideal and pragmatic performance. Similarly, method D has a smaller magnitude of the difference than method E when compared to Slope_{obs} . However, when comparing to the larger validation set method C has a smaller magnitude of the difference than method B when compared with Slope_{target} ($\text{Slope}_{target,obs}$ or $\text{Slope}_{target,imputed}$). Methods J and K have similar ideal performance when their estimated slopes are compared with Slope_{obs} or $\text{Slope}_{target,obs}$. For pragmatic performance, the magnitude of the difference between the estimated slope and Slope_{obs} is smaller for method J than method K. However, when the estimated slopes are compared to $\text{Slope}_{target,imputed}$, the magnitude of the difference is smaller for method K, while method J has the largest magnitude of the difference across all methods being compared.

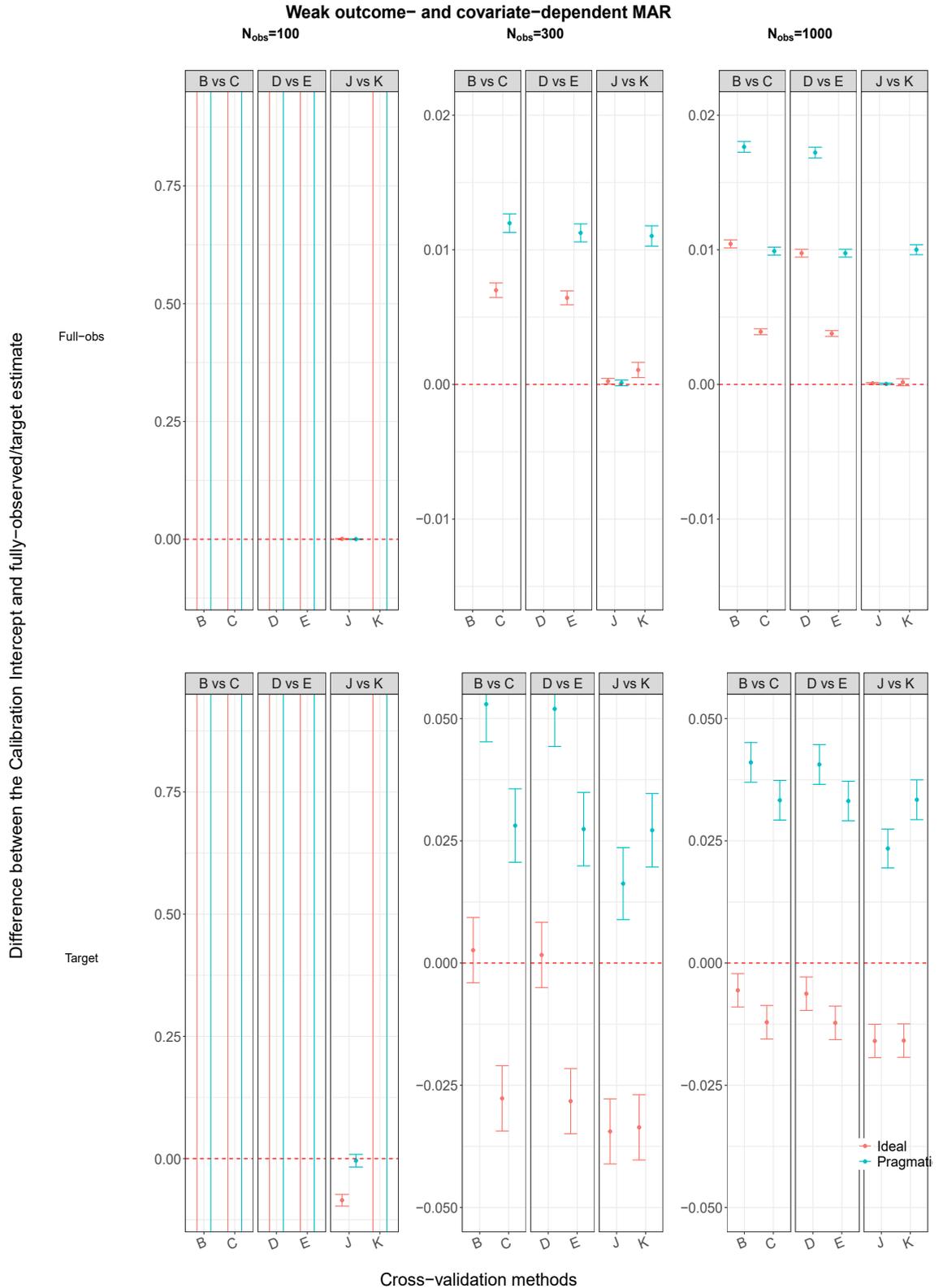


Figure 5.27: Assessing data leakage within the imputation process for cross-validation. The differences $\text{Intercept}_{imp} - \text{Intercept}_{obs}$ and $\text{Intercept}_{imp} - \text{Intercept}_{target}$ are compared when data are weak outcome- and strong covariate-dependent MAR. Methods are compared to both the calibration intercept estimate when data are fully-observed (Full-obs, row 1) and the target estimate (Target, row 2) from a larger validation set. CC (complete-case); methods A-K are described in Table 2.3 and summarised in Table 4.4.

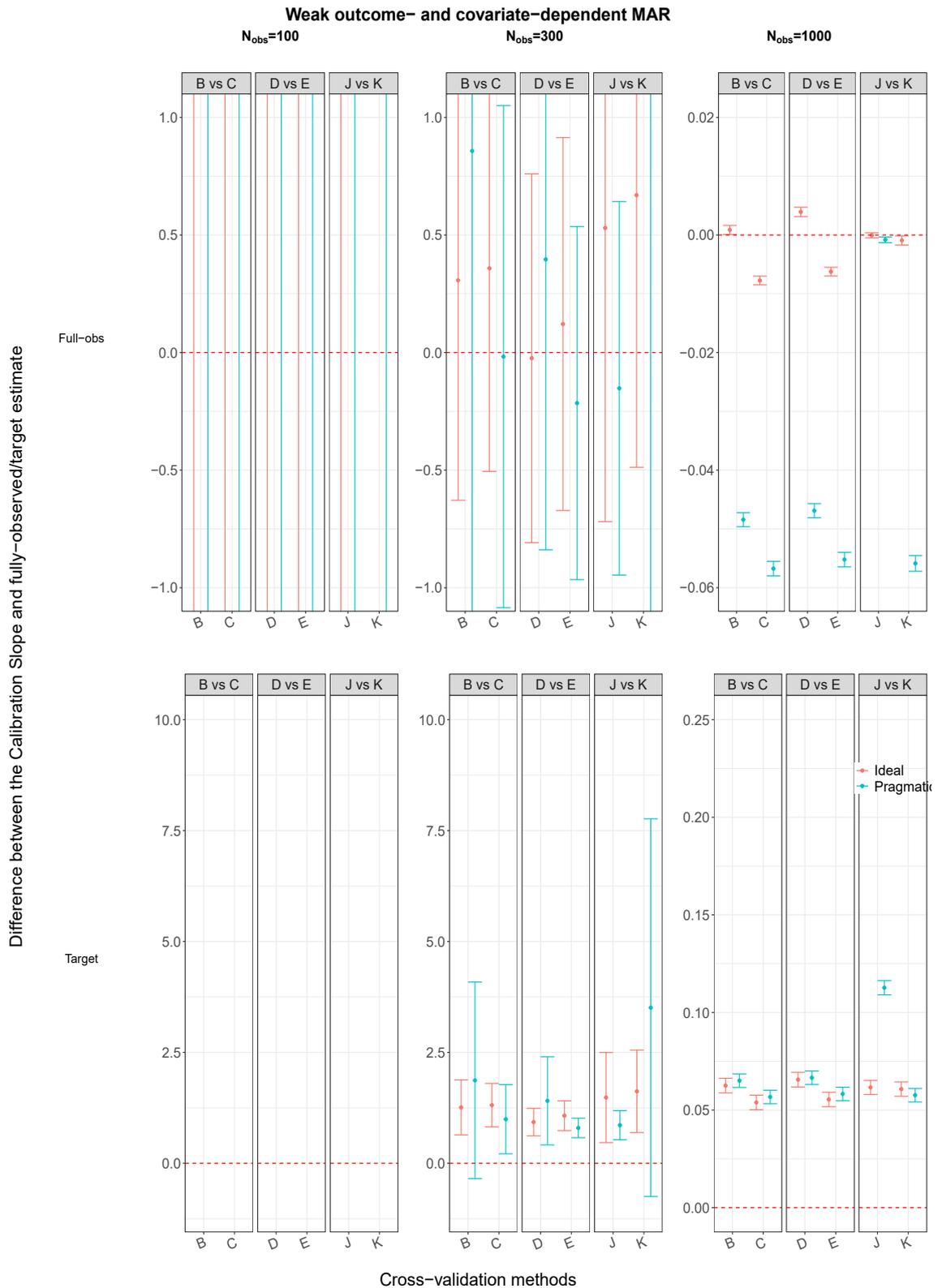


Figure 5.28: Assessing data leakage within the imputation process for cross-validation. The differences $Slope_{imp} - Slope_{obs}$ and $Slope_{imp} - Slope_{target}$ are compared when data are weak outcome- and strong covariate-dependent MAR. Methods are compared to both the calibration slope estimate when data are fully-observed (Full-obs, row 1) and the target estimate (Target, row 2) from a larger validation set. CC (complete-case); methods A-K are described in Table 2.3 and summarised in Table 4.4.

5.8 Discussion of results for the binary outcome

The aim of this study was to investigate the appropriate ways to combine MI and cross-validation for a binary outcome when the three performance measures of interest were the AUC, Brier score and Calibration intercept and slope. In addition, an analysis of the impact of data leakage on the imputation process by comparing various methods was also investigated.

Overall, all imputation methods had a tendency to underestimate the AUC estimate when data are fully-observed. For the AUC, a higher value closer to 1 usually indicates good model performance. An exception to the underestimation of AUC_{obs} was the *MI-then-CV* methods which occasionally over-estimated the performance or were over-optimistic (i.e. stated that the model performed better after imputation than it did when data were fully-observed). For the Brier score, a smaller score indicates better model performance. In general, the various *CV-then-MI* methods tended to overestimate the Brier score when data are fully-observed. The *MI-then-CV* methods tended to underestimate the Brier score for small sample sizes or were over-optimistic i.e. the imputed model states better performance than if data had been fully-observed. Methods J and K have the highest levels of data leakage, the missing values have been imputed using knowledge of the outcome and covariates in the test folds. This increases any correlation between the imputed values and the values of the outcome in the test fold. Therefore, any prediction model trained using these imputed values will have an unfair advantage when it is evaluated in the test fold - hence the model having a more optimistic performance measure estimate after imputation, than if the data had never been missing to begin with.

For the calibration intercept and slope, deviations away from 0 and 1, respectively, can indicate poor performance. For large sample sizes the majority of the imputation methods overestimated the intercept by less than 0.02 and underestimated the slope by less than 0.08. For small and moderate sample sizes, calibration performance was poor even when data were fully-observed, with an average slope value of 80 (Table 5.1), however this decreased to values between 1 and 2 for larger sample sizes. The issue of unstable calibration results appeared to result from small sample sizes. This is supported by Van Calster et al. [52] and Riley et al. [30] who both concluded that small sample sizes can lead to miscalibration of predictions and that shrinkage methods do not help to resolve this issue.

The ideal performance of methods C and E performed poorly when assessing the MSE for the continuous outcome. However, their performance measures in the binary scenario did not react similarly. As C and E had to additionally impute Y , which was set to be 90% missing, the imputed value for Y of 0 or 1 was far less variable than in the continuous case

(see variance of continuous Y in Table 4.1) which may have resulted in better imputations.

5.9 Conclusions

This chapter aimed to assess methods for combining MI and cross-validation. It was shown that for all performance measures, complete-case analysis performs poorly in certain MAR scenarios. The consequences of using a complete-case analysis could be over-optimistic estimates of performance for the AUC or Brier score when data are covariate-dependent MAR or outcome- and covariate-dependent MAR. It may also result in a larger magnitude of the calibration intercept or slope than the best performing imputation methods.

The effects of data leakage in the imputation process were assessed for the AUC, Brier score and calibration intercept and slope. Data leakage was not an issue for method A or B (*CV-then-MI methods*) while methods J and K (*MI-then-CV*) had the most leakage out of all the methods. Methods C-H had some form of data leakage through the use of the training or test folds in the imputation process.

Method A had better performance than method B for small and moderate sample sizes, while both had comparable methods of performance for larger sample sizes and tended to perform similarly to the methods that had an “advantage” due to data leakage.

For small sample sizes, the ideal performance of *MI-then-CV* (methods J and K) tended to underestimate the MSE suggesting that the prediction model performed better post imputation than if the data had been fully-observed (i.e. the methods are over-optimistic) and I have suggested that this is a direct result of data leakage in the imputation process. This over-optimism was similarly seen for a small sample size when the performance measure of interest is the AUC or Brier score. All other imputation methods tended to state that the prediction model did not perform as well post-imputation than if the data had been fully-observed.

In agreement with the previous literature, I propose that *MI-then-CV* methods are at the highest risk of data leakage and should be avoided. Method A is at no risk of data leakage and is the best method to combine MI and cross-validation for moderate sample sizes. With larger sample sizes, method B performs similarly to method A and also avoids data leakage.

6 Simulation study results for the bootstrap: continuous outcome

In Chapter 3 I described the design of a simulation study to investigate the performance of various methods which combined MI with an internal validation method. Results for combining MI with cross-validation were then presented and discussed in Chapters 4 and 5.

6.1 Introduction

This chapter will present the results from combining MI with the optimism-corrected bootstrap algorithm, including the *standard* and *0.632* versions. As in the previous chapters, the impact of data leakage from the imputation process will be assessed. In addition, the reuse of imputed datasets to estimate performance will be assessed. The findings from the simulation study for the continuous outcome will be presented and, due to the quantity of results produced, all graphs are available in the supplementary plot chapter (Section S3), in addition to the graphs presented in this chapter. The simulation results for the binary outcome are presented in Appendix C.

The methods which will be evaluated in this chapter were fully described in Section 2.7 but are re-summarised in Table 6.1. The methods fall into two broad classes- those in which the imputation is performed first (*MI-then-BS*), and those in which the bootstrap samples are obtained first (*BS-then-MI*). The methods detailed in Table 6.1 focus on the estimation of the bootstrap and test performance. To estimate the apparent performance, for all methods, the original dataset is imputed using a training imputation model. A prediction model is fitted to each imputed dataset m_{train} for $m_{train} = 1, \dots, M_{train}$. Each prediction model is then evaluated in M_{test} imputed datasets which were imputed using a test imputation model (which will include the outcome or not depending on whether ideal or pragmatic performance is of interest). Rubin's first rule will get an overall performance estimate for each of the prediction models. Rubin's first rule is used again to average across the performance estimate for the M_{train} prediction models to get the apparent performance estimate.

The methods detailed in Table 6.1 will use training and test imputation models (Section 2.5). The training imputation model will include any relevant covariates and the outcome. The test imputation model will always include any relevant covariates but may, or may not, include the outcome depending on whether the ideal or pragmatic performance is of interest.

Table 6.1: A brief summary of the various methods under consideration to combine the bootstrap (BS) algorithms with multiple imputation (MI). The methods below describe how to estimate the bootstrap and test performance of the internal validation for a single bootstrap sample, they will be repeated B times.

| Methods | Description |
|-----------------------|---|
| BS first | |
| BS-then-MI (default) | Draw a BS sample. Impute the BS sample using separate training and test imputation models. Using the training imputed BS sample, fit a prediction model. Evaluate the prediction model using the test imputed BS sample to estimate the BS performance (for the <i>standard</i> method). Evaluate the prediction model in the test imputed original dataset (<i>standard</i>) or those who were not selected to be in the BS sample (0.632) to estimate the test performance. |
| BS-then-MI reuse imps | Same process as <i>BS-then-MI</i> . However, instead of imputing the BS sample using a training and test imputation model, reuse the imputed datasets used to estimate the apparent performance and sample the observations from these imputed datasets that were selected to be in the BS sample. Train a prediction model within the BS sample and evaluate it in the same way as <i>BS-then-MI</i> . |
| Impute first | |
| MI-then-BS (default) | The original dataset is imputed using a training and test imputation model (these are used to first estimate the apparent performance). Take a bootstrap sample of one of the training imputed datasets m_{train} and fit a prediction model. Impute the BS sample using the test imputation model to estimate the BS performance of the prediction model. Evaluate the prediction model in the original test imputed dataset (<i>standard</i>) or those who were not selected to be in the BS sample (0.632) to estimate the test performance. This will be repeated for B bootstrap samples of the training imputed dataset m . This will be iterated for $m_{train} = 1, \dots, M_{train}$ |
| MI-then-BS fixed BS | Same process as <i>MI-then-BS</i> , except that the same B BS samples are used within each of the training imputed datasets. |

Table 6.1: A brief summary of the various methods under consideration to combine the bootstrap (BS) algorithms with multiple imputation (MI). Continued.

| Methods | Description |
|----------------------------|--|
| Impute first | |
| MI-then-BS reuse test imps | Same process as <i>MI-then-BS</i> . However, instead of imputing the BS sample using the test imputation model to estimate the BS performance, reuse the test imputed datasets of the original dataset, select the relevant BS observations and estimate the BS performance. |
| MI-then-BS re-impute | Same process as <i>MI-then-BS</i> except instead of reusing the original training imputed datasets, re-impute the original dataset using a training imputation model. Apply the <i>MI-then-BS</i> procedure to the second set of training imputed datasets. |
| MI-then-BS impute once | Use one set of imputed datasets to estimate the apparent, bootstrap and test performance. Use an imputation model which will either include or exclude the outcome, depending on whether ideal or pragmatic performance is of interest. |

Recall from Section 1.9.4 that the *standard* bootstrap algorithm has three performance measures (apparent, bootstrap and test) whereas the 0.632 method has two (apparent and test). The test performance estimated within the *standard* algorithm uses all observations from the original dataset, whereas the 0.632 test performance is calculated in those observations not selected for the BS sample.

The methods in Table 6.1 detail how to train a prediction model in a bootstrap sample and, if using the *standard* bootstrap algorithm, how to estimate the bootstrap performance. In addition to the apparent and bootstrap performance estimates, the test performance must also be estimated. I initially proposed in Section 2.7.1 to impute the original dataset a second time using a test imputation model (including or excluding the outcome dependent on whether the ideal or pragmatic performance is of interest) to get a second set of imputed test datasets i.e. re-imputing test datasets. This second set would be used to evaluate the prediction model trained in the bootstrap sample. An alternative to using a second set of imputed test datasets for the methods described in Table 6.1 is to reuse the imputed test datasets which were used to estimate the apparent performance.

In the following section, I will show results that compare the reuse of imputed test datasets versus re-imputing the dataset a second time using a test imputation model. I will then briefly describe the results for one *BS-then-MI* method and one *MI-then-BS* method as

a gentle introduction for the reader before presenting the results for all methods in the simulation study. These results will include comparing the MSE obtained using each method with the MSE estimated when data are fully-observed and also with a ‘target value’ estimated from a larger validation set. The use of an increased number of imputed datasets will be assessed, as will the impact of an increased percentage of missing values. Finally, I will compare data leakage through the imputation process for the *BS-then-MI* and *MI-then-BS* methods before presenting a discussion of the results.

Results from the simulation study

Similarly to the simulation study that combined MI and cross-validation, several factors were varied for the continuous outcome setting (including sample size, levels of R-squared and dependence of missingness). The same simulated data used when evaluating cross-validation techniques were used to evaluate the bootstrap methods. The summary information on the outcome and MSE for the 2000 repetitions using the fully-observed data can be found in Tables 4.1 and 4.2. Recall from section 3.5 the notation for the averaged estimate (across the 2000 repetitions) of the MSE in the full data (MSE_{obs}) and the larger validation set (MSE_{target}). In addition, $MSE_{prag,imp}$ denotes the pragmatic performance of a proposed method *imp* and $MSE_{ideal,imp}$ denotes the ideal performance where *imp* denotes various methods such as the complete-case analysis (CC), *BS-then-MI* (BS-MI) or *MI-then-BS* (MI-BS) methods.

6.2 A brief overview of the *BS-then-MI* and *MI-then-BS* methods for the *standard* bootstrap algorithm

Due to the large number of results from the simulation studies presented in this chapter which assess the various methods under multiple data-generating scenarios, I will first present results for two methods when $R^2 = 0.1$. The aim is to introduce the reader to how the results are being displayed and interpreted as well as introducing the impact that data leakage can have on the results. The analysis is similar to that in Section 4.4.

I will briefly compare method *BS-then-MI* (the default version which has no data leakage) to method *MI-then-BS impute once* (a method considered to have the most opportunity for leakage). Figure 6.1 displays results for these two methods for the various missing data scenarios when $R^2=0.1$. The MSE results from each method are compared to the estimates of the MSE from applying the *standard* bootstrap algorithm to the fully-observed data i.e. $MSE_{imp} - MSE_{obs}$.

For all sample sizes and missing data scenarios, the estimated pragmatic performance of both *BS-then-MI* and *MI-then-BS impute once* overestimates the MSE value estimated by the *standard* bootstrap algorithm when data are fully-observed i.e. $MSE_{prag,imp} - MSE_{obs} > 0$. Method *BS-then-MI* tends to overestimate the MSE to a greater degree ($|MSE_{prag,BS-MI} - MSE_{obs}|$) than *MI-then-BS impute once* for all sample sizes. However, with increasing sample size the magnitude of the difference ($|MSE_{prag,imp} - MSE_{obs}|$) for both methods decreases and the difference becomes more similar between the two methods. This can be seen across all missing data scenarios for $R^2 = 0.1$.

The estimated ideal performance of method *BS-then-MI* tends to overestimate the MSE when data are fully-observed for all sample sizes. However, *MI-then-BS impute once* underestimates MSE_{obs} for all sample sizes. This means that the results from *MI-then-BS impute once* are over-optimistic for ideal performance i.e. the method gives better performance post-imputation than what would have been observed if missing data were not present. The magnitudes of under- or overestimation of the two methods are similar across all sample sizes and missing data scenarios.

In the following section, I will present a summary of results for all methods in a similar manner as above. Recall that a ‘good’ prediction model would have a lower MSE score. Therefore, over-estimation of MSE_{obs} implies worse performance after handling missing data than if the data had all been observed to begin with. Under-estimation of MSE_{obs} suggests that the method is over-optimistic; that is, it is performing better than if we had observed the data. In the following results, it will be shown that many of the methods which are subject to data leakage tend to have over-optimistic ideal performance.

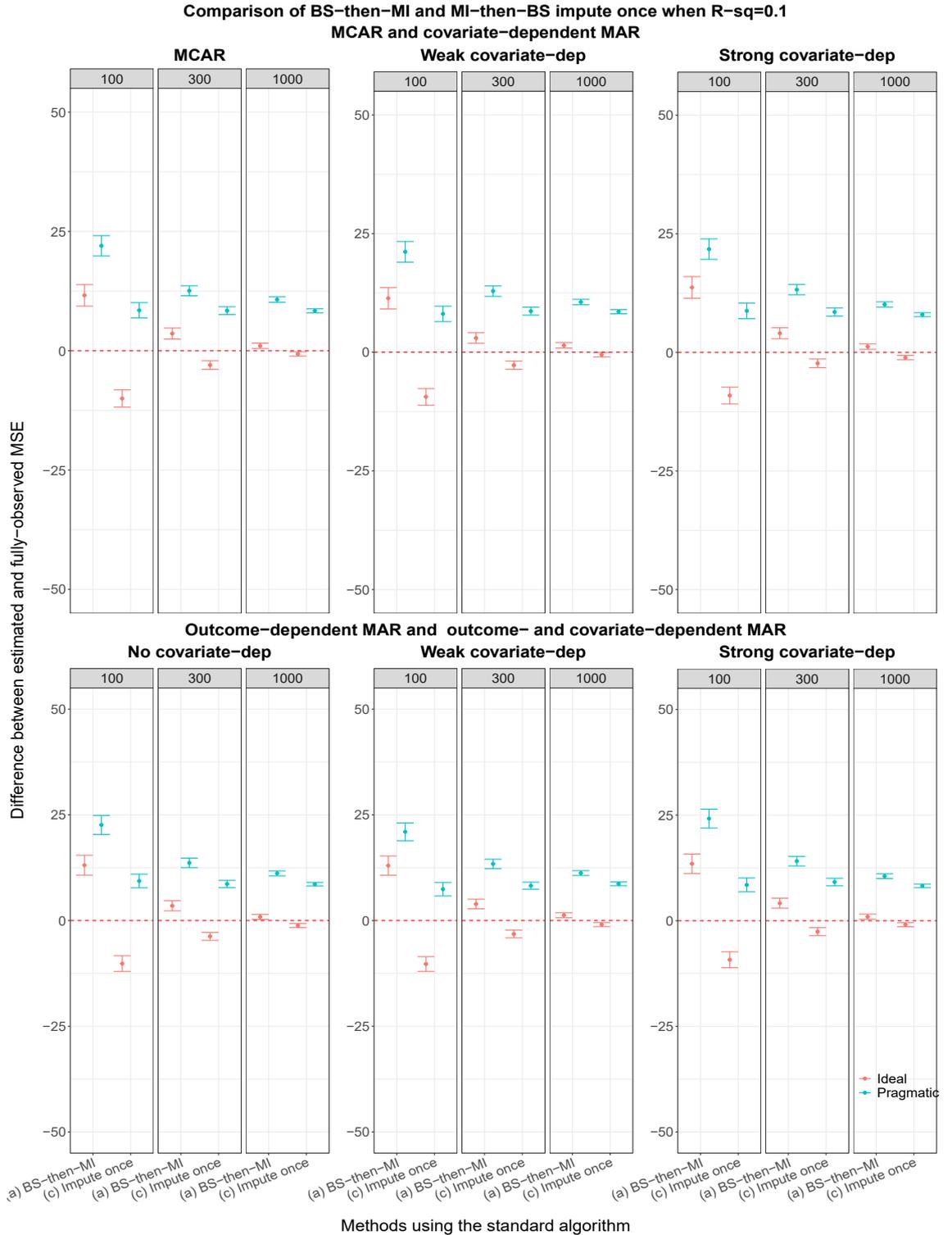


Figure 6.1: The difference $MSE_{imp} - MSE_{obs}$ when $R^2 = 0.1$ for $M = 5$ when 25% of values are missing in X_1 . Each sub-graph displays results for a sample size of 100, 300 and 1000. Row 1 presents results when data are MCAR or covariate-dependent MAR. Row 2 presents results when data are outcome-dependent MAR or outcome- and covariate-dependent MAR. Ideal performance is in red and pragmatic performance is in blue. The error bars summarise results from the 2000 repetitions and the limits represent the Monte Carlo 95% confidence interval of $MSE_{imp} - MSE_{obs}$. CC (complete-case); methods are described in Section 2.7 or Table 6.1.

6.3 A comparison of reusing versus re-imputing test datasets: *standard* bootstrap algorithm

The variations of reusing imputed datasets or imputing bootstrap samples in Table 6.1 focused on ensuring the bootstrap sample has imputed values with which to train a prediction model in the bootstrap sample and estimate the bootstrap performance. This section will focus on reusing imputed test datasets versus re-imputation when estimating the test performance of the bootstrap’s prediction model (this uses all observations from the original dataset). This comparison is specifically for the *standard* bootstrap internal validation algorithm as the test performance of the 0.632 algorithm uses the observations which were not selected for the bootstrap sample. The default methods of *MI-then-BS* and *BS-then-MI* take the non-sampled observations and impute them independently for the test performance of the 0.632 algorithm.

When combining the *standard* bootstrapping algorithm with MI, I initially proposed (Section 2.7.1) obtaining two imputed versions of the original dataset using the test imputation model. The first set of imputed datasets would be used to estimate the apparent performance while the second set would be used to estimate the test performance. However, the test imputed datasets used to estimate the apparent performance could also be used for estimating the test performance, instead of re-imputing the original dataset again. This is investigated for all *standard* bootstrap methods and involves using a training and test imputation model (i.e. all methods except for *MI-then-BS impute once* which involves one set of imputed datasets).

Figure 6.2 shows the simulation results for the ideal performance of the methods for the simulation scenario in which missingness is outcome-dependent and $R^2 = 0.1$. The figure shows the performance of the same bootstrap methods which have either reused the test imputed datasets used to estimate apparent performance or imputed the entire dataset a second time. The results in the figure are representative of the results for ideal and pragmatic performance across all scenarios (available in supplementary plots section S3.1). There is no difference between reusing and re-imputing the test imputed datasets for the *standard* bootstrap algorithm, and this holds for all scenarios.

The reuse of test imputed datasets for estimating the test performance is more computationally efficient than re-imputing, therefore all subsequent results for the *standard* bootstrap algorithm presented below are based on reusing the test imputed datasets.

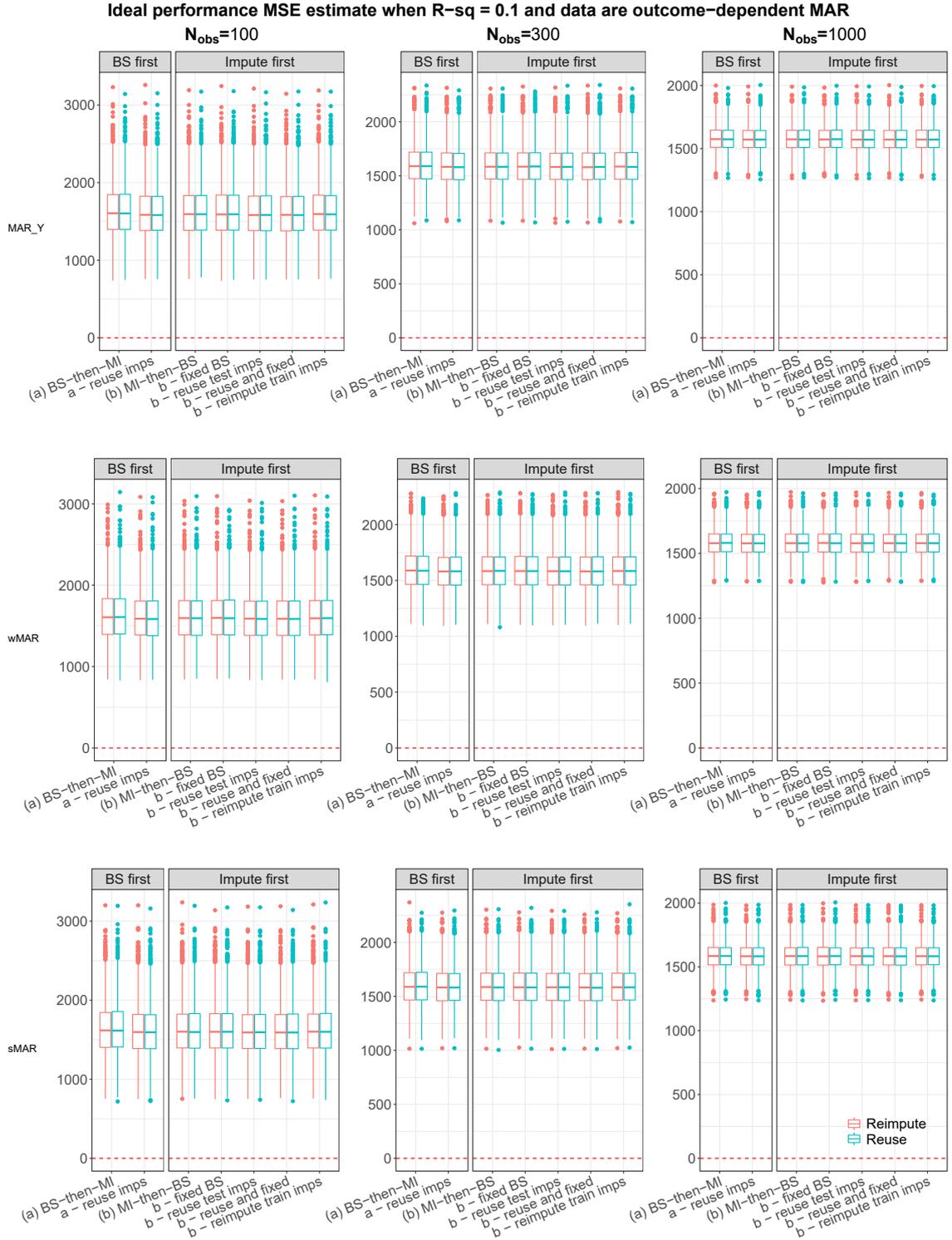


Figure 6.2: A comparison of reusing versus re-imputing test datasets on the MSE estimates for the *standard* bootstrap algorithm. The boxplots display the estimates of the MSE from 2000 repetitions for each method which are compared to the MSE estimated when data are fully-observed ($MSE_{imp} - MSE_{obs}$). The results are for data which are outcome-dependent MAR (row 1), weak outcome- and covariate-dependent MAR (row 2) and weak outcome- and strong covariate-dependent MAR. CC (complete-case); methods are described in Section 2.7 or Table 6.1.

6.4 Detailed results for the *standard* bootstrap algorithm

In this section I will summarise the results of the simulation study for the *standard* bootstrap algorithm. I will explore how the methods perform using the estimated MSE, which is compared to both the MSE when data are fully-observed and also a target MSE value from a larger validation set when 25% of values in X_1 are missing and 5 imputed datasets are used. In addition, I will analyse whether increasing the number of imputed datasets improves results and also examine how the methods perform with an increased percentage of missingness. The methods presented were summarised in full in Section 2.7 and briefly resummaries above in Table 6.1.

6.4.1 Comparing results to the MSE estimate when data are fully-observed

MCAR and covariate-dependent MAR

Figure 6.3 presents results for the various methods to handle missing data alongside bootstrap validation when compared to the MSE estimate when data are fully-observed ($\text{MSE}_{imp} - \text{MSE}_{obs}$) when data are weak covariate-dependent MAR. The results in the graph are representative of those scenarios in which missing data are not outcome-dependent (MCAR, weak and strong covariate-dependent MAR, Supplementary plot section S3.2.1).

When data are MCAR or covariate-dependent MAR and for a sample size of 100, the complete-case analysis tends to overestimate MSE_{obs} ($\text{MSE}_{CC} - \text{MSE}_{obs} > 0$). The estimates of performance from the complete-case analysis are more variable than those from the MI based methods. With increasing sample size, variability is reduced and the difference, $\text{MSE}_{CC} - \text{MSE}_{obs}$, tends to zero.

For pragmatic performance of imputation methods, when sample size is small *BS-then-MI* tends to have the largest difference between its MSE and MSE_{obs} from the *standard* algorithm ($\text{MSE}_{prag,BS-MI} - \text{MSE}_{obs}$) for all values of R^2 . *MI-then-BS impute once* tends to have the smallest difference for pragmatic performance while all other imputation methods tend to have similar performance when compared to MSE_{obs} for all values of R^2 . With increasing sample size all imputation methods tend to have similar pragmatic performance, with method *MI-then-BS impute once* having the smallest difference ($\text{MSE}_{prag,MI-BS-once} - \text{MSE}_{obs}$) throughout all scenarios.

For ideal performance, when sample size is small all methods which involve reusing the imputed datasets of the original training and test datasets (used to estimate apparent performance) to fit a model to the bootstrap sample and evaluate it tended to underestimate MSE_{obs} . In other words, $\text{MSE}_{ideal,imp} - \text{MSE}_{obs} < 0$ when *imp* = *BS-then-MI reuseimps*, *MI-then-BS reuse testimps*, *MI-then-BS reuse testimps with fixed BS samples* and *MI-then-BS impute once*. Method *BS-then-MI* tends to overestimate MSE_{obs}

($\text{MSE}_{ideal,BS-MI} - \text{MSE}_{obs} > 0$) while *MI-then-BS* and its variations, which include using fixed bootstrap samples and re-imputing the train imputed datasets which are bootstrap sampled, tend to give MSEs that are close to MSE_{obs} . This is true for all values of R^2 that I considered. With increasing sample size, all methods tend to give an MSE that is similar to MSE_{obs} ($\text{MSE}_{ideal,imp} - \text{MSE}_{obs} \xrightarrow{n_{obs} \rightarrow \infty} 0$ where *imp* represents any one of the methods considered).

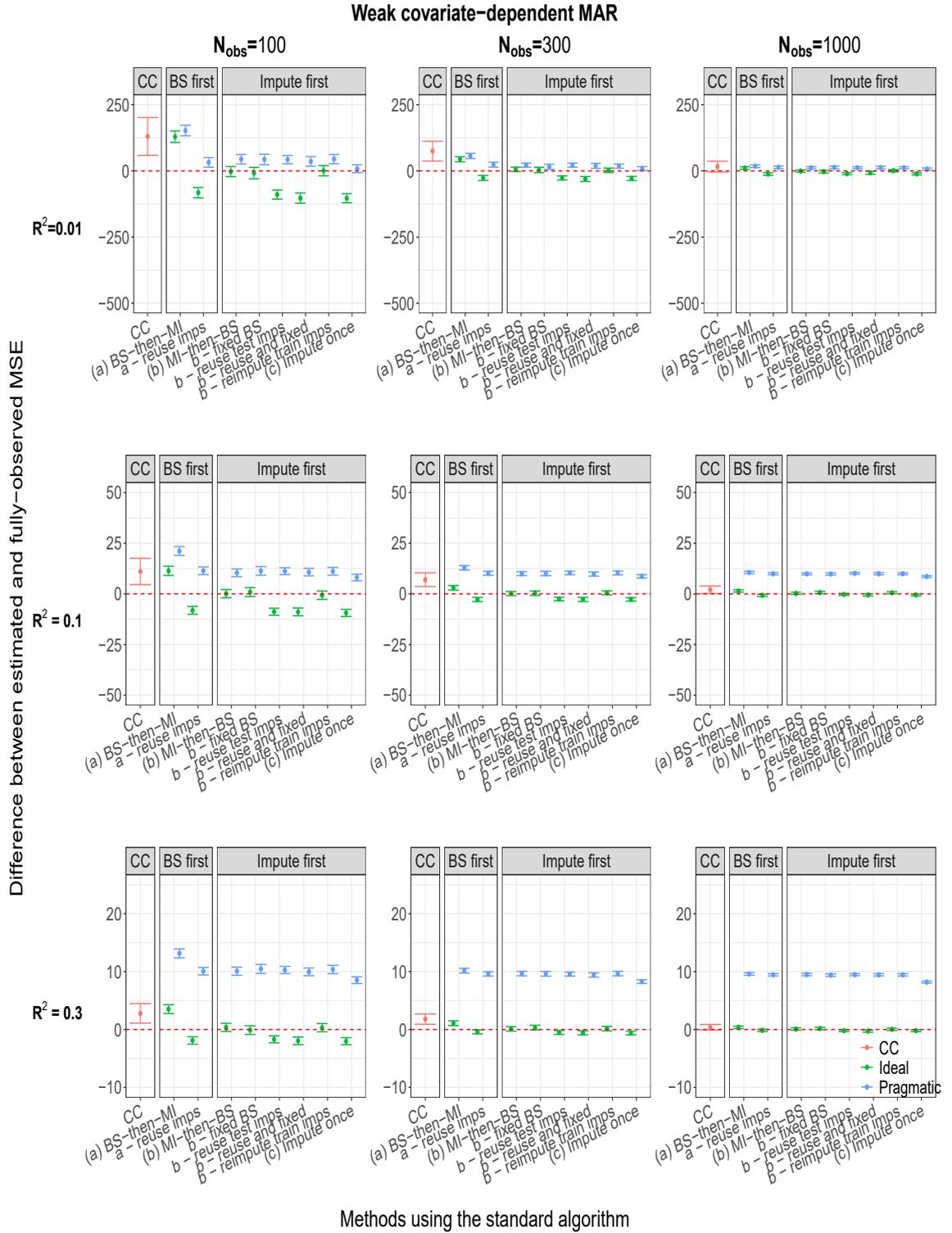


Figure 6.3: The difference $MSE_{imp} - MSE_{obs}$ when data are weakly covariate-dependent MAR for $M = 5$ when 25% of values are missing in X_1 . The error bars summarise results from the 2000 repetitions and the limits represent the Monte Carlo 95% confidence interval of $MSE_{imp} - MSE_{obs}$. CC (complete-case); methods are described in Section 2.7 or Table 6.1.

Outcome-dependent MAR

Figure 6.4 presents results for the complete-case analysis and various proposed methods when data are weakly outcome- and covariate-dependent MAR. The results in the graph are representative of all results when missingness is dependent on the outcome (additional graphs available in Supplementary Plots Section S3.2).

For all scenarios, the complete-case analysis tends to underestimate MSE_{obs} ($MSE_{CC} - MSE_{obs} < 0$), and has increased variability compared to the imputation-based methods. For R^2 values of 0.01 and 0.1, the magnitude of the difference between the complete-case and MSE_{obs} is larger than that for the imputation methods, and this is true for both ideal and pragmatic performance ($|MSE_{CC} - MSE_{obs}| > |MSE_{imp} - MSE_{obs}|$). For $R^2 = 0.3$ the magnitude of the difference for the complete-case analysis is greater than the magnitude of the difference for all methods between the estimated ideal performance MSE and MSE_{obs} ($|MSE_{CC} - MSE_{obs}| > |MSE_{imp,ideal} - MSE_{obs}|$). However, the complete-case analysis magnitude of the difference is less than the magnitude of the difference for pragmatic performance when $R^2 = 0.3$ ($|MSE_{CC} - MSE_{obs}| < |MSE_{imp,prag} - MSE_{obs}|$).

For pragmatic performance when the sample size is 100, *BS-then-MI* overestimates MSE_{obs} ($MSE_{prag,BS-MI} - MSE_{obs} > 0$); the magnitude of this difference is smaller when reusing the imputed datasets used to estimate the apparent performance (method *BS-then-MI reuse imps*) for all values of R^2 . With increasing sample size both methods (*BS-then-MI* and *BS-then-MI reuse imps*) perform similarly. For all sample sizes and values of R^2 all variations of *MI-then-BS* perform similarly when their MSE estimates are compared to MSE_{obs} . The magnitude of the difference ($|MSE_{prag,imp} - MSE_{obs}|$) for the *MI-then-BS* various methods is similar to method *BS-then-MI reuse imps*. For all scenarios, *MI-then-BS impute once* results in an MSE that is closest to MSE_{obs} .

For ideal performance, *BS-then-MI* overestimates MSE_{obs} ($MSE_{ideal,BS-MI} - MSE_{obs} > 0$) for all scenarios. The magnitude of this difference decreases with increasing sample size or increasing R^2 . For a sample size of 100, *BS-then-MI* and *MI-then-BS* methods which involve reusing imputed datasets, as well as *MI-then-BS impute once*, underestimate MSE_{obs} ($MSE_{ideal,imp} - MSE_{obs} < 0$). However, as the sample size increases, this underestimation starts to disappear, with the difference $MSE_{ideal,imp} - MSE_{obs}$ tending to zero. For all scenarios, the ideal performance of methods *MI-then-BS*, *MI-then-BS fixed BS* and *MI-then-BS re-impute* is such that these methods give MSEs with the smallest difference $MSE_{ideal,imp} - MSE_{obs}$ while the other methods have similar magnitudes for the difference ($|MSE_{ideal,imp} - MSE_{obs}|$).

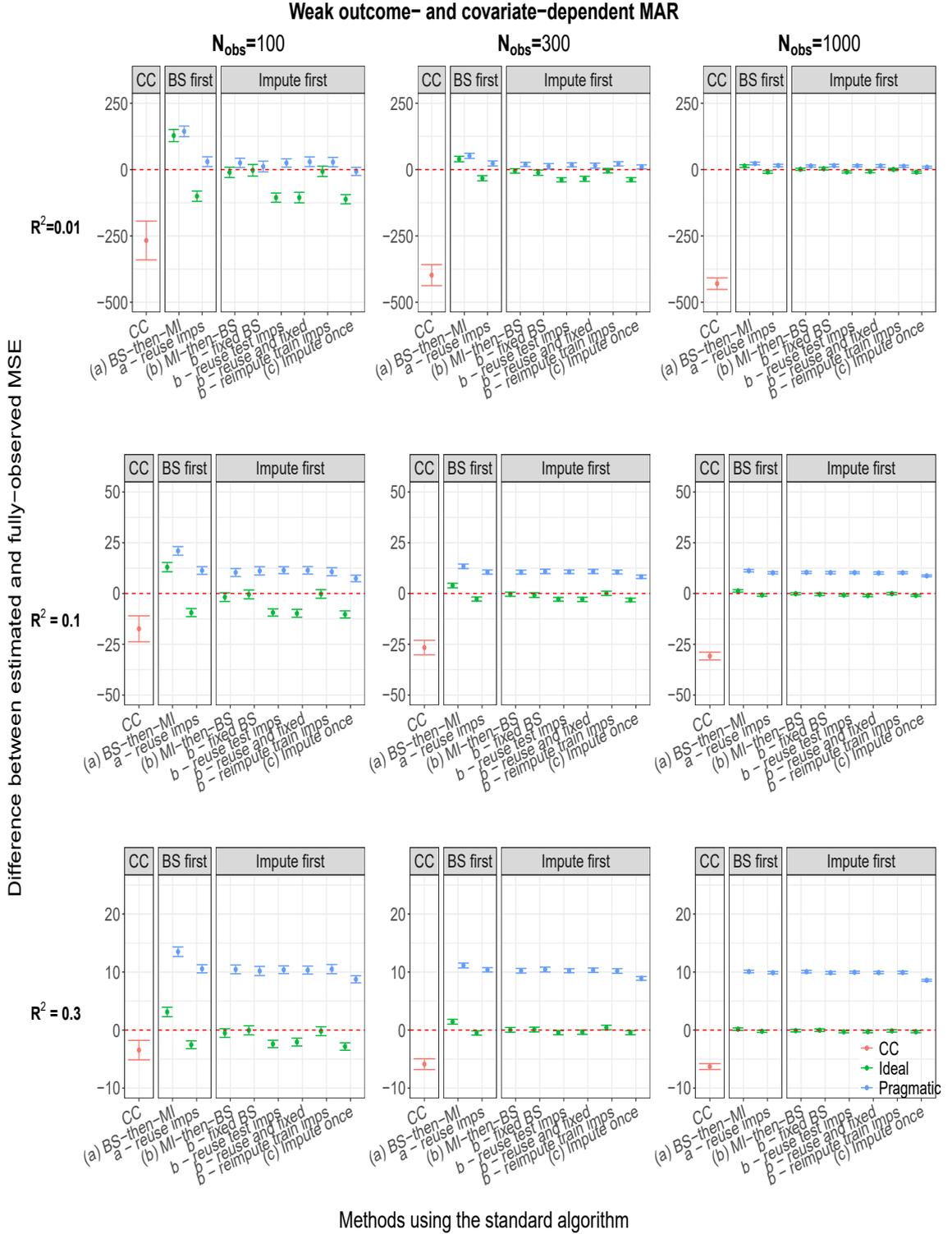


Figure 6.4: The difference $MSE_{imp} - MSE_{obs}$ when data are weakly outcome- and covariate-dependent MAR for $M = 5$ when 25% of values are missing in X_1 . The error bars summarise results from the 2000 repetitions and the limits represent the Monte Carlo 95% confidence interval of $MSE_{imp} - MSE_{obs}$. CC (complete-case); methods are described in Section 2.7 or Table 6.1.

6.4.2 Increasing the number of imputed datasets from 5 to 25

Figure 6.5 displays the results for comparing the various imputation-based methods when using 5 or 25 imputed datasets. The MSE estimate for each method using 5 or 25 imputed datasets is compared to MSE_{obs} . The results in the graph are for the scenario when data are weak covariate-dependent MAR but are representative of results from all simulation scenarios (additional graphs available in Section S3.2.3 in the supplementary plots).

Due to the increased computation time when using 25 imputed datasets the comparison was performed for a reduced number of repetitions (1000 repetitions are used in the Figure) and for a reduced number of methods (methods *MI-then-BS fixed BS* and *MI-then-BS re-impute* are not available for comparison here).

The increased number of imputed datasets appears to make little to no difference for the ideal or pragmatic performance. For all sample sizes and levels of R^2 , the results for each method when compared to MSE_{obs} are similar regardless of whether 5 or 25 imputed datasets are used ($\text{MSE}_{imp,M=5} - \text{MSE}_{obs} \cong \text{MSE}_{imp,M=25} - \text{MSE}_{obs}$).

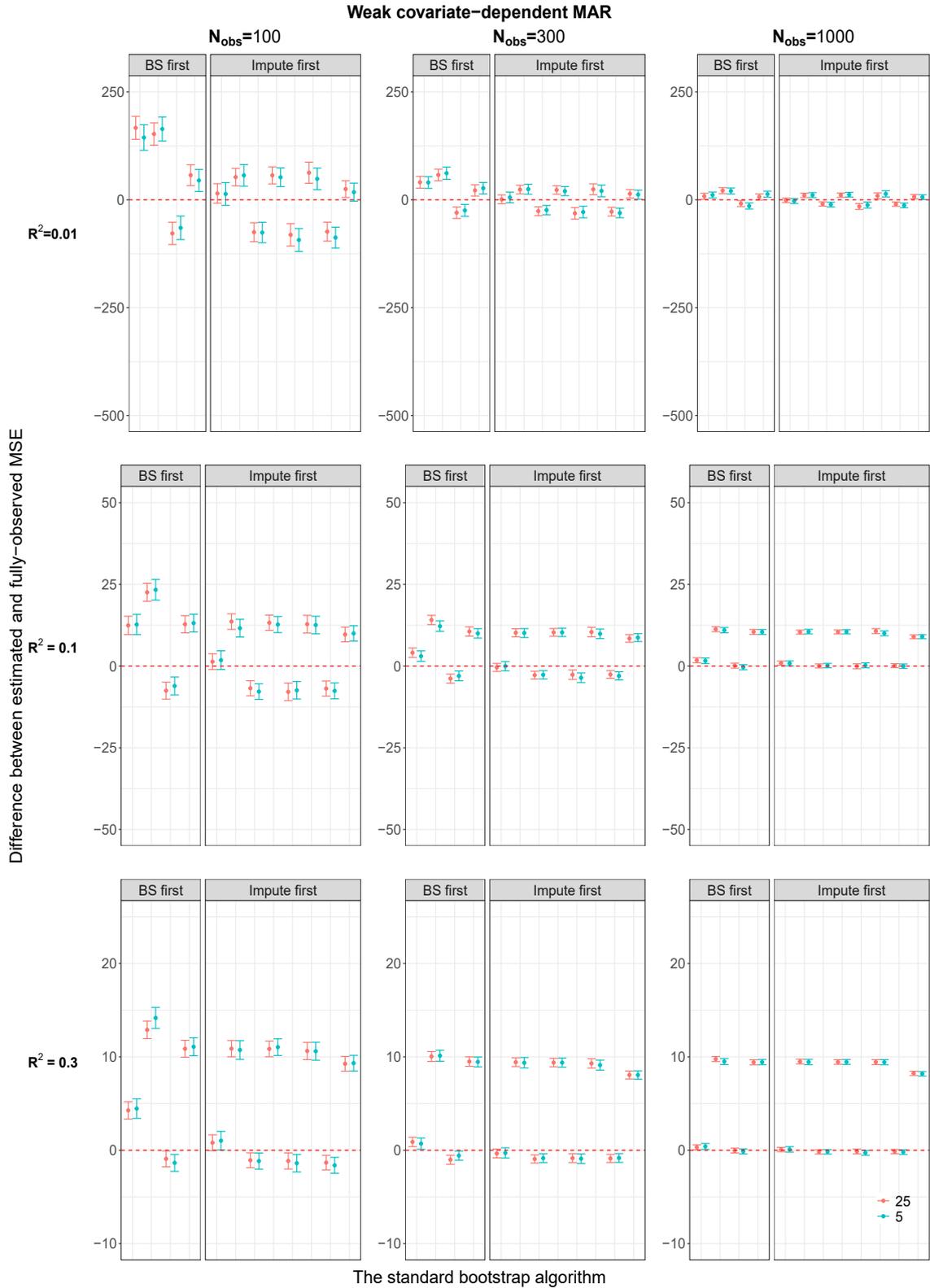


Figure 6.5: The difference $MSE_{imp} - MSE_{obs}$ when data are weakly outcome- and covariate-dependent MAR for $M = 25$ when 25% of values are missing in X_1 . The error bars summarise results from the 2000 repetitions and the limits represent the Monte Carlo 95% confidence interval of $MSE_{imp} - MSE_{obs}$. CC (complete-case); methods are described in Section 2.7 or Table 6.1.

6.4.3 Increasing the percentage of missingness to 40%

Figure 6.6 displays the results for comparing how the various methods handle an increased percentage of missing values in X_1 from 25% to 40%. The MSE estimate from these methods is compared to MSE_{obs} . The graph presents results for the scenario when data are weakly outcome- and covariate- dependent MAR and $R^2 = 0.1$ and the results shown are representative of those for the ideal and pragmatic performance in all other scenarios (additional graphs available in Supplementary plots section S3.2.2).

When data are MCAR or covariate-dependent MAR, the magnitude of the difference for the complete case analysis when compared to MSE_{obs} increases with an increased percentage of missingness ($|MSE_{CC,40\%} - MSE_{obs}| > |MSE_{CC,25\%} - MSE_{obs}|$). With increasing sample size, the magnitude of the MSE difference when 40% of values are missing tends to decrease ($|MSE_{CC,40\%} - MSE_{obs}| \rightarrow |MSE_{CC,25\%} - MSE_{obs}|$). When data are outcome-dependent MAR (with or without dependence of missingness on covariate X_2), the magnitude of the MSE difference for the complete-case analysis is increased when 40% of the values of X_1 are missing, compared to when the percentage of missingness is 25%.

For pragmatic performance, the MSE estimate when 40% of the values of X_1 are missing tends to overestimate MSE_{obs} ($MSE_{imp,40\%} - MSE_{obs} > MSE_{imp,25\%} - MSE_{obs}$) for all scenarios. The variability of the MSE estimates across repetitions when 40% of values are missing is comparable to that when 25% of values are missing.

For ideal performance, the difference between the imputation methods' MSE estimates and MSE_{obs} when 40% of values are missing tends to be slightly greater than when 25% of values are missing ($|MSE_{imp,40\%} - MSE_{obs}| > |MSE_{imp,25\%} - MSE_{obs}|$), with *BS-then-MI* having the largest difference. With increasing sample size, the MSE estimates from the imputation-based methods when 25% of the data are missing are comparable to when 40% of X_1 values are missing.

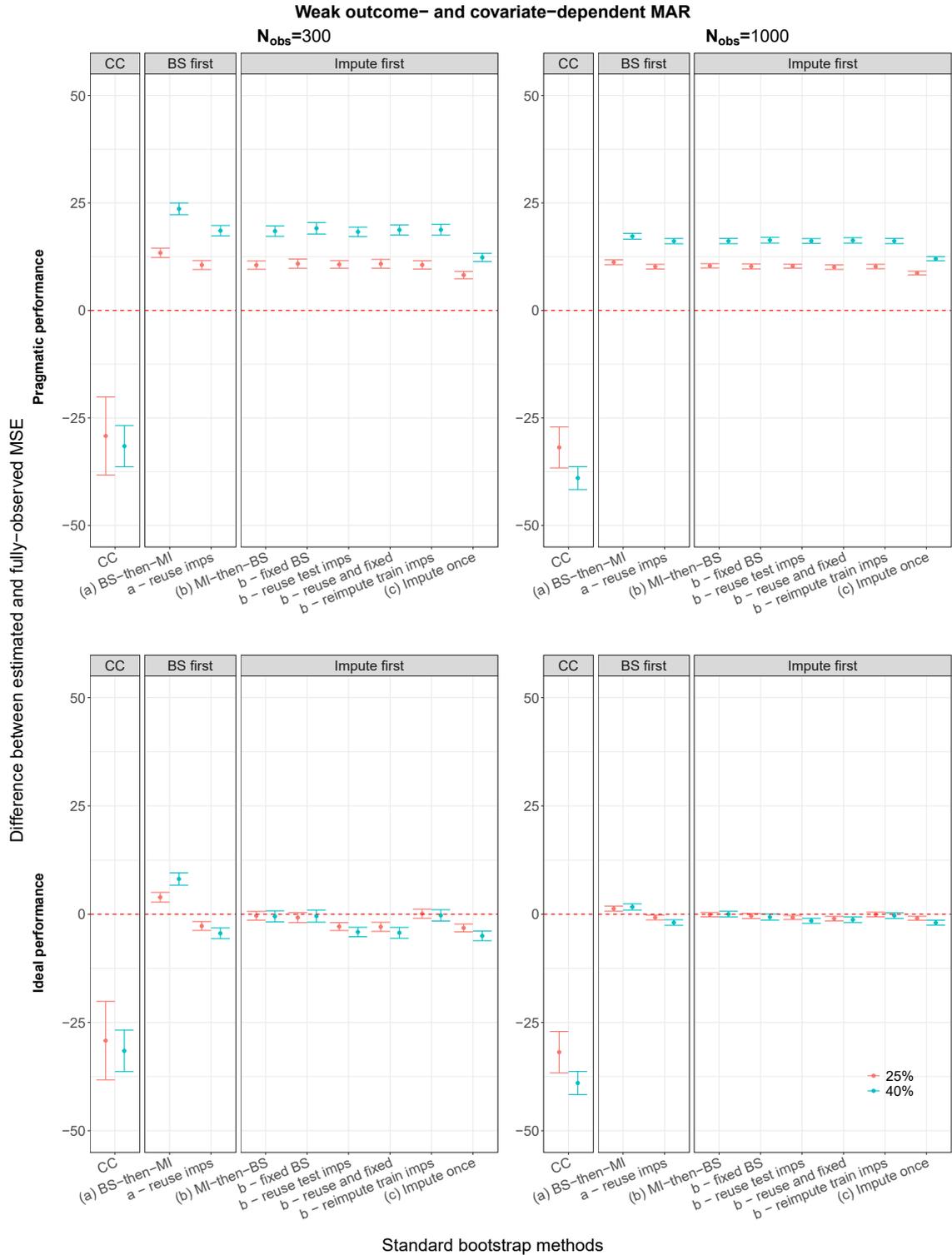


Figure 6.6: Comparing the impact of increasing the percentage of missingness on the difference $MSE_{imp} - MSE_{obs}$ when data are weakly outcome- and covariate-dependent MAR and $R^2 = 0.1$ for the *standard* bootstrap algorithm when $M = 5$. The error bars summarise results from the 2000 repetitions and the limits represent the Monte Carlo 95% confidence interval of $MSE_{imp} - MSE_{obs}$. Red denotes $MSE_{imp} - MSE_{obs}$ when 25% of X_1 values are missing and blue denotes $MSE_{imp} - MSE_{obs}$ when 40% of X_1 values are missing. The top row presents the results for pragmatic performance and the bottom row presents results for ideal performance. CC (complete-case); methods are described in Section 2.7 or Table 6.1.

6.4.4 Comparing results to the target performance

As previously detailed in the comparison of the cross-validation results (Section 4.5.4), the ideal performance of the imputation-based methods and MSE_{obs} were compared to the ideal target MSE estimate ($MSE_{target,obs}$). This is estimated by applying a prediction model, developed using all data, to the fully-observed data in the larger test set to get an MSE estimate (Section 3.6). The pragmatic performance of the imputation methods is compared to applying a prediction model, developed using all data, to the imputed datasets of the larger test set ($MSE_{target,imputed}$). The complete-case estimate of the MSE is obtained from applying a prediction model to the observed cases of the larger test set ($MSE_{target,CC}$). Graphs from all scenarios are available in the supplementary plot section S3.2.4.

MCAR and covariate-dependent MAR

For many of the scenarios assessed when 25% of the values are missing, MSE_{obs} tends to over- or underestimate the MSE performance in the fully-observed larger test set. The results from the various methods involving MI (MSE_{imp}) tend to over- or underestimate the larger test set MSE estimate when MSE_{obs} does. For example, if ($MSE_{obs} - MSE_{target,obs} < 0$) then we typically also see that $MSE_{imp} - MSE_{target,obs}$ or $MSE_{imp} - MSE_{target,imputed}$ is less than zero). This can be seen in Figure 6.7 which presents results for comparing MSE estimates to the target MSE ($MSE_{imp} - MSE_{target}$) when data are weak covariate-dependent MAR. For a sample size of 1000 and $R^2 = 0.1$, MSE_{obs} underestimates the target estimate, as do all other methods.

When sample size is 100 and for all R^2 values, both the estimated ideal and pragmatic performance of the methods tends to either underestimate or perform similarly to the ideal and pragmatic target MSE of the larger test set. When the methods underestimate the target MSE, the default *BS-then-MI* tends to have a smaller difference when compared to the target MSE estimate than the other methods for both ideal and pragmatic performance. When increasing the sample size to 300 or 1000 the estimated ideal and pragmatic performance of the imputation-based methods all perform similarly when compared to the ideal or pragmatic target MSE.

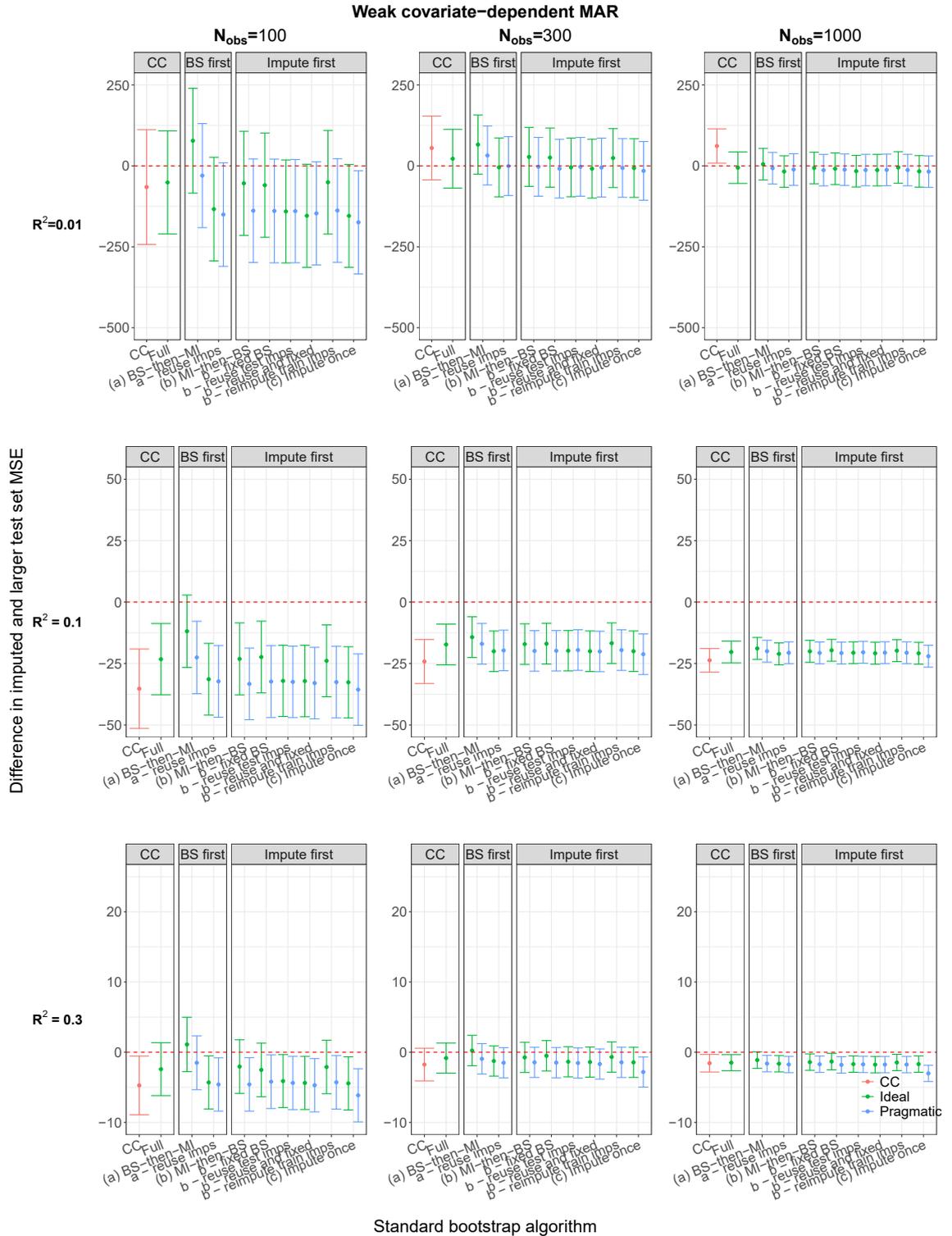


Figure 6.7: The difference $MSE_{imp} - MSE_{target}$ when data are weakly covariate-dependent MAR for $M = 5$ when 25% of values are missing in X_1 . The error bars summarise results from the 2000 repetitions and the limits represent the Monte Carlo 95% confidence interval of $MSE_{imp} - MSE_{target}$. CC (complete-case); methods are described in Section 2.7 or Table 6.1.

Outcome-dependent MAR

Figure 6.7 presents results for comparing MSE estimates to the target MSE ($MSE_{imp} - MSE_{target}$) when data are weakly outcome- and covariate-dependent MAR. For all scenarios, all methods have overlapping confidence intervals. For the majority of scenarios, all methods tend to approximate the ideal and pragmatic target MSE well. When the majority of the methods tend to under-estimate the ideal or pragmatic target MSE, method *BS-then-MI* tends to either have the smallest magnitude of the difference for underestimation or tends to over-estimate the target MSE estimate. When increasing the sample size to 300 or 1000 the methods all perform similarly when compared to their respective target MSE for ideal and pragmatic performance.

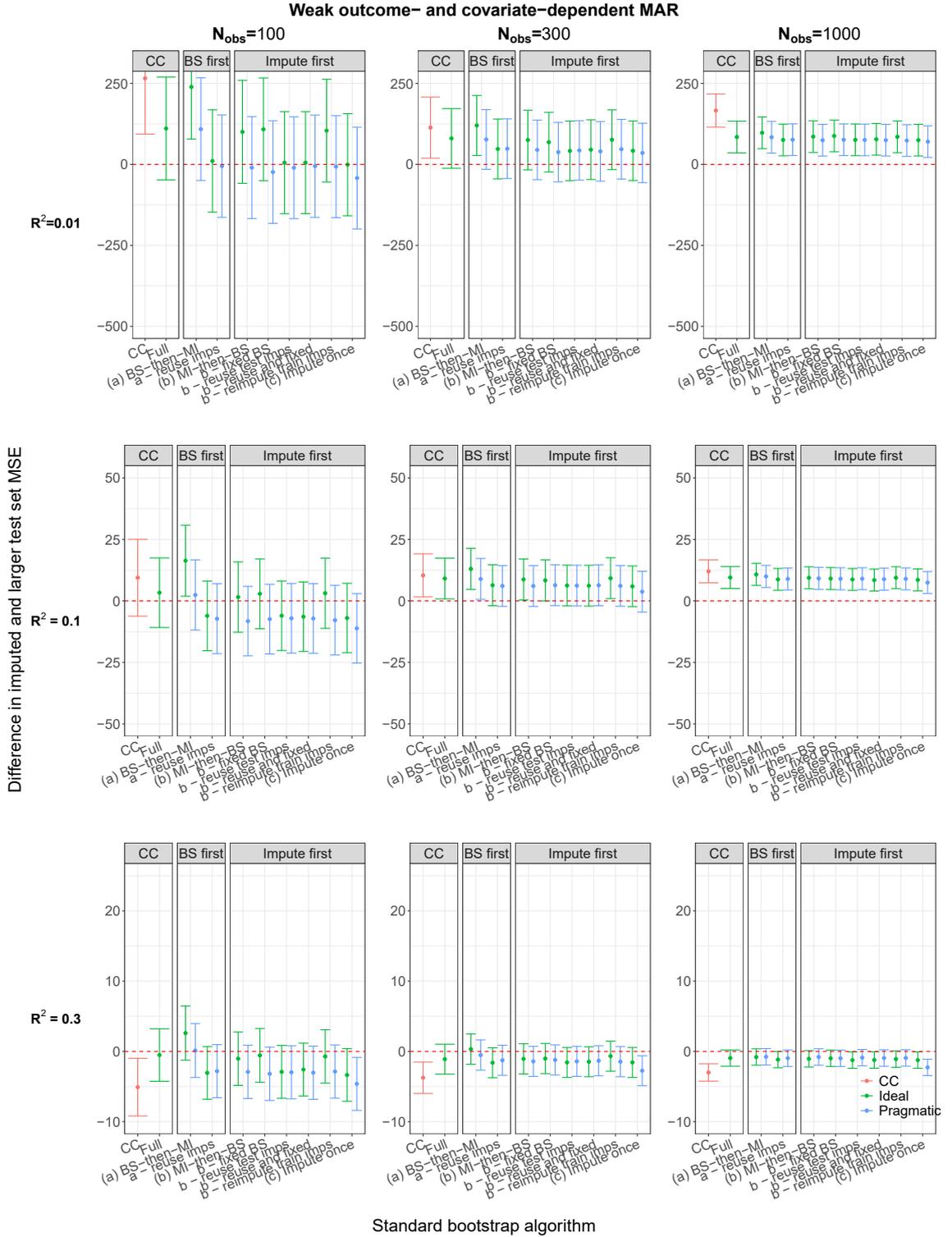


Figure 6.8: The difference $MSE_{imp} - MSE_{target}$ when data are weakly outcome- and covariate-dependent MAR for $M = 5$ when 25% of values are missing in X_1 . The error bars summarise results from the 2000 repetitions and the limits represent the Monte Carlo 95% confidence interval of $MSE_{imp} - MSE_{target}$. CC (complete-case); methods are described in Section 2.7 or Table 6.1.

6.5 Detailed results for the 0.632 bootstrap algorithm

In this section I will summarise the results from the simulation study when the outcome is continuous for the 0.632 bootstrap algorithm. I will explore how the methods perform when their estimated MSEs are compared to MSE_{obs} and also when compared to a target MSE value from a larger validation set with 25% of values in X_1 missing and 5 imputed datasets. In addition, I will analyse whether increasing the number of imputed datasets improves results and also examine how the methods perform with an increased percentage of missingness. The methods presented were summarised in full in Section 2.7 and were briefly resummarised at the start of this chapter in Table 6.1.

The results presented in this section are nearly identical to the results presented for the *standard* bootstrap internal validation algorithm in Section 6.4. I have included the 0.632 results here so that the reader may reassure themselves that the results are similar to the *standard* algorithm, if so desired. Otherwise, the reader may skip to Section 6.6 for a discussion of data leakage in the imputation process.

6.5.1 Comparing results to the MSE estimate when data are fully-observed

MCAR and covariate-dependent MAR

Figure 6.9 compares the MSE estimates from various missing data methods when data are weakly covariate-dependent MAR. The methods' MSE estimates are compared to MSE_{obs} . The results in Figure 6.9 are similar for MCAR and strong covariate-dependent MAR scenarios. When sample size is 100 or 300 the complete-case analysis tends to overestimate MSE_{obs} ($\text{MSE}_{CC} - \text{MSE}_{obs} > 0$) and tends to be more variable than the imputation methods. With increased sample size to 1000, the variability and magnitude of the difference decreases.

For the pragmatic performance of the imputation methods when sample size is 100 and for all R^2 values, method *BS-then-MI* tends to overestimate MSE_{obs} . For all values of R^2 , methods *BS-then-MI reuseimps* overestimates MSE_{obs} but the magnitude of this difference is less than *BS-then-MI*. Pragmatic performance for all *MI-then-BS* methods overestimates MSE_{obs} while method *MI-then-BS impute once* has the smallest difference overall across all methods. For all values of R^2 when the sample size increases to 300, the magnitude of overestimation for all imputation methods decreases and the methods perform similarly.

For ideal imputation and sample size of 100, all methods which involve reusing the imputed datasets used to estimate apparent performance, and method *MI-then-BS impute once*, tend to underestimate MSE_{obs} . That is for all values of R^2 , $\text{MSE}_{ideal,imp} - \text{MSE}_{obs} < 0$ for $imp = \text{BS-then-MI reuseimps}, \text{MI-then-BS reuse testimps}, \text{MI-then-BS reuse and fixed}$ and *MI-then-BS impute once*. For all values of R^2 , the other imputation methods tend

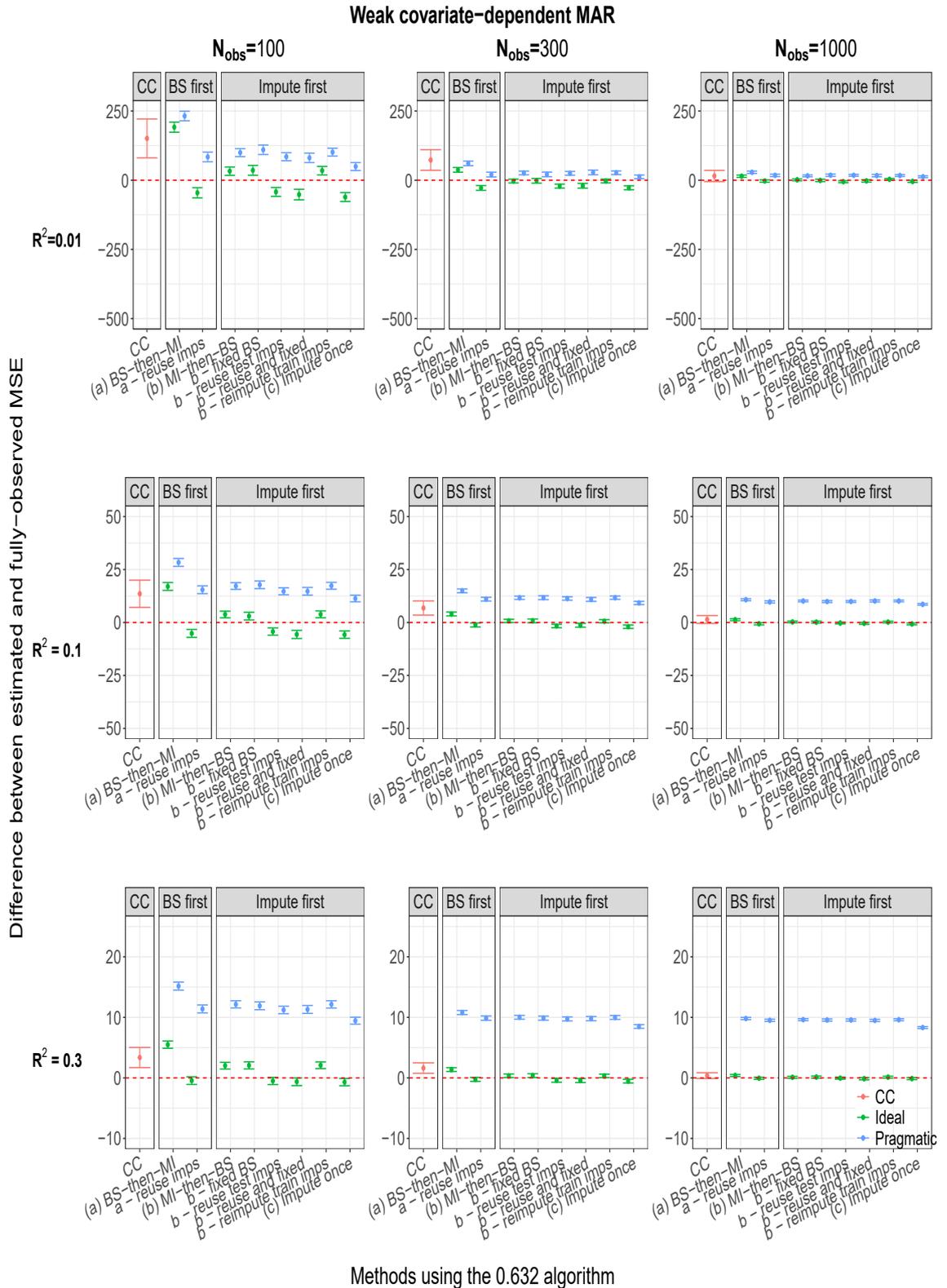


Figure 6.9: The difference $MSE_{imp} - MSE_{obs}$ when data are weakly covariate-dependent MAR for $M = 5$ when 25% of values are missing in X_1 . The error bars summarise results from the 2000 repetitions and the limits represent the Monte Carlo 95% confidence interval of $MSE_{imp} - MSE_{obs}$. CC (complete-case); methods are described in Section 2.7 or Table 6.1.

to overestimate MSE_{obs} ($MSE_{ideal,imp} - MSE_{obs} > 0$ for $imp = BS\text{-then-}MI, MI\text{-then-}BS, MI\text{-then-}BS\text{ fixed }BS$ and $MI\text{-then-}BS\text{ re-impute}$). With increasing sample size all methods tend to perform similarly and the difference between their MSE estimates and MSE_{obs} tends to zero. With increasing sample size the Monte Carlo standard error of the MSE estimates also decreases.

Outcome-dependent MAR

Figure 6.10 displays summary MSE estimates for the complete-case analysis and imputation methods which are compared to MSE_{obs} when data are weak outcome- and covariate-dependent MAR. The graph is representative of all results when missingness is dependent on the outcome (weak outcome-dependent MAR and weak outcome- and strong covariate-dependent MAR).

For all scenarios, the complete-case analysis underestimates MSE_{obs} ($MSE_{CC} - MSE_{obs} < 0$). Across all scenarios, the magnitude of this difference is larger than any of the imputation-based methods ideal performance ($|MSE_{CC} - MSE_{obs}| > |MSE_{imp,ideal} - MSE_{obs}|$). The magnitude of the difference between the complete-case analysis MSE and MSE_{obs} is larger than the imputation methods pragmatic performance when $R^2 = 0.01$ or for sample sizes of 300 or 1000 when $R^2 = 0.1$ ($|MSE_{CC} - MSE_{obs}| > |MSE_{imp,prag} - MSE_{obs}|$). For all other scenarios, the pragmatic performance of the imputation methods has a larger difference than the complete-case analysis when compared to MSE_{obs} .

For a small sample size of 100, the pragmatic performance of all imputation methods overestimates MSE_{obs} . *BS-then-MI* overestimates MSE_{obs} the most for all values of R^2 . Method *BS-then-MI reuse imps*, which reuses the imputed datasets used to estimate the apparent performance, has a smaller difference ($MSE_{prag,BS-MI-reuse} - MSE_{obs}$) than *BS-then-MI* and performs similarly to the various *MI-then-BE* methods. For all values of R^2 , method *MI-then-BE impute once* has the smallest difference overall. With increasing sample size to 300 and 1000 and for all values of R^2 , all imputation methods perform similarly when compared to MSE_{obs} .

For ideal performance, all imputation methods which involve reusing imputations tend to underestimate MSE_{obs} (methods *BS-then-MI reuse imps*, *MI-then-BE reuse test imps*, *MI-then-BE reuse and fixed*, *MI-then-BE impute once*). Methods *BS-then-MI*, *MI-then-BE*, *MI-then-BE fixed BE* and *MI-then-BE re-impute* overestimate MSE_{obs} . For sample size of 100, the magnitude of the difference between imputation methods' MSE estimate and MSE_{obs} is largest for *BS-then-MI* while all other methods have similar magnitudes for all values of R^2 . With increasing sample size to 300 and 1000, all methods have similar performance when compared to MSE_{obs} .

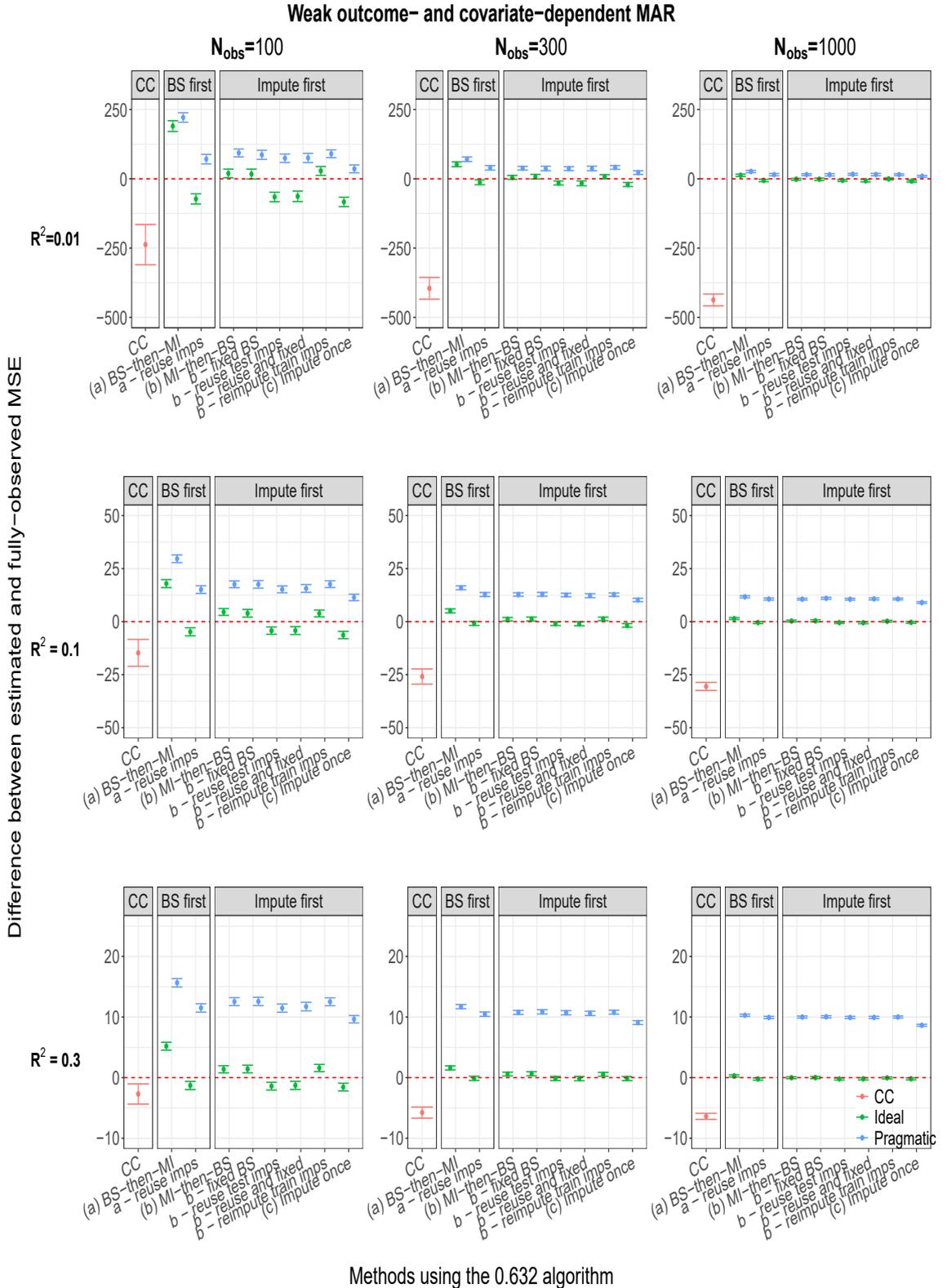


Figure 6.10: The difference $MSE_{imp} - MSE_{obs}$ when data are weakly outcome- and covariate-dependent MAR for $M = 5$ when 25% of values are missing in X_1 . The error bars summarise results from the 2000 repetitions and the limits represent the Monte Carlo 95% confidence interval of $MSE_{imp} - MSE_{obs}$. CC (complete-case); methods are described in Section 2.7 or Table 6.1.

6.5.2 Increasing the number of imputed datasets from 5 to 25

Figure 6.11 displays the results for comparing the various imputation-based methods when using 5 or 25 imputed datasets. The MSE estimate for each method using 5 or 25 imputed datasets is compared to MSE_{obs} . The results in the graph are for the scenario when data are weak covariate-dependent MAR but are representative of results from all simulation scenarios (additional graphs available in Section S3.3.3 in the supplementary plots).

Due to the increased computation time when using 25 imputed datasets the comparison was performed for a reduced number of repetitions (1000 repetitions are used for the simulation study presented in the Figure) and for a reduced number of methods (methods *MI-then-BS fixed BS* and *MI-then-BS re-impute* are not available for comparison).

The increased number of imputed datasets appears to make little to no difference for the ideal or pragmatic performance. For all sample sizes and levels of R^2 , the results for each method when compared to MSE_{obs} are similar regardless of whether 5 or 25 imputed datasets are used ($\text{MSE}_{imp,M=5} - \text{MSE}_{obs} \cong \text{MSE}_{imp,M=25} - \text{MSE}_{obs}$).

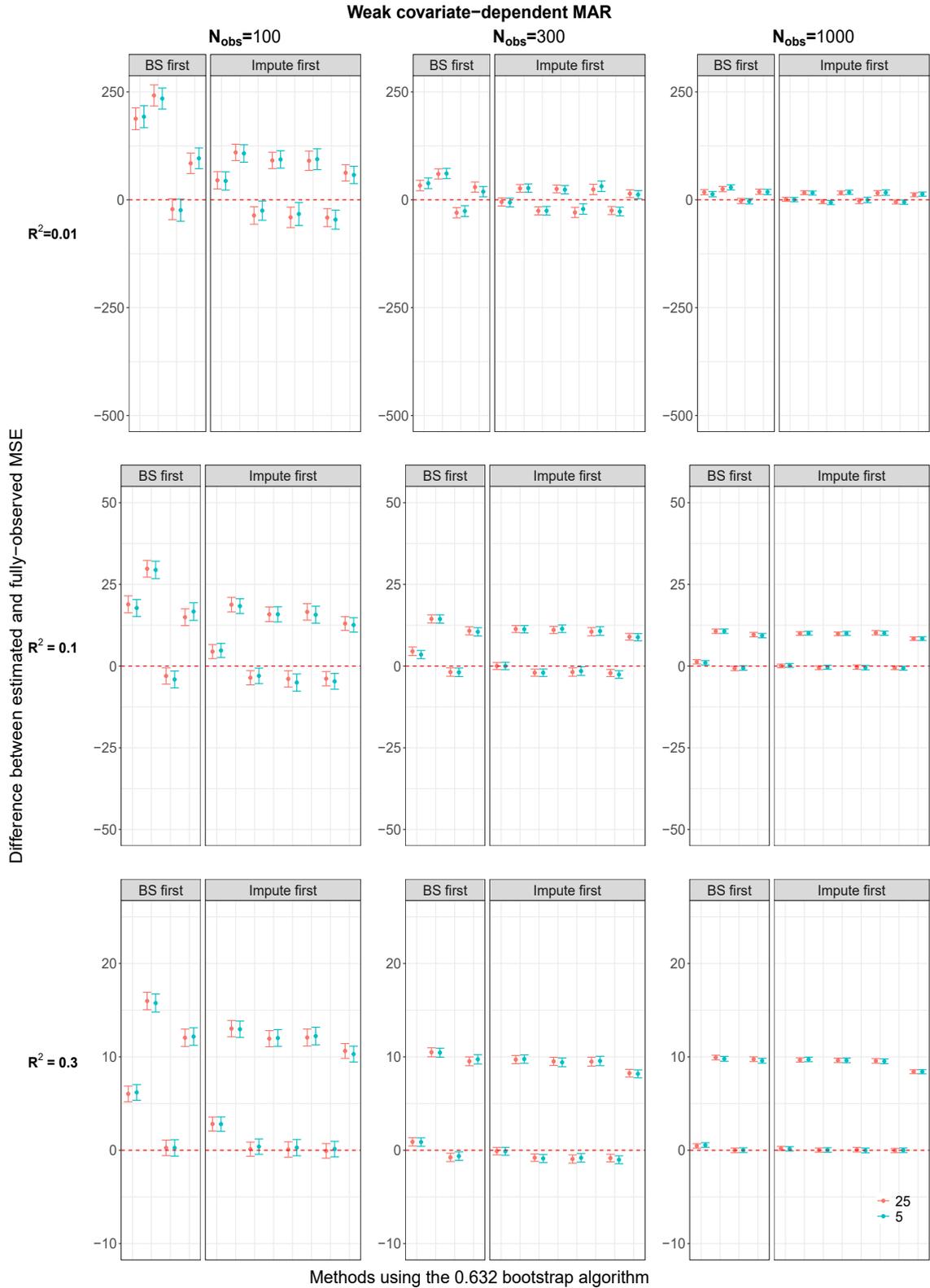


Figure 6.11: The difference $MSE_{imp} - MSE_{obs}$ when data are weakly covariate-dependent MAR for $M = 25$ when 25% of values are missing in X_1 . The error bars summarise results from the 2000 repetitions and the limits represent the Monte Carlo 95% confidence interval of $MSE_{imp} - MSE_{obs}$. CC (complete-case); methods are described in Section 2.7 or Table 6.1.

6.5.3 Increasing the percentage of missingness to 40%

Figure 6.12 compares the impact of the percentage of missingness on the various missing data methods and how well they perform when compared to the fully-observed MSE estimate (MSE_{obs}). The graph displays results for ideal and pragmatic performance when data are weakly outcome- and covariate-dependent MAR and $R^2 = 0.1$ but the results shown are representative of those for the ideal and pragmatic performance for all scenarios (additional graphs in Supplementary plots section S3.3.2).

When data are MCAR or covariate-dependent MAR with 40% of data missing, the complete-case analysis tends to be similar or have a larger difference compared to when 25% of the data are missing ($|MSE_{CC,40\%} - MSE_{obs}| \geq |MSE_{CC,25\%} - MSE_{obs}|$). The magnitude of this difference decreases with increasing sample size tending to zero. When data are outcome-dependent MAR, as in Figure 6.12, with increasing sample size there is a tendency for the complete-case analysis estimate to have increased under-estimation of MSE_{obs} with increasing percentage of missing values.

In Figure 6.12 the estimated pragmatic performance of the MSE estimate when 40% of the X_1 values are missing overestimates MSE_{obs} more than when data are 25% missing for all imputation methods in all scenarios ($|MSE_{imp,40\%} - MSE_{obs}| > |MSE_{imp,25\%} - MSE_{obs}|$). With increasing sample size the magnitude of the difference for 40% of missing values decreases but is still greater than the magnitude of the difference when 25% of the values are missing.

For ideal performance when sample size is 300, the difference between the imputation methods' MSE estimates, when 40% of X_1 values are missing, compared to MSE_{obs} is similar or slightly larger than the imputation methods MSE estimates when 25% of values are missing. When the sample size is increased to 1000, the MSE estimates of the imputation methods are similar for both 25% and 40% of X_1 values are missing.

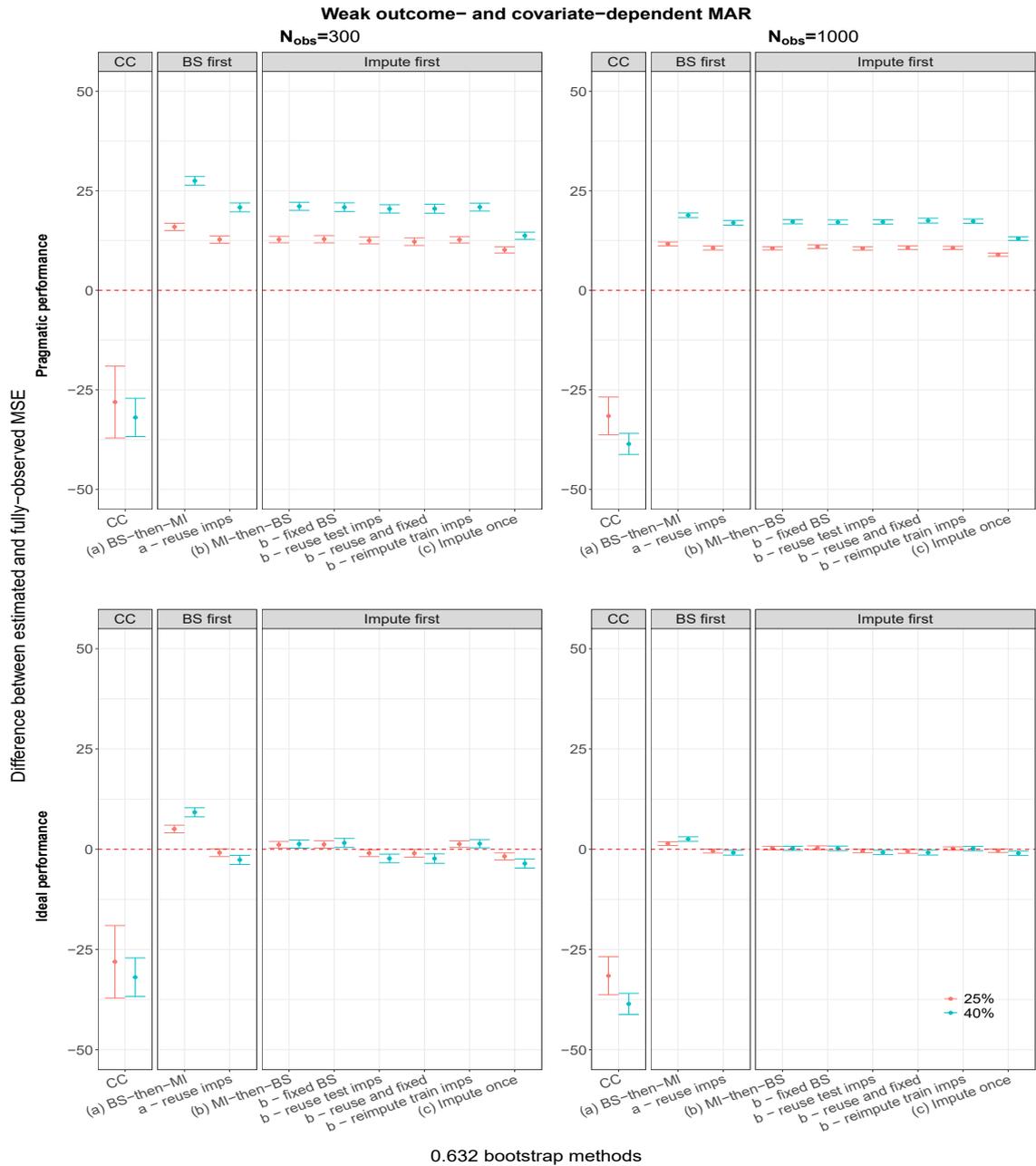


Figure 6.12: Comparing the impact of increasing the percentage of missingness on the difference $MSE_{imp} - MSE_{obs}$ when data are weakly outcome- and covariate-dependent MAR and $R^2 = 0.1$ for the 0.632 bootstrap algorithm when $M = 5$. The error bars summarise results from the 2000 repetitions and the limits represent the Monte Carlo 95% confidence interval of $MSE_{imp} - MSE_{obs}$. Red denotes $MSE_{imp} - MSE_{obs}$ when 25% of X_1 values are missing and blue denotes $MSE_{imp} - MSE_{obs}$ when 40% of X_1 values are missing. The top row presents the results for pragmatic performance and the bottom row presents results for ideal performance. CC (complete-case); methods are described in Section 2.7 or Table 6.1.

6.5.4 Comparing results to the target performance

Briefly as a reminder, for the target MSE, the ideal performance of the bootstrap imputation methods and MSE_{obs} were compared to the ideal target MSE estimate ($MSE_{target,obs}$). This is estimated by applying a prediction model, developed using all data, to the fully-observed data in the larger test set to get an MSE estimate (Section 3.6). The pragmatic performance of the imputation methods is compared to applying a prediction model, developed using all data, to the imputed datasets of the larger test set ($MSE_{target,imputed}$). The complete-case estimate of the MSE is obtained from applying a prediction model to the observed cases of the larger test set ($MSE_{target,CC}$). Figure 6.13 displays results for comparing the various methods MSE estimate with their respective ideal, pragmatic or CC target MSE when data are weak covariate-dependent MAR. Graphs from all scenarios are available in the supplementary plot section S3.3.4.

MCAR and covariate-dependent MAR

When $R^2 = 0.01$, the confidence intervals for the difference between the complete-case analysis MSE estimate and $MSE_{target,CC}$ overlaps with zero. For all other R^2 and sample size values, the complete-case analysis tends to under- or overestimate the target MSE ($|MSE_{CC} - MSE_{target,CC}| > 0$).

When data are MCAR, weak or strong covariate-dependence MAR and for sample sizes of 300 or 1000, all imputation methods tend to perform similarly to each other when their ideal or pragmatic performance is compared, respectively, to the ideal or pragmatic target performance. The confidence intervals for the difference between the estimated and target MSE for all methods tends to overlap.

For a sample size of 100, the majority of imputation methods tends to underestimate the target MSE for ideal or pragmatic performance. In instances where all imputation methods underestimate or approximate the target MSE, *BS-then-MI* tends to underestimate the performance the least, as seen in Figure 6.13 for $R^2 = 0.1$ and 0.3 and performs well across the majority of scenarios when sample size is 100. When all imputation methods underestimate the target performance, the ideal performance of *BS-then-MI reuseimps*, *MI-then-BS reuse testimps* and *MI-then-BS impute once* underestimate the ideal target performance the most.

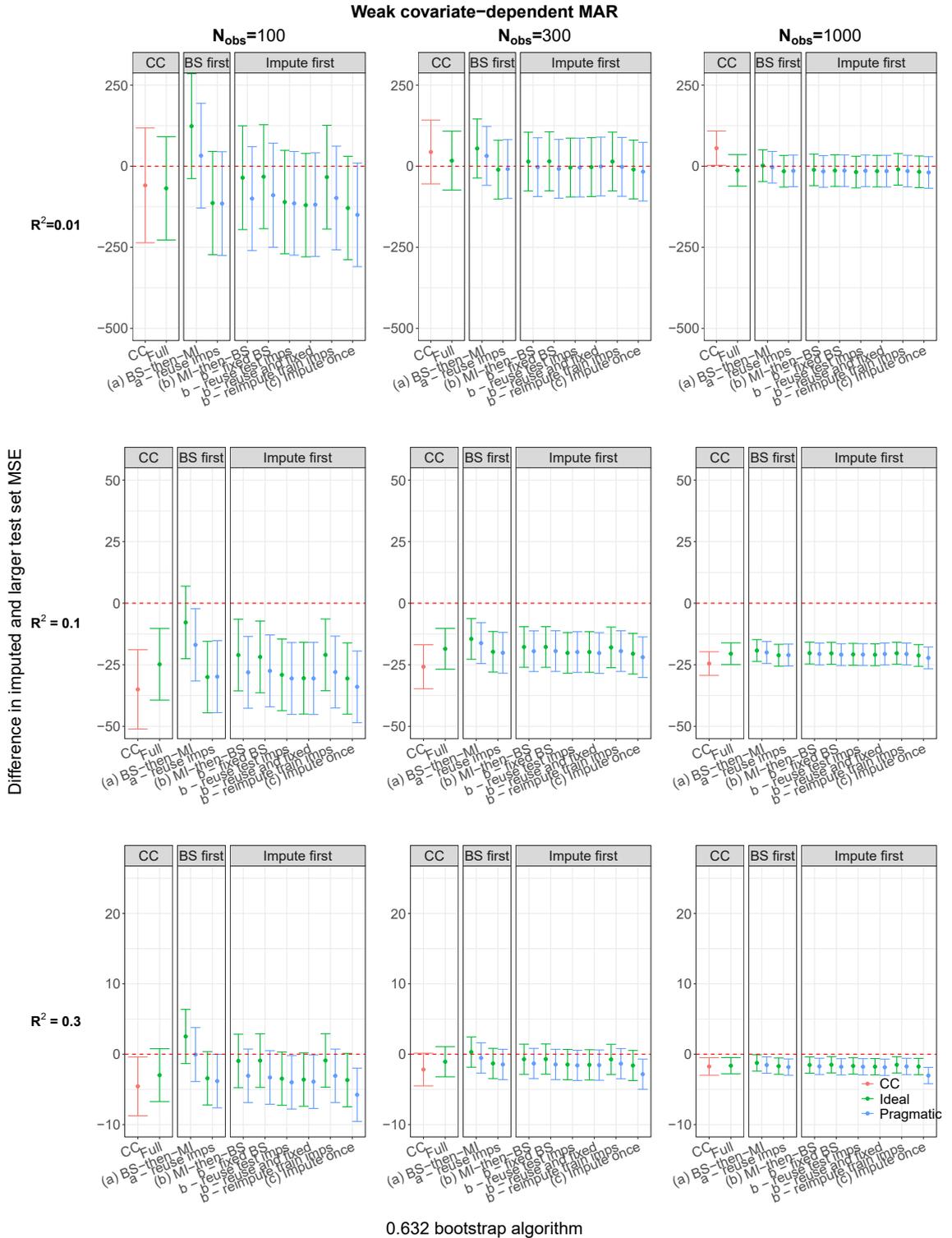


Figure 6.13: The difference $MSE_{imp} - MSE_{target}$ when data are weakly covariate-dependent MAR for $M = 5$ when 25% of values are missing in X_1 . The error bars summarise results from the 2000 repetitions and the limits represent the Monte Carlo 95% confidence interval of $MSE_{imp} - MSE_{target}$. CC (complete-case); methods are described in Section 2.7 or Table 6.1.

Outcome-dependent MAR

When the sample size is 100 and data are either weakly outcome-dependent MAR when $R^2 = 0.1$ or weakly outcome-dependent and strongly covariate-dependent MAR when $R^2 = 0.1, 0.3$, *BS-then-MI* overestimates the target MSE for ideal and pragmatic performance with a larger magnitude of the difference than the other methods. For all other scenarios when the sample size is 100 the *BS-then-MI* method either approximates the target performance well for ideal and pragmatic performance or it approximates the target MSE best when compared to the other imputation methods (i.e. it has the smallest magnitude of the difference). Graphs are available in Section [S3.3.4](#) of the Supplementary plots section. With increasing sample size all *BS-then-MI* and *MI-then-BS* methods begin to perform similarly to each other when compared to their respective ideal or pragmatic target performance.

6.6 Is data leakage an issue within the imputation process for the standard and 0.632 bootstrap algorithms?

The introduction of data leakage by imputing data was discussed in Section 2.8 and this was previously assessed for cross-validation when the outcome is continuous in Section 4.6. The bootstrap imputation methods range from method *MI-then-BS impute once*, which has the highest amount of leakage, to *BS-then-MI*, which has no data leakage for the 0.632 bootstrap version. All methods for the *standard* bootstrap algorithm have some leakage, however, this leakage is a natural (and intended) part of the algorithm and cannot be avoided. Data leakage discussed here will be in terms of any leakage introduced through the imputation process.

The default version of *BS-then-MI* has no data leakage compared to the other methods. The bootstrap sample is imputed using a training imputation model (including the outcome, introduced in Section 2.5), and the imputed bootstrap sample is then used to train M bootstrap prediction models. The same bootstrap sample is then imputed again using a test imputation model (which may include the outcome depending on whether the ideal or pragmatic performance is of interest) to estimate the BS performance of the M bootstrap prediction models (for the *standard* bootstrap algorithm). A variation of this method is *BS-then-MI reuseimps* which is subject to data leakage, unlike the default method *BS-then-MI*. Method *BS-then-MI reuseimps* reuses the imputed datasets which were originally used to train and evaluate prediction models in order to estimate the apparent performance i.e. these imputed datasets contain all observations from the original dataset and all of these observations were used to impute any missing values. As seen in Figures 6.3 and 6.4 (for the *standard* bootstrap) and Figures 6.9 and 6.10 (for the 0.632 bootstrap), this has caused the magnitude of both the pragmatic and ideal performances' MSE estimate for *BS-then-MI reuseimps* to be smaller than that for *BS-then-MI* ($|\text{MSE}_{BS-MI \text{ reuse}} - \text{MSE}_{obs}| < |\text{MSE}_{BS-MI} - \text{MSE}_{obs}|$ for ideal or pragmatic performance). This difference in magnitude is most prominently seen for a sample size of 100, where in some scenarios the magnitude of *BS-then-MI reuseimps* is half that of *BS-then-MI*. For ideal performance, this leakage causes *BS-then-MI reuseimps* to become over-optimistic. With increasing sample size the two methods begin to perform similarly.

All variations of the *MI-then-BS* approach inherently suffer from data leakage as the bootstrap prediction models are trained on a bootstrap sample taken from data imputed using all available observations in the original dataset. The pragmatic performance of all *MI-then-BS* variations is similar to that of the *BS-then-MI reuseimps* method, which, as discussed above, is also subject to data leakage. Method *MI-then-BS* is similar to *BS-then-MI reuseimps* in that both methods reuse imputed training datasets to fit the bootstrap prediction model. However, *BS-then-MI reuseimps* also reuses imputed test datasets

to evaluate the bootstrap prediction model, whereas *MI-then-BS* imputes the bootstrap sample separately using a test imputation model to evaluate bootstrap performance. For pragmatic performance, both perform similarly. However, for ideal performance (where the outcome is included in the imputation model), *BS-then-MI reuseimps* tends to underestimate the fully-observed estimate of the MSE while *MI-then-BS* tends to give an MSE that is similar to MSE_{obs} . This underestimation can also be seen when comparing *MI-then-BS* to *MI-then-BS reuse testimps* (reuses the imputed test datasets to estimate the bootstrap performance in the same way that *BS-then-MI reuseimps* does). This can be seen in Figures 6.3, 6.4, 6.9 and 6.10 and in the additional plots available in the supplementary plot section (Sections S3.2.1, S3.3.1) for a sample size of 100. Therefore, using all available covariates and outcome data to impute missing values in the bootstrap sample (*MI-then-BS reuseimps*, *BS-then-MI reuseimps*) results in a suggestion that the prediction model is doing better than it would have done if all the data were observed. These two methods are therefore over-optimistic, and the over-optimism arises due to this further increased leakage from knowledge of the outcome, when compared to *MI-then-BS* (where the bootstrap sample to evaluate the bootstrap prediction model is imputed based only on the observations which were sampled). Similarly, *MI-then-BS impute once* uses one set of imputed datasets to train and evaluate the prediction models and is subject to data leakage, performing similarly to *BS-then-MI reuseimps* and *MI-then-BS reuse testimps*.

Comparing *BS-then-MI* versus *BS-then-MI reuseimps* shows that reusing the imputed datasets (to train and evaluate bootstrap prediction models) in order to estimate the bootstrap performance leads to more optimistic performance. This is regardless of whether pragmatic or ideal performance is of interest. Comparing *MI-then-BS* with *MI-then-BS reuse testimps* shows that reusing imputed test datasets (imputed using all available observations) versus imputing the bootstrap sample with the test imputation model can lead to more optimistic results for ideal performance due to data leakage. With increasing sample size, while the direction of either under- or overestimating the fully-observed MSE remains the same for each method, the magnitude of the difference decreases and the methods can be seen to perform similarly for a sample size of 1000.

Figures 6.7 and 6.13 display results when comparing the ideal and pragmatic performance of the imputation methods to the ideal or pragmatic target MSE, additional graphs are available in supplementary plots Section S3.2.4 and S3.3.4. As with the comparison with the fully-observed MSE estimate, for larger sample sizes the methods tend to perform similarly and data leakage appears to have little effect. For a sample size of 100, the methods which tend to underestimate the ideal target performance the most are *BS-then-MI reuseimps*, *MI-then-BS reuse testimps* and *MI-then-BS impute once* i.e. methods which are subject to data leakage. Across the majority of scenarios, *BS-then-MI* (which

has no data leakage) tends to perform well. The magnitude $|\text{MSE}_{BS-MI} - \text{MSE}_{target}|$ tends to either be smaller than the other imputation methods or it performs similarly to the other methods, all of which approximate the target MSE well.

6.7 Comparing internal validation algorithms

Up to this point, the methods involving cross-validation, the *standard* bootstrap and the 0.632 bootstrap were primarily comparing the methods when data were missing to the methods when no missing data were present. This in turn makes it difficult to compare cross-validation performance ($\text{MSE}_{CV} - \text{MSE}_{CV,obs}$) to bootstrap performance ($\text{MSE}_{BS} - \text{MSE}_{BS,obs}$) as the reference values (either $\text{MSE}_{CV,obs}$, $\text{MSE}_{BS_{0.632,obs}}$ or $\text{MSE}_{BS_{Std,obs}}$) are different (Table 6.2). For example, when $R^2 = 0.1$ and the sample size is 300, the average estimated MSE for cross-validation across the 2000 repetitions is 1595.43, while for the *standard* bootstrap estimate, the MSE is 1594.28. Recall that the cross-validation methods are detailed in Table 2.3 and the bootstrap algorithms are described in Section 2.7.

Table 6.2: Summary of MSE_{obs} for cross-validation (CV) and the *standard* (Std) and 0.632 bootstrap algorithms. The point estimates are averaged across the 2000 repetitions.

| R^2 | Method | N_{100} | N_{300} | N_{1000} |
|-------|--------|-----------|-----------|------------|
| 0.01 | CV | 17931.20 | 17571.10 | 17357.90 |
| | Std | 17829.00 | 17546.50 | 17352.50 |
| | 0.632 | 17811.20 | 17534.90 | 17347.80 |
| 0.1 | CV | 1620.81 | 1595.43 | 1577.43 |
| | Std | 1618.57 | 1594.28 | 1576.99 |
| | 0.632 | 1630.10 | 1597.37 | 1577.99 |
| 0.3 | CV | 422.62 | 414.13 | 409.11 |
| | Std | 420.14 | 413.63 | 409.002 |
| | 0.632 | 419.92 | 413.27 | 408.83 |

However, the target performance of the methods is method-agnostic as the target performance is based on the performance across 2000 repetitions, each of which fits a prediction model using all data in the repetition and then evaluates the model in a larger internal validation set. Therefore, the target performance could be used to compare cross-validation ($\text{MSE}_{CV} - \text{MSE}_{target}$) to any of the bootstrap methods ($\text{MSE}_{BS} - \text{MSE}_{target}$).

Figure 6.14 presents results for the cross-validation, 0.632 and *standard* bootstrap when compared to the ideal or pragmatic target performance. The graph is representative of the various missing data scenarios (all graphs are available in Section S3.4 of the Supplementary Plots).

For all sample sizes the various methods combining MI with the bootstrap tend to perform similarly for the *standard* and 0.632 variations across all scenarios. The K -fold cross-validation methods tend to be more variable than the bootstrap methods for smaller sample sizes with a tendency to overestimate the ideal or pragmatic performance of the target MSE. However, with increasing sample size, the three internal validation approaches tend to perform similarly when compared to the target performance of the MSE.

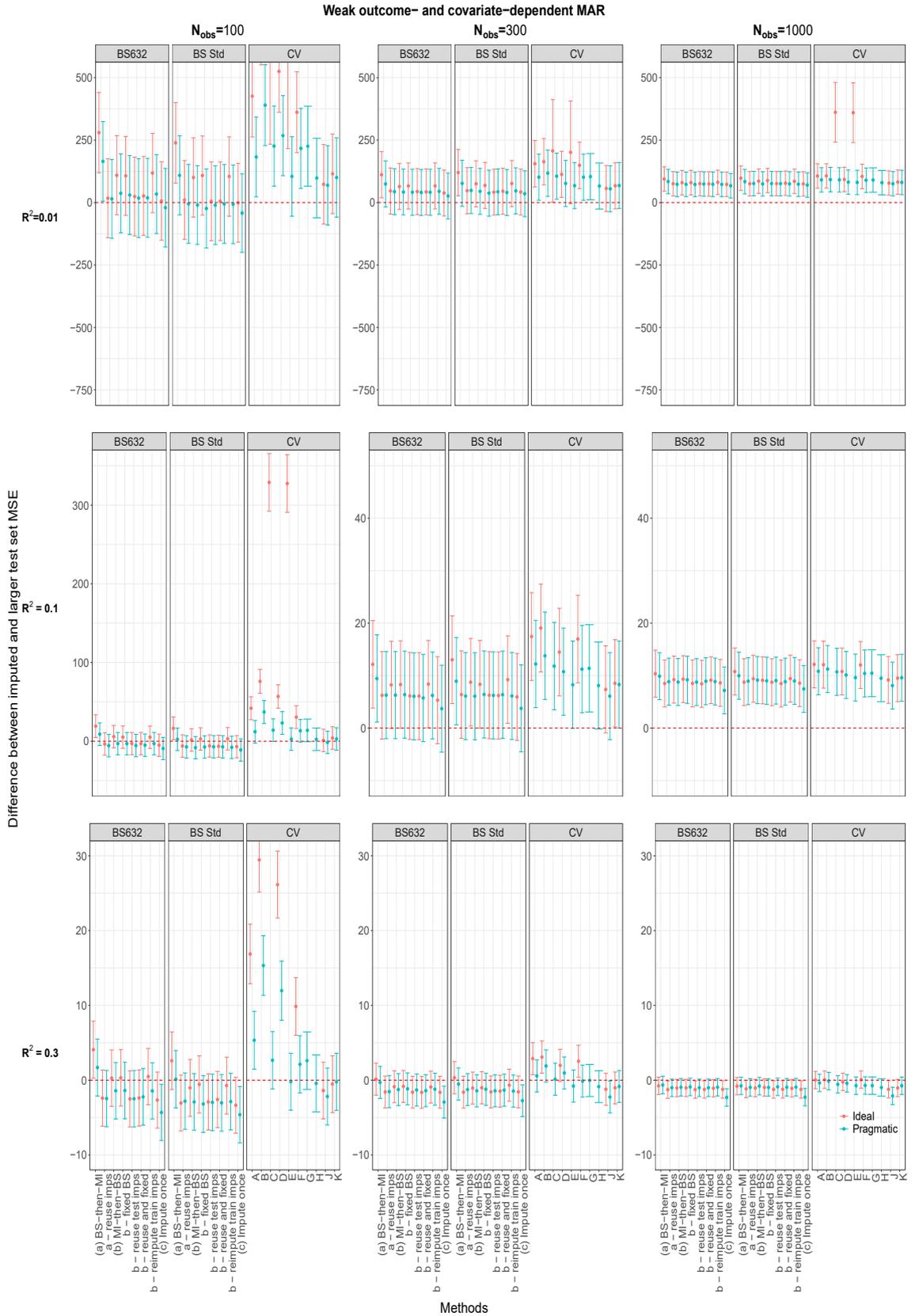


Figure 6.14: Comparing cross-validation, the 0.632 bootstrap (BS632) and the *standard* bootstrap (BS Std) using the target MSE. Error bars of the difference in the imputed MSE and the MSE estimate from a larger validation set are presented for the weak outcome- and covariate-dependent MAR scenario. CC (complete-case); CV methods A-K are described in Table 2.3; bootstrap methods are described in Section 2.7 or Table 6.1.

6.8 Discussion of the results when the outcome is continuous

The aim of the simulation study discussed in this chapter was to identify the most appropriate way to combine the 0.632 or the *standard* bootstrap optimism-corrected algorithm with MI.

In general, the impact of data leakage was most noticeable for smaller sample sizes, with the method with no data leakage (*BS-then-MI*) having a larger difference between the imputed and fully-observed MSE than all other imputation methods, which were subject to data leakage. The methods with the most data leakage (*BS-then-MI reuseimps*, *MI-then-BS reuse testimps*, *MI-then-BS impute once*) tended to underestimate the MSE when data were fully-observed for ideal performance i.e. they were over-optimistic. However, their pragmatic performance was similar to those methods subject to moderate data leakage such as method *MI-then-BS*.

With increased sample size all methods tended to perform similarly when compared to the fully-observed MSE suggesting that the impact of data leakage through the imputation process is lessened with increasing sample size. With a larger sample size there are more observations available when fitting an imputation model and I suggest that the influence of each observation on the posterior predictive distribution from which the imputed value will be sampled from is lessened.

For the *standard* bootstrap algorithm, the reuse of testing imputed datasets (imputed using all available data in the original dataset and originally used to estimate the apparent performance) was shown to perform similarly to imputing the original dataset a second time using a testing imputation model for ideal or pragmatic performance. However, the MSE estimate from reusing the testing imputed datasets and restricting the imputed datasets to those that were bootstrap sampled in order to estimate the bootstrap performance was shown to underestimate the MSE when data were fully-observed i.e. they were over-optimistic. This underestimation/over-optimism could also be seen for the 0.632 algorithm when reusing test imputed datasets, restricting to those observations who were not bootstrap sampled and evaluating the bootstrap prediction model to estimate the test performance.

The impact of data leakage was seen for smaller sample sizes in the simulation study which focused on the simple scenario of two covariates related to an outcome. Real datasets are typically larger and have more complex relationships relating an outcome to covariates and also relating covariates to each other. In more complex situations, it is possible that data leakage could impact studies of much larger sample sizes.

In addition, the simulation study showed that increasing the number of imputed datasets from 5 to 25 had little to no impact on the results. It was also seen that an increase in the percentage of missingness tended to lead to a larger over- or underestimation of the MSE estimate when data were fully-observed.

Overall, we have learned that when the outcome is continuous the *standard* and 0.632 bootstrap algorithms have similar ideal and pragmatic performance and that data leakage is present through the imputation process.

6.9 Summary and discussion of results when the outcome is binary

The aim of the simulation study, when the outcome is binary, is to determine an appropriate method to combine MI and the bootstrap optimism-corrected algorithms (*standard* and 0.632). The methods were assessed for several performance measures (AUC, Brier score and calibration intercept and slope) and were compared to both the estimate when data were fully-observed and also to a ‘target value’ from a larger validation set (Section 3.6). Factors such as increasing the strength of missingness, the percentage of missingness and the number of imputed datasets were also assessed. In addition, the impact of data leakage was investigated. A fully-written report of the results for the *standard* and 0.632 methods is available in Appendix C. Here, I shall briefly summarise and discuss the results.

Overall, it was found that the number of imputed datasets had minimal effect on the performance measure estimates of the AUC, Brier score or calibration intercept and slope which supports the findings found in Sections 6.4.2 and 6.5.2 when the outcome is continuous.

For pragmatic performance, increasing the percentage of missingness was shown to increase the magnitude of the under- or overestimation of the various performance measures when compared to the performance when data are fully-observed ($|\text{Perf}_{imp} - \text{Perf}_{obs}|$). For ideal performance, the majority of the methods performed similarly or had a slightly increased magnitude of the difference suggesting that the percentage of missingness has less impact on the ideal performance than it does on the pragmatic performance. This is most likely due to the inclusion of the outcome in the test imputation model used to estimate the ideal performance (which will help reduce bias in the regression model coefficients [53] and, therefore, produce ‘better’ predictions). The outcome is not included in the test imputation model for the pragmatic performance, hence why the pragmatic performance tends to have a larger magnitude of the difference, when compared to Perf_{obs} , than the ideal performance.

The performance of the calibration intercept and slope tended to be unstable for small

sample sizes. As previously discussed in Section 5.8 (cross-validation results when the outcome is binary), the estimation of calibration can be unstable for small sample sizes which may lead to miscalibration, this holds true even if shrinkage or penalisation methods are implemented [30,52].

The impact of data leakage on the various performance measures was examined when the outcome is binary. The impact of data leakage has previously been explored for cross-validation when the outcome is continuous or binary and for the bootstrap algorithms when the outcome is continuous. Data leakage through the imputation process was most noticeable for small sample sizes when the estimates of performance from the methods were compared to the fully-observed or target estimates of performance. With increasing sample size, the impact of data leakage through the imputation process appeared to decrease. However, given that the simulation study was based on the simple scenario of two covariates related to an outcome and lacks the complexity seen in real-life observational data, the impact of data leakage could be more serious in larger sample sizes.

Similarly, to the exploration of data leakage when the outcome is continuous (Section 6.6) it was noted that the ideal performance of the methods subject to data leakage (*BS-then-MI reuse imps*, *MI-then-BS reuse test imps*, *MI-then-BS impute once*) tended to be over-optimistic for the MSE, AUC and Brier score when compared to the estimate of performance if data had not been missing. I argue that in practice, when using real data where the impact of data leakage could be more severe, this over-optimism could lead to the implementation of prediction models in a clinical setting with an anticipated performance which is unrealistic. It may perhaps be better to have a prediction model which underestimates (but tends towards) the performance if no data had been missing than a model which is over-optimistic. In this scenario, at least if the model does not have ‘acceptable’ performance the researcher can always retune the model to improve it, which they may not do if they think the model is already performing well.

Based on this reasoning, in addition to previous literature [40,44,54], I recommend the method *BS-then-MI* as it avoids data leakage and performs similarly to the other methods which have the ‘advantage’ of data leakage for moderate and large sample sizes. In addition, the reuse of train imputed datasets (*BS-then-MI reuse*) to increase computational efficiency when training a bootstrap prediction model (and estimating bootstrap performance if using the *standard* bootstrap algorithm) is not recommended.

6.10 Conclusions

Overall, in this chapter we have learned that increasing the number of imputed datasets has little effect on the performance estimate and that data leakage through the imputa-

tion process could potentially be an issue in an analysis, if not carefully considered. With increasing sample size, the various methods tended to perform similarly when compared to the performance if data had not been missing. As such, I recommend the method *BS-then-MI* which also has the advantage of avoiding data leakage.

Internal validation not only assesses the evaluation of a prediction model, but also evaluates the entire analysis procedure which is used to arrive at a final model. As such, it is necessary that internal validation in the presence of missing data should be able to handle considerations such as transformations of continuous covariates or covariate selection. Chapter 7 will extend the methods developed for combining cross-validation or bootstrapping with MI using fractional polynomials to address these issues.

7 Multiple imputation and internal validation with fractional polynomial terms in the prediction model

7.1 Introduction

Previously, I assessed methods to combine MI with internal validation methods when the linear predictor of the prediction model was of a fixed, known form. In this chapter, I will explore how to combine the multivariable fractional polynomial algorithm with internal validation in the presence of missing data.

I will first summarise previous published literature on combining MI with fractional polynomials, before proposing how it might be combined with the previously proposed cross-validation or bootstrap algorithms. For cross-validation, I previously recommended methods A (impute each fold separately) or B (impute $k - 1$ training folds and k^{th} test fold separately). For the bootstrap algorithms, the default *BS-then-MI* was previously recommended. As it was shown in my investigations in earlier chapters that the 0.632 and *standard* bootstrap algorithms have similar performance (Sections 6.7 and C.9), in this section I will focus on adaptations for the 0.632 algorithm; the adaptation for the *standard* algorithm should follow similarly and would be expected to have similar performance to the 0.632 bootstrap (as this was seen in earlier chapters).

7.2 Background

Fractional polynomials and the multivariable fractional polynomial (MFP) algorithm were introduced in Section 1.7. Recall the usual set of fractional polynomial exponents is $S = (-2, -1, -\frac{1}{2}, 0, \frac{1}{2}, 1, 2, 3)$ where 0 represents a log transformation. The MFP algorithm allows for both the selection of covariates into a model and also for the transformation of continuous covariates. The MFP procedure has previously been adapted by Morris et al. to handle missing data using MI [21], in the context of parameter estimation. The two main parts to the paper focused on (i) the selection of an exponent to impute the missing data and (ii) estimation of the exponents and selection of covariates in multiply imputed data.

Part (i) of the paper aims to select the functional form of a continuous partially-observed covariate. The covariate can then be appropriately transformed before being imputed with respect to the other covariates and/or outcome in the dataset. Imputing a covariate with the correct functional form ensures that the association between the covariate and all other covariates and/or the outcome is maintained [18]. Part (ii) of the paper primarily focused on which exponent or covariates to include into a final ‘pooled’ model. A single choice of exponent for covariate transformation is required so that the covariate parameters are on the same scale to be combined. The focus of Morris et al. was on appropriate estimation

of parameters (i.e. an inference setting) where one final ‘pooled’ model is of interest to explore whether covariates are associated with an outcome.

However, as discussed in Wood, Royston and White’s paper [38], pooling the parameters of prediction models fitted to each imputed dataset (using Rubin’s rules) could lead to imprecise estimates of performance. In keeping prediction models unpooled, this therefore means that each imputed dataset’s prediction model can have different covariates or exponents selected. Therefore, only part (i) from Morris et al. is relevant when adapting MI and the MFP algorithm to handle internal validation.

When the transformation of a covariate involves fractional polynomials of degree 1, the imputation process recommended by Morris et al. involves using the approximate Bayesian bootstrap (ABB) method [21]. This samples individuals with fully-observed values in order to get a draw of the exponent which will be used to transform the covariate, in order to impute the partially-observed covariate.

The ABB method and how to impute using the selected exponent will now be detailed. I will outline the algorithm using a similar setting as the previous simulation studies, which I will extend upon: an outcome Y , a partially-observed covariate X_1 and a fully-observed covariate X_2 . The covariate X_1 has a transformation X_1^E , in relation to the outcome Y , where E is an exponent selected from the set S of fractional polynomials (Section 1.7). In the case of linear regression, $\mathbb{E}(Y | X_1, X_2) = \beta_0 + \beta_1 X_1^E + \beta_2 X_2$.

7.2.1 Approximate Bayesian bootstrapping

If observations from a dataset can be assumed to be independently and identically distributed then the ABB can be considered as a method for MI [55] (as a way to allow for uncertainty about exponents at the imputation stage). Therefore, it is possible to use the ABB to get a ‘complete’ dataset (sampled from those who have observed values of the partially-observed covariate) which can then be used to select an exponent which fits the observed data well. This exponent could then be used to transform the covariate for imputation.

The ABB has a few simple steps. To produce one imputed dataset:

1. Take a dataset \mathbf{D} and split it into those who have fully-observed records for the variable being imputed (if there are multiple incomplete covariates) \mathbf{D}_{obs} and those who do not \mathbf{D}_{miss}
2. Take a bootstrap sample with replacement from \mathbf{D}_{obs} to create a donor sample, \mathbf{D}^* .

7.2.2 Selecting an exponent in order to impute missing values

The imputation of missing values using a selected exponent was outlined in [21, Section 5.4] but will be detailed here for convenience. For a dataset $\mathbf{D} = \{Y, X_1, X_2\}$ which has fully-observed Y and X_2 . X_1 is partially-observed and may potentially have a FP transformation. The steps to impute X_1 while accounting for the uncertainty in the selection of an exponent for one imputed dataset are:

1. Apply the ABB algorithm, detailed in Section 7.2.1 Steps 1 to 2, to \mathbf{D} to get a ‘complete’ dataset \mathbf{D}^* as above.
2. Fit a regression imputation model to \mathbf{D}^* , regressing X_1^E on the outcome Y and the other relevant covariate X_2 for $E \in S$ i.e. the set fractional polynomial transformations. This may be done incrementally (-2, -1.9, -1.8, ...), although I have chosen to only use the exponents in set S to reduce computation time.
 - (a) Estimate the log likelihood, $\log(L)$ for each of the regression models fitted under various choices of E
 - (b) Calculate the value of the Jacobian adjustment, J , of the transformation of X_1 to X_1^E . Note: this is the log of the absolute derivative of the transformation and it is summed over all records used to fit the models. For example, if $E = 0$, the transformation of X_1 is $t(X_1) = \log(X_1)$. Then $|t'| = X_1^{-1}$ and $\log(|t'|) = -\log X_1$. Therefore, $J = \sum_{i=1}^{N_{\mathbf{D}^*}} -\log X_{1,i}$
3. Choose the exponent, draw E^* which has the largest value of $\log(L) + J$
4. For \mathbf{D} , multiply impute $X_1^{E^*}$ using linear regression of $X_1^{E^*}$ on Y and X_2
5. Finally, back-transform to get an imputed value of $X_1 = \sqrt[E^*]{X_1^{E^*}}$

7.3 Adapting the exponent selection and imputation process to handle prediction models

Section 7.2.2 detailed how to draw an exponent to impute a continuous covariate with missing values. This focused on using all available data in a dataset and using an imputation model (in Step 2) which included the outcome.

However, when adapting the algorithm for internal validation a few small alterations are needed. Above, \mathbf{D} represented the entire analysis dataset, but when extending the approach to a prediction setting consideration is required as to which observations should be included in the dataset used to fit a prediction model. In the prediction setting, \mathbf{D} could indicate a subset of the complete analysis data set. For example, when imputing the $k - 1$ training folds in cross-validation, \mathbf{D} could represent the observations from the

training folds.

For *MI-then-Validate* methods \mathbf{D} would be all records in the dataset. For *Validate-then-MI* methods where the ‘training’ and ‘test’ datasets are imputed separately, there will be two ‘versions’ of \mathbf{D} : $\mathbf{D}_{\text{training}}$ and \mathbf{D}_{test} . In cross-validation, $\mathbf{D}_{\text{training}}$ is the $k - 1$ training folds and \mathbf{D}_{test} is the k^{th} test fold. For the 0.632 bootstrap validation method, $\mathbf{D}_{\text{training}}$ contains those observations who were selected into the bootstrap to train a bootstrap prediction model. \mathbf{D}_{test} contains the observations which were not selected in the bootstrap sample (as I am only focusing on the 0.632 version in this chapter) and which will be used to evaluate the bootstrap prediction model.

Secondly, step 2 applies an imputation model to impute the transformed X_1 using the outcome and other relevant covariates. Extending this to the prediction setting requires considering whether an ideal or pragmatic setting is of interest. The inclusion of the outcome in a training imputation model to produce training imputed datasets of $\mathbf{D}_{\text{training}}$ is fine. However, the inclusion of the outcome in the test imputation model (to impute \mathbf{D}_{test}) will depend on whether ideal or pragmatic performance is of interest.

The previously proposed methods to combine MI and cross-validation will be extended to handle fractional polynomials. It is important to note that there will be two stages of exponent selection. Exponents will be selected when using ABB, in order to help impute the partially-observed covariate. Once the data have been multiply imputed, the FPS or MFP algorithms (first introduced in Section 1.7) can be applied to select the appropriate functional form of continuous covariates when fitting a prediction model.

In Section 1.11 I presented a data leakage example which highlighted how the choice of clusters can be affected based on using all the data or just the data that should be used for training a model. In a similar manner, the choice of exponents using all the data versus the data to be used for training or evaluating could cause data leakage.

In the following sections, I will incorporate fractional polynomial algorithms with the previously proposed *MI-then-Validate* and *Validate-then-MI* methods. While the *MI-then-Validate* methods are prone to data leakage and are not recommended, it is possible that a user may still wish to use this version of imputing and validating (as it is less computationally intensive than *Validate-then-MI*).

7.4 *MI-then-validate* methods

In this section I propose how the *MI-then-Validate* methods can be adapted for use when the analysis model includes FPs. The steps discussed in this section can be used for both

MI-then-CV and *MI-then-BS* methods. The methods detailed in this section are similar to those detailed in Chapter 2. The key difference is the use of the ABB algorithm to select the fractional polynomial exponent to be used in the imputation process. The *MI-then-CV* and *MI-then-BS* algorithms follow similarly, but allowing for the selection of fractional polynomials when fitting a prediction model.

1. \mathbf{D} is first defined to include all observations of the entire dataset. The ABB algorithm (Section 7.2.1) is applied to \mathbf{D} in order to select FP exponents (Steps 1-3 from Section 7.2.2). This is repeated M times in order to get a vector of exponents \mathbf{E} for X_1 of length M , $\mathbf{E}=\{E_1, \dots, E_M\}$
2. To obtain an imputed dataset $m = 1, \dots, M$:
 - (a) Using all of the available data, \mathbf{D} , transform covariate X_1 using the m^{th} element of \mathbf{E} , E_m i.e. $X_1^{E_m}$.
 - (b) Use a training and test imputation model to impute the observations with missing values $X_1^{E_m}$. This will produce one training imputed dataset and one test imputed dataset.
 - (c) Back transform $X_1^{E_m}$ to X_1 in both the training and test imputed datasets

Repeat this step until there are M training and test imputed datasets.

3. After obtaining M training and test imputed datasets, apply the internal validation algorithm of interest. As per the previous *MI-then-BS* or *MI-then-CV* methods, any prediction models should be trained on the training imputed datasets and evaluated on the test imputed datasets, as this will reduce correlation between the training and test imputed datasets. Please see Sections 7.4.1 and 7.4.2 for specific details.

Step 3 will be explained in more detail below for both cross-validation (method K: impute the dataset using a training and test imputation model, Section 2.6) and the bootstrap (default method *MI-then-BS*, Section 2.7).

7.4.1 *MI-then-CV*

For full details of method K, please refer back to Table 2.4. Briefly, this involved imputing the entire dataset twice, using both a training and test imputation model. Due to the nature of imputing first, followed by applying cross-validation, ABB selection was performed once to estimate M exponents for imputing X_1 . All observations in the imputed training datasets are divided randomly into K folds of equal size. The prediction model is fitted using the $k - 1$ folds from the training imputed datasets. The k^{th} holdout fold selected the observations belonging to fold k from the test imputed datasets in order to evaluate the prediction model trained on the $k - 1$ training imputed datasets.

Steps 1 and 2 to *MI-then-CV* were defined above. For $k = 1, \dots, K$, use the k^{th} fold as the holdout test fold and the remaining $k - 1$ folds as the training dataset for which to train a prediction model.

3. Within each training imputed dataset m_{train} :

(a) take the $k - 1$ training folds and apply the fractional polynomial selection algorithm (or MFP algorithm) to select the functional form of X_1 and train a prediction model $P_{m_{train},k}$

(b) Evaluate model $P_{m_{train},k}$ on the observations belonging to the k^{th} fold in the M_{test} imputed datasets to obtain M_{test} estimates of performance. Use Rubin's first rule to average across the performance estimates to get one overall estimate of performance for $P_{m_{train},k}$

4. Repeat Step (3) K times, holding out each fold $k = 1, \dots, K$ as the holdout test fold. This will produce K estimates of performance which are averaged (using cross-validation averaging rule, Section 1.9.3).

5. Repeat Steps (3)-(4), for each training imputed dataset $m_{train} = 1, \dots, M_{train}$ to get M_{train} estimates of performance. Use Rubin's first rule to get an overall estimate of performance across the M_{train} imputed datasets (i.e. take the average of the averaged performances obtained in step 4).

7.4.2 *MI-then-BS*

Recall that the 0.632 algorithm involves calculating the *apparent* and *test* performance. When data are fully-observed, a prediction model is trained and evaluated using all available observations in a dataset. The entire dataset is then bootstrap sampled, with replacement, B times. For a bootstrap sample b , a prediction model is fitted to those observations who were sampled. This 'bootstrap sample prediction model' is then evaluated in those observations who were not selected into bootstrap sample b , to estimate the *test* performance. Full details of the application of the bootstrap algorithm when using MI were given in Section 2.7.2.

Below I briefly summarise the extension of the algorithm detailed in Section 2.7.2 to incorporate fractional polynomials. Steps 1 and 2 for *MI-then-BS* were defined above.

3. First, the *apparent* performance will be estimated. Using the imputed datasets from Step 2, train M_{train} prediction models (fitted using either the FPS or MFP algorithm). These prediction models are fitted using all observations from the M_{train} training imputed datasets. Evaluate each prediction model in the M_{test} test imputed datasets and use Rubin's first rule to average across the M_{test} estimates to get one overall performance estimate for each model. Then apply Rubin's first rule again to

average across the performance for the M_{train} models to produce an overall estimate of the *apparent* performance.

4. Take a bootstrap sample b from the training imputed dataset m_{train} . Train a prediction model (using either the FPS or MFP algorithm) on the bootstrap sample b , allowing for the selection of fractional polynomials, and evaluate it in the observations which were not sampled in the M_{test} test imputed datasets from Step 2. Use Rubin's first rule to average across the M_{test} estimates (from each of the M_{test} test imputed dataset) to get an overall estimate for the prediction model.
5. Repeat Step 4 for $m_{train} = 1, \dots, M_{train}$ and use Rubin's first rule to average across the M_{train} prediction models' estimates of performance. This will give an estimate of the *test* performance.
6. The optimism-corrected performance may then be calculated as $(0.368 \times Apparent) + (0.632 \times \frac{1}{B} \sum_{b=1}^B Test_b)$

7.5 Validate-then-MI methods

In this section I propose how the *Validate-then-MI* methods can be adapted for use when the analysis model includes FPs. This shall adapt cross-validation methods A and B (Table 2.4) and the default *BS-then-MI* method for the 0.632 bootstrap algorithm (Section 2.7.1). In the 'validate first' methods involving cross-validation, we split the data into $k-1$ training and k^{th} test folds first before imputing. In the 'validate first' method using the 0.632 bootstrap, after estimating the apparent performance, take a bootstrap sample from the dataset which is partially-observed, and then impute the bootstrap sample. Recall that validating first is the recommended approach to combining missing data methods and internal validation, as this approach avoids data leakage.

7.5.1 CV-then-MI

I shall describe the adaption of method B to accommodate fractional polynomials, and detail the steps which need to be altered to instead use method A.

All observations in the data are divided randomly into K folds of equal size. Each fold k is used in turn as a holdout test fold, for $k = 1, \dots, K$:

1. To impute the $k-1$ training folds:
 - (a) Apply the ABB algorithm to the $k-1$ training folds ($\mathbf{D}_{training}$) in order to select M_{train} exponents for X_1 . This results in a vector of exponents \mathbf{E}_{train} of size M_{train} .
 - (b) To obtain the m_{train}^{th} imputed dataset of the $k-1$ training folds:

- i. Transform covariate X_1 using the m_{train}^{th} element of \mathbf{E}_{train} , $X_1^{E_{m_{train}}}$
 - ii. Use a training imputation model to impute $X_1^{E_{m_{train}}}$ (the imputation model will include covariate X_2 and outcome Y)
 - iii. Back transform $X_1^{E_{m_{train}}}$ to X_1 . One training imputed dataset has now been obtained for the $k - 1$ training folds.
- (c) Repeat Step 1b for $m_{train} = 1, \dots, M_{train}$ to obtain M_{train} training imputed datasets $\{Y, X_1, X_2\}$. These imputed datasets only contain the observations belonging to the $k - 1$ training folds.
2. Repeat Step 1 in order to impute the k^{th} test fold (\mathbf{D}_{test}). However, for Step 1(b)ii, the test imputation model (either including or excluding the outcome depending on whether ideal or pragmatic performance is of interest) should be used in place of the training imputation model. This test imputation model will be fit to the observations in the k^{th} test fold. This will produce M_{test} test imputed datasets.
3. For each training imputed dataset $m = 1, \dots, M_{train}$
- (a) Fit a prediction model (using either the FPS or MFP algorithm) to the m_{train}^{th} training imputed dataset.
 - (b) This prediction model will then be evaluated in the M_{test} test imputed datasets from Step 2 and use Rubin's first rule to average the performance estimates to get one overall estimate for the prediction model.
 - (c) These 3 steps are then repeated and an averaged performance estimate is obtained for each prediction model fitted to training imputed dataset $m = 1, \dots, M_{train}$.
4. Use Rubin's first rule to average across the M_{train} overall estimates of performance for each of the M prediction models from Step 3c.

Steps 1-3 are repeated, iteratively holding-out each fold as the test fold to produce K estimates of performance. These are then averaged using cross-validation averaging rules to get a performance estimate.

Method A

The Steps for Method A are similar to those detailed for method B. However for Step 1b, each fold is imputed separately using the ABB algorithm, instead of imputing the $k - 1$ folds together. The imputed datasets from all K folds are then combined together to produce M_{train} and M_{test} imputed datasets (i.e. an imputed dataset contains all K folds).

7.5.2 BS-then-MI

In this section, I will extend the 0.632 algorithm to handle fractional polynomials in the imputation process. Recall that the 0.632 algorithm estimates the *apparent* performance (where a model is trained and evaluated using all of the data) and the *test* performance (a prediction model trained in a bootstrap sample is evaluated in those who were not selected in the sample).

The steps are as follows:

1. The *apparent* performance is estimated using the same steps as those in Step 3 of Section 7.4.2.
2. Sample from the original partially-observed data, \mathbf{D} , with replacement to get a bootstrap sample b , $\mathbf{D}_{training}$.
 - (a) Use the ABB algorithm on $\mathbf{D}_{training}$ to select M_{train} FP exponents. This will output a vector of exponents \mathbf{E}_{train} of size M_{train} .
 - (b) To obtain the m_{train}^{th} imputed bootstrap sample b
 - i. Transform covariate X_1 using the m_{train}^{th} element of \mathbf{E} , $X_1^{E_{m_{train}}}$
 - ii. Use a training imputation model (including covariate X_2 and outcome Y) to impute $X_1^{E_{m_{train}}}$
 - iii. Back transform $X_1^{E_{m_{train}}}$ to X_1
 - iv. Repeat for $m_{train} = 1, \dots, M_{train}$
 - (c) M_{train} imputed training datasets $\{Y, X_1, X_2\}$ of bootstrap sample b have now been obtained. Train a bootstrap prediction model (using either the FPS or MFP algorithm) in each bootstrap training imputed dataset, $P_{b, m_{train}}$.
 - (d) For the observations which were not sampled in bootstrap sample b , repeat Steps 2a and 2b, using a test imputation model in Step 2(b)ii to obtain M_{test} test imputed datasets.
 - (e) Evaluate each prediction model $P_{b, m_{train}}$ in the M_{test} test imputed datasets and use Rubin's first rule to get an overall estimate of performance for the model.
 - (f) Repeat Step 2e for each prediction model $P_{b, m_{train}}$ for $m_{train} = 1, \dots, M_{train}$.
 - (g) Use Rubin's first rule again to average the estimates of performance from the M_{train} prediction models from Step 2f, to get an overall estimate of *test* performance, $Test_b$.
3. Repeat Step 2 for $b = \dots, B$ to get B estimates of the *test* performance
4. The optimism-corrected performance is then estimated as:

$$OCP = (0.368 \times \text{Apparent}) + \left(0.632 \times \frac{1}{B} \sum_{b=1}^B Test_b \right)$$

7.6 Additional considerations when applying the proposed methods

7.6.1 Application of an origin-shift transformation before applying the FPS or MFP procedure

One final consideration which affects both *MI-then-validate* and *Validate-then-MI* in these extensions to accommodate FPs is the use of preliminary transformations on a covariate. Extreme covariate values can have an impact on the selection of the FP exponent. Royston and Sauerbrei (2007) [56] give the example that a FP2 model could be accepted over an FP1 model in order to improve the fit at a small number of extreme points, which is a type of overfitting. A linear transformation of the covariate to which the FP transformation is to be applied can be used to modify low values which are influential [56,57]. This also controls the range of X_1 in a way that improves precision (in the computing sense).

$$w_\delta(X_1) = \delta + (1 - \delta) \frac{X_1 - X_{1,min}}{X_{1,max} - X_{1,min}} \quad (7.1)$$

Equation 7.1 can be used to either transform a covariate which has negative values or one which has influential low values close to zero which could unduly affect the selection of a fractional polynomial exponent. A default value of $\delta = 0.2$ is recommended [56], and the transformed range of the values of the covariate will be $[\delta, 1]$. The use of a transformation can result in more sensible final models [56] which may overall reduce overfitting and improve predictions from a model.

7.6.2 Choice of α -levels in the FPS and MFP procedure

In both the FPS and MFP algorithms, a significance level α_E (denoted α_2 in [11]) controlling the significance level for choosing a FP exponent for a covariate must be decided. For example, the best fitting FP1 model will be compared to a linear model (a model which does not include a FP transformation) at the α_E level. Typically, $\alpha_E = 0.05$ but if $\alpha_E = 1$ then the best fitting FP transformation will be used.

Recall that the details of the MFP algorithm are available in Appendix A. Essentially, MFP has an additional significance level which must be decided in the algorithm. This is α_β (denoted α_1 in [11]) which controls the significance level for including any covariate into a model. Typically $\alpha_\beta = 0.05$ but it can be set to 1 to force a covariate into a model.

7.7 Conclusion

In this chapter I have summarised relevant current literature for combining MI with fractional polynomials. I have proposed how fractional polynomials can be incorporated into the previously proposed methods to combine internal validation with MI. Chapter 8 will detail a simulation study which will be used to evaluate the methods when using fractional polynomials.

8 Designing a simulation study to evaluate methods for combining MI and internal validation techniques while incorporating fractional polynomials and covariate selection

8.1 Introduction

This chapter describes the design of simulation studies that aim to evaluate methods for combining internal validation techniques with MI while handling the incorporation of the functional form of a continuous covariate and covariate selection. The simulation design is similar to the design in Chapter 3. I use this chapter to describe the simulation setup, and the simulation results will be described in subsequent chapters. The outline of this chapter follows the ADEMP structure recommended by Morris et al. [10] for clear reporting of simulation studies. I will assess the proposed methods for a continuous outcome.

Chapter 2 outlined proposed methods for combining MI with cross-validation (Section 2.6) and bootstrap algorithms (Section 2.7). These methods were then assessed using a simulation study and the recommended methods were *BS-then-MI* when using the bootstrap algorithm (Section 6.9). For cross-validation, it was recommended to cross-validate first and then apply MI. For smaller sample sizes, method A which involves imputing each fold separately was recommended, while for larger sample sizes method B which imputes the $k - 1$ training folds together and then imputes the k^{th} test set separately (Section 5.9).

8.2 Aim

In Chapters 2 to 6 I explored how to combine MI with internal validation techniques for a fixed prediction model i.e. the covariates included into the prediction model were already determined. I aim to extend these previous methods to a setting where the prediction model is not fixed but instead includes decisions based on the dataset available. I have incorporated fractional polynomials to the previously proposed methods in order to allow for covariate selection and the flexible transformation of continuous covariates when fitting a prediction model.

The aim of the simulation study in this chapter is to identify how the proposed methods combining internal validation with MI can be incorporated with fractional polynomial selection, for the incorporation of covariate transformation and selection into prediction models. This will be investigated across a range of different settings, including different exponents for variable transformation and multiple missing data mechanisms. The simulation studies will be used to assess the proposed methods for both cross-validation and the bootstrap optimism-corrected algorithms (detailed in Sections 7.4 and 7.5).

8.3 Data-generating mechanisms (DGM)

Similarly to the simulation study set-up in Section 3.2, a linear model will be used as the data-generating model when the outcome is continuous. The linear predictor will have two correlated Normally distributed covariates, X_1 and X_2 . They will be generated with correlation $\rho = 0.5$.

$$\begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \sim \mathcal{N} \left(\begin{bmatrix} 0.5 \\ 4 \end{bmatrix}, \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix} \right)$$

A FP transformation will be applied to covariate X_1 when generating the outcome. A requirement of fractional polynomials is that $X_1 \in \mathbb{R}_{>0}^+$. If the simulated X_1 value was less than 0.1, it was rejected and the sampling was repeated using the bivariate distribution.

The continuous outcome, Y , was generated using $Y \sim \mathcal{N}(\mu, \sigma^2)$, where

- $\mu = \beta_0 + \beta_1 X_1^E + \beta_2 X_2$
- $\sigma^2 = (\beta_1^2 \text{Var}(X_1^E) + \beta_2^2 \text{Var}(X_2) + 2\beta_1\beta_2 \text{Cov}(X_1^E, X_2)) \times \frac{1-R^2}{R^2}$

The outcome will be generated for three exponent choices: 2, 0 (log) and -2. The model is written like this in order to allow the variance of the outcome (σ^2) to be adjusted for varying levels of R-squared (R^2) while keeping the values of β constant. Two values were considered for R^2 : 0.1 and 0.3. The derivation of the adjustment to σ^2 to allow for different values of R^2 is available in Appendix B. The variance of X_1^E and its covariance with X_2 were estimated using simulated data based on one million sampled observations from the same distributions as described above. This was used to estimate the variance of Y (σ^2) and the values used for different combinations of exponent E and the R^2 are presented in Table 8.1.

Table 8.1: Variance of the outcome (σ^2) used in the simulation study for each choice of the exponent and level of R^2

| Exponent (E) | R^2 | SD(Y) |
|------------------|-------|-----------|
| 2 | 0.1 | 7.23 |
| | 0.3 | 3.68 |
| log | 0.1 | 4.15 |
| | 0.3 | 2.11 |
| -2 | 0.1 | 37.70 |
| | 0.3 | 19.20 |

In all simulation scenarios the linear line was set to intercept the origin ($\beta_0 = 0$). The β_1 value was set to one and β_2 was set to either zero or one. Allowing β_2 to vary will allow

for covariate selection assessment (when covariate X_2 should or should not be included in the prediction model) when using the MFP algorithm with internal validation.

In the simulation study, the dataset shall contain $\{Y, X_1, X_2\}$ where X_1 will be included in the analysis model while allowing for a fractional polynomial transformation.

8.3.1 Introducing missingness

As in the previous simulation set-up, missingness was induced in one covariate, X_1 , for the continuous outcome. Scenarios in which the missingness in X_1 does and does not depend on X_2 and/or on the outcome Y are considered. The probability of X_1 being missing for patient j is the same as in equation 3.1:

$$\pi_{X_1,j} = \frac{\exp(\psi_0 + \psi_2 X_{2,j} + \psi_3 Y_j)}{1 + \exp(\psi_0 + \psi_2 X_{2,j} + \psi_3 Y_j)}$$

Using this equation, three missing data scenarios were considered:

1. MCAR ($\psi_2 = 0, \psi_3 = 0$)
2. Covariate-dependent MAR ($\psi_2 \neq 0, \psi_3 = 0$)
3. Covariate- and outcome-dependent MAR ($\psi_2 \neq 0, \psi_3 \neq 0$)

For the two MAR mechanisms non-zero values of ψ_2 and ψ_3 were selected to produce weak and strong MAR. This strength was calibrated based on the area under a ROC curve (AUC) from regressing the missing indicator on the covariates related to missingness. Values for ψ_0 were then selected such that approximately 25% of observations in X_1 were set as missing.

Table 8.2 shows the finalised ψ parameter values and the AUC of missingness when the outcome is continuous. When missingness is MCAR or covariate-dependent MAR, missingness does not depend on the outcome and therefore the values of ψ are unaffected by the R^2 values. For covariate- and outcome-dependent MAR, the values of ψ are selected to maintain a similar missingness AUC for the three R^2 values.

Table 8.2: Specification of parameter values ψ_0, ψ_2, ψ_3 to ensure MCAR, weak MAR and strong MAR with approximately 25% ($\psi_{0,25}$) of observations induced to be missing.

| Mechanism | Exponent E | R^2 | ψ_3 | ψ_2 | $\psi_{0,25}$ | AUC |
|--|--------------|-----------|----------|----------|---------------|-------|
| MCAR | All E | All R^2 | 0 | 0 | -1.1 | 0.501 |
| weak covariate-dependent MAR | All E | All R^2 | 0 | 0.6 | -3.75 | 0.65 |
| strong covariate-dependent MAR | All E | All R^2 | 0 | 1.1 | -6.02 | 0.744 |
| weak outcome-dependent MAR | 2 | 0.1 | 0.07 | 0 | -1.58 | 0.644 |
| | | 0.3 | 0.12 | 0 | -1.85 | 0.643 |
| | log | 0.1 | 0.13 | 0 | -1.71 | 0.652 |
| | | 0.3 | 0.22 | 0 | -2.07 | 0.649 |
| | -2 | 0.1 | 0.015 | 0 | -1.32 | 0.659 |
| | | 0.3 | 0.025 | 0 | -1.40 | 0.651 |
| weak outcome- and covariate-dependent MAR | 2 | 0.1 | 0.07 | 0.35 | -3.11 | 0.678 |
| | | 0.3 | 0.12 | 0.35 | -3.40 | 0.687 |
| | log | 0.1 | 0.13 | 0.35 | -3.25 | 0.688 |
| | | 0.3 | 0.15 | 0.45 | -4.74 | 0.679 |
| | -2 | 0.1 | 0.015 | 0.42 | -3.15 | 0.683 |
| | | 0.3 | 0.02 | 0.5 | -3.25 | 0.667 |
| weak outcome- and strong covariate-dependent MAR | 2 | 0.1 | 0.15 | 0.35 | -3.80 | 0.785 |
| | | 0.3 | 0.25 | 0.35 | -4.38 | 0.787 |
| | log | 0.1 | 0.25 | 0.35 | -3.925 | 0.779 |
| | | 0.3 | 0.35 | 0.45 | -4.74 | 0.766 |
| | -2 | 0.1 | 0.028 | 0.42 | -3.45 | 0.772 |
| | | 0.3 | 0.05 | 0.5 | -4.00 | 0.776 |

8.3.2 Factors to vary in the simulation

I specified above (Section 8.3) that different simulation scenarios will be considered for two values of R^2 for the continuous outcome, six missing data mechanisms, and three exponent values E . I also considered two sample sizes ($n_{obs} = 300, 1000$). The proposed methods are therefore assessed across 288 different simulated scenarios. Each scenario was assessed with 2000 repetitions in order to minimise Monte Carlo error. The factors varied (factorially) across scenarios and their values are found in Table 8.3.

Table 8.3: Factors which will be varied for the continuous outcome simulations

| Factors | Notation | Values |
|--------------------------------------|------------|--------------------|
| Number of individuals | n_{obs} | {300, 1000} |
| Number of repetitions used | n_{sim} | {2000} |
| Proportion of missingness | p_{miss} | {25%} |
| Level of R-squared | R^2 | {0.1, 0.3} |
| Number of imputed datasets | M | {5} |
| Choice of FP exponent | E | {-2, 0, 2} |
| Inclusion of X_2 in analysis model | β_2 | {0, 1} |
| Dependence of missingness on X_2 | ψ_2 | Refer to Table 8.2 |
| Dependence of missingness on Y | ψ_3 | Refer to Table 8.2 |

8.4 Estimands

In each simulation scenario and using each analysis method (see below) I assess the ideal and pragmatic estimates of performance measures. The ideal and pragmatic performance measure estimates for each repetition will be compared to the performance measure estimated from the same repetition but with fully-observed X_1 (Perf_{obs}), similarly to the previous simulation study. We expect pragmatic estimates to underestimate those of ideal performance [38].

For all methods in either the ideal or pragmatic setting, an overall performance measure ($\widehat{\text{Perf}}_{imp}$) will be estimated. This will be compared to the performance measure Perf_{obs} :

$$\widehat{\text{Perf}}_{imp} - \text{Perf}_{obs}$$

While the main estimands of interest will be about the performance of predictions, I will also investigate how often the correct exponent is selected (i.e. the type I error rates) for both

- the selection of exponents (via ABB) which are used to impute missing values
- the selection of exponents from the MFP process when building the prediction model.

8.5 Methods

The proposed methods for combining MI with cross-validation and bootstrapping were initially proposed in Sections 2.6 and 2.7. After assessment in Chapters 4-6, only a small number of methods will be considered. The adaption of the cross-validation and bootstrapping methods to also handle fractional polynomials, in addition to missing data, was described in Sections 7.4 and 7.5. For cross-validation, *CV-then-MI* methods A (impute each fold separately) and B (impute the $k - 1$ training folds together and impute the

k^{th} test fold separately) and *MI-then-CV* method K (impute the entire dataset using a training and test imputation model) will be assessed. Methods A and B were found to be the most promising methods for *CV-then-MI* and method K has slightly reduced data leakage than if the dataset were imputed using the same imputation model and one set of M imputed datasets. As it was previously seen that the 0.632 and *standard* bootstrap algorithms have similar performance, I shall only assess the 0.632 algorithm here. For the 0.632 bootstrap, the default method for *BS-then-MI* and *MI-then-BS* will be assessed.

As I will only be simulating data with a continuous outcome, multivariate imputation by chained equations will be used. While this presents a possibility of negative imputed values, which would be problematic when back-transforming X_1^E to X_1 , this was not found to be an issue here due to careful selection of the mean and standard deviation of covariate X_1 .

As detailed in Section 7.6, an origin-shift transformation may reduce overfitting of models to the data. The application of this shift transformation will also be explored within each of the methods detailed above. Similarly, I will adjust the level of α_E to compare which exponents are selected (for Type I error assessment). I will also compare the impact of covariate selection for X_2 (when β_2 is either 0 or 1) by adjusting α_β , this will assess how often it is correctly added into a model or not.

Table 8.4 outlines the factors which will be varied in the assessment of the MFP procedure on imputed data.

Table 8.4: Factors which will be varied in the MFP analysis

| Factors | Notation | Values |
|---|----------------|-----------|
| Apply an origin-shift transformation (equation 7.1) | I_δ | {No, Yes} |
| Significance level for exponent selection | α_E | {0.05, 1} |
| Significance level for covariate selection | α_β | {0.05, 1} |

8.6 Performance Measures

In Section 8.4, I detailed the estimands of interest for assessing the various methods in this simulation study. This includes assessing the estimated MSE from the methods, how often the correct exponent is selected for imputing missing data, and also in the MFP procedure when fitting a prediction model. While the exponent selection is interesting to assess, a procedure which systematically selects the ‘wrong’ exponent (i.e. it is biased), may still have good performance. In addition, how often a covariate is correctly included into the prediction model will be assessed. Here, I shall detail how they will be estimated.

8.6.1 Assessment of the predictions

The performance measure for the prediction models when the outcome is continuous is the mean-squared error. For simulated replication, $r = 1, \dots, n_{sim}$, the mean MSE for each DGM is:

$$\widehat{\text{MSE}}_{imp} = \frac{1}{n_{sim}} \sum_{r=1}^{n_{sim}} \widehat{\text{MSE}}_{r,imp}$$

The fully-observed MSE will also be estimated:

$$\text{MSE}_{obs} = \frac{1}{n_{sim}} \sum_{r=1}^{n_{sim}} \widehat{\text{MSE}}_{r,obs}$$

with $\widehat{\text{MSE}}_{r,s} = \frac{1}{n_{obs,r}} \sum_{i=1}^{n_{obs,r}} (\hat{Y}_{i,r} - Y_{i,r})^2$ for $s = imp, obs$.

As outlined in Section 8.4, this averaged estimate will be compared with the averaged MSE when data are fully-observed ($\widehat{\text{MSE}}_{imp} - \text{MSE}_{obs}$). These are equivalent to the Perf_{obs} and Perf_{imp} notation outlined in Section 8.4.

8.6.2 Assessment of the exponent selection for imputation or the MFP procedure

The proportion of exponents selected for imputing data using ABB, or the exponents selected using the MFP procedure to build a prediction model will be estimated. The proportions for the various exponents will be compared to see whether certain exponents are favoured or whether the true underlying exponent, E , is selected.

For cross-validation, each repetition will have selected exponents for the K iterations of the training or test folds within M imputed datasets (i.e. $K * M * n_{sim}$ exponents). For bootstrapping, this will involve analysing exponents from the 2000 repetitions each of which will have selected exponents in B bootstrap samples and M imputed datasets ((i.e. $B * M * n_{sim}$) exponents).

8.6.3 Assessment of covariate selection in the MFP procedure

This will be estimated as the proportion of times that the covariate X_2 will be either correctly (when $\beta_2 = 1$) or incorrectly (when $\beta_2 = 0$) included in the trained prediction model.

Again, for cross-validation this will include assessment of K prediction models within M training imputed datasets for each repetition. For bootstrapping, this will involve assessing the inclusion of X_2 into a prediction model for B bootstrap samples and M imputed datasets for each repetition.

8.7 The ‘Target’ performance measure

In the previous simulation studies detailed in Chapter 3, I compared the estimated MSEs with both the MSE value when data were fully-observed. I also made comparisons with a ‘target’ value, which was estimated using a larger validation set (Section 3.6). Due to the number of comparisons to be made in the current simulation study, I will only compare the estimated MSE to the MSE value when data are fully-observed. As noted by Wood et al. (2015) [38, Section 4.1], comparing the estimated ideal performance with the fully-observed performance estimate is the gold standard. However, comparing the estimated pragmatic performance to the fully-observed performance estimate is not considered to be the correct comparison. The estimated pragmatic performance is expected to be ‘lower than ideal model performance’ [38, Section 6.1].

8.8 Conclusion

In this chapter I have discussed the set-up of the simulation study to be used for both cross-validation and the bootstrap internal validation algorithms when incorporating fractional polynomials. I have outlined the comparison of the performance estimates from the proposed methods in the simulation study to the estimate from the fully-observed data. In addition, I have stated how I will compare the selection of the exponents.

9 Simulation study results for combining multiple imputation, internal validation and fractional polynomials

9.1 Introduction

In Chapter 1 I introduced fractional polynomials and in Chapter 7 I summarised existing work on how they can be combined with MI. I proposed how fractional polynomials can be used for covariate selection and the flexible transformation of continuous covariates when using internal validation to assess prediction models. In Chapter 8 I described the design of a simulation study to investigate the performance of various methods which combined MI, fractional polynomials and internal validation algorithms. In this chapter I present the results from the simulation study. The results in this chapter will aim to answer the following questions:

1. Do the proposed methods for combining MI and fractional polynomials with internal validation perform well in terms of the MSE?
2. Is the application of an origin-shift transformation, to remove the influence of small extreme values, useful when multiply imputing?
3. Will the correct functional form be selected for covariate X_1 and, if not, how will this affect performance?
4. Will data leakage (present in the *MI-then-Validate* methods) lead to over-optimistic results?

The methods to be assessed were previously described in Chapter 7 and the estimated MSE from the methods will be compared to the MSE when data are fully-observed i.e. $MSE_{imp} - MSE_{obs}$. This will be assessed for various data-generating mechanisms which involve several values of R-squared and sample size for three choices (-2, 0, 2) of a ‘true’ underlying exponent relationship between X_1 and Y (the various factors varied in the simulations are found in Table 8.3). The model used in the simulation study to predict values of Y , is a linear regression model of the form $\mathbb{E}(Y | X_1, X_2) = \beta_0 + \beta_1 X_1^E + \beta_2 X_2$.

In addition to various data-generating mechanisms, there are several parameters which will be varied in the analysis procedure which are presented in Table 9.1. For $\beta_1 = \beta_2 = 1$, the impact of an origin-shift (which removes the influence of low values of X_1 on exponent selection) on $MSE_{imp} - MSE_{obs}$ will be assessed (Section 7.6). The impact of using an origin-shift transformation for X_1 on the selection of exponents will also be investigated. The origin-shift is a decision made during the analysis procedure i.e. whether to use an origin-shift before fitting a prediction model which allows for fractional polynomials.

The selection of exponents for covariate X_1 in the FPS procedure (introduced in Section 1.7) will be assessed for two significance levels (α_E). When $\alpha_E = 1$, this will assess which exponent from set S ($(-2, -1, -\frac{1}{2}, 0, \frac{1}{2}, 1, 2, 3)$) was selected as the best fitting exponent in the prediction model. $\alpha_E = 0.05$ will assess whether the best selected exponent improves the fit of the prediction model to the data, instead of the covariate being included linearly. The impact of α_E will also be assessed when comparing $MSE_{imp} - MSE_{obs}$. A similar assessment will be conducted for the MFP algorithm (which assesses both exponent and covariate selection) when $\beta_1 = 1$ and $\beta_2=1$. Covariate selection will also be assessed when covariate X_2 is not related to the outcome ($\beta_2=0$). Finally, how often the correct exponent is selected for imputing the missing covariates (via the Approximate Bayesian Bootstrap) will also be analysed.

Table 9.1: Parameters in the analysis procedure which will be assessed. Fractional polynomial selection (FPS) focuses on exponent selection while multivariable fractional polynomial (MFP) selection assesses exponent and covariate selection.

| Algorithm | β_2 | Analysis parameters |
|-----------|-------------------|---|
| FPS | $\beta_2=1$ | $I_\delta = \text{No, Yes}$ $\alpha_E = 0.05, 1$ |
| MFP | $\beta_2=\{0,1\}$ | $I_\delta = \text{No, Yes}$ $\alpha_E = 0.05, 1$ $\alpha_\beta = 0.05, 1$ |

I_δ = origin-shift transformation

α_E = exponent selection significance level

α_β = covariate selection significance level

This analysis will be conducted for both cross-validation and the 0.632 bootstrap algorithm. Due to the large number of scenarios, this chapter will focus on overall messages related to the various data-generating scenarios and analysis parameters. The full set of results are provided in Supplementary plot sections [S5-S9](#).

I shall first assess the proposed methods combining MI, internal validation and fractional polynomials when using FPS (exponent selection but no covariate selection). This will assess the performance of the methods when using the MSE as a performance measure. I will also investigate how often the correct exponent is selected via ABB (to impute X_1) and via FPS (exponent selection in the prediction model). Secondly, I will then assess the performance of the proposed methods when using MFP (this involves both exponent and covariate selection). Due to the similarity of of the MFP results to the FPS results, I will present and discuss figures for the covariate selection element of the MFP algorithm.

The methods which will be assessed for cross-validation are methods A (impute each fold

separately), method B (impute the $k-1$ training folds together and impute the k^{th} test fold separately) and method K (impute the entire dataset using a training and test imputation model, then apply cross-validation). For the 0.632 bootstrap algorithm, I will assess *BS-then-MI* and *MI-then-BS*. Methods A, B and *BS-then-MI* belong to the class of methods titled *Validate-then-MI* and methods K and *MI-then-BS* belong to the *MI-then-Validate* group of methods.

9.2 Summary of the simulated fully-observed data

I begin by summarising the fully-observed data, which is the simulated data before missing values are introduced to covariate X_1 . Table 9.2 presents a summary of the outcome, Y , for the simulated data. The distribution of Y changes depending on the exponent generating it. The outcome tends to have the largest standard deviation, as well as minimum and maximum values of Y when the true underlying exponent is -2. The mean and standard deviation approximately match the underlying values which were used to simulate the data (Table 8.1).

Table 9.2: The mean and variance of the outcome Y across the 2000 simulated datasets. The min and max values of Y are the minimum and maximum across all repetitions.

| E | R^2 | N_{obs} | Mean Y | SD(Y) | Min(Y) | Max(Y) |
|-----|-------|-----------|----------|-----------|------------|------------|
| -2 | 0.1 | 300 | 9.68 | 39.71 | -165.34 | 231.35 |
| | | 1000 | 9.76 | 39.73 | -188.57 | 242.34 |
| | 0.3 | 300 | 9.69 | 22.89 | -81.31 | 167.76 |
| | | 1000 | 9.74 | 22.92 | -92.53 | 174.44 |
| 0 | 0.1 | 300 | 4.10 | 4.37 | -16.66 | 23.94 |
| | | 1000 | 4.10 | 4.38 | -17.29 | 25.88 |
| | 0.3 | 300 | 4.10 | 2.52 | -8.04 | 15.77 |
| | | 1000 | 4.10 | 2.53 | -8.46 | 16.42 |
| 2 | 0.1 | 300 | 5.87 | 7.62 | -28.95 | 48.95 |
| | | 1000 | 5.87 | 7.62 | -30.54 | 52.79 |
| | 0.3 | 300 | 5.87 | 4.40 | -13.05 | 44.69 |
| | | 1000 | 5.87 | 4.40 | -14.26 | 41.26 |

E : the true exponent used to generate Y ; R^2 : R-squared value;

N_{obs} : the number of observations in a simulated dataset

Table 9.3 displays the estimates of the MSE for the various data-generating scenarios and analysis parameters when the data are fully-observed. When the true underlying exponent is 0 (log) or 2, the choice of α_E or the application of an origin-shift has little impact on the estimated MSE. However, when E is -2, the MSE is much larger. This is most likely due to the large variation in the outcome, Y . Increasing the level of R^2 results in a lower

MSE for all choices of exponent.

Table 9.3: The MSE estimates when data are fully-observed (MSE_{obs}) for cross-validation and the 0.632 bootstrap

| E | α_E | origin-shift (I_δ) | 0.632 bootstrap | | Cross-validation | |
|-------------|------------|--------------------------------|-----------------|------------------|------------------|------------------|
| | | | $N_{obs} = 300$ | $N_{obs} = 1000$ | $N_{obs} = 300$ | $N_{obs} = 1000$ |
| $R^2 = 0.1$ | | | | | | |
| -2 | 0.05 | No | 1439.87 | 1426.77 | 1441.48 | 1427.53 |
| | 0.05 | Yes | 1540.30 | 1522.33 | 1546.11 | 1523.22 |
| | 1 | No | 1438.08 | 1426.77 | 1440.32 | 1427.53 |
| | 1 | Yes | 1532.85 | 1522.20 | 1535.02 | 1521.20 |
| 0 | 0.05 | No | 17.47 | 17.33 | 17.52 | 17.36 |
| | 0.05 | Yes | 17.46 | 17.38 | 17.51 | 17.36 |
| | 1 | No | 17.45 | 17.32 | 17.50 | 17.33 |
| | 1 | Yes | 17.42 | 17.32 | 17.46 | 17.33 |
| 2 | 0.05 | No | 53.04 | 52.62 | 53.26 | 52.69 |
| | 0.05 | Yes | 53.07 | 52.61 | 53.23 | 52.67 |
| | 1 | No | 52.96 | 52.57 | 53.11 | 52.61 |
| | 1 | Yes | 52.90 | 52.55 | 53.01 | 52.58 |
| $R^2 = 0.3$ | | | | | | |
| -2 | 0.05 | No | 373.02 | 369.78 | 373.21 | 369.98 |
| | 0.05 | Yes | 469.34 | 466.44 | 470.29 | 466.76 |
| | 1 | No | 373.01 | 369.78 | 373.21 | 369.98 |
| | 1 | Yes | 469.07 | 466.44 | 470.15 | 466.76 |
| 0 | 0.05 | No | 4.54 | 4.49 | 4.56 | 4.49 |
| | 0.05 | Yes | 4.55 | 4.50 | 4.56 | 4.50 |
| | 1 | No | 4.53 | 4.49 | 4.54 | 4.49 |
| | 1 | Yes | 4.53 | 4.50 | 4.54 | 4.50 |
| 2 | 0.05 | No | 13.81 | 13.64 | 13.87 | 13.64 |
| | 0.05 | Yes | 13.80 | 13.66 | 13.85 | 13.66 |
| | 1 | No | 13.80 | 13.66 | 13.80 | 13.63 |
| | 1 | Yes | 13.75 | 13.65 | 13.77 | 13.66 |

9.3 Comparing results from the FPS procedure to the MSE estimate when data are fully-observed

In this section I will summarise results from the simulation study for both cross-validation and the 0.632 bootstrap algorithms. I will assess the impact of α_E (significance level for exponent selection) and using an origin-shift transformation (I_δ) on the various methods' estimated MSE. This estimated MSE is compared to the MSE when data are fully-observed for ideal or pragmatic performance (Sections 8.6 and 8.7) i.e. $\text{MSE}_{imp,perf,\alpha_E,I_\delta} - \text{MSE}_{obs,\alpha_E,I_\delta}$ [38].

Due to the large number of results from the simulation study, a small selection of graphs will be presented in this chapter. All graphs for the MSE comparison assessment are available in Supplementary Section S5.

9.3.1 The impact of α_E when an origin-shift transformation is not used

Recall that when α_E is 1, the best exponent for X_1 is selected. When α_E equals 0.05 a decision is made between including the covariate in the model with the best exponent and including it linearly, based on a hypothesis test with α_E set to 0.05. I will first assess the impact of α_E when no origin-shift transformation has been applied ($I_\delta = \text{No}$), followed by assessing its impact when an origin-shift has been implemented.

Results for the various methods are compared to the fully-observed data's MSE estimate when an origin-shift transformation has not been applied ($\text{MSE}_{imp,perf,\alpha_E,I_\delta=\text{No}} - \text{MSE}_{obs,\alpha_E,I_\delta=\text{No}}$). The results are available in Supplementary plot sections S5.1.1 (for cross-validation) and S5.2.1 (for the 0.632 bootstrap). A plot is not presented here as it is not very informative.

For both *Validate-then-MI* (cross-validation methods A and B; *BS-then-MI*) and *MI-then-Validate* (cross-validation method K; *MI-then-BS*), the magnitude of the MSE difference and the Monte Carlo confidence intervals are very large. For example, when $E = -2$, $\alpha_E = 0.05$ and sample size is 1000, the estimated MSE difference for method A is 81.60 for ideal performance and 109.23 for pragmatic performance. These estimates are similar for $\alpha_E = 1$ (with differences in the third decimal place). Methods B and K give much larger differences for both the ideal and pragmatic performance when sample size is 300 or 1000. The estimated differences between the MSE and fully-observed MSE are at least $1.00e+26$. The proposed bootstrap imputation methods have similarly large estimates of the MSE difference. An explanation for why such large estimates occur is available, with an example, in Section 9.3.2. Of all the proposed imputation methods, method A (impute each fold separately) tends to have the smallest differences between the estimated and fully-observed MSE for ideal or pragmatic imputation ($|\text{MSE}_{A,perf,\alpha_E,I_\delta=\text{No}} - \text{MSE}_{obs,\alpha_E,I_\delta=\text{No}}|$)

across all scenarios when an origin-shift has not been used.

Using the complete-case analysis with either cross-validation or the 0.632 bootstrap tends to result in much smaller magnitudes of the MSE difference. The difference between the estimated MSE and $\text{MSE}_{obs, \alpha_E, I_{\delta} = N_0}$ decreases when α_E is increased from 0.05 to 1. For example, when sample size is 300, $E = -2$ and when data are weak covariate-dependent MAR, the magnitude of the mean difference is 9.93 when $\alpha_E = 0.05$ and 7.48 when $\alpha_E = 1$. When sample size is 1000 both values of α_E result in a similar difference of 1.60 between the estimated and fully-observed MSE. This is similar for all MCAR or covariate-dependent MAR scenarios when using the complete-case analysis. When data are outcome-dependent or outcome- and covariate-dependent MAR and for both values of α_E , the magnitude of the difference increased for all exponents. The complete-case analysis tended to have a small magnitude of the difference ($|\text{MSE}_{CC, \alpha_E, I_{\delta} = N_0} - \text{MSE}_{obs, \alpha_E, I_{\delta} = N_0}|$) for the majority of data-generating scenarios but tends to underestimate the fully-observed estimate of the MSE (i.e. it is over-optimistic) when data are outcome-dependent MAR.

Overall in this section, $\alpha_E = 1$ tends to provide similar or slightly worse performance to $\alpha_E = 0.05$. However, due to the large magnitudes of the difference for the majority of the methods, a comparison of the MSEs was not very informative. An explanation for why such large differences are seen for the proposed methods involving imputation is available in the next section.

9.3.2 An explanation for the large MSE estimates when an origin-shift transformation is not used

In Section 9.3.1 it was stated that the difference between the methods' estimated MSE and the MSE when data are fully-observed is very large when an origin-shift transformation is not used. These large mean MSE estimates (the MSE is averaged across the 2000 repetitions) are due to a few repetitions having very large MSE estimates. The question then arises, what is happening in some of these repetitions that is leading to large estimated MSE value?

These large MSE estimates are due to missing observations being imputed with incredibly small values. The prediction model is then developed in the imputed dataset and the FPS procedure is applied. In some instances, the FPS procedure selects a negative exponent. This can lead to a large predicted value for a data row with missing X_1 . I shall demonstrate with an example when $E = -2$, $N_{obs} = 300$, $R^2 = 0.1$, data are weak covariate-dependent MAR and the pragmatic performance is of interest. Across the 2,000 repetitions generated under this scenario, 139 had an MSE estimate greater than 100 when applying *MI-then-CV* method K. The mean of the MSE across the 2000 repetitions is $3.42e+10$ while the median is 17.78.

Within each repetition when applying method K, the data were first imputed and then cross-validation was applied. The folds were iteratively used as a test fold and the prediction model was fitted on the remaining $k - 1$ folds. For one of the 2,000 repetitions with a large estimated MSE, the estimated MSE was small when test fold $k = 1, \dots, 7$ (ranging from 9.59 to 24.69).

When fold 8 is used as a test fold, the prediction model is fitted to the $k - 1$ training set (folds 1-7, 9 and 10). The FPS procedure has selected exponent -1 for covariate X_1 when fitting the prediction model: $\mathbb{E}[Y | X_1, X_2] = -0.301 - 0.481X_1^{-1} + 1.251X_2$. In Table 9.4, I show the impact of small imputed values when predicting the outcome for two observations which had missing values for X_1 using this prediction model.

Table 9.4: An example of the impact of low imputed values on the performance of a prediction model

| Observation | Imputed value for X_1 | Predicted Y | Observed Y |
|-------------|-------------------------|---------------|--------------|
| i | 1.015 | 5.05 | -0.10 |
| j | 0.006255 | -73.43 | 7.81 |

For observation i , X_1 was imputed with a 'large' value 1.015 and has a predicted value of 5.051, while the observed value of Y is -0.098. For observation j , X_1 was imputed with

a small value 0.006255 which has led to a large predicted value of -73.434. Overall, when fold 8 is used as the test set, the estimated MSE is 2.606e+12.

The above example demonstrates how small imputed values in combination with negative exponents selected via FPS can lead to large predicted values. However, these extreme estimates can potentially be rectified by using an origin-shift transformation. Table 9.5 displays the imputed values of X_1 after applying an origin-shift. For both observations i and j , the difference between the predicted and observed Y has decreased which will lead to a lower estimate of the MSE. Overall, when fold 8 is used as the test set and an origin-shift transformation has been used, the estimated MSE has decreased from 2.606e+12 to 25.06.

Table 9.5: Applying an origin-shift transformation to X_1 to improve the performance of a prediction model

| Observation | Imputed value for X_1 with an origin-shift | Predicted Y | Observed Y |
|-------------|---|---------------|--------------|
| i | 0.2773 | 4.74 | -0.10 |
| j | 0.2005 | 0.61 | 7.81 |

In the following section, I will assess the impact of α_E when an origin-shift transformation has been used. All subsequent results that are presented in this chapter will be based on applying an origin-shift transformation.

9.3.3 The impact of α_E when an origin-shift transformation is used

Figure 9.1 presents results for the various methods to handle missing data alongside cross-validation with FP selection when compared to the MSE estimate when data are fully-observed when an origin-shift transformation has been applied ($\text{MSE}_{imp,perf,\alpha_E,I_\delta=Yes} - \text{MSE}_{obs,\alpha_E,I_\delta=Yes}$). The results in the graph are for the scenario when the underlying true exponent is -2, $R^2 = 0.1$ and data are MCAR or covariate-dependent MAR. Results from another data-generating scenario are presented in Table 9.6. The table presents point estimates and the Monte Carlo 95% confidence interval estimates for each of the methods from Figure 9.1 for the scenario when data are MCAR and sample size is 300. Figure 9.1 is also representative of the scenarios in which missing data are outcome-dependent (outcome-dependent MAR or outcome- and covariate-dependent MAR) or R^2 is 0.3. Additional graphs for these scenarios are available in Supplementary plot sections S5.1.2 (for cross-validation) and S5.2.2 (for the 0.632 bootstrap).

For the complete-case analysis when $E = -2$ and the sample size is 300, the magnitude of the difference between the estimated MSE and $\text{MSE}_{obs,\alpha_E,I_\delta=Yes}$ tends to decrease when α_E is increased from 0.05 to 1, as seen in Table 9.6. This is the case for both levels of R^2 and all missing data scenarios. When the sample size is increased to 1000, the performance for both values of α_E is similar when compared to the fully-observed estimate. When data are outcome-dependent MAR, the complete-case analysis estimate tends to underestimate $\text{MSE}_{obs,\alpha_E,I_\delta=Yes}$ (i.e. it becomes over-optimistic). These trends are similar when the true underlying exponent is 0 or 2, with the complete-case estimate having similar performance for both values of α_E when sample size is 300 or 1000. The similar performance for both values of α_E can be seen for both cross-validation and the 0.632 bootstrap (Supplementary plot sections S5.1.2 and S5.2.2).

For *MI-then-Validate* (Method K and *MI-then-BS*), the MSE difference $\text{MSE}_{obs,\alpha_E,I_\delta=Yes}$ when $\alpha_E = 0.05$ tends to be similar or slightly larger for both ideal and pragmatic performance than when $\alpha_E = 1$ across all exponent values (-2, 0, 2). This is illustrated in Figure 9.1 and Table 9.6 when the exponent is -2. When $\alpha_E = 1$, the magnitude of the difference for method *MI-then-BS* ($|\text{MSE}_{MI-Val,\alpha_E=1,I_\delta=Yes} - \text{MSE}_{obs,\alpha_E=1,I_\delta=Yes}|$) tends to be slightly larger than when $\alpha_E = 0.05$. In general, the estimated MSE difference tends to be similar for both values of α_E . This holds for all sample sizes, levels of R^2 and values of the exponent.

For *Validate-then-MI* (Method A, B, *BS-then-MI*) when sample size is 300, the magnitude of the difference when $\alpha_E = 1$, $|\text{MSE}_{Val-MI,\alpha_E=1,I_\delta=Yes} - \text{MSE}_{obs,\alpha_E=1,I_\delta=Yes}|$, tends to be larger than the magnitude of difference when $\alpha_E = 0.05$. This can be seen in Table 9.6 for methods A, B and *BS-then-MI*. However, as seen in Figure 9.1, with increasing sample

size when the exponent is -2 (or 0) the estimated performance of the methods (in relation to $\text{MSE}_{obs, \alpha_E, I_{\delta}=\text{Yes}}$) tends to become similar when using either value of α_E for the 0.632 bootstrap. When the exponent is 2, the magnitude of the difference tends to be larger for $\alpha_E = 1$ than for $\alpha_E = 0.05$. When the sample size is 1000 and we use $\alpha_E = 1$, the Monte Carlo confidence intervals (based on 2000 repetitions) tend to be slightly larger when using the cross-validation methods compared to using the 0.632 bootstrap i.e. the variance of the MSE across 2000 repetitions is larger for cross-validation than 0.632, potentially due to the lower number of folds used compared to the number of bootstrap samples. This can be seen for all data-generating mechanisms in Supplementary plot sections [S5.1.2](#) and [S5.2.2](#) for cross-validation and the 0.632 bootstrap.

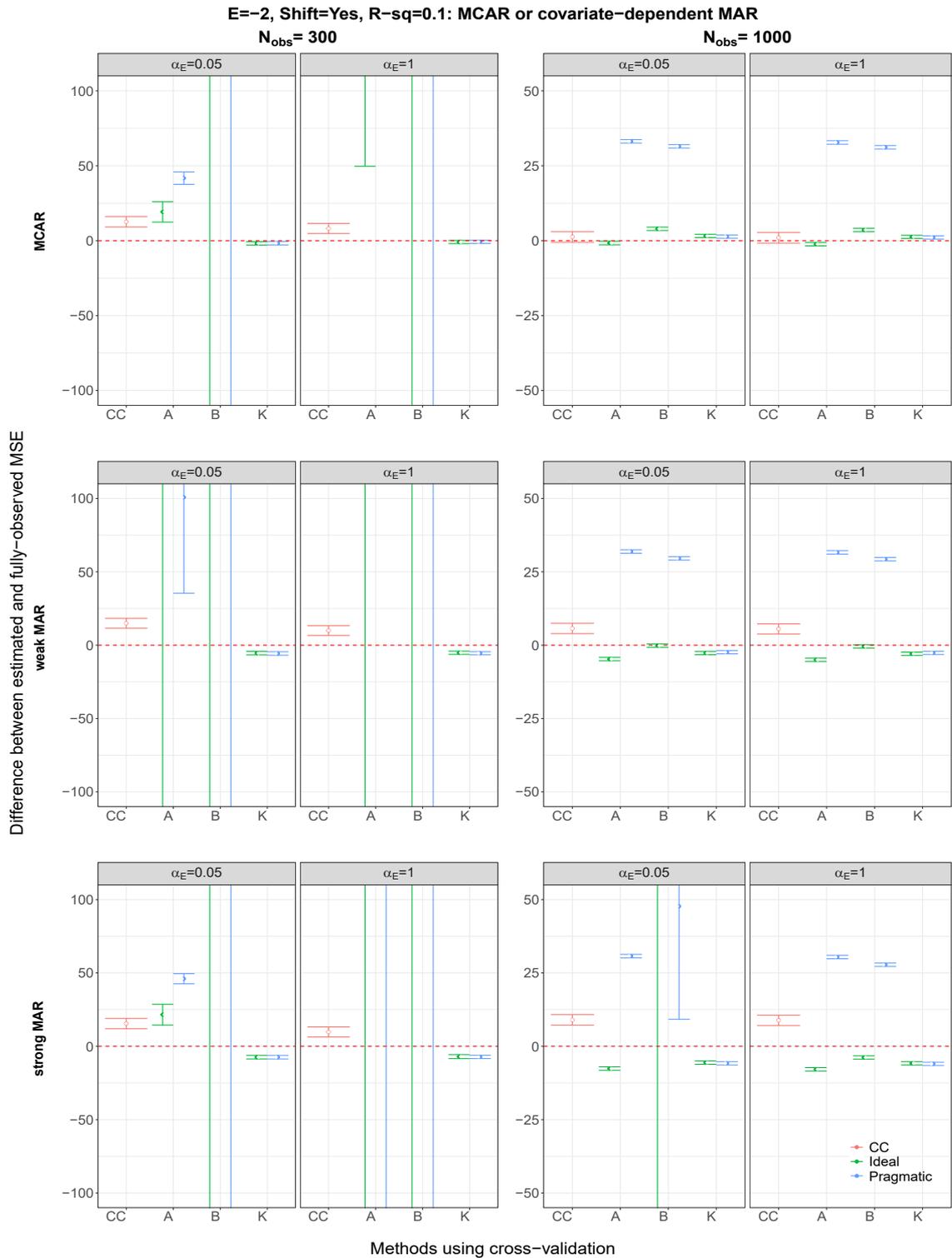


Figure 9.1: The difference $MSE_{imp} - MSE_{obs}$ when data are MCAR or covariate-dependent MAR for $M = 5$ when 25% of values are missing in X_1 . The true exponent, E , is -2, an origin-shift transformation has been applied and $R^2 = 0.1$. The error bars summarise results from the 2000 repetitions and the limits represent the Monte Carlo 95% confidence interval of $MSE_{imp} - MSE_{obs}$. CC (complete-case); methods A-K are described in Table 2.3.

Table 9.6: The estimated MSE and Monte Carlo 95% confidence interval (CI) when the exponent is -2 and an origin-shift has been applied. The presented results are for the scenario when sample size is 300, $R^2 = 0.1$ and data are MCAR.

| Method | Estimand | α_E | $MSE_{imp} - MSE_{obs}$ | Monte Carlo 95% CI | |
|---------------------|-----------|------------|-------------------------|--------------------|----------|
| Cross-validation | | | | | |
| CC | | 0.05 | 12.62 | 9.12 | 16.12 |
| | | 1.00 | 8.13 | 4.75 | 11.51 |
| A | Ideal | 0.05 | 19.20 | 12.38 | 26.02 |
| | | 1.00 | 225.41 | 49.63 | 401.19 |
| | Pragmatic | 0.05 | 41.75 | 37.60 | 45.90 |
| | | 1.00 | 664.23 | 162.38 | 1,166.08 |
| B | Ideal | 0.05 | 2.28e+09 ¹ | -9.47e+08 | 5.51e+09 |
| | | 1.00 | 4.93e+24 ² | -4.72e+24 | 1.46e+25 |
| | Pragmatic | 0.05 | 9.17e+11 ³ | -8.80e+11 | 2.71e+12 |
| | | 1.00 | 1.33e+16 ⁴ | -1.20e+16 | 3.86e+16 |
| K | Ideal | 0.05 | -1.81 | -2.98 | -0.65 |
| | | 1.00 | -0.09 | -1.98 | 0.14 |
| | Pragmatic | 0.05 | -1.70 | -2.85 | -0.55 |
| | | 1.00 | -0.8 | -1.85 | 0.24 |
| The 0.632 bootstrap | | | | | |
| CC | | 0.05 | 10.56 | 7.08 | 14.04 |
| | | 1.00 | 7.66 | 4.23 | 11.09 |
| BS-then-MI | Ideal | 0.05 | 25.60 | -4.99 | 56.185 |
| | | 1.00 | 210.26 | -43.31 | 463.84 |
| | Pragmatic | 0.05 | 36.10 | 32.33 | 39.88 |
| | | 1.00 | 2651.15 | -1035.68 | 6337.98 |
| MI-then-BS | Ideal | 0.05 | -20.24 | -21.38 | -19.10 |
| | | 1.00 | -18.75 | -20.04 | -17.46 |
| | Pragmatic | 0.05 | 8.56 | 7.31 | 9.81 |
| | | 1.00 | 11.07 | 9.88 | 12.26 |

¹ median (25th, 75th percentile): 31.35 (7.41, 102.97)

² median (25th, 75th percentile): 21.21 (2.66, 47.86)

³ median (25th, 75th percentile): 55.26 (23.70, 118.20)

⁴ median (25th, 75th percentile): 44.34 (24.52, 76.64)

9.3.4 The impact of the origin-shift transformation on the estimated MSE when using fractional polynomials

The performance of the various methods has been compared in Section 9.3.1 when assessing the impact of α_E . It was found overall that in small sample sizes, using $\alpha_E = 1$ compared with using $\alpha_E = 0.05$ had little effect when an origin-shift was not applied, but could result in slightly worse performance for *Validate-then-MI* methods when a shift had been used. In this section, I shall briefly summarise the impact of using an origin-shift transformation on the estimate MSE for the various methods.

Table 9.7 displays the estimated difference in MSE for the various methods when $E = 0$ and data are MCAR (although the results are representative of the various data-generating scenarios) in order to show the impact of using an origin-shift transformation. For the complete-case analysis the application of an origin-shift transformation has little effect on the magnitude of the difference between the estimated MSE and $\text{MSE}_{obs, \alpha_E, I_\delta}$, regardless of the value of α_E .

For all methods that involve MI (*Validate-then-MI* and *MI-then-Validate*) the use of an origin-shift transformation can help reduce the magnitude of the difference $|\text{MSE}_{imp, \alpha_E, I_\delta = \text{Yes}} - \text{MSE}_{obs, \alpha_E, I_\delta = \text{Yes}}|$. In Table 9.7 when $E = 0$ and the sample size is 300, for both the bootstrap and cross-validation, applying an origin-shift transformation has reduced the magnitude of the MSE difference. For example, the estimated ideal performance of method K for the difference $\text{MSE}_{K, \alpha_E = 0.05, I_\delta = \text{No}} - \text{MSE}_{obs, \alpha_E = 0.05, I_\delta = \text{No}}$ is $6.65e+11$ when an origin-shift transformation has not been applied. This decreased to -0.03 after applying the transformation.

As seen in Table 9.7 when sample size is 300 and an origin-shift transformation was not used, the magnitude of the MSE difference can be very large for methods B, K, *BS-then-MI* and *MI-then-BS*. This large magnitude increases with increased strength of the underlying missing mechanism. For the *Validate-then-MI* cross-validation methods, method A tends to provide much smaller estimates of the difference than method B i.e. the difference $\text{MSE}_{imp, \alpha_E, I_\delta = \text{Yes}} - \text{MSE}_{obs, \alpha_E, I_\delta = \text{Yes}}$ is smaller for method A, an explanation for this will be discussed in Section 9.7. All proposed methods for the 0.632 bootstrap provide smaller estimates of the difference in the estimated and fully-observed MSE for a sample size of 300 when an origin-shift transformation had been used (Table 9.7, Supplementary plot section S5.2.2). However, as we saw in Table 9.7 when $E = 0$, even the 0.632 bootstrap methods can have large estimates of the difference (the pragmatic performance of method *BS-then-MI* when $I_\delta = \text{No}$).

Table 9.7: Estimated MSE results before and after applying an origin-shift transformation (I_δ) when $E = 0$ and $\alpha_E = 0.05$. The presented results are for the scenario when sample size is 300, $R^2 = 0.1$ and data are MCAR. The averaged difference and Monte Carlo confidence interval (CI) are based on 2000 repetitions.

| Method | Estimand | I_δ | $MSE_{imp} - MSE_{obs}$ | Monte Carlo 95% CI | |
|---------------------|-----------|------------|-------------------------|--------------------|----------|
| Cross-validation | | | | | |
| CC | | No | 0.10 | 0.06 | 0.14 |
| | | Yes | 0.09 | 0.06 | 0.13 |
| A | Ideal | No | 2.39 | 0.60 | 4.17 |
| | | Yes | 0.39 | 0.29 | 0.50 |
| | Pragmatic | No | 3.61 | -1.77 | 8.98 |
| | | Yes | 0.85 | 0.09 | 1.61 |
| B | Ideal | No | 5.62e+16 | -5.39e+16 | 1.66e+17 |
| | | Yes | 1.62e+08 | -9.06e+07 | 4.14e+08 |
| | Pragmatic | No | 1.06e+15 | -1.01e+15 | 3.13e+15 |
| | | Yes | 7.12e+06 | -6.63e+06 | 2.09e+07 |
| K | Ideal | No | 6.65e+11 | -6.06e+11 | 1.94e+12 |
| | | Yes | -0.03 | -0.04 | -0.025 |
| | Pragmatic | No | 8.17e+13 | -7.85e+13 | 2.42e+14 |
| | | Yes | -0.04 | -0.05 | -0.03 |
| The 0.632 bootstrap | | | | | |
| CC | | No | 0.09 | 0.05 | 0.13 |
| | | Yes | 0.09 | 0.05 | 0.13 |
| BS-then-MI | Ideal | No | 3.87e+27 | -3.69e+27 | 1.14e+28 |
| | | Yes | 3.12 | -2.30 | 8.54 |
| | Pragmatic | No | 2.88e+27 | -2.76e+27 | 8.53e+27 |
| | | Yes | 2.21e+08 | -2.12e+08 | 6.55e+08 |
| MI-then-BS | Ideal | No | 4.49e+27 | -4.31e+27 | 1.33e+28 |
| | | Yes | -0.23 | -0.24 | -0.22 |
| | Pragmatic | No | 7.38e+27 | -7.07e+27 | 2.18e+28 |
| | | Yes | 0.03 | 0.03 | 0.05 |

9.3.5 An explanation for the large MSE estimates when an origin-shift transformation is used

In Section 9.3.2 I explained how the influence of small imputed values, paired with negative exponent selection could lead to large performance estimates. An origin-shift transformation was used post-imputation which removed the influence of the smaller extreme values. This was noted in Section 9.3.4 to have improved performance methods for the majority of methods across most data-generating scenarios. However, large performance estimates are still present for some methods across various data-generating scenarios.

Opposite to the explanation discussed in Section 9.3.2, the large performance estimates are due to large imputed values combined with a positive exponent selected via FPS when fitting a prediction model. I shall provide an example for the ideal performance of cross-validation method B when $E = -2$, $R^2 = 0.1$, sample size is 300 and data are MCAR (i.e. the scenario presented in Table 9.6). From the 2000 repetitions generated under this scenario and when $\alpha_E = 0.05$, the mean of the MSE across the 2000 repetitions when an origin-shift transformation is used is $2.28e+09$ and the median estimate of the MSE is 1605. Out of 2000 repetitions, 255 observations have an MSE estimate greater than 2000. Within each repetition when applying method B, the $k - 1$ training folds were imputed together and the k^{th} test fold is imputed separately. The folds are iteratively used as a test fold and the prediction model is fitted on the remaining $k - 1$ training folds. For one of the repetitions with a large estimated MSE, the MSE was small (MSE=1602) when fold 1 is used as the test fold but this increases to 464,589,635 when fold 2 is used as the test fold.

When fold 2 is used as the test fold, the prediction model is fitted to the $k - 1$ training set (folds 1, 3-10). When $\alpha_E = 0.05$, the fitted prediction model when using FPS is $\mathbb{E}[Y | X_1, X_2; \alpha_E = 0.05] = -9.29 + 51.49X_1 + 1.78X_2$ and when $\alpha_E = 1$ the prediction model is $\mathbb{E}[Y | X_1, X_2; \alpha_E = 1] = 1.50 + 267.81X_1^3 + 1.08X_2$. Table 9.8 demonstrates how large imputed values are poorly handled when using fractional polynomials, outputting large predicted values.

Table 9.8: An example demonstrating how large imputed values of X_1 can lead to poor predictions

| Observation | Imputed value for X_1 with an origin-shift | Observed Y | α_E | Predicted Y |
|-------------|---|--------------|------------|---------------|
| i | 5127 | -5.905 | 0.05 | 263,979 |
| | | | 1.00 | 3.61e+13 |
| j | 0.22 | 25.69 | 0.05 | 7.85 |
| | | | 1.00 | 7.84 |

9.3.6 Overall summary of the estimated MSE results

I have assessed the impact of α_E (α -level for FP exponent selection) and I_δ (origin-shift transformation) on the estimated MSE compared to the MSE when data are fully-observed for ideal and pragmatic performance. In general, $\alpha_E = 1$ tends to result in similar or slightly worse performance (i.e. a larger difference between the estimated and fully-observed MSE) for small sample sizes. In larger sample sizes, the value of α_E had little impact on the performance of the methods.

When the FPS algorithm was used and an origin-shift transformation was not implemented, for all sample sizes and when the exponent is -2 or 0, method A tended to have the lowest magnitude of the difference compared to methods B and K. When the exponent is 2 and sample size is 300, method A has a smaller magnitude than methods B and K. However, when sample size is increased to 1000, the ideal and pragmatic performance of method K tends to be closer to the MSE estimate when data are fully-observed. When sample size is 300, method *BS-then-MI* tended to have a larger magnitude of the difference than *MI-then-BS*. With increasing sample size, the ideal performance of method *BS-then-MI* tends to be smaller than *MI-then-BS* across the majority of simulation scenarios.

When an origin-shift transformation was used, method A generally performed well. The ideal and pragmatic performance of method K tended to have the smallest difference between its estimated MSE and the MSE when data are fully-observed. However, for some scenarios (for example: $n_{obs} = 1000$, $R^2 = 0.1$ and the true exponent E is 0 or -2), method K tended to underestimate the MSE estimate when data are fully-observed (i.e. it became over-optimistic) for both ideal and pragmatic performance. For these same scenarios, the ideal performance of both *CV-then-MI* methods (A and B) under-estimated the fully-observed MSE estimate when the exponent was -2. However, the pragmatic performance of methods A and B over-estimated the fully-observed MSE estimate (compared to the pragmatic performance of method K which was over-optimistic). When the exponent was 0, neither method A nor B was over-optimistic.

When an origin-shift transformation was used, the proposed 0.632 bootstrap methods tended to perform well with small magnitudes of the difference with the fully-observed estimate of the MSE. For ideal or pragmatic performance, method *MI-then-BS* tended to have a slightly smaller magnitude of the difference than *BS-then-MI* when both methods tended to over-estimate the fully-observed estimate of the MSE. However, the ideal performance of method *MI-then-BS* was over-optimistic in several data-generating scenarios for the three choices of the true underlying exponent (-2, 0, 2).

Overall, while the *MI-then-Validate* methods (cross-validation method K and *MI-then-BS*)

tended to have a smaller MSE difference in some scenarios, compared to *Validate-then-MI*, they also were more likely to have over-optimistic ideal or pragmatic performance. Preferred methods for use in practice will be discussed in Section 9.7.

9.4 Selection of exponents for imputation

In this section, I will present results for the selection of exponents (via the approximate Bayesian bootstrap method) which will be used to transform covariate X_1 before using MI (this was introduced in Section 7.2.2).

Figure 9.2 presents results for the exponents which are selected to transform X_1 to X_1^E in order to multiply impute the partially-observed covariate when using cross-validation to internally validate the prediction model. The results presented are for the scenario when the true exponent E is 0, $R^2 = 0.1$ and when data are MCAR or covariate-dependent MAR. However, the results in the figure are representative of all data-generating scenarios (Supplementary plot section S6.1). The selection of exponents occurs before MI is applied, therefore the selected exponents are similar regardless of whether ideal or pragmatic performance is of interest. The results presented in the graphs in this section and in Supplementary plot section S6.1 are based on the averaged results (across the 2000 repetitions) of the selected exponents for imputing.

In Figure 9.2 methods B (the $k - 1$ training folds) and K (which use 90% and 100% of the data, respectively), an exponent of 0.5 was selected in the majority of the repetitions, the reasoning for this will be discussed in Section 9.7. Across all data-generating scenarios, 0.5 was selected at least 92% of the time when the true exponent was 2, 84% when the true exponent was zero and at least 55% when E is -2. For method B (for the training folds) and method K, if 0.5 was not selected then the alternate exponent which was chosen tended to be zero. This was similarly seen for exponent selection of the 0.632 bootstrap. The exponent 0.5 was selected in over 80% of the $2000 * B * M$ datasets when imputing either all of the data, a bootstrap sample or those who were not sampled (Supplementary plot section S6.1).

When choosing the exponents using a smaller number of observations, either in method A (where each fold is imputed separately) or method B (when imputing the k^{th} test fold), more variability was introduced in the exponent which was selected. While 0.5 is still the most commonly selected exponent (see Figure 9.2), when the sample size is 300 exponents 0 and 1 are also selected. However, when the sample size is increased to 1000 and the number of observations in one fold increases from 30 to 100, 0.5 tends to dominate as the selected exponent.

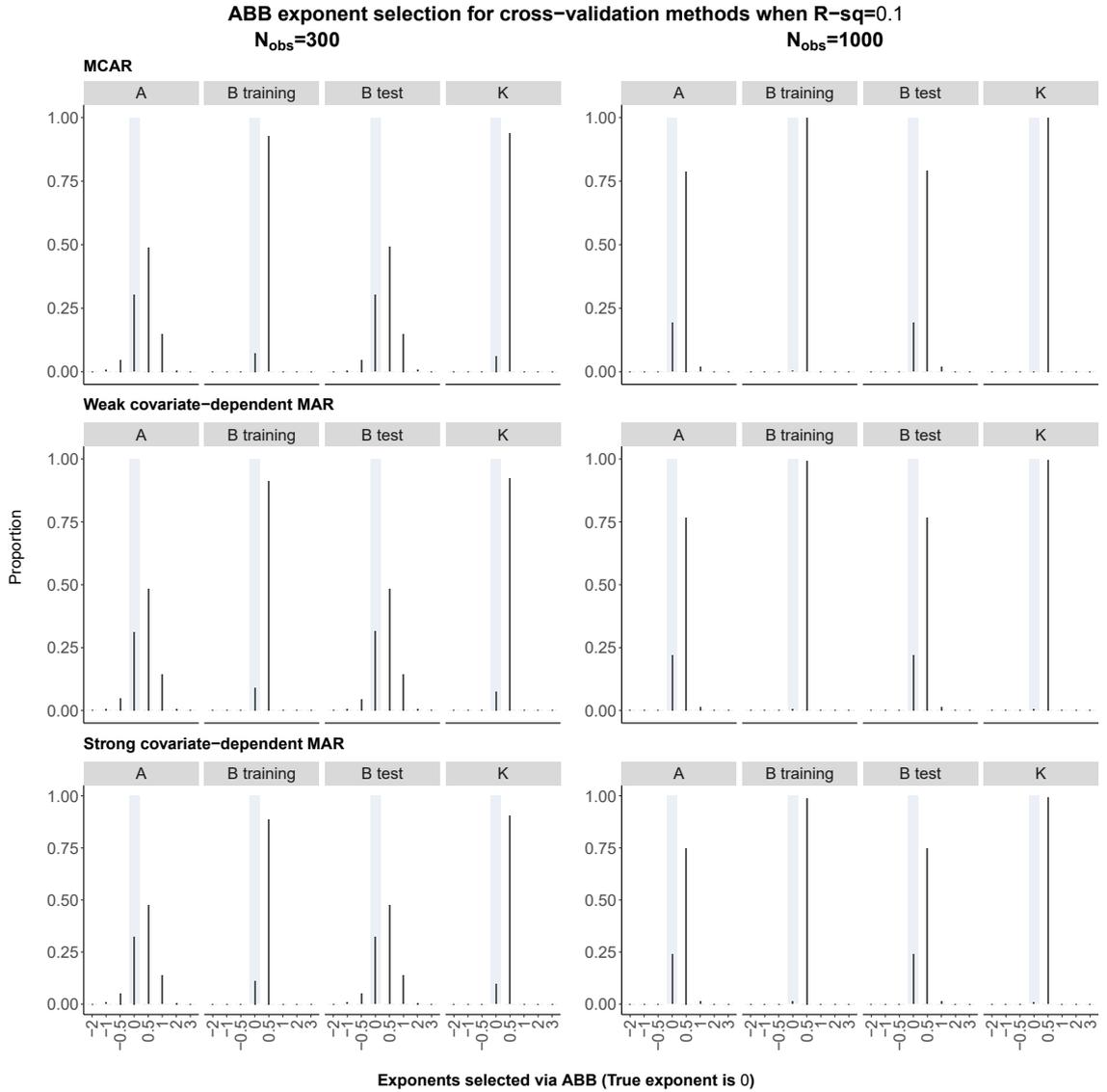


Figure 9.2: The proportion of times an exponent was selected to impute missing values ($M = 5$) using the ABB exponent selection when 25% of values are missing in X_1 . The results are for the scenario where data are MCAR or covariate-dependent MAR, the true exponent E is 0, and $R^2 = 0.1$. The pale blue bar highlights the underlying ‘true’ exponent, while the black bars represent the proportion of times across the 2000 repetitions, K folds and M imputed datasets, that each exponent was selected in order to impute X_1^E . CC (complete-case); methods A-K are described in Table 2.3.

9.5 Selection of exponents from the fractional polynomial selection algorithm

In this section I will assess the selection of exponents from the fractional polynomial selection procedure in the prediction model. For cross-validation, this involves assessing the selection of an exponent for X_1 when training a prediction model on the $k - 1$ training folds. For the 0.632 bootstrap, this involves assessing exponent selection when using all of the data to train a prediction model (in order to estimate apparent performance), and also when using the bootstrap sample observations to train a model (in order to estimate test performance). The frequency with which the true underlying exponent (either 2, 0 or -2) is selected will be assessed across the various data-generating scenarios. This section will present a few selected figures which are generally representative of the majority of the scenarios. All plots are available in Supplementary Plots Section S6.2 for cross-validation and the 0.632 bootstrap algorithms.

The exponent selection will be assessed in two ways. Firstly, how often the correct exponent is selected when the ‘best-fitting’ fractional polynomial is included in the prediction model (i.e. $\alpha_E = 1$). Secondly, how often will the correct exponent be selected when the ‘best-fitting’ fractional polynomial will be compared against the inclusion of the covariate as a linear term in the prediction model (i.e. $\alpha_E = 0.05$). For both values of α_E , the exponent selection will be assessed after the application of an origin-shift transformation as this was shown to improve the estimated MSE for ideal and pragmatic performance.

9.5.1 Selection of exponents when the best-fitting fractional polynomial is selected ($\alpha_E = 1$)

Figure 9.3 presents the proportion of times exponents were selected across the 2000 repetitions, K iterations of cross-validation and M imputed datasets. The results presented are for the scenario where the true exponent is 0, an origin-shift transformation has been applied, $R^2 = 0.1$ and data are MCAR or covariate-dependent MAR.

From Figure 9.3 it can be seen that applying an origin-shift transformation has resulted in the correct exponent, 0, to be correctly selected less than 10% of the time regardless of whether the data were fully-observed, imputed or a complete-case analysis was applied. Exponent -2 is incorrectly selected over 50% of the time when sample size is 300, this increases to over 75% when the sample size is 1000. When the true exponent is -2, it is correctly selected over 90% of the time when data are fully-observed and for all methods used to handle missing data (complete-case analysis and methods A, B and K). When the true exponent is 2, the exponent 3 is selected the majority of the time for methods B or K, when the data are fully-observed and when complete-case analysis is used. When sample size is 300, method A tends to select -2 approximately 25% of the time. When sample size

is 1000, method A tends to select -0.5 or 0 across the majority of the $2000 \times K \times M$ datasets (both are selected less than 25% of the time).

These results are similar for the 0.632 bootstrap methods (plots available in Supplementary Plot Section S6.2). When the true exponent is 0, exponent -2 is most likely to be selected across the $2000 \times B \times M$ datasets. Similarly, when the true exponent is 2, exponent 3 has the highest selection rate. When the true exponent is -2, it has the highest percentage for being correctly selected across all of the potential exponents.

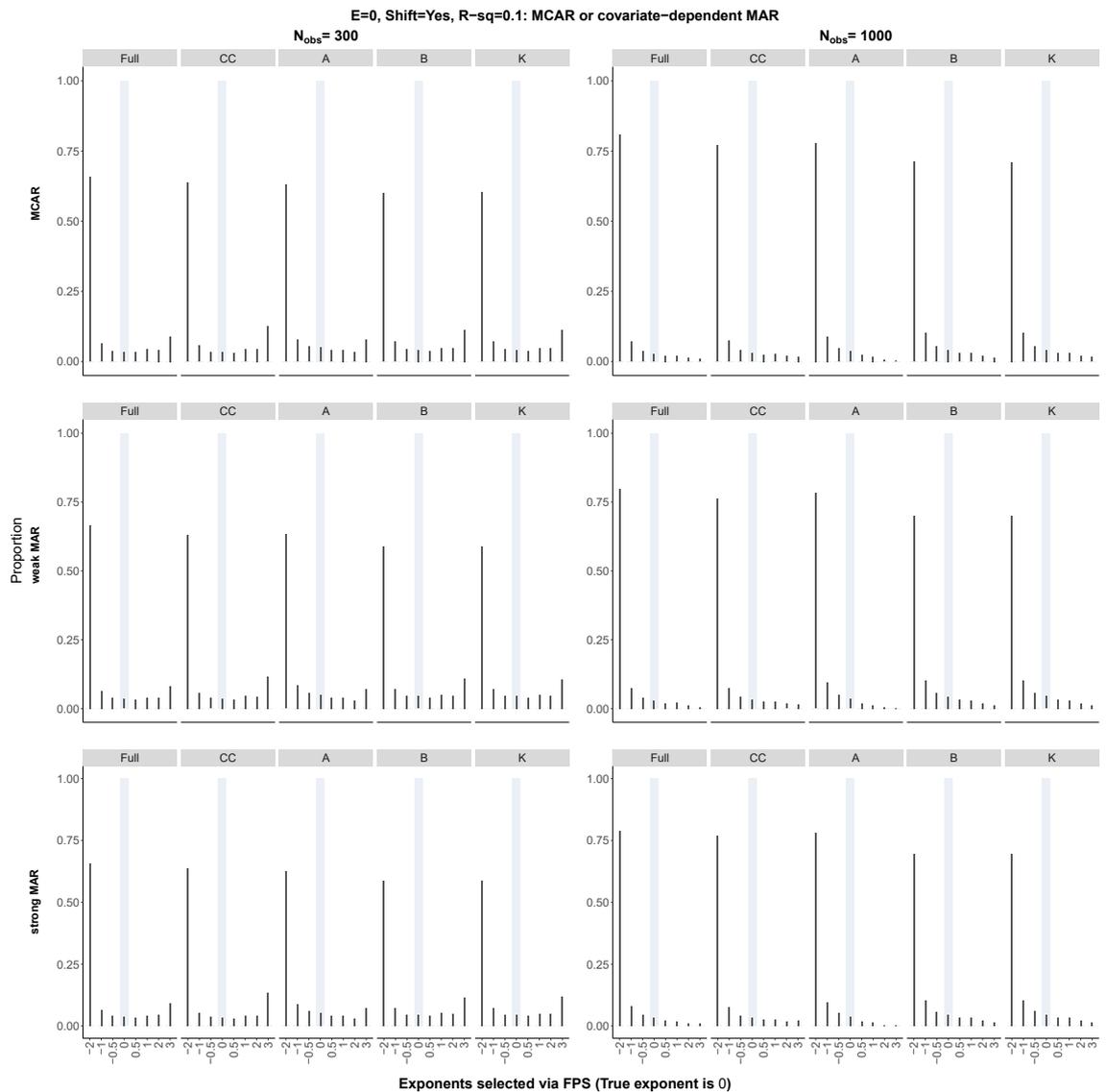


Figure 9.3: The proportion of times an exponent was selected via FPS post-imputing when $\alpha_E = 1$ and an origin-shift transformation was used. The results are for the scenario where data are MCAR or covariate-dependent MAR, the true exponent E is 0 and $R^2 = 0.1$. The pale blue bar highlights the underlying ‘true’ exponent, while the black bars represent the proportion of times across the 2000 repetitions, K folds and M imputed datasets, that each exponent was selected via the FPS algorithm. CC (complete-case); methods A-K are described in Table 2.3.

9.5.2 Selection of exponents when the best-fitting fractional polynomial is compared against the default inclusion as a linear covariate ($\alpha_E = 0.05$)

Figure 9.4 presents results for the scenario where the true exponent is 0, $\alpha_E = 0.05$, an origin-shift transformation has been applied, $R^2 = 0.1$ and data are MCAR or covariate-dependent MAR. Selection of exponent 1 still dominates for all methods, except method A. When sample size is 300, exponent 1 is selected across 90% of the datasets for fully-observed data, the complete-case analysis and methods B or K. If exponent 1 is not selected, the next common exponent is -2. All other exponents are rarely selected. With increasing sample size, the percentage of occasions on which the selected exponent value is 1 decreases to over 50% and the proportion of times exponent -2 is selected increases. For method A when sample size is 300, exponent 1 is selected approximately 70% of the time and exponent -2 is selected approximately 25% of the time. With increasing sample size, exponent -2 tends to be selected more frequently than exponent 1.

When the true exponent is -2 and sample size is 300, -2 has the highest proportion when data are fully-observed, when the complete-case analysis is applied or for method A, with exponent 1 having the next highest proportion. For methods B and K exponent 1 has a higher proportion than -2. However, with increasing sample size -2 dominates, being selected over 90% of the time for all methods (Supplementary Plots Section S6.2). When the true exponent is 2 and sample size is 300, the proportion of times exponent 1 is selected to at least 75%. However, with increasing sample size the proportion of times exponent 1 is selected tends to decrease and the proportion tends to increase for exponent 3. The true exponent 2 is not selected by any method, even when data are fully-observed.

Similarly, for the 0.632 bootstrap (plots available in Supplementary Plot Section S6.2) when the true exponent is 0 and the sample size is 300, an exponent of 1 is primarily selected in at least 90% of the $2000B^*M$ datasets for both *BS-then-MI* and *MI-then-BS* across the various missing data scenarios. With increasing sample size, the percentage of times 1 is selected decreases to 46% when data are fully-observed and approximately 55% for the missing data method, in favour of selecting exponent -2. When the true exponent is 2 and sample size is 300, exponent 1 is most likely to be selected across all missing data methods. However, for sample size 1000 exponent 3 is most likely to be selected when data are fully-observed or a complete-case analysis is used (74% and 55%, respectively). For *BS-then-MI* or *MI-then-BS*, exponent 3 is selected in approximately 27% of the $2000B^*M$ datasets while exponent 1 is selected in approximately 72% of the datasets. When the true exponent is -2 and sample size is 300, it is correctly selected in 64% of the datasets when data are fully-observed and 48% when a complete-case analysis is used. For the MI methods, -2 and 1 are selected across approximately 50% of the datasets. When sample size is increased to 1000, exponent -2 is correctly selected in 99% of the datasets.

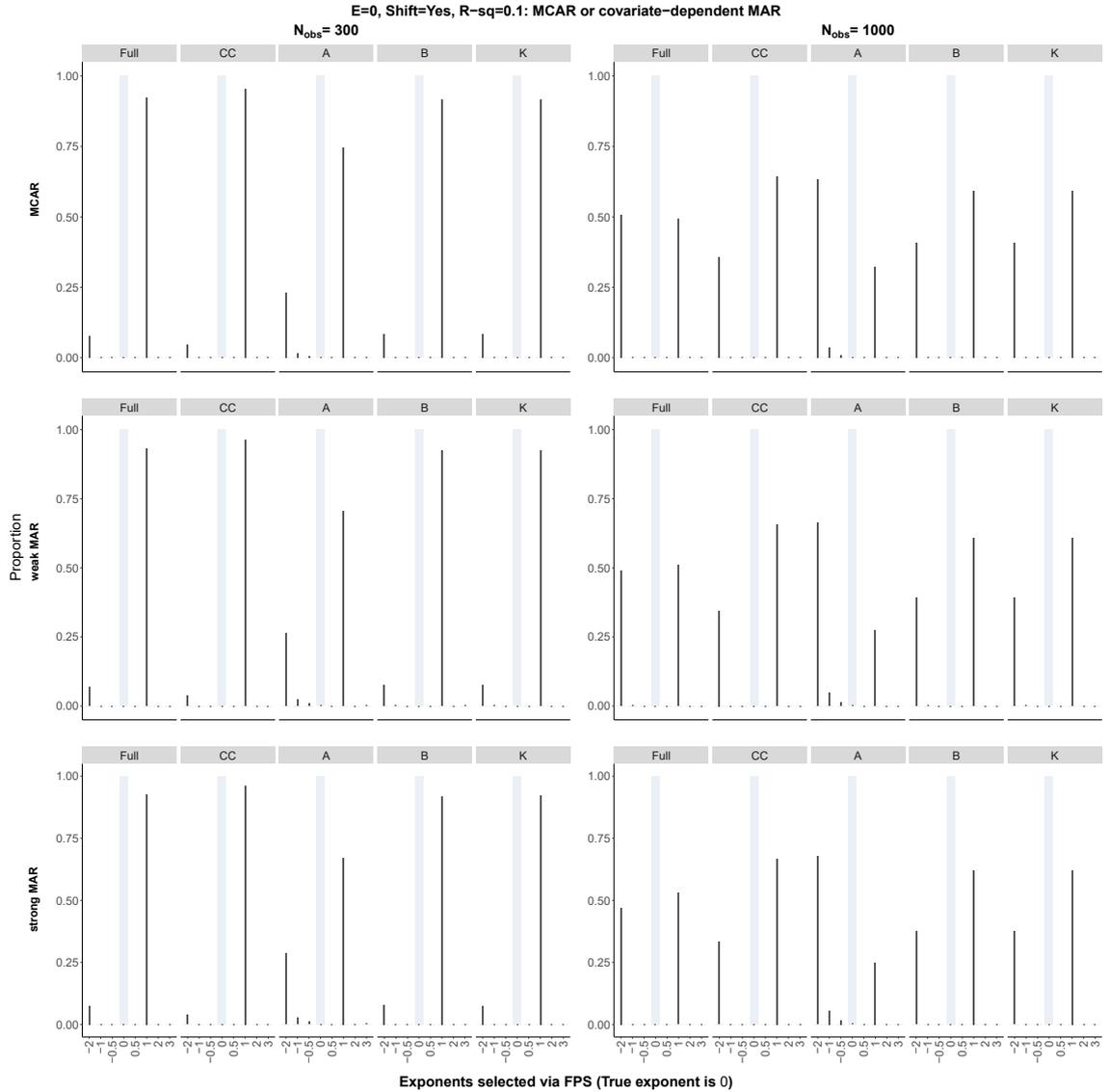


Figure 9.4: The proportion of times an exponent was selected via FPS post-imputing when $\alpha_E = 0.05$ and an origin-shift transformation was used. The results are for the scenario where data are MCAR or covariate-dependent MAR, the true exponent E is 0 and $R^2 = 0.1$. The grey bar highlights the underlying ‘true’ exponent, while the black bars represent the proportion of times across the 2000 repetitions, K folds and M imputed datasets, that each exponent was selected via the FPS algorithm. CC (complete-case); methods A-K are described in Table 2.3.

9.6 Results for the Multivariable fractional polynomial (MFP) algorithm

In sections 9.3 to 9.5, I assessed how well the proposed methods (for handling internal validation with fractional polynomial selection in the presence of missing data) performed. This involved assessment of the estimated MSE when compared to the MSE obtained when data are fully-observed, as well as examining the exponents selected via ABB when imputing missing values and the exponents selected by the FPS algorithm.

The FPS algorithm focuses solely on exponent selection while the MFP allows for both exponent and covariate selection. For the MFP algorithm, the selection of exponents via ABB or FPS and the performance of the methods are comparable to the results from FPS (sections 9.3 to 9.5). MFP results are presented using graphs which are available in Supplementary plots section S8 for ABB and FP exponent selection. Graphs are available for the comparison of the estimated MSE with the MSE when data are fully-observed for the MFP procedure in Supplementary plots section S7.

In this section, I will therefore evaluate the covariate selection process of the MFP algorithm. Recall from Chapter 8 that two values are used for the coefficient for covariate X_2 ($\beta_2 = 0$ or 1) in the prediction model. When $\beta_2 = 0$, covariate X_2 should not be selected into the prediction model and when $\beta_2 = 1$, it should be included in the prediction model. Recall from Table 9.1 that two values were selected for α_β (significance level for exponent selection). When $\alpha_\beta = 1$ all covariates are forced into the model, this is equivalent to the previous results for FPS (which focused solely on exponent selection). When $\alpha_\beta = 0.05$ each covariate is considered for inclusion into the model based on a hypothesis test with significance level 0.05 for $\beta_p = 0$.

9.6.1 Covariate selection of X_1 when using the MFP algorithm

In this section, I will assess the covariate selection process of the MFP algorithm for covariate X_1 . Figure 9.5 presents the cross-validation results for covariate selection of X_1 when using the MFP algorithm. The figure displays results for the scenario when the true exponent of X_1 is 0, an origin-shift transformation has been used, $R^2 = 0.1$, data are MCAR or covariate-dependent MAR and α_E (significance level for exponent selection of X_1) is 0.05. These results are, generally, representative of the varying data-generating scenarios. All covariate selection graphs are available in Supplementary plot section S9.

When the sample size is 300 and the true exponent is 0, covariate X_1 is selected into the prediction model across approximately 75% of the 2000 repetition's K fold iterations of cross-validation when data are fully-observed. When a complete-case analysis was used, this decreased to approximately 61%. For the methods which involved MI (methods A, B

and K), covariate X_1 was selected in at least 70% of the $2000 * K * M$ datasets.

These results are similar when the true exponent is -2. For all missing data scenarios when the true exponent is 2 and sample size is 300, X_1 has a higher selection percentage of at least 75% for the complete-case analysis, method A, method B and method K. The selection of X_1 into the prediction model is similar for the 0.632 algorithm (graphs available in Supplementary plot section S9). Changing the value of α_E from 0.05 to 1 had little effect on the results.

For both the 0.632 algorithm and cross-validation when sample size is increased to 1000, selection of X_1 increases to at least 99%.

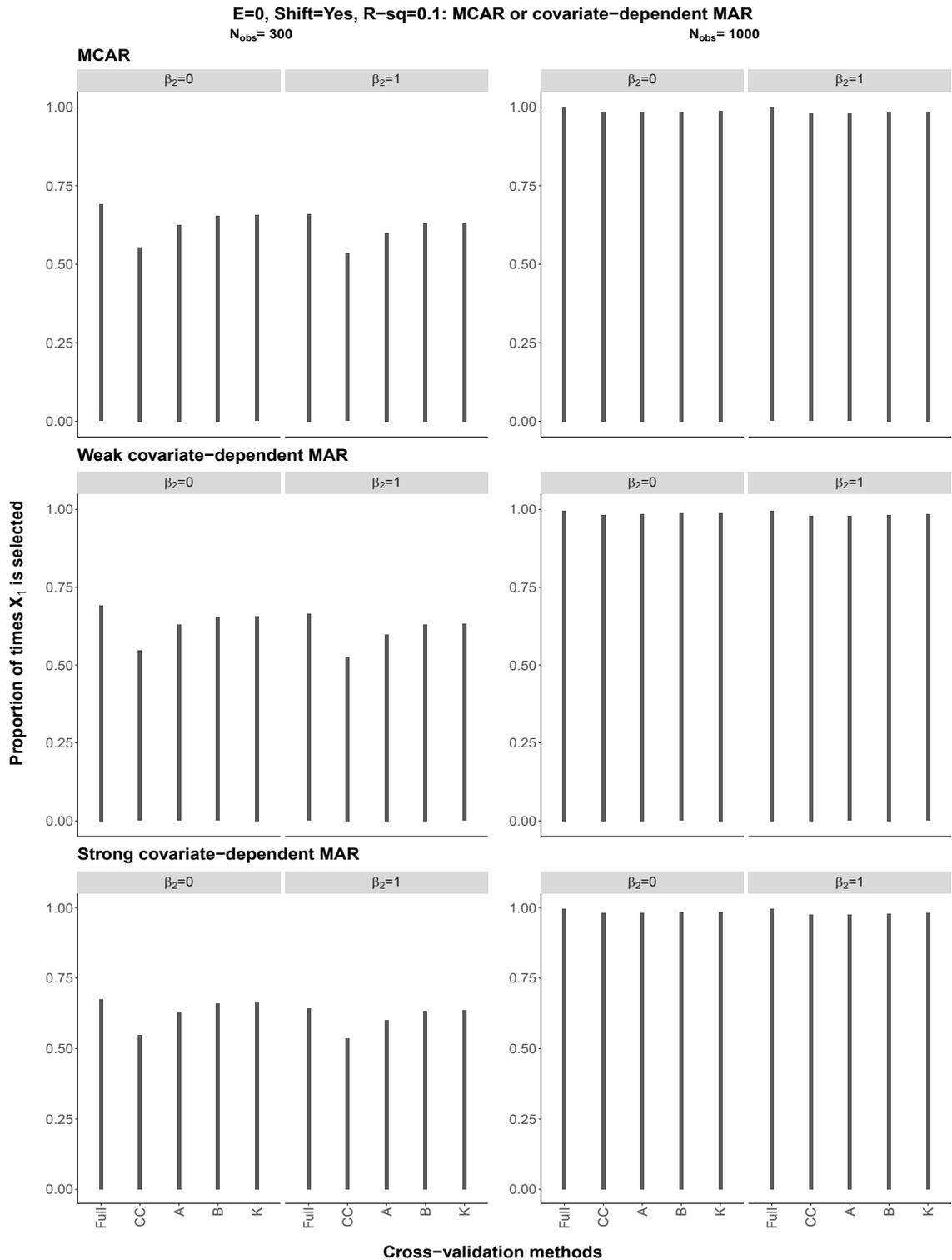


Figure 9.5: The proportion of times covariate X_1 is selected for inclusion to the prediction model using the MFP algorithm. The results presented are for the scenario where $\alpha_E = 0.05$ and an origin-shift transformation was used. The results are for the scenario where data are MCAR or covariate-dependent MAR, the true exponent E is 0 and $R^2 = 0.1$. The black bars represent the proportion of times across the 2000 repetitions, K folds and M imputed datasets, that X_1^E was selected into a prediction model when the parameter for covariate X_2 , β_2 is 0 or 1. CC (complete-case); methods A-K are described in Table 2.3.

9.6.2 Covariate selection of X_2 when using the MFP algorithm

Figure 9.6 presents the cross-validation results for covariate selection of X_2 when using the MFP algorithm. The figure displays results for the scenario when the true exponent of X_2 is 0, an origin-shift transformation has been used, $R^2 = 0.1$, data are MCAR or covariate-dependent MAR and α_E (significance level for exponent selection of X_1) is 0.05. These results are, generally, representative of the varying scenarios when the exponent is 0. I will also discuss results for when the true exponent is -2 and 2, although the relevant graphs will not be presented here. All covariate selection graphs are available in Supplementary plot section S9.

For the various missing data mechanisms when the true exponent is 0, sample size is 300 and $\beta_2 = 0$, covariate X_2 tends to be selected for inclusion in the prediction model in, at most, 10% of the datasets for the various cross-validation methods. With increasing sample size, the inclusion of X_2 into the prediction model decreases to at most 6% across the $2000 \times K \times M$ datasets. For the same scenarios when $\beta_2 = 1$ and sample size is 300, X_2 is selected across at least 86% of the $2000 \times K$ datasets when a complete-case analysis is applied and in at least 92% of the $2000 \times K \times M$ datasets for methods A, B and K. These percentages increase to at least 99% when sample size is 1000. These results are similar for the 0.632 algorithm (when data are fully-observed or a complete-case analysis, *BS-then-MI* or *MI-then-BS* are applied).

For the various missing data mechanisms when the true exponent is 2, sample size is 300 and $\beta_2 = 0$, covariate X_2 tends to be selected for inclusion in the prediction model in, at most, 11% of the datasets for the various cross-validation methods. With increasing sample size, the inclusion of X_2 into the prediction model decreases to at most 9% across the $2000 \times K \times M$ datasets. For the same scenarios when $\beta_2 = 1$ and sample size is 300, X_2 is selected across at least 40% of the $2000 \times K$ datasets when a complete-case analysis is applied and in at least 53% of the $2000 \times K \times M$ datasets for methods B and K and at least 67% for method A. These percentages increase to at least 86% (for complete-case analysis) and 94% (for methods A, B, K or when data are fully-observed) when sample size is 1000. Again, these results are similar for the 0.632 bootstrap algorithm.

For the various missing data mechanisms when the true exponent is -2, sample size is 300 and $\beta_2 = 0$, covariate X_2 tends to be selected for inclusion in the prediction model in, at most, 9% of the datasets for the various cross-validation methods. With increasing sample size, the inclusion of X_2 into the prediction model decreases to at most 8% across the $2000 \times K \times M$ datasets. For the same scenarios when $\beta_2 = 1$ and sample size is 300, X_2 is selected across at most 8% of the $2000 \times K$ datasets when a complete-case analysis is applied and in at least 10% of the $2000 \times K \times M$ datasets for methods A, B and K. These

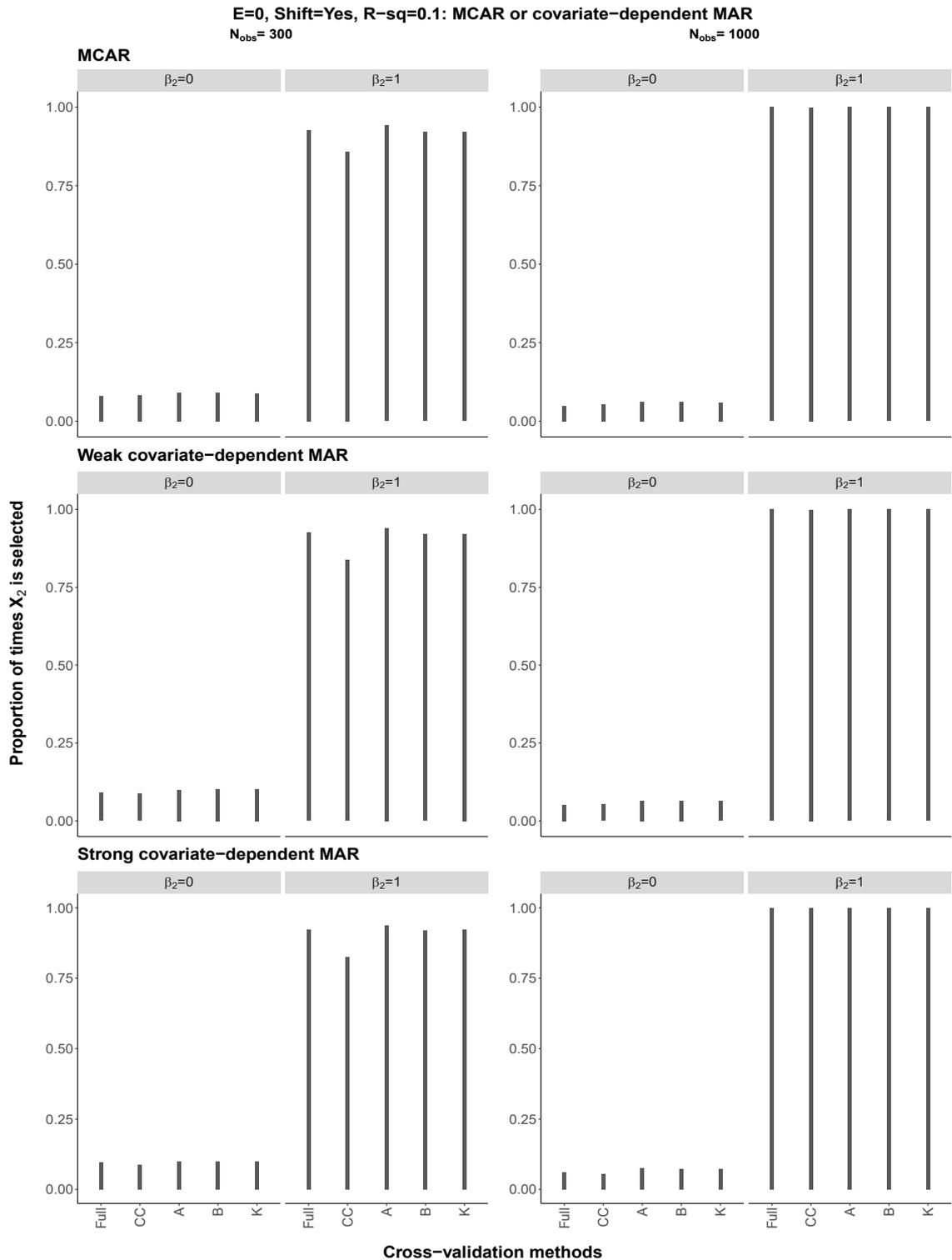


Figure 9.6: The proportion of times covariate X_2 is selected for inclusion to the prediction model using the MFP algorithm. The results presented are for the scenario where $\alpha_E = 0.05$ and an origin-shift transformation was used. The results are for the scenario where data are MCAR or covariate-dependent MAR, the true exponent E is 0 and $R^2 = 0.1$. The black bars represent the proportion of times across the 2000 repetitions, K folds and M imputed datasets, that X_2 was selected into a prediction model when its parameter, β_2 is 0 or 1. CC (complete-case); methods A-K are described in Table 2.3.

percentages increase to at most 15% (for complete-case analysis) and 20% (for methods A, B, K or when data are fully-observed) when sample size is 1000. The inclusion of X_2 into the prediction model is similarly low for the 0.632 bootstrap algorithm when the true exponent is -2.

Overall, the choice of exponent for covariate X_1 impacts covariate selection of X_2 into the prediction model.

9.7 Discussion

The aim of the simulation study discussed in this chapter was to identify how well the various proposed methods performed when adapted to handle flexible transformation of continuous covariates and covariate selection using fractional polynomials. The results were assessed for both cross-validation and the 0.632 bootstrap across varying data-generating scenarios while also investigating the impact of analysis decisions (the choice of α_E or using an origin-shift transformation).

Overall, it was found that the results could be highly unstable when an origin-shift transformation was not used. This led to large differences between the estimated MSE from the proposed methods and the MSE when data are fully-observed, with large Monte Carlo confidence intervals around point estimates. This did not improve with increasing sample size. However, the majority of the methods became increasingly stable with increasing sample size and when an origin-shift transformation was used to remove the impact of low values close to zero on the prediction model. However, while the use of an origin-shift transformation improved the estimated MSE when low values were present in the data, the MSE was affected by large imputed values. This could perhaps be improved by either standardising covariates or applying a transformation which will remove the influence of large values.

While the use of an origin-shift transformation improved the estimated MSE, it could result in the incorrect exponent being selected when fitting a prediction model. However, the focus here is on prediction modelling (as opposed to an exploratory or causal analysis) and I argue that the correct exponent does not need to be selected. It is the exponent which produces the ‘best’ prediction model that is of interest. The other investigated analysis parameter was α_E , which controlled the significance level for the inclusion of the best-fitting fractional polynomial when compared to a default inclusion as a linear covariate. The performance of the estimated MSE, in relation to the fully-observed MSE, was similar with an increased value of α_E . A large value of α_E should be avoided when using fractional polynomials to develop prediction models. This is due to an increased risk of over-fitting the prediction model to the data when selecting the best exponent where inclu-

sion as a linear term would have fitted a prediction model which performed equally as well.

Covariate selection was assessed for all methods when using the MFP procedure and the parameter of covariate X_2 changed from 1 to 0 to assess whether X_2 was correctly selected or not when training a prediction model during internal validation. When the exponent was 0 or 2 and the sample size was sufficient to fit a stable model, covariate X_2 was correctly selected or rejected from the prediction model depending on whether β_2 was 0 or 1. However, when the exponent was -2, X_2 was rarely included in the prediction model, even when $\beta_2 = 1$, while X_1 was nearly always included into the prediction model. This is potentially due to the large variation in the outcome rendering it difficult to select X_2 in the model, as when the R^2 increased from 0.1 to 0.3, the proportion increased for X_2 being correctly selected into the prediction model.

For all values of the true exponent, when selecting an exponent via ABB to impute the missing values, 0.5 tended to be the dominant exponent selected, followed by 0 which is potentially due to the way the simulated data were initially generated (X_1 was simulated from a Normal distribution before it was transformed to generate the outcome Y). Arguably, we are looking for the best predicted value, rather than the best imputed value. Therefore, the ‘incorrect’ exponent being commonly selected is not disadvantageous, provided the predicted value, based on the imputed value, is improved.

This chapter concludes the final simulation study investigating how to internally validate in the presence of missing data. For cross-validation, results were presented for method A (impute each fold separately), method B (impute the $k - 1$ folds together, impute the k^{th} test fold separately) and method K (impute first, then cross-validate). In Chapters 4 and 5, I concluded that *MI-then-CV* methods, such as method K, were not advisable due to data leakage through the imputation process and that methods A and B should be preferred.

Due to the results from this chapter, I conclude that method A is the preferred *CV-then-MI* method. For all sample sizes, method A tended to produce more stable results (i.e. a smaller magnitude when compared to the MSE when data are fully-observed, with less Monte Carlo variation across the 2000 repetitions) than method B. In Chapter 4, I concluded that imputing each fold individually may result in more variable imputed values than imputing $k - 1$ folds together. This in turn could potentially lead to a more robust prediction model. In this chapter, imputing each fold separately also resulted in a more diverse selection of exponents via ABB, in order to impute the missing X_1 values.

For the 0.632 bootstrap algorithm, results were presented for *BS-then-MI* and *MI-then-BS*. I previously concluded in chapters 6 and C that *MI-then-BS* was not recommended

due to data leakage. In this chapter, *BS-then-MI* performed well and tended to have slightly more stable results than cross-validation, when sample size was small.

In this simulation study, I investigated a FP transformation for one covariate, X_1 , when in reality a FP transformation may also have been investigated for another covariate such as X_2 . In more complex and realistic data, it is not uncommon for more than one covariate to be partially-observed. The methods assessed in this chapter can easily be extended to imputing multiple partially-observed covariates (using the likes of MICE described in Section 1.6.2) and allowing for more than one continuous covariate to have a fractional polynomial.

Overall, a large sample size is required when using internal validation, and applying covariate selection or transforming continuous covariates. *Validate-then-MI* methods are preferred for combining MI, internal validation and fractional polynomials.

10 Rotterdam breast cancer data

10.1 Introduction

In this chapter I will apply the proposed methods from Chapter 7 to illustrate how they can be implemented in practice. The Rotterdam dataset is taken from the Rotterdam tumour bank and concerns 2,982 patients with primary breast cancer. This dataset is publicly available and was used in [11] to demonstrate an extension of the MFP algorithm to handle time-varying effects. The statistical analysis in this chapter would not be conducted in practice, as the Rotterdam data available for download does not contain missing values. Therefore, missing values will be induced in the dataset.

Using this dataset will demonstrate how the proposed methods can be adapted for a survival setting and explore how the methods perform when a more complex and real data structure is used. The increased complexity comes from multiple covariates which should be considered for inclusion into the prediction model and several continuous covariates which may need to be transformed. Previously, the methods were assessed in ‘cleaner’ simulated data which included an outcome and two covariates. The simulation studies had the benefit of knowing which covariates should be included in both the prediction and imputation models, which is not the case in practice.

10.2 Background to the Rotterdam data

The Rotterdam data, from the Rotterdam tumour bank, contains patients who were receiving treatment for primary breast cancer between 1978 and 1993. The dataset originally had 3,001 patients, however 11 patients were excluded due to missing information about the number of positive nodes. Eight more were excluded due to non-standard treatment of node negative patients. Therefore, 2,982 patients remained for the analysis [58]. Missing data were present in tumour size (1.1%), tumour grade (26.6%), number of positive lymph nodes (2.2%), progesterone receptor (5.4%) and oestrogen receptor (3.6%) ([58, Table 1]). The missing values were imputed using MICE and it is this imputed dataset which is available to download. The dataset does not include missing indicators so it is impossible to know which patients were imputed previously. For the demonstration of the various methods, I will treat the dataset as fully-observed and introduce missing values under a MAR mechanism in Section 10.5.

The time scale for the analysis is the number of years since surgery for removal of primary tumours. Follow-up time ranged from 1 to 231 months and 1,518 (50.91%) patients died due to breast cancer or had a recurrence of the disease by the end of follow-up. Those who died from other causes were censored, as were those who had not had an event by the end of follow-up, as in [58]. By 8 years approximately 50% of patients had died due

to breast cancer or had a recurrence.

The dataset contains information on the time to relapse or death, age at surgery, menopausal status, size and grade of the tumour, the number of positive nodes, progesterone receptors (PgR), estrogen receptors, hormonal therapy and chemotherapy. While information is also available on the overall survival and metastasis free survival this will not be considered in this analysis. Tumour size has previously been split into two binary dummy covariates [58]. Size 1 for those with a tumour size ≤ 20 mm or size 2 for those with a tumour size ≤ 50 mm. Due to a small proportion of patients having a grade 1 tumour (2%), grade 1 and grade 2 were previously collapsed together. Monotonic transformations were previously chosen for the number of positive nodes ($\text{Enodes}=\exp(-0.12\text{Nodes})$) and progesterone receptor ($\text{Pr}_1=\log(\text{PgR}+1)$).

10.3 Developing a prediction model in the Rotterdam dataset

A proportional hazards Cox model [59] will be used as the prediction model of interest in this chapter. The fitted prediction model will be used to estimate the risk of having a recurrence or dying from breast cancer by the end of follow-up. The performance measure that will be used to assess the performance of the prediction model is the *C*-statistic [33] available in the R package ‘survcomp’ [60].

The same handling of covariates, including creation of dummy covariates for size (size 1 and size 2) and transformation of continuous covariates, will be used here. This includes monotonic transformations for the number of positive nodes and progesterone receptors, and collapsing tumour grade 1 and 2 together.

Following previous analyses [11,58,61], I will focus on age, tumour size 1 and 2, tumour grade, the number of positive nodes, hormonal therapy, chemotherapy and progesterone receptors. As progesterone receptor (Pr_1) contains zero values, I will use the same origin-shift transformation which was used previously in Chapters7to9:

$$0.2 + 0.8 \frac{\text{Pr}_1 - \min(\text{Pr}_1)}{\min(\text{Pr}_1) - \max(\text{Pr}_1)}$$

Previously in Chapters7to9, I assessed the proposed methods while allowing for fractional polynomial transformations of degree 1 (FP1: X^E). A FP1 transformation will be assessed for continuous covariates in the Rotterdam dataset. The MFP algorithm will be used for covariate selection and the flexible transformation of the continuous covariates (enodes, Pr_1 and age). The default significance level of 0.05 will be used for the selection of exponents and covariates into the prediction model.

10.4 Splitting data into a training and external validation dataset

As I am using the Rotterdam data to demonstrate how to apply the proposed methods in practice, I will split the data into two sections, a ‘training and internal validation’ set and an external validation ‘holdout’ set. I will ensure that the training dataset will contain the minimum sample size required to develop a prediction model. The minimum sample size recommended assumes that all observations in the collected data are fully-observed.

10.4.1 Sample size calculation for building a prediction model

Sauerbrei, Royston and Look (2007, [58]) developed a MFP model as part of their MFP Time algorithm. The covariates included in their MFP model are available in Table 2 of [58]. This model was replicated using the available data, and an apparent performance estimate of the C -statistic was estimated as 0.69. The estimated C -statistic which would be used to estimate the minimum sample size, in practice, would be taken from another source. This C -statistic value, which was estimated using all available data, is expected to be optimistic.

Riley et al. (2019) [26] provides formulas to convert the C -statistic to an estimate of the Cox-Snell R^2 ($R_{CS,adj}^2 = 0.1804$). This can be used alongside the R package ‘pmsampsize’ [62] to obtain a minimum sample size for the development of a prediction model based on an anticipated R^2 value and taking into consideration the number of parameters which will be fitted in the prediction model.

With eight covariates potentially being considered for inclusion into the model, a minimum sample size of approximately 360 patients is required to develop a prediction model. As three covariates are continuous and will have a fractional polynomial exponent to estimate, this will account for three additional parameters to be estimated, in addition to the eight coefficient parameters, β . Increasing the number of parameters to estimate to eleven increases the required minimum sample size to 492.

10.4.2 Creating a training and holdout set

The dataset will be sorted into ascending order by year. Of those patients included into the study last, 10% of patients with the outcome and 10% of patients without the outcome will be selected (without replacement) for inclusion to an external holdout set. Taking 10% of observations from those who did and did not have the event will ensure the holdout set has the same proportion of patients with and without the outcome. Using the patients who were last to enter into the study will replicate the scenario of data from a later time period becoming available which can be used to create a ‘temporal’ external validation set (although these patients will potentially have lower follow-up times). The holdout set will include 298 observations, of which 152 will have experienced the outcome. The use of this

holdout set is similar to the generation of a larger validation set (Section 3.6.4) which was used in the previous simulation studies, in order to obtain a target estimate of performance.

The training and validation data will consist of the 2,684 patients not selected for inclusion into the holdout set, of which 1,366 (50.9) will have had the outcome.

10.5 Introducing missing data to covariates

Values will be set as ‘missing’ for tumour grade, enodes and progesterone receptor (Pr_1) under a MAR mechanism. The generated missing data will be non-monotone. Each covariate will have approximately the same proportion of missingness that was reported to be in the original Rotterdam data before it was imputed using MICE (Grade 26.6%; Enodes 2.2%; Pr_1 5.4%, [58, Table 1]). Two missing data scenarios will be considered. Scenario 1 will focus on covariate-dependent MAR and scenario 2 will assess outcome- and covariate-dependent MAR. The outcome, in this setting, will be the event indicator. In both scenarios, the covariates which will be used to induce missingness are age and chemotherapy (both of which are fully-observed). For patient j the probability of covariate X_p being missing is:

$$\pi_{X_p,j} = \frac{\exp(\psi_0 + \psi_A \text{Age}_j + \psi_C \text{Chemo}_j + \psi_E \text{Event}_j)}{1 + \exp(\psi_0 + \psi_A \text{Age}_j + \psi_C \text{Chemo}_j + \psi_E \text{Event}_j)}$$

For the two missing data scenarios, non-zero values of ψ_A and ψ_C were selected to produce a moderately strong MAR mechanism. This strength was calibrated based on an AUC value estimated by regressing the missing indicator of X_p on the covariates used to generate the missing data in the covariate. Values for ψ_0 were then selected to ensure the desired proportion of observations in X_p were set as missing. Table 10.1 shows the finalised ψ parameter values and the AUC of missingness. Values in the external holdout set will be set to missing using the same parameters in Table 10.1.

Table 10.1: Specification of the ψ parameter values to ensure covariate-dependent MAR (scenario 1) and outcome- and covariate-dependent MAR. For each covariate, the strength of the MAR mechanism (AUC) and percentage of observations induced to be missing (%) are given.

| Partially-observed covariate (X_p) | Scenario 1 | | | | | | Scenario 2 | | | | | |
|--|------------|----------|----------|----------|-------|-------|------------|----------|----------|----------|-------|-------|
| | ψ_0 | ψ_A | ψ_C | ψ_E | AUC | % | ψ_0 | ψ_A | ψ_C | ψ_E | AUC | % |
| Grade | -3.92 | 0.05 | 0.06 | 0 | 0.692 | 26.64 | -4.43 | 0.05 | 0.06 | 1 | 0.706 | 26.53 |
| Enodes | -6.85 | 0.05 | 0.06 | 0 | 0.707 | 2.27 | -7.60 | 0.05 | 0.06 | 1 | 0.751 | 2.19 |
| Progesterone | -5.67 | 0.05 | 0.06 | 0 | 0.682 | 5.40 | -6.32 | 0.05 | 0.06 | 1 | 0.74 | 5.44 |

10.6 The imputation model used when multiply imputing

Three covariates are now partially-observed (Enodes, progesterone receptors and tumour grade). To use MI, covariates which should be included in the training and test imputa-

tion models (Section 2.5) must be determined. In addition, the training imputation model and the test imputation model (when estimating ideal performance) must also include the ‘outcome’. In a survival setting, this will involve the inclusion of the Nelson-Aalen estimate of the cumulative hazard and the event/censoring indicator [20].

Table 10.2 presents the covariates which will be included in the imputation model for each partially-observed covariate, in addition to potentially including the ‘outcome’ covariates.

Table 10.2: Covariates which will be included in the imputation models used to impute each of the partially-observed covariates.

| Partially-observed covariate (X_p) | Covariates to be included in imputation model | | | | | | | |
|--|---|--------|------|-----|-------------------|-------------------|-------|----|
| | Grade | Enodes | Pr_1 | Age | Size ₁ | Size ₂ | Chemo | HT |
| Grade | | ✓ | | ✓ | ✓ | ✓ | ✓ | ✓ |
| Enodes | ✓ | | | ✓ | ✓ | ✓ | ✓ | ✓ |
| Pr_1 | ✓ | ✓ | | ✓ | ✓ | | ✓ | ✓ |

Pr_1: progesterone receptor; HT: hormonal therapy

Covariates were included in the imputation model based on being:

1. associated with the covariate X_p
2. included in the final prediction model when using all available training data (Table 10.4)
3. associated with the missing data indicator for each partially-observed covariate.

The covariates selected in steps 1 and 2 will be selected using the fully-observed data to ensure the ‘best-possible’ imputation model will be selected for demonstrating the methods. Therefore, the form of the imputation model (i.e. which covariates are included in the imputation model for each partially-observed covariate) for ideal and pragmatic performance is fixed for all methods using MI in both missing data scenarios.

As several covariates contain missing values, multivariate imputation by chained equations (MICE) will be used here. Enodes and progesterone receptors are heavily skewed and tumour grade is a binary covariate. I shall use predictive mean matching MI [14, p.77-84] to impute each covariate, using ten imputed datasets ($M = 10$) and a donor pool of size 15.

10.7 Evaluating performance when data are fully-observed

When data are fully-observed, a ‘final’ prediction model will be fitted to the training dataset as described in Section 10.3. The performance of this model will then be evaluated using both internal and external validation.

10.7.1 Internal validation

The performance of the prediction model (fitted using all observations in the training dataset) will be evaluated using the training dataset. This performance will be estimated using apparent performance, cross-validation and the 0.632 bootstrap algorithm. All of these algorithms were explicitly detailed for fully-observed data in Section 1.9.

The apparent performance (Section 1.9) will be estimated by evaluating the prediction model in the training dataset it was fitted to. Both cross-validation and the 0.632 will be stratified by the outcome. This ensures that the same proportions of those with or without the outcome are available in each cross-validation fold or bootstrap sample. For cross-validation, the data will be split into 10 folds ($K = 10$) and for the 0.632 bootstrap algorithm, 200 bootstrap samples will be used ($B = 200$) as this is often found to produce stable estimates [23, p.334].

The training dataset will then be bootstrap sampled (with replacement) 200 times in order to estimate 95% confidence intervals for the performance estimates of the apparent performance, cross-validation and 0.632 algorithms. Within each of these bootstrap samples, the apparent performance will be estimated and the cross-validation and 0.632 bootstrap algorithms will be applied. In total, there will be 200 estimates of the C -statistic for the apparent performance, cross-validation and the 0.632 algorithm. For each internal validation algorithm, the lower and upper estimates of the confidence interval will be estimated using the 2.5th and 97.5th percentiles of the 200 bootstrap estimates.

10.7.2 External validation

When data are fully-observed, a prediction model is fitted using all available data in the training set and evaluated in the fully-observed holdout set to get an estimate of the C -statistic.

The training dataset will be bootstrap sampled and in each sample a prediction model will be fitted using the same analysis procedure described in Section 10.3. This prediction model will then be evaluated in the holdout set to get an estimate of the C -statistic. In total 200 bootstrap samples will be taken, and 95% confidence intervals will be estimated using the 2.5th and 97.5th percentiles of the 200 C -statistic values estimated in the holdout set.

10.8 Evaluating performance when data are partially-observed

When data are partially-observed, a missing data method must be applied before a ‘final’ prediction model can be obtained from the training dataset. This missing data method could be applying a complete-case analysis and fitting a prediction model to those patients with fully-observed data. Another approach is to use MI (using the training imputation models described in Section 10.6), to generate M training imputed datasets. A prediction model will be fitted to each of these imputed datasets using the analysis procedure described in Section 10.3. We will have M ‘final’ prediction models.

10.8.1 Internal validation

Both the cross-validation and 0.632 bootstrap algorithms will be demonstrated. Results for cross-validation *CV-then-MI* method A (impute each fold separately) and *MI-then-CV* method K (impute all data twice using a training and test imputation model) will be presented. For the 0.632 bootstrap method, both *BS-then-MI* and *MI-then-BS* algorithms will be used. The bootstrap and cross-validation methods will be stratified to ensure that the same proportions of those who had the outcome or were censored will be in the bootstrap samples or cross-validation folds. Recall that these algorithms, adapted for fractional polynomials, are detailed in Sections 7.4 and 7.5. For cross-validation, the data will be split into 10 folds ($K = 10$) and for the 0.632 bootstrap algorithm, 200 bootstrap samples will be used ($B = 200$) as this is often found to produce stable estimates [23, p.334].

10.8.2 External validation

The complete-case analysis involves using complete-cases to fit a prediction model in the training dataset. This is then evaluated in the complete-cases of the holdout dataset.

If missing values are present in the training dataset and ideal performance is of interest, the missing values in the training dataset are multiply imputed M times and a prediction model is fitted to each of these imputed datasets (as detailed at the start of Section 10.8 above). Each of these M ‘final’ prediction models are evaluated in the fully-observed holdout set. An overall performance measure is estimated by using Rubin’s first rule to average across the M estimates of performance.

Similarly for pragmatic performance M ‘final’ prediction models are obtained, as detailed at the start of Section 10.8 above. However, pragmatic performance anticipates a scenario where future observations will be partially-observed and how this data will be handled needs to be considered. The missing values in the holdout set are imputed using a test imputation model (this includes the covariates from Table 10.2 for each covariate to be imputed but excludes the ‘outcome’ covariates: the event indicator and Nelson-Aalen

estimate). Each of the M ‘final’ prediction models are evaluated in the test imputed datasets of the holdout set. For each prediction model, Rubin’s first rule is used to get an overall performance estimate across the M test imputed datasets. Each of the M prediction models will have an overall performance estimate from the holdout set, these M estimates are then averaged using Rubin’s first rule to get a final performance estimate.

10.9 Results

In this section, I will briefly present the baseline characteristics of the patients selected into the training and holdout datasets. I will then present results from the application of the MFP procedure (for fractional polynomial exponent selection and covariate selection) when data are fully-observed. Finally, I will present the results from the application of the proposed methods for cross-validation and the 0.632 bootstrap algorithms in the presence of missing data.

10.9.1 Baseline characteristics of the training and holdout datasets

Baseline characteristics for the covariates can be found in Table10.3. For the covariates included in the analysis, age, enodes and progesterone are continuous and will need to be assessed for an appropriate functional form.

The training dataset contains 2,684 patients who entered the study between 1978 to 1992, of which 1,366 (50.9%) had the outcome. The ‘temporal’ external validation holdout set contains 298 patients who entered into the study between 1992 and 1993, of which 152 (51.0%) had the outcome. Those in the holdout set tended to have smaller relapse free intervals and were more likely to have received hormonal therapy than those in the training set.

Table 10.3: Baseline Characteristics stratified by those who had the event or were censored within 231 months of follow-up

| Covariate | Training data ($n = 2,684$) | Holdout set ($n = 298$) |
|--|----------------------------------|------------------------------|
| Entry into study (year) | | |
| Min-Max | 1978-1992 | 1992-1993 |
| Relapse Free interval (months) | | |
| Mean (SD) | 71.0 (46.9) | 50.1 (29.7) |
| Median (quartiles) | 65.8 (27.4, 108.3) | 46.5 (21.9,80.5) |
| Age at surgery (years) | | |
| Mean (SD) | 55.2 (13.0) | 59.9 (12.4) |
| Median (quartiles) | 55.0 (45.0, 66.0) | 52.0 (44.0, 62.75) |
| Enodes* | | |
| Mean (SD) | 0.8 (0.3) | 0.7 (0.3) |
| Median (quartiles) | 0.9 (0.6, 1.0) | 0.9 (0.5, 1.0) |
| Progesterone Receptors (fmol/l)* | | |
| Mean (SD) | 3.5 (2.2) | 2.9 (2.3) |
| Median (quartiles) | 3.8 (1.8, 5.3) | 3.1 (0, 4.9) |
| Tumour size | | |
| $\leq 20\text{mm}$ (%) | 1,242 (46.3) | 145 (48.7) |
| $>20\text{-}50\text{mm}$ (%) | 1,186 (44.2) | 105 (35.2) |
| $>50\text{mm}$ (%) | 256 (9.5) | 48 (16.1) |
| Hormonal Therapy | | |
| No (%) | 2,442 (91.0) | 201 (67.4) |
| Yes (%) | 242 (9.0) | 97 (32.6) |
| Chemotherapy | | |
| No (%) | 2,179 (81.2) | 223 (74.8) |
| Yes (%) | 505 (18.8) | 75 (25.2) |
| Menopausal Status | | |
| Pre (%) | 1,170 (43.6) | 142 (47.7) |
| Post (%) | 1,514 (56.4) | 156 (52.3) |
| Differentiation grade | | |
| 1, 2 (%) | 524 (26.6) | 56 (24.7) |
| 3 (%) | 1,445 (73.4) | 171 (75.3) |
| Event: Death or recurrence of breast cancer | | |
| No (%) | 1,318 (49.1) | 146 (49.0) |
| Yes (%) | 1,366 (50.9) | 152 (51.0) |

* Covariates have had their monotonic transformation applied (Section 10.2)

SD: standard deviation; quartiles: 25th and 75th percentiles

10.9.2 Training a prediction model when the data are fully-observed

In this section, I will present estimates of the C -statistic from fitting a prediction model to the fully-observed training dataset, found in Table10.4.

The covariates which were selected into the prediction model when data are fully-observed were Enodes (with a squared transformation), age (included as a linear term), grade, chemotherapy, hormonal therapy and Size 1.

Table 10.4: Results from the MFP algorithm on the Rotterdam dataset when data are fully-observed ($N = 2,684$)

| Covariates | Exponent | $\hat{\beta}$ |
|-------------------|----------|---------------|
| Enodes | 2 | -1.71 |
| Age | 1 | -1.36 |
| Grade | - | 0.36 |
| Chemotherapy | - | -0.49 |
| Hormonal therapy | - | -0.51 |
| Size 1 | - | -0.29 |
| Size 2 | - | - |
| Menopausal Status | - | - |
| Progesterone | - | - |

| Validation | C -statistic (95% CI) ¹ |
|----------------------|--------------------------------------|
| Apparent performance | 0.6888 (0.6780, 0.7017) |
| Cross-validation | 0.6792 (0.6664, 0.6931) |
| 0.632 bootstrap | 0.6871 (0.6624, 0.6965) |
| External validation | 0.6873 (0.6792, 0.7001) |

¹ 95% confidence interval based on 200 bootstrap samples

When the data are fully-observed, the apparent performance estimate of the C -statistic is 0.6888. The estimates of performance when using cross-validation and the 0.632 bootstrap are 0.6792 and 0.6871, respectively. When the prediction model in Table10.4 is evaluated in the fully-observed external holdout dataset (with patients recruited in 1992 and 1993), the C -statistic is estimated as 0.6873.

10.9.3 Applying the proposed methods to the partially-observed Rotterdam dataset

The results from applying the various proposed methods to the partially-observed Rotterdam data are presented in Table10.5. Recall that in scenario 1 the missing values of the training dataset are covariate-dependent MAR. For scenario 2, the missing values of the

training dataset are outcome- and covariate-dependent MAR. A complete-case analysis of the training data reduces the sample size from 2,684 to 1,849 (with 810 events).

Table 10.5: The estimated C -statistic performance when using the proposed methods on the Rotterdam data. Internal validation results are based on a sample size of 2,684 and the external holdout set contained 298 observations.

| Methods | Estimand | Scenario 1 | Scenario 2 |
|----------------------------|-----------|--------------------------|------------|
| <i>Internal validation</i> | | | |
| Cross-validation | | | |
| Fully-observed | | 0.6792 (0.6664, 0.6931)* | |
| Complete-case | | 0.6818 | 0.6922 |
| <i>CV-then-MI</i> | Ideal | 0.6756 | 0.6791 |
| | Pragmatic | 0.6750 | 0.6759 |
| <i>MI-then-CV</i> | Ideal | 0.6900 | 0.6886 |
| | Pragmatic | 0.6805 | 0.6853 |
| The 0.632 bootstrap | | | |
| Fully-observed | | 0.6871 (0.6624, 0.6965)* | |
| Complete-case | | 0.6812 | 0.6919 |
| <i>BS-then-MI</i> | Ideal | 0.6861 | 0.6855 |
| | Pragmatic | 0.6798 | 0.6816 |
| <i>MI-then-BS</i> | Ideal | 0.6875 | 0.6858 |
| | Pragmatic | 0.6784 | 0.6830 |
| <i>External validation</i> | | | |
| Methods | Estimand | Scenario 1 | Scenario 2 |
| Fully-observed | | 0.6873 (0.6792, 0.7001)* | |
| Complete-case | | 0.7044 | 0.7109 |
| Multiple imputation | Ideal | 0.6915 | 0.6872 |
| | Pragmatic | 0.6824 | 0.6814 |

* 200 bootstrap samples used to estimate the 95% confidence intervals.

These are the same estimates provided in Table 10.4.

Overall, the estimates of the C -statistic, \hat{C} , are all very similar for the various internal and external validation used. When using a complete-case analysis with cross-validation, the estimated C -statistics for missing data scenario 1 and 2 (0.6816 and 0.6922) tends to be slightly larger than the point estimate 0.6792, when data are fully-observed. Similarly, the ideal and pragmatic performance of *MI-then-CV* tends to be higher than the fully-observed estimate of the C -statistic. The ideal and pragmatic performance *CV-then-MI* tends to lower than the fully-observed estimate of the C -statistic. However, the point estimates from applying the various missing data methods are all contained within the

95% confidence interval for the fully-observed estimate.

When applying the complete-case analysis with the 0.632 bootstrap algorithm, \hat{C}_{CC} is slightly lower than the fully-observed estimate for scenario 1 (covariate-dependent MAR) but tends to be larger in scenario 2 (outcome- and covariate-dependent MAR). Similarly, the estimated ideal performance of *MI-then-BS* is larger than the fully-observed C -statistic estimate for missing data scenario 1. By comparison for *BS-then-MI*, the ideal and pragmatic performance estimate of \hat{C}_{BS-MI} tends to underestimate the fully-observed estimate in both missing data scenarios. However, the estimated \hat{C} from applying the various missing data methods are all contained within the 95% confidence interval for the fully-observed estimate.

The fully-observed estimates of the C -statistic for cross-validation (0.6792) and the 0.632 bootstrap (0.6871) underestimate the external validation estimate when data are fully-observed (in the training and external holdout datasets) with a C -statistic estimate of 0.6873 (95%CI: 0.6792, 0.7109). Applying a complete-case analysis to the missing values in the holdout set leads to over-optimistic estimates of the C -statistic for both missing data scenarios; 0.7044 and 0.7109 for scenarios 1 and 2, respectively. The estimated ideal and pragmatic performance of the holdout set, when using MI to impute the training set, are similar to the fully-observed estimate of 0.6873.

Overall, the complete-case analysis tends to over-estimate the fully-observed estimates of the C -statistic for internal and external validation. *MI-then-Validate* tends to be slightly over-optimistic but there is no evidence to suggest that there is a difference between the methods.

10.10 Discussion

In this Chapter I presented the application of the proposed methods from Chapters 7 to 9 in the Rotterdam dataset. This demonstrates that the proposed methods can be applied in practice to a more complex dataset (compared to the simpler scenarios used in the simulation studies) and also illustrates their extension to a survival setting.

Confidence intervals were obtained for the internal and external C -statistic estimates by replicating each procedure in a bootstrap sample. While it is possible to obtain confidence intervals for the C -statistics for *Validate-then-MI* or *MI-then-Validate*, this was too computationally intensive. For example, obtaining confidence intervals for the fully-observed C -statistic when using the 0.632 algorithm took several hours on my personal laptop. Bootstrapping to obtain confidence intervals when also using MI within cross-validation or the 0.632 algorithm may be difficult for researchers in practice, particularly on large

datasets which are often used in prediction.

Overall, I found that both *Validate-then-MI* and *MI-then-Validate* performed well when compared to the fully-observed estimates from internal and external validation. While *MI-then-Validate* tended to be slightly optimistic in comparison to the fully-observed estimates of the *C*-statistic, there was no evidence to suggest this optimism was statistically significant. I am only able to conclude that no optimism is present in this instance due to the availability of the fully-observed Rotterdam data, in practice we can never know whether optimism will be an issue with *MI-then-Validate* methods.

11 How are missing data in covariates handled in observational time-to-event studies in oncology? A systematic review

This chapter details a systematic review [9] which was conducted in the first year of the PhD and published in *BMC Medical Research Methodology* .

[22]O. U. Carroll, T. P. Morris, and R. H. Keogh, “How are missing data in covariates handled in observational time-to-event studies in oncology? A systematic review”, *BMC Medical Research Methodology*, vol. 20, pp.1-15, May 2020.

The aim of the systematic review discussed in this chapter is to understand how researchers approach and handle missing covariate values in time-to-event analyses. For a study to be included into the review, it had to use a proportional hazards or an extended Cox model, while also making reference to missing data. Studies which aimed to investigate risk factors or which aimed to develop a prediction model were included. Particular focus was given to covariate selection, the functional forms of continuous covariates, assessment of the proportional hazards assumption and the handling of missing data and the assumptions made as this aligned with the initial aim of the PhD. Also of interest was whether time-varying effects or time-dependent covariates were included in the analysis and whether they were affected by missing data. Finally, the systematic review provides recommendations for using MI in time-to-event analyses. It also provides references to papers which address common issues that can arise in statistical analyses.

Conducting this systematic review gave me insights into the poor handling and reporting of missing data in observational time-to-event studies. The review demonstrated poor adherence to published guidelines, which focus on conducting and reporting an observational study. The default method to handle missing data is complete-case analysis and more modern methods are often overlooked.

In addition to the published paper, supplementary results concerning prediction modelling, which were not included in the final publication are available at the end of the chapter.

After completing this systematic review I decided to focus on the handling of missing data in prediction studies, focusing specifically on internal validation.

RESEARCH PAPER COVER SHEET

Please note that a cover sheet must be completed for each research paper included within a thesis.

SECTION A – Student Details

| | | | |
|---------------------|--|-------|----|
| Student ID Number | lsh1600418 | Title | Ms |
| First Name(s) | Orlagh | | |
| Surname/Family Name | Carroll | | |
| Thesis Title | Strategies for imputing missing covariate values in observational data | | |
| Primary Supervisor | Prof. Ruth Keogh | | |

If the Research Paper has previously been published please complete Section B, if not please move to Section C.

SECTION B – Paper already published

| | | | |
|--|----------------------------------|---|-----|
| Where was the work published? | BMC Medical Research Methodology | | |
| When was the work published? | 29/05/2020 | | |
| If the work was published prior to registration for your research degree, give a brief rationale for its inclusion | N/A | | |
| Have you retained the copyright for the work?* | Yes | Was the work subject to academic peer review? | Yes |

*If yes, please attach evidence of retention. If no, or if the work is being included in its published format, please attach evidence of permission from the copyright holder (publisher or other author) to include this work.

SECTION C – Prepared for publication, but not yet published

| | |
|---|-----------------|
| Where is the work intended to be published? | |
| Please list the paper's authors in the intended authorship order: | |
| Stage of publication | Choose an item. |

SECTION D – Multi-authored work

| | |
|---|--|
| <p>For multi-authored work, give full details of your role in the research included in the paper and in the preparation of the paper. (Attach a further sheet if necessary)</p> | <p>I undertook the data extraction and analysis of the papers included in the systematic review. I wrote the first draft of the manuscript and edited it following comments from my supervisors.</p> |
|---|--|

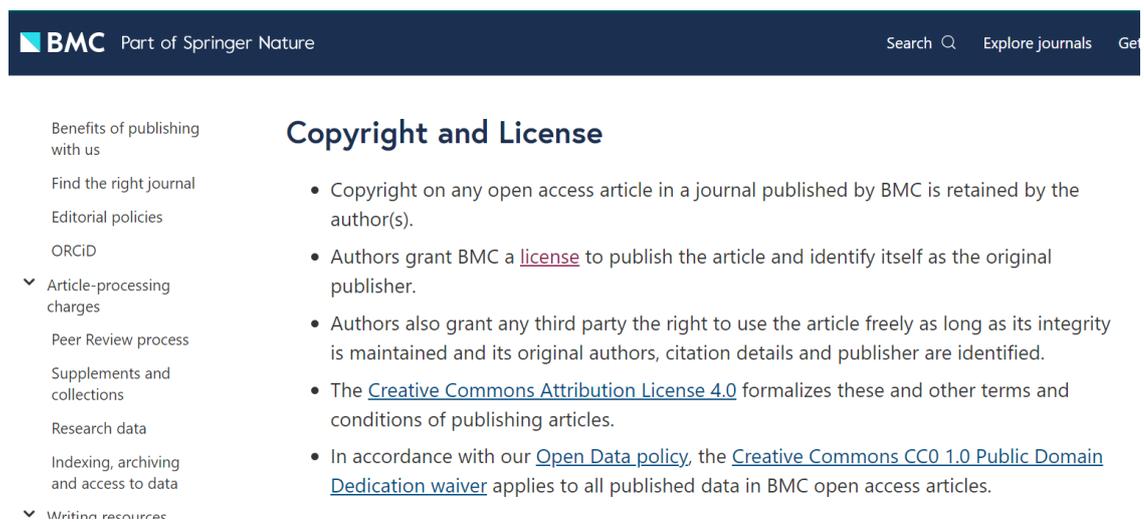
SECTION E

| | |
|--------------------------|------------|
| Student Signature | [Redacted] |
| Date | 18/08/21 |

| | |
|-----------------------------|------------|
| Supervisor Signature | [Redacted] |
| Date | 26/8/21 |

Evidence of copyright retention

The following research paper was published as an open-access paper with BMC Medical Research Methodology. All BMC journals publish under The Creative Commons Attribution License (CC BY). In addition, the Copyright and License webpage (currently available on the BMC website at <https://www.biomedcentral.com/getpublished/copyright-and-license>) states that “Copyright on any open access article in a journal published by BMC is retained by the author(s).



Copyright and License

- Copyright on any open access article in a journal published by BMC is retained by the author(s).
- Authors grant BMC a [license](#) to publish the article and identify itself as the original publisher.
- Authors also grant any third party the right to use the article freely as long as its integrity is maintained and its original authors, citation details and publisher are identified.
- The [Creative Commons Attribution License 4.0](#) formalizes these and other terms and conditions of publishing articles.
- In accordance with our [Open Data policy](#), the [Creative Commons CC0 1.0 Public Domain Dedication waiver](#) applies to all published data in BMC open access articles.

Screenshot of the BMC Copyright and License page stating that author’s of open-access articles retain copyright

Article relevant information

All numbered references in the following text can be found in the reference section at the end of this chapter, and are separate to the bibliography at the end of the thesis.

In the following article I mention two supplementary files. Supplementary file 1 can be found in AppendixD of the thesis. Supplementary file 2 is available online at the BMC Medical Research Methodology website where the article is published: <https://doi.org/10.1186/s12874-020-01018-7>.

The following pages contain the text from the published article.

How are missing data in covariates handled in observational time-to-event studies in oncology? A systematic review

Authors: Orlagh U Carroll, Tim P Morris, Ruth H Keogh

Abstract

Background: Missing data in covariates can result in biased estimates and loss of power to detect associations. It can also lead to other challenges in time-to-event analyses including the handling of time-varying effects of covariates, selection of covariates and their flexible modelling. This review aims to describe how researchers approach time-to-event analyses with missing data.

Methods: Medline and Embase were searched for observational time-to-event studies in oncology published from January 2012 to January 2018. The review focused on proportional hazards models or extended Cox models. We investigated the extent and reporting of missing data and how it was addressed in the analysis. Covariate modelling and selection, and assessment of the proportional hazards assumption were also investigated, alongside the treatment of missing data in these procedures.

Results: 148 studies were included. The mean proportion of individuals with missingness in any covariate was 32%. 53% of studies used complete-case analysis, and 22% used multiple imputation. In total, 14% of studies stated an assumption concerning missing data and only 34% stated missingness as a limitation. The proportional hazards assumption was checked in 28% of studies, of which, 17% did not state the assessment method. 58% of 144 multivariable models stated their covariate selection procedure with use of a pre-selected set of covariates being the most popular followed by stepwise methods and univariable analyses. Of 69 studies that included continuous covariates, 81% did not assess the appropriateness of the functional form.

Conclusion: While guidelines for handling missing data in epidemiological studies are in place, this review indicates that few report implementing recommendations in practice. Although missing data are present in many studies, we found that few state clearly how they handled it or the assumptions they have made. Easy-to-implement but potentially biased approaches such as complete-case analysis are most commonly used despite these relying on strong assumptions and where often more appropriate methods should be employed. Authors should be encouraged to follow existing guidelines to address missing data, and increased levels of expectation from journals and editors could be used to improve practice.

Keywords: Missing data, Time-to-event, Observational studies, Survival, Epidemiology, Oncology, Multiple imputation

Background

Time-to-event or survival studies focus on the analysis of times to an outcome or event. Missing data in covariates is a problem in many such investigations. It can render estimators biased if applied to the complete-cases or using an ad hoc approach to handling missingness, and a loss of power to detect associations between explanatory variables and times-to-event. The presence of missing data can also lead to further challenges in a survival setting such as the handling of time-varying effects or dealing with time-dependent covariates when values are missing in the covariates in question. Additionally, it can lead to questions about how best to approach the checking of model assumptions, for example, the proportional hazards assumption when using a Cox model. Missing data also brings further challenges not specific to time-to-event scenarios, such as how to address the selection of covariates into a model or the flexible modelling of covariates. Another type of missingness concerning time-to-event scenarios is missing observations of the event time due to patients being censored, for example, due to administrative censoring or loss to follow-up. This is typically addressed in the analyses for right-censored data, making the assumption that the censoring is uninformative. Missingness in the outcome is not assessed in this review which instead focuses on missing data in the explanatory covariates only.

Complete-case analysis is both a simple and popular method for dealing with missing data, which involves restricting the analysis to individuals with no missing data. Other simple approaches involve replacing missing observations in a covariate with the mean, median or modal value or the use of a missing indicator category for categorical covariates. While popular, these methods can be biased, inefficient or underestimate the variance of estimates. Multiple imputation is an increasingly popular method for handling missing data which involves replicating the original dataset multiple times and in each replication replacing the missing values with plausible observations drawn from the posterior predictive distribution [1]. It is typically conducted using the ‘missing at random’ (MAR) assumption [2], which also subsumes ‘missing completely at random’ (MCAR). MCAR means that missingness does not depend on the observed or missing values while MAR means that missingness is conditionally independent of the missing values given those which have been observed. Further methodology has been developed to adapt the use of multiple imputation in a survival setting. White and Royston in 2009 [3] focused on the Cox model and recommend including the Nelson-Aalen estimate and event indicator in the imputation model. Bartlett et al in 2015 [4] described an alternative imputation approach suitable for several analysis models including the Cox model and Keogh and Morris (2018) [5] adapted both approaches to handle time-varying covariate effects - that is, non-proportionality of hazards.

In addition to developed methodology, there have been several published guidelines focusing on how to conduct and report an observational study with some recommendations pertinent to reporting with incomplete covariate data, summarised in Table 1. Some guidelines, such as Sterne et al. [6] focus purely on the handling and reporting of missing data while using multiple imputation, whereas STROBE [7,8] and ROBINS-I [9] focus more generally on reporting of observational studies. Examples of recommendations range from providing detail on eligibility criteria of patients to clearly stating the selection process for the final analysis model to reporting the amount of missingness in each covariate and which method was chosen to deal with the missing observations. Sensitivity analyses are also recommended to investigate plausibility of any assumptions assumed and the robustness of results. These published guidelines aim to introduce transparency as well as replicability of results if another analyst were to conduct the same investigation.

Table 1: Summary of recommendations or considerations from STROBE, ROBINS-I and Sterne et al. guidelines

| Recommendation | Explanation | STROBE | ROBINS-I | Sterne |
|---|--|--------|----------|--------|
| Patient Selection | | | | |
| State eligibility criteria | <ul style="list-style-type: none"> • State inclusion and exclusion criteria of study participants, including criteria concerning missing data | ✓ | | ✓ |
| Report the number of individuals at each stage of the study | <ul style="list-style-type: none"> • Give reasons for exclusion at each stage | ✓ | | |
| | <ul style="list-style-type: none"> • Indicate the amount of individuals discarded due to missingness at each stage of the study | ✓ | | ✓ |
| | <ul style="list-style-type: none"> • Give consideration to selection bias introduced by exclusion criteria | | ✓ | |
| | <ul style="list-style-type: none"> • May use a flowchart to summarise | ✓ | | |

Table 1: continued

| Recommendation | Explanation | STROBE | ROBINS-I | Sterne |
|--|---|--------|----------|--------|
| Modelling and Covariate Selection | | | | |
| Covariates | <ul style="list-style-type: none"> • Detail whether included as continuous or categorical and, if relevant, detail how the quantitative covariate was categorised | ✓ | ✓ | |
| | <ul style="list-style-type: none"> • Consider departures from linearity for continuous covariates and state which transformation, if any, was used | ✓ | ✓ | |
| State analysis model | <ul style="list-style-type: none"> • make it clear which method will be used to model the data | ✓ | ✓ | |
| Covariate Selection | <ul style="list-style-type: none"> • describe the procedure used to reach the final model | ✓ | ✓ | |
| | <ul style="list-style-type: none"> • this includes, but is not restricted to, missing data imputation, transformation of covariates, interactions between covariates or inclusion of covariates for a priori reasons | ✓ | ✓ | |
| Results | <ul style="list-style-type: none"> • Provide unadjusted estimates and the final adjusted model | ✓ | ✓ | |
| | <ul style="list-style-type: none"> • State the number of participants included in unadjusted and adjusted analyses | ✓ | | |

Table 1: continued

| Recommendation | Explanation | STROBE | ROBINS-I | Sterne |
|---|---|--------|----------|--------|
| Missing Data | | | | |
| Report the number of participants with missing data | <ul style="list-style-type: none"> • Report this for each covariate of interest or the number of complete data for the important covariates | ✓ | | ✓ |
| | <ul style="list-style-type: none"> • Give reasons for missing values | ✓ | ✓ | ✓ |
| | <ul style="list-style-type: none"> • Investigate if there are key differences between those observed and those with missing data - this may be compared across exposure/intervention groups. | | ✓ | ✓ |
| <i>Missing data methods (general)</i> | | | | |
| Which method was used to handle missing data? | <ul style="list-style-type: none"> • State clearly the method used | ✓ | ✓ | ✓ |
| State any missing data assumptions that were made | <ul style="list-style-type: none"> • Such as whether the data are MCAR, MAR or MNAR | ✓ | ✓ | ✓ |
| Sensitivity analysis | <ul style="list-style-type: none"> • Should investigate robustness of findings | ✓ | ✓ | |
| | <ul style="list-style-type: none"> • Compare method with a complete-case analysis | | ✓ | |
| Sensitivity analysis | <ul style="list-style-type: none"> • If necessary, assess validity of methods if there are differences | ✓ | ✓ | |
| | <ul style="list-style-type: none"> • Assess plausibility of missing data assumptions | | ✓ | |

Table 1: continued

| Recommendation | Explanation | STROBE | ROBINS-I | Sterne |
|--|---|--------|----------|--------|
| <i>Multiple Imputation</i> | | | | |
| Give details of the imputation model | <ul style="list-style-type: none"> • State the software used and key settings for imputation model | | | ✓ |
| | <ul style="list-style-type: none"> • State the number of imputations used | | | ✓ |
| | <ul style="list-style-type: none"> • State variables included in imputation model | | | ✓ |
| | <ul style="list-style-type: none"> • State how non-normal or binary covariates were handled | | | ✓ |
| | <ul style="list-style-type: none"> • Were interactions in analysis model included in imputation model? | | | ✓ |
| If a large fraction of data are imputed, compare observed and imputed values | | | | ✓ |
| Missing data assumptions | <ul style="list-style-type: none"> • Discuss if variables included in the imputation model make MAR assumption plausible | | | ✓ |
| Sensitivity analyses | <ul style="list-style-type: none"> • Compare MI results with CC results | | | ✓ |
| | <ul style="list-style-type: none"> • Investigate departures from MAR assumption | | | ✓ |
| | <ul style="list-style-type: none"> • If necessary, suggest explanations for why there are differences in results across sensitivity analyses | | | ✓ |

Time-to-event studies are commonly conducted in oncology with a search for time-to-event or survival studies on Web of Science indicating oncology to be the most popular category at approximately 30% of journal articles. As such, this review focused on studies conducted in any area of oncology. Common scenarios involve assessing the risk factors of patients developing a specific cancer or investigating factors associated with survival post-diagnosis. Proportional hazards models and Cox regression, in particular, continues to be the dominant analysis technique in time-to-event studies. As such, the review focuses on proportional hazards models while allowing for the extension of the Cox model to include time-varying effects.

Given the developed methodology in this field and the detailed recommendations in place, this review aims to:

- understand which methods researchers are using in time-to-event analyses when missing data are present
- assess if methods used are being carried out appropriately and the relevant assumptions stated
- assess how other challenges such as covariate selection, choice of functional forms (i.e. whether the covariate should be included as a linear term or be more flexibly modelled) for continuous covariates and checking of model assumptions are handled, particularly in the presence of missing data.

Methods

Databases, search strategy and screening

Medline and Embase databases were searched for studies published between January 2012 and January 2018 to allow time for developed methods and guidelines to be used in practice. The search strategy for observational studies consisted of three main components: oncology, missing data and time-to-event analyses; additional details can be found in additional file 1.

For inclusion, studies had to use a proportional hazards or an extended Cox model (includes an interaction between a covariate and time) in a cancer setting. The study also had to have a reference to missing data (either ‘complete’ or ‘missing’) in the abstract or in the full-text. Studies involving only competing risks, frailty models, accelerated failure time models or excess hazards in the abstract or full-text were excluded from the review. If the abstract mentioned a time-to-event outcome but did not specify the analysis models used, the paper proceeded to a full-text review. Papers not written in English or which focused on methodology, meta-analyses, validations of previously created models, and primary or

secondary trial outcomes were excluded. However, retrospective observational analyses of a trial cohort were included.

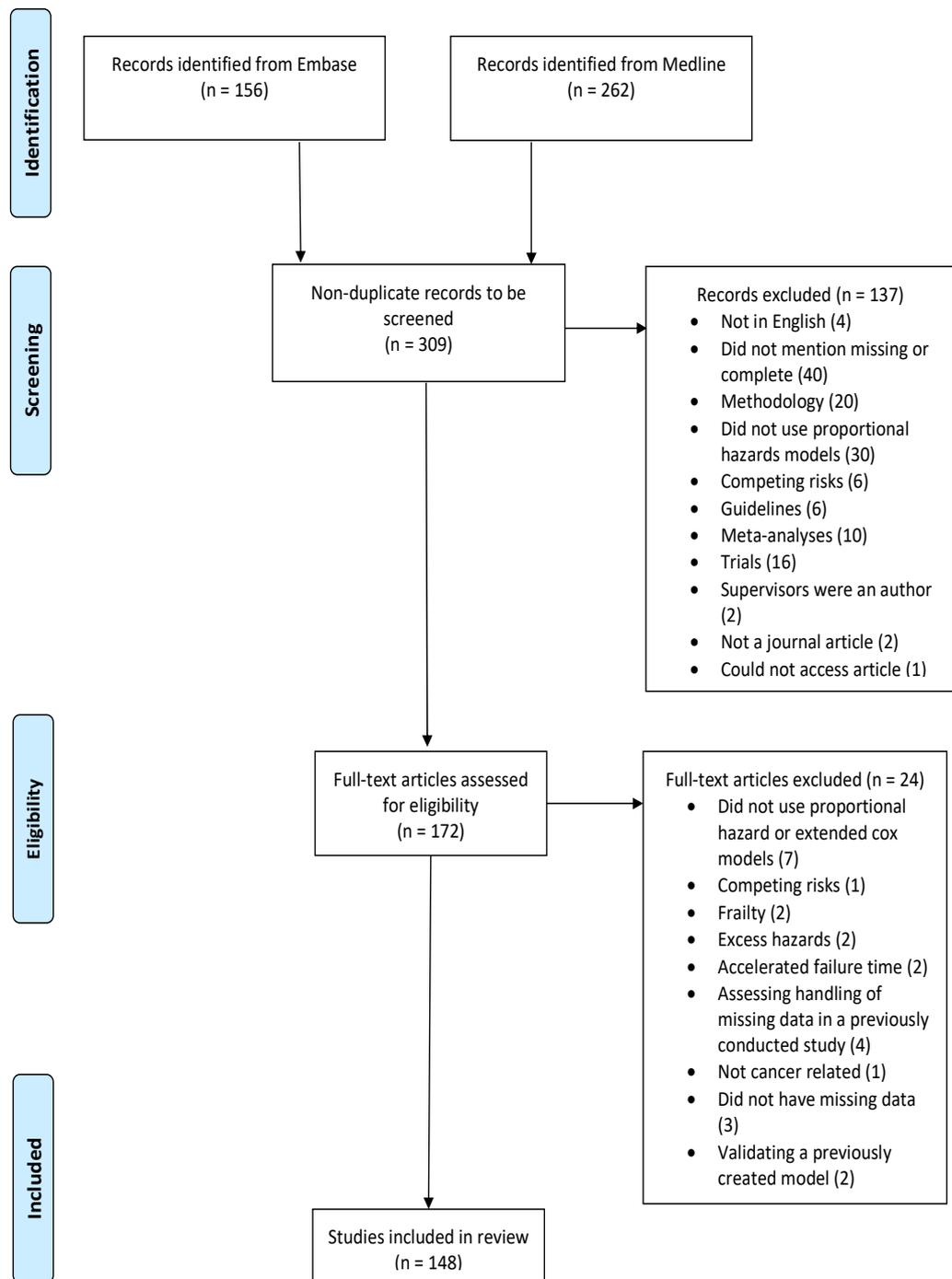
Data extraction

The information extracted focused on two key areas: missing data and features of the time-to-event analysis. The missing data component assessed the sample size used in the study, how much missing data had been discarded, if assumptions about the treatment of missing data in the analysis were stated and how any missing data were handled in the analysis: complete-case analysis, single imputation techniques or multiple imputation. Where multiple imputation was used, the choice of univariate or multivariate imputation was recorded, the number of imputations used and which covariates were included in the imputation model. Online supplementary materials were accessed only when referenced with regards to the handling of missing data in the text. The features of the time-to-event analysis assessed were whether the proportional hazards assumption was investigated, how covariates were selected for model inclusion and the assessment of the functional form (if continuous covariates were included). We also assessed, where relevant, how missing data were treated in the context of these features. In addition, the software used for the analysis was also extracted by searching for ‘Stata’, ‘SAS’, ‘SPSS’, ‘R’ and ‘plus’ (for S-plus and Mplus). Papers which did not mention one of these six programs were then searched for the software used. A detailed list of the information extracted can be found in additional file 1 which was motivated by the guideline recommendations found in Table 1 and are evaluated in the Results section.

A pilot investigation consisting of 10 randomly selected papers was carried out by OUC, TPM and RHK to assess the consistency of data extraction, refine the data extraction checklist and agree on how to extract information when answers were ambiguous. Data extraction was then carried out by OUC.

Results

The PRISM diagram [10] summarising the review inclusion process is shown in Figure 1. Four hundred and eighteen papers were identified from Embase and Medline, of which 309 were non-duplicates and proceeded to the screening step. One hundred and thirty-seven studies did not meet the inclusion criteria during screening and were therefore excluded. After a full-text assessment, a further 24 studies were excluded with a total of 148 studies included within the review. The studies included came from 110 journals, of which the most prominent were BMC Cancer (5), International Journal of Radiation Oncology, Biology, Physics (4) and Journal of the National Cancer Institute (4).



From: Moher D, Liberati A, Tetzlaff J, Altman DG, The PRISMA Group (2009). Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. PLoS Med 6(7): e1000097. doi:10.1371/journal.pmed1000097

For more information, visit www.prisma-statement.org.

Figure 1: Flowchart of the inclusion process for studies into the review [10]

Missing data

Reporting extent of missing data

In the pilot stage we noticed that many studies excluded individuals due to missing data on key covariates in an initial phase (in which the study population was determined using inclusion or exclusion criteria). That is, having certain covariates observed was used as part of the inclusion criteria. One hundred and six (72%) studies excluded missing data while determining their study population in this initial phase. Of the 106 studies which excluded observations, 66 (62%) reported the number of individuals excluded. On average, 14% of individuals were discarded during this stage. After inclusion criteria had been applied, 102 (69%) studies contained patients with missing data. Table 2 shows the breakdown of missing data during the initial phase and analysis stage of the study.

Table 2: Breakdown of the number of individuals with missing data.

| Description | Number | (%) |
|--|--------|----------------|
| Excluded missing data in initial phase (N=106) | | |
| Excluded individuals with missing data in any covariate ¹ | 44 | (42) |
| Excluded individuals with missing data in a subset of covariates | 62 | (58) |
| Reported the number of individuals excluded | 66 | (62) |
| Percentage (%) of individuals excluded ($n = 66$) | | |
| Mean (SD) | 14.14 | (12.40) |
| Median (IQR) | 10.22 | (4.73, 18.34) |
| Min, Max | 0.11, | 47.38 |
| Missing data present for the analysis stage (N=102) | | |
| Reported missing data in baseline table for incomplete covariates | 82 | (80) |
| Used a complete-case analysis ² | 35 | (34) |
| Used other missing data methods | 36 | (35) |
| Quantified the complete-case sample size | 25 | (25) |
| Percentage (%) of individuals excluded ($n = 25$) | | |
| Mean (SD) | 31.65 | (21.90) |
| Median (IQR) | 31.34 | (13.67, 37.76) |
| Min, Max | 1.77, | 94.16 |

The initial phase is the stage when defining the study population using inclusion exclusion criteria.

¹ 1 potentially used a complete-case in initial phase but did not clearly state their methods

² A further 31 were not clear on whether they used a complete-case during the analysis

In the demographics table (often considered to be ‘Table 1’ in publications), 87 (59%) studies summarised the missing data in covariates, 47 (32%) reported the breakdown of missingness in incomplete covariates and two (1%) used missing data pattern plots. Thirty-four (23%) used both the text and a Table to report the extent of missingness. For the 48 (32%) who did not use a plot, use a table or explicitly break down the missing values in each covariate a general statement was typically made stating which variables were incomplete or that variables or patients were excluded due to having incomplete data.

Analyses performed

Table 3 summarises the methods used for the analysis in the presence of missing data. Complete-case analysis was the most popular and was used in 79 (53%) studies either in the initial phase or at the analysis stage (either as the primary method used to deal with missing data or as a sensitivity analysis). This was followed in popularity by removing individuals with missing values in certain key covariates (62, 42%) and multiple imputation (33, 22%). Some studies used multiple methods for handling missing data with 18 (12%) using both complete-case and multiple imputation.

Table 3: Methods used in studies for the handling of missing data.

| Missing data Methods | Count | (%)* |
|---|-------|------|
| Complete-case | 79 | (53) |
| Removed individuals with incomplete data for a subset of covariates | 67 | (45) |
| Multiple Imputation | 33 | (22) |
| Missing indicator | 10 | (7) |
| Worst or best case scenario ¹ | 2 | (1) |
| Stochastic imputation | 1 | (1) |
| Mean value imputation | 1 | (1) |
| Mode value imputation | 1 | (1) |
| Growth models | 1 | (1) |
| Bayesian model incorporating handling of missing data | 1 | (1) |
| Full-information maximum likelihood estimation ² | 1 | (1) |
| Selection procedure ³ | 1 | (1) |
| Unclear | 33 | (22) |

* Percentages do not sum to 100 as there is overlap with some studies using more than one method.

¹ [11, 12]

² [11]

³ A selection model to account for missing data and time-varying covariates [13]

68 (50%) of all studies used a complete-case analysis as their primary analysis method and 24 (16%) reported multiple imputation as their main analysis. Of those using complete-case analysis as the main analysis, nine (13%) also used MI or other methods. Of those using MI as the main analysis, 12 (50%) used complete-case analysis or another method as a secondary analysis.

Missing data assumptions

Of the 148 studies, 128 (86%) did not state the assumptions that their chosen analysis made regarding the missing data. Eighteen (12%) stated the MAR assumption, of which 16 (89%) gave a general statement such as ‘MAR was assumed’, with no further explanation. One (0.7%) study stated MCAR and another stated ‘missing not at random’.

Sensitivity analyses and stating missing data as a limitation

Ninety-eight (66%) studies did not mention the presence of missing data as a limitation to their analysis. Twenty-six (18%) used sensitivity analyses to check the robustness of their final results to either different assumptions concerning the missingness or comparing results with other techniques to handle missing data.

Description of complete-case analysis

Thirty-five (34%) used a complete-case analysis and a further 31 (30%) were suspected to have used complete-case during the analysis stage based on the information provided but did not state this clearly in their paper. On average, 32% of individuals were discarded by applying a complete-case analysis, the maximum being 94% where complete-case was used as a sensitivity analysis for comparison with the main analysis using multiple imputation. Figure 2 summarises the reporting of missing data in the 79 studies that used a complete-case analysis (either during the initial phase or analysis stage). Seven (16%) of the 44 (56%) studies using the initial phase complete-case stated missing data as a limitation. Of the two (5%) studies using a sensitivity analysis, one compared with multiple imputation and the other compared the initial complete-case results pre and post propensity score matching and therefore with different sample sizes. Thirty-five (44%) studies used complete-case during the analysis stage, of which 18 (51%) stated missing data as a limitation. In addition, we presumed based on the information provided that a further 33 studies used a complete-case in the initial phase or analysis stage but did not clearly state this as their method to handle missing data.

Sensitivity analyses with complete-case: Eighteen (51%) studies used a sensitivity analysis, of which 14 (78%) involved multiple imputation versus complete-case analysis, two (11%) used complete-case analysis where individuals with missing data in any covariate were excluded versus excluded if there were missing data in a specific subset of covariates

(known as available case analysis), one (6%) tested various missing data assumptions and one (6%) did not specify.

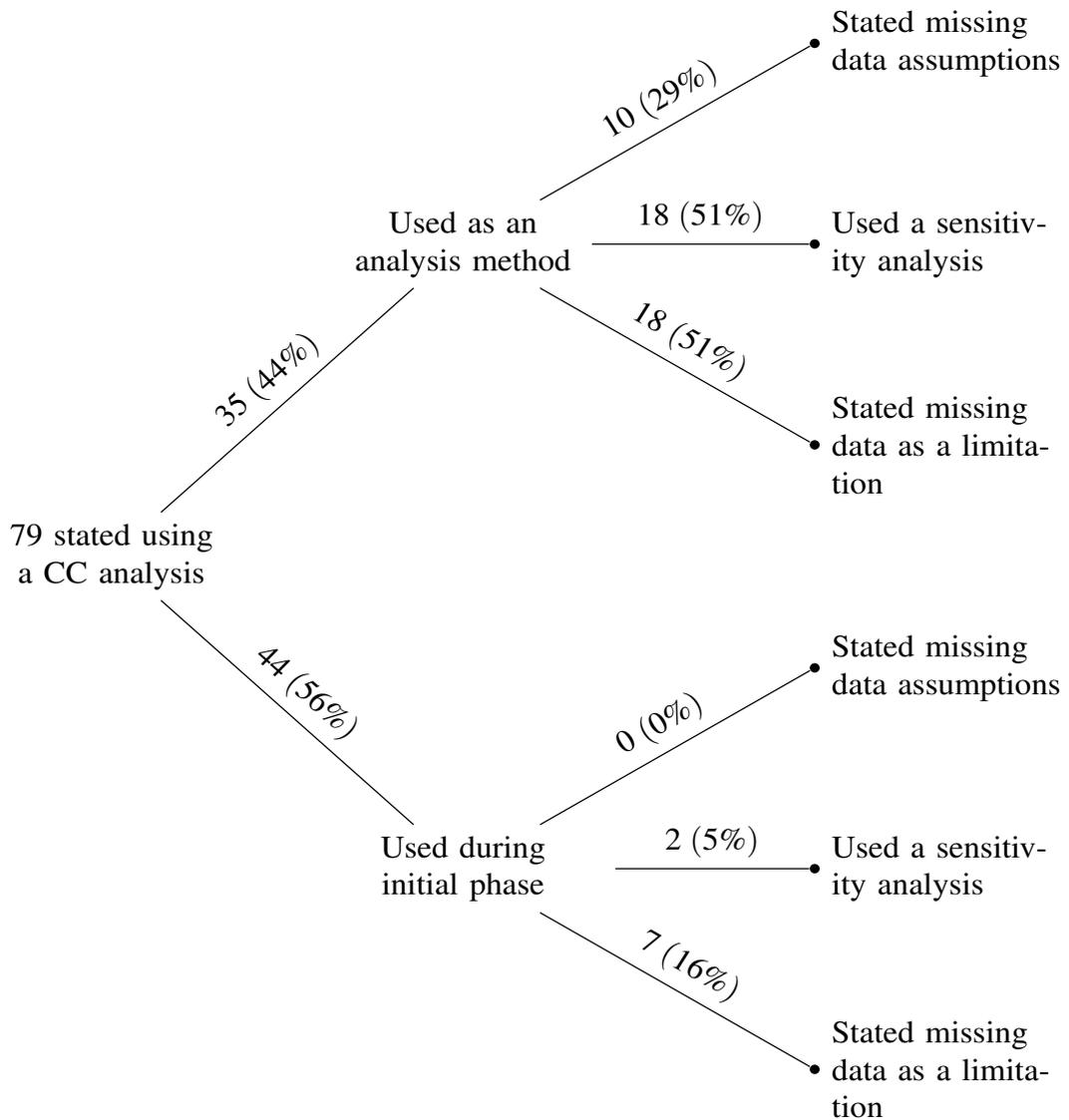


Figure 2: Breakdown of complete-case (CC) usage. The initial phase refers to those who used complete-case analysis when determining inclusion/exclusion of individuals to the study population.

Description of multiple imputation

The breakdown of multiple imputation usage can be seen in Figure 3. Thirty-three (22%) studies used multiple imputation, of which 24 (73%) reported the multiple imputation estimates as their main study results. Fourteen (42%) stated a missing data assumption and 25 (76%) described whether a multivariate or univariate approach was taken. For those using a multivariate imputation approach (22, 88%), multivariate imputation by chained equations (MICE) was the most popular method (19, 86%). In total, 14 (42%) studies included a component of the time-to-event outcome in their imputation model. These

included the baseline hazard (7, 50%), the event indicator (9, 14%) or both (2, 14%). Twenty-six (79%) studies using multiple imputation stated the number of imputations. One (3%) used a single imputation, five (15%) used five, six (18%) used 10, seven (21%) used 20, two (6%) used 25 and five (15%) used 50. Some studies (example: [14]) cited the White, Royston and Wood paper [15] which suggests that the rule of thumb for choosing the number of imputations should be at a minimum the percentage of cases that are incomplete while other studies (example: [16]) stated the number of imputations with no justification.

Sensitivity analyses with multiple imputation: Of the 21 (64%) studies that conducted a sensitivity analysis, 18 (90%) compared complete-case and multiple imputation (three of which did not explicitly state complete-case) and 10 (56%) used multiple imputation as the main analysis method and one (6%) was unclear on the main strategy while reporting both multiple imputation and complete-case results.

Missing data assumptions with methods

Of the 18 studies which stated the MAR assumption, 11 (61%) used multiple imputation, two (11%) used complete-case and one (16%) was not clear on whether they used complete-case or multiple imputation, two (11%) were suspected to have used complete-case but did not clearly state, one (16%) used a stochastic single regression imputation model and one (16%) used a fully Bayesian model. The one study stating MCAR used complete-case analysis and the other stating ‘missing not at random’ performed an analysis using a selection model for the joint distribution of the missing covariates, the outcome and the probability that covariate data are missing [13].

None of the 44 (56%) studies using the initial phase complete-case stated a missing data assumption and for the 35 using complete-case analysis during the analysis stage 10 (29%) stated a missing data assumption. This consists of one (10%) study stating MCAR and nine (90%) MAR, of which seven (78%) used multiple imputation and complete-case analysis together for sensitivity analyses (six (86%) of these used multiple imputation as the main method for handling missing data and complete-case analysis used as a comparison). For the 14 studies using multiple imputation having stated a missing data assumption, 13 (93%) used MAR and one (7%) used MCAR.

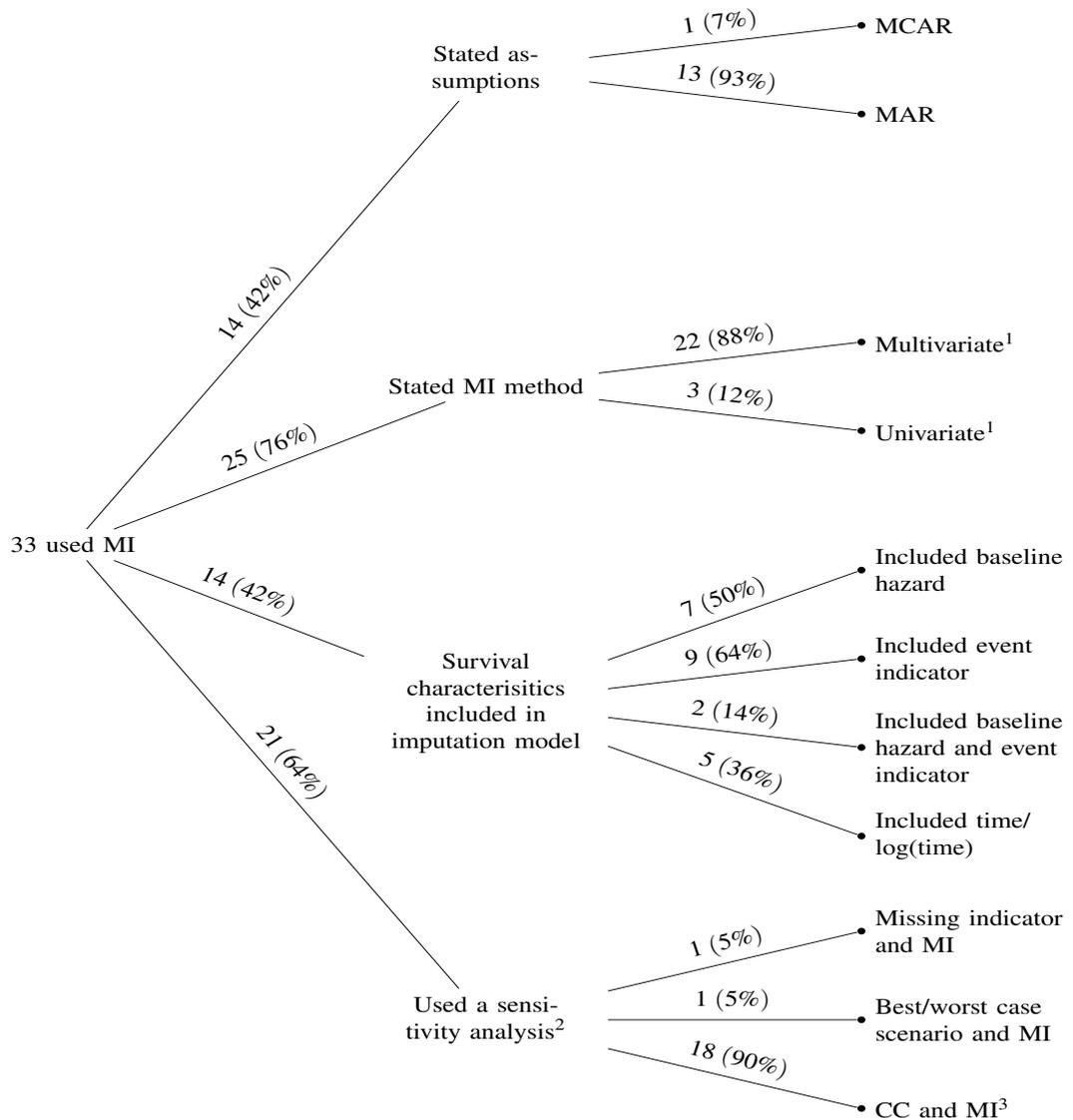


Figure 3: Breakdown of multiple imputation (MI) usage. ¹ 2 did not specify what type of multivariate MI model used, similarly 1 for univariate. ² 1 study ensured the sample size stayed the same for different models. ³ 3 studies did not clearly state that they were using complete-case.

Features of the analysis

Selection of covariates into the model

One hundred and forty-four (97%) studies used a multivariable model and therefore used some selection method or criteria to select which covariates should be included. Of these, 85 (59%) stated a clear selection procedure. The use of a predefined set of covariates (33, 39%), stepwise methods (31, 37%) and univariable analyses (32, 38%) were most commonly used, with eight (10%) studies using both a predefined set and univariable analyses. Of the 31 studies using stepwise methods, backwards elimination was used in 18 (58%), six (19%) used forwards selection and seven (23%) did not state which type of stepwise method was used.

Eleven (35%) used complete-case as the main method to handle missing data. Five (45%) out of the 31 studies using stepwise methods stated excluding individuals with missing data on key covariates in the initial phase, and were left with no additional missing data at the analysis phase. Six (55%) studies that used a stepwise procedure did so in a complete-case analysis. We suspect that an additional six (19%) studies used complete-case analysis but this was not clearly stated.

For the 31 studies using stepwise methods, 10 (32%) used them in combination with multiple imputation. Of this, eight (80%) used multiple imputation as the main method to handle missing data, one (10%) used a missing indicator as the main method with multiple imputation as a sensitivity analysis and one (10%) used a sensitivity analysis but did not state whether multiple imputation or complete-case was the main method. For those who used multiple imputation as the primary method to handle missing data, seven (88%) did not state how they combined it with the stepwise methods and the other is suspected to have applied the stepwise procedure in a complete-case analysis to determine the set of covariates to be included, before using this set in the model fitted in each imputed dataset, however this was not clearly stated.

For the 32 studies using univariable analyses, four (13%) studies used multiple imputation as the main analysis method, 14 (44%) stated that they used a complete-case as the main method for handling missing data, 12 (38%) were presumed to have used a complete-case analysis based on the information available, one (3%) used stochastic single regression imputation model and one (3%) did not include incomplete covariates in the analysis model (available case analysis).

Functional form of continuous covariates

Sixty-nine (47%) studies included continuous covariates in their model, of which 57 (83%) did not report considering whether any form other than linear was required, i.e. its appropriate functional form. For those that did consider it, splines were the most popular way of transforming covariates in the model (8, 12%), followed by fractional polynomials (2, 3%) and Martingale residuals (2, 3%). Including a quadratic term or a 'flexible non-linear model' were each used once. One study used Martingale residuals, cubic splines and fractional polynomials to investigate evidence of non-linear associations [17]. For a further 11 (7%) studies, it was not clear whether included covariates were continuous or categorical. For the 12 studies which reported assessing the functional form of covariates, three (25%) used multiple imputation as the main method for handling missing data, three (25%) used initial phase complete-case, three (25%) presumably used complete-case analysis but this was not clearly stated, one (8%) used stochastic single regression imputation, one (8%)

used a study-specific model to impute missing values and one (8%) used available case analysis by restricting to individuals with complete data in the covariates to be included in the analysis model.

Proportional hazards assumption and time-varying effects of covariates

The primary analysis method in 142 (96%) studies was the Cox model and the remaining six (4%) stated the use of a proportional hazards model. When investigating the proportional hazards assumption, the covariates included within the analysis model should be assessed. Forty-one (28%) studies stated that the proportional hazards assumption was assessed either using a general statement (example: [18]) or specifically detailing how to handle the covariates which violated it (example: [19]). Of those who checked, seven (17%) did not state the method used to assess the assumption. Schoenfeld residuals were most frequently used (18, 44%), followed by visual inspection of plots of Kaplan-Meier estimates of survivor curves, or functions thereof (12, 29%). Of these two methods, seven (17%) studies used both. Ten (24%) studies tested the assumption by including an interaction between covariates and follow-up time in the model.

For the studies that checked the assumption, 13 (32%) used multiple imputation as the main method to handle missing data, three (7%) used a missing indicator, 11 (27%) used complete-case analysis, seven (17%) presumably used complete-case analysis but did not clearly state, two (5%) had no missing data in covariates chosen for inclusion in the analysis model, one (2%) used both multiple imputation and complete-case but did not state which was the main method, one (2%) excluded incomplete covariates from the analysis model and one (2%) removed individuals with missing data in specific covariates. For the 18 studies using Schoenfeld residuals, six (33%) used multiple imputation as the main method for handling missing data. For the 12 studies using visual inspection of survivor curve plots four (33%) used multiple imputation and for the 10 including an interaction with time, three (30%) used multiple imputation as the main method.

Five studies discovered evidence for time-varying effects, of which three (60%) had incomplete covariates associated with time-varying effects. Two (67%) of these used multiple imputation to impute the missing values in the covariate, of which one took into account the time-varying effect using methods developed by Keogh and Morris [5] and the other stated using MICE while the third study was unclear on how they handled the missing data.

Software

Forty-four (30%) studies used SPSS, 41 (28%) used SAS, 36 (24%) used Stata, 11 (7%) used R, two (1%) used winbugs, one (1%) used XL-stat life and 17 (11%) did not state. Of

these, three (2%) used both SAS and SPSS, one (1%) used SAS and Stata together, one (1%) used SAS and S-plus and one (1%) used SAS and Mplus. Of the 11 studies using R, four (36%) used multiple imputation with three using the MICE package and one using Hmisc. Examples of other potential packages that could have been used are Amelia [20], jomo [21] or smcfcs [22].

Discussion

Missing data is a pervasive problem in observational time-to-event studies. However, this review has found that few studies appropriately report this issue. Whether this is due to a lack of appreciation of the potential implications of missing data from the researcher, or to the handling of missing data not being deemed of high enough importance to be described in the ‘Methods’ section is unclear. There are general guidelines in place such as STROBE [7,8] and Sterne’s specific multiple imputation recommendations [6] from 2007 and 2009, respectively, but it appears that many of their recommendations are still not being implemented. By considering literature from 2012 onwards, all papers we reviewed came after the publication of these guidelines. Over half of papers considered (53%) were from 2016 onwards. A surprising finding was that in many studies it was not clear how the study population was selected and what the extent of missing data was. We recommend that authors provide clear and comprehensive information on these aspects including detailing the finalisation of the study population, and stating the sample size used in each model when missing data are present. These recommendations would aid in the transparency of research findings.

Methods for handling missing data such as the multiple imputation approach of White and Royston [3] were implemented by two studies in 2014 [23] and 2016 [24], five and seven years respectively after the method was published. Although valid methods have been developed to handle missing data, the easier-to-implement approach of complete-case analysis is still the most popular method used. However, the studies suggest that little or no consideration is being given to the missing data assumptions needed for this method and whether they are introducing bias to their results. It is plausible that some authors had not noticed the missing data, since software by default runs complete-case analysis without flagging that some individuals were dropped from the analysis. Also of note are the studies which have a ‘fully’ observed dataset and therefore had no need to consider any missing data assumptions or methods. However, this ‘fully’ observed dataset originated from using a complete-case inclusion/exclusion criteria for individuals entering their study. These studies gave no consideration to missing data assumptions and only seven (16%) considered the missing data excluded to be a limitation.

Several systematic reviews have been conducted to assess the handling of missing data in

studies, most of which have focus on randomized trials. Wood et al. [25] reviewed the handling of missing outcome data in randomized control trials published in 2001. They found that missing data are typically handled inadequately and that there was almost no use of modern data methods with complete-case used in 46% of studies. Similar findings were made in other reviews covering trials published between 2005 and 2014 [26-28]. Karahalios et al. [29] focused on missing data in cohort studies published between 2000 and 2009 and found inconsistent reporting of missing data and inappropriate methods used with 66% of studies using complete-case analysis. With regards to missing data, these reviews collectively looked at studies published between 2001 and 2014. They, along with our own review looking at papers from 2012 to 2018, highlight the lack of progress that has been made in appropriate handling of missing data in both trials and observational studies.

Our review revealed a lack of rigour in other aspects of a study investigation. 42% of studies did not state how the covariates were selected for their final model. When conducting a covariate selection procedure, thought should also be given to continuous covariates and whether categorising is worth a loss of power to detect associations or the occurrence of residual confounding [30]. Clarity should also be required for how the selection procedure is combined when using multiple imputation. For example, [31] states using multiple imputation for multivariable analyses and goes on to detail that univariable models and backward selection were used. However, no discussion is given as to if this process was repeated across the multiple imputed datasets and, if so, what happened when there were disagreements across them regarding the selection process? In 2008, Wood et al. [32] discusses methods to handle covariate selection with multiple imputation. Studies included in the review tended to be exploratory or predictive in nature and consideration should be given to the selection procedure for including covariates into these models. Stepwise methods were used in 37% of studies which stated a covariate selection procedure despite the disadvantages being well-known [33, p.68]. These include underestimating standard errors of parameter estimates, narrow confidence intervals, low p-values and parameter estimates biased away from zero. VanderWeele also discusses the use of stepwise methods and their drawbacks in a causal setting [34].

For the 41 studies that checked the PH assumption, it was not clear how the 13 studies using multiple imputation incorporated the use of Schoenfeld residuals or inspection of survivor curves as these details were not provided. For those using a time-interaction and multiple imputation, only one did not make it clear how they were incorporating the two methods. Using again the example of [31] it is possible that they checked the assumption using scaled Schoenfeld residuals over time in a complete-case scenario or individually in each imputed dataset but without specification it is difficult to say whether the assumption diagnostics were carried out appropriately. It is important to note that when considering compatibility between the analysis and imputation model thought should also be given

to allowing for time-varying effects in the imputation process, in order to allow for valid tests of the proportional hazards assumption. Further thought should also be provided on whether there is sufficient statistical power to detect violations of the proportional hazards assumption [35].

This review demonstrates poor adherence to guidelines already in place and further drives the need for clear reporting. Ideally, an external analyst should be able to rerun the study analysis from the information published which is currently not possible in many studies. Finally, Table 4 provides some related references for consideration of different aspects of missing data and time-to-event features in a study.

Table 4: Selected papers describing methods for addressing common issues arising in the analysis of time-to-event data when there is missing covariate data

| Consideration | Some recommended references |
|-----------------------------------|---|
| Missing data (general) | |
| General recommendations | [6] <i>Sterne et al.</i> : Recommendations for missing data and multiple imputation |
| Simple imputation | [36] <i>Zhang</i> : Mean, median, mode, regression imputations |
| Complete-case bias considerations | [37] <i>Bartlett et al.</i> : When CC is valid |
| | [38] <i>Carpenter & Kenward</i> : When CC is valid |
| Multiple imputation | |
| Number of imputations to use | [15] <i>White et al.</i> : at least the percentage of incomplete cases [39] <i>von Hippel</i> : two-stage quadratic rule |
| Covariate selection procedures | [32] <i>Wood et al.</i> : Repeated use of Rubin’s rules or stacking approach [40] <i>Morris et al.</i> : Adapted for MFP including selection procedure and functional form |
| Non-linear effects | [40] <i>Morris et al.</i> : Adapted for MFP including selection procedure and functional form [41] <i>Seaman et al.</i> : recommend just another variable (JAV) approach |
| Using a Cox model | [3] <i>White & Royston</i> : inclusion of Nelson-Aalen estimate and event indicator in imputation model [4] <i>Bartlett & Seaman</i> : full conditional specification adjusting for the analysis model of choice |

MFP: Multivariable fractional polynomials

Table 4: Continued

| Consideration | Some recommended references |
|--|--|
| Multiple imputation | |
| Testing the Proportional hazards assumption and modelling time-varying effects of covariates | [5] <i>Keogh & Morris</i> : adapting White & Royston and Bartlett & Seaman approaches for time-varying effects |
| Time-dependent covariates | [42] <i>De Silva et al.</i> : Investigating performance of two-fold fully conditional specification for time-dependent covariates [43] <i>Moreno-Betancur et al.</i> : Use of joint modelling for time-dependent covariates |
| Time-to-event features not concerning missing data | |
| Functional form | [44] <i>Sauerbrei et al.</i> : multivariable fractional polynomial time i.e. MFP in survival setting accounting for time-varying effects [45] <i>Buchholz & Sauerbrei</i> : comparison of procedures for assessing time-varying effects and functional form [46] <i>Heinzel & Kaider</i> : Using cubic spline functions to assess functional form [47] <i>Wynant & Abrahamowicz</i> : Importance of assessing time-varying effects and functional form [48] <i>Abrahamowicz & MacKenzie</i> : Joint estimation of time-varying effects and functional form using splines |
| Covariate selection procedures | [44] See above [49] <i>Yan & Huang</i> : Assessing time-varying effects using an adaptive lasso method |
| Testing the Proportional hazards assumption | [35] <i>Austin</i> : Assessing power of tests to assess proportional hazards assumption [50] <i>Bellera et al.</i> : Recommend assessing proportional hazards assumption and inclusion of time-varying effects where necessary [51] <i>Abrahamowicz et al.</i> : use of regression splines to model time-varying effects [52] <i>Hess</i> : use of cubic splines to model time-varying effects |

MFP: Multivariable fractional polynomials

Table 4: Continued

| Consideration | Some recommended references |
|---|--|
| Time-to-event features not concerning missing data | |
| Time-varying effects | [44] See above |
| | [45] See above |
| | [46] See above |
| | [47] See above |
| | [48] See above |
| | [49] See above |
| | [50] See above |
| | [52] See above |
| General study considerations | |
| Categorising of covariates | [53] <i>MacCallum et al.</i> : Discussion on dichotomising continuous covariates |
| Non-linear effects | [54] <i>Royston & Sauerbrei</i> : Text book providing overview of model selection with a focus on MFP procedures |
| | [33] <i>Harrell</i> : Text book providing overview of strategies for regression modelling |
| Covariate selection procedures | [54] See above |
| | [55] <i>Heinze et al.</i> : Review of methods for covariate selection |

MFP: Multivariable fractional polynomials

Limitations of review

A large number of search terms were used to extract the relevant studies. However, it is possible that some time-to-event studies did not mention how they handled missing data in the title, abstract or keywords and therefore were not included in the review. The search also focused solely on oncology, it is possible that in other medical setting studies could be performed differently in terms of reporting or methods used. A further limitation stems from only one reviewer identifying, screening and extracting information from the studies which may have introduced bias from the selection and interpretation of papers. An agreement check was conducted with RHK and TPM and initially found poor agreement in the collection of sample size of studies and the amount of missing data. The data collection check-list was reviewed and amended to improve discrepancies.

Many journals have a page or word limit which restricts the study analysts from fully

detailing methods conducted and results. It is possible that studies were unable to detail information such as checking the PH assessment or conducting a sensitivity analysis. However, most journals also allow for online supplementary materials which could have been used.

For this review we focused on methods used in the oncology field. It is possible that the handling of missing data may be better or worse in other medical fields or study designs.

Recommendations for multiple imputation in time-to-event analyses

While it is difficult to recommend a gold standard method as it can depend on the context of the study, for time-to-event studies involving the Cox model we would recommend using the substantive model compatible fully conditional specification (SMC-FCS) of Bartlett et al. [4] as the gold standard method for multiple imputation. It allows for compatibility between the study analysis model and the imputation model. This method is available in both Stata and R software. Keogh and Morris [5] have adapted SMC-FCS to allow for the presence of time-varying effects and proposed an algorithm to allow for model selection with time-varying effects.

White and Royston [3] recommend the inclusion of the event indicator and the Nelson-Aalen estimator in the imputation model for an approximately compatible model. While this is simpler and more straightforward using widely available MI software, the approximation can perform badly in ‘extreme’ scenarios such as strong covariate effects and a high event rate. The approximation also has weaker statistical properties (estimators will generally be inconsistent) than SMC-FCS due to semi-compatibility of the imputation and analysis model. Keogh and Morris have also adapted White and Royston’s method to handle time-varying effects.

Conclusions

More consideration is required for observational time-to-event analyses with missing data, including clear reporting of how the missing data were handled and how any selection procedures or assumption checks were conducted in conjunction with the missing data method implemented. Wider thought should be given to the limitations the missing data introduces to the observational study, such as bias of parameter estimates, and which methods can be used to help deal with this. While methods such as complete-case analysis are well ingrained in the community there are more modern methods which should also be considered when conducting a study. There appears to be a delay between methodology publication and uptake into the applied research field [56] or, rather, a delay in departing from simpler favoured methods of the field. There are many published guidelines readily available to help researchers conduct and report their study and these should be

consulted, alongside a statistician. All recommendations that came from conducting the review were found to have already been emphasised in the published guidance discussed in the Introduction section of this paper. Finally, we recommend that journal editors have requirements for appropriate reporting in the presence of missing data to ensure high quality studies are published and that their results are robust.

List of Abbreviations

CC: Complete-case; MCAR: Missing completely at random; MAR: Missing at random; MFP: Multivariable fractional polynomials; MI: Multiple Imputation; MICE: Multivariate imputation by chained equations.

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Competing interests

The authors declare that they have no competing interests.

Funding

OUC was funded by the Economic and Social Research Council Doctoral Training Partnership (ES/P000592/1).

Author's contributions

OUC performed data extraction and analysis, and wrote the manuscript. RHK and TPM contributed to the design of the data extraction checklist, conducted a pilot check of data extraction involving 10 papers and provided feedback on all versions of the paper.

Availability of data and material

The dataset supporting the conclusions of this article is included within the additional files.

Acknowledgements

We thank Ian White for providing helpful comments on the initial results of this review and the Economic and Social Research Council Doctoral Training Partnership for funding OUC.

References

1. Rubin DB. Multiple Imputation for Nonresponse in Surveys. United States of America: Wiley; 1987.
2. Little RJA, Rubin DB. Statistical Analysis with Missing Data, 2nd edn. United States of America: Wiley; 2002.
3. White IR, Royston P. Imputing missing covariate values for the Cox model. *Stat Med*. 2009;28(15):1982–98.
4. Bartlett JW, Seaman SR, White IR, Carpenter JR. Multiple imputation of covariates by fully conditional specification: Accommodating the substantive model. *Stat Methods Med Res*. 2015;24(4):462–87.
5. Keogh Ruth H., Morris Tim P. Multiple imputation in Cox regression when there are time-varying effects of covariates. *Statistics in Medicine*. 2018;37(25):3661–3678.
6. Sterne JAC, White IR, Carlin JB, Spratt M, Royston P, Kenward MG, Wood AM, Carpenter JR. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *BMJ (Clin Res Ed)* 2009;338:2393.
7. Von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies for the STROBE initiative. *Lancet*. 2007;370(9596):1453–57.
8. Vandenbroucke JP, von Elm E, Altman DG, Gøtzsche PC, Mulrow CD, Pocock SJ, Poole C, Schlesselman JJ, Egger M, Initiative ftS. Strengthening the Reporting of Observational Studies in Epidemiology (STROBE): Explanation and Elaboration. *PLoS Med*. 2007;4(10):297.
9. Sterne JA, Hernán MA, Reeves BC, Savović J, Berkman ND, Viswanathan M, Henry D, Altman DG, Ansari MT, Boutron I, Carpenter JR, Chan A-W, Churchill R, Deeks JJ, Hróbjartsson A, Kirkham J, Jüni P, Loke YK, Pigott TD, Ramsay CR, Regidor D, Rothstein HR, Sandhu L, Santaguida PL, Schünemann HJ, Shea B, Shrier I, Tugwell P, Turner L, Valentine JC, Waddington H, Waters E, Wells GA, Whiting PF, Higgins JP. ROBINS-I: a tool for assessing risk of bias in non-randomised studies of interventions. *BMJ (Clin Res Ed)*. 2016; 355. 10.1136/BMJ.I4919.

10. Moher D, Liberati A, Tetzlaff J, Altman DG, Group TP. Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. *PLoS Med.* 2009;6(7):1000097.
11. Elaidi R, Harbaoui A, Beuselinck B, Eymard J-C, Bamias A, De Guillebon E, Porta C, Vano Y, Linassier C, Debruyne PR, Gross-Goupil M, Ravaud A, Aitelhaj M, Marret G, Oudard S. Outcomes from second-line therapy in long-term responders to first-line tyrosine kinase inhibitor in clear-cell metastatic renal cell carcinoma. *Ann Oncol Off J Eur Soc Med Oncol.* 2015;26(2):378–85.
12. Clive AO, Kahan BC, Hooper CE, Bhatnagar R, Morley AJ, Zahan-Evans N, Bintlcliffe OJ, Boshuizen RC, Fysh ETH, Tobin CL, Medford ARL, Harvey JE, Van Den Heuvel MM, Lee YCG. Predicting survival in malignant pleural effusion: Development and validation of the LENT prognostic score. *Thorax.* 2014;69(12):1098–104.
13. Bradshaw PT, Ibrahim JG, Stevens J, Cleveland R, Abrahamson PE, Satia JA, Teitelbaum SL, Neugut AI, Gammon MD. Postdiagnosis change in bodyweight and survival after breast cancer diagnosis. *Epidemiology.* 2012;23(2):320–7.
14. Lukic M, Licaj I, Lund E, Skeie G, Weiderpass E, Braaten T. Coffee consumption and the risk of cancer in the Norwegian Women and Cancer (NOWAC) Study. *Eur J Epidemiol.* 2016;31(9):905–16. [PubMed] [Google Scholar]
15. White IR, Royston P, Wood AM. Multiple imputation using chained equations: Issues and guidance for practice. *Stat Med.* 2010;30(4):377–99.
16. Vogiatzoglou A, Mulligan AA, Bhaniani A, Lentjes MAH, McTaggart A, Luben RN, Heiss C, Kelm M, Merx MW, Spencer JPE, Schroeter H, Khaw K-T, Kuhnle GGC. Associations between flavan-3-ol intake and CVD risk in the Norfolk cohort of the European Prospective Investigation into Cancer (EPIC-Norfolk) *Free Radic Biol Med.* 2015;84:1–10.
17. Thompson EM, Hielscher T, Bouffet E, Remke M, Luu B, Gururangan S, McLendon RE, Bigner DD, Lipp ES, Perreault S, Cho Y-J, Grant G, Kim S-K, Lee JY, Rao AAN, Giannini C, Li KKW, Ng H-K, Yao Y, Kumabe T, Tominaga T, Grajkowska WA, Perek-Polnik M, Low DCY, Seow WT, Chang KTE, Mora J, Pollack IF, Hamilton RL, Leary S, Moore AS, Ingram WJ, Hallahan AR, Jouvett A, Fevre-Montange M, Vasiljevic A, Faure-Contier C, Shofuda T, Kagawa N, Hashimoto N, Jabado N, Weil AG, Gayden T, Wataya T, Shalaby T, Grotzer M, Zitterbart K, Sterba J, Kren L, Hortobagyi T, Klekner A, Laszlo B, Poczta T, Hauser P, Schuller U, Jung S, Jang W-Y, French PJ, Kros JM, van Veelen M-LC, Massimi L, Leonard JR, Rubin JB, Vibhakar R, Chambless LB, Cooper MK, Thompson RC, Faria CC, Carvalho A, Nunes S, Pimentel J, Fan X, Muraszko KM, Lopez-Aguilar E, Lyden D, Garzia L,

Shih DJH, Kijima N, Schneider C, Adamski J, Northcott PA, Kool M, Jones DTW, Chan JA, Nikolic A, Garre ML, Van Meir EG, Osuka S, Olson JJ, Jahangiri A, Castro BA, Gupta N, Weiss WA, Moxon-Emre I, Mabbott DJ, Lassaletta A, Hawkins CE, Tabori U, Drake J, Kulkarni A, Dirks P, Rutka JT, Korshunov A, Pfister SM, Packer RJ, Ramaswamy V. Prognostic value of medulloblastoma extent of resection after accounting for molecular subgroup: a retrospective integrated clinical and molecular analysis. *Lancet Oncol.* 2016;17(4):484–95. [PMC free article] [PubMed] [Google Scholar]

18. Renfro LA, Grothey A, Xue Y, Saltz LB, Andre T, Twelves C, Labianca R, Allegra CJ, Alberts SR, Loprinzi CL, Yothers G, Sargent DJ, Group ACCEA. ACCENT-based web calculators to predict recurrence and overall survival in stage III colon cancer. *J Natl Cancer Inst.* 2014; 106(12). 10.1093/jnci/dju333.
19. Ali HR, Dawson S-J, Blows FM, Provenzano E, Leung S, Nielsen T, Pharoah PD, Caldas C. A Ki67/BCL2 index based on immunohistochemistry is highly prognostic in ER-positive breast cancer. *J Pathol.* 2012;226(1):97–107.
20. Honaker J, King G, Blackwell M. Amelia II: A program for missing data. *J Stat Softw.* 2011;45(7):1–47.
21. Quartagno M, Carpenter J. jomo: A Package for Multilevel Joint Modelling Multiple Imputation. 2019. <https://CRAN.R-project.org/package=jomo>. Accessed 17 Feb 2020.
22. Bartlett J, Keogh R. smcfcs: Multiple Imputation of Covariates by Substantive Model Compatible Fully Conditional Specification. 2019. R package version 1.4.0. <https://CRAN.R-project.org/package=smcfcs>. Accessed 17 Feb 2020.
23. Ali HR, Provenzano E, Dawson S-J, Blows FM, Liu B, Shah M, Earl HM, Poole CJ, Hiller L, Dunn JA, Bowden SJ, Twelves C, Bartlett JMS, Mahmoud SMA, Rakha E, Ellis IO, Liu S, Gao D, Nielsen TO, Pharoah PDP. Association between CD8+ T-cell infiltration and breast cancer survival in 12 439 patients. *Ann Oncol.* 2014;25(8):1536–43.
24. McCabe EL, Larson MG, Lunetta KL, Newman AB, Cheng S, McCabe EL, Larson MG, Lunetta KL, Newman AB, Cheng S, Murabito JM. Association of an Index of Healthy Aging With Incident Cardiovascular Disease and Mortality in a Community-Based Sample of Older Adults. *J Gerontol Ser A Biol Sci Med Sci.* 2016;71(12):1695–701.
25. Wood AM, White IR, Thompson SG. Are missing outcome data adequately handled? A review of published randomized controlled trials in major medical journals. *Clin Trials.* 2004;1:368–76.

26. Fiero MH, Huang S, Bell ML. Statistical analysis and handling of missing data in cluster randomized trials: a systematic review. *Trials*. 2016; 17(72). 10.1186/s13063-016-1201-z.
27. Bell ML, Fiero M, Horton NJ, Hsu C-H. Handling missing data in RCTs; a review of the top medical journals. *BMC Med Res Methodol*. 2014;14(1):118.
28. Powney M, Williamson P, Kirkham J, Kolamunnage-Dona R. A review of the handling of missing longitudinal outcome data in clinical trials. BioMed Central Ltd. 2014. 10.1186/1745-6215-15-237.
29. Karahalios A, Baglietto L, Carlin JB, English DR, Simpson JA. A review of the reporting and handling of missing data in cohort studies with repeated assessment of exposure measures. *BMC Med Res Methodol*. 2012;12(1):96.
30. Altman Douglas G, Royston Patrick. The cost of dichotomising continuous variables. *BMJ*. 2006;332(7549):1080.1.
31. van Maaren MC, de Munck L, Jobsen JJ, Poortmans P, de Bock GH, Siesling S, Strobbe LJA. Breast-conserving therapy versus mastectomy in T1-2N2 stage breast cancer: a population-based study on 10-year overall, relative, and distant metastasis-free survival in 3071 patients. *Breast Cancer Res Treat*. 2016;160(3):511–21.
32. Wood AM, White IR, Royston P. How should variable selection be performed with multiply imputed data? *Stat Med*. 2008;27:3227–46.
33. HARRELL F. *Regression Modeling Strategies: with Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis*. Switzerland: Springer International Publishing; 2016.
34. VanderWeele TJ. Principles of confounder selection. *Eur J Epidemiol*. 2019;34(3):211–9.
35. Austin PC. Statistical power to detect violation of the proportional hazards assumption when using the Cox regression model. *J Stat Comput Simul*. 2018;88(3):533–52.
36. Zhang Z. Missing data imputation: focusing on single imputation. *Ann Trans Med*. 2016;4(1):9.
37. Bartlett JW, Harel O, Carpenter JR. Asymptotically Unbiased Estimation of Exposure Odds Ratios in Complete Records Logistic Regression. *Am J Epidemiol*. 2015;182(8):730–6.
38. Carpenter JR, Kenward MG. *Multiple Imputation and Its Application*: Wiley; 2013. <https://www.wiley.com/en-gb/Multiple+Imputation+and+its+Application+-p-97804707>
39. von Hippel PT. How Many Imputations Do You Need? A Two-stage Calculation Using a Quadratic Rule. *Sociol Methods Res*. 2018; 004912411774730. 10.1177/0049124117747303.

40. Morris TP, White IR, Carpenter JR, Stanworth SJ, Royston P. Combining fractional polynomial model building with multiple imputation. *Stat Med.* 2015;34(25):3298–317.
41. Seaman SR, Bartlett JW, White IR. Multiple imputation of missing covariates with non-linear effects and interactions: an evaluation of statistical methods, *BMC Med Res Methodol.* 2012;12(1):46.
42. De Silva AP, Moreno-Betancur M, De Livera AM, Lee KJ, Simpson JA. A comparison of multiple imputation methods for handling missing values in longitudinal data in the presence of a time-varying covariate with a non-linear association with time: a simulation study. *BMC Med Res Methodol.* 2017;17(1):114.
43. Moreno-Betancur M, Carlin JB, Brilleman SL, Tanamas SK, Peeters A, Wolfe R. Survival analysis with time-dependent covariates subject to missing data or measurement error: Multiple Imputation for Joint Modeling (MIJM) *Biostatistics.* 2018;19(4):479–96.
44. Sauerbrei W, Royston P, Look M. A new proposal for multivariable modelling of time-varying effects in survival data based on fractional polynomial time-transformation, *Biom J.* 2007;49(3):453–73.
45. Buchholz A, Sauerbrei W. Comparison of procedures to assess non-linear and time-varying effects in multivariable models for survival data. *Biom J.* 2011;53(2):308–31.
46. Heinzl H, Kaider A. Gaining more flexibility in Cox proportional hazards regression models with cubic spline functions. *Comput Methods Prog Biomed.* 1997;54(3):201–8.
47. Wynant W, Abrahamowicz M. Impact of the model-building strategy on inference about nonlinear and time-dependent covariate effects in survival analysis. *Stat Med.* 2014;33(19):3318–37.
48. Abrahamowicz M, MacKenzie TA. Joint estimation of time-dependent and non-linear effects of continuous covariates on survival. *Stat Med.* 2007;26(2):392–408.
49. Yan J, Huang J. Model Selection for Cox Models with Time-Varying Coefficients. *Biometrics.* 2012;68(2):419–28.
50. Bellera CA, MacGrogan G, Debled M, de Lara CT, Brouste V, Mathoulin-Pélissier S. Variables with time-varying effects and the Cox model: some statistical concepts illustrated with a prognostic factor study in breast cancer. *BMC Med Res Methodol.* 2010;10:20.
51. Abrahamowicz M, Mackenzie T, Esdaile JM. Time-Dependent Hazard Ratio: Modeling and Hypothesis Testing With Application in Lupus Nephritis. *J Am Stat Assoc.* 1996;91(436):1432–39.

- 52.Hess KR. Assessing time-by-covariate interactions in proportional hazards regression models using cubic spline functions. *Stat Med.* 1994;13(10):1045–62.
- 53.MacCallum RC, Zhang S, Preacher KJ, Rucker DD. On the Practice of Dichotomization of Quantitative Variables. *Psychol Methods.* 2002;7(1):19–40.
- 54.Schemper Michael. *Multivariable model-building: A pragmatic approach to regression analysis based on fractional polynomials for modelling continuous variables.* Patrick Royston and Willi Sauerbrei, Wiley, Chichester, 2008. No. of pages: 322. Price: \$130.00. ISBN: 978-0-470-02842-1. *Statistics in Medicine.* 2009;28(3):537–538.
- 55.Heinze G, Wallisch C, Dunkler D. Variable selection-A review and recommendations for the practicing statistician. *Biom J Biom Z.* 2018;60(3):431–49.
- 56.Binder N. The gap between development of time-to-event methods and their application in epidemiology. In: *Survival Analysis for Junior Researchers: 2019.* <https://publicifsv.sund.ku.dk/~safjr2019/NadineBinderAbstract.pdf>.

Supplementary information: Additional file 1

This file contains the search terms used, the checklist for data extraction and the list of papers included in the review. The contents of this file are available in AppendixD.

Supplementary information: Additional file 2

This file is stored as an excel spreadsheet and is available from the supplementary information section on the BMC Medical Research Methodology website (<https://doi.org/10.1186/s12874-020-01018-7>).

11.1 Relevant material not included in the finalised journal article

While the above publication summarised the majority of the results, additional information was extracted from the articles included in the review. This included whether the aim of the study was to develop a prediction model for the outcome or to explore associations between exposures and outcome. I refer to the second of these study aims as ‘inference modelling’. While this additional information did not make it into the finalised paper, it is still relevant to the PhD due to the changed focus of the thesis towards prediction.

11.1.1 Prediction modelling and Inference modelling

Of the 148 studies included in the review, 18 focused on prediction modelling and 130 focused on inference modelling. The aims of the ‘inference modelling’ studies may have been to estimate causal effects (focusing on one of the exposures with adjustments for confounders), or may have been more exploratory in nature. Many traditional epidemiological studies are of this second type, and their aim is not always clear.

Of the 18 papers focusing on prediction models, 17 used multivariable modelling and, therefore, had to decide on a covariate selection method. The one other prediction paper was focused on the assessment of an already developed prediction model. Of the 130 papers focusing on inference, 127 used multivariable modelling and therefore had to determine how to include covariates into their final model.

For the 17 studies building a multivariable prediction model, 59% of studies used stepwise selection to select covariates to be included in their prediction model compared to 17% of studies for inference. Three (19%) studies did not state clearly how they determined the set of covariates to be included in the model compared to 56 (44%) of inference studies.

Table 11.1: Comparing covariate selection for predictive and inference modelling

| Selection | Inference models ($N = 127$) ¹ | Predictive models ($N = 17$) ^{1,2} |
|---------------------------------------|--|--|
| Stated selection procedure (%) | | |
| Stepwise regression | 71 (56) | 14 (82) |
| Univariate analysis | 21 (17) | 10 (59) |
| A priori | 29 (23) | 4 (25) |
| Log-rank test | 30 (23) | 3 (19) |
| | 5 (4) | 0 (0) |

¹ Due to overlap of methods, percentages do not sum to 100

² 1 study kept significant covariates in multivariable analysis

11.1.2 Validation methods used in prediction modelling

Of the 18 prediction models included in the study, seven (39%) used apparent validation. Six (33%) used external validation, of which two also used internal validation methods, one used the *standard* bootstrap optimism-corrected method and the other used the 0.632+ algorithm. Five (28%) studies in total used internal validation using a bootstrap algorithm (four *standard* and one 0.632+). Two (11%) studies split the data into training and test sets. One repeatedly split the data randomly 1000 times while the other split the data into two, one split for training and testing, the other for validation.

Of the 18 prediction models, five also used multiple imputation either as the main missing data method or a sensitivity analysis. One study pooled each of the imputed datasets' analysis model to get one overall model. Predictions were then obtained from the pooled model when using apparent validation. One paper used multiple imputation as a sensitivity analysis to the model originally developed (presumably using complete-case analysis) using 0.632+ validation and external validation. Two papers developed a prediction model using multiple imputation and then validated the model on an external cohort. However, details on whether a final pooled prediction model was used versus pooling the predicted values of the M prediction models were not available for either study. One study used multiple imputation (including the event indicator in the imputation model) and used the bootstrap *standard* algorithm for validation. They combined multiple imputation and the *standard* bootstrap algorithm using *MI-then-BS* as they stated that they bootstrapped each of the imputed datasets but gave no reasoning for why they had combined MI and bootstrapping in this manner.

11.1.3 Discussion of the additional information

The primary aim of this systematic review was not focused solely on prediction modelling, hence why only a small number of studies included used prediction modelling. Even with this small number of studies, it is possible to see that a range of approaches were used to handle missing data.

Of the papers included in the review which focused on prediction, the majority used an automated way of including covariates into their model such as stepwise regression (either forward or backward selection). In addition, the majority of papers (61%) used a validation method to assess their model, either external or internal validation. Clear guidance on the best way to combine missing data methods with validation, including when there is a need to account for other aspects of the study such as covariate selection, will help to improve prediction model development. This will help to avoid any potential data leakage which could arise from the methods selected to handle the missingness.

12 Discussion

12.1 Summary of the key results

Missing data is a common issue faced by researchers in many different health research settings. When developing a prediction model the presence of missing values can lead not only to challenges in fitting a prediction model, but also in validating it.

12.1.1 Combining internal validation with multiple imputation

The main aim of this thesis was to propose and assess how to combine multiple imputation, a method used to handle missing data, with internal validation algorithms in order to assess the performance of prediction models. The two main internal validation algorithms used here are 10-fold cross-validation and the optimism-corrected bootstrap algorithms.

This project was motivated by a missing data discussion proposed by Professor Angela Wood regarding the best way to combine multiple imputation with cross-validation. Limited literature in this field was available (Section 2.3) but lacking in detail and breadth. For example, it was not immediately clear how to impute the data folds used in cross-validation and questions such as “Should all folds be used when drawing imputed values for the training folds?”, “Which folds should be used in order to draw values to impute the test fold?” or “Should the same imputation model be used when drawing plausible values to impute the training folds and test folds?” were unanswered. Similarly for the bootstrap algorithms, questions arose regarding the reuse of imputed datasets at different stages of the bootstrap process, when to apply Rubin’s rules or how the methods would perform in an ideal or pragmatic scenario.

The work conducted in this thesis aimed to thoroughly investigate how to combine multiple imputation with cross-validation or the bootstrap algorithms. I proposed the use of a training and test imputation model. The training imputation model should include all relevant covariates and the outcome. The test imputation model should include all relevant covariates and, potentially, the outcome if the ideal performance is of interest. A primary comparison of interest was whether it is better to validate then impute or impute then validate. Further, I assessed which observations should be included when drawing imputations for the training or test datasets, when to apply Rubin’s rules and how the methods performed in either an ideal or pragmatic setting. Simulation studies with a continuous and binary outcome were used to assess the proposed methods using various performance measures such as the MSE (continuous), AUC (binary), Brier score (binary) and weak calibration (binary).

The proposed methods initially focused on a setting where the covariates included in

the prediction model were ‘known’ and fixed. They were then extended to handle both covariate selection and the flexible transformation of continuous covariates using fractional polynomials. A ‘known’ imputation model is never the case in practice but, here it makes it possible to compare the proposed methods in a clean way and remove unpromising methods for later, more realistic simulation studies. Due to software limitations, not all methods could be examined (Method I in Section 2.6). This was mainly due to an inability to extract the training imputation model to impute missing values in the test set when using non-multivariate normality MI methods.

MI and cross-validation

For cross-validation, I have proposed that the data should first be split into K folds (Section 5.8). Each fold should separately be imputed M times using both a training and test imputation model (Section 2.5). The m^{th} training imputed dataset from each fold $1, \dots, k - 1$ are then combined, and this is repeated across all M imputations, to give M imputed training data sets (containing observations in folds $1, \dots, K$). A prediction model can be fitted in each imputed dataset and then evaluated in each of the M imputed datasets of test fold k . This was known as *Method A* throughout the thesis (described in more detail in Table 2.3).

This method was shown to perform well across many simulated scenarios when the outcome was continuous or binary. I suggest this performance is due to the smaller number of observations available in each fold when imputing which results in the imputed values having increased variability. The prediction model trained on the $k - 1$ training folds is therefore more robust to the increased variability that comes from imputing the k^{th} fold separately. When the proposed methods were evaluated in the context of fractional polynomial selection, method A was also found to have a more variable selection of exponents via ABB compared to the other cross-validation proposed methods. These exponents were then used to impute missing values. This increased variability in either the imputed values or the selection of exponents used for imputation may lead to a more robust prediction model which is better at predicting observations in the test set.

MI and the bootstrap for optimism-correction

For the bootstrapping algorithms (both the *standard* and 0.632 versions), I have proposed that the ‘default’ *BS-then-MI* method should be used (Section 6.9). This imputes the full dataset using a training and test imputation model (Section 2.5) to get training and test imputed datasets which are used to estimate the *apparent* performance. A bootstrap sample is then taken from the partially-observed dataset and imputed using a training imputation model. A prediction model is then fitted to the M bootstrap training imputed datasets. The subsequent steps differ depending on whether the *standard* or 0.632 variation is being used. For (i) the *standard* algorithm the bootstrap sample is imputed again

using a test imputation model. Each bootstrap prediction model fitted to the M bootstrap training imputed datasets is then evaluated in the M bootstrap test imputed datasets in order to estimate the *bootstrap* performance. The imputed test datasets, originally used to estimate apparent performance, can be reused to estimate the test performance. For (ii) the 0.632 algorithm, the partially-observed observations which were not selected into the bootstrap sample are then imputed using a test imputation model. Each bootstrap prediction model fitted to the M bootstrap training imputed datasets is then evaluated in the M not-selected imputed datasets and used to estimate the *test* performance. More thorough details are available in Section 2.7.1.

The *BS-then-MI* method performed well across the majority of the simulation scenarios considered and had performance comparable to *MI-then-BS* when the sample size was large ($n_{obs} = 1000$). Method *MI-then-BS* also had the advantage of data leakage. For *BS-then-MI* and *MI-then-BS* methods, re-using the test imputed datasets (used to estimate *apparent* performance) and restricting them to a smaller dataset (either those in the bootstrap sample to estimate the *bootstrap* performance for the *standard* algorithm or the observations not selected into the bootstrap sample to estimate the *test* performance for the 0.632 algorithm) is not recommended. This is perhaps due to reduced variability of the imputed values (as more observations are used to fit the imputation model) than if the bootstrap sample, or those who were not selected into the bootstrap sample, had been imputed. This reduction in variability may be too great and may provide more optimistic estimates of the *bootstrap* performance for the *standard* bootstrap algorithm or a more optimistic estimate of the *test* performance for the 0.632 version. The estimates of performance post-imputing should not be ‘better’ than the performance estimated if data were fully-observed. This is similar to the comparison of applying Rubin’s rules to the predicted values or to performance estimates in Wood et al. (2015) [38]. Wood et al. found that by applying Rubin’s rules to the predicted values (thus reducing the variance of the predicted value), the estimates of performance became optimistic.

12.1.2 How are missing data handled in practice?

In addition to proposing methods to combine MI and internal validation, I reviewed how researchers handle missing data in practice. The systematic review focused on a survival setting in oncology (survival data was the initial focus of this PhD project) and assessed how researchers conducted and reported analyses in the presence of missing data. This covered both prediction and descriptive/exploratory settings. The review assessed the handling of missing data, covariate selection, the assessment of the functional form and the assessment of the proportional hazards assumption.

Although valid methods have been developed to handle missing data, simpler methods

such as complete-case analysis remained popular. It can be very easy to unknowingly conduct a complete-case analysis as it is the default approach to missing data in many statistical programmes (such as R, Stata and SAS). In my experience of R and Stata, the only indication of the application of a complete-case analysis is the reduced sample size in the model summary. This default can lead to strong assumptions unintentionally being made. A potential improvement to software could be a default message warning users that a complete-case analysis has been used and stating how many observations were dropped from the dataset.

Overall, the reporting and handling of missing data in observational research was found to be poor. It became clear through the course of conducting the review that, while guidelines were in place, adherence was poor. It is easy to make statements saying that researchers should read guideline statements, improve the transparency of their research or give more consideration to the handling of missing data. However, as shown in many other systematic reviews [63,64,65,66] which range from observational studies to clinical trials (involving studies published from 2001-2014) little progress has been made in the handling of missing data. The complete-case analysis has remained the automatic default method. This is despite methods such as MI being made more easily available in statistical software.

It is easy to say “do better!” but this has been said for a long time and as indicated from numerous systematic reviews, nothing is getting better. To quote Altman (1994) [67] “We need less research, better research, and research done for the right reasons”.

If publishers do not enforce the use of guidelines to improve reporting there is little requirement for change by researchers in practice, not only for the handling of missing data but also for the improved reporting and transparency of research. Publishers taking recommended guidelines seriously is the best way to incentivise authors to improve the transparency and quality of their research.

12.1.3 Training and validating a prediction model in practice

The recommended methods for cross-validation and the bootstrap algorithm (0.632 variation) were implemented using a real dataset - the Rotterdam breast cancer dataset (Chapter10). This demonstrated that the methods could be not only implemented in practice but also extended to a survival setting.

The work in this thesis has focused on handling missing data when validating a clinical prediction model (Stage 2 from Section2.1). More research has been conducted for stage 1, which concerns the development of a model when missing data are present, than stages 2 and 3. Previous literature incorporating missing data with other analysis decisions such

as covariate selection [36] or using FPs to assess functional forms and covariate selection [21] are available. These papers have primarily focused on an inference setting where focus is based on the bias of parameters and having one ‘finalised’ model from which to explore associations between covariates and an outcome. In an inference setting, the entire dataset would be imputed and a model developed in each of the M imputed datasets. These models would then be collapsed to get one overall model [21,36].

However, in a prediction setting there is no need to collapse the models into one overall prediction model [38]. Instead, the M prediction models can be used to get a predicted value for new observations, which can then be averaged using Rubin’s first rule. As Rubin’s second rule is not used for this, multivariate normality is less important. The prognostic setting can therefore avoid some of the difficulties that arise in the parameter estimation setting, such as how to get an overall model when models in the M imputed datasets have different covariates or functional forms selected.

Overall, MI can be used to handle missing values in a dataset which is being used to develop a prediction model. Predicted values for new observations will come from a ‘final model’ or algorithm which will consist of M prediction models. The procedure used to develop these M prediction models will then be internally-validated using the recommended methods for cross-validation or the optimism-corrected bootstrap.

Recommendations are now in place for stage 1 (model development), and now also stage 2 (model assessment). At this point, a researcher may wish to publish their prediction model for others to use, for example in a journal. As previously seen in the systematic review in Chapter 11, reporting and transparency of studies are not always optimal. Specifically for prediction models, the TRIPOD statement [7] can be used to guide researchers in clear reporting. With regards to missing data, the TRIPOD statement recommends stating how missing data are handled in the study (such as a complete-case analysis or imputation methods), stating the number of patients with missing data in the baseline table and noting missing data to be a limitation. However, improvement to the TRIPOD statement could be made by also requiring that researchers state any missing data assumptions they have made. In addition, making an explicit statement requiring researchers to explain how they combined their missing data methods with their internal validation algorithms would be helpful for not only study reproducibility but also for assessing how optimistic internal validation results may be.

The remaining difficulty now lies with stage 3 (Section 2.1) which focuses on how to handle missing values in new patients. This is not an issue for the ideal setting in which, by definition, all future patients have fully-observed data, but is a concern for the pragmatic setting (where we expect future observations to have missing data). As stated above, and

in Section 2.6, it is not currently possible to extract the training imputation model and conduct an out-of-sample imputation outside of multivariate normal multiple imputation in R or Stata. This makes it difficult to impute factor or non-normal continuous covariates (a transformation could be considered for the continuous covariates to make them more normally distributed but this is not generally advisable [14, p.74]).

Due to software limitations, stage 3 will be difficult to handle in practice. Current options could involve:

- waiting until sufficient patients have accumulated within the patient population in which new predictions are to be made. This decision is, however, not optimal in healthcare situations in which real-time predictions are needed.
- if access is still available to the training data, the new patients could be ‘added’ to the training dataset. A test imputation model could then be fitted to all of this data to obtain imputed data for these new patients. These ‘fully-observed’ measurements can then be fed into the prediction model. This may not always be appropriate, depending on where the data come from (for example, would one do this if the new individuals were from a different country than the training data?).
- Fletcher and Blume (2020) [46] and Hoogland et al. (2020) [47], previously discussed in Section 2.3, focused on stage 3 and suggested the use of submodels. Hoogland et al. (2020) create many submodels of the original prediction model. These submodels will range from containing only one covariate to containing all but one covariate. In this way, for any situation involving missing values in covariates, there should be a developed submodel which will only require the observed covariates of a new patient to predict an outcome. Fletcher and Blume use pattern mixture modelling which fits a submodel using data from a specific missing data pattern. Submodels were not explored in this thesis so their performance when combined with internal validation algorithms is currently unknown.
- Nijman et al. (2021) [68] have most recently investigated the use of MI for real-time imputation of missing values in newly observed patients. Appendix A details how Nijman et al. used MI in R to impute new patients and Appendix D provides sample code.

12.1.4 What does this thesis contribute to the field?

This thesis has extended the existing literature in several ways:

- I conducted a systematic review assessing the handling of missing data in a survival setting. The review demonstrated poor adherence to published guidelines. I also provided a summary of various issues faced in a statistical analysis and provided references to relevant published literature to help researchers.
- I investigated how to combine MI with internal validation algorithms. This included proposing several different methods for both cross-validation and the optimism-corrected bootstrap algorithm.
- I have recommended that two imputation models should be used (training - includes the outcome; test - may include the outcome if an ideal setting is of interest). These two models should be used even if *MI-then-Validate* is chosen by the researcher (where traditionally the data are imputed using one imputation model).
- I have provided recommendations on the best way to combine MI with cross-validation (impute each fold separately) and with the bootstrap (*BS-then-MI*). Both recommendations belong to the class *Validate-then-MI* which is also supported by previous published literature [40,44]
- I linked the concept of ‘data leakage’, which is commonly used in the prediction setting, to the concept of missing data and the imputation of missing values.

12.2 Limitations

The majority of this thesis focused on extensive simulation studies to evaluate the various proposed methods. However, despite the large number of data-generating mechanisms used, it is not possible to cover all potential settings that could be faced in practice. Further, because of the large number of methods investigated, the data-generating mechanisms I was able to explore were limited. I was able to identify the methods that are most promising out of a large range of potential methods. However, simulation studies involving a range of more complex, realistic scenarios are needed to evaluate how the methods perform and where they can potentially perform badly.

I coded the simulation studies, including the data-generation and coding of methods, by myself which could lead to undetected coding errors. However, the code for the combination of MI with the *standard* bootstrap optimism-correction algorithm were double-checked by a fellow PhD student, Patrick Rockenschaub, who I collaborated with regarding internal-validation in the presence of missing data [69]. In addition, there is no one way to use a simulation study to investigate particular methods. The same investigation by

another researcher could have had a different data-generating structure than those considered in this thesis [70] i.e. another researcher may have made different choices than the ones I made in this thesis.

The missing data scenarios evaluated in this thesis assumed data are either MCAR or MAR - these are scenarios in which MI can perform well. Robustness of the proposed methods will depend to some extent on these assumptions (i.e. missing not at random), but the extent is unknown. In addition, due to the nature of simulated data, I was able to choose which covariates should be included in the imputation model to give the best imputed values. In practice, an imputation model must be specified by the researcher. How well the method will perform when using a misspecified imputation model has not been tested. However, this is the nature of the pragmatic setting and cannot be avoided. Note that the test imputation model excluding the outcome is misspecified and uncongenial to the analysis procedure, though this concept was developed for the setting of parameter estimation and it is not clear that this matters in the prediction context. In referencing misspecification here, I refer to other covariates which should have been included but were not, or other aspects of model misspecification (for example: transformations, interactions).

In some situations, the complete-case analysis performed well compared to the proposed imputation models (such as MCAR or covariate-dependent MAR scenarios when selecting fractional polynomials, Section 9.3). The proposed methods are recommended for scenarios in which complete-case analysis is known to perform poorly (or where a prediction will be made for someone with partially observed data). However, given that in practice we will never truly know the underlying mechanism behind missing data, the decision of when to use MI is left to the researcher.

A previously discussed limitation concerns software issues. Throughout the simulation study analysis and initial write-up stages of my PhD, it was not possible to extract the parameters of imputation models from the ‘mice’ package in R [49], as discussed in Section 2.6. This meant that not all potential methods could be evaluated in this thesis. This issue was previously discussed with Patrick Rockenschaub (PhD student/a collaborator) when discussing how to combine MI and the optimism-corrected bootstrap for his dataset of interest. Patrick took recommended steps from Stev Van Buuren on extracting imputation parameters and produced a rough draft of code [ongithub\[71\]](#). As of 2021, Patrick has amended this code and it is now a feature of the ‘mice’ package to allow out-of-sample MI.

For cross-validation, I investigated combining MI with K -fold cross-validation. In practice, it is recommended to use repeated K -fold cross-validation [23, p.301]. This involves repeat-

ing the cross-validation procedure multiple times but is more computationally-intensive, especially when combined with MI. A method which works well for K -fold cross-validation should translate well, in terms of performance, when repeating cross-validation multiple times. Additionally, another computationally taxing issue, involving both the bootstrapping and cross-validation algorithms, is estimating confidence intervals for the performance estimate. In this thesis, I focused on proposing methods to obtain an overall point estimate of performance. To estimate confidence intervals, bootstrapping can be used [72] and the proposed methods can be repeated within each bootstrap sample. Depending on the number of bootstrap samples to be used to estimate the confidence interval, in addition to the number of imputed datasets used when multiply imputing, and either the number of bootstrap samples for optimism-correction or the number of times cross-validation will be repeated could all lead to long run-times on computers and difficult to use in practice.

Finally, when developing a prediction model an appropriate sample size is required [25, 26,28]. This in turn is also important for internal validation algorithms, in particular the recommended cross-validation method A, which involves imputing each fold separately. In Chapters4and5, method A (impute each fold separately) was “stress” tested when the sample size was 100. As 10-folds were used, each fold contained 10 observations, which initially lead to errors when, for example, 9 out of the 10 observations in a fold contained missing data. A sufficiently large sample size is required to ensure there are sufficient fully-observed individuals in each fold to use MI.

12.3 Future extensions

The finding presented in this thesis suggest a number of avenues for future research. During my investigation into assessing which methods performed well using simulation studies, it was not possible to evaluate all methods due to software limitations. The MICE package in R [49] now allows a previously fitted imputation model to be fitted to new data. Methods which previously could not be examined (such as Section2.6, Method I) can now be evaluated and compared to the methods examined in this thesis. Future work can involve a comparison of the currently recommended methods with methods which involve using the imputation model fitted to the training data to impute missing data in the test data. Due to the number of methods considered in this thesis, the simulation studies I used involved simple analysis scenarios with only two covariates. Future work can involve evaluating the methods in more realistic and complex situations.

Chapter9showed that the lack of an origin-shift post-imputation could lead to large estimates of the MSE performance. This was due to small values in covariate X_1 post-imputation. The application of an origin-shift transformation helped to improve the fit of the prediction models and produced smaller values of the MSE. Currently the application

of an origin-shift transformation is manual when using fractional polynomials i.e. it must be considered and applied by the user. Arguably, the application of an origin-shift transformation to imputed data could be considered as a tuning parameter concerning the fit of a prediction model. Future work could involve investigating the tuning of an origin-shift parameter post-imputation. This would involve investigating the values of δ (equation 7.1, Section 7.6) which are used to shift the origin and whether the minimum and maximum values of the covariate should be allowed to change or remain fixed across imputed datasets. In addition, further exploration concerning the influence of large imputed values in the fitted prediction models is required. This should also involve assessing how to reduce the influence of these large values, for example, by applying a transformation that will rescale the covariate.

When the proposed methods were extended using fractional polynomials to handle covariate and functional form selection, the methods were restricted to using fractional polynomials of degree 1 [21]. Additional future work could involve extending the proposed methods to handle fractional polynomials of degree 2 or to handle time-varying effects in a survival setting (by adapting the MFP Time (MFPT) algorithm to handle MI and internal validation).

For the pragmatic setting in this thesis, I investigated how to combine internal validation with MI. Another potential method which could work well is the use of partial prediction models (or submodels) [38,47]. This has a resemblance to the missing indicator method in that each missing data pattern will have a separate model. An interesting future project could involve comparing the methods provided in this thesis with methods combining internal validation with partial prediction models.

Another direct extension of the work conducted in this thesis would be to investigate the best way to impute missing values when using all available observations to develop the M prediction models for use in future patients (Stage 1, Section 2.1). I have previously suggested that method A has performed well across many of the data-generating scenarios due to increased variability in either the exponent selection via ABB (for fractional polynomials) or in the imputed values themselves. If this does lead to a more robust prediction model when internally-validating, it is worth considering whether to impute the dataset in a similar manner when fitting the final prediction model for future use.

Future sensitivity analyses could also be investigated to assess how well the methods will perform when the MCAR or MAR assumptions are violated.

In the Rotterdam demonstration chapter (Chapter 10), I noted that the sample size calculations for training a prediction model were based on having fully-observed data (Section

10.4.1). In clinical trials, sample size calculations can be adjusted for issues such as loss to follow-up. Assuming MCAR, the estimated sample size is multiplied by $\frac{1}{1-p}$ where p is the proportion of patients who will not contribute to the analysis (due to loss to follow-up), if a complete-case analysis is to be used. Extensions to the work conducted by Riley et al. [25,26,28] on the minimum sample size for developing a clinical prediction model could include accounting for missing data. For example, would the loss to follow-up adjustment be sufficient? An additional extension could also account for the internal validation algorithm which will be used. An example for why this may be necessary was discussed in the limitations, Section 12.2. I noted that a minimum sample size within each fold was required for *CV-then-MI* method A (imputing each fold separately) when missing data are present.

As noted in Chapter 11 (the systematic review chapter) and Section 12.1.2 above, adherence to guidelines and the handling of missing data were found to be poor in practice. An interesting project, outside of methodological work, could involve discussions with journals surrounding recommended minimum requirements for publication of applied research. In addition, specifically with regards to missing data, improvements could be made in statistical software such as R, Stata and SAS. This could involve outputting a warning when fitting a model after applying a complete-case analysis. The warning could involve the sample size before and after using a complete-case analysis and stating the assumptions that are being made from applying this method. For example: “Complete-case analysis assumes that missingness is independent of <outcome variable> given <list of covariates>.”

12.4 Conclusions

Overall, in this thesis I have investigated how to best combine MI with internal validation algorithms. I have concluded that *Validate-then-MI* is the appropriate way to combine the handling of missing data with cross-validation or bootstrapping. It avoids data leakage and therefore removes any optimism in performance estimates when imputing missing values.

It is my overall hope that the work in this thesis will provide useful guidance to researchers who are looking to assess the performance of a prediction model when values are missing. In addition, I hope that it will act as a guide to any future guidelines put in place for the handling of missing data in predictive research when validating a prediction model.

Bibliography

- [1] J. A. C. Sterne, I. R. White, J. B. Carlin, M. Spratt, P. Royston, M. G. Kenward, A. M. Wood, and J. R. Carpenter, “Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls,” *BMJ (Clinical research ed.)*, vol. 338, p. 2393, jun 2009.
- [2] G. S. Collins, S. Mallett, O. Omar, and L. M. Yu, “Developing risk prediction models for type 2 diabetes: A systematic review of methodology and reporting,” *BMC Medicine*, vol. 9, pp. 1–14, sep 2011.
- [3] C. L. A. Navarro, J. A. A. Damen, T. Takada, S. W. J. Nijman, P. Dhiman, J. Ma, G. S. Collins, R. Bajpai, R. D. Riley, K. G. M. Moons, and L. Hooft, “Risk of bias in studies on prediction models developed using supervised machine learning techniques: systematic review,” *BMJ*, vol. 375, p. n2281, oct 2021.
- [4] J. C. Jakobsen, C. Gluud, J. Wetterslev, and P. Winkel, “When and how should multiple imputation be used for handling missing data in randomised clinical trials - A practical guide with flowcharts,” *BMC Medical Research Methodology*, vol. 17, pp. 1–10, dec 2017.
- [5] A. Tsvetanova, M. Sperrin, N. Peek, I. Buchan, S. Hyland, and G. P. Martin, “Missing data was handled inconsistently in UK prediction models: a review of method used,” *Journal of Clinical Epidemiology*, vol. 140, pp. 149–158, dec 2021.
- [6] S. Nijman, A. Leeuwenberg, I. Beekers, I. Verkouter, J. Jacobs, M. Bots, F. Asselbergs, K. Moons, and T. Debray, “Missing data is poorly handled and reported in prediction model studies using machine learning: a literature review,” *Journal of Clinical Epidemiology*, vol. 142, pp. 218–229, feb 2022.
- [7] G. S. Collins, J. B. Reitsma, D. G. Altman, and K. G. Moons, “Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement,” *BMJ*, vol. 350, jan 2015.
- [8] K. L. Masconi, T. E. Matsha, J. B. Echouffo-Tcheugui, R. T. Erasmus, and A. P. Kengne, “Reporting and handling of missing data in predictive research for prevalent undiagnosed type 2 diabetes mellitus: A systematic review,” *EPMA Journal*, vol. 6, pp. 1–11, mar 2015.
- [9] O. U. Carroll, T. P. Morris, and R. H. Keogh, “How are missing data in covariates handled in observational time-to-event studies in oncology? A systematic review,” *BMC Medical Research Methodology*, vol. 20, pp. 1–15, may 2020.
- [10] T. P. Morris, I. R. White, and M. J. Crowther, “Using simulation studies to evaluate statistical methods,” *Statistics in Medicine*, vol. 38, pp. 2074–2102, may 2019.

- [11]P. Royston and W. Sauerbrei, *Multivariable model-building : a pragmatic approach to regression analysis based on fractional polynomials for modelling continuous variables*. John Wiley, 2008.
- [12]J. R. Carpenter and M. G. Kenward, *Multiple imputation and its application*. John Wiley & Sons, 2013.
- [13]R. J. A. Little and D. B. Rubin, *Statistical analysis with missing data*. United States of America: John Wiley & Sons, 2 ed., 2002.
- [14]S. van Buuren, “Flexible Imputation of Missing Data, Second Edition,” *Flexible Imputation of Missing Data, Second Edition*, jul 2018.
- [15]D. B. Rubin, J. Wiley, N. York, C. Brisbane, and T. Singapore, *Multiple Imputation for Nonresponse in Surveys*. Wiley Series in Probability and Statistics, Hoboken, NJ, USA: John Wiley & Sons, Inc., jun 1987.
- [16]I. R. White, P. Royston, and A. M. Wood, “Multiple imputation using chained equations: Issues and guidance for practice,” *Statistics in Medicine*, vol. 30, no. 4, pp. 377–399, 2010.
- [17]D. M. Tompsett, F. Leacy, M. Moreno-Betancur, J. Heron, and I. R. White, “On the use of the not-at-random fully conditional specification (NARFCS) procedure in practice,” *Statistics in Medicine*, vol. 37, pp. 2338–2353, jul 2018.
- [18]J. W. Bartlett, S. R. Seaman, I. R. White, and J. R. Carpenter, “Multiple imputation of covariates by fully conditional specification: Accommodating the substantive model,” *Statistical Methods in Medical Research*, vol. 24, no. 4, pp. 462–487, 2015.
- [19]X. Xie and X.-L. Meng, “Dissecting Multiple Imputation From A Multi-phase Inference Perspective: What Happens When God’s, Imputer’s And Analyst’s Models Are Uncongenial?,” *Statistica Sinica*, vol. 27, pp. 1485–1594, 2017.
- [20]I. R. White and P. Royston, “Imputing missing covariate values for the Cox model,” *Statistics in Medicine*, vol. 28, no. 15, pp. 1982–1998, 2009.
- [21]T. P. Morris, I. R. White, J. R. Carpenter, S. J. Stanworth, and P. Royston, “Combining fractional polynomial model building with multiple imputation,” *Statistics in Medicine*, vol. 34, no. 25, pp. 3298–3317, 2015.
- [22]D. G. Altman, B. Lausen, W. Sauerbrei, and M. Schumacher, “Dangers of Using “Optimal” Cutpoints in the Evaluation of Prognostic Factors,” *JNCI: Journal of the National Cancer Institute*, vol. 86, pp. 829–835, jun 1994.
- [23]E. W. Steyerberg, *Clinical prediction models : a practical approach to development, validation, and updating*. Springer International Publishing, 2 ed., 2019.

- [24]W. Sauerbrei, A. Perperoglou, M. Schmid, M. Abrahamowicz, H. Becher, H. Binder, D. Dunkler, F. E. Harrell, Jr, P. Royston, G. Heinze, and f. T. o. t. S. Initiative, “State of the art in selection of variables and functional forms in multivariable analysis—outstanding issues,” *Diagnostic and Prognostic Research*, vol. 4, dec 2020.
- [25]R. D. Riley, K. I. Snell, J. Ensor, D. L. Burke, F. E. Harrell, K. G. Moons, and G. S. Collins, “Minimum sample size for developing a multivariable prediction model: Part I – Continuous outcomes,” *Statistics in Medicine*, vol. 38, no. 7, pp. 1262–1275, 2019.
- [26]R. D. Riley, K. I. Snell, J. Ensor, D. L. Burke, F. E. Harrell, K. G. Moons, and G. S. Collins, “Minimum sample size for developing a multivariable prediction model: PART II - binary and time-to-event outcomes,” *Statistics in Medicine*, vol. 38, no. 7, pp. 1276–1296, 2019.
- [27]M. van Smeden, K. G. Moons, J. A. de Groot, G. S. Collins, D. G. Altman, M. J. Eijkemans, and J. B. Reitsma, “Sample size for binary logistic prediction models: Beyond events per variable criteria,” *Statistical Methods in Medical Research*, vol. 28, pp. 2455–2474, aug 2019.
- [28]R. D. Riley, J. Ensor, K. I. Snell, F. E. Harrell, G. P. Martin, J. B. Reitsma, K. G. Moons, G. Collins, and M. Van Smeden, “Calculating the sample size required for developing a clinical prediction model,” *The BMJ*, vol. 368, no. March, pp. 1–12, 2020.
- [29]G. Heinze, C. Wallisch, and D. Dunkler, “Variable selection – A review and recommendations for the practicing statistician,” *Biometrical Journal. Biometrische Zeitschrift*, vol. 60, p. 431, may 2018.
- [30]R. D. Riley, K. I. Snell, G. P. Martin, R. Whittle, L. Archer, M. Sperrin, and G. S. Collins, “Penalization and shrinkage methods produced unreliable clinical prediction models especially when sample size was small,” *Journal of Clinical Epidemiology*, vol. 132, pp. 88–96, apr 2021.
- [31]B. Efron and R. Tibshirani, *An introduction to the bootstrap*. Chapman and Hall/CRC, 1st ed., may 1994.
- [32]T. A. Gerds, T. Cai, and M. Schumacher, “The performance of risk prediction models,” *Biometrical Journal*, vol. 50, pp. 457–479, aug 2008.
- [33]F. E. Harrell, K. L. Lee, and D. B. Mark, “Prognostic/Clinical Prediction Models: Multivariable Prognostic Models: Issues in Developing Models, Evaluating Assumptions and Adequacy, and Measuring and Reducing Errors,” *Tutorials in Biostatistics, Statistical Methods in Clinical Studies*, vol. 1, pp. 223–249, 2005.

- [34]B. Van Calster, D. Nieboer, Y. Vergouwe, B. De Cock, M. J. Pencina, and E. W. Steyerberg, “A calibration hierarchy for risk models was defined: from utopia to empirical data,” *Journal of Clinical Epidemiology*, vol. 74, pp. 167–176, jun 2016.
- [35]R. Tibshirani, “Regression Shrinkage and Selection Via the Lasso,” *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 58, pp. 267–288, jan 1996.
- [36]A. M. Wood, I. R. White, and P. Royston, “How should variable selection be performed with multiply imputed data?,” *Statistics in Medicine*, vol. 27, pp. 3227–3246, 2008.
- [37]S. R. Seaman, J. W. Bartlett, and I. R. White, “Multiple imputation of missing covariates with non-linear effects and interactions: an evaluation of statistical methods,” *BMC medical research methodology*, vol. 12, no. 1, p. 46, 2012.
- [38]A. M. Wood, P. Royston, and I. R. White, “The estimation and use of predictions for the assessment of model performance using large samples with multiply imputed data,” *Biometrical journal.*, vol. 57, pp. 614–32, jul 2015.
- [39]A. Miles, “Obtaining Predictions from Models Fit to Multiply Imputed Data,” *Sociological Methods & Research*, vol. 45, pp. 175–185, oct 2015.
- [40]Y. Vergouwe, P. Royston, K. G. Moons, and D. G. Altman, “Development and validation of a prediction model with missing predictor data: a practical approach,” *Journal of Clinical Epidemiology*, vol. 63, pp. 205–214, feb 2010.
- [41]E. W. Steyerberg and Y. Vergouwe, “Towards better clinical prediction models: seven steps for development and an ABCD for validation,” *European Heart Journal*, vol. 35, pp. 1925–1931, aug 2014.
- [42]B. C. Jaeger, N. J. Tierney, and N. R. Simon, “When to Impute? Imputation before and during cross-validation,” *arXiv*, vol. 2010.00718, oct 2020.
- [43]J. Z. Musoro, A. H. Zwinderman, M. A. Puhan, G. ter Riet, and R. B. Geskus, “Validation of prediction models based on lasso regression with multiply imputed data,” tech. rep., 2014.
- [44]S. Wahl, A.-L. Boulesteix, A. Zierer, B. Thorand, and M. Avan De Wiel, “Assessment of predictive performance in incomplete data by combining internal validation and multiple imputation,” *BMC Medical Research Methodology*, vol. 16, p. 144, 2016.
- [45]B. J. A. Mertens, E. Banzato, and L. C. Wreede, “Construction and assessment of prediction rules for binary outcome in the presence of missing predictor data using multiple imputation and cross-validation: Methodological approach and data-based evaluation,” *Biometrical Journal*, vol. 62, pp. 724–741, may 2020.

- [46]S. Fletcher Mercaldo and J. D. Blume, “Missing data and prediction: the pattern submodel,” *Biostatistics*, vol. 21, pp. 236–252, apr 2020.
- [47]J. Hoogland, M. Barreveld, T. P. A. Debray, J. B. Reitsma, T. E. Verstraelen, M. G. W. Dijkgraaf, and A. H. Zwinderman, “Handling missing predictor values when validating and applying a prediction model to new patients,” *Statistics in Medicine*, vol. 39, pp. 3591–3607, nov 2020.
- [48]R. J. A. Little, “Regression With Missing X’s: A Review,” *Journal of the American Statistical Association*, vol. 87, p. 1227, dec 1992.
- [49]S. van Buuren and K. Groothuis-Oudshoorn, “mice: Multivariate Imputation by Chained Equations in R,” *Journal of Statistical Software*, vol. 45, pp. 1–67, dec 2011.
- [50]J. Honaker, G. King, and M. Blackwell, “{Amelia II}: A Program for Missing Data,” *Journal of Statistical Software*, vol. 45, no. 7, pp. 1–47, 2011.
- [51]J. F. Salgado, “Transforming the area under the normal curve (AUC) into cohen’s d, pearson’s rpb, odds-ratio, and natural log odds-ratio: Two conversion tables,” *European Journal of Psychology Applied to Legal Context*, vol. 10, pp. 35–47, jan 2018.
- [52]B. V. Calster, M. van Smeden, B. D. Cock, and E. W. Steyerberg, “Regression shrinkage methods for clinical prediction models do not guarantee improved performance: Simulation study:,” <https://doi.org/10.1177/0962280220921415>, vol. 29, pp. 3166–3178, may 2020.
- [53]K. G. Moons, R. A. Donders, T. Stijnen, and F. E. Harrell, “Using the outcome for imputation of missing predictor values was preferred,” *Journal of Clinical Epidemiology*, vol. 59, pp. 1092–1101, oct 2006.
- [54]J. W. Bartlett and R. A. Hughes, “Bootstrap inference for multiple imputation under uncongeniality and misspecification:,” *Statistical Methods in Medical Research*, vol. 29, pp. 3533–3546, jun 2020.
- [55]J. K. Kim, “A Note on Approximate Bayesian Bootstrap Imputation,” *Biometrika*, vol. 89, no. 2, pp. 470–477, 2002.
- [56]P. Royston and W. Sauerbrei, “Improving the robustness of fractional polynomial models by preliminary covariate transformation: A pragmatic approach,” *Computational Statistics & Data Analysis*, vol. 51, pp. 4240–4253, may 2007.
- [57]G. Ambler and P. Royston, “Fractional polynomial model selection procedures: investigation of type i error rate,” *Journal of Statistical Computation and Simulation*, vol. 69, no. 1, pp. 89–108, 2007.

- [58]W. Sauerbrei, P. Royston, and M. Look, “A new proposal for multivariable modelling of time-varying effects in survival data based on fractional polynomial time-transformation.,” *Biometrical Journal*, vol. 49, no. 3, pp. 453–473, 2007.
- [59]D. R. Cox, “Regression Models and Life-Tables,” *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 34, no. 2, pp. 187–220, 1972.
- [60]M. S. Schröder, A. C. Culhane, J. Quackenbush, and B. Haibe-Kains, “survcomp: An R/Bioconductor package for performance assessment and comparison of survival models,” *Bioinformatics*, vol. 27, pp. 3206–3208, nov 2011.
- [61]R. H. Keogh and T. P. Morris, “Multiple imputation in Cox regression when there are time-varying effects of exposures,” jun 2017.
- [62]E. C. M. Joie Ensor and R. D. Riley, “pmsampsize: Calculates the Minimum Sample Size Required for Developing a Multivariable Prediction Model,” 2020.
- [63]M. H. Fiero, S. Huang, and M. L. Bell, “Statistical analysis and handling of missing data in cluster randomized trials: a systematic review,” *Trials*, vol. 17, no. 72, 2016.
- [64]M. L. Bell, M. Fiero, N. J. Horton, and C.-H. Hsu, “Handling missing data in RCTs; a review of the top medical journals,” *BMC Medical Research Methodology*, vol. 14, p. 118, dec 2014.
- [65]M. Powney, P. Williamson, J. Kirkham, and R. Kolamunnage-Dona, “A review of the handling of missing longitudinal outcome data in clinical trials,” jun 2014.
- [66]A. M. Wood, I. R. White, and S. G. Thompson, “Are missing outcome data adequately handled? A review of published randomized controlled trials in major medical journals,” *Clinical Trials*, vol. 1, pp. 368–376, 2004.
- [67]D. G. Altman, “The scandal of poor medical research,” *BMJ*, vol. 308, pp. 283–284, jan 1994.
- [68]S. W. J. Nijman, T. K. J. Groenhof, J. Hoogland, M. L. Bots, M. Brandjes, J. J. Jacobs, F. W. Asselbergs, K. G. Moons, and T. P. Debray, “Real-time imputation of missing predictor values improved the application of prediction models in daily practice,” *Journal of Clinical Epidemiology*, vol. 134, pp. 22–34, jun 2021.
- [69]P. Rockenschaub, M. J. Gill, D. McNulty, O. Carroll, N. Freemantle, and L. Shallcross, “Development of risk prediction models to predict urine culture growth for adults with suspected urinary tract infection in the emergency department: protocol for an electronic health record study from a single UK university hospital,” *Diagnostic and Prognostic Research 2020 4:1*, vol. 4, pp. 1–9, sep 2020.

- [70]A. L. Boulesteix, S. Hoffmann, A. Charlton, and H. Seibold, “A replication crisis in methodological research?,” *Significance*, vol. 17, pp. 18–21, oct 2020.
- [71]S. van Buuren, “Get at the final model used in the MICE iterations? · Discussion #346 · amices/mice · GitHub.”
- [72]H. Noma, T. Shinozaki, K. Iba, S. Teramukai, and T. A. Furukawa, “Confidence intervals of prediction accuracy measures for multivariable prediction models based on the bootstrap-based optimism correction methods,” *Statistics in Medicine*, vol. 40, pp. 5691–5701, nov 2021.
- [73]R. R. Andridge and R. J. Little, “A Review of Hot Deck Imputation for Survey Non-response,” *International statistical review = Revue internationale de statistique*, vol. 78, p. 40, apr 2010.
- [74]M. L. Wallace, S. J. Anderson, and S. Mazumdar, “A stochastic multiple imputation algorithm for missing covariate data in tree-structured survival analysis,” *Statistics in Medicine*, 2010.
- [75]R. H. Keogh and T. P. Morris, “Multiple imputation in Cox regression when there are time-varying effects of covariates,” *Statistics in Medicine*, jul 2018.
- [76]A. Miles, “Obtaining Predictions from Models Fit to Multiply Imputed Data,” *Sociological Methods and Research*, vol. 45, pp. 175–185, feb 2016.
- [77]J. A. Sterne, M. A. Hernán, B. C. Reeves, J. Savović, N. D. Berkman, M. Viswanathan, D. Henry, D. G. Altman, M. T. Ansari, I. Boutron, J. Carpenter, A.-W. Chan, R. Churchill, A. Hróbjartsson, J. Kirkham, P. Jüni, Y. Loke, T. Pigott, C. Ramsay, D. Regidor, H. Rothstein, L. Sandhu, P. Santaguida, H. J. Schünemann, B. Shea, I. Shrier, P. Tugwell, L. Turner, J. C. Valentine, H. Waddington, E. Waters, P. Whiting, and J. P. Higgins, “The Risk Of Bias In Non-randomized Studies – of Interventions (ROBINS-I) assessment tool (version for cohort-type studies) ROBINS-I tool (Stage I): At protocol stage Specify the review question,” *BMJ*, vol. 355, no. i4919, 2016.
- [78]E. W. Steyerberg, F. E. Harrell, G. J. Borsboom, M. Eijkemans, Y. Vergouwe, and J. F. Habbema, “Internal validation of predictive models: Efficiency of some procedures for logistic regression analysis,” *Journal of Clinical Epidemiology*, vol. 54, pp. 774–781, aug 2001.
- [79]M. Abrahamowicz and T. A. MacKenzie, “Joint estimation of time-dependent and non-linear effects of continuous covariates on survival,” *Statistics in Medicine*, vol. 26, pp. 392–408, jan 2007.

- [80]P. Royston and P. C. Lambert, *Flexible parametric survival analysis using Stata : beyond the Cox model*. Stata Press, 2011.
- [81]M. Schomaker and C. Heumann, “Bootstrap Inference when Using Multiple Imputation,” feb 2016.
- [82]B. Van Calster, D. J. McLernon, M. Van Smeden, L. Wynants, E. W. Steyerberg, P. Bossuyt, G. S. Collins, P. MacAskill, D. J. McLernon, K. G. Moons, E. W. Steyerberg, B. Van Calster, M. Van Smeden, and A. J. Vickers, “Calibration: The Achilles heel of predictive analytics,” *BMC Medicine*, vol. 17, p. 230, dec 2019.
- [83]X. L. Meng, “Multiple-Imputation Inferences with Uncongenial Sources of Input,” *Statistical Science*, vol. 9, pp. 538–558, nov 1994.
- [84]A. Buchholz and W. Sauerbrei, “Comparison of procedures to assess non-linear and time-varying effects in multivariable models for survival data,” *Biometrical Journal*, vol. 53, no. 2, pp. 308–331, 2011.
- [85]J. P. Vandenbroucke, E. von Elm, D. G. Altman, P. C. Gøtzsche, C. D. Mulrow, S. J. Pocock, C. Poole, J. J. Schlesselman, M. Egger, and f. t. S. Initiative, “Strengthening the Reporting of Observational Studies in Epidemiology (STROBE): Explanation and Elaboration,” *PLoS Medicine*, vol. 4, p. e297, oct 2007.
- [86]A. Karahalios, L. Baglietto, J. B. Carlin, D. R. English, and J. A. Simpson, “A review of the reporting and handling of missing data in cohort studies with repeated assessment of exposure measures,” *BMC Medical Research Methodology*, vol. 12, p. 96, dec 2012.
- [87]J. A. Sterne, M. A. Hernán, B. C. Reeves, J. Savović, N. D. Berkman, M. Viswanathan, D. Henry, D. G. Altman, M. T. Ansari, I. Boutron, J. R. Carpenter, A.-W. Chan, R. Churchill, J. J. Deeks, A. Hróbjartsson, J. Kirkham, P. Jüni, Y. K. Loke, T. D. Pigott, C. R. Ramsay, D. Regidor, H. R. Rothstein, L. Sandhu, P. L. Santaguida, H. J. Schünemann, B. Shea, I. Shrier, P. Tugwell, L. Turner, J. C. Valentine, H. Waddington, E. Waters, G. A. Wells, P. F. Whiting, and J. P. Higgins, “ROBINS-I: a tool for assessing risk of bias in non-randomised studies of interventions.,” *BMJ (Clinical research ed.)*, vol. 355, oct 2016.
- [88]M. Quartagno and J. Carpenter, “ {jomo}: A package for Multilevel Joint Modelling Multiple Imputation,” 2018.
- [89]E. Von Elm, D. G. Altman, M. Egger, S. J. Pocock, P. C. Gøtzsche, and J. P. Vandenbroucke, “The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies for the STROBE initiative,” *Lancet*, vol. 370, no. 9596, pp. 1453–57, 2007.

- [90]W. Wynant and M. Abrahamowicz, “Impact of the model-building strategy on inference about nonlinear and time-dependent covariate effects in survival analysis,” *Statistics in Medicine*, vol. 33, no. 19, pp. 3318–3337, 2014.
- [91]P. T. von Hippel, “How Many Imputations Do You Need? A Two-stage Calculation Using a Quadratic Rule,” *Sociological Methods & Research*, p. 004912411774730, jan 2018.
- [92]J. M. Bland, “Sample size for a study of agreement between two methods of measurement,” 2004.
- [93]M. Abrahamowicz, T. Mackenzie, and J. M. Esdaile, “Time-Dependent Hazard Ratio: Modeling and Hypothesis Testing With Application in Lupus Lupus Nephritis,” *Journal of the American Statistical Association*, vol. 91, no. 436, pp. 1432–1439, 1996.
- [94]original by Gareth Ambler and modified by Axel Benner, “mfp: Multivariable Fractional Polynomials,” 2015.
- [95]D. Moher, A. Liberati, J. Tetzlaff, D. G. Altman, and PRISMA Group, “Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement.,” *BMJ (Clinical research ed.)*, vol. 339, p. b2535, jul 2009.
- [96]J. W. Bartlett and R. H. Keogh, “Bayesian correction for covariate measurement error: A frequentist evaluation and comparison with regression calibration,” *Statistical Methods in Medical Research*, vol. 27, pp. 1695–1708, jun 2018.
- [97]J. Yan and J. Huang, “Model Selection for Cox Models with Time-Varying Coefficients,” *Biometrics*, vol. 68, no. 2, pp. 419–428, 2012.
- [98]R Development Core Team, “R: A Language and Environment for Statistical Computing,” 2008.
- [99]P. C. Austin, “Statistical power to detect violation of the proportional hazards assumption when using the Cox regression model,” *Journal of Statistical Computation and Simulation*, vol. 88, no. 3, pp. 533–552, 2018.
- [100]P. Royston and D. G. Altman, “Regression Using Fractional Polynomials of Continuous Covariates: Parsimonious Parametric Modelling,” Tech. Rep. 3, 1994.
- [101]P. Royston and M. K. B. Parmar, “Flexible parametric proportional-hazards and proportional-odds models for censored survival data, with application to prognostic modelling and estimation of treatment effects,” *Statistics in Medicine*, vol. 21, pp. 2175–2197, aug 2002.

- [102]M. Abrahamowicz and T. A. MacKenzie, “Joint estimation of time-dependent and non-linear effects of continuous covariates on survival,” *Statistics in Medicine*, vol. 26, no. 2, pp. 392–408, 2007.
- [103]C. A. Bellera, G. MacGrogan, M. Debled, C. T. de Lara, V. Brouste, and S. Mathoulin-Pélissier, “Variables with time-varying effects and the Cox model: some statistical concepts illustrated with a prognostic factor study in breast cancer.,” *BMC medical research methodology*, vol. 10, p. 20, 2010.
- [104]P. Royston and W. Sauerbrei, “Stability of multivariable fractional polynomial models with selection of variables and transformations: a bootstrap investigation,” *Statistics in Medicine*, vol. 22, pp. 639–659, feb 2003.
- [105]I. Bou-Hamad, D. Larocque, and H. Ben-Ameur, “Discrete-time survival trees and forests with time-varying covariates,” *Statistical Modelling: An International Journal*, vol. 11, pp. 429–446, oct 2011.
- [106]W. Wynant and M. Abrahamowicz, “Impact of the model-building strategy on inference about nonlinear and time-dependent covariate effects in survival analysis,” *Statistics in Medicine*, vol. 33, pp. 3318–3337, aug 2014.
- [107]H. Heinzl and A. Kaider, “Gaining more flexibility in Cox proportional hazards regression models with cubic spline functions,” *Computer Methods and Programs in Biomedicine*, vol. 54, no. 3, pp. 201–208, 1997.

Appendices

A Chapter1: The Introduction

A.1 Multivariable FPs (MFP): Selecting a FP for multiple covariates

For P covariates in a model, the following algorithm states the selection procedure for FPs. All significance tests are from likelihood ratio tests with degrees of freedom dependent on the degree of the models being compared.

1. Choose nominal p-values for α_β , α_E where α_β controls whether a covariate is included in the model and α_E selects the FP function to be fitted for a continuous covariate.
2. Choose maximum degree D for fractional polynomials.
3. The full linear model is fitted with linear predictor: $\beta_0 + \sum_{p=1}^P \beta_p X_p$. The covariates are considered in order of significance from most significant to least: $X_{(1)}, \dots, X_{(P)}$
4. Set cycle counter $c = 0$ and variable counter within each cycle $j = 0$
5. For **categorical/binary** X_j : The joint significance of its dummy variable(s) is tested at α_β significance level. If significant retain in the model, otherwise remove it.
6. For **continuous** X_j the selection procedure is as follows while all other variables, \mathbf{X}_{-j} , to be included in the model are adjusted for:
 - (a) Test the best FP2 model against null model for X_j and tested at α_β level. If significant, keep.
 - (b) Apply Steps (3) - (5) of the FP algorithm. For FP algorithm step (3) $\alpha = \alpha_\beta$ and for step (4) and (5) $\alpha = \alpha_E$. Test the best FP2 model against linear model and against FP1 model with significance α_E .

For subsequent covariates being considered, $X_{>j}$, the current form of X_j is kept.

7. Including or dropping X_j applies until it is reconsidered in cycle $c + 1$.
8. Increment covariate counter, $++ j$. If $j \leq P$ return to step 5/6 to process next variable until iterated through all covariates i.e. $j = P$, then reset $j = 0$ and proceed to step (9).
9. Increment cycle counter $++ c$. The c^{th} cycle is complete.
 - (a) If $c = 1$ return to step (5) - (8).

- (b) If $c \neq$ maximum number of cycles (typically 5) and $c > 1$, check if included covariates and chosen FP functions have changed from cycle $c - 1$ to current cycle c .
- i. If they have changed return to step (5) - (8)
 - ii. If they have not, stop.
- (c) If reached the maximum number of cycles, stop and report that the algorithm has not converged and report current model.

B Chapter3: The simulation set-up

Adjusting the variance to allow for varying R-squared

For $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$, the variance of Y is:

$$\begin{aligned}\text{Var}(Y) &= \beta_1^2 \text{Var}(X_1) + \beta_2^2 \text{Var}(X_2) + 2\beta_1\beta_2 \text{Cov}(X_1, X_2) + \text{Var}(\epsilon) \\ &= \beta_1^2 \text{Var}(X_1) + \beta_2^2 \text{Var}(X_2) + 2\beta_1\beta_2 \text{Cov}(X_1, X_2) + \sigma^2 \\ &= \zeta + \sigma^2\end{aligned}\quad (\text{B0})$$

See (B0). We know that $R^2 = 1 - \frac{\sigma^2}{\text{Var}(Y)}$ and multiplying through by $\text{Var}(Y)$ we get

$$\begin{aligned}R^2 \text{Var}(Y) &= \text{Var}(Y) - \sigma^2 \\ R^2(\zeta + \sigma^2) &= \zeta \\ \sigma^2 R^2 &= \zeta - R^2 \zeta \\ \sigma^2 &= \frac{\zeta - R^2 \zeta}{R^2} \\ \implies \sigma^2 &= \zeta \frac{1 - R^2}{R^2}\end{aligned}$$

where $\zeta = \beta_1^2 \text{Var}(X_1) + \beta_2^2 \text{Var}(X_2) + 2\beta_1\beta_2 \text{Cov}(X_1, X_2)$.

C Chapter6: Write-up of the simulations results when the outcome is binary

In Chapter3I described the design of a simulation study to investigate the performance of various methods which combined MI with an internal validation method. Results for combining MI with cross-validation were then presented and discussed in Chapter4. Results for combining MI with the bootstrap validation algorithms were then presented and discussed in Chapter6for a continuous outcome scenario.

C.1 Introduction

This Appendix chapter will present the results from combining MI with the bootstrap optimism-corrected algorithm for the binary outcome scenario. As investigated in the previous chapter, the impact of data leakage from the imputation process will be assessed, as will the reuse of imputed datasets. The output from the simulation study for the binary outcome will be presented and, due to the quantity of results produced, all graphs will be available in the supplementary plot chapter (SectionS4), in addition to the graphs presented in this chapter.

The bootstrap methods which will be evaluated in this chapter were fully described in Section2.7and were re-summarised in Table6.1. Recall that the methods involved either imputing first (*MI-then-BS*) or taking a bootstrap sample first (*BS-then-MI*). Variations of these methods include reusing previously imputed datasets, re-imputing training imputed datasets or using a fixed bootstrap sample.

In the following section, I will present results for all methods in the simulation study when the AUC is the performance measure of interest. These results will include comparing the estimated AUC from each method with the AUC estimated using the 0.632 algorithm when data are fully-observed. In addition, the methods' estimated AUCs will also be compared with a 'target value' estimated from a larger validation set (details available in Section3.6). The use of an increased number of imputed datasets will be investigated, as will the impact of an increased percentage of missing values. This analysis will then be repeated for the Brier score and calibration intercept and slope. Finally, I will compare data leakage through the imputation process for the *BS-then-MI* and *MI-then-BS* methods before presenting a discussion of the results.

Results from the simulation study

Similarly to the simulation study that combined MI and cross-validation, several factors were varied for the binary outcome setting (including sample size and dependence of missingness). The same simulated data used when evaluating the cross-validation techniques

were used here to evaluate the bootstrap methods. The summary information on the outcome and performance measures for the fully-observed 2000 repetitions can be found in Table 5.1.

Recall from section 3.5 the notation for the averaged estimate of the performance in the fully-observed data (Perf_{obs}) and the larger validation set (Perf_{target}) where Perf denotes the AUC, Brier score, Calibration intercept and slope (previously detailed in Section 1.10). In addition, $\text{Perf}_{prag,imp}$ denotes the pragmatic performance of a proposed method and $\text{Perf}_{ideal,imp}$ denotes the ideal performance where imp denotes various methods such as the complete-case analysis (CC), *BS-then-MI* (BS-MI) or *MI-then-BS* (MI-BS) methods.

The results will be presented for the 0.632 and *standard* algorithms in a similar manner as Chapter 6. For a quick reminder on how results will be compared, a brief overview of the analysis comparing methods *BS-then-MI* and *MI-then-BS impute once* is available in Section 6.2 when the performance measure of interest is the MSE.

C.2 Detailed results for the AUC performance of the 0.632 algorithm

A higher AUC estimate generally suggests the model is performing well. Therefore, if a method overestimates AUC_{obs} (the AUC estimate when data are fully-observed) then the method is considered to be over-optimistic i.e. the model performs better when data have been imputed than if the data had not been missing to begin with.

C.2.1 Comparing results to the AUC estimate when data are fully-observed

MCAR and covariate-dependent MAR

Figure C1 presents results for the various methods to handle missing data alongside bootstrap validation when compared to the AUC estimated when data are fully-observed (AUC_{obs}) i.e. $AUC_{imp} - AUC_{obs}$. The results presented are for the scenario when data are MCAR (top row, $\psi_2 = 0$) or covariate-dependent MAR ($\psi_2 > 0$).

When data are MCAR or covariate-dependent MAR, the AUC estimate from the complete-case analysis tends to underestimate AUC_{obs} ($AUC_{CC} - AUC_{obs} < 0$). The magnitude of this underestimation decreases with increasing sample size. With increasing strength of missingness, the magnitude of the underestimation decreases as can be seen in Figure C1. When data are MCAR and sample size is 100, the magnitude of the difference is approximately 0.003. When data are weak covariate-dependent MAR this increases to 0.005 and the magnitude increases further to 0.013 when data are strong covariate-dependent MAR.

The pragmatic performance of all imputation based methods for all sample sizes when data are MCAR or covariate-dependent MAR underestimates AUC_{obs} with a magnitude of approximately 0.01 ($|AUC_{prag,imp} - AUC_{obs}| \approx 0.01$). When the sample size is 100, method *BS-then-MI* has the largest magnitude of underestimation amongst the imputation methods while method *MI-then-BS impute once* has the smallest. The other imputation methods perform similarly when compared to AUC_{obs} . With increasing sample size all methods tend to perform similarly.

For ideal performance, when data are MCAR or covariate-dependent MAR methods and the sample size is 100 *BS-then-MI*, *MI-then-BS*, *MI-then-BS fixed BS* and *MI-then-BS re-impute* underestimate AUC_{obs} . Method *BS-then-MI* underestimates AUC_{obs} more than the other methods. Methods *BS-then-MI reuseimps*, *MI-then-BS reuse* (with or without fixed bootstrap samples) and *MI-then-BS impute once* overestimate the fully-observed estimate of the AUC, although they tend to underestimate it when sample size increases to 1000. With increasing sample size the magnitude of the under- or overestimation for all methods decreases. For a sample size of 1000 the methods all perform similarly and approximate AUC_{obs} well.

MCAR and covariate-dependent MAR

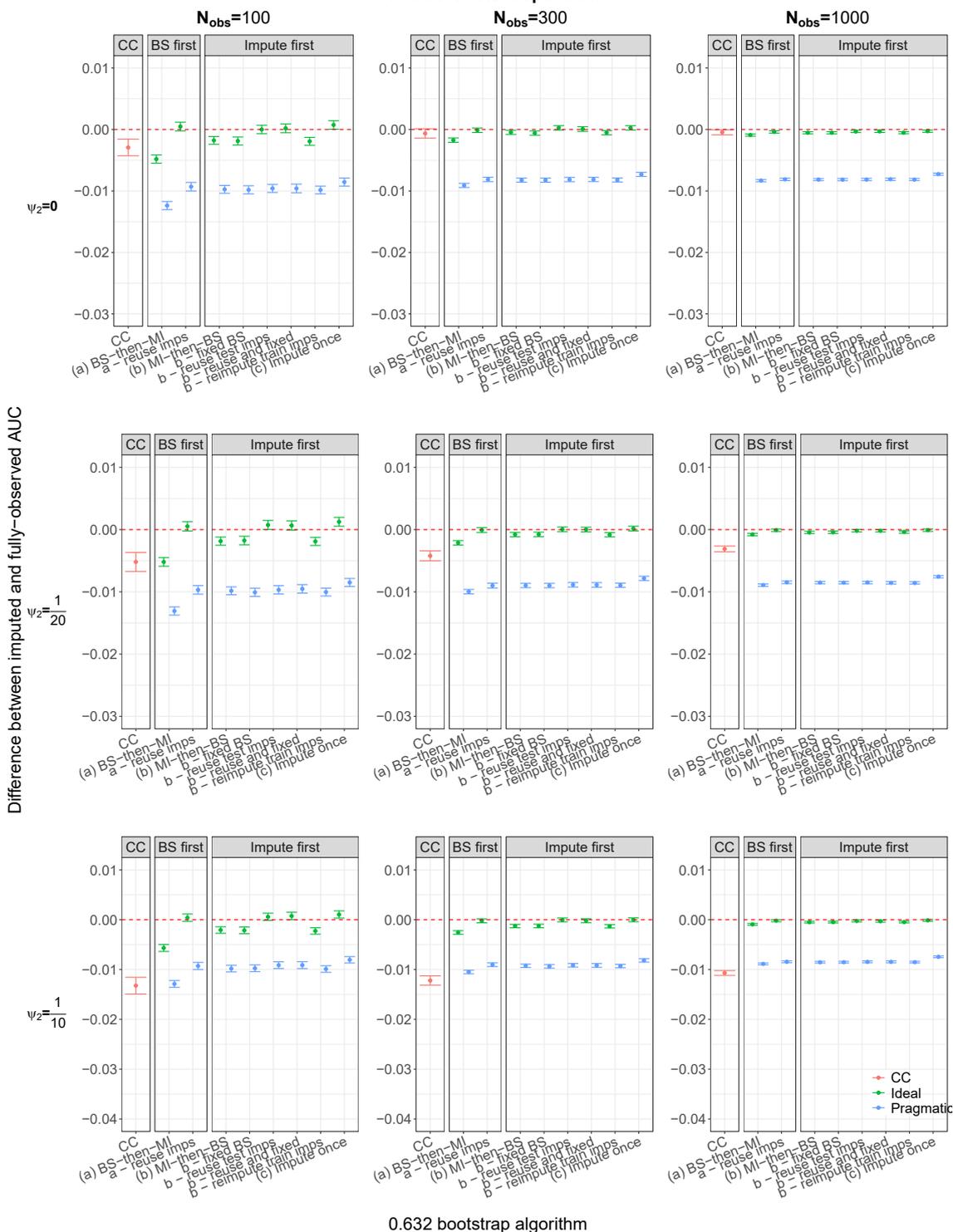


Figure C1: The difference $AUC_{imp} - AUC_{obs}$ when data are MCAR or covariate-dependent MAR for $M = 5$ when 25% of values are missing in X_1 . The error bars summarise results from the 2000 repetitions and the limits represent the Monte Carlo 95% confidence interval of $AUC_{imp} - AUC_{obs}$. CC (complete-case); methods are described in Section 2.7 or Table 6.1.

Outcome-dependent MAR

Figure C2 presents results for $AUC_{imp} - AUC_{obs}$ for the scenario when data are outcome-dependent or outcome- and covariate-dependent MAR.

For all sample sizes and MAR scenarios, the complete-case analysis AUC estimate underestimates AUC_{obs} . When data are outcome-dependent MAR and sample size is 100, the magnitude of the underestimation ($|AUC_{CC} - AUC_{obs}|$) is approximately 0.0025 and this decreases further with increasing sample size. For weak outcome- and covariate-dependent MAR the magnitude is approximately 0.02 and this increases to 0.0375 when data are weak outcome-dependent and strong covariate-dependent MAR.

The pragmatic performance of all methods underestimates the estimate of the AUC when data are fully-observed. Similarly to the MCAR and covariate-dependent MAR scenarios, method *BS-then-MI* tends to underestimate the fully-observed AUC estimate the most while method *MI-then-BS impute once* underestimates it the least. With increasing sample size the methods tend to perform similarly when compared to AUC_{obs} , with method *MI-then-BS impute once* still having the smallest magnitude of the difference.

For ideal performance when sample size is 100 and data are outcome-dependent MAR, methods *BS-then-MI*, *MI-then-BS*, *MI-then-BS fixed BS* and *MI-then-BS re-impute* underestimate the fully-observed AUC estimate. Similarly to the MCAR and covariate-dependent MAR scenarios, methods *BS-then-MI reuse imps*, *MI-then-BS reuse test imps* (with or without fixed bootstrap samples) and *MI-then-BS impute once* overestimate the fully-observed AUC. However, for the outcome- and covariate-dependent MAR when sample size is 100, these methods approximate the fully-observed estimate of the AUC well. With increasing sample size all methods tend to perform similarly and underestimate AUC_{obs} .

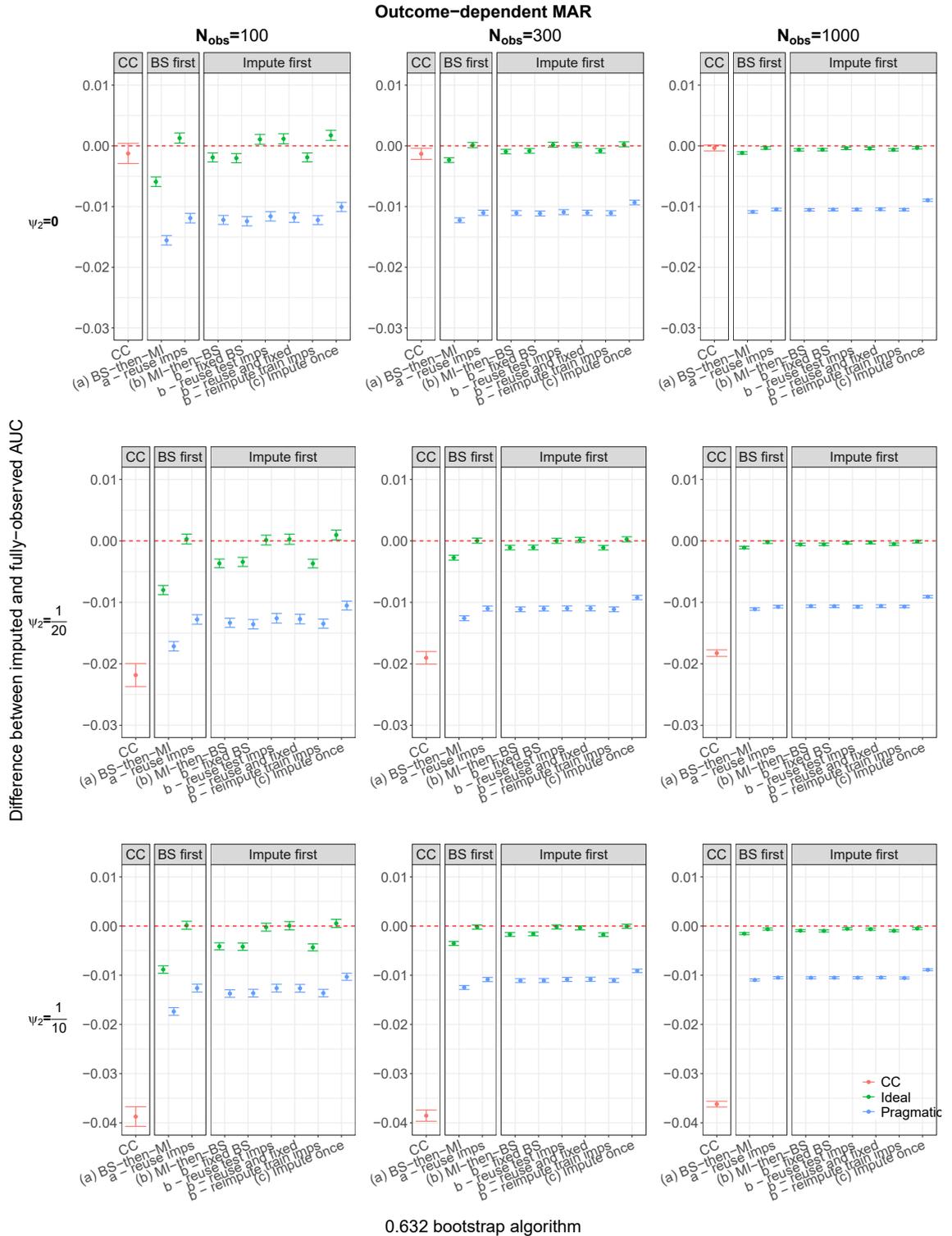


Figure C2: The difference $AUC_{imp} - AUC_{obs}$ when data are outcome-dependent MAR for $M = 5$ when 25% of values are missing in X_1 . The error bars summarise results from the 2000 repetitions and the limits represent the Monte Carlo 95% confidence interval of $AUC_{imp} - AUC_{obs}$. CC (complete-case); methods are described in Section 2.7 or Table 6.1.

C.2.2 Increasing the number of imputed datasets from 5 to 25

Figure C3 displays the results for comparing the various imputation based methods when using 5 or 25 imputed datasets. The results in the graph are for the scenario when data are outcome-dependent MAR but are representative of the results when data are MCAR or covariate-dependent MAR (available in Supplementary Plots S4.4.3). Due to increased computation time when using 25 imputed datasets the comparison a reduced set of methods were assessed. Results are available for methods *BS-then-MI*, *MI-then-BS* and *MI-then-BS impute once* which are based on 1000 repetitions.

The estimates of the AUC for the various methods when using 25 imputed datasets are similar to the performance when only 5 imputed datasets are used.

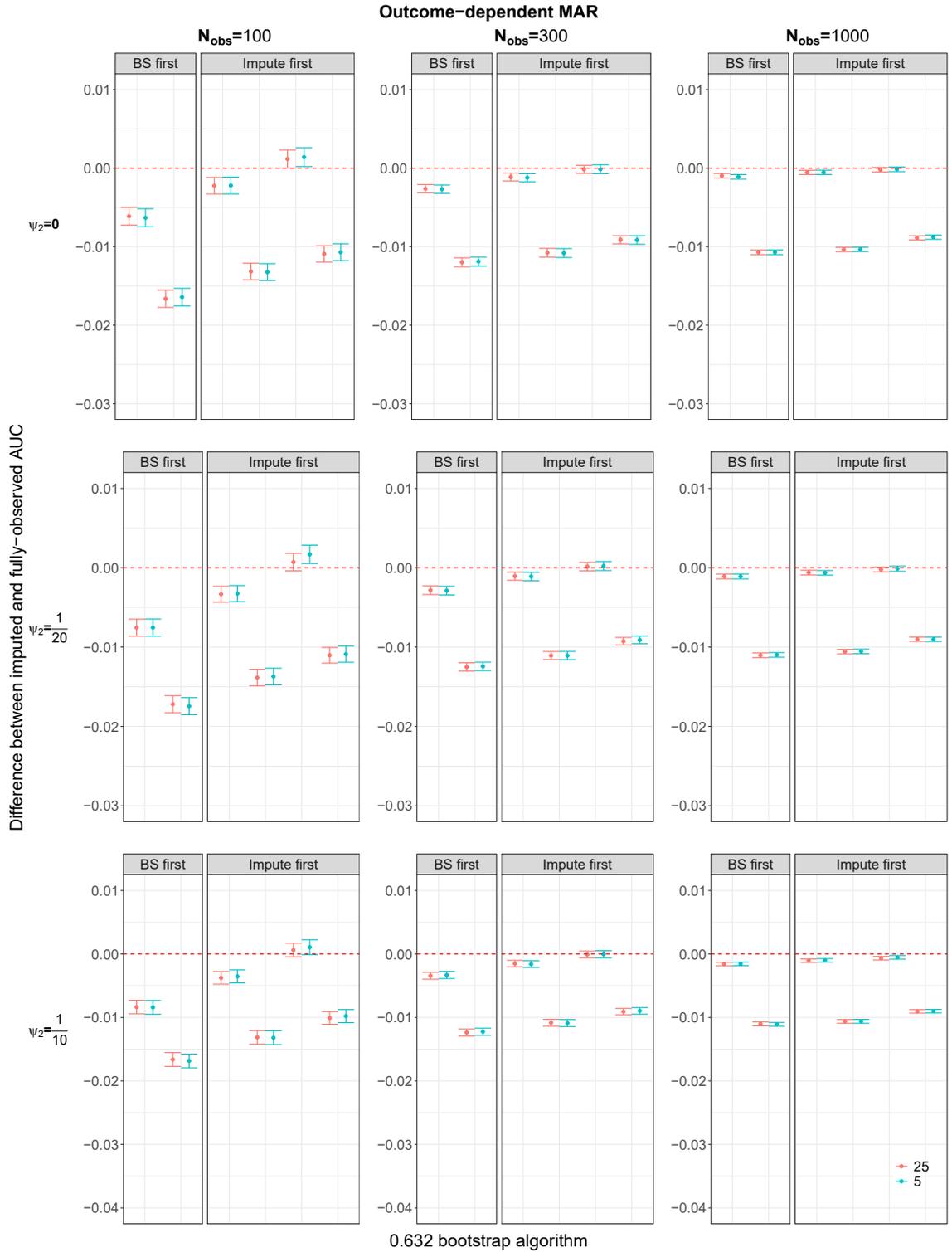


Figure C3: The difference $AUC_{imp} - AUC_{obs}$ when data are outcome-dependent or outcome- and covariate-dependent MAR for $M = 25$ when 25% of values are missing in X_1 . The error bars summarise results from the 2000 repetitions and the limits represent the Monte Carlo 95% confidence interval of $AUC_{imp} - AUC_{obs}$. CC (complete-case); methods are described in Section 2.7 or Table 6.1.

C.2.3 Increasing the percentage of missingness to 40%

Figure C4 displays the results for comparing how the various methods handle an increased percentage of missing values in X_1 from 25% to 40%. The graph presents results for the scenario when data are weak outcome- and covariate-dependent MAR but is representative of all missing data scenarios, additional graphs for ideal and pragmatic performance comparing the percentage of missing values increasing from 25% to 40% can be found in Supplementary Plots Section S4.4.2.

When data are MCAR or outcome-dependent MAR, the complete-case analysis estimate of the AUC when 40% of X_1 values are missing performs similarly to the complete-case analysis when 25% of values are missing. For all other missing scenarios, the complete-case analysis with 40% missing tends to underestimate AUC_{obs} more than the complete-case analysis when 25% of values are missing ($|AUC_{CC,25\%} - AUC_{obs}| < |AUC_{CC,40\%} - AUC_{obs}|$). Similarly, the pragmatic performance of all methods when 40% of values are missing tends to underestimate AUC_{obs} more than the methods do when 25% of values are missing ($|AUC_{prag,imp,25\%} - AUC_{obs}| < |AUC_{prag,imp,40\%} - AUC_{obs}|$) when data are MCAR, or outcome- or covariate-dependent MAR.

For ideal performance, the increased percentage of missingness has caused a slight increase in the magnitude of over- or underestimation ($|AUC_{ideal,imp,25\%} - AUC_{obs}| < |AUC_{ideal,imp,40\%} - AUC_{obs}|$), this increase is at most 0.008. Method *BS-then-MI* tends to have the largest increase in magnitude with increased percentage of missingness across all methods. The Monte Carlo 95% confidence intervals are much larger when the percentage of missingness is 40% and the intervals tend to overlap or encompass the confidence intervals when the percentage of missingness is 25%.

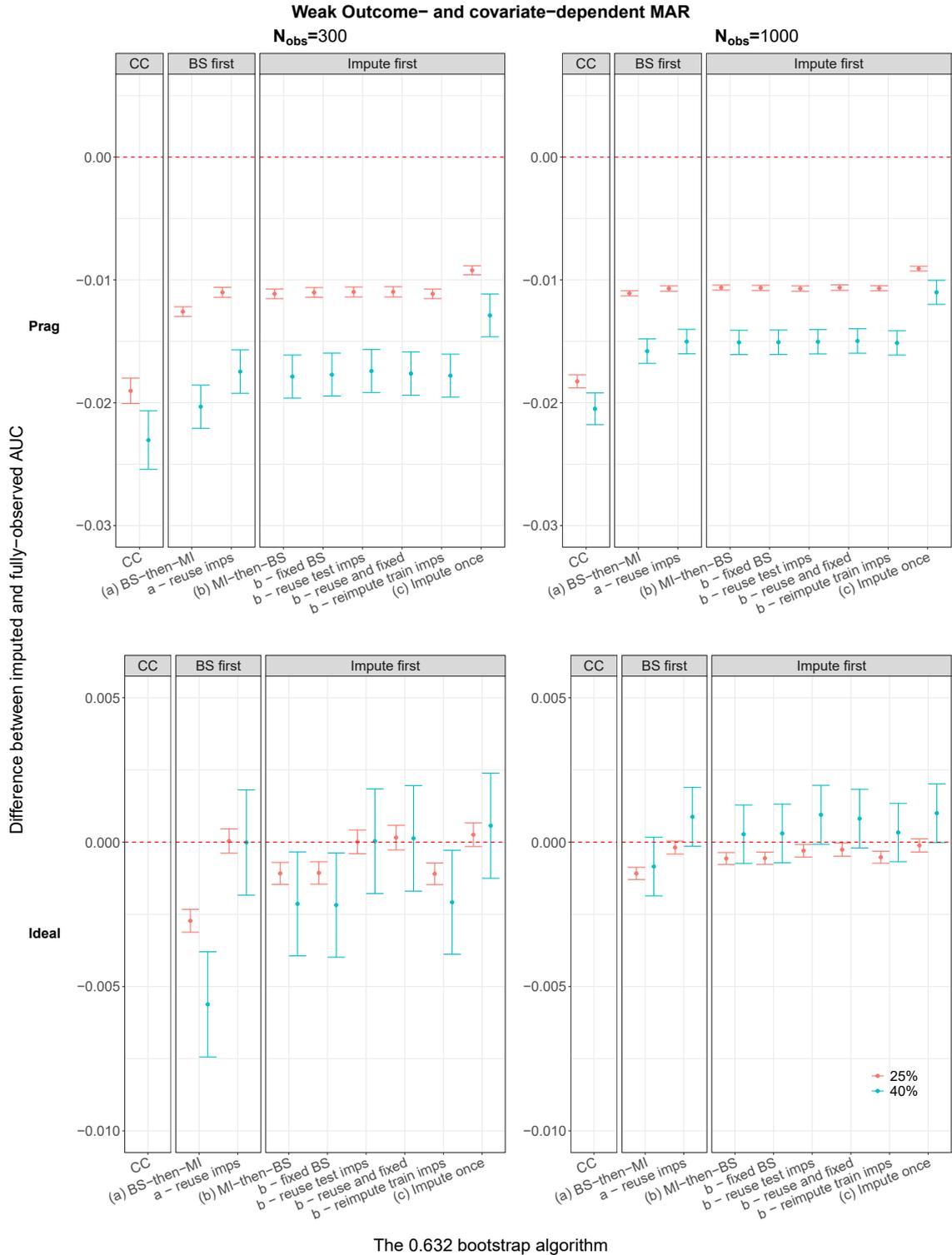


Figure C4: Comparing the impact of increasing the percentage of missingness on the difference $AUC_{imp} - AUC_{obs}$ when data are outcome- and covariate-dependent MAR when $M = 5$. The error bars summarise results from the 2000 repetitions and the limits represent the Monte Carlo 95% confidence interval of $AUC_{imp} - AUC_{obs}$. Red denotes $AUC_{imp} - AUC_{obs}$ when 25% of X_1 values are missing and blue denotes $AUC_{imp} - AUC_{obs}$ when 40% of X_1 values are missing. The top row presents the results for pragmatic performance and the bottom row presents results for ideal performance. CC (complete-case); methods are described in Section 2.7 or Table 6.1.

C.2.4 Comparing results to the target performance

As similarly detailed in the comparison of the bootstrap results for a continuous outcome (Section 6.4.4), the ideal performance of the bootstrap imputation methods and AUC_{obs} were compared to the ideal target AUC estimate. This is estimated by applying a prediction model, developed using all data, to the fully-observed data in the larger test set ($AUC_{target,obs}$). The pragmatic performance of the imputation methods is compared to applying a prediction model, developed using all data, to the imputed datasets of the larger test set ($AUC_{target,imputed}$). The complete-case estimate of the AUC is compared to applying a prediction model to the observed cases of the larger test set ($AUC_{target,CC}$).

MCAR and covariate-dependent MAR

Figure C5 presents results for comparing AUC estimates to their respective ideal, pragmatic or complete-case target AUC estimate when data are MCAR or covariate-dependent MAR. All methods perform well when compared to their target estimate with a magnitude less than 0.005 ($|AUC_{imp} - AUC_{target}| < 0.005$).

When data are MCAR, the complete-case analysis estimate of the AUC tends to overestimate the target complete-case estimate ($AUC_{CC} - AUC_{target,CC} < 0$). When data are covariate-dependent MAR and the sample size is 100, the complete-case analysis approximates the target complete-case estimate of the AUC well but with increasing sample size, it tends to overestimation.

When data are covariate-dependent MAR and sample size is 100, the pragmatic performance of the various methods tends to underestimate the target pragmatic estimate of the AUC ($AUC_{prag,imp} - AUC_{target,imputed} < 0$). Method *BS-then-MI* underestimates the target estimate the most and method *MI-then-BS impute once* overestimates it (i.e. becomes over-optimistic). With increasing sample size the methods perform similarly. When data are weak covariate-dependent MAR and sample size is 300 or 1000, all imputation methods overestimate the target AUC estimate (i.e. become over-optimistic). Method *BS-then-MI* has the smallest magnitude of overestimation when compared to the target AUC and its comparison for ideal performance overlaps with zero when sample size is 300.

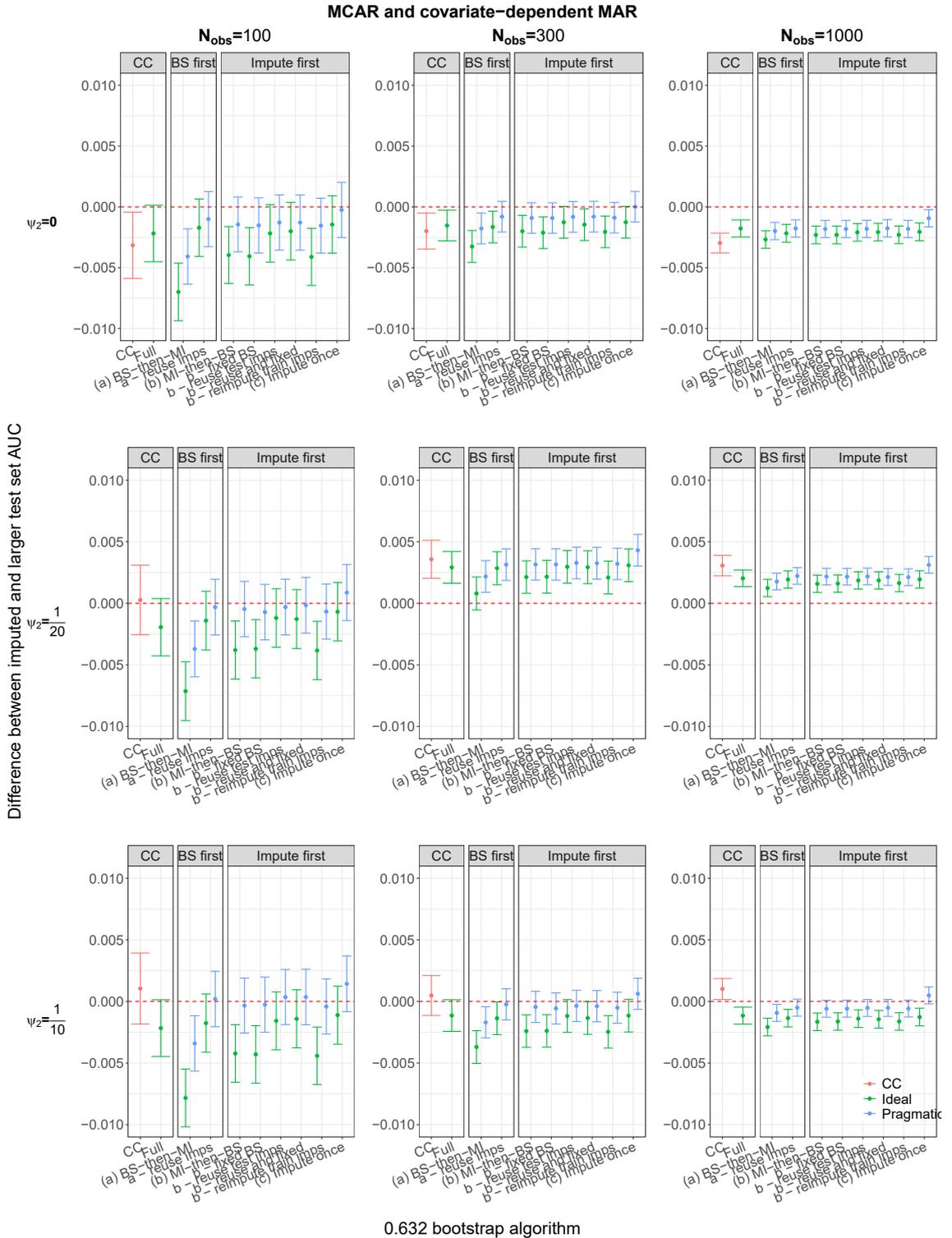


Figure C5: The difference $AUC_{imp} - AUC_{target}$ when data are MCAR or covariate-dependent MAR for $M = 5$ when 25% of values are missing in X_1 . The error bars summarise results from the 2000 repetitions and the limits represent the Monte Carlo 95% confidence interval of $AUC_{imp} - AUC_{target}$. CC (complete-case); methods are described in Section 2.7 or Table 6.1.

Outcome-dependent MAR

Figure C6 presents results for comparing AUC estimates to their respective ideal, pragmatic or complete-case target AUC estimate when data are outcome-dependent or outcome- and covariate-dependent MAR. Similarly to the MCAR and covariate-dependent MAR scenarios, for all sample sizes the methods perform well when compared to their target estimate with a magnitude of the difference in AUCs less than 0.01 ($|AUC_{imp} - AUC_{target}| < 0.01$).

The complete-case analysis tends to overestimate the target complete-case estimate of the AUC for the various outcome-dependent MAR scenarios. With increasing sample size the magnitude of the overestimation tends to decrease and is less than 0.0025 for sample size of 1000 ($|AUC_{CC} - AUC_{target,CC}| < 0.0025$).

When data are outcome-dependent MAR, the pragmatic performance of method *BS-then-MI* tends to underestimate $AUC_{target,imputed}$ while the other methods overestimate $AUC_{target,imputed}$ (i.e. they are over-optimistic); all methods have a similar magnitude when compared to $AUC_{target,imputed}$ ($|AUC_{prag,imp} - AUC_{target,imputed}|$). When data are weak outcome- and covariate-dependent MAR and sample size is 100 method *BS-then-MI* underestimates $AUC_{target,imputed}$ with a smaller magnitude than the other imputations which overestimate $AUC_{target,imputed}$. Method *MI-then-BS impute once* overestimates $AUC_{target,imputed}$ and has the largest magnitude across all methods. With increased sample size all methods tend to overestimate $AUC_{target,imputed}$ except method *BS-then-MI* which approximates $AUC_{target,imputed}$ well. When data are weak outcome-dependent and strong covariate-dependent MAR, all methods tend to underestimate $AUC_{target,imputed}$, except for method *MI-then-BS impute once* which overestimates (i.e. it is over-optimistic). Method *BS-then-MI* underestimates $AUC_{target,imputed}$ more than the other methods. With increasing sample size all methods tend to perform similarly.

The ideal performance of the methods subject to the most amount of data leakage tend to overestimate $AUC_{target,obs}$ (methods *BS-then-MI reuse*, *MI-then-BS reuse test imps* with or without fixed BS, *MI-then-BS impute once*). The ideal performance of the other methods tends to underestimate $AUC_{target,obs}$. With increasing sample size, the magnitude of the difference decreases ($|AUC_{ideal,imp} - AUC_{target,obs}| \rightarrow 0$). With increasing sample size, all methods tend to perform similarly.

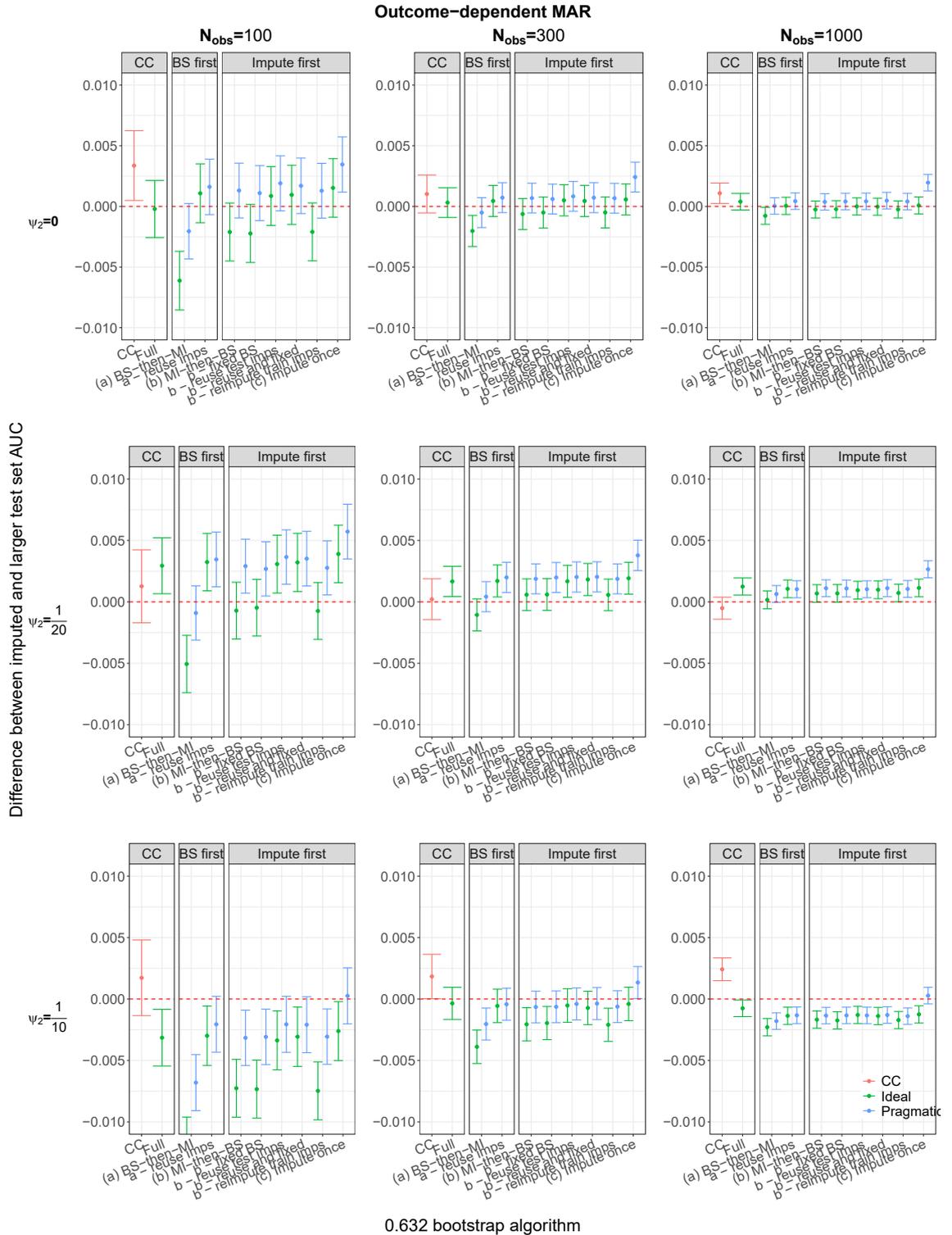


Figure C6: The difference $AUC_{imp} - AUC_{target}$ when data are outcome-dependent MAR for $M = 5$ when 25% of values are missing in X_1 . The error bars summarise results from the 2000 repetitions and the limits represent the Monte Carlo 95% confidence interval of $AUC_{imp} - AUC_{target}$. CC (complete-case); methods are described in Section 2.7 or Table 6.1.

C.3 Detailed results for the Brier score performance of the 0.632 algorithm

A lower Brier score estimate generally suggests the model is performing well. Therefore, if a method underestimates the Brier score estimated when data are fully-observed the method is considered to be over-optimistic i.e. the model performs better when data have been imputed than if the data had not been missing to begin with.

C.3.1 Comparing results to the Brier score estimate when data are fully-observed

MCAR and covariate-dependent MAR

Figure C7 presents results for the Brier score when data are MCAR or covariate-dependent MAR. The Brier score estimates from the various missing data methods are compared to the Brier score estimated when data are fully-observed ($\text{Brier}_{imp} - \text{Brier}_{obs}$).

When data are MCAR and the sample size is 100, the complete-case analysis estimate of the Brier score tends to overestimate Brier_{obs} with a magnitude less than 0.0025 ($|\text{Brier}_{CC} - \text{Brier}_{obs}| < 0.0025$). With increasing sample size the magnitude of the overestimation decreases further. For covariate-dependent MAR, the complete-case analysis tends to underestimate Brier_{obs} with a magnitude of 0.005 for weak covariate-dependent MAR and a magnitude of approximately 0.01 for strong covariate-dependent MAR i.e. the Brier score estimate becomes over-optimistic.

For all sample sizes when data are MCAR or covariate-dependent MAR, the pragmatic performance of all imputation methods overestimates Brier_{obs} ($\text{Brier}_{prag,imp} - \text{Brier}_{obs} > 0$) with a magnitude less than or approximately equal to 0.005. When sample size is 100, method *BS-then-MI* has the largest magnitude of overestimating Brier_{obs} . With increasing sample size when data are MCAR or covariate-dependent MAR, the methods have similar performance to each other when compared with Brier_{obs} . The exception to the overestimation of Brier_{obs} is the pragmatic performance of method *MI-then-BS impute once*. This method underestimates the fully-observed with an estimate of approximately -0.069 and therefore does not fit the scale of Figure C7.

The ideal performance of methods *BS-then-MI*, *MI-then-BS* with or without fixed bootstrap samples and *MI-then-BS re-impute* overestimate Brier_{obs} . Methods *BS-then-MI reuseimps*, *MI-then-BS reuse testimps* and *MI-then-BS impute once* underestimate Brier_{obs} . For a sample size of 100, method *BS-then-MI* has the largest magnitude of overestimation but with increasing sample size it performs similarly to the other methods. Similarly to the pragmatic performance, the ideal performance of method *MI-then-BS impute once* underestimates Brier_{obs} with a magnitude greater than 0.06.

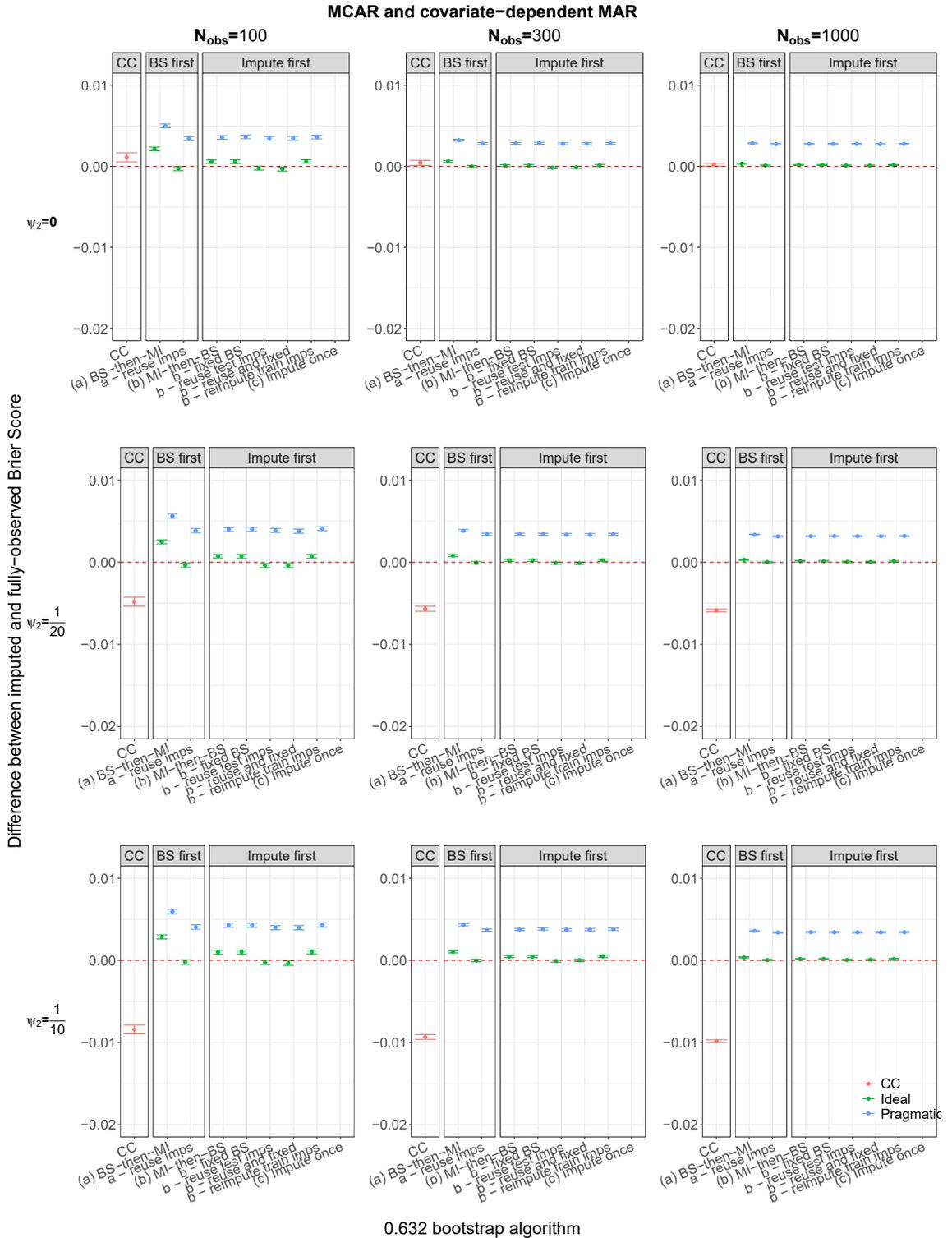


Figure C7: The difference $\text{Brier}_{imp} - \text{Brier}_{obs}$ when data are MCAR or covariate-dependent MAR for $M = 5$ when 25% of values are missing in X_1 . The error bars summarise results from the 2000 repetitions and the limits represent the Monte Carlo 95% confidence interval of $\text{Brier}_{imp} - \text{Brier}_{obs}$. CC (complete-case); methods are described in Section 2.7 or Table 6.1.

Outcome-dependent MAR

Figure C8 presents the results $\text{Brier}_{imp} - \text{Brier}_{obs}$ for the scenario when data are outcome-dependent MAR or outcome- and covariate-dependent MAR.

The complete-case analysis underestimates Brier_{obs} for all sample sizes when data are outcome-dependent or outcome- and covariate-dependent MAR. The magnitude of this underestimation has increased to being greater than 0.01 whereas previously, when data were MCAR or covariate-dependent MAR, the magnitude of underestimation was less than 0.01.

The pragmatic performance of all methods (except *MI-then-BS impute once*) overestimates Brier_{obs} . For a sample size of 100, method *BS-then-MI* has the largest magnitude of overestimation (approximately 0.005) and with increasing sample size it performs similarly to the other imputation methods. The exception is method *MI-then-BS impute once* which underestimates Brier_{obs} and does not fit onto the scale of the Figure.

Similarly again, for the ideal performance methods *BS-then-MI*, *MI-then-BS* (with or without fixed bootstrap samples) and *MI-then-BS re-impute* tend to overestimate Brier_{obs} ($\text{Brier}_{ideal,imp} - \text{Brier}_{obs} > 0$). The other methods which involve reusing imputed datasets or *MI-then-BS impute once* tend to underestimate Brier_{obs} ($\text{Brier}_{ideal,imp} - \text{Brier}_{obs} < 0$). Method *BS-then-MI* tends to have the largest magnitude of the difference ($|\text{Brier}_{ideal,imp} - \text{Brier}_{obs}|$) but with increasing sample size all methods perform similarly with a magnitude less than 0.005 ($|\text{Brier}_{imp} - \text{Brier}_{obs}| < 0.005$).

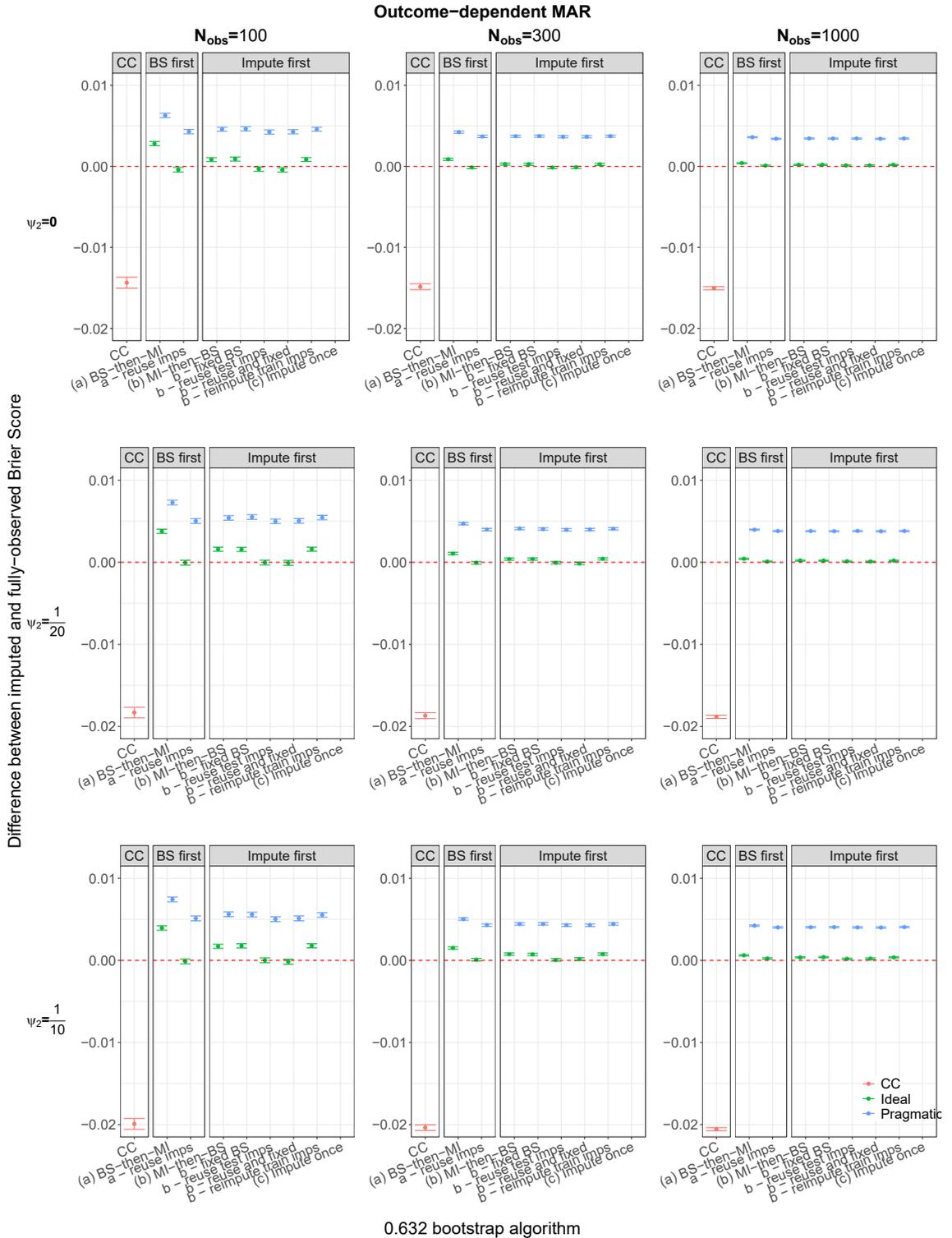


Figure C8: The difference $Brier_{imp} - Brier_{obs}$ when data are outcome-dependent or outcome- and covariate-dependent MAR for $M = 5$ when 25% of values are missing in X_1 . The error bars summarise results from the 2000 repetitions and the limits represent the Monte Carlo 95% confidence interval of $Brier_{imp} - Brier_{obs}$. CC (complete-case); methods are described in Section 2.7 or Table 6.1.

C.3.2 Increasing the number of imputed datasets from 5 to 25

Figure C9 displays the results for comparing the various imputation based methods when using 5 or 25 imputed datasets. The results in the graph are for the scenario when data are outcome-dependent MAR but are representative of the results when data are MCAR or covariate-dependent MAR (available in Supplementary Plots S4.5.3). Due to increased computation time when using 25 imputed datasets the comparison a reduced set of methods were assessed. Results are available for methods *BS-then-MI*, *MI-then-BS* and *MI-then-BS impute once* which are based on 1000 repetitions.

The estimates of the Brier score perform similarly in relation to Brier_{obs} , for all methods regardless of whether 5 or 25 imputed datasets are used.

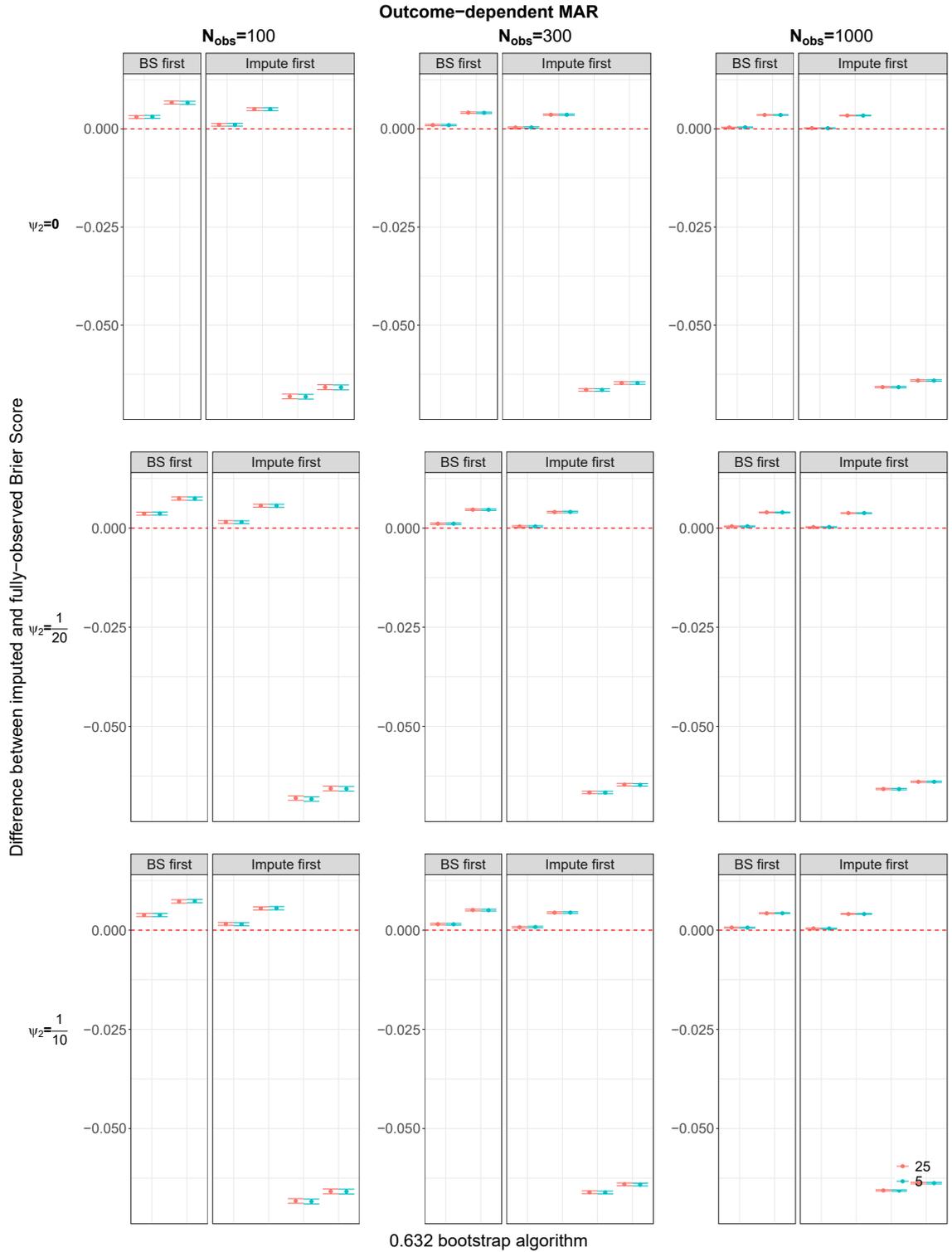


Figure C9: The difference $\text{Brier}_{imp} - \text{Brier}_{obs}$ when data are outcome-dependent or outcome- and covariate-dependent MAR for $M = 25$ when 25% of values are missing in X_1 . The error bars summarise results from the 2000 repetitions and the limits represent the Monte Carlo 95% confidence interval of $\text{Brier}_{imp} - \text{Brier}_{obs}$. CC (complete-case); methods are described in Section 2.7 or Table 6.1.

C.3.3 Increasing the percentage of missingness to 40%

Figure C10 displays the results for comparing how the various methods handle an increased percentage of missing values in X_1 from 25% to 40%. The graph presents results for the scenario when data are weak outcome- and covariate-dependent MAR but are representative of all missing data scenarios, additional graphs for ideal and pragmatic performance comparing the percentage of missing values increasing from 25% to 40% can be found in Supplementary Plots Section S4.2.3.

For the complete-case analysis when data are MCAR the estimate of the Brier score when 40% of X_1 values are missing performs similarly to the estimate of the Brier score when 25% of X_1 values are missing. For all other missing data scenarios, the larger percentage of missingness tends to underestimate Brier_{obs} more than the estimate of the Brier score when 25% of X_1 values are missing ($\text{Brier}_{CC,25\%} - \text{Brier}_{obs} < \text{Brier}_{CC,40\%} - \text{Brier}_{obs}$).

For pragmatic performance, the estimate of the Brier score when 40% of X_1 values tends to overestimate Brier_{obs} more than the estimate of the Brier score when 25% of X_1 values are missing. There is also an increase in the variability across the 2000 repetitions for the 40% missing case when compared to 25% of values missing in X_1 .

For ideal performance when data are MCAR or covariate-dependent MAR, the magnitude of the difference between the Brier score when 40% of X_1 and Brier_{obs} tends to be similar or greater than the magnitude for 25% of X_1 values being missing ($|\text{Brier}_{ideal,imp,25\%} - \text{Brier}_{obs}| \leq |\text{Brier}_{ideal,imp,40\%} - \text{Brier}_{obs}|$). In general, when data are outcome-dependent or outcome- and covariate-dependent MAR the magnitude of the difference between the Brier score estimated when 40% of values are missing and Brier_{obs} is similar to or greater than the Brier estimate comparison when 25% of the values are missing ($|\text{Brier}_{ideal,imp,25\%} - \text{Brier}_{obs}| \leq |\text{Brier}_{ideal,imp,40\%} - \text{Brier}_{obs}|$). The Monte Carlo 95% confidence intervals when 40% of values are missing are wider and tend to encompass or overlap the confidence intervals when 25% of values are missing. The exception is method *BS-then-MI* whose confidence intervals do not overlap for the majority of scenarios when sample size is 300.

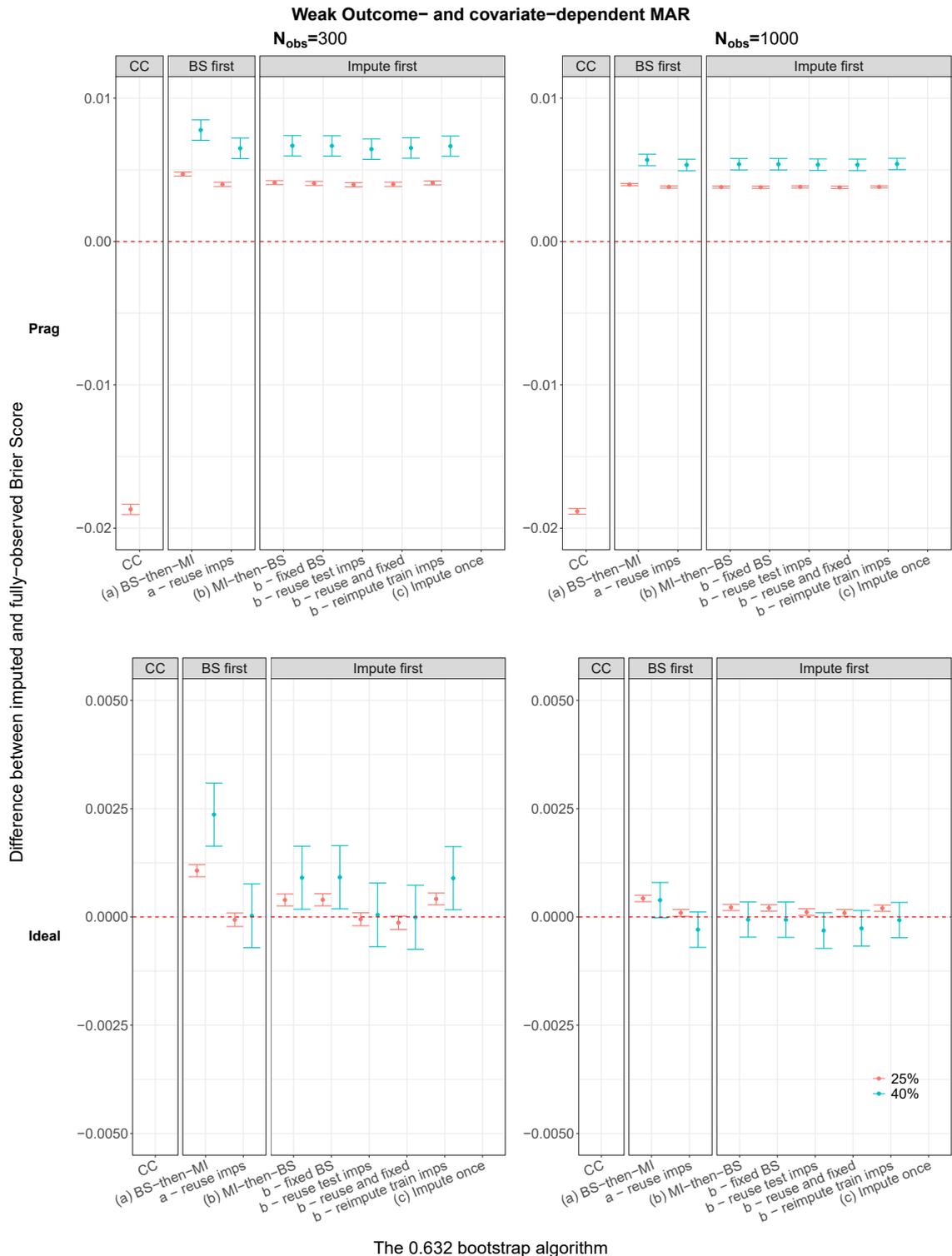


Figure C10: Comparing the impact of increasing the percentage of missingness on the difference $Brier_{imp} - Brier_{obs}$ when data are outcome- and covariate-dependent MAR when $M = 5$. The error bars summarise results from the 2000 repetitions and the limits represent the Monte Carlo 95% confidence interval of $Brier_{imp} - Brier_{obs}$. Red denotes $Brier_{imp} - Brier_{obs}$ when 25% of X_1 values are missing and blue denotes $Brier_{imp} - Brier_{obs}$ when 40% of X_1 values are missing. The top row presents the results for pragmatic performance and the bottom row presents results for ideal performance. CC (complete-case); methods are described in Section 2.7 or Table 6.1.

C.3.4 Comparing results to the target performance

As similarly detailed in the comparison of the bootstrap results for a continuous outcome (Section 6.4.4), the ideal performance of the bootstrap imputation methods and Brier_{obs} were compared to the ideal target Brier score estimate. This is estimated by applying a prediction model, developed using all data, to the fully-observed data in the larger test set ($\text{Brier}_{target,obs}$). The pragmatic performance of the imputation methods is compared to applying a prediction model, developed using all data, to the imputed datasets of the larger test set ($\text{Brier}_{target,imputed}$). The complete-case estimate of the Brier score is compared to applying a prediction model to the observed cases of the larger test set ($\text{Brier}_{target,CC}$).

MCAR and covariate-dependent MAR

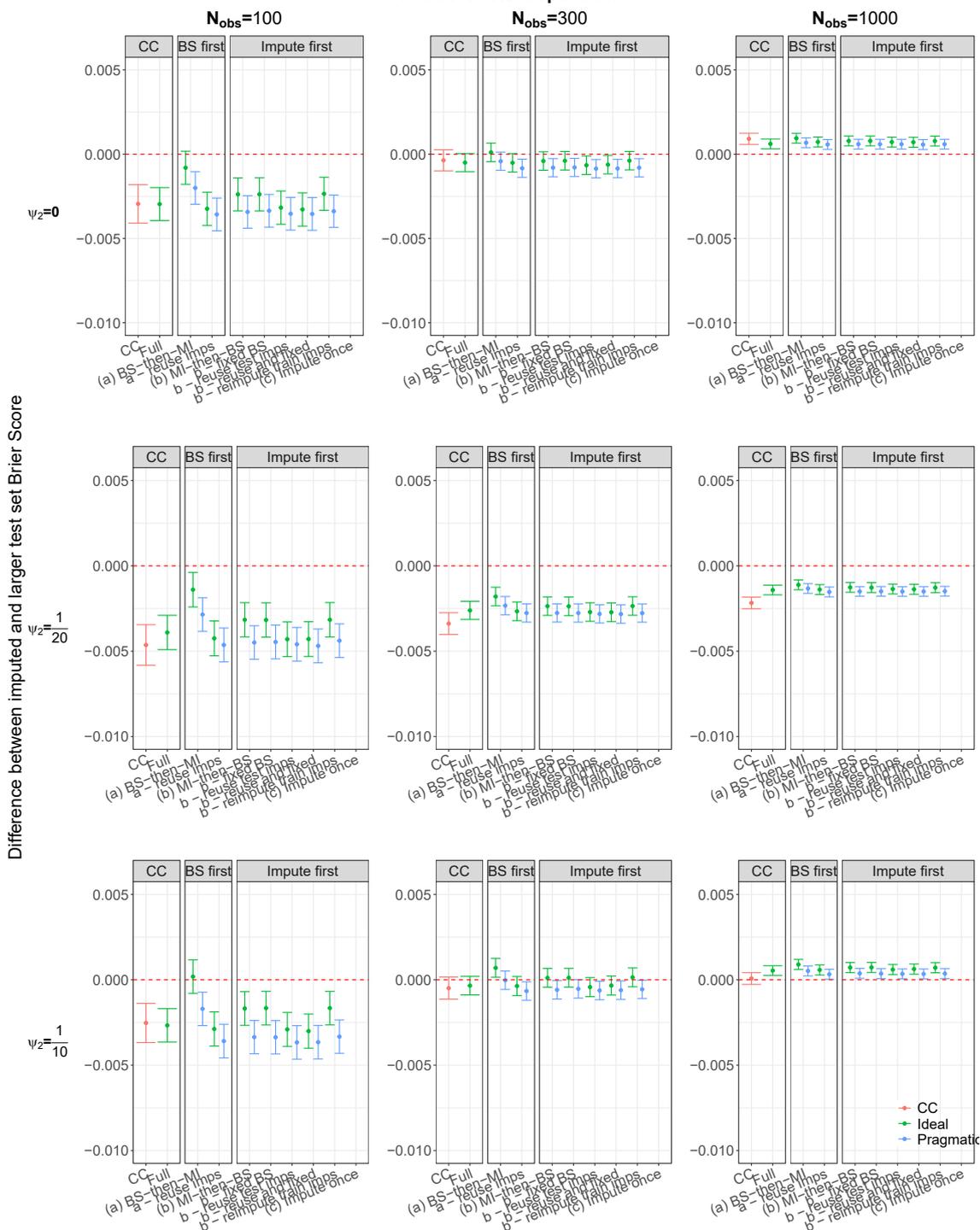
Figure C11 displays the results for comparing the ideal and pragmatic performance Brier score estimates to their respective target estimate when data are MCAR or covariate-dependent MAR.

When sample size is 100, the complete-case analysis underestimates $\text{Brier}_{target,CC}$ ($\text{Brier}_{CC} - \text{Brier}_{target,CC} < 0$). When sample size is increased to 300 and data are MCAR or strong covariate-dependent MAR, the difference between the complete-case estimate and $\text{Brier}_{target,CC}$ is centred around zero. With increasing sample size, the magnitude of the difference between the complete-case analysis estimate and $\text{Brier}_{target,CC}$ tends to decrease.

For a sample size of 100 and 300, the pragmatic performance of all imputation based methods tends to underestimate $\text{Brier}_{target,imputed}$ ($\text{Brier}_{prag,imp} - \text{Brier}_{target,imputed} < 0$). The magnitude of this difference is smallest for method *BS-then-MI* while the other methods tend to perform similarly. With increasing sample size, all methods perform similarly when compared to $\text{Brier}_{target,imputed}$.

For a sample size of 100 and 300, the ideal performance of all imputation methods tends to underestimate Brier_{obs} . Method *BS-then-MI* tends to either have the smallest magnitude of the difference with $\text{Brier}_{target,obs}$ or approximate it well. With increasing sample size, the methods all tend to perform similarly.

MCAR and covariate-dependent MAR



0.632 bootstrap algorithm

Figure C11: The difference $Brier_{imp} - Brier_{target}$ when data are MCAR or covariate-dependent MAR for $M = 5$ when 25% of values are missing in X_1 . The error bars summarise results from the 2000 repetitions and the limits represent the Monte Carlo 95% confidence interval of $Brier_{imp} - Brier_{target}$. CC (complete-case); methods are described in Section 2.7 or Table 6.1.

Outcome-dependent MAR

Figure C12 presents results for comparing the ideal and pragmatic performance Brier score estimates to their respective target estimate when data are outcome-dependent or outcome- and covariate-dependent MAR.

The complete-case analysis tends to underestimate $\text{Brier}_{target,CC}$ for all sample sizes and missing data mechanisms. When the sample size is 100, the magnitude of this underestimation is greater than 0.005 ($|\text{Brier}_{CC} - \text{Brier}_{target,CC}| > 0.005$) but with increasing sample size to 1000, the magnitude decreases to approximately 0.0025.

For pragmatic performance the results are similar to the MCAR and covariate-dependent MAR scenarios. The pragmatic performance of all imputation based methods tend to underestimate $\text{Brier}_{target,imputed}$ for sample sizes of 100 and 300. Method *BS-then-MI* tends to underestimate $\text{Brier}_{target,imputed}$ the least while all other methods perform similarly.

For ideal performance and all sample sizes, method *BS-then-MI* either has the lowest magnitude of underestimation when compared to $\text{Brier}_{target,obs}$, or estimates it well. When sample size is 100, methods *BS-then-MI reuseimps*, *MI-then-BS reuse testimps* and *MI-then-BS impute once* have the largest magnitude of underestimation when compared to the ideal target estimate. For a sample size of 1000, all methods tend to perform similarly when compared to the ideal target estimate of the Brier score, except method *MI-then-BS impute once*.

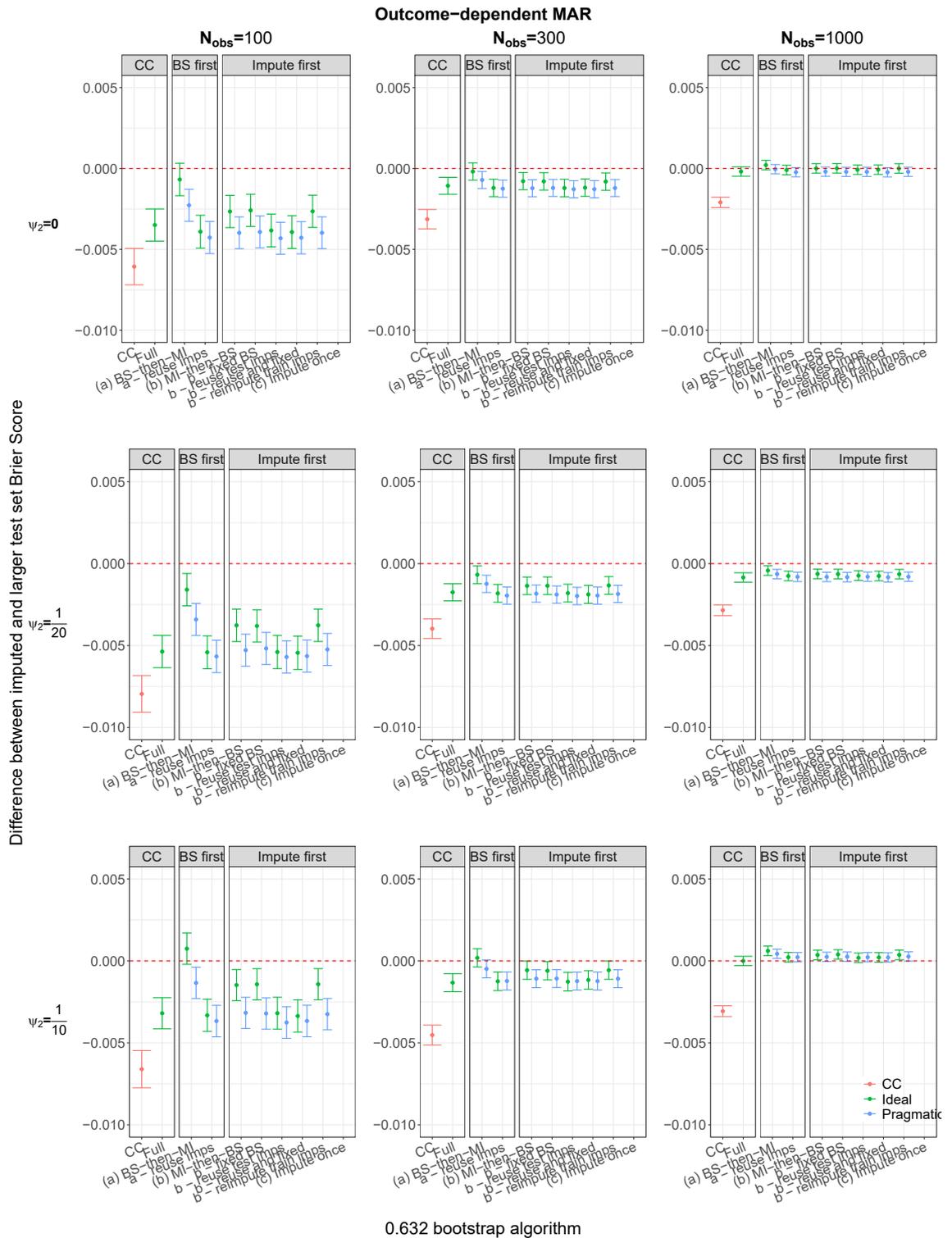


Figure C12: The difference $\text{Brier}_{imp} - \text{Brier}_{target}$ when data are outcome-dependent or outcome- and covariate-dependent MAR for $M = 5$ when 25% of values are missing in X_1 . The error bars summarise results from the 2000 repetitions and the limits represent the Monte Carlo 95% confidence interval of $\text{Brier}_{imp} - \text{Brier}_{target}$. CC (complete-case); methods are described in Section 2.7 or Table 6.1.

C.4 Detailed results for the calibration intercept performance of the 0.632 algorithm

C.4.1 Comparing results to the calibration intercept estimate when data are fully-observed

MCAR and covariate-dependent MAR

Figure C13 presents results for the various methods to handle missing data alongside bootstrap validation when compared to the calibration intercept estimated when data are fully-observed (Intercept_{obs}) i.e. $\text{Intercept}_{imp} - \text{Intercept}_{obs}$. The results presented are for the scenario when data are MCAR (top row, $\psi_2 = 0$) or covariate-dependent MAR ($\psi_2 > 0$).

For a sample size of 100, the various performance estimates of the calibration intercept estimate are very unstable with a large magnitude when compared to the intercept estimated when data are fully-observed. The estimates of the calibration intercept for a sample size of 100 when data are fully-observed were noted to vary widely (Chapter 4, Table 5.1). Here, we will focus on results for a sample size of 300 and 1000.

For MCAR and weak covariate-dependent MAR when sample size is 300 or 1000, the complete-case estimate tends to approximate Intercept_{obs} well. When data are strong covariate-dependent MAR, the complete-case analysis tends to underestimate the fully-observed estimate ($\text{Intercept}_{CC} - \text{Intercept}_{obs} < 0$) and with increasing sample size the difference approaches zero.

When data are MCAR, the pragmatic performance of all imputation methods performs well when compared to Intercept_{obs} . When data are weak or covariate-dependent MAR, all imputation methods tend to overestimate the fully-observed estimate with a magnitude less than 0.005 ($|\text{Intercept}_{prag,imp} - \text{Intercept}_{obs}| < 0.005$). Methods *BS-then-MI*, *MI-then-BS* (with or without fixed bootstrap samples) and *MI-then-BS re-impute* tend to overestimate the fully-observed estimate with a larger magnitude than methods *BS-then-MI reuse*, *MI-then-BS reuse test imps* and *MI-then-BS impute once*. With increasing sample size the methods tend to perform similarly when compared to Intercept_{obs} .

The ideal performance of all methods performs well when compared to Intercept_{obs} when data are MCAR. When data are covariate-dependent MAR, methods *BS-then-MI reuse*, *MI-then-BS reuse test imps* and *MI-then-BS impute once* estimates Intercept_{obs} well. Methods *BS-then-MI*, *MI-then-BS* (with or without fixed bootstrap samples) and *MI-then-BS re-impute* tend to overestimate Intercept_{obs} but still have a magnitude less than 0.005 when sample size is 300. With increasing sample the magnitude of the methods decreases further.

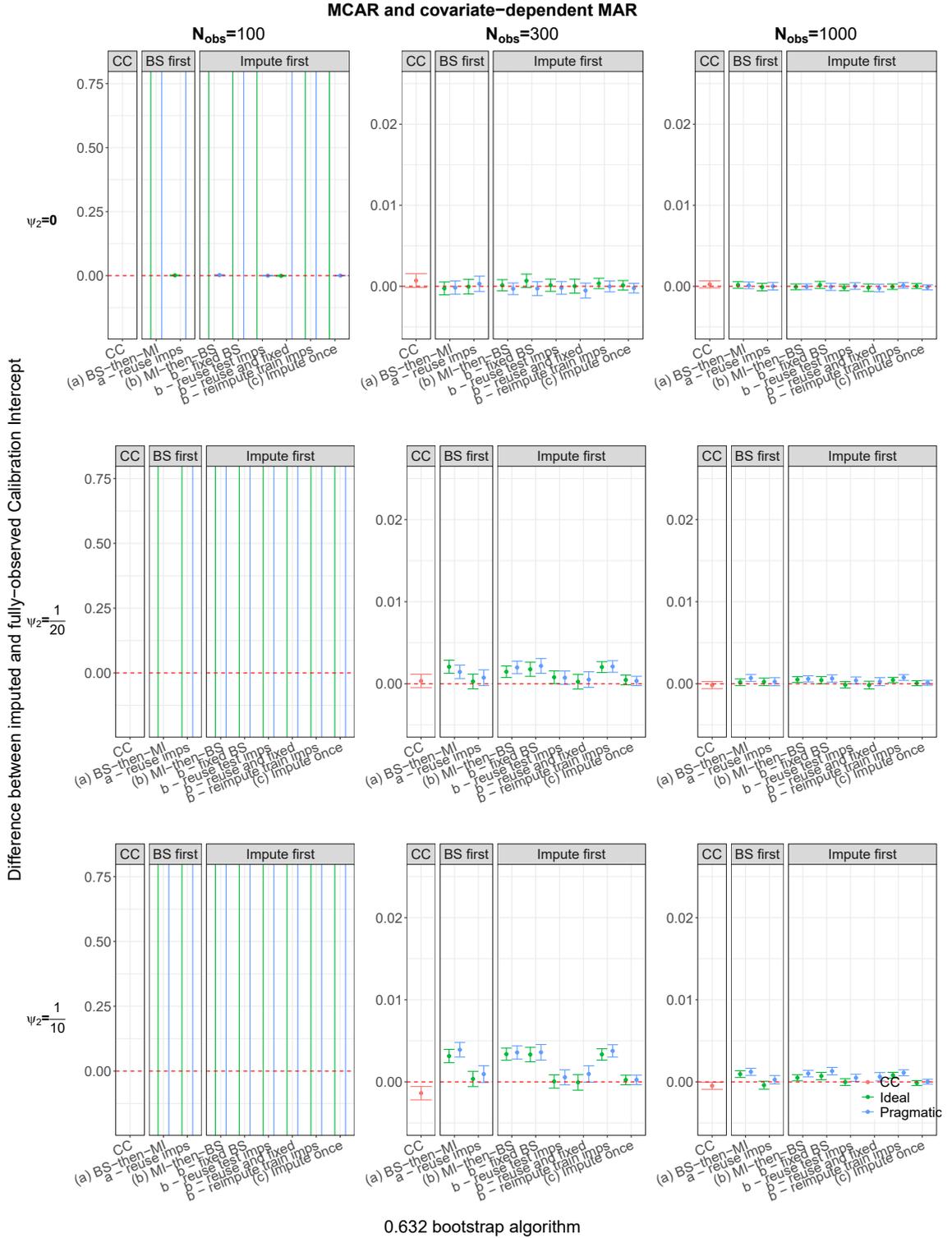


Figure C13: The difference $\text{Intercept}_{imp} - \text{Intercept}_{obs}$ when data are MCAR or covariate-dependent MAR for $M = 5$ when 25% of values are missing in X_1 . The error bars summarise results from the 2000 repetitions and the limits represent the Monte Carlo 95% confidence interval of $\text{Intercept}_{imp} - \text{Intercept}_{obs}$. CC (complete-case); methods are described in Section 2.7 or Table 6.1.

Outcome-dependent MAR

Figure C14 displays the results for the various missing data methods' estimate of the calibration intercept when data are outcome-dependent or outcome- and covariate-dependent MAR. These estimates are compared to the estimate of the calibration intercept when data are fully-observed ($\text{Intercept}_{imp} - \text{Intercept}_{obs}$).

The complete-case estimate tends to underestimate Intercept_{obs} ($\text{Intercept}_{CC} - \text{Intercept}_{obs} < 0$). The magnitude of this difference when sample size is 300 is less than 0.005 and decreases further with increasing sample size. However, the magnitude of this difference has increased when compared to the scenario where data are MCAR or covariate-dependent MAR.

The pragmatic performance of the various methods tends to overestimate Intercept_{obs} for sample sizes of 300 or 1000 ($\text{Intercept}_{prag,imp} - \text{Intercept}_{obs} > 0$). Method *MI-then-BS impute once* tends to have the smallest magnitude ($|\text{Intercept}_{MI-BS-once} - \text{Intercept}_{obs}|$), performing similarly to Intercept_{obs} . The pragmatic performance of methods *BS-then-MI*, *MI-then-BS* (with or without fixed bootstrap samples) and *MI-then-BS re-impute* have the largest magnitudes of overestimation ($0.01 < |\text{Intercept}_{prag,MI-BS-once} - \text{Intercept}_{obs}| < 0.02$) while the remaining methods perform similarly to each other. The magnitudes of all methods, except *MI-then-BS impute once*, become similar with increasing sample size.

The ideal performance of the various methods tends to overestimate Intercept_{obs} for sample sizes of 300 or 1000 ($\text{Intercept}_{ideal,imp} - \text{Intercept}_{obs} > 0$). The magnitude of all ideal performance estimates is less than 0.01. Method *BS-then-MI* tends to have the largest magnitude of overestimation when sample size is 300 and methods *MI-then-BS* (with or without fixed bootstrap samples) and *MI-then-BS re-impute* have the next largest magnitude while the remaining methods all perform similarly in relation to Intercept_{obs} . For a sample size of 1000 all methods tend to perform similarly.

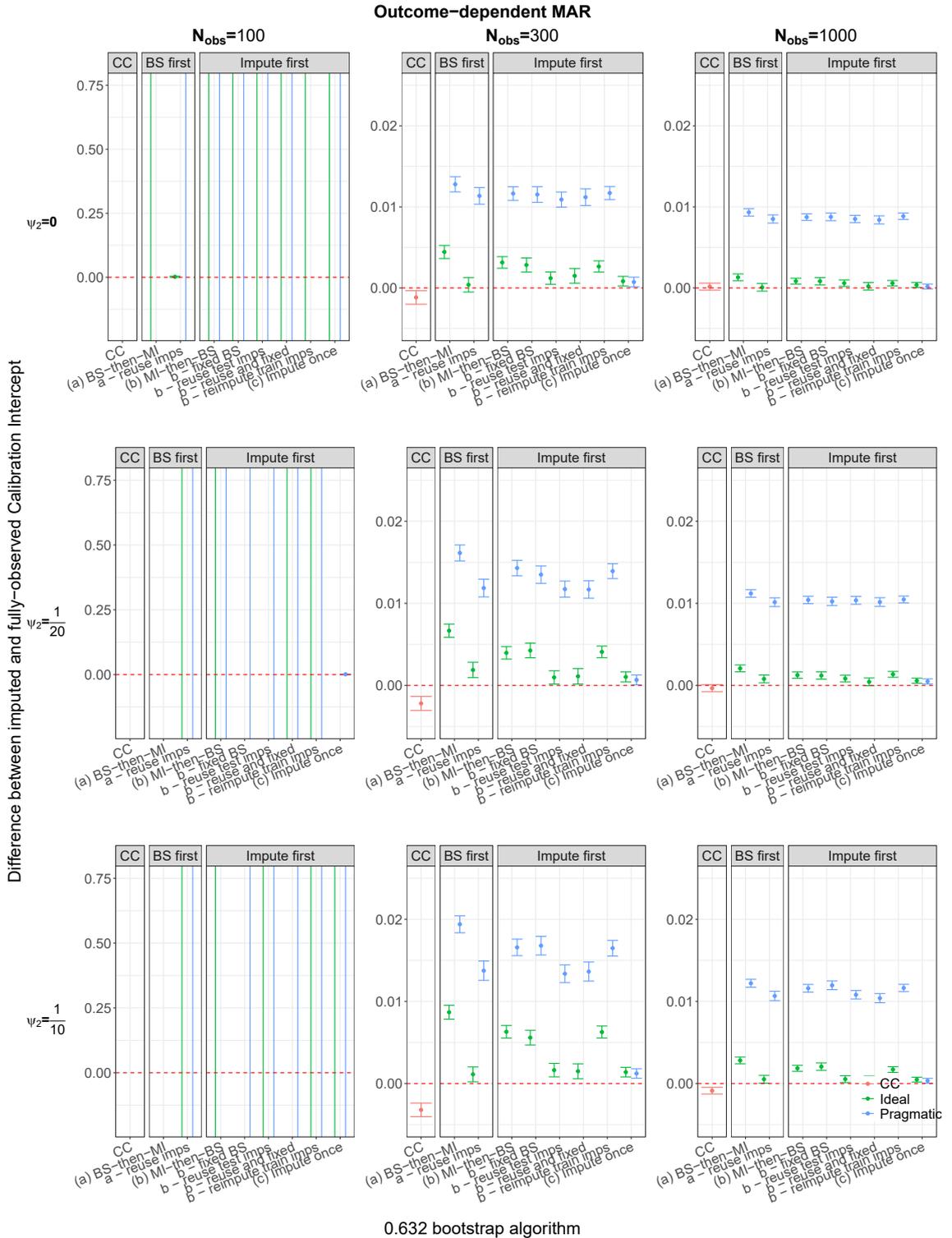


Figure C14: The difference $\text{Intercept}_{imp} - \text{Intercept}_{obs}$ when data are outcome-dependent or outcome- and covariate-dependent MAR for $M = 5$ when 25% of values are missing in X_1 . The error bars summarise results from the 2000 repetitions and the limits represent the Monte Carlo 95% confidence interval of $\text{Intercept}_{imp} - \text{Intercept}_{obs}$. CC (complete-case); methods are described in Section 2.7 or Table 6.1.

C.4.2 Increasing the number of imputed datasets from 5 to 25

FigureC15 displays the results for comparing the various imputation based methods when using 5 or 25 imputed datasets. The results in the graph are for the scenario when data are outcome-dependent MAR but are representative of the results when data are MCAR or covariate-dependent MAR (available in Supplementary PlotsS4.6.3). Due to increased computation time when using 25 imputed datasets the comparison a reduced set of methods were assessed. Results are available for methods *BS-then-MI*, *MI-then-BS* and *MI-then-BS impute once* which are based on 1000 repetitions.

The estimates for the comparison of the calibration intercept for the various methods to Intercept_{obs} , perform similarly regardless of whether 5 or 25 imputed datasets are used.

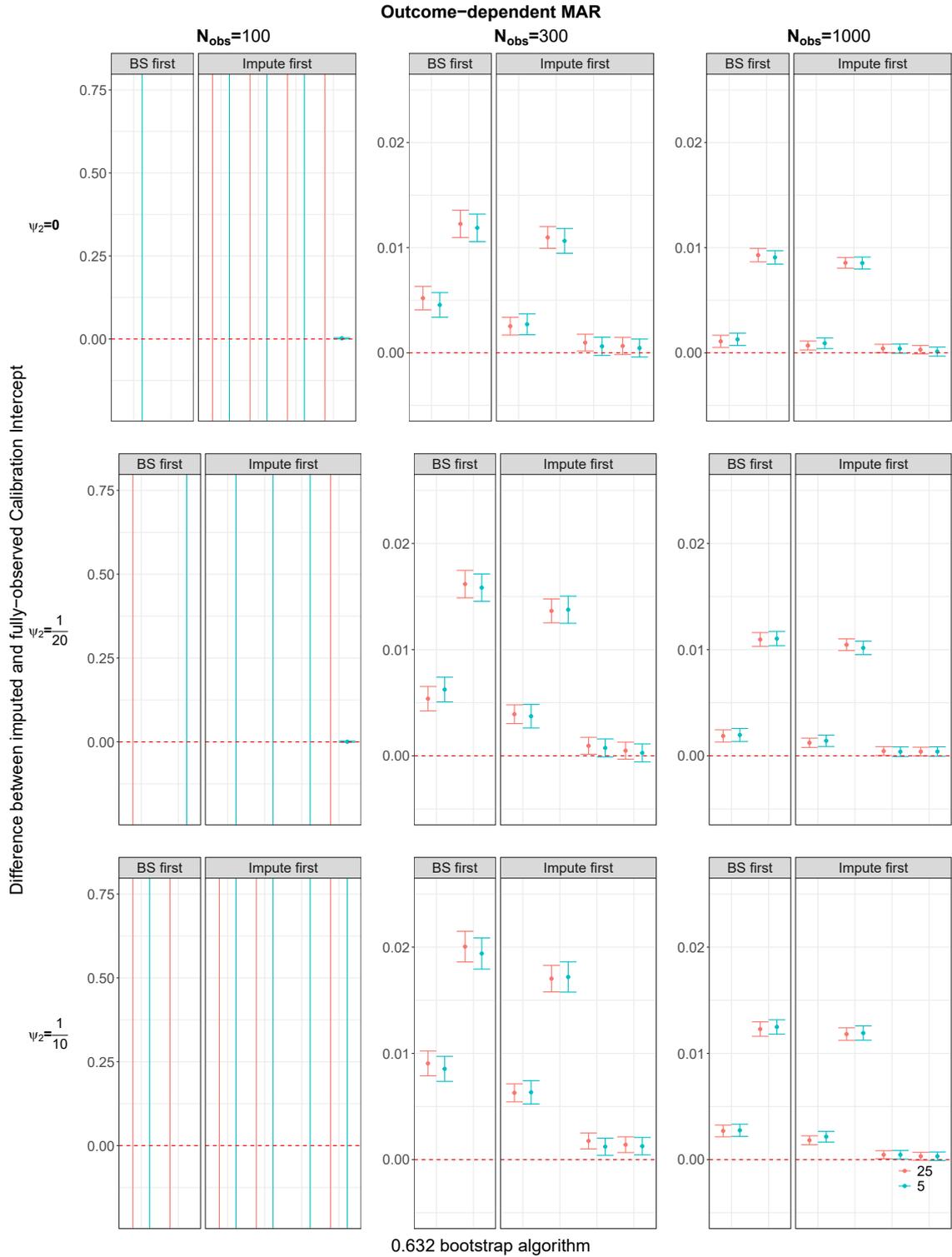


Figure C15: The difference $\text{Intercept}_{imp} - \text{Intercept}_{obs}$ when data are outcome-dependent or outcome- and covariate-dependent MAR for $M = 25$ when 25% of values are missing in X_1 . The error bars summarise results from the 2000 repetitions and the limits represent the Monte Carlo 95% confidence interval of $\text{Intercept}_{imp} - \text{Intercept}_{obs}$. CC (complete-case); methods are described in Section 2.7 or Table 6.1.

C.4.3 Increasing the percentage of missingness to 40%

Figure C16 displays the results for comparing how the various methods handle an increased percentage of missing values in X_1 from 25% to 40%. The graph presents results for the scenario when data are weak outcome- and covariate-dependent MAR but are representative of all missing data scenarios, additional graphs for ideal and pragmatic performance comparing the percentage of missing values increasing from 25% to 40% can be found in Supplementary Plots Section S4.6.2.

When data are MAR, the complete-case analysis estimate of the calibration intercept when 40% of X_1 values are missing tends to be larger in magnitude in relation to Intercept_{obs} than the comparison when 25% of values are missing. When data are MCAR, the estimate when 40% of values are missing compares similarly to when 25% of values are missing. When sample size increases from 300 to 1000, the estimate when 40% of values are missing tends towards the estimate when 25% of values are missing ($|\text{Intercept}_{CC,40\%} - \text{Intercept}_{obs}| \rightarrow |\text{Intercept}_{CC,25\%} - \text{Intercept}_{obs}|$).

For pragmatic performance when data are MCAR, the estimate of the calibration intercept when 40% of X_1 values are missing performs similarly to the estimate when the percentage of missingness is 25%. When data are MAR, for all methods the estimate of the calibration intercept when 40% of X_1 values are missing tends to be similar to or larger in magnitude than the comparison when 25% of values are missing ($|\text{Intercept}_{prag,imp,40\%} - \text{Intercept}_{obs}| \geq |\text{Intercept}_{prag,imp,25\%} - \text{Intercept}_{obs}|$). When data are outcome-dependent or outcome and covariate-dependent MAR, the calibration intercept estimate when 40% of values are missing tends to have a larger magnitude than when 25% of values are missing ($|\text{Intercept}_{prag,imp,40\%} - \text{Intercept}_{obs}| > |\text{Intercept}_{prag,imp,25\%} - \text{Intercept}_{obs}|$). The exception to this is *MI-then-BS impute once* whose estimates perform similarly regardless of 25% or 40% of X_1 values are missing.

For ideal performance, the calibration intercept estimate when 40% of values are missing tends to perform similarly to when 25% of values are missing when data are MCAR. For covariate-dependent, outcome-dependent or outcome- and covariate-dependent MAR the estimate when 40% of values are missing tends to have a similar or larger magnitude than the estimate when 25% of values are missing ($|\text{Intercept}_{ideal,imp,40\%} - \text{Intercept}_{obs}| \geq |\text{Intercept}_{ideal,imp,25\%} - \text{Intercept}_{obs}|$). When data are outcome-dependent or outcome- and covariate-dependent MAR, methods *BS-then-MI*, *MI-then-BS* (with or without fixed BS samples) and *MI-then-BS re-impute* tend to have non-overlapping confidence intervals when sample size is 300 or 1000 while the other methods tend to perform similarly or have overlapping confidence intervals.

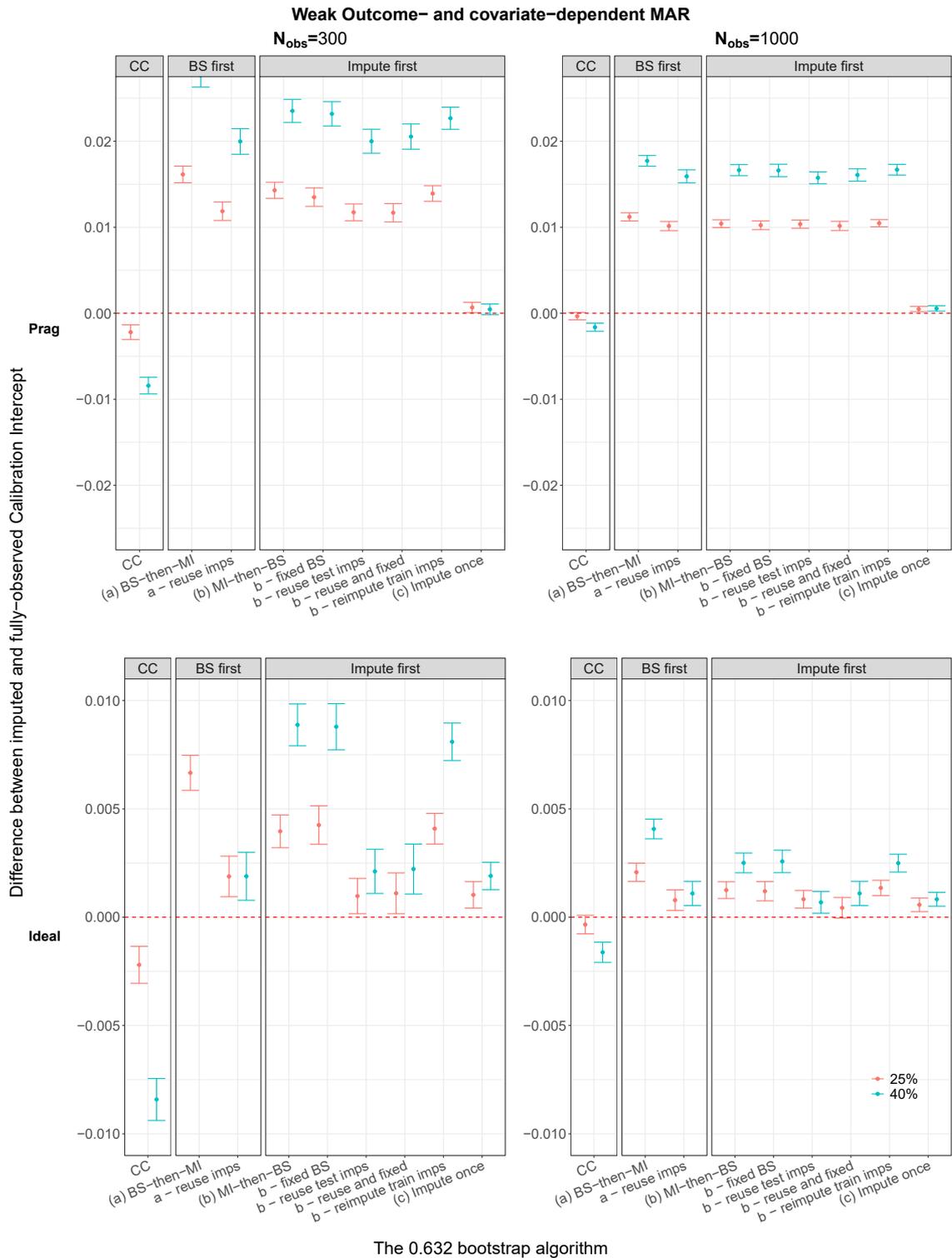


Figure C16: Comparing the impact of increasing the percentage of missingness on the difference $\text{Intercept}_{imp} - \text{Intercept}_{obs}$ when data are outcome- and covariate-dependent MAR when $M = 5$. The error bars summarise results from the 2000 repetitions and the limits represent the Monte Carlo 95% confidence interval of $\text{Intercept}_{imp} - \text{Intercept}_{obs}$. Red denotes $\text{Intercept}_{imp} - \text{Intercept}_{obs}$ when 25% of X_1 values are missing and blue denotes $\text{Intercept}_{imp} - \text{Intercept}_{obs}$ when 40% of X_1 values are missing. The top row presents the results for pragmatic performance and the bottom row presents results for ideal performance. CC (complete-case); methods are described in Section 2.7 or Table 6.1.

C.4.4 Comparing results to the target performance

As similarly detailed for the AUC and Brier score, the ideal performance of the bootstrap imputation methods and Intercept_{obs} were compared to the ideal target calibration intercept estimate. This is estimated by applying a prediction model, developed using all data, to the fully-observed data in the larger test set ($\text{Intercept}_{target,obs}$). The pragmatic performance of the imputation methods is compared to applying a prediction model, developed using all data, to the imputed datasets of the larger test set ($\text{Intercept}_{target,imputed}$). The complete-case estimate of the calibration intercept is compared to applying a prediction model to the observed cases of the larger test set ($\text{Intercept}_{target,CC}$).

As seen when previously comparing the methods' calibration estimates to the intercept estimate when data are fully-observed, the results for sample size of 100 are very unstable for all missing data scenarios and the results will not be further analysed until the Discussion section.

MCAR and covariate-dependent MAR

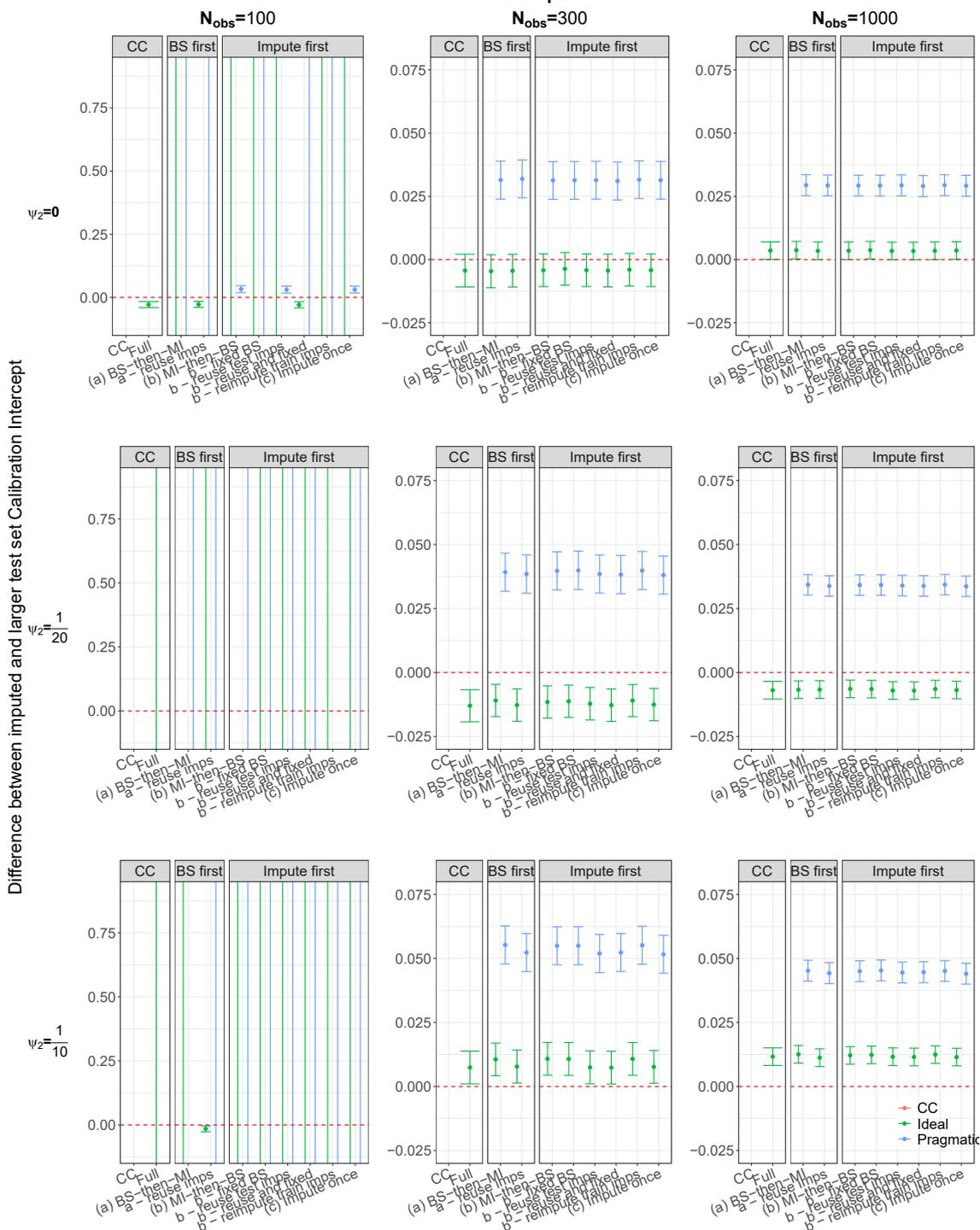
Figure C17 presents results for $\text{Intercept}_{imp} - \text{Intercept}_{target}$ for the scenario when data are MCAR or covariate-dependent MAR.

The complete-case analysis tends to underestimate $\text{Intercept}_{target,CC}$, with an approximate value of -0.48, and does not fit onto the scale of the results presented in Figure C17.

The pragmatic performance of the methods involving MI (*imp*) tends to overestimate the target pragmatic performance ($\text{Intercept}_{prag,imp} - \text{Intercept}_{target,imputed} > 0$). The magnitude of this difference is less than 0.05 for MCAR and weak covariate-dependent MAR, and less than 0.0625 for strong covariate-dependent MAR. All methods have similar pragmatic performance when compared to $\text{Intercept}_{target,imputed}$. With increasing sample size the magnitude of the difference decreases, as does the Monte Carlo standard error across the 2000 repetitions.

The ideal performance of the various imputation based methods tends to approximate $\text{Intercept}_{target,obs}$ well when data are MCAR. For weak or strong covariate-dependent MAR, the methods tend to under- or overestimate $\text{Intercept}_{target,obs}$ with a similar magnitude.

MCAR and covariate-dependent MAR



0.632 bootstrap algorithm

Figure C17: The difference $\text{Intercept}_{imp} - \text{Intercept}_{target}$ when data are MCAR or covariate-dependent MAR for $M = 5$ when 25% of values are missing in X_1 . The error bars summarise results from the 2000 repetitions and the limits represent the Monte Carlo 95% confidence interval of $\text{Intercept}_{imp} - \text{Intercept}_{target}$. CC (complete-case); methods are described in Section 2.7 or Table 6.1.

Outcome-dependent MAR

FigureC18 presents the comparison of the calibration intercept estimates to the complete-case, ideal and pragmatic target estimate for the scenario when data are outcome-dependent or outcome- and covariate-dependent MAR.

Similarly to the MCAR and covariate-dependent MAR, the complete-case analysis tends to underestimate $\text{Intercept}_{target,CC}$ and does not fit on the scale used in FigureC18.

The pragmatic performance of the methods involving MI (*imp*) tends to overestimate the target pragmatic performance ($\text{Intercept}_{prag,imp} - \text{Intercept}_{target,imputed} > 0$). The magnitude of this difference is less than 0.05 for MCAR and weak covariate-dependent MAR, and less than 0.0625 for strong covariate-dependent MAR. All methods have similar pragmatic performance when compared to $\text{Intercept}_{target,imputed}$, although method *MI-then-BS impute once* has the smallest magnitude of the difference overall. With increasing sample size the magnitude of the difference decreases, as does the Monte Carlo standard error across the 2000 repetitions.

The ideal performance of the various imputation based methods tends to underestimate $\text{Intercept}_{target,obs}$. The magnitude of the mean difference tends to be less than 0.025 ($\text{Intercept}_{ideal,imp} - \text{Intercept}_{target,imputed} < 0.025$). With increasing sample size the magnitude of the difference decreases. All methods tend to perform similarly when compared to $\text{Intercept}_{target,obs}$.

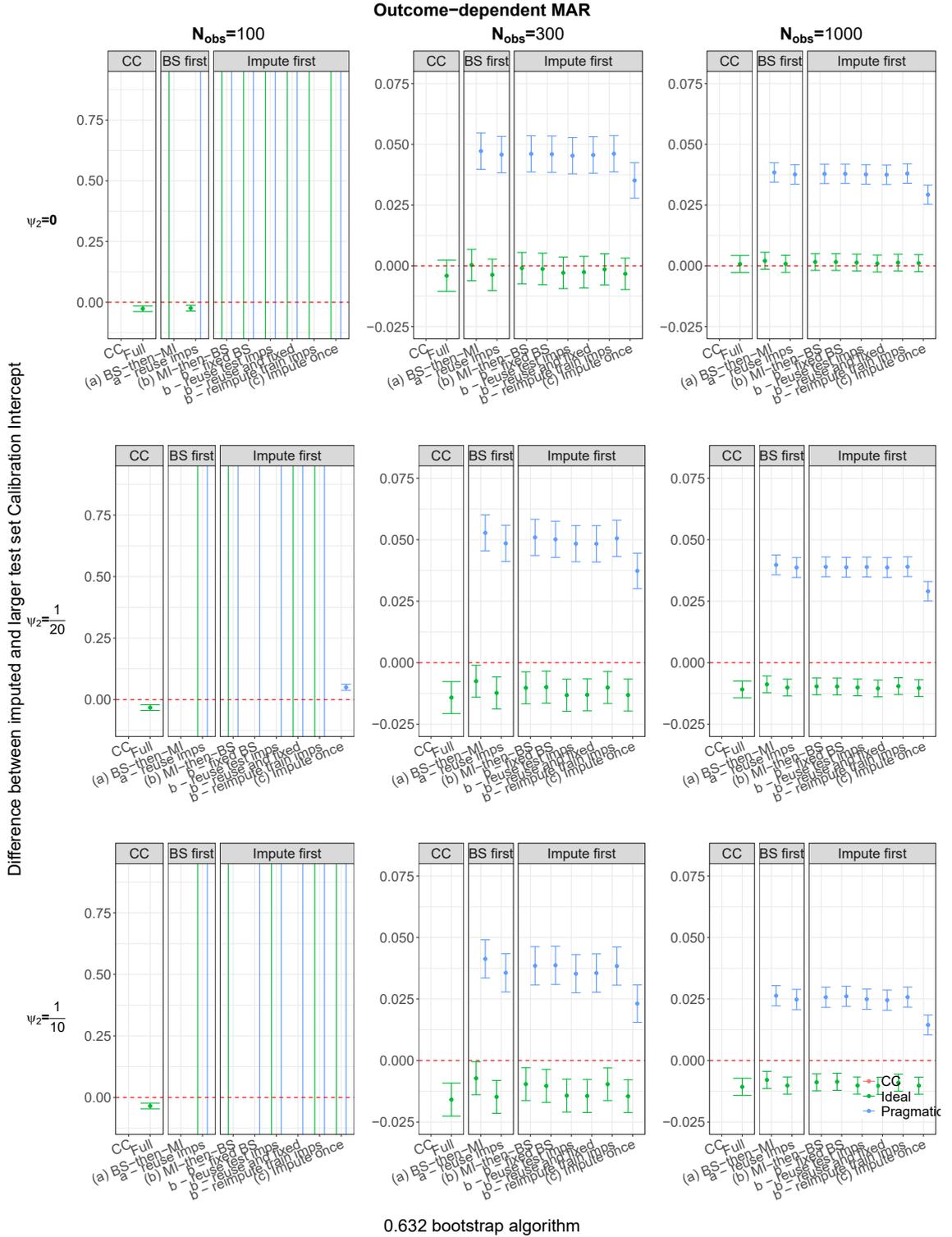


Figure C18: The difference $\text{Intercept}_{imp} - \text{Intercept}_{target}$ when data are outcome-dependent or outcome- and covariate-dependent MAR for $M = 5$ when 25% of values are missing in X_1 . The error bars summarise results from the 2000 repetitions and the limits represent the Monte Carlo 95% confidence interval of $\text{Intercept}_{imp} - \text{Intercept}_{target}$. CC (complete-case); methods are described in Section 2.7 or Table 6.1.

C.5 Detailed results for the calibration slope performance of the 0.632 algorithm

C.5.1 Comparing results to the calibration slope estimate when data are fully-observed

MCAR and covariate-dependent MAR

Figure C19 presents results for the various missing data methods' estimate of the calibration slope. These results are compared to the estimate of the calibration slope when data are fully-observed i.e. $\text{Slope}_{imp} - \text{Slope}_{obs}$. The Figure displays results for the scenario when data are MCAR or covariate-dependent MAR.

When data are MCAR or covariate-dependent MAR, the complete-case analysis tends to overestimate Slope_{obs} . When the sample size is 100, the magnitude of the difference tends to be large ($|\text{Slope}_{CC} - \text{Slope}_{obs}| > 0.5$). Increasing the sample size to 300 or 1000 decreases the magnitude to less than 0.01.

Similarly to the calibration intercept, the ideal and pragmatic performance of the methods tends to be slightly unstable when sample size is small. The performance either under- or overestimates Slope_{obs} . The majority of estimates tend to have large magnitudes ($|\text{Slope}_{imp} - \text{Slope}_{obs}|$) of underestimation and large 95% confidence intervals. This improved with increasing sample size.

When the sample size is 300 or 1000, the pragmatic performance of all methods tends to underestimate Slope_{obs} . Method *BS-then-MI* tends to underestimate the calibration slope the most while method *MI-then-BS impute once* tends to underestimate it the least. The other methods perform similarly in relation to Slope_{obs} with an average difference of -0.05. With increased sample size method *BS-then-MI* tends to perform similarly to the other methods, except for method *MI-then-BS impute once* which has a magnitude less than 0.01 ($|\text{Slope}_{frag,MI-BS-once} - \text{Slope}_{obs}| < 0.01$).

The ideal performance of the imputation based methods underestimates Slope_{obs} ($-0.0125 < |\text{Slope}_{ideal,imp} - \text{Slope}_{obs}| < 0$). Method *BS-then-MI* tends to have the largest magnitude of underestimation when compared to the fully-observed estimate ($|\text{Slope}_{ideal,BS-MI} - \text{Slope}_{obs}|$). With increased sample size the ideal performance of all the imputation methods tends to decrease and performs similarly to Slope_{obs} .

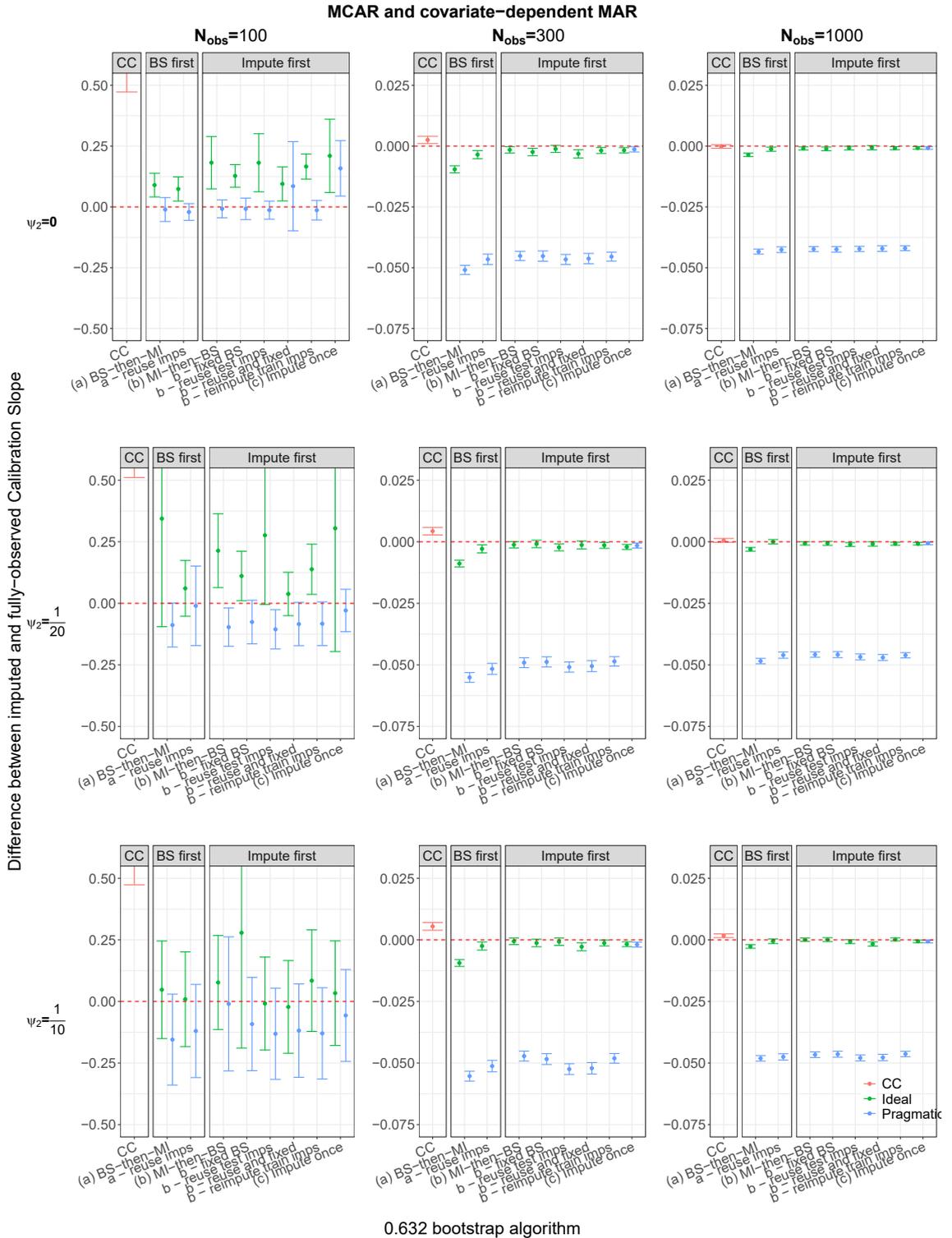


Figure C19: The difference $\text{Slope}_{imp} - \text{Slope}_{obs}$ when data are MCAR or covariate-dependent MAR for $M = 5$ when 25% of values are missing in X_1 . The error bars summarise results from the 2000 repetitions and the limits represent the Monte Carlo 95% confidence interval of $\text{Slope}_{imp} - \text{Slope}_{obs}$. CC (complete-case); methods are described in Section 2.7 or Table 6.1.

Outcome-dependent MAR

Figure C20 displays the results for the various missing data methods' estimate of the calibration slope which is compared to the calibration slope estimated when data are fully-observed ($\text{Slope}_{imp} - \text{Slope}_{obs}$). The Figure displays the results for the outcome-dependent and outcome- and covariate-dependent MAR scenarios.

Similarly to the MCAR and covariate-dependent MAR scenario, the estimates of performance when sample size is 100 tend to be unstable with large magnitudes and wide confidence intervals. This improves with increasing sample size.

The complete-case analysis tends to overestimate Slope_{obs} . The magnitude of the difference tends to be less than 0.0125 ($|\text{Slope}_{CC} - \text{Slope}_{obs}| < 0.0125$) when sample size is 300 or 1000.

The pragmatic performance of all imputation methods tends to underestimate Slope_{obs} . The magnitude of this difference ($|\text{Slope}_{prag,imp} - \text{Slope}_{obs}|$) tends to be between 0.05 and 0.075 for all methods except method *MI-then-BS impute once* whose magnitude tends to be less than 0.01. For a sample size of 300 the method *BS-then-MI* tends to have the largest magnitude of underestimation while the other methods tend to perform similarly when data are outcome-dependent or outcome- and covariate-dependent MAR. With increasing sample size all methods perform similarly when compared to Slope_{obs} , with the exception of method *MI-then-BS impute once*.

The ideal performance of all imputation methods underestimates Slope_{obs} for outcome-dependent and outcome- and covariate-dependent MAR. For a sample size of 300, method *BS-then-MI* has the largest magnitude when compared to Slope_{obs} ($0.0125 \leq |\text{Slope}_{ideal,BS-MI} - \text{Slope}_{obs}| < 0.025$). All other methods tend to perform similarly with a magnitude less than 0.0125. With increasing sample size the magnitude of all methods decreases to be less than 0.006.

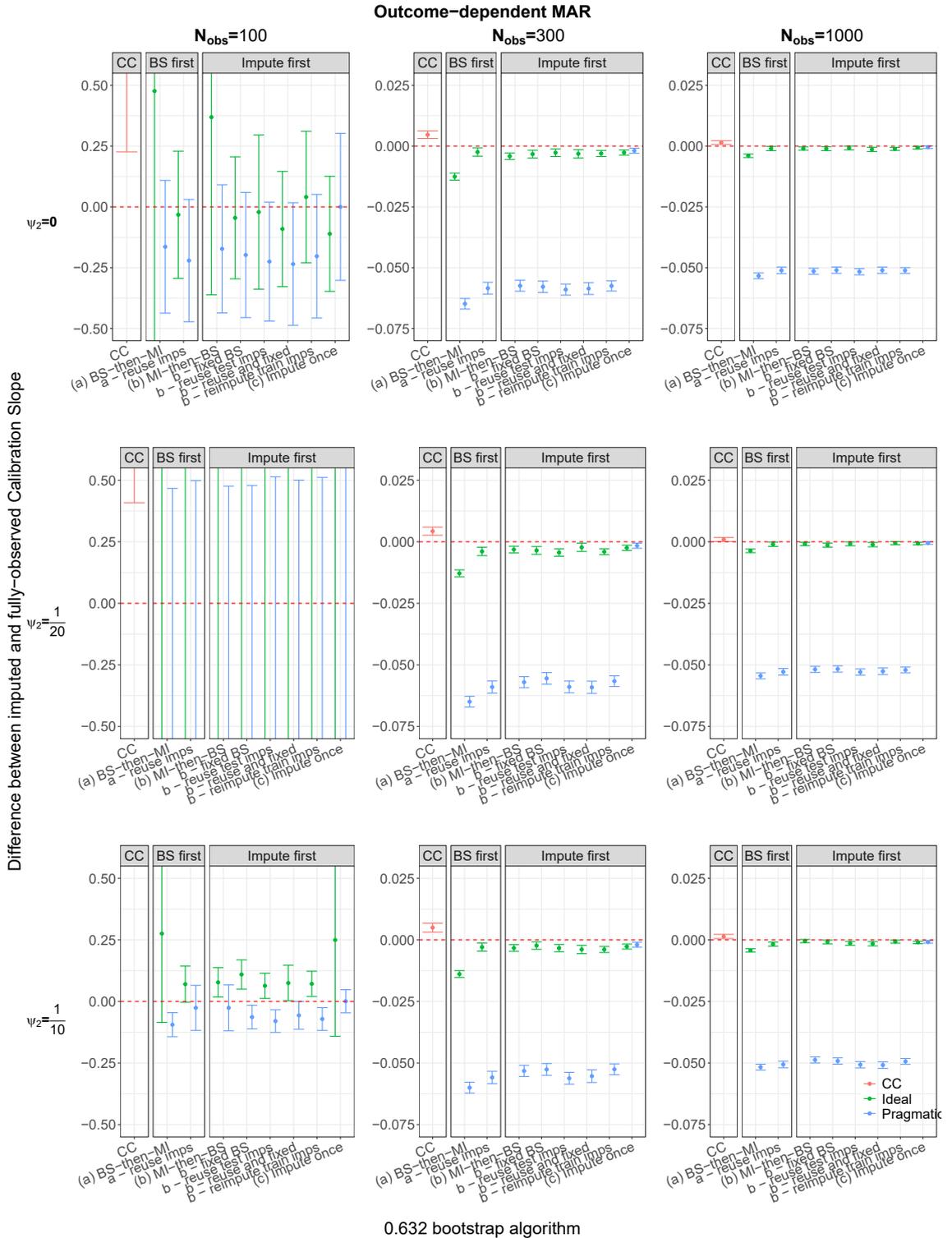


Figure C20: The difference $\text{Slope}_{imp} - \text{Slope}_{obs}$ when data are outcome-dependent or outcome- and covariate-dependent MAR for $M = 5$ when 25% of values are missing in X_1 . The error bars summarise results from the 2000 repetitions and the limits represent the Monte Carlo 95% confidence interval of $\text{Slope}_{imp} - \text{Slope}_{obs}$. CC (complete-case); methods are described in Section 2.7 or Table 6.1.

C.5.2 Increasing the number of imputed datasets from 5 to 25

FigureC21 displays the results for comparing the various imputation based methods when using 5 or 25 imputed datasets. The results in the graph are for the scenario when data are outcome-dependent MAR but are representative of the results when data are MCAR or covariate-dependent MAR (available in Supplementary PlotsS4.6.3). Due to increased computation time when using 25 imputed datasets the comparison a reduced set of methods were assessed. Results are available for methods *BS-then-MI*, *MI-then-BS* and *MI-then-BS impute once* which are based on 1000 repetitions.

The estimates for the comparison of the calibration slope for the various methods to Slope_{obs} , perform similarly regardless of whether 5 or 25 imputed datasets are used.

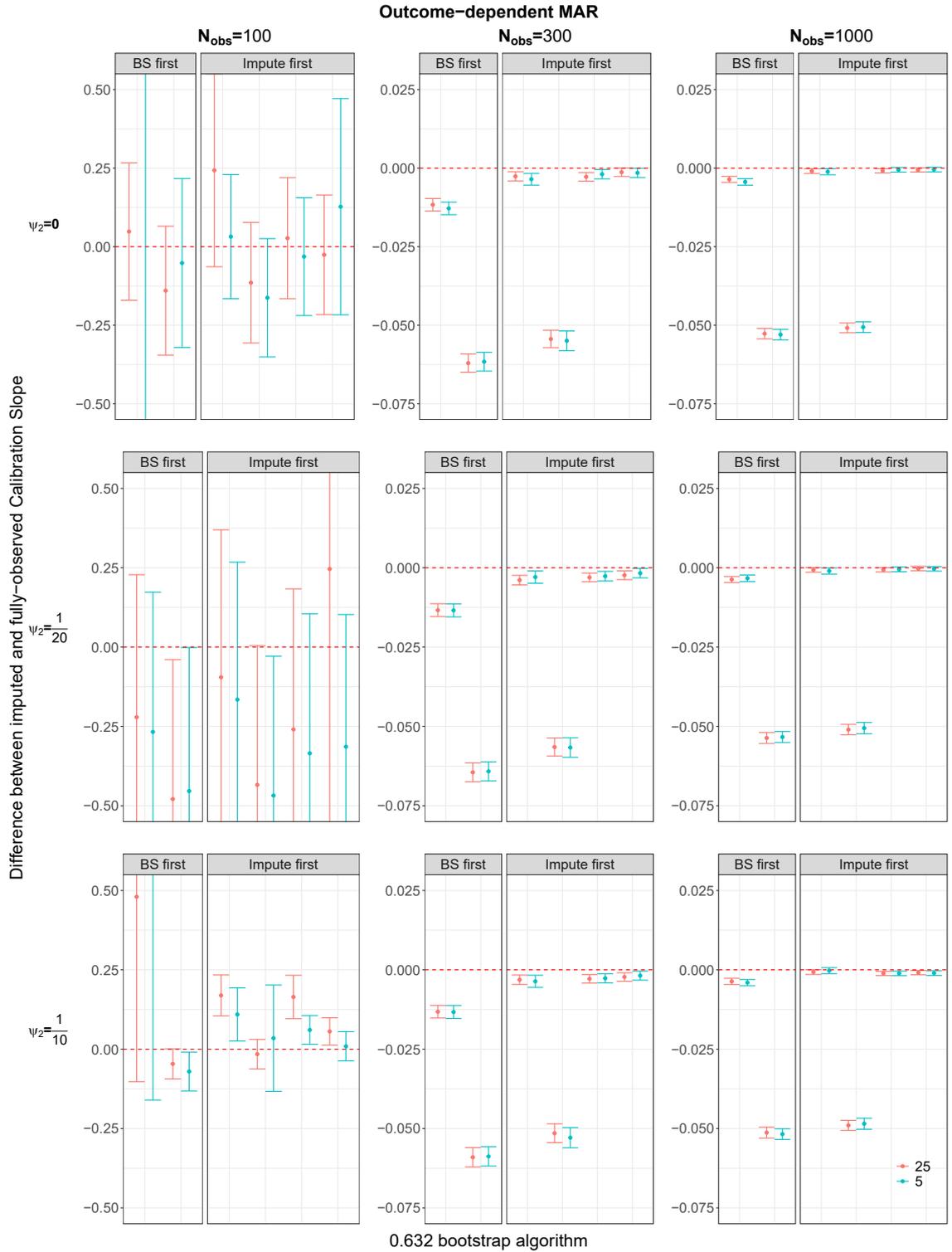


Figure C21: The difference $\text{Slope}_{imp} - \text{Slope}_{obs}$ when data are outcome-dependent or outcome- and covariate-dependent MAR for $M = 25$ when 25% of values are missing in X_1 . The error bars summarise results from the 2000 repetitions and the limits represent the Monte Carlo 95% confidence interval of $\text{Slope}_{imp} - \text{Slope}_{obs}$. CC (complete-case); methods are described in Section 2.7 or Table 6.1.

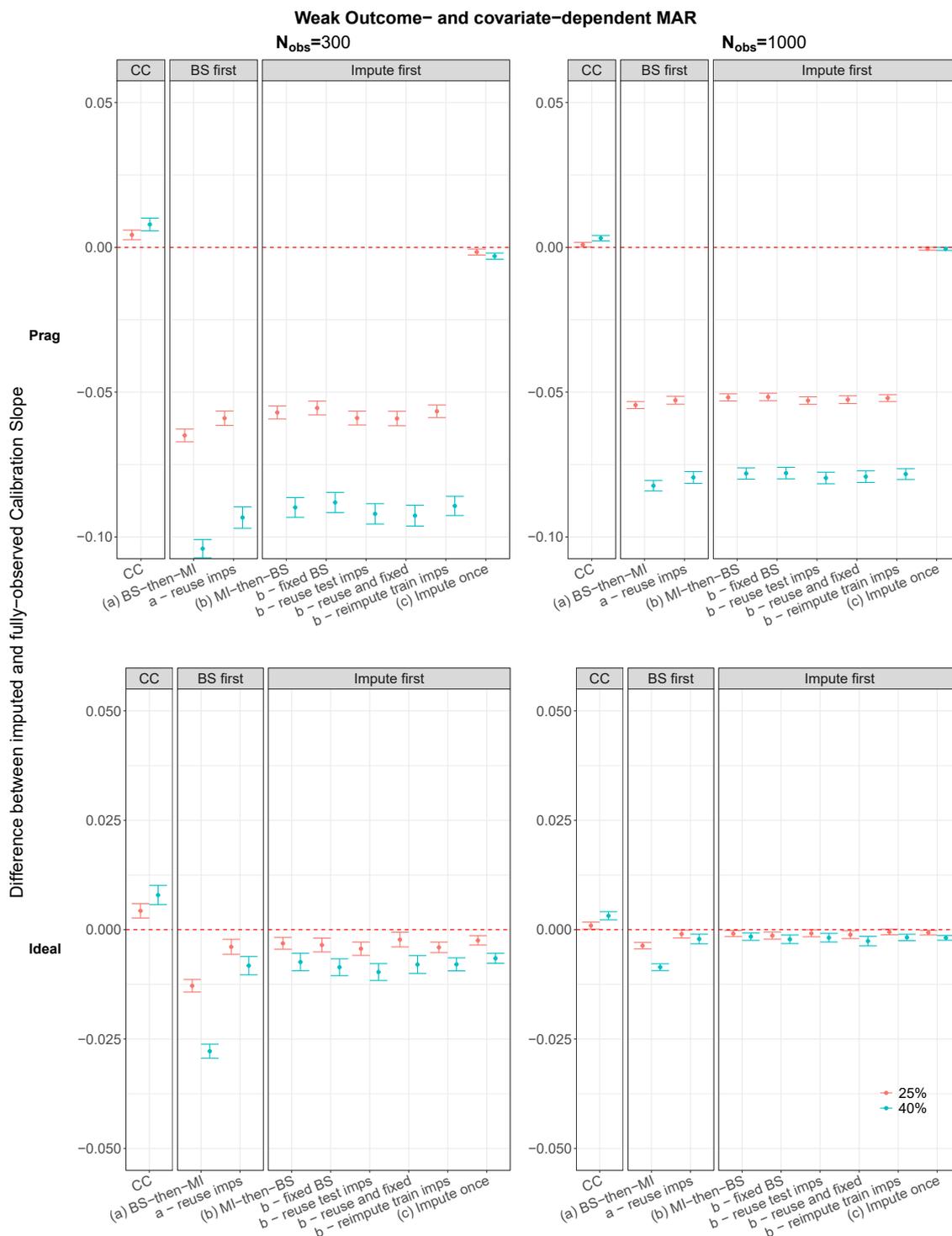
C.5.3 Increasing the percentage of missingness to 40%

Figure C22 displays the results for comparing how the various methods handle an increased percentage of missing values in X_1 from 25% to 40%. The graph presents results for the scenario when data are weak outcome- and covariate-dependent MAR but are representative of all missing data scenarios, additional graphs for ideal and pragmatic performance comparing the percentage of missing values increasing from 25% to 40% can be found in Supplementary Plots Section S4.6.2.

For all missing data scenarios, the complete-case analysis when 40% of X_1 values are missing tends to overestimate Slope_{obs} more than when 25% of X_1 values are missing. With increasing sample size, the magnitude of the complete-case analysis when 40% of values are missing decreases and tends towards the magnitude when 25% of values are missing ($|\text{Slope}_{CC,40} - \text{Slope}_{obs}| \rightarrow |\text{Slope}_{CC,25} - \text{Slope}_{obs}|$) for all missing data scenarios.

For pragmatic performance the estimates of the calibration slope when 40% of X_1 values were set as missing underestimates the calibration slope when 25% of X_1 values are missing ($|\text{Slope}_{prag,imp,40} - \text{Slope}_{obs}| > |\text{Slope}_{prag,imp,25} - \text{Slope}_{obs}|$) for all missing data scenarios. This holds true for all imputation methods except *MI-then-BS impute once* which tend to perform similarly in relation to Slope_{obs} .

Similarly for the ideal performance, the calibration slope estimate when 40% of values are missing tends to underestimate the slope estimate when 25% of values are missing ($\text{Slope}_{ideal,imp,40} - \text{Slope}_{obs} < \text{Slope}_{ideal,imp,25} - \text{Slope}_{obs}$) for methods *BS-then-MI* and *BS-then-MI reuse imps*. The various *MI-then-BS* methods tend to have similar or larger magnitudes for a larger percentage of missingness, but the confidence intervals tend to overlap. With increasing sample size from 300 to 1000, the magnitude of the performance tends to be similar to the percentage when 25% of values are missing for all methods i.e. $|\text{Slope}_{ideal,imp,40} - \text{Slope}_{obs}| \rightarrow |\text{Slope}_{ideal,imp,25} - \text{Slope}_{obs}|$.



The 0.632 bootstrap algorithm

Figure C22: Comparing the impact of increasing the percentage of missingness on the difference $Slope_{imp} - Slope_{obs}$ when data are outcome- and covariate-dependent MAR when $M = 5$. The error bars summarise results from the 2000 repetitions and the limits represent the Monte Carlo 95% confidence interval of $Slope_{imp} - Slope_{obs}$. Red denotes $Slope_{imp} - Slope_{obs}$ when 25% of X_1 values are missing and blue denotes $Slope_{imp} - Slope_{obs}$ when 40% of X_1 values are missing. The top row presents the results for pragmatic performance and the bottom row presents results for ideal performance. CC (complete-case); methods are described in Section 2.7 or Table 6.1.

C.5.4 Comparing results to the target performance

As similarly detailed for the AUC and Brier score, the ideal performance of the bootstrap imputation methods and Slope_{obs} were compared to the calibration slope estimate from applying a prediction model, developed using all data, to the fully-observed data in the larger test set ($\text{Slope}_{target,obs}$). The pragmatic performance of the imputation methods is compared to applying a prediction model, developed using all data, to the imputed datasets of the larger test set ($\text{Slope}_{target,imputed}$). The complete-case estimate of the calibration slope is compared to applying a prediction model to the observed cases of the larger test set ($\text{Slope}_{target,CC}$).

MCAR and covariate-dependent MAR

FigureC23 presents the calibration slope estimates to the complete-case, ideal and pragmatic target estimate when data are MCAR or covariate-dependent MAR.

The complete-case analysis estimate performs poorly and tends to overestimate $\text{Slope}_{target,CC}$ by at least 1.5 when data are MCAR or covariate-dependent MAR ($\text{Slope}_{CC} - \text{Slope}_{target,CC} > 1.5$). It does not fit onto the scale of the graph in FigureC23.

The pragmatic performance of all methods tends to overestimate $\text{Slope}_{target,imputed}$ ($\text{Slope}_{prag,imp} - \text{Slope}_{target,imputed} > 0$). The magnitude of this difference is approximately 0.25 when data are MCAR or covariate-dependent MAR and sample size is 100. When the sample size is 300, method *MI-then-BS impute once* tends to have a magnitude greater than 0.05 ($\text{Slope}_{prag,MI-BS-once} - \text{Slope}_{target,imputed} > 0.05$). The other imputation methods have magnitudes less than 0.05 and method *BS-then-MI* tends to have the smallest magnitude. Increasing the sample size to 1000, the magnitude of the difference tends to decrease and all methods perform similarly, except for method *MI-then-BS impute once* which still overestimates $\text{Slope}_{target,imputed}$ by approximately 0.05.

The ideal performance of all methods overestimates the target ideal estimate of the calibration slope ($\text{Slope}_{ideal,imp} - \text{Slope}_{target,obs} > 0$). The magnitude of the difference is less than 0.05 for all methods when sample size is 300 or 1000. Method *BS-then-MI* results in a slope estimate that is closest to the ideal target estimate while all other methods perform similarly. With increasing sample, all methods tend to perform similarly and either estimate $\text{Slope}_{target,obs}$ well or tend to slightly overestimate it.

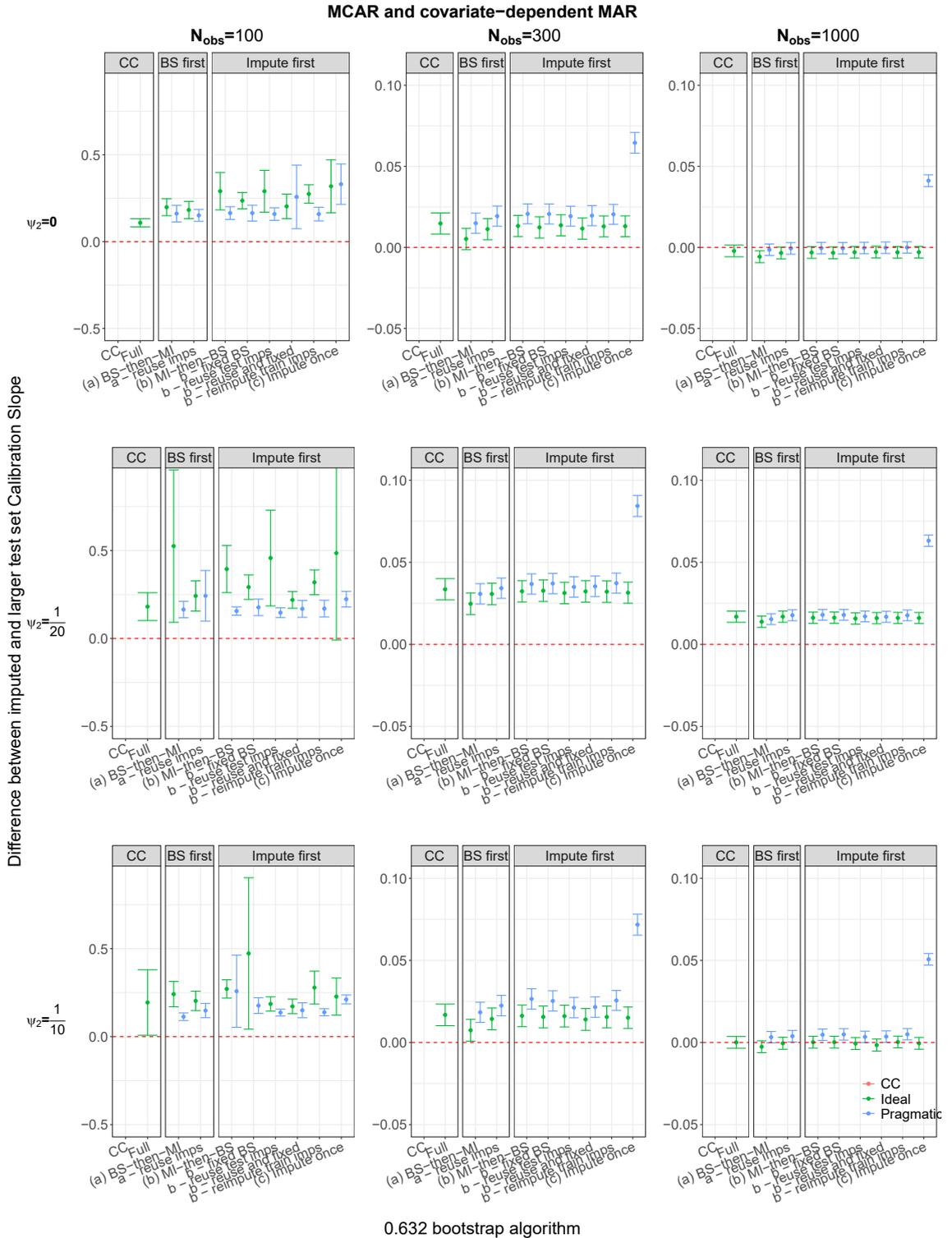


Figure C23: The difference $\text{Slope}_{imp} - \text{Slope}_{target}$ when data are MCAR or covariate-dependent MAR for $M = 5$ when 25% of values are missing in X_1 . The error bars summarise results from the 2000 repetitions and the limits represent the Monte Carlo 95% confidence interval of $\text{Slope}_{imp} - \text{Slope}_{target}$. CC (complete-case); methods are described in Section 2.7 or Table 6.1.

Outcome-dependent MAR

Figure C24 presents the various missing data methods results for the calibration slope when compared to the complete-case, ideal and pragmatic target estimate when data are outcome-dependent or outcome- and covariate-dependent MAR.

Similarly to when data were MCAR or covariate-dependent MAR, for all sample sizes the complete-case analysis tends to overestimate $\text{Slope}_{\text{target},CC}$ ($\text{Slope}_{CC} - \text{Slope}_{\text{target},CC} > 1.5$) and does not fit onto the scale of Figure C24.

The pragmatic performance of all methods overestimates $\text{Slope}_{\text{target},\text{imputed}}$ for all scenarios. When sample size is 300 or 1000, *MI-then-BS impute once* has the largest magnitude $|\text{Slope}_{\text{prag},MI-BS-once} - \text{Slope}_{\text{target},\text{imputed}}|$ (greater than 0.05) of all the methods' pragmatic performance while method *BS-then-MI* tends to have the smallest magnitude. Increasing the sample size to 1000, all methods tend to perform similarly with a magnitude less than 0.025, except for method *MI-then-BS impute once*.

Similarly to the pragmatic performance, the ideal performance of all methods tends to overestimate $\text{Slope}_{\text{target},\text{obs}}$ for all scenarios. Method *BS-then-MI* tends to either overestimate $\text{Slope}_{\text{target},\text{obs}}$ the least across all methods or approximates $\text{Slope}_{\text{target},\text{obs}}$ well. With increasing sample size all methods tend to perform similarly.

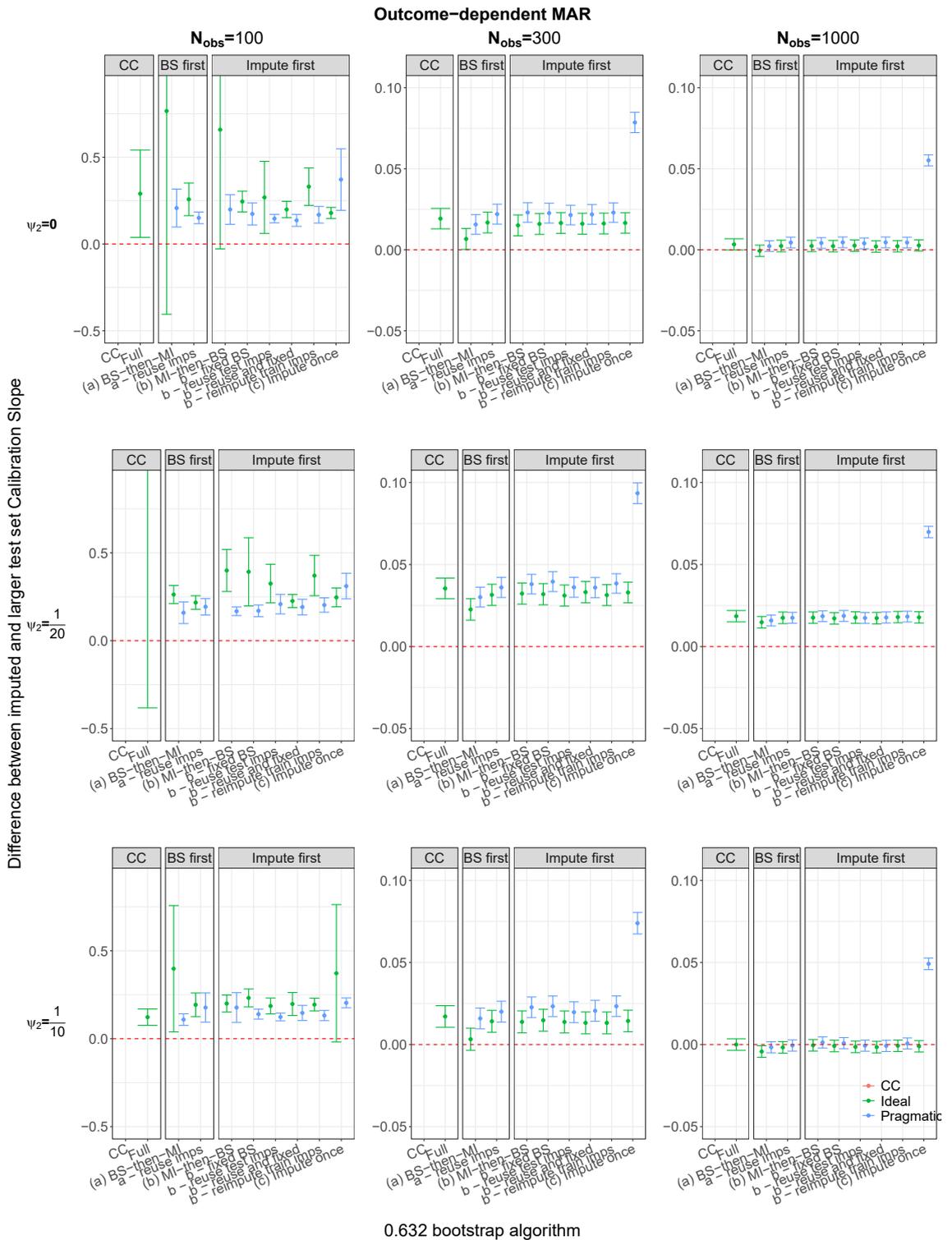


Figure C24: The difference $\text{Slope}_{imp} - \text{Slope}_{target}$ when data are outcome-dependent or outcome- and covariate-dependent MAR for $M = 5$ when 25% of values are missing in X_1 . The error bars summarise results from the 2000 repetitions and the limits represent the Monte Carlo 95% confidence interval of $\text{Slope}_{imp} - \text{Slope}_{target}$. CC (complete-case); methods are described in Section 2.7 or Table 6.1.

C.6 Comparing reusing and re-imputing test imputed datasets of the original dataset

Previously, in Section 6.3I presented results for the continuous outcome case which compared reusing the test imputed datasets (used to estimate the apparent performance) in order to estimate the test performance against re-imputing the original dataset a second time to then use to estimate test performance. It was shown for the MSE that both methods performed similarly. Figure C25 displays the results for the AUC, Brier score and Calibration intercept and slope for pragmatic performance when the sample size is 1000 and data are weakly outcome- and covariate-dependent MAR. The results in Figure C25 are generally representative for ideal and pragmatic performance for all scenarios, all graphs are available in the Supplementary plot sections S4.1.1, S4.2.1 and S4.3.1.

For the AUC and Brier score, both reusing the test imputed datasets or reimputing datasets perform similarly across all scenarios, as seen in Figure C25. For the calibration intercept and slope, reimputing versus reusing test datasets tend to have the same median values but the reusing option tends to be slightly more variable. However, reusing test imputed datasets is more computationally efficient and therefore, all subsequent results for the standard bootstrap algorithm presented below are based on reusing the test imputed datasets.

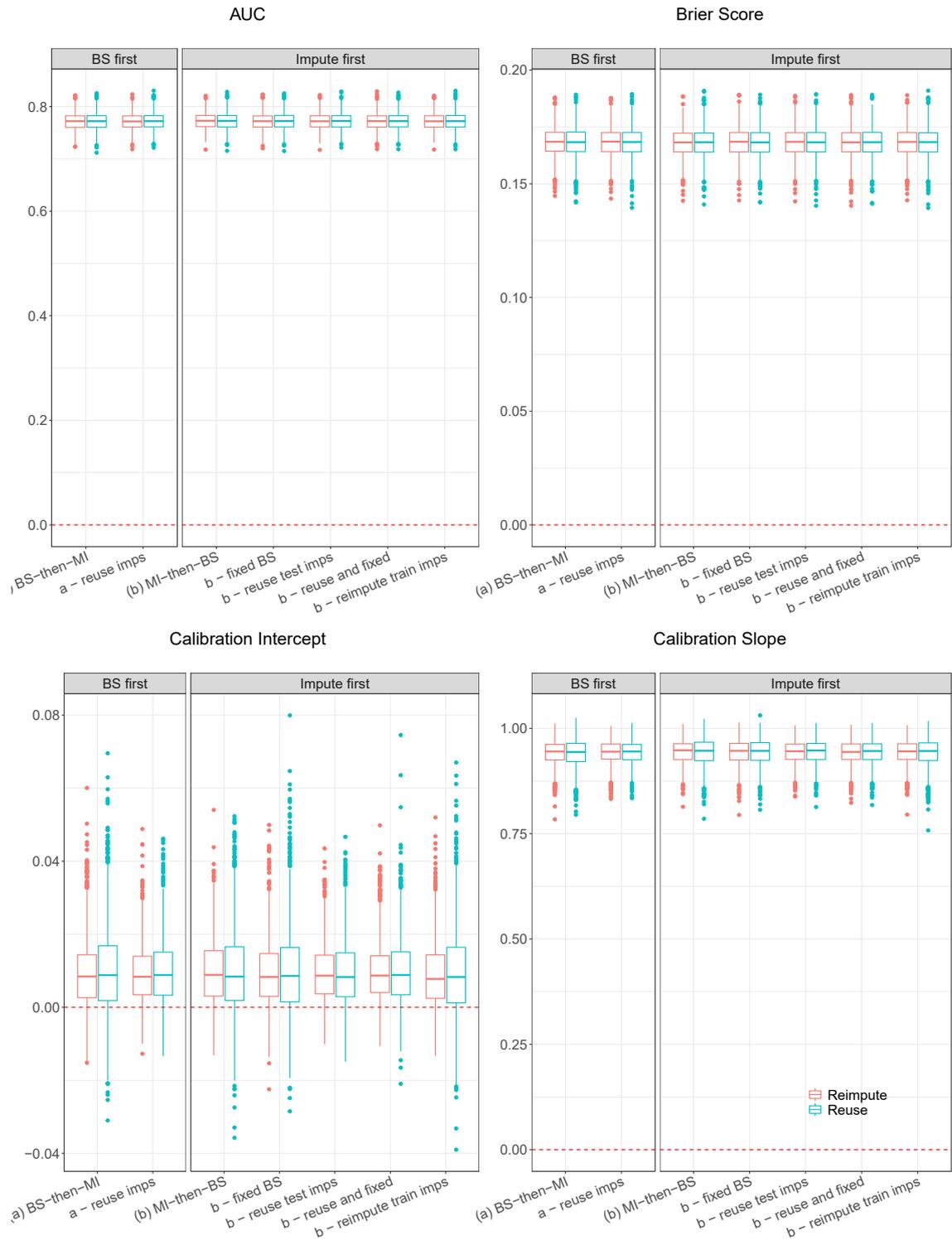


Figure C25: Comparing reusing test imputations for test performance in the standard BS algorithm with reimputing the original dataset using the test imputation model for the AUC, Brier Score and Calibration intercept and slope. The above scenario is for a sample size of 1000 when data are weakly outcome- and covariate-dependent. CC (complete-case); methods are described in Section 2.7 or Table 6.1.

C.7 Overview of results for the standard bootstrap algorithm

C.7.1 Area under the ROC curve

MCAR and covariate-dependent MAR

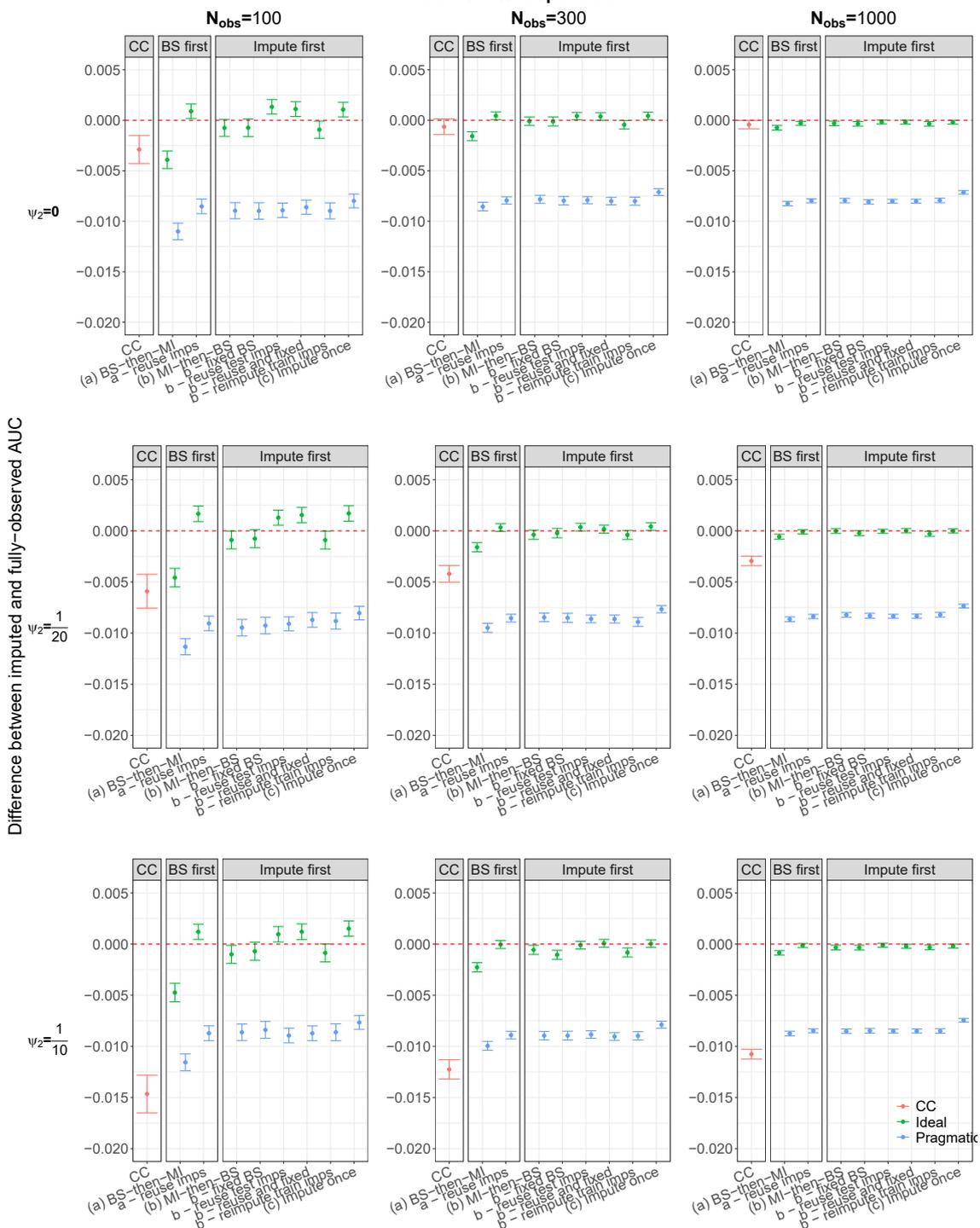
Figure C26 displays the results of the AUC when data are MCAR (top row, $\psi_2 = 0$) or covariate-dependent MAR ($\psi_2 > 0$). Each of the AUCs from the various missing data methods (AUC_{imp}) are compared to the AUC estimated when data are fully-observed (AUC_{obs}) i.e. $AUC_{imp} - AUC_{obs}$.

When data are MCAR and for a sample size of 100, the AUC estimate from the complete-case analysis tends to underestimate the AUC estimate when data are fully-observed ($AUC_{CC} - AUC_{obs} < 0$). With increasing sample size the AUC estimate tends towards the estimate when data are fully-observed ($AUC_{CC} - AUC_{obs} \xrightarrow{n_{obs} \rightarrow \infty} 0$). When missingness is weakly covariate-dependent MAR, the AUC estimate from the complete-case analysis underestimates the AUC estimate when data are fully-observed. With increasing strength of missingness the magnitude of the difference between the complete-case analysis' AUC and the fully-observed estimate increases and it is outperformed by the other missing data methods.

For a sample size of 100, when data are MCAR or weak or strong covariate-dependent MAR, the pragmatic performance of *BS-then-MI* underestimates the AUC estimate when data are fully-observed more than the other imputation based methods do. *MI-then-BS reuseimps* performs similarly to the *MI-then-BS* various methods with method *MI-then-BS impute once* having the smallest difference between its AUC estimate and the fully-observed AUC estimate. With increasing sample size the pragmatic performance of method *BS-then-MI* performs similarly to the other imputation based methods in relation to the fully-observed AUC estimate. For all sample sizes, method *MI-then-BS impute once* has the smallest difference in AUC between its estimate and the estimate when data are fully-observed.

For ideal performance when data are MCAR or covariate-dependent MAR methods *BS-then-MI*, *MI-then-BS*, *MI-then-BS* with fixed bootstrap samples and *MI-then-BS reimpute* training imputed datasets underestimate the fully-observed AUC estimate for a sample size of 100. Method *BS-then-MI* underestimates the fully-observed AUC more than the other methods. Methods *BS-then-MI reuseimps*, *MI-then-BS reuse testimps* (with or without fixed bootstrap samples) and *MI-then-BS impute once* overestimate the fully-observed AUC. With increasing sample size the magnitude of the under- or overestimation decreases and with a sample size of 1000 the methods all perform similarly and approximate the fully-observed AUC estimate well.

MCAR and covariate-dependent MAR



The standard bootstrap algorithm

Figure C26: Error bars of the difference in the AUC from the imputation methods and the AUC estimate when data are fully-observed, with Monte Carlo 95% confidence intervals, when data are MCAR or covariate-dependent MAR. CC (complete-case); methods are described in Section 2.7 or Table 6.1.

Outcome-dependent MAR

When data are outcome-dependent MAR with no dependence of missingness on covariate X_2 , the complete-case analysis tends to underestimate the AUC when data are fully-observed. With increasing sample size from 100 to 1000, the complete-case analysis approximates the fully-observed AUC well. When data are weakly outcome- and covariate-dependent MAR, the magnitude of the difference between the complete-case analysis' AUC estimate and the AUC when data are fully-observed increases. When the strength of missingness on the covariate X_2 increases, the magnitude of the bias also increases. The complete-case analysis is outperformed by all other imputation based methods.

For pragmatic performance, all imputation based methods tend to underestimate the AUC when data are fully-observed ($AUC_{imp} - AUC_{obs} < 0$ where *imp* represents the various methods). For a small sample size of 100, method *BS-then-MI* underestimates the fully-observed AUC estimate the most while the method which underestimates the fully-observed AUC the least is *MI-then-BS impute once*. All other methods perform similarly when compared to the fully-observed AUC. With increasing sample size, method *BS-then-MI* performs similarly to the other imputation based methods.

For ideal performance, similarly to the MCAR and covariate-dependent MAR scenario, methods *BS-then-MI*, *MI-then-BS* (with or without fixed bootstrap samples) and *MI-then-BS reimpute* training imputed datasets underestimate the fully-observed AUC estimate when sample size is 100. Method *BS-then-MI* tends to underestimate the fully-observed AUC estimate the most. Methods *BS-then-MI reuse imps*, *MI-then-BS reuse test imps* (with or without fixed bootstrap samples) and *MI-then-BS impute once* overestimate the fully-observed AUC. With increasing sample size, the magnitude of the difference between the imputation methods' AUC estimate and the fully-observed AUC estimate decreases and by sample size of 1000 all methods perform similarly and approximate the fully-observed AUC estimate ($AUC_{imp} - AUC_{obs} \xrightarrow{n_{obs} \rightarrow \infty} 0$).

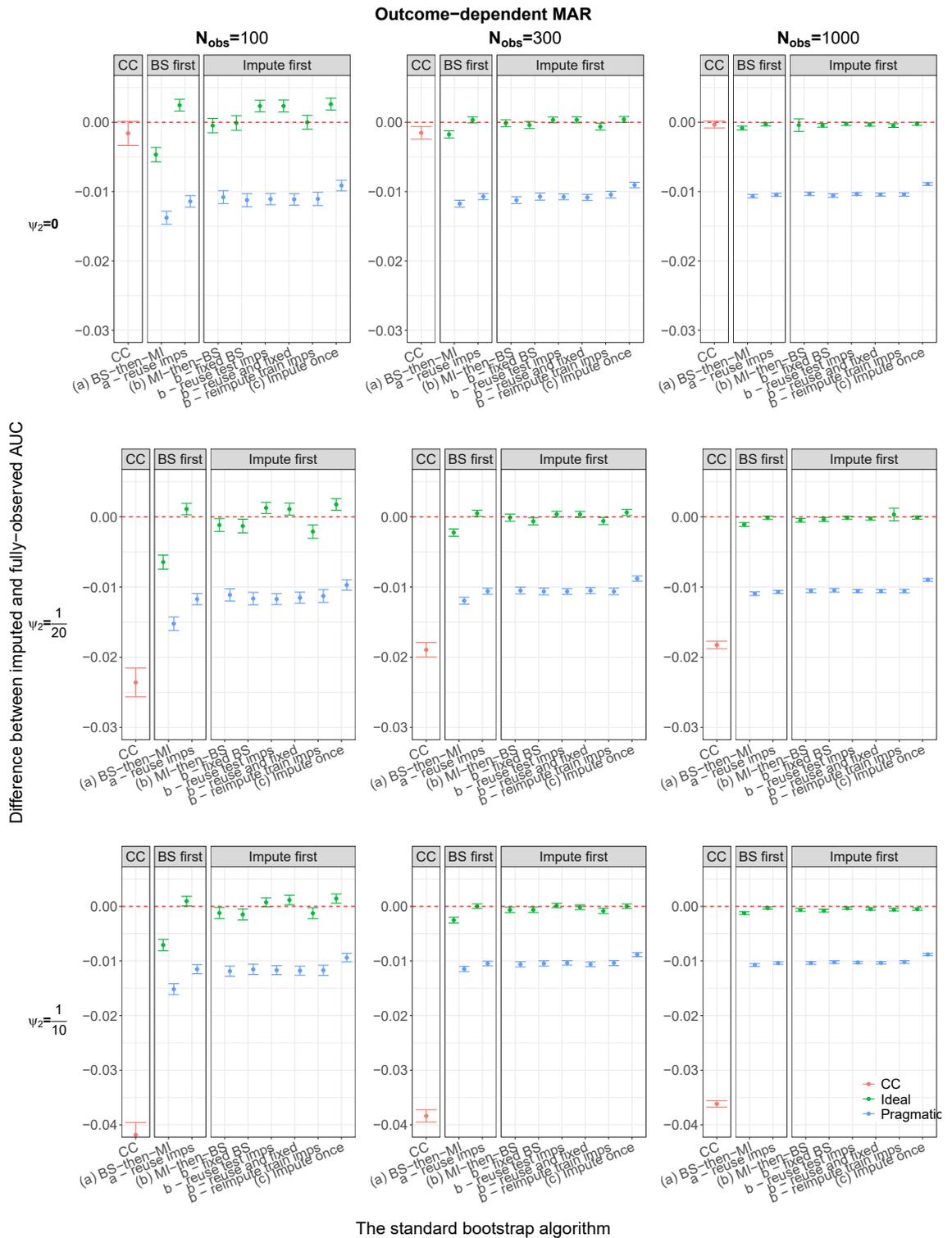


Figure C27: Error bars of the difference in the AUC from the imputation methods and the AUC estimate when data are fully-observed, with Monte Carlo 95% confidence intervals, when data are outcome-dependent MAR. CC (complete-case); methods are described in Section 2.7 or Table 6.1.

Increasing the number of imputed datasets from 5 to 25

Figure C28 displays the results for comparing the various imputation based methods when using 5 or 25 imputed datasets. The results in the graph are for the scenario when data are outcome-dependent MAR but are representative of the results when data are MCAR or covariate-dependent MAR (available in Supplementary Plots S4.1.4). Due to increased computation time when using 25 imputed datasets the comparison a reduced set of methods were assessed. Results are available for methods *BS-then-MI*, *MI-then-BS* and *MI-then-BS impute once* which are based on 1000 repetitions.

The estimates of the AUC for the various methods when using 25 imputed datasets are similar to the performance when only 5 imputed datasets are used.

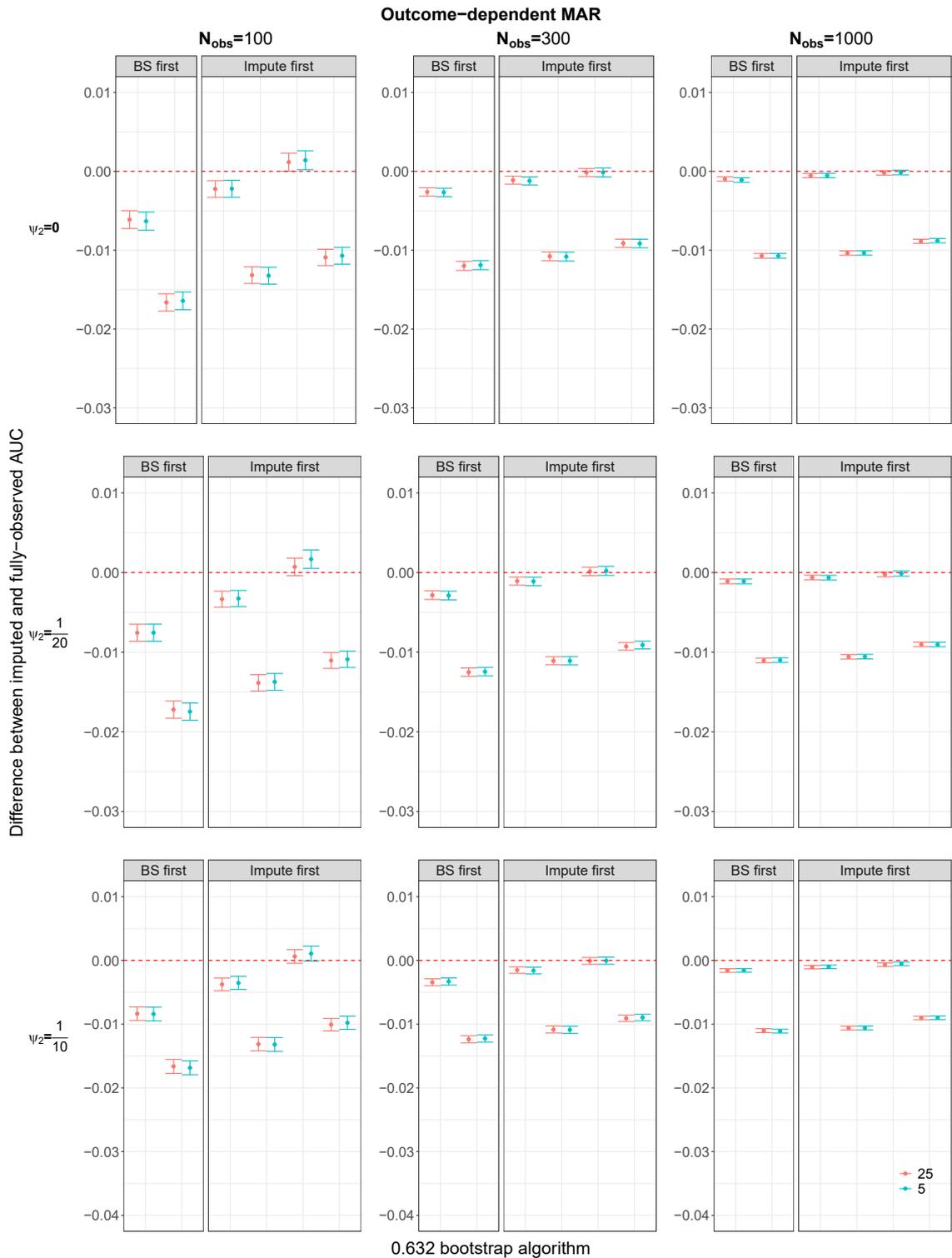


Figure C28: The difference $AUC_{imp} - AUC_{obs}$ when data are outcome-dependent or outcome- and covariate-dependent MAR for $M = 25$ when 25% of values are missing in X_1 . The error bars summarise results from the 2000 repetitions and the limits represent the Monte Carlo 95% confidence interval of $AUC_{imp} - AUC_{obs}$. CC (complete-case); methods are described in Section 2.7 or Table 6.1.

Increasing the proportion of missingness to 40%

Figure C29 compares the pragmatic and ideal performance when 25% of X_1 values are missing to 40% missingness when data are weak outcome- and covariate-dependent MAR. The results in the figure are representative of all missing data scenarios, additional graphs for ideal and pragmatic performance comparing the proportion of missing values increasing from 25% to 40% can be found in supplementary plots Section S4.1.3.

For pragmatic performance, an increase in the proportion of missing values causes increased underestimation of the fully-observed AUC estimate for all missing data methods ($|AUC_{imp,25\%} - AUC_{obs}| < |AUC_{imp,40\%} - AUC_{obs}|$). In addition, the variability across the 2000 repetitions has increased with increasing proportion of missingness, as can be seen when comparing the Monte Carlo 95% confidence intervals in Figure Figure C29. An exception to this is the complete-case analysis method when data are MCAR, the difference between the complete-case analysis' AUC estimate is the same for 25% or 40% missingness, but the variability has increased for 40% missingness.

For ideal performance, in general the increased proportion of missingness has caused an increase in the magnitude of the difference between the various methods' AUC and the AUC estimate when data are fully-observed ($|AUC_{imp,25\%} - AUC_{obs}| < |AUC_{imp,40\%} - AUC_{obs}|$). When data are weakly outcome- and covariate-dependent and the sample size is 1000 *BS-then-MI* when 25% of X_1 values are missing underestimates the fully-observed AUC more than when 40% of values are missing. Similarly, *MI-then-BS* with fixed bootstrap samples when the sample size is 300 underestimates the fully-observed AUC less when 40% of the values are missing. However, the Monte Carlo 95% confidence intervals when 40% of the data are wider and either overlap or encompass the confidence intervals when 25% of values are missing.

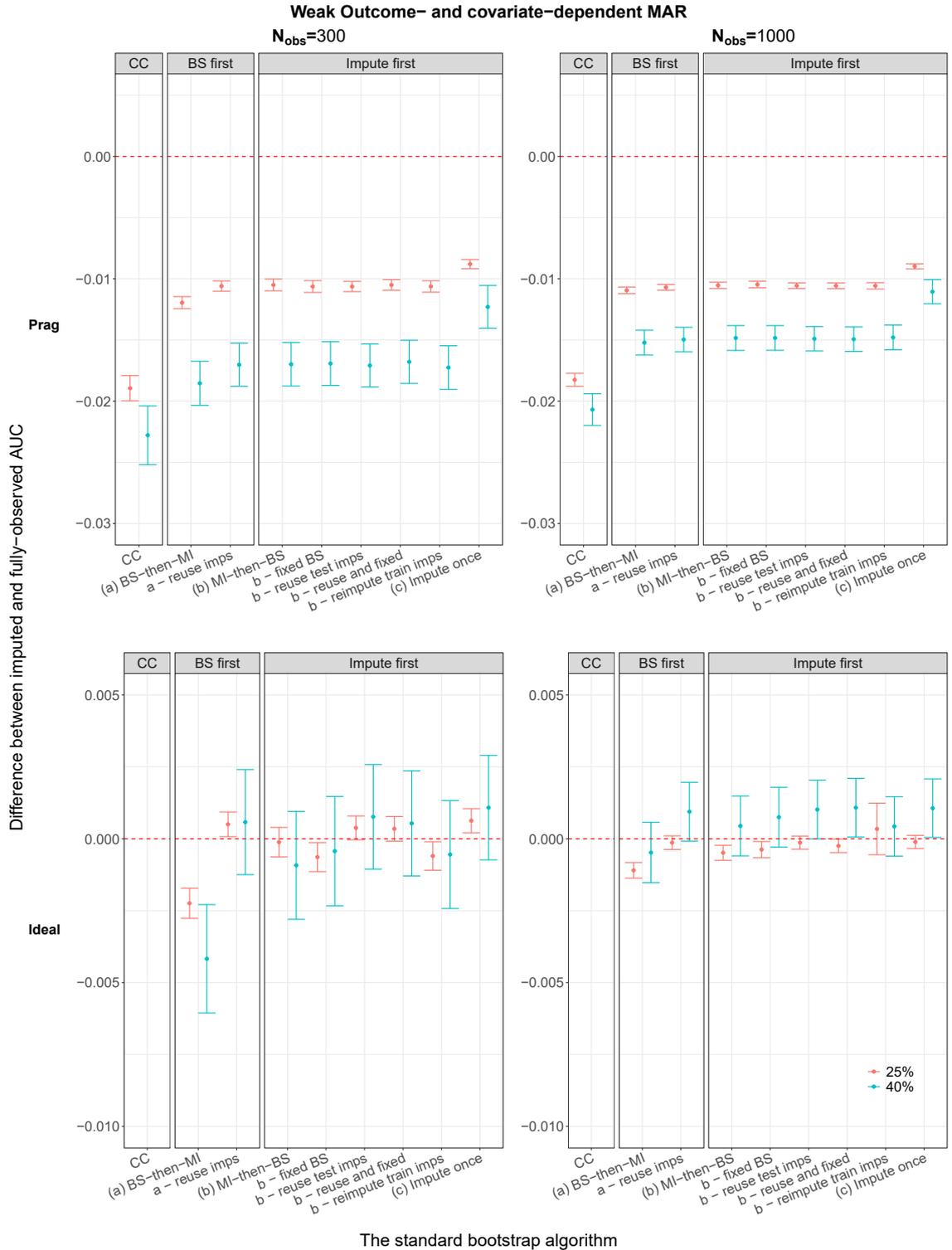


Figure C29: Error bars of the difference in the AUC from the imputation methods and the AUC estimate when data are fully-observed, with Monte Carlo 95% confidence intervals, when data are outcome- and covariate-dependent MAR. The graph compares the AUC estimates when 25% of X_1 values are missing versus 40% missing for ideal and pragmatic performance. CC (complete-case); methods are described in Section 2.7 or Table 6.1.

Comparing to the target performance

For the target AUC, the ideal performance of the bootstrap imputation methods and the AUC estimate when data are fully-observed were compared to the AUC estimate from applying a prediction model, based on all data in a repetition, to the fully-observed data in the larger test set ($AUC_{target,obs}$). The pragmatic performance of the imputation methods is compared to applying a repetition's prediction model to the imputed datasets of the larger test set ($AUC_{target,imputed}$). The complete-case estimate is compared to applying a repetition's prediction model to the observed cases of the larger test set ($AUC_{target,CC}$).

MCAR and covariate-dependent MAR: Figure C30 displays results for comparing the various methods' AUC estimate with their respective ideal, pragmatic or complete-case target AUC, estimated from a larger validation set, when the data are MCAR or covariate-dependent MAR. When data are MCAR, the AUC estimate from the complete-case analysis tends to underestimate $AUC_{target,CC}$ for all sample sizes, although with increasing sample size the variability decreases. For weak covariate-dependent MAR, when sample size is 100 the complete-case analysis approximates $AUC_{target,CC}$ well but with increasing sample size, it tends to overestimate $AUC_{target,CC}$ ($AUC_{CC} - AUC_{target,CC} > 0$). For strong covariate-dependent MAR, for sample sizes of 100 and 300 the complete-case analysis approximates $AUC_{target,CC}$ well, but when increasing sample size to 1000 there is a tendency to overestimate $AUC_{target,CC}$.

When sample size is 100, the ideal performance of methods *BS-then-MI reuseimps*, *MI-then-BS reuse testimps* (with or without fixed bootstrap samples) and *MI-then-BS impute once* tend to underestimate $AUC_{target,obs}$ ($AUC_{imp} - AUC_{target,obs} > 0$). The ideal performance of method *BS-then-MI* tends to underestimate $AUC_{target,obs}$ the most ($AUC_{BS-MI} - AUC_{target,obs} < 0$). The ideal performance of methods *MI-then-BS* (with or without fixed bootstrap samples) and *MI-then-BS reimpute trainimps* tend to underestimate $AUC_{target,obs}$ also. With increasing sample size, the ideal performance of the various imputation methods tends to perform similarly to each other when compared to $AUC_{target,obs}$.

For pragmatic performance, when sample size is 100 method *BS-then-MI* tends to underestimate $AUC_{target,imputed}$ for MCAR or covariate-dependent MAR, although its confidence intervals overlap with zero for MCAR and sample size if 100 and all samples sizes when data are strong covariate-dependent MAR. The other imputation based methods estimate $AUC_{target,imputed}$ well when data are MCAR or strong covariate-dependent MAR but tend to overestimate $AUC_{target,imputed}$ when data are weak covariate-dependent MAR. With increasing sample size all the imputation based methods tend to perform similarly when compared to $AUC_{target,imputed}$.

MCAR and covariate-dependent MAR

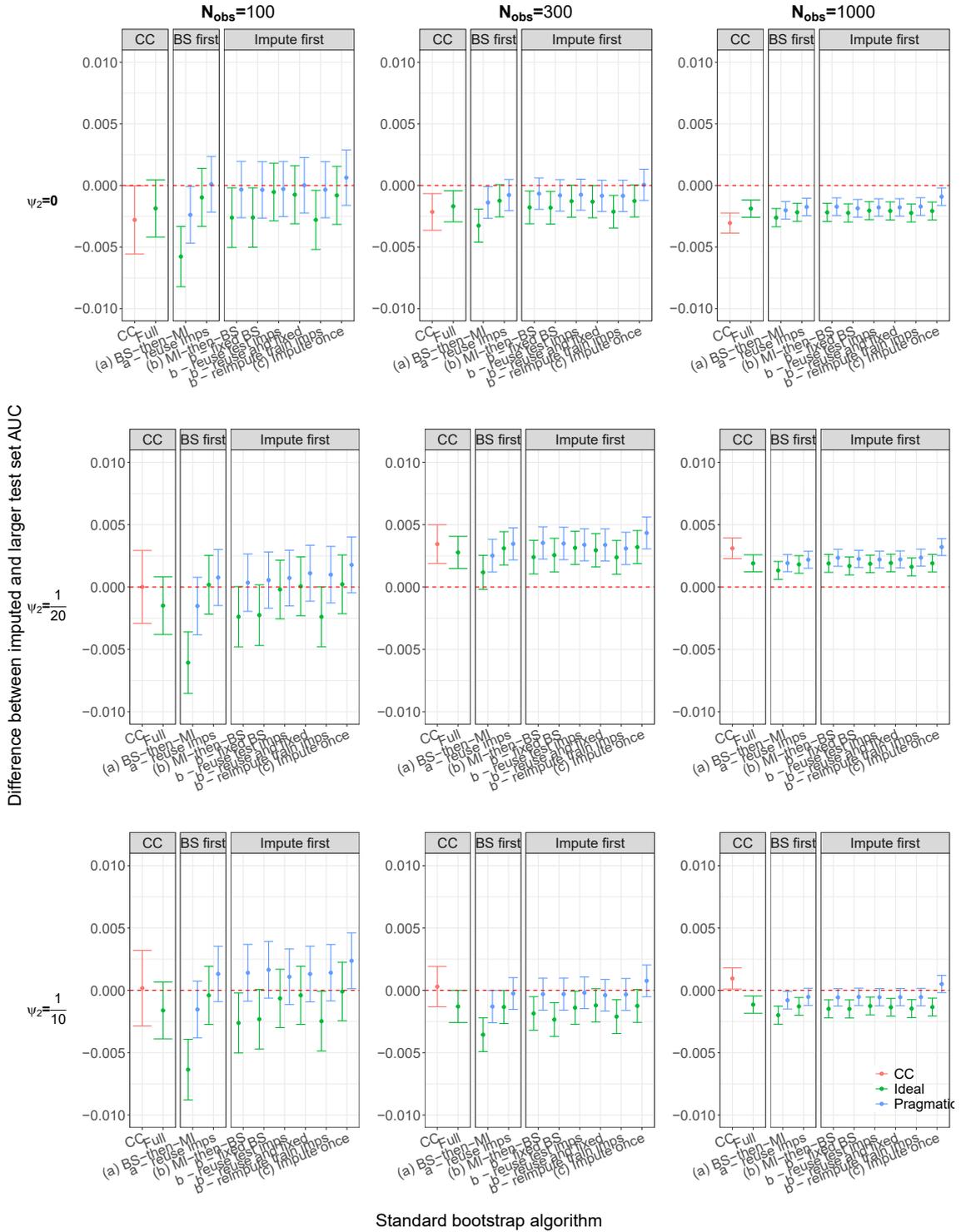


Figure C30: The difference between the imputed AUC and the AUC estimate from a larger test set when data are weakly covariate-dependent MAR. Errorbars represent Monte Carlo 95% confidence intervals. CC (complete-case); methods are described in Section 2.7 or Table 6.1.

Outcome-dependent MAR: Figure C31 displays results for comparing the various methods' AUC estimate with their respective ideal, pragmatic or complete-case target AUC, estimated from a larger validation set, when the data are outcome-dependent or outcome- and covariate-dependent MAR. When data are outcome-dependent MAR, the complete-case analysis AUC estimate tends to underestimate $AUC_{target,CC}$ ($AUC_{CC} - AUC_{target,CC} > 0$). When data are weakly outcome- and covariate-dependent MAR and sample size is 100 or data are weakly outcome- and strongly covariate-dependent MAR, the complete-case analysis approximates $AUC_{target,CC}$ well.

For ideal performance, the AUC estimate from method *BS-then-MI* tends to underestimate the $AUC_{target,obs}$ for the majority of scenarios, even when all other methods overestimate $AUC_{target,obs}$. When the other methods tend to overestimate the target AUC, methods *BS-then-MI reuse imps*, *MI-then-BS reuse test imps* and *MI-then-BS impute once* overestimate $AUC_{target,obs}$ more than the other *MI-then-BS* methods. For a sample size of 1000, all methods tend to perform similarly.

For pragmatic performance, method *BS-then-MI* approximates $AUC_{target,imputed}$ well for outcome-dependent MAR and weak outcome- and covariate-dependent MAR. The pragmatic performance of all other imputation methods tends to overestimate $AUC_{target,imputed}$, this is most noticeable for a sample size of 100 but with increasing sample size the pragmatic performance tends to overestimate $AUC_{target,imputed}$ less. For weak outcome- and strong covariate-dependent MAR the AUC estimate for method *BS-then-MI* tends to underestimate $AUC_{target,imputed}$ while all other imputation based methods tend to approximate the target AUC well for a sample size of 100 and 300. With increasing sample size, the AUC estimate for method *BS-then-MI* tends to underestimate the target AUC estimate less and for a sample size of 1000 all methods perform similarly for pragmatic performance.

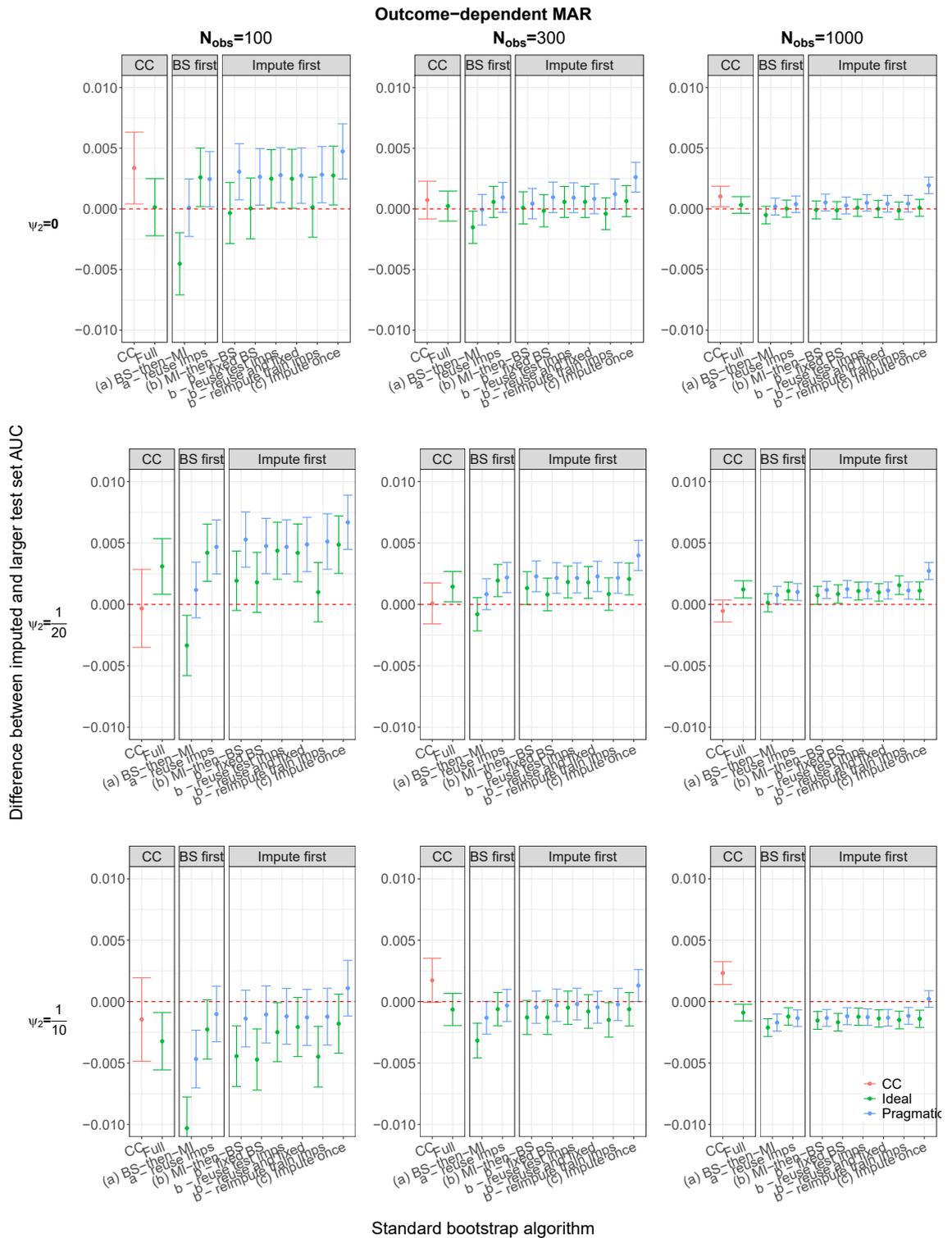


Figure C31: The difference between the imputed AUC and the AUC estimate from a larger test set when data are outcome-dependent or outcome- and covariate-dependent MAR. Errorbars represent Monte Carlo 95% confidence intervals. CC (complete-case); methods are described in Section 2.7 or Table 6.1.

C.7.2 Brier Score

MCAR and covariate-dependent MAR

Figure C32 displays the results of the Brier score when data are MCAR or covariate-dependent MAR. Each of the Brier scores from the various missing data methods (Brier_{imp}) are compared to the Brier score estimated when data are fully-observed (Brier_{obs}) i.e. $\text{Brier}_{imp} - \text{Brier}_{obs}$.

When data are MCAR, the complete-case analysis tends to overestimate the Brier score estimate when data are fully-observed ($\text{Brier}_{CC} - \text{Brier}_{obs} > 0$). However, this overestimate is less than 0.005, and decreases with increasing sample size. For weak covariate-dependent MAR, the complete-case analysis estimate of the Brier score tends to underestimate the Brier score estimated when data are fully-observed ($\text{Brier}_{CC} - \text{Brier}_{obs} < 0$) i.e. the complete-case analysis estimate tends to be over-optimistic compared to the estimate that would have been observed had there been no missing data. With increasing strength of missingness, the magnitude of the underestimation of the fully-observed Brier score estimate increases.

The pragmatic performance of all imputation based methods tends to overestimate the Brier score estimate when data are fully-observed for MCAR and weak or strong covariate-dependent MAR. For a small sample size of 100 method *BS-then-MI* tends to have the largest magnitude of overestimation of the Brier score when data are fully-observed, while *MI-then-BS* has the smallest magnitude. With increasing sample size to 300 and 1000 all methods tend to perform similarly.

For ideal performance, when sample size is 100 methods *BS-then-MI*, *MI-then-BS* (with or without fixed bootstrap samples) and *MI-then-BS reimpute train imps* tend to overestimate the Brier score estimated when data are fully-observed ($\text{Brier}_{imp} - \text{Brier}_{obs} > 0$) while the other methods which involve reusing imputed datasets or *MI-then-BS impute once* tend to underestimate the fully-observed Brier score estimate. The magnitude of the difference between all methods' Brier score estimates and the estimate when data are fully-observed is less than 0.005 for all sample sizes ($|\text{Brier}_{imp} - \text{Brier}_{obs}| < 0.005$).

MCAR and covariate-dependent MAR

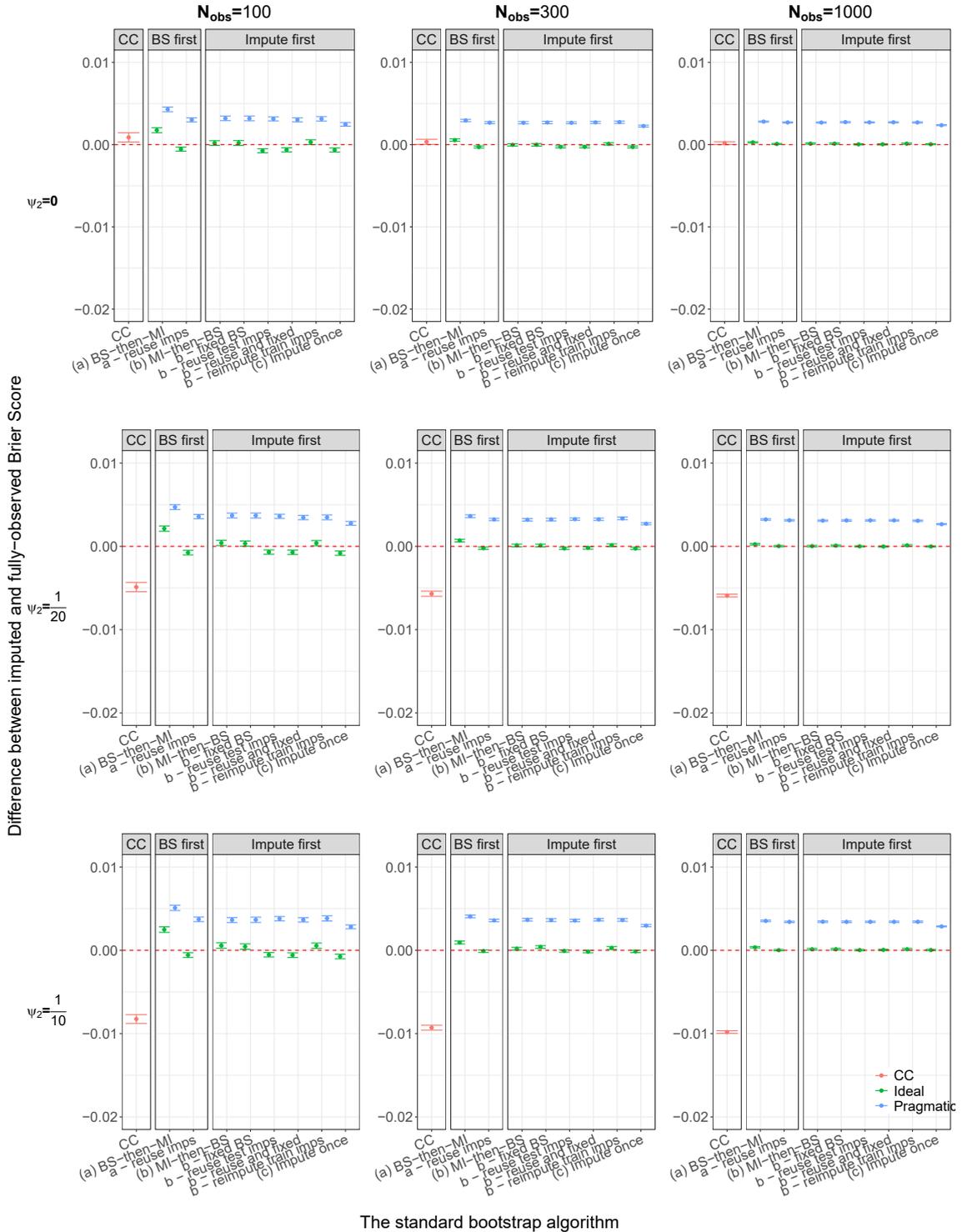


Figure C32: Error bars of the difference in the Brier score from the imputation methods and the Brier score estimate when data are fully-observed, with Monte Carlo 95% confidence intervals, when data are MCAR or covariate-dependent MAR. CC (complete-case); methods are described in Section 2.7 or Table 6.1.

Outcome-dependent MAR

When data are outcome-dependent MAR or outcome- and covariate-dependent MAR, the complete-case analysis tends to underestimate the Brier score estimate when data are fully-observed ($\text{Brier}_{CC} - \text{Brier}_{obs} < -0.01$). With increased strength of missingness the magnitude of the underestimation increases i.e. the complete-case analysis estimate becomes more optimistic.

Similarly to the MCAR and covariate-dependent MAR scenarios, the pragmatic performance of the imputation based methods tends to overestimate the Brier score estimated when data are fully-observed. As before, for a sample size of 100 the method *BS-then-MI* tends to overestimate the fully-observed Brier score estimate the most, while method *MI-then-BS impute once* tends to overestimate it the least. With increasing sample size all methods tend to perform similarly in relation to the fully-observed estimate with a magnitude less than 0.0075 ($|\text{Brier}_{imp} - \text{Brier}_{obs}| < 0.0075$).

Similarly again, for the ideal performance methods *BS-then-MI*, *MI-then-BS* (with or without fixed bootstrap samples) and *MI-then-BS reimpute train imps* tend to overestimate the Brier score estimated when data are fully-observed ($\text{Brier}_{imp} - \text{Brier}_{obs} > 0$), while the other methods which involve reusing imputed datasets or *MI-then-BS impute once* tend to underestimate the fully-observed Brier score estimate ($\text{Brier}_{imp} - \text{Brier}_{obs} < 0$). Method *BS-then-MI* tends to have the largest magnitude ($|\text{Brier}_{imp} - \text{Brier}_{obs}|$) but with increasing sample size all methods perform similarly with a magnitude less than 0.005 ($|\text{Brier}_{imp} - \text{Brier}_{obs}| < 0.005$).

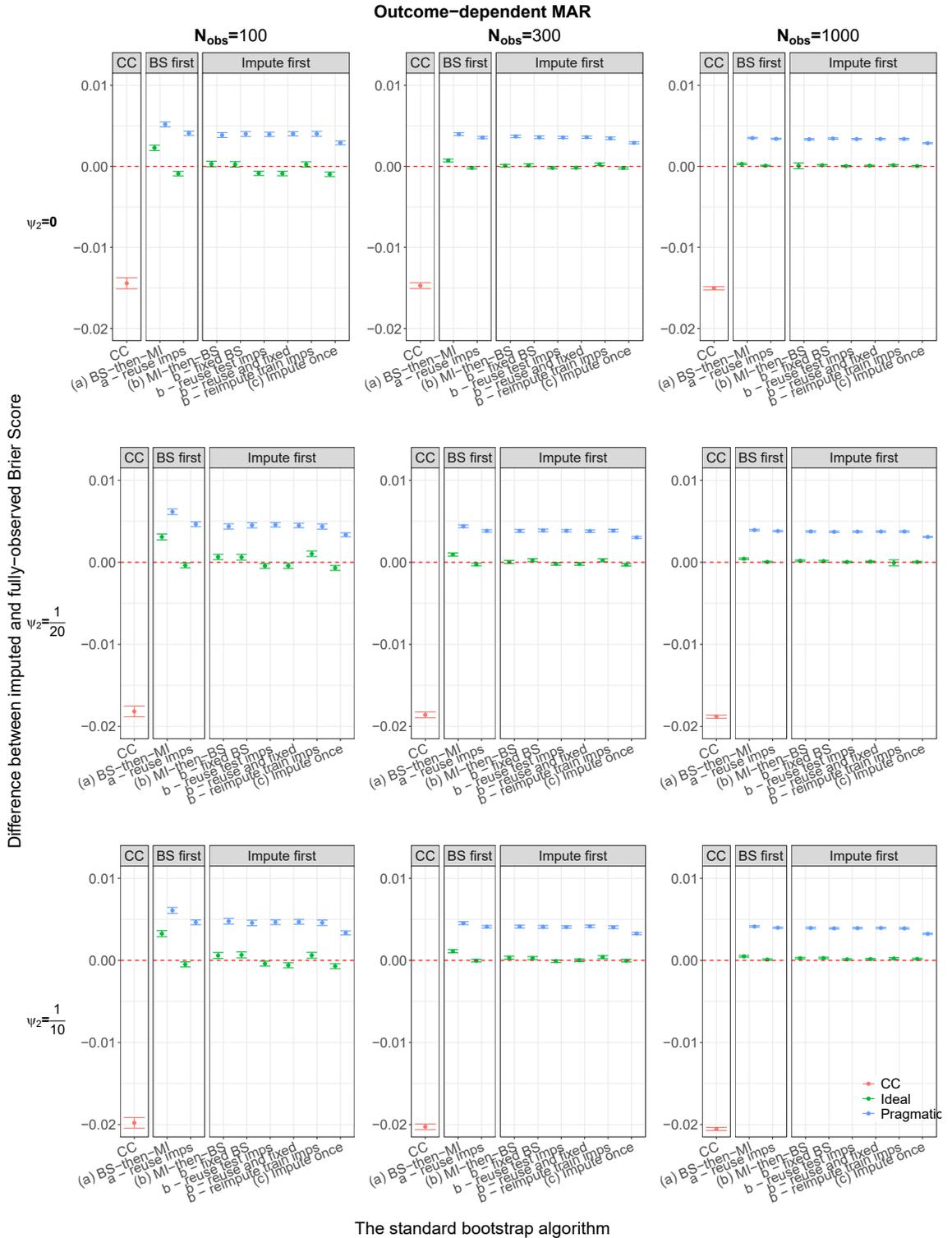


Figure C33: Error bars of the difference in the Brier score from the imputation methods and the Brier score estimate when data are fully-observed, with Monte Carlo 95% confidence intervals, when data are outcome-dependent or outcome- and covariate-dependent MAR. CC (complete-case); methods are described in Section 2.7 or Table 6.1.

Increasing the number of imputed datasets from 5 to 25

Figure C34 displays the results for comparing the various imputation based methods when using 5 or 25 imputed datasets. The results in the graph are for the scenario when data are outcome-dependent MAR but are representative of the results when data are MCAR or covariate-dependent MAR (available in Supplementary Plots S4.2.4). Due to increased computation time when using 25 imputed datasets the comparison a reduced set of methods were assessed. Results are available for methods *BS-then-MI*, *MI-then-BS* and *MI-then-BS impute once* which are based on 1000 repetitions.

The estimates of the Brier score perform similarly in relation to Brier_{obs} , for all methods regardless of whether 5 or 25 imputed datasets are used.

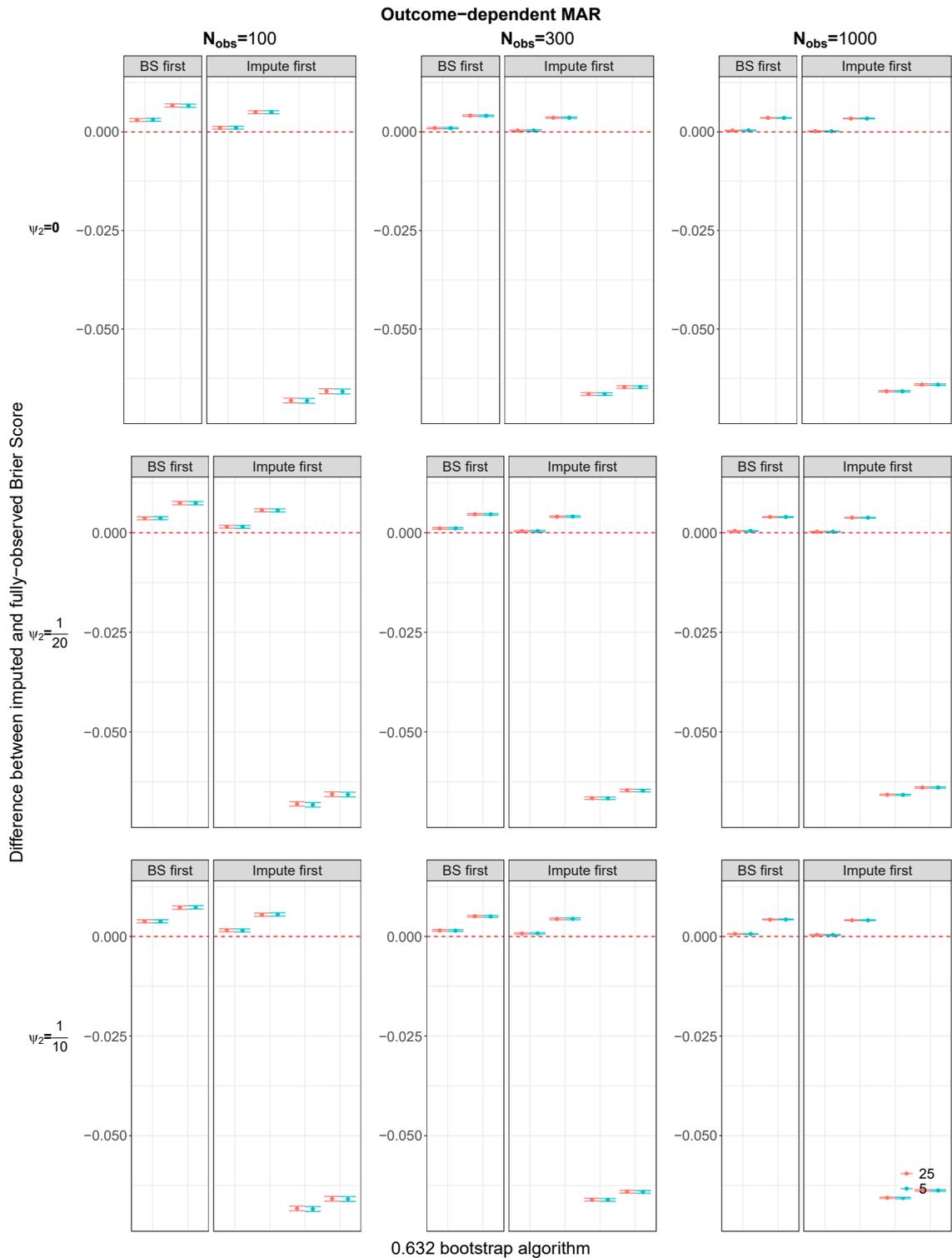


Figure C34: The difference $\text{Brier}_{imp} - \text{Brier}_{obs}$ when data are outcome-dependent or outcome- and covariate-dependent MAR for $M = 25$ when 25% of values are missing in X_1 . The error bars summarise results from the 2000 repetitions and the limits represent the Monte Carlo 95% confidence interval of $\text{Brier}_{imp} - \text{Brier}_{obs}$. CC (complete-case); methods are described in Section 2.7 or Table 6.1.

Increasing the proportion of missingness to 40%

Figure C35 compares the pragmatic and ideal performance when 25% of X_1 values are missing to 40% missingness when data are weak outcome- and covariate-dependent MAR. The results in the figure are representative of all missing data scenarios, additional graphs for ideal and pragmatic performance comparing the proportion of missing values increasing from 25% to 40% can be found in supplementary plots Section S4.2.3.

For the complete-case analysis, the estimate of the Brier score when 40% of X_1 values tends to underestimate the fully-observed estimate of the Brier score more than the estimate of the Brier score when 25% of X_1 values are missing.

For pragmatic performance, the estimate of the Brier score when 40% of X_1 values tends to overestimate the fully-observed estimate of the Brier score more than the estimate of the Brier score when 25% of X_1 values are missing. There is also an increase in the variability across the 2000 repetitions for the 40% missing case when compared to 25% of values missing in X_1 .

For ideal performance, the magnitude of the difference between the Brier score when 40% of X_1 and the Brier score estimate when data are fully-observed is greater than the magnitude for 25% of X_1 values being missing ($|\text{Brier}_{imp,25\%} - \text{Brier}_{obs}| < |\text{Brier}_{imp,40\%} - \text{Brier}_{obs}|$) when data are MCAR or covariate-dependent MAR. In general, when data are outcome-dependent or outcome- and covariate-dependent MAR the magnitude of the difference between the Brier score estimated when 40% of values are missing and the fully-observed estimate is similar to or greater than the Brier estimate comparison when 25% of the values are missing ($|\text{Brier}_{imp,25\%} - \text{Brier}_{obs}| \leq |\text{Brier}_{imp,40\%} - \text{Brier}_{obs}|$). The Monte Carlo 95% confidence intervals when 40% of values are missing are wider and tend to encompass or overlap the confidence intervals when 25% of values are missing. The exception is method *BS-then-MI* whose confidence intervals do not overlap for the majority of scenarios when sample size is 300.

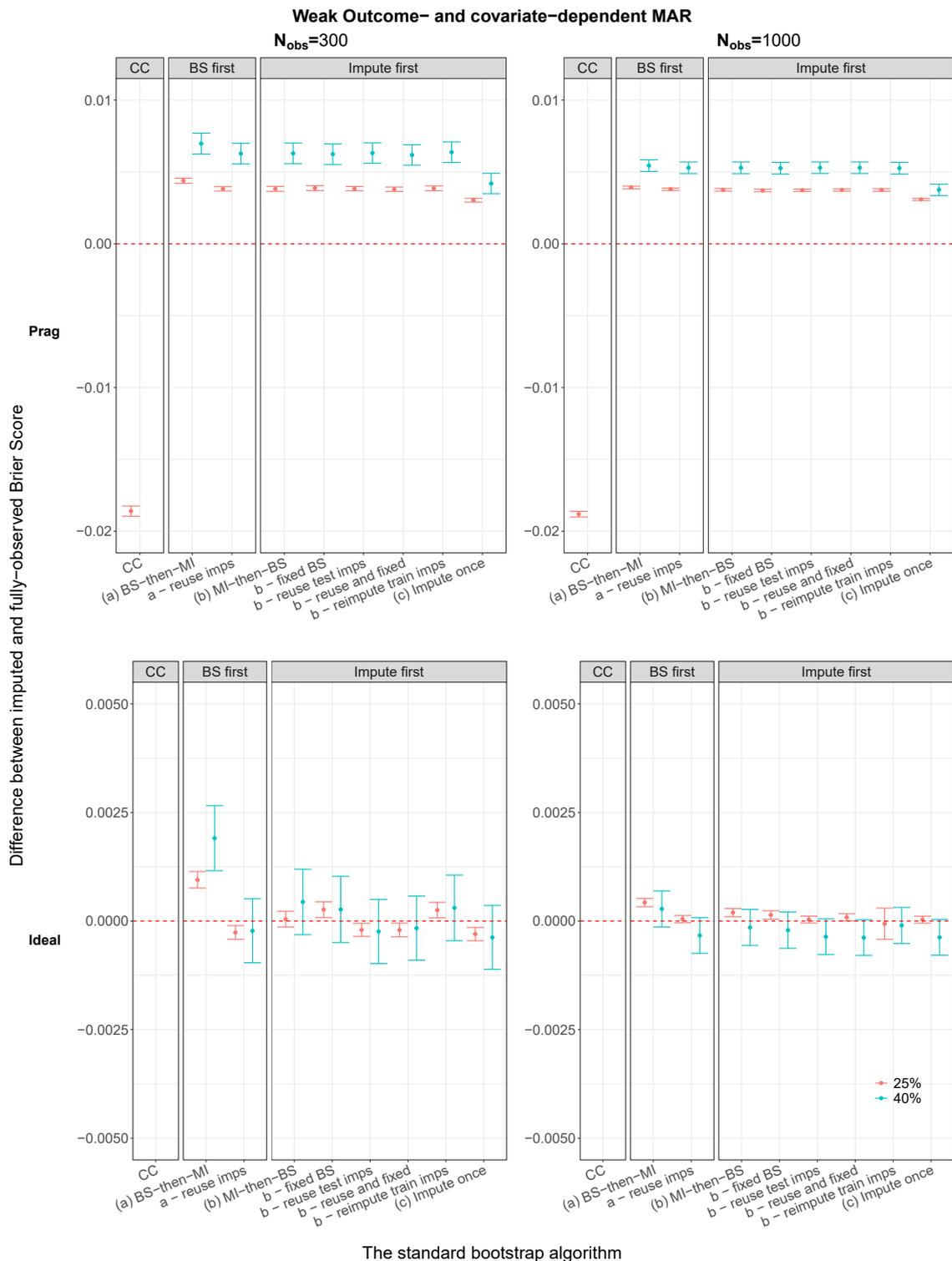


Figure C35: Error bars of the difference in the Brier score from the imputation methods and the Brier score estimate when data are fully-observed, with Monte Carlo 95% confidence intervals, when data are outcome- and covariate-dependent MAR. The graph compares the Brier score estimates when 25% of X_1 values are missing versus 40% missing for ideal and pragmatic performance. CC (complete-case); methods are described in Section 2.7 or Table 6.1.

Comparing to the target performance

As previously described for the MSE and AUC, the ideal performance of the bootstrap imputation methods and the Brier score estimate when data are fully-observed were compared to the ideal performance of the target Brier score which is estimated from applying a prediction model, based on all data in a repetition, to the fully-observed data in the larger test set ($\text{Brier}_{\text{target,obs}}$). The pragmatic performance of the imputation methods is compared to applying a repetition's prediction model to the imputed datasets of the larger test set ($\text{Brier}_{\text{target,imputed}}$). The complete-case estimate is compared to applying a repetition's prediction model to the observed cases of the larger test set ($\text{Brier}_{\text{target,CC}}$).

MCAR and covariate-dependent MAR

When data are MCAR, the complete-case analysis underestimates $\text{Brier}_{\text{target,CC}}$ when the sample size is 100 ($\text{Brier}_{\text{CC}} - \text{Brier}_{\text{target,CC}} < 0$). When sample size is increased to 300, the difference between the complete-case estimate and $\text{Brier}_{\text{target,CC}}$ is centred around zero when data are MCAR or strong-covariate dependent MAR. When data are weak covariate-dependent MAR, the complete-case estimate underestimates $\text{Brier}_{\text{target,CC}}$. With increasing sample size the difference between the complete-case analysis estimate and $\text{Brier}_{\text{target,CC}}$ decreases.

For a sample size of 100 and 300, the pragmatic performance of all imputation-based methods tends to underestimate $\text{Brier}_{\text{target,imputed}}$ ($\text{Brier}_{\text{imp}} - \text{Brier}_{\text{target,imputed}} < 0$). The magnitude of this difference is smallest for method *BS-then-MI* and largest for method *MI-then-BS impute once*. All other imputation based methods (*BS-then-MI reuseimps*, *MI-then-BS* with or without fixed bootstrap samples, *MI-then-BS reuse testimps*, *MI-then-BS reimpute*) perform similarly. For a sample size of 1000, all methods perform similarly when compared to $\text{Brier}_{\text{target,imputed}}$, either overestimating (MCAR or strong covariate-dependent MAR) or underestimating (weak covariate-dependent MAR) $\text{Brier}_{\text{target,imputed}}$.

For a sample size of 100 and 300, the ideal performance of all imputation methods tends to underestimate $\text{Brier}_{\text{target,obs}}$. The exception is when the sample size is 300 and data are strong covariate-dependent MAR as the ideal performance of method *BS-then-MI* tends to slightly overestimate $\text{Brier}_{\text{target,obs}}$. When comparing the ideal performance of the methods with $\text{Brier}_{\text{target,obs}}$, the underestimation for method *BS-then-MI* tends to have the smallest magnitude while method *MI-then-BS impute once* tends to have the largest magnitude. When sample size is 1000, all methods perform similarly when compared to $\text{Brier}_{\text{target,obs}}$, either under- or overestimating the ideal target estimate in the same direction as the pragmatic performance discussed previously.

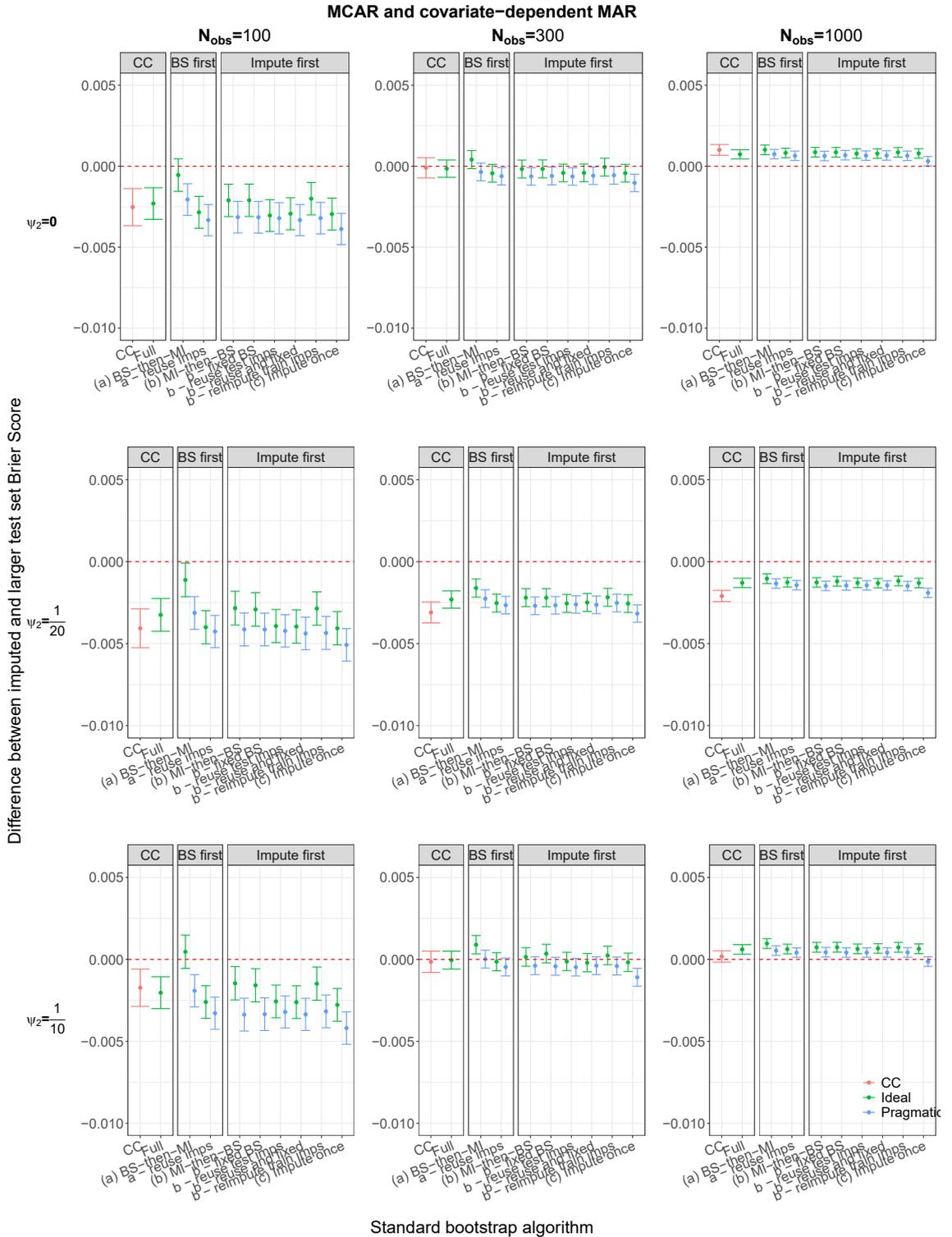


Figure C36: The difference between the imputed Brier score and the Brier score estimate from a larger test set when data are MCAR or covariate-dependent MAR. Errorbars represent Monte Carlo 95% confidence intervals. CC (complete-case); methods are described in Section 2.7 or Table 6.1.

Outcome-dependent MAR

The complete-case analysis tends to underestimate $\text{Brier}_{target,CC}$ for all sample sizes and missing data scenarios. When sample size is 100, the magnitude of this underestimation is greater than 0.005 ($|\text{Brier}_{CC} - \text{Brier}_{target,CC}| > 0.005$) but with increasing sample size to 1000, the magnitude decreases to approximately 0.0025.

For pragmatic performance the results are similar to the MCAR and covariate-dependent MAR scenarios. The pragmatic performance of all imputation based methods tend to underestimate $\text{Brier}_{target,imputed}$ for sample sizes of 100 and 300. Method *BS-then-MI* tends to underestimate $\text{Brier}_{target,imputed}$ the least while method *MI-then-BS impute once* tends to underestimate it the most. For a sample size of 1000 the methods tend to perform similarly, although Method *MI-then-BS impute once* still has the largest magnitude of underestimation. When sample size is 1000 and data are weak outcome-dependent and strong covariate-dependent MAR, the pragmatic performance of method *MI-then-BS impute once* underestimates $\text{Brier}_{target,imputed}$ (i.e. it is over-optimistic), while all other imputation based methods overestimate the target estimate.

For ideal performance when sample size is 100 or 300, method *BS-then-MI* has the lowest magnitude of underestimation when compared to $\text{Brier}_{target,obs}$. Methods *BS-then-MI reuse imps*, *MI-then-BS reuse test imps* and *MI-then-BS impute once* have the largest magnitude of underestimation when compared to the ideal target estimate. For a sample size of 1000, all methods tend to perform similarly when compared to $\text{Brier}_{target,obs}$.

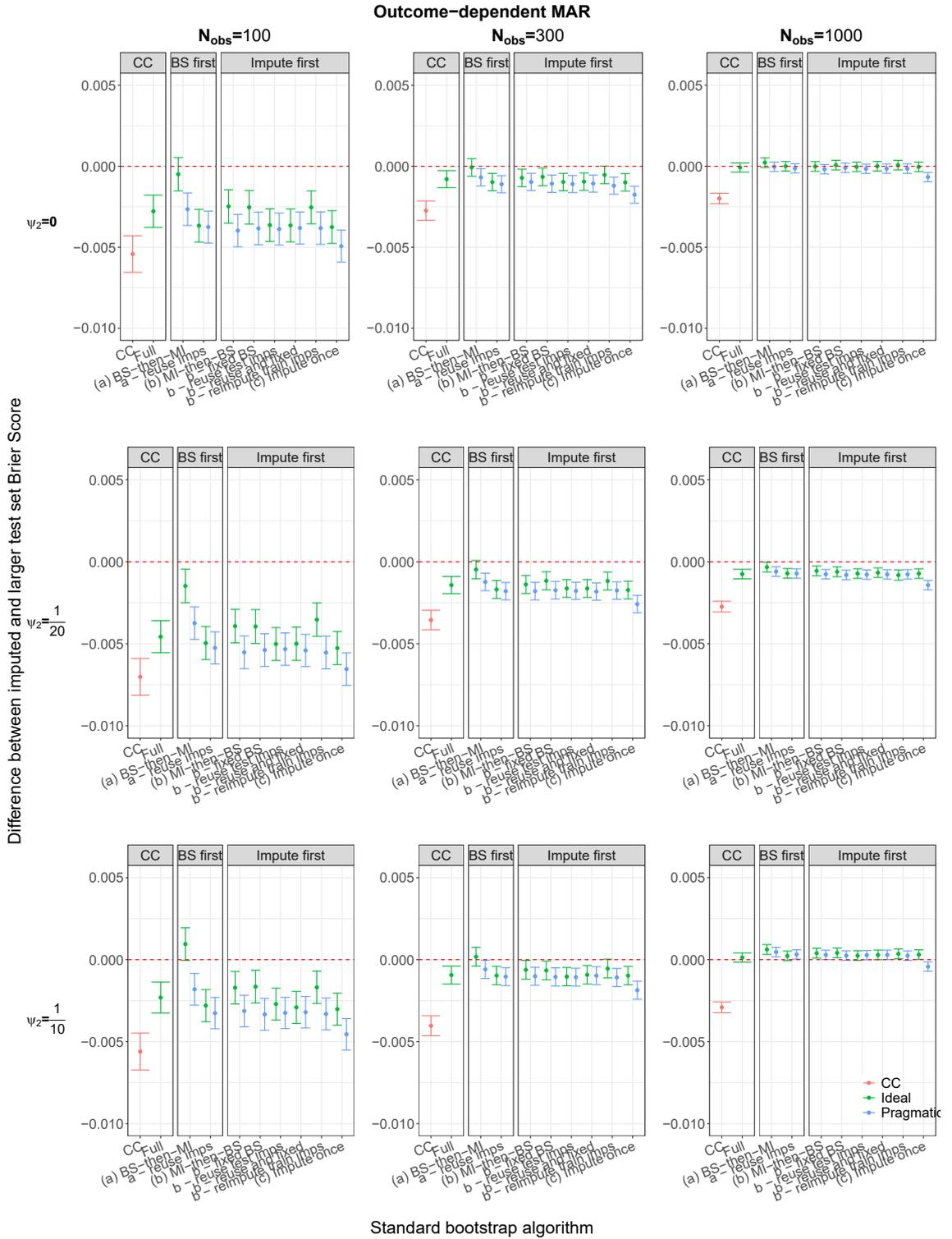


Figure C37: The difference between the imputed Brier score and the Brier score estimate from a larger test set when data are outcome-dependent or outcome- and covariate-dependent MAR. Errorbars represent Monte Carlo 95% confidence intervals. CC (complete-case); methods are described in Section 2.7 or Table 6.1.

C.7.3 Calibration intercept

MCAR and covariate-dependent MAR

Figure C38 displays the results for the various missing data methods' estimate of the calibration intercept which is compared to the estimate of the calibration intercept when data are fully-observed ($\text{Intercept}_{imp} - \text{Intercept}_{obs}$).

For MCAR and covariate-dependent MAR, the complete-case estimate tends to underestimate the calibration intercept estimated when data are fully-observed ($\text{Intercept}_{CC} - \text{Intercept}_{obs} < 0$). With increasing sample size to 1000, the magnitude of this difference tends to decrease.

For a sample size of 100, the ideal and pragmatic performance estimates of the calibration intercept estimate are very unstable with a large magnitude when compared to the intercept estimated when data are fully-observed. The estimates of the calibration intercept for a sample size of 100 when data are fully-observed were noted to vary widely (Chapter 4, Table 5.1).

When sample size is 300 or 1000, the ideal and pragmatic estimates of the various imputation based methods tend to perform similarly to the estimate of the calibration intercept when data are fully-observed. They either under- or overestimate the estimate when data are fully-observed but the Monte-Carlo confidence intervals overlap with zero. The magnitude of the mean estimate of the difference between the ideal or pragmatic performance with the fully-observed estimate tends to be less than 0.001 for all imputation based methods ($|\text{Intercept}_{imp} - \text{Intercept}_{obs}| < 0.001$).

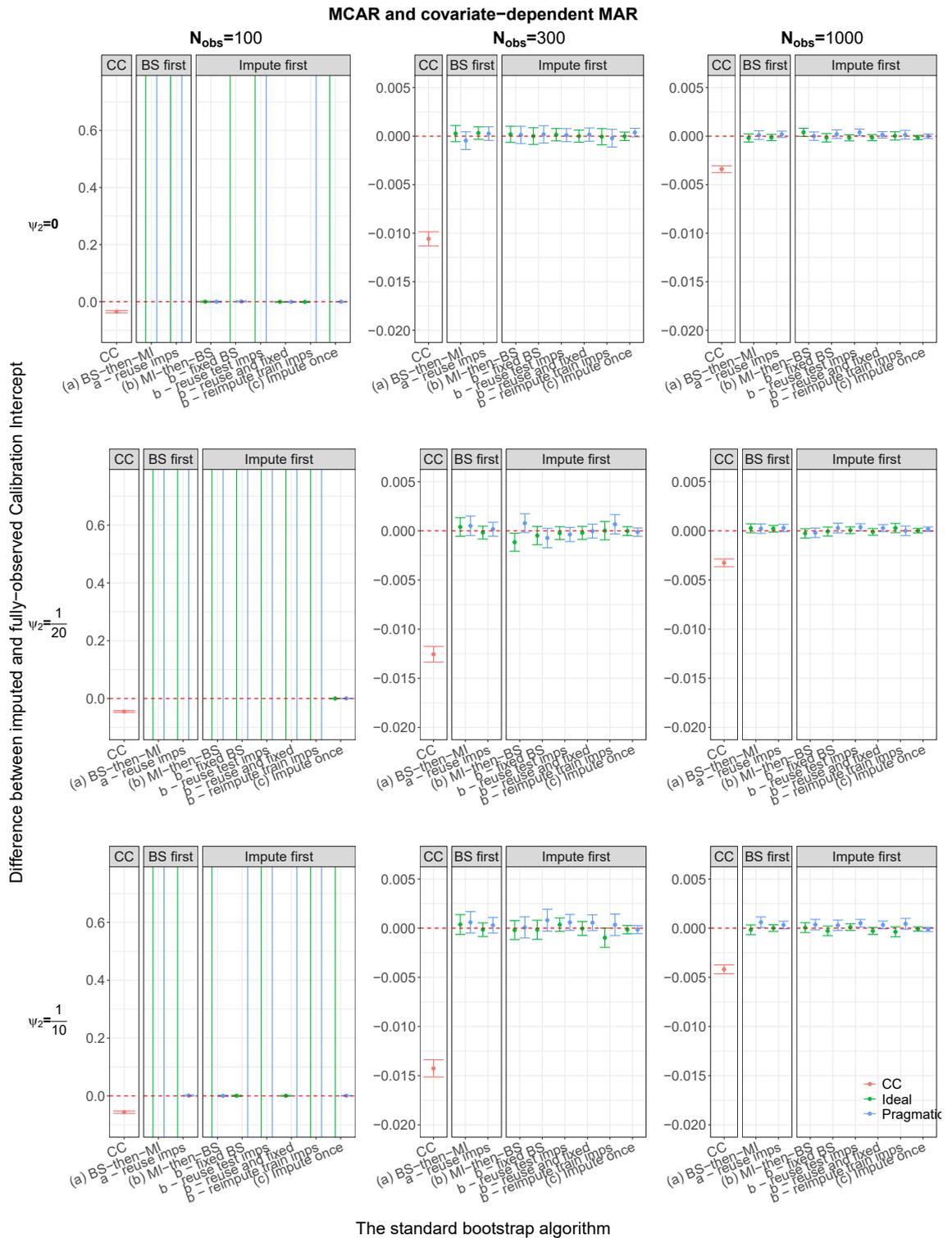


Figure C38: Error bars of the difference in the Calibration intercept estimate from the imputation methods and the calibration intercept estimate when data are fully-observed, with Monte Carlo 95% confidence intervals, when data are MCAR or covariate-dependent MAR. CC (complete-case); methods are described in Section 2.7 or Table 6.1.

Outcome-dependent MAR

Figure C39 displays the results for the various missing data methods' estimate of the calibration intercept when data are outcome-dependent or outcome- and covariate-dependent MAR. These estimates are compared to the estimate of the calibration intercept when data are fully-observed ($\text{Intercept}_{imp} - \text{Intercept}_{obs}$).

The complete-case estimate tends to underestimate the calibration intercept estimated when data are fully-observed ($\text{Intercept}_{CC} - \text{Intercept}_{obs} < 0$). With increasing sample size to 1000, the magnitude of this difference tends to decrease but is still larger than the magnitude when data were MCAR or covariate-dependent MAR.

The pragmatic performance of the various imputation based methods tends to overestimate the estimate of the calibration intercept when data are fully-observed for sample sizes of 300 or 1000 ($\text{Intercept}_{imp} - \text{Intercept}_{obs} > 0$). Method *MI-then-BS impute once* tends to have the smallest magnitude ($|\text{Intercept}_{MI-BS-once} - \text{Intercept}_{obs}|$), performing similarly to the fully-observed estimate of the calibration intercept. The pragmatic performance of all other imputation based methods tends to perform similarly to each other and overestimate the fully-observed estimate by 0.01, approximately.

The ideal performance of the various imputation methods tends to overestimate the the calibration intercept estimate when data are fully-observed for sample sizes of 300 or 1000 ($\text{Intercept}_{imp} - \text{Intercept}_{obs} > 0$). Method *BS-then-MI* tends to have the largest magnitude of overestimation when sample size is 300 and the other methods all perform similarly in relation to the intercept estimated when data are fully-observed. For a sample size of 1000 all methods tend to perform similarly. The magnitude for all methods tends to be less than 0.005 ($|\text{Intercept}_{imp} - \text{Intercept}_{obs}| < 0.005$).

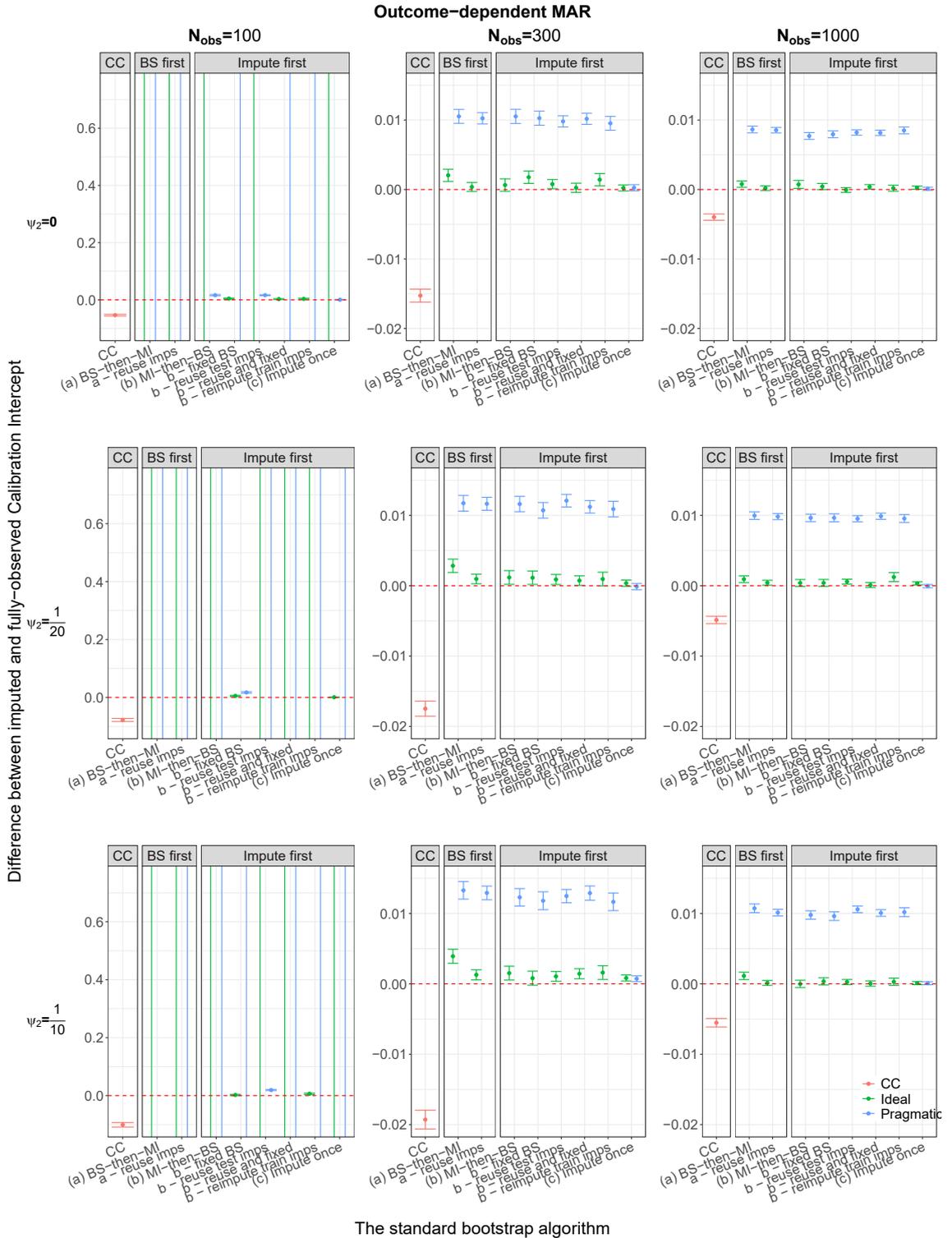


Figure C39: Error bars of the difference in the Calibration intercept estimate from the imputation methods and the calibration intercept estimate when data are fully-observed, with Monte Carlo 95% confidence intervals, when data are outcome-dependent or outcome- and covariate-dependent MAR. CC (complete-case); methods are described in Section 2.7 or Table 6.1.

Increasing the number of imputed datasets from 5 to 25

Figure C40 displays the results for comparing the various imputation based methods when using 5 or 25 imputed datasets. The results in the graph are for the scenario when data are outcome-dependent MAR but are representative of the results when data are MCAR or covariate-dependent MAR (available in Supplementary Plots S4.3.4). Due to increased computation time when using 25 imputed datasets the comparison a reduced set of methods were assessed. Results are available for methods *BS-then-MI*, *MI-then-BS* and *MI-then-BS impute once* which are based on 1000 repetitions.

The estimates for the comparison of the calibration intercept for the various methods to Intercept_{obs} , perform similarly regardless of whether 5 or 25 imputed datasets are used.

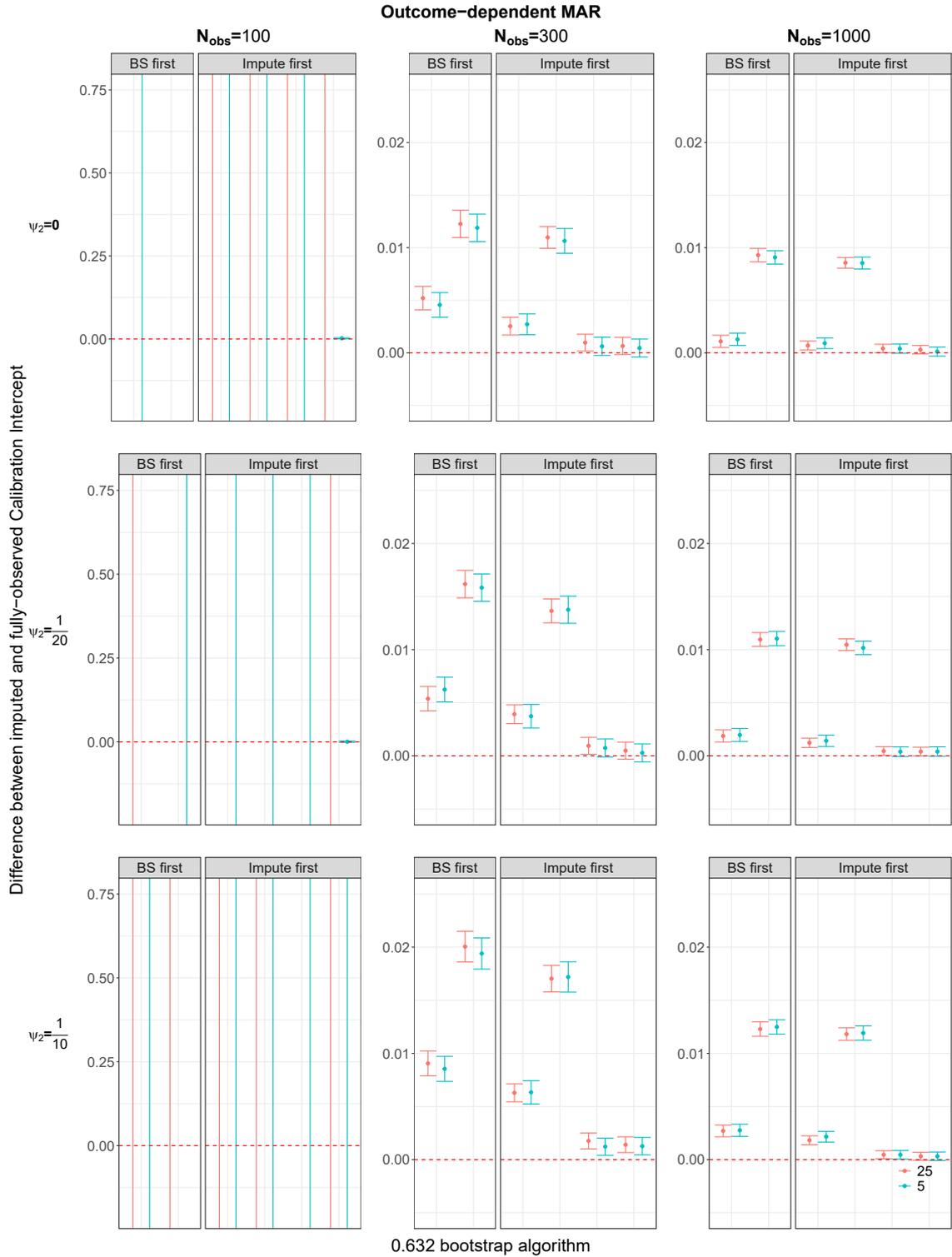


Figure C40: The difference $\text{Intercept}_{imp} - \text{Intercept}_{obs}$ when data are outcome-dependent or outcome- and covariate-dependent MAR for $M = 25$ when 25% of values are missing in X_1 . The error bars summarise results from the 2000 repetitions and the limits represent the Monte Carlo 95% confidence interval of $\text{Intercept}_{imp} - \text{Intercept}_{obs}$. CC (complete-case); methods are described in Section 2.7 or Table 6.1.

Increasing the proportion of missingness to 40%

Figure C41 compares the pragmatic and ideal performance when 25% of X_1 values are missing to 40% missingness when data are weak outcome- and covariate-dependent MAR. The results in the figure are generally representative of all missing data scenarios, additional graphs for ideal and pragmatic performance comparing the proportion of missing values increasing from 25% to 40% can be found in supplementary plots Section S4.3.3.

For complete-case analysis, the estimate of the calibration intercept when 40% of X_1 values are missing tends to be larger in magnitude than the comparison when 25% of values are missing when data are MAR. For MCAR, the estimate when 40% of values are missing compares similarly to when 25% of values are missing. When sample size increases from 300 to 1000, the estimate when 40% of values are missing tends towards the estimate when 25% of values are missing ($|\text{Intercept}_{imp,40\%} - \text{Intercept}_{obs}| \rightarrow |\text{Intercept}_{imp,25\%} - \text{Intercept}_{obs}|$).

For pragmatic performance, the estimate of the calibration intercept when 40% of X_1 values are missing tends to be similar to or larger in magnitude than the comparison when 25% of values are missing when data are MCAR or covariate-dependent MAR for all methods ($|\text{Intercept}_{imp,40\%} - \text{Intercept}_{obs}| \geq |\text{Intercept}_{imp,25\%} - \text{Intercept}_{obs}|$). When data are outcome-dependent or outcome and covariate-dependent the calibration intercept estimate when 40% of values are missing tends to have a larger magnitude than when 25% of values are missing ($|\text{Intercept}_{imp,40\%} - \text{Intercept}_{obs}| > |\text{Intercept}_{imp,25\%} - \text{Intercept}_{obs}|$). The exception to this is *MI-then-BS impute once* whose estimates perform similarly regardless of 25% or 40% of X_1 values are missing.

For ideal performance, the calibration intercept estimate when 40% of values are missing tends to perform similarly to when 25% of values are missing when data are MCAR. For covariate-dependent, outcome-dependent or outcome- and covariate-dependent MAR the estimate when 40% of values are missing tends to have a similar or larger magnitude than the estimate when 25% of values are missing ($|\text{Intercept}_{imp,40\%} - \text{Intercept}_{obs}| \geq |\text{Intercept}_{imp,25\%} - \text{Intercept}_{obs}|$).

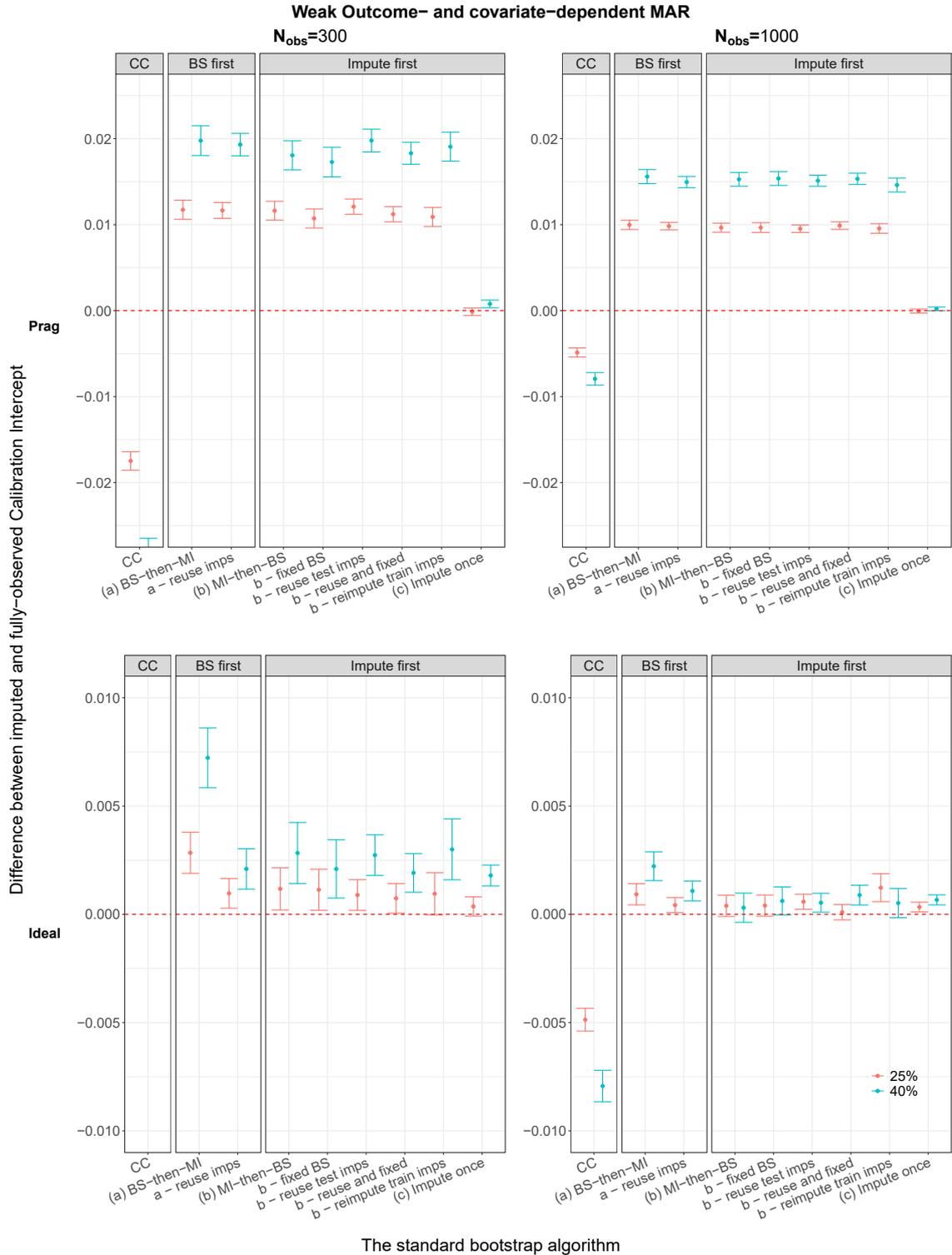


Figure C41: Error bars of the difference in the Calibration intercept from the imputation methods and the intercept estimate when data are fully-observed, with Monte Carlo 95% confidence intervals, when data are outcome- and covariate-dependent MAR. The graph compares the intercept estimates when 25% of X_1 values are missing versus 40% missing for ideal and pragmatic performance. CC (complete-case); methods are described in Section 2.7 or Table 6.1.

Comparing to the target performance

As previously described for the MSE, AUC and Brier score, the ideal performance of the bootstrap imputation methods and the calibration intercept estimate when data are fully-observed were compared to the ideal performance of the target calibration intercept which is estimated from applying a prediction model, based on all data in a repetition, to the fully-observed data in the larger test set ($\text{Intercept}_{\text{target,obs}}$). The pragmatic performance of the imputation methods is compared to applying a repetition's prediction model to the imputed datasets of the larger test set ($\text{Intercept}_{\text{target,imputed}}$). The complete-case estimate is compared to applying a repetition's prediction model to the observed cases of the larger test set ($\text{Intercept}_{\text{target,CC}}$).

As seen when previously comparing the methods' calibration estimates to the intercept estimate when data are fully-observed, the results for sample size of 100 are very unstable for all missing data scenarios and the results will not be further analysed until the Discussion section.

MCAR and covariate-dependent MAR

Figure C42 presents results for the comparison of the various methods' calibration intercept estimate to the target calibration intercept when data are MCAR or covariate-dependent MAR.

The complete-case analysis tends to underestimate $\text{Intercept}_{\text{target,CC}}$, with an approximate value of -0.48, and does not fit onto the scale of the results presented in Figure C42.

The pragmatic performance of the methods involving imputation (*imp*) tends to overestimate $\text{Intercept}_{\text{target,imputed}}$ ($\text{Intercept}_{\text{imp,prag}} - \text{Intercept}_{\text{target,imputed}} > 0$). The magnitude of this difference is less than 0.05 for MCAR and weak covariate-dependent MAR, and less than 0.0625 for strong covariate-dependent MAR. All methods have similar pragmatic performance when compared to $\text{Intercept}_{\text{target,imputed}}$. With increasing sample size the magnitude of the difference decreases, as does the variability across the 2000 repetitions.

The ideal performance of the various imputation based methods tends to approximate or overestimate $\text{Intercept}_{\text{target,obs}}$ when data are MCAR or strong covariate-dependent MAR. When data are weak covariate-dependent MAR, the methods tend to underestimate $\text{Intercept}_{\text{target,obs}}$. The magnitude of the mean difference tends to be less than 0.025 ($\text{Intercept}_{\text{imp,ideal}} - \text{Intercept}_{\text{target,imputed}} < 0.025$). With increasing sample size the magnitude of the difference decreases. All methods tend to perform similarly when compared to $\text{Intercept}_{\text{target,obs}}$.

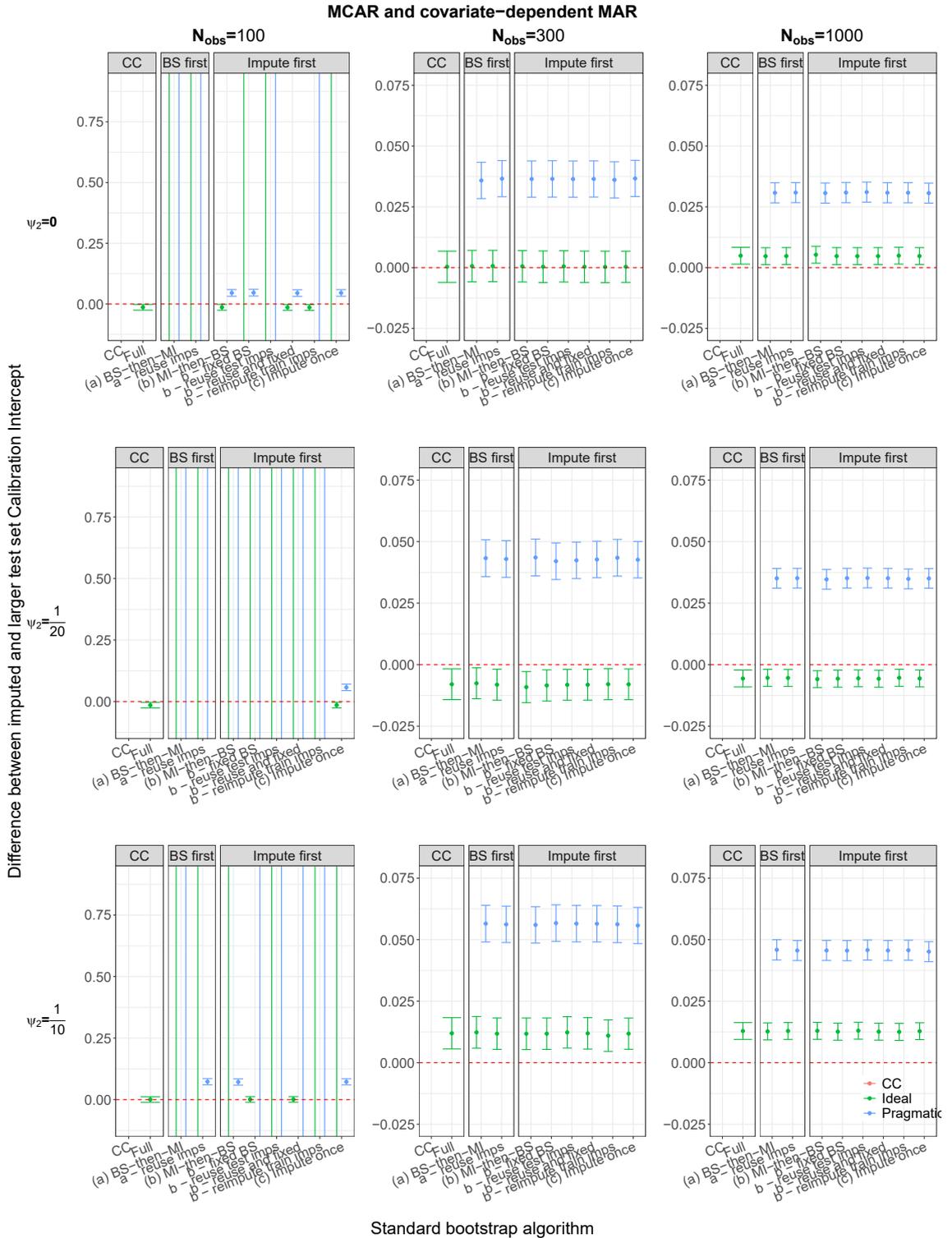


Figure C42: Error bars of the difference in the Calibration intercept estimate from the imputation methods and the target estimate of the calibration intercept, with Monte Carlo 95% confidence intervals, when data are MCAR or covariate-dependent MAR. CC (complete-case); methods are described in Section 2.7 or Table 6.1.

Outcome-dependent MAR

FigureC43 presents the comparison of the various missing data methods calibration intercept estimate with the complete-case, ideal and pragmatic target estimate.

Similarly to the MCAR and covariate-dependent MAR, the complete-case analysis tends to underestimate $\text{Intercept}_{target,CC}$ and does not fit on the scale used in FigureC43.

The pragmatic performance of the various imputation based methods tends to overestimate $\text{Intercept}_{target,imputed}$ ($0.025 < \text{Intercept}_{imp,prag} - \text{Intercept}_{target,imputed} < 0.062$). The pragmatic performance of the various *BS-then-MI* and *MI-then-BS* methods all perform similarly in relation to $\text{Intercept}_{target,imputed}$, except for method *MI-then-BS impute once* which tends to have a smaller magnitude ($|\text{Intercept}_{imp,prag} - \text{Intercept}_{target,imputed}|$).

The ideal performance of the imputation based methods tends to underestimate $\text{Intercept}_{target,obs}$ when data are weak outcome-dependent and either weak or strong covariate-dependent MAR. When data are weak outcome-dependent MAR, all methods tend to approximate $\text{Intercept}_{target,obs}$ well. The magnitude of the difference tends to be less than 0.025 ($|\text{Intercept}_{imp,prag} - \text{Intercept}_{target,imputed}| < 0.025$) when sample size is 300 and this decreases with increased sample size.

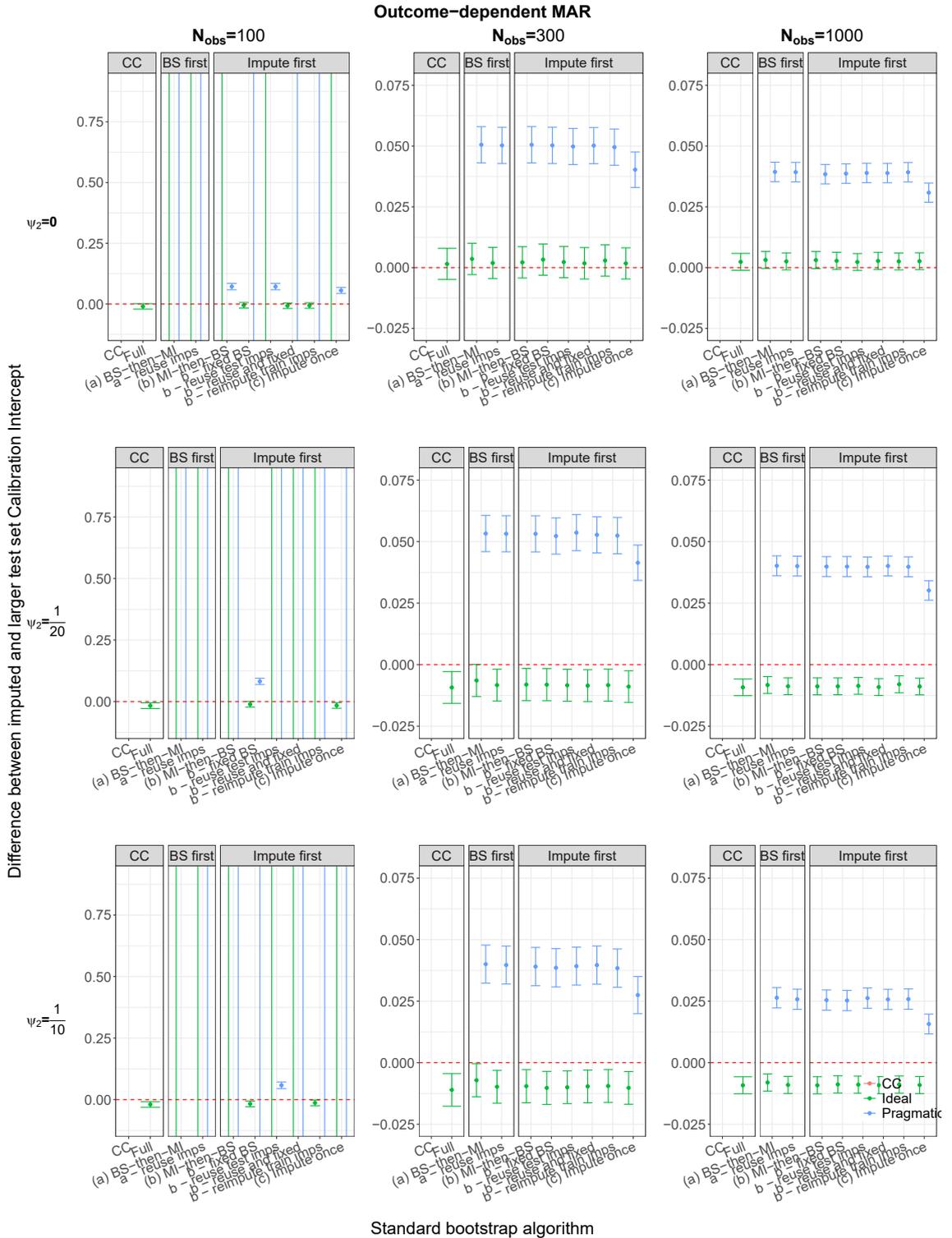


Figure C43: Error bars of the difference in the Calibration intercept estimate from the imputation methods and the target estimate of the calibration intercept, with Monte Carlo 95% confidence intervals, when data are outcome-dependent or outcome- and covariate-dependent MAR. CC (complete-case); methods are described in Section 2.7 or Table 6.1.

C.7.4 Calibration slope

MCAR and covariate-dependent MAR

Figure C44 displays the results for the various missing data methods' estimate of the calibration slope when data are MCAR or covariate-dependent MAR. These results are compared to the estimate of the calibration slope when data are fully-observed ($\text{Slope}_{imp} - \text{Slope}_{obs}$).

The complete-case analysis tends to underestimate the estimate of the calibration slope when data are fully-observed. When the sample size is 100, the magnitude of the difference tends to be between 0.025 and 0.05 ($0.025 \leq |\text{Slope}_{CC} - \text{Slope}_{obs}| < 0.05$). Increasing the sample size to 300 or 1000 decreases the magnitude to less than 0.025 when data are MCAR or covariate-dependent MAR.

The pragmatic performance of the methods tends to underestimate the estimate of the calibration slope when data are fully-observed. When sample size is 100, method *BS-then-MI* tends to underestimate the calibration slope when data are fully-observed the most while method *MI-then-BS impute once* tends to underestimate it the least. The other methods perform similarly in relation to the estimate when data are fully-observed with an average difference of -0.05. With increased sample size method *BS-then-MI* tends to perform similarly to the other methods, except for method *MI-then-BS impute once* which has a magnitude less than 0.01 ($|\text{Slope}_{MI-BS-once} - \text{Slope}_{obs}| < 0.01$).

The ideal performance of the imputation based methods underestimates the calibration slope estimate when data are fully-observed. When sample size is 100 method *BS-then-MI* tends to underestimate the fully-observed calibration slope estimate the most with a magnitude of approximately 0.025. Methods *MI-then-BS* with or without fixed bootstrap samples and *MI-then-BS reimpute* perform similarly in relation to the fully-observed estimate with a magnitude of around 0.02. Methods *BS-then-MI reuse imps*, *MI-then-BS reuse test imps* (with or without fixed bootstrap samples) and *MI-then-BS impute once* perform most similarly to the fully-observed estimate of the calibration slope. With increased sample size the ideal performance of all the imputation methods tends to decrease and perform similarly to the fully-observed estimate when data are MCAR or covariate-dependent MAR.

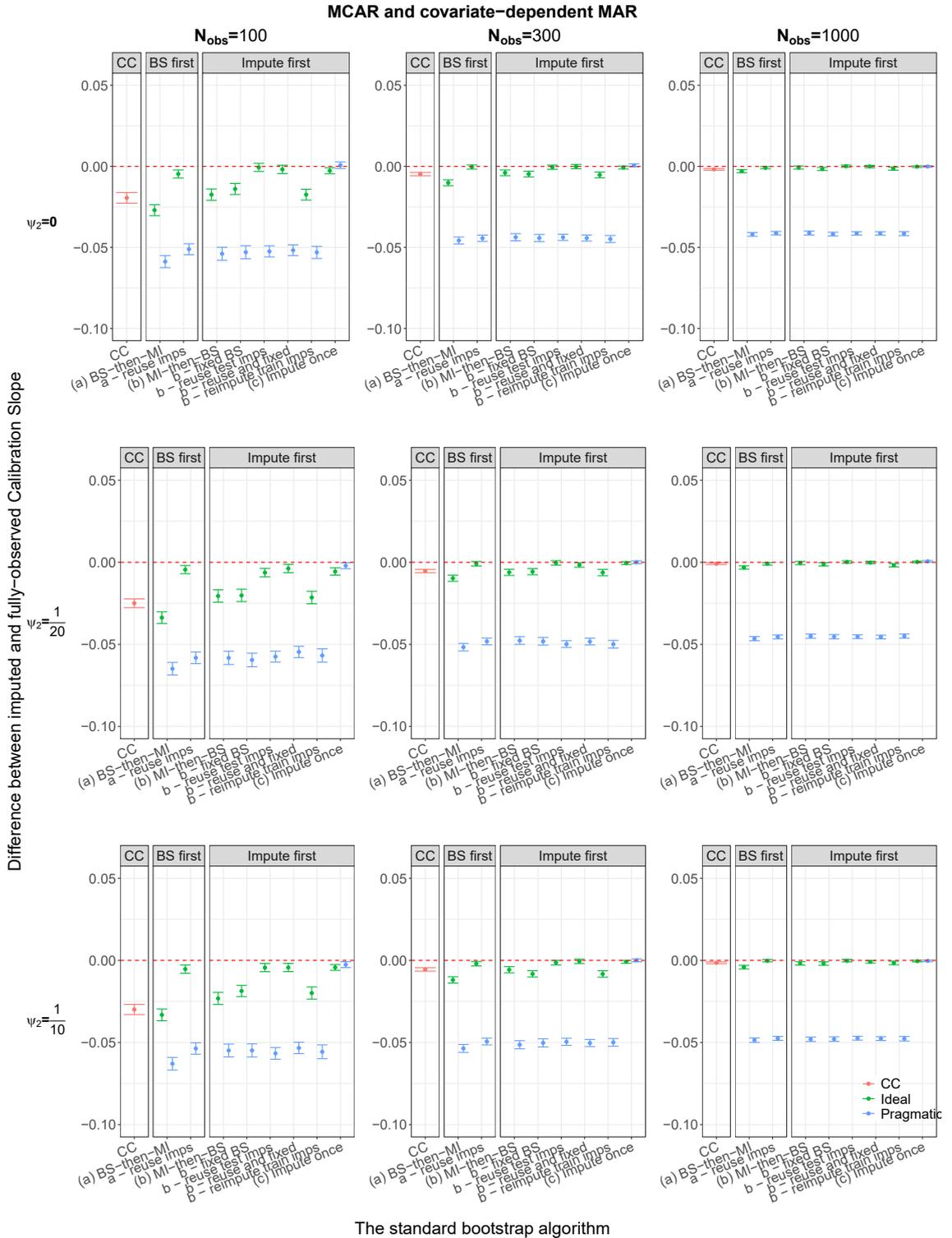


Figure C44: Error bars of the difference in the Calibration intercept estimate from the imputation methods and the calibration intercept estimate when data are fully-observed, with Monte Carlo 95% confidence intervals, when data are MCAR or covariate-dependent MAR. CC (complete-case); methods are described in Section 2.7 or Table 6.1.

Outcome-dependent MAR

Figure C45 displays the results for the various missing data methods' estimate of the calibration slope which is compared to the calibration slope estimated when data are fully-observed ($\text{Slope}_{imp} - \text{Slope}_{obs}$). The Figure displays the results for the outcome-dependent and outcome- and covariate-dependent MAR scenarios.

The complete-case analysis tends to underestimate the estimate of the calibration slope when data are fully-observed. When the sample size is 100, the magnitude of the difference tends to be between 0.025 and 0.05 ($0.025 < |\text{Slope}_{CC} - \text{Slope}_{obs}| < 0.055$). Increasing the sample size to 300 or 1000 decreases the magnitude to less than 0.025.

The pragmatic performance of all imputation methods tends to underestimate the estimate of the calibration slope when data are fully-observed. The magnitude of this difference ($|\text{Slope}_{imp} - \text{Slope}_{obs}|$) tends to be between 0.05 and 0.08. For a sample size of 100 the method *BS-then-MI* tends to have the largest magnitude of underestimation while the other methods tend to perform similarly. With increasing sample size all methods perform similarly when compared to the calibration slope estimate when data are fully-observed. The exception is method *MI-then-BS impute once* which has a magnitude less than 0.01 for all sample sizes when data are outcome-dependent or outcome- and covariate-dependent MAR.

The ideal performance of all imputation methods underestimates the calibration slope estimate when data are fully-observed for outcome-dependent and outcome- and covariate-dependent MAR. For a sample size of 100 method *BS-then-MI* has the largest magnitude when compared to the slope estimate when data are fully-observed ($0.03 < |\text{Slope}_{BS-MI} - \text{Slope}_{obs}| < 0.05$). Methods *MI-then-BS* with or without fixed bootstrap samples and *MI-then-BS reimpute train imps* have the next highest magnitudes of underestimation which tends to approximately be 0.025. Methods *BS-then-MI reuse imps*, *MI-then-BS reuse test imps* (with or without fixed bootstrap samples) and *MI-then-BS impute once* have the smallest magnitude of underestimation. With increasing sample size the magnitude of all methods decreases to be less than 0.025 for a sample size of 300 and less than 0.01 for a sample size of 1000.

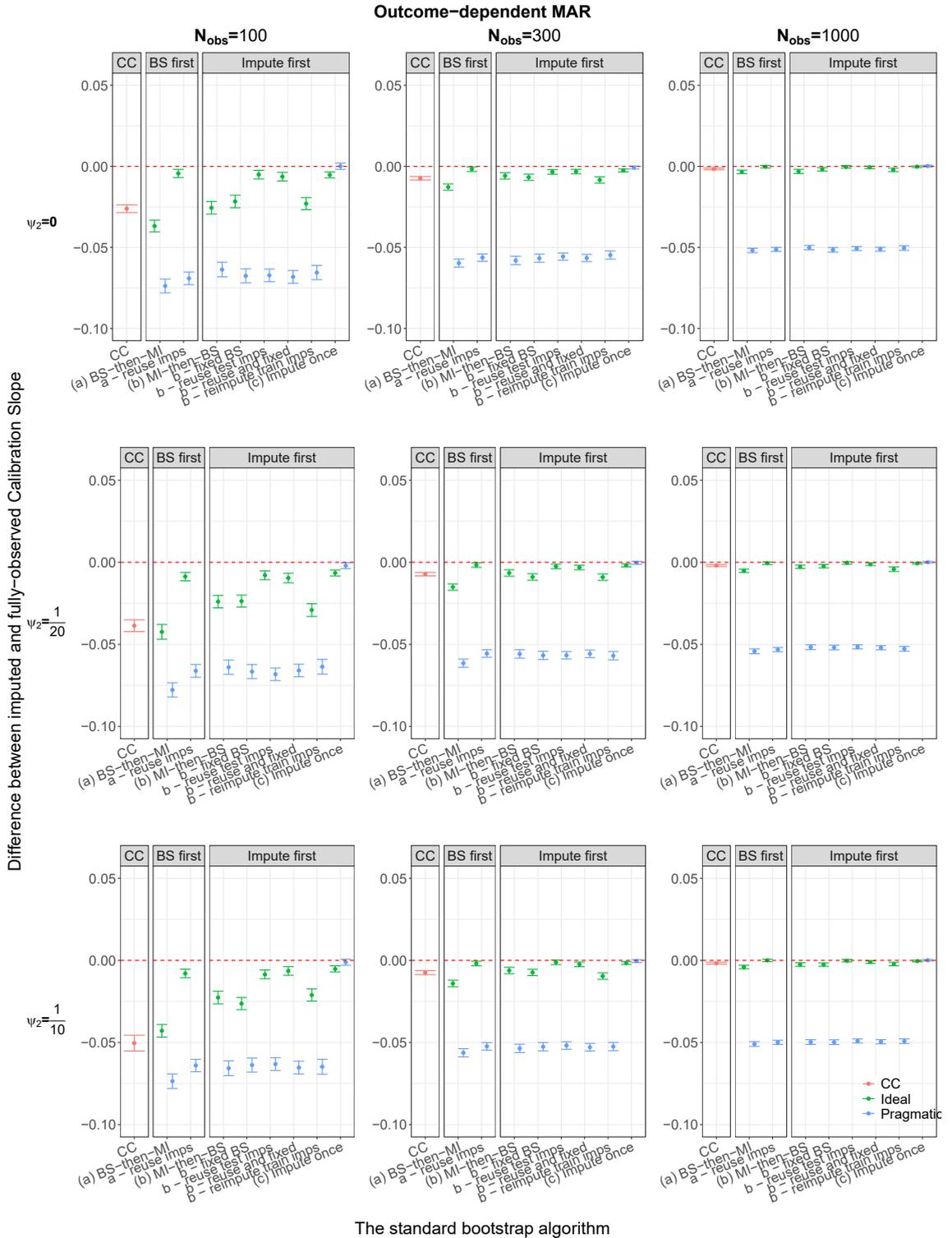


Figure C45: Error bars of the difference in the Calibration intercept estimate from the imputation methods and the calibration intercept estimate when data are fully-observed, with Monte Carlo 95% confidence intervals, when data are outcome-dependent or outcome- and covariate-dependent MAR. CC (complete-case); methods are described in Section 2.7 or Table 6.1.

Increasing the number of imputed datasets from 5 to 25

Figure C46 displays the results for comparing the various imputation based methods when using 5 or 25 imputed datasets. The results in the graph are for the scenario when data are outcome-dependent MAR but are representative of the results when data are MCAR or covariate-dependent MAR (available in Supplementary Plots S4.3.4). Due to increased computation time when using 25 imputed datasets the comparison a reduced set of methods were assessed. Results are available for methods *BS-then-MI*, *MI-then-BS* and *MI-then-BS impute once* which are based on 1000 repetitions.

The estimates for the comparison of the calibration slope for the various methods to Slope_{obs} , perform similarly regardless of whether 5 or 25 imputed datasets are used.

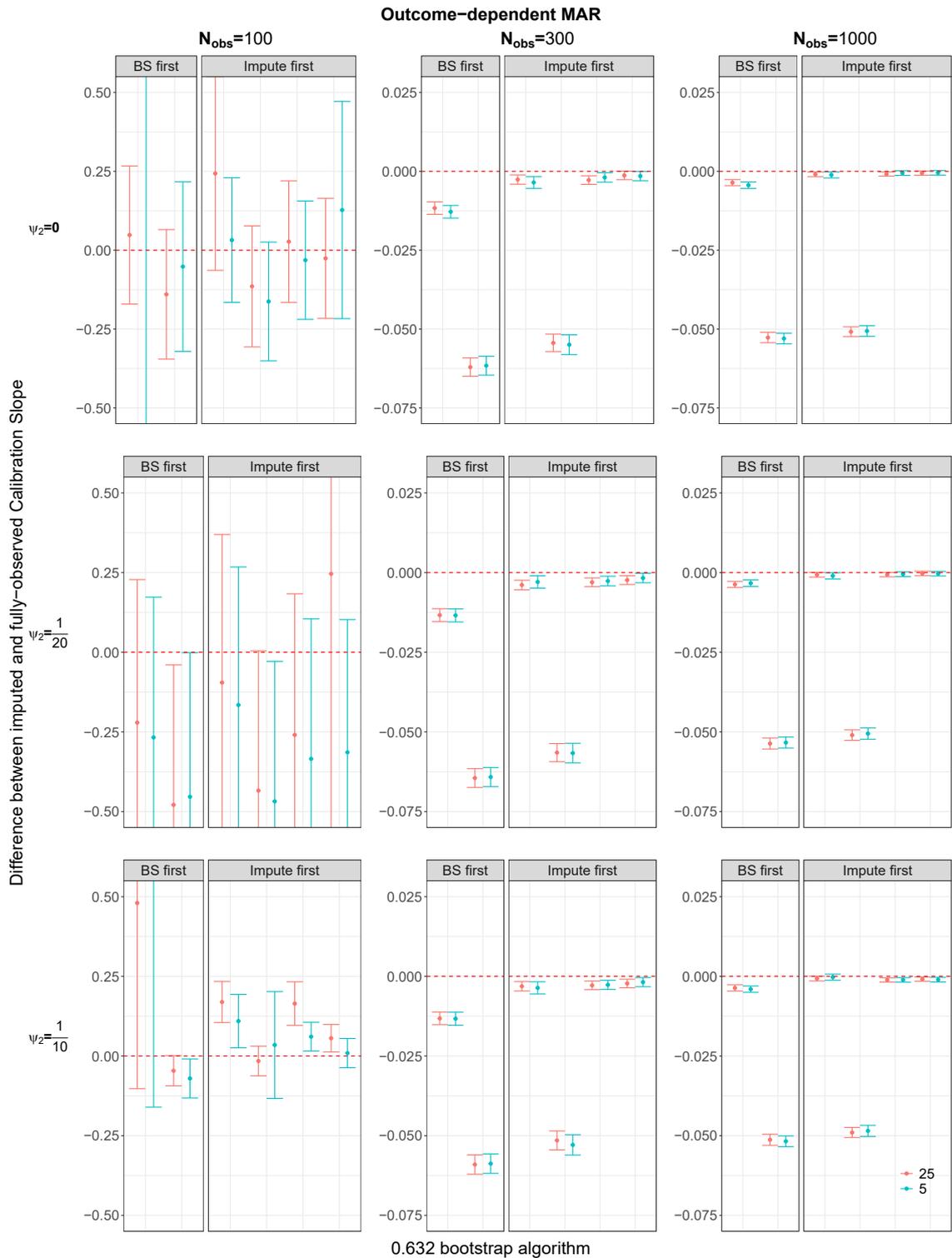


Figure C46: The difference $\text{Slope}_{imp} - \text{Slope}_{obs}$ when data are outcome-dependent or outcome- and covariate-dependent MAR for $M = 25$ when 25% of values are missing in X_1 . The error bars summarise results from the 2000 repetitions and the limits represent the Monte Carlo 95% confidence interval of $\text{Slope}_{imp} - \text{Slope}_{obs}$. CC (complete-case); methods are described in Section 2.7 or Table 6.1.

Increasing the proportion of missingness to 40%

Figure C47 compares the pragmatic and ideal performance when 25% of X_1 values are missing to 40% missingness when data are weak outcome- and covariate-dependent MAR. The results in the figure are generally representative of all missing data scenarios, additional graphs for ideal and pragmatic performance comparing the proportion of missing values increasing from 25% to 40% can be found in supplementary plots Section S4.3.3.

The complete-case analysis when 40% of X_1 values are missing tends to underestimate the fully-observed slope estimate more than when 25% of X_1 values are missing for all missing data scenarios. With increasing sample size the magnitude of the complete-case analysis when 40% of values are missing decreases and tends towards the magnitude when 25% of values are missing ($|\text{Slope}_{CC,40} - \text{Slope}_{obs}| \rightarrow |\text{Slope}_{CC,25} - \text{Slope}_{obs}|$).

For pragmatic performance the estimates of the calibration slope when 40% of X_1 values were set as missing underestimates the calibration slope when 25% of X_1 values are missing ($|\text{Slope}_{imp,40} - \text{Slope}_{obs}| \rightarrow |\text{Slope}_{imp,25} - \text{Slope}_{obs}|$) for all missing data scenarios. This holds true for all imputation methods except *MI-then-BS impute once* which tend to perform similarly in relation to the calibration slope estimate when data are fully-observed.

Similarly for the ideal performance of all imputation methods, the calibration slope estimate when 40% of values are missing tends to underestimate the slope estimate when 25% of values are missing ($\text{Slope}_{imp,40} - \text{Slope}_{obs} < \text{Slope}_{imp,25} - \text{Slope}_{obs}$) for all missing data scenarios. With increasing sample size from 300 to 1000, the magnitude of the underestimation decreases i.e. $|\text{Slope}_{imp,40} - \text{Slope}_{obs}| \rightarrow |\text{Slope}_{imp,25} - \text{Slope}_{obs}|$.

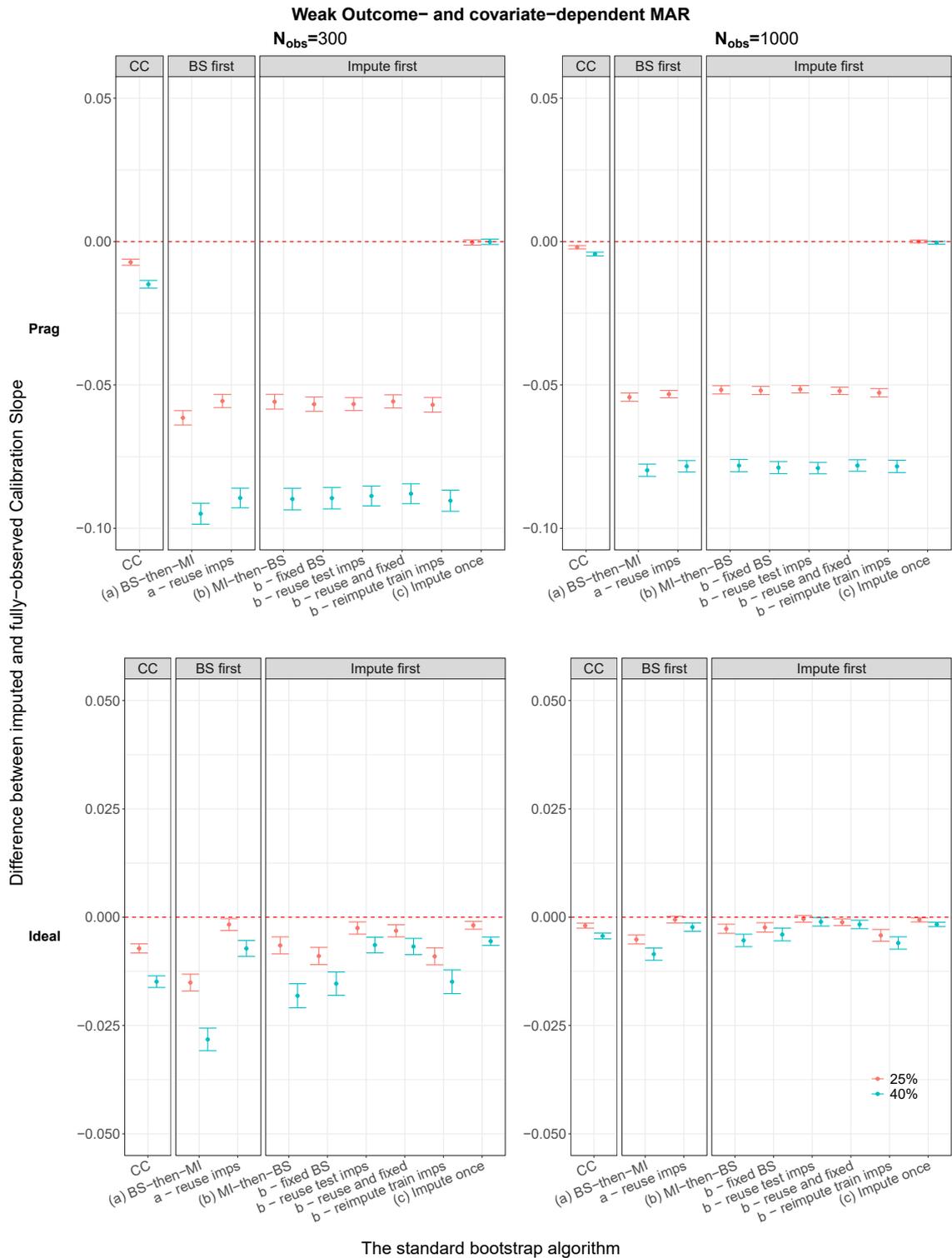


Figure C47: Error bars of the difference in the Calibration slope from the imputation methods and the slope estimate when data are fully-observed, with Monte Carlo 95% confidence intervals, when data are outcome- and covariate-dependent MAR. The graph compares the intercept estimates when 25% of X_1 values are missing versus 40% missing for ideal and pragmatic performance. CC (complete-case); methods are described in Section 2.7 or Table 6.1.

Comparing to the target performance

As previously described for the MSE, AUC, Brier score and calibration intercept, the ideal performance of the bootstrap imputation methods and the calibration slope estimate when data are fully-observed were compared to the ideal performance of the target calibration slope which is estimated from applying a prediction model, based on all data in a repetition, to the fully-observed data in the larger test set ($\text{Slope}_{\text{target,obs}}$). The pragmatic performance of the imputation methods is compared to applying a repetition's prediction model to the imputed datasets of the larger test set ($\text{Slope}_{\text{target,imputed}}$). The complete-case estimate is compared to applying a repetition's prediction model to the observed cases of the larger test set ($\text{Slope}_{\text{target,CC}}$).

MCAR and covariate-dependent MAR

Figure C48 presents the various missing data methods results for the calibration slope when compared to the complete-case, ideal and pragmatic target estimate when data are MCAR or covariate-dependent MAR.

The complete-case analysis estimate tends to overestimate $\text{Slope}_{\text{target,CC}}$ by approximately 0.75 ($\text{Slope}_{\text{CC}} - \text{Slope}_{\text{target,CC}} \geq 0.75$). It does not fit onto the scale of the graph for the weak and strong covariate-dependent MAR in Figure C48 (second and third row of the figure).

When sample size is 100, the pragmatic performance of method *BS-then-MI* tends to approximate $\text{Slope}_{\text{target,imputed}}$ well. Method *MI-then-BS impute once* performs the worst across all imputation methods, overestimating $\text{Slope}_{\text{target,imputed}}$ by at least 0.05. All other imputation methods perform similarly, overestimating $\text{Slope}_{\text{target,imputed}}$ by at most 0.01. With increasing sample size, all methods tend to perform similarly to each other in relation to $\text{Slope}_{\text{target,imputed}}$, except method *MI-then-BS impute once* which continues to overestimate the target estimate by around 0.05 ($\text{Slope}_{\text{MI-BS-once}} - \text{Slope}_{\text{target,CC}} \geq 0.4$).

For ideal performance when sample size is 100, method *BS-then-MI* tends to underestimate $\text{Slope}_{\text{target,obs}}$. Method *BS-then-MI* tends to have the largest magnitude of the difference with $\text{Slope}_{\text{target,obs}}$ across all methods but has overlapping confidence intervals with *MI-then-BS*, *MI-then-BS fixed BS* and *MI-then-BS reimpute*. With increasing sample size, all imputation methods tend to overestimate the ideal target estimate of the calibration slope and perform similarly. When data are weak outcome- and covariate-dependent MAR, methods *BS-then-MI*, *MI-then-BS*, *MI-then-BS fixed BS* and *MI-then-BS reimpute* approximate $\text{Slope}_{\text{target,obs}}$ well while the other methods overestimate $\text{Slope}_{\text{target,obs}}$.

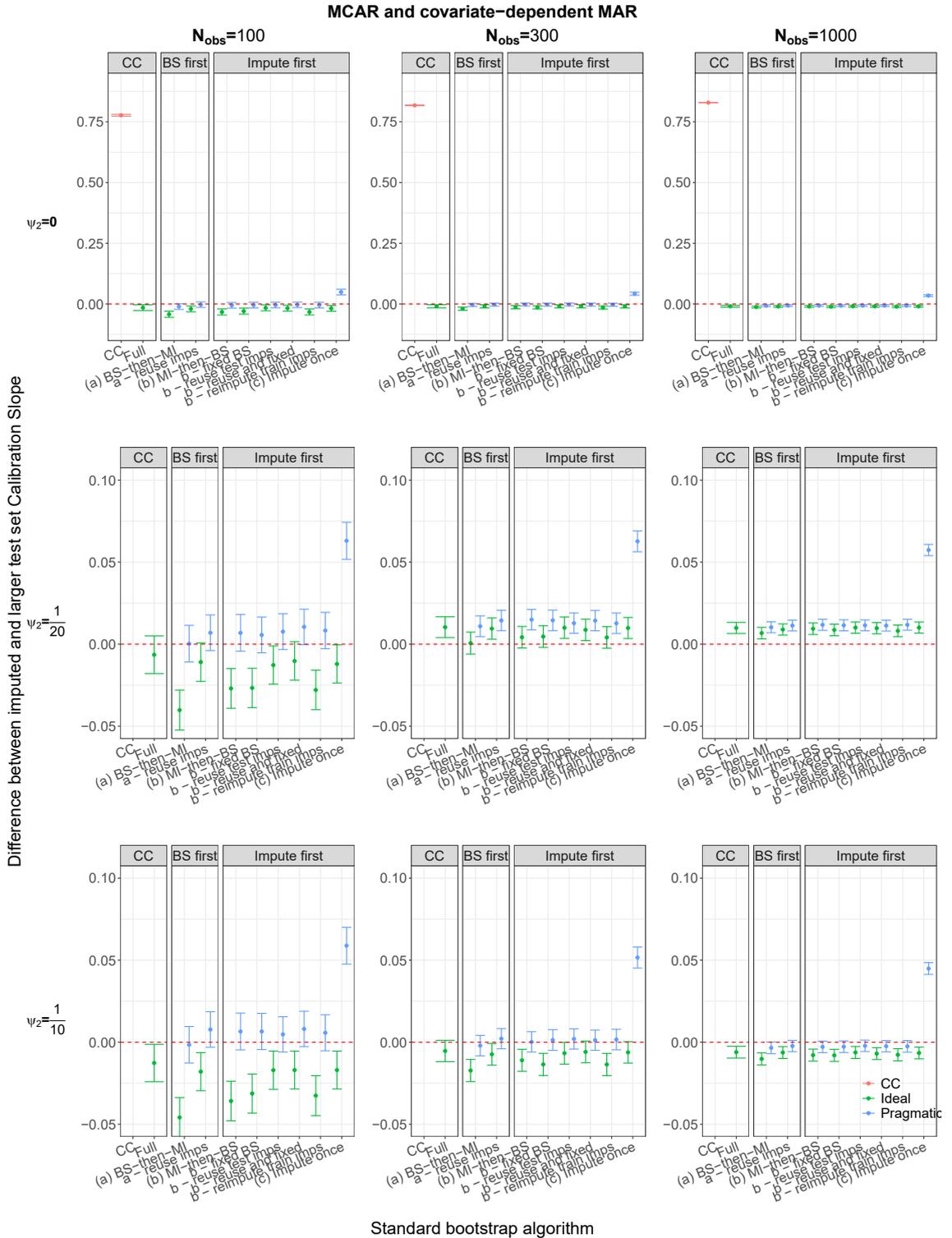


Figure C48: Error bars of the difference in the Calibration intercept estimate from the imputation methods and the target estimate of the calibration intercept, with Monte Carlo 95% confidence intervals, when data are MCAR or covariate-dependent MAR. CC (complete-case); methods are described in Section 2.7 or Table 6.1.

Outcome-dependent MAR

Figure C49 presents the various missing data methods results for the calibration slope when compared to the complete-case, ideal and pragmatic target estimate when data are outcome-dependent or outcome- and covariate-dependent MAR.

Similarly to when data were MCAR or covariate-dependent MAR, the complete-case analysis tends to overestimate $\text{Slope}_{\text{target},CC}$ by over 0.75 ($\text{Slope}_{CC} - \text{Slope}_{\text{target},CC} > 0.75$) for all sample sizes.

The pragmatic performance of method *MI-then-BS impute once* overestimates $\text{Slope}_{\text{target},\text{imputed}}$ and has the largest magnitude of all the methods' pragmatic performance ($\text{Slope}_{MI-BS-once} - \text{Slope}_{\text{target},\text{imputed}} > 0.4$). With increasing sample size the magnitude of its overestimation tends to decrease but it still has the largest magnitude across all missing data scenarios. When data are weakly outcome- and covariate-dependent MAR for all sample sizes, all methods tend to overestimate $\text{Slope}_{\text{target},\text{imputed}}$ ($\text{Slope}_{\text{imp}} - \text{Intercept}_{\text{target},\text{imputed}} > 0$). Method *BS-then-MI* has the lowest magnitude across all methods, method *MI-then-BS impute once* has the largest and all other methods perform similarly. Increasing the sample size to 1000 all methods tend to perform similarly with a magnitude of approximately 0.01, except for method *MI-then-BS impute once*. When data are weakly outcome- and strongly covariate-dependent MAR, the pragmatic performance of method *BS-then-MI* tends to underestimate $\text{Slope}_{\text{target},\text{imputed}}$ more than the other imputation methods (*BS-then-MI reuse imps* and the *MI-then-BS* variations, excluding *MI-then-BS impute once*). Although with increasing sample size to 300 or 1000 it tends to perform similarly.

Similarly to the pragmatic performance, the ideal performance of method *BS-then-MI* tends to underestimate $\text{Slope}_{\text{target},\text{obs}}$ and has the largest magnitude ($|\text{Slope}_{BS-MI} - \text{Slope}_{\text{target},\text{obs}}|$) when sample size is 100 or 300. When data are weakly outcome- and covariate-dependent MAR and sample size is 300 or 1000, the ideal performance of method *BS-then-MI*, either approximates $\text{Slope}_{\text{target},\text{obs}}$ well or has the smallest magnitude of the difference with it. With increasing sample size all methods tend to perform similarly to each other in relation to $\text{Slope}_{\text{target},\text{obs}}$.

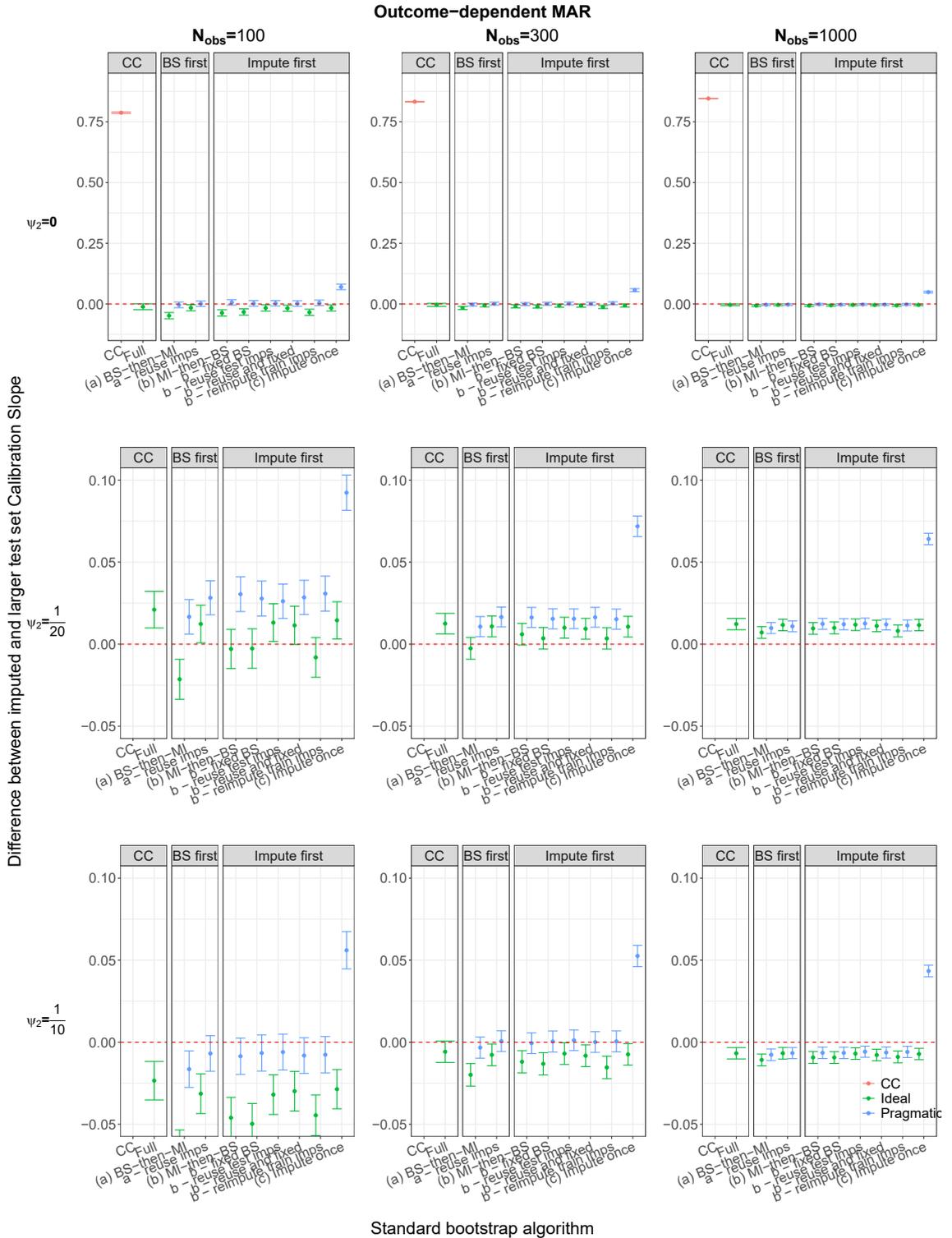


Figure C49: Error bars of the difference in the Calibration intercept estimate from the imputation methods and the target estimate of the calibration intercept, with Monte Carlo 95% confidence intervals, when data are outcome-dependent or outcome- and covariate-dependent MAR. CC (complete-case); methods are described in Section 2.7 or Table 6.1.

C.8 Is data leakage an issue for the *standard* and 0.632 bootstrap algorithms?

The aim of the binary outcome simulation study was to identify the most appropriate way to combine the 0.632 or the *standard* bootstrap validation algorithm with multiple imputation. Data leakage for the 0.632 and *standard* bootstrap validation algorithm was previously discussed for a continuous outcome in Section 6.6 and was initially introduced in Section 4.6.

Generally, the impact of data leakage was more noticeable for small sample sizes, as previously noted in the data leakage discussion for cross-validation (Sections 4.6 and 5.7) and the continuous outcome scenario for the bootstrap algorithms (Section 6.6). This was seen for the AUC, Brier score and calibration intercept and slope as method *BS-then-MI* (the method which is not subject to data leakage through the imputation process) would have the largest magnitude of over- or underestimation compared to the other methods in relation to Perf_{obs} .

When comparing a methods' performance to the fully-observed performance measure, method *BS-then-MI reuse* imputed datasets tended to have a smaller magnitude than method *BS-then-MI* $|\text{Perf}_{BS-MI-reuse} - \text{Perf}_{obs}| < |\text{Perf}_{BS-MI} - \text{Perf}_{obs}|$. Method *BS-then-MI reuse imps* is subject to data leakage as observations from the imputed datasets are reused to form an imputed training and testing dataset for the bootstrap samples. For both ideal and pragmatic performance, this leakage has caused method *BS-then-MI reuse imps* to perform better than method *BS-then-MI* when compared to the fully-observed estimate of the performance measure.

The impact of reusing imputed datasets which were imputed using information from the entire dataset can also be seen when comparing the various *MI-then-BS* methods. Similarly to the MSE performance measure scenarios, reusing imputed datasets (methods *MI-then-BS reuse test imps* with or without fixed bootstrap samples) lead to over-optimistic estimates of ideal performance for the AUC and Brier score when compared to method *MI-then-BS*. As previously stated in Section 6.6, using all covariate and outcome data to impute missing values in the bootstrap sample can lead to an estimate of performance which is better than if the data had been fully-observed. This over-optimism of the fully-observed estimate appears to come from the increased leakage due to knowledge of the outcome. The over-optimism can be seen when comparing the ideal performance of method *MI-then-BS reuse test imps* with *MI-then-BS* for the AUC or Brier score performance measures but not for the pragmatic performance (Figures C1, C2, C7, C8).

C.9 Comparing internal validation algorithms

As discussed in Section 6.7, we can use the target estimate of the AUC, Brier score and calibration intercept and slope to compare the various internal validation algorithms.

Figure C50 presents results for the various performance measures assessed for the binary outcome for the cross-validation, 0.632 and *standard* bootstrap algorithms when data are weak outcome- and covariate-dependent MAR. The results presented in the graph are generally representative of the results for each performance measure across all scenarios (all graphs are available in the Supplementary Plots Section S4.7).

Previously for the MSE in Section 6.7 the cross-validation results were more variable than the bootstrap 0.632 and *standard* bootstrap algorithms. Here, for the AUC both bootstrap methods perform similarly while the cross-validation methods tend to be more variable when the sample size is small. With increasing sample size, all internal validation methods tend to perform similarly.

For the Brier score, the cross-validation methods tend to overestimate the target estimate while the bootstrap methods tend to underestimate the target Brier score when the sample size is small. For a large sample size, the internal validation methods perform similarly.

Due to the instability of the calibration intercept when the sample size is small it is impossible to compare the various internal validation methods. For larger sample sizes, the methods tend to perform similarly with cross-validation for the pragmatic performance slightly outperforming the pragmatic performance of both the *standard* and 0.632 bootstrap algorithms.

For the calibration slope, the cross-validation methods are outperformed by the bootstrap methods for all sample sizes and missing data scenarios. For small sample sizes the *standard* bootstrap method performs better than the 0.632 variation and by sample size of 1000 the two variations perform similarly for the various methods.

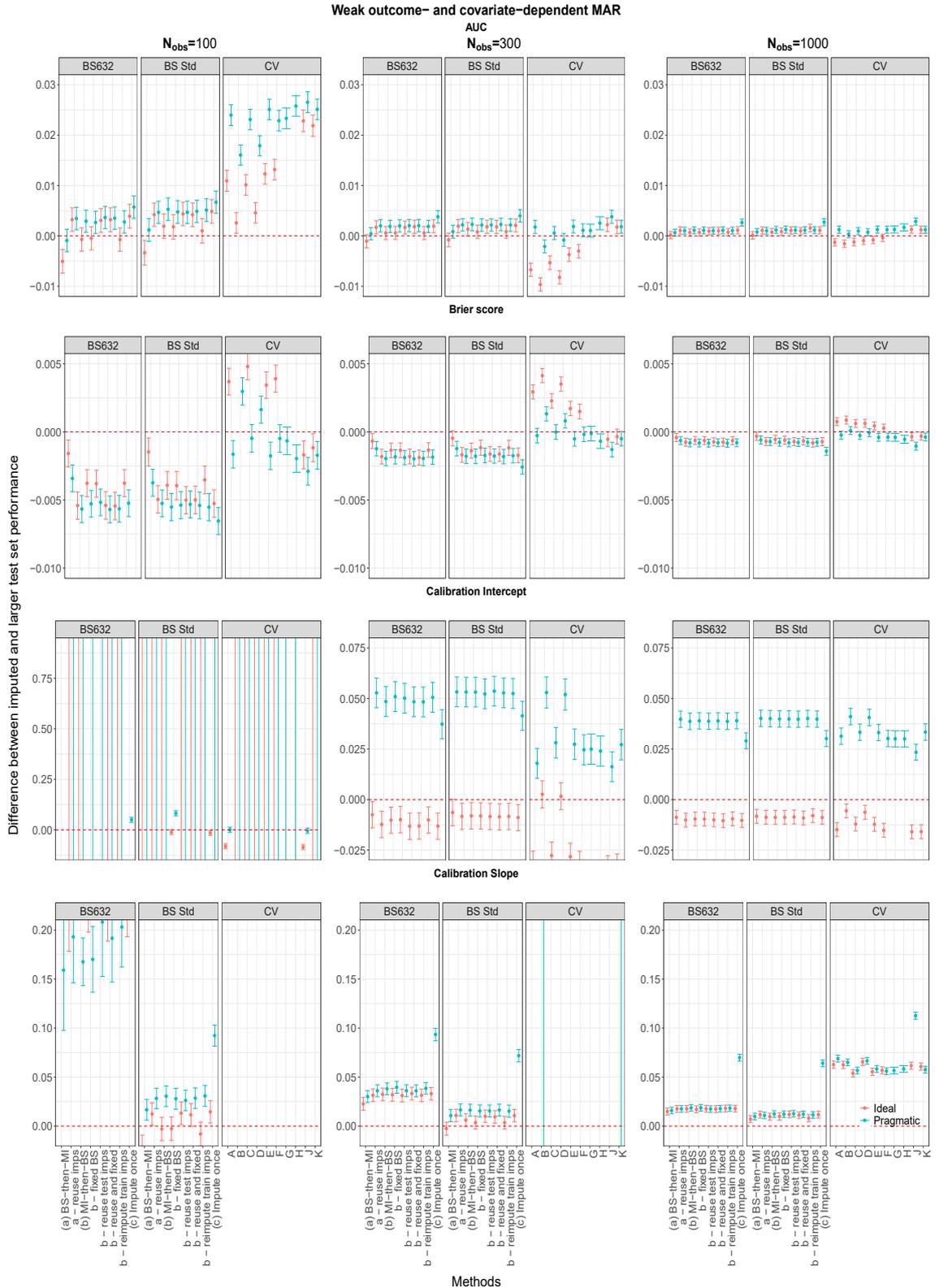


Figure C50: Comparing cross-validation and the 0.632 and *standard* bootstrap algorithms using the target performance for the AUC, Brier score and calibration intercept and slope. Error bars of the difference in the imputed performance estimate and the estimate from a larger test set are presented for the weak outcome- and covariate-dependent MAR scenario. CC (complete-case); CV methods A-K are described in Table 2.3; bootstrap methods are described in Section 2.7 or Table 6.1.

D Chapter11: Systematic review - Supplementary file 1

D.1 Search Terms

Table C1: Search Terms used in Medline and Embase databases

| # | Terms |
|---|--|
| 1 | exp neoplasm/ |
| 2 | exp RADIATION ONCOLOGY/ or exp MEDICAL ONCOLOGY/ or exp PSYCHO-ONCOLOGY/ or exp SURGICAL ONCOLOGY/ |
| 3 | tumo?r.mp. |
| 4 | cancer.mp. |
| 5 | exp pancreas cancer/ or exp female genital tract cancer/ or exp central nervous system cancer/ or exp nasopharynx cancer/ or exp hypopharynx cancer/ or exp cancer antibody/ or exp ovarian cancer cell line/ or exp breast cancer resistance protein/ or exp cancer palliative therapy/ or exp cancer size/ or exp endometrial cancer cell line/ or exp cancer diagnosis/ or exp "HCC cell line (colorectal cancer)" / or exp cancer staging/ or exp cancer cell/ or exp lung cancer/ or exp experimental pancreatic cancer/ or exp vulva cancer/ or exp uterine cervix cancer/ or exp cancer graft/ or exp cancer immunology/ or exp brain cancer cell line/ or exp colorectal cancer cell line/ or exp germ cell cancer/ or exp larynx cancer/ or exp cancer therapy/ or exp cancer genetics/ or exp skin cancer/ or exp mouth cancer/ or exp cancer statistics/ or exp breast cancer-related lymphedema/ or exp disseminated cancer/ or exp second cancer/ or exp cancer surgery/ or exp childhood cancer/ or exp human epidermal growth factor receptor 2 positive breast cancer/ or exp colon cancer cell line/ or exp esophageal cancer cell line/ or exp cancer survivor/ or exp cancer transplantation/ or exp small cell lung cancer/ or exp pelvis cancer/ or exp cancer adjuvant therapy/ or exp advanced cancer/ or exp lung cancer cell line/ or exp early cancer/ or exp urogenital tract cancer/ or exp rectum cancer/ or exp breast cancer molecular subtype/ or exp metastatic colon cancer/ or exp castration resistant prostate cancer/ or exp triple negative breast cancer/ or exp cancer recurrence/ or exp cancer chemotherapy/ or exp small intestine cancer/ or exp oropharynx cancer/ or exp Institute for Cancer Research mouse/ or exp poorly differentiated thyroid cancer/ or exp non muscle invasive bladder cancer/ or exp differentiated thyroid cancer/ or exp abdominal cancer/ or exp cancer center/ or exp "head and neck cancer" / or exp ovary cancer/ or exp cancer prognosis/ or exp blood cancer cell line/ or exp "hereditary breast and ovarian cancer syndrome" / or exp cancer incidence/ or exp cancer pain/ or exp muscle invasive bladder cancer/ or exp biliary tract cancer/ or exp cancer growth factor/ or exp cancer susceptibility/ or exp cancer cell culture/ or exp bladder cancer cell line/ or exp digestive system cancer/ |

Table C1: Search Terms (continued)

| # | Terms |
|----|---|
| 5 | or exp cancer control/ or exp nervous system cancer/ or exp hereditary colorectal cancer/ or exp cervical cancer cell line/ or exp childhood cancer survivor/ or exp cancer survival/ or exp liver cancer cell line/ or exp multimodality cancer therapy/ or exp gastric cancer cell line/ or exp inflammatory breast cancer/ or exp brain cancer/ or exp metastatic colorectal cancer/ or cancer*.mp. or exp testis cancer/ or exp gallbladder cancer/ or exp liver cancer/ or exp cancer radiotherapy/ or exp esophagus cancer/ or exp colorectal cancer/ or exp prostate cancer/ or exp uterus cancer/ or exp occult cancer/ or exp cancer registry/ or exp cancer localization/ or exp cancer testis antigen/ or exp heart cancer/ or exp cancer mortality/ or exp breast cancer/ or exp cancer resistance/ or exp cancer specific survival/ or exp cancer tissue/ or exp cancer patient/ or exp cancer associated fibroblast/ or exp cancer screening/ or exp cecum cancer/ or exp cancer vaccine/ or exp adrenal cancer/ or exp respiratory tract cancer/ or exp multiple cancer/ or exp estrogen receptor positive breast cancer/ or exp tongue cancer/ or exp stomach cancer/ or exp cancer grading/ or exp hereditary nonpolyposis colorectal cancer/ or exp pancreatic cancer cell line/ or exp cancer classification/ or exp non melanoma skin cancer/ or exp cancer fatigue/ or exp prostate cancer cell line/ or exp early cancer diagnosis/ or exp colon cancer/ or exp cancer model/ or exp cancer stem cell/ or exp vagina cancer/ or exp penis cancer/ or exp cancer risk/ or exp non small cell lung cancer/ or exp thyroid cancer/ or exp cancer cell line/ or exp "cancer of unknown primary site"/ or exp cancer prevention/ or exp breast cancer cell line/ or exp kidney cancer/ or exp bladder cancer/ or exp endometrium cancer/ or exp anus cancer/ or exp cancer immunotherapy/ or exp "HCC cell line (cervical cancer)"/ or exp bone marrow cancer/ or exp cancer research/ or exp metastatic breast cancer/ or exp urinary tract cancer/ or exp cancer inhibition/ or exp basal like breast cancer/ or exp cancer immunization/ or exp progesterone receptor positive breast cancer/ |
| 6 | 1 or 2 or 3 or 4 or 5 |
| 7 | missing data.mp. |
| 8 | drop out.mp. |
| 9 | non-response.mp. |
| 10 | incomplete data.mp. |
| 11 | exclude* data.mp. |
| 12 | 7 or 8 or 9 or 10 or 11 |

Table C1: Search Terms (continued)

| # | Terms |
|----|---|
| 13 | 6 and 12 |
| 14 | exp survival analysis/ |
| 15 | hazard ratio.mp. |
| 16 | time-to-event analysis.mp. |
| 17 | cox model.mp. or Proportional Hazards Models/ |
| 18 | time-dependent.mp. |
| 19 | time varying.mp. |
| 20 | relative ratio.mp. |
| 21 | 14 or 15 or 16 or 17 or 18 or 19 or 20 |
| 22 | 13 and 21 |
| 23 | 22 and 2012:2018.(sa_year). |

D.2 Data extraction checklist

Table C2: Checklist for data extraction

| Heading | Checklist | More detail |
|--------------|--|--|
| Journal | | |
| Year | | |
| First Author | | |
| Analysis | <p>Models used</p> <p>Complexity of analysis model</p> <p>Functional form</p> <p>PH assumption</p> | <ul style="list-style-type: none"> • Kaplan-Meier • Log rank test • Cox model • Exponential • Weibull • Univariable • Multivariable • Checked? • If yes, Martingale residuals? • Other method? • Checked? • Schoenfeld Residuals • KM or log-log plots • Interaction with time • If other, specify • If checked, was there an attempt to handle missing data other than a CC analysis? |

PH= proportional hazards

Table C2: Checklist for data extraction (continued)

| Heading | Checklist | More detail |
|--------------|---|---|
| Analysis | Covariate selection | <ul style="list-style-type: none"> • Explicitly stated? • A priori involved? • Univariable model with $p < \alpha$ • Chi-square • T-test • Fishers exact test • Likelihood ratio test • Forward selection • Backward selection • Other, specify |
| Missing data | <p>Outside of CC analysis, was extent of missing specified</p> <p>Assumptions stated?</p> <p>Methods to handle missing data</p> | <ul style="list-style-type: none"> • in text • in table • shown in a plot • If yes, which one? • If no, presumably which one? • Were methods declared in full text or supplementary material? • Initial size of sample (before excluding due to missing data or other criteria) • After applying exclusion criteria not relating to missing data, was it unclear whether there were any missing covariate data? |

Table C2: Checklist for data extraction (continued)

| Heading | Checklist | More detail |
|--------------|--------------------------------|---|
| Missing data | Methods to handle missing data | <ul style="list-style-type: none"> • Out of the initial sample size, how many subjects were excluded in an initial phase (prior to any descriptive statistics or analysis) because they had missing data on one or more covariates? • Out of the initial sample size, how many subjects were excluded in an initial phase (prior to any descriptive statistics or analysis) due to other exclusion criteria? • Did the reporting make it possible to ascertain the numbers excluded due to missing data or due to other exclusion criteria? • If the reporting did not make it possible to ascertain the numbers excluded due to missing data or due to other exclusion criteria, what was the total number excluded in the initial phase for any reason? • Of which, was it possible to determine at least some of the exclusion was due to missing data? • If yes, how many were confirmed to be excluded due to missing data? • Final sample size after all exclusion criteria applied to be used for analysis? |

Table C2: Checklist for data extraction (continued)

| Heading | Checklist | More detail |
|--------------|--------------------------------|--|
| Missing data | Methods to handle missing data | <ul style="list-style-type: none"> • If excluded missing data, what would sample size for analysis have been if missing data had been kept? • After exclusions in an initial phase due to either missing data and/or other criteria, are there missing values in any additional study covariates? • If missing data still present in analysis sample, were covariates containing NAs used in model? • Did initial phase include removing those with incomplete data in some covariates? • Did initial phase include a CC analysis? • Was complete-case analysis used during or post initial phase? • Stated no. of people with complete records? • If no, could no. of complete records be worked out from information in paper? • Removed covariates from analysis due to large amount of missing data? • If removed, are they thought to be highly predictive? |

Table C2: Checklist for data extraction (continued)

| Heading | Checklist | More detail |
|---------------|--------------------------------|---|
| Missing data | Methods to handle missing data | <ul style="list-style-type: none"> • Included a missing indicator in model? • Minimum value imputation • Maximum value imputation • Mean value imputation • Mode value imputation • LOCF • Was MI used? • If yes, uni or multi? • Specify if Joint MVN, FCS etc. • If yes, included time, log time, event indicator or Nelson-Aalen estimate? • If yes, specify no. of imputations? • Was a sensitivity analysis used? • Were there any Time-dependent covariates or time-varying effects with missing data? |
| Software used | | State whether SAS, SPSS, Stata, R, S-plus or Mplus |

D.3 Papers included in the review

1. Laura Bredow, Lisa Stutzel, Daniel Bohringer, Enken Gundlach, Thomas Reinhard, and Claudia Auw-Haedrich. Progesterone and estrogen receptors in conjunctival melanoma and nevi. *Graefe's archive for clinical and experimental ophthalmology = Albrecht von Graefes Archiv fur klinische und experimentelle Ophthalmologie*, 252(2):359–365, 2014.
2. Benjamin Kasenda, Annatina Bass, Dieter Koeberle, Bernhard Pestalozzi, Markus Borner, Richard Herrmann, Lorenz Jost, Andreas Lohri, and Viviane Hess. Survival in overweight patients with advanced pancreatic carcinoma: a multicentre cohort study. *BMC cancer*, 14:728, 2014.
3. Rajesh Sehgal, Mohamed Alsharedi, Chris Larck, Phyllis Edwards, and Todd Gress. Pancreatic cancer survival in elderly patients treated with chemotherapy. *Pancreas*, 43(2):306–310, 2014.
4. Bradshaw P.T., Ibrahim J.G., Khankari N., Cleveland R.J., Abrahamson P.E., Stevens J., Satia J.A., Teitelbaum S.L., Neugut A.I., and Gammon M D. Post-diagnosis physical activity and survival after breast cancer diagnosis: The Long Island Breast Cancer Study. *Breast Cancer Research and Treatment*, 145(3):735–742, 2014.
5. Daniel M Halperin, Chan Shen, Arvind Dasari, Ying Xu, Yiyi Chu, Shouhao Zhou, Ya-Chen Tina Shih, and James C Yao. Frequency of carcinoid syndrome at neuroendocrine tumour diagnosis: a population-based study. *The Lancet. Oncology*, 18(4):525–534, apr 2017.
6. Milan Risteski, Simonida Crvenkova, Zoran Atanasov, and Rozalinda Isjanovska. Epidemiological analysis of progression-free survival (PFS) and overall survival (OS) in non-small-cell lung cancer patients in Republic of Macedonia. *Prilozi (Makedonska akademija na naukite i umetnostite. Oddelenie za medicinski nauki)*, 34(3):49–61, 2013.
7. Annukka Pasanen, Taru Tuomi, Jorma Isola, Synnove Staff, Ralf Butzow, and Mikko Loukovaara. L1 Cell Adhesion Molecule as a Predictor of Disease-Specific Survival and Patterns of Relapse in Endometrial Cancer. *International journal of gynecological cancer : official journal of the International Gynecological Cancer Society*, 26(8):1465–1471, 2016.
8. Andrew J Kaufman, Justin Palatt, Mark Sivak, Peter Raimondi, Dong-Seok Lee, Andrea Wolf, Fouad Lajam, Faiz Bhora, and Raja M Flores. Thymectomy for Myasthenia Gravis: Complete Stable Remission and Associated Prognostic Factors

in Over 1000 Cases. *Seminars in thoracic and cardiovascular surgery*, 28(2):561–568, 2016.

9. Patrick T Bradshaw, Joseph G Ibrahim, June Stevens, Rebecca Cleveland, Page E Abrahamson, Jessie A Satia, Susan L Teitelbaum, Alfred I Neugut, and Marilie D Gammon. Postdiagnosis change in bodyweight and survival after breast cancer diagnosis. *Epidemiology*, 23(2):320–327, 2012.
10. Makatsoris T., Tsamandas A.C., Strimpakos A., Alexopoulou Z., Dionysopoulos D., Pervana S., Konstantara A., Papakostas P., Samantas E., Rallis G., Dimou A., Pentheroudakis G., Papaparaskeva K., Psyrris A., Kalogeras K.T., Syrigos K., and Scopa C.D. HER family protein expression in a Greek population with gastric cancer. A retrospective hellenic cooperative oncology group study. *Anticancer Research*, 36(4):1581–1590, 2016.
11. Keto C.J., Aronson W.J., Terris M.K., Presti J.C., Kane C.J., and Amling C.L. Obesity is associated with castration-resistant disease and metastasis in men treated with androgen deprivation therapy after radical prostatectomy: Results from the SEARCH database. *BJU International*, 110(4):492–498, 2012.
12. Lubna Alhalabi, Matthew J Singleton, Abdullahi O Oseni, Amit J Shah, Zhu-Ming Zhang, and Elsayed Z Soliman. Relation of Higher Resting Heart Rate to Risk of Cardiovascular Versus Noncardiovascular Death. *The American journal of cardiology*, 119(7):1003–1007, apr 2017.
13. R Elaidi, A Harbaoui, B Beuselinck, J-C Eymard, A Bamias, E De Guillebon, C Porta, Y Vano, C Linassier, P R Debruyne, M Gross-Goupil, A Ravaud, M Aitelhaj, G Marret, and S Oudard. Outcomes from second-line therapy in long-term responders to first-line tyrosine kinase inhibitor in clear-cell metastatic renal cell carcinoma. *Annals of oncology : official journal of the European Society for Medical Oncology*, 26(2):378–385, 2015.
14. Amini A., Rusthoven C.G., Jones B.L., Armstrong H., and Raben D. Survival outcomes of radiotherapy with or without androgen-deprivation therapy for patients with intermediate-risk prostate cancer using the National Cancer Data Base. *Urologic Oncology: Seminars and Original Investigations*, 34(4):165, 2016.
15. Joseph C. Y. Chan, Connie I. Diakos, David L. H. Chan, Alexander Engel, Nick Pavlakakis, Anthony Gill, and Stephen J. Clarke. A Longitudinal Investigation of Inflammatory Markers in Colorectal Cancer Patients Perioperatively Demonstrates Benefit in Serial Remeasurement. *Annals of Surgery*, page 1, apr 2017.
16. Michael E Egger, Brittany L Tabler, Erik M Dunki-Jacobs, Glenda G Callender, Charles R Scoggins, Robert C G 2nd Martin, Amy R Quillo, Arnold J Stromberg,

and Kelly M McMasters. Clinicopathologic and survival differences between upper and lower extremity melanomas. *The American surgeon*, 78(7):779–787, 2012.

17. Alper Biler, Ulas Solmaz, Selcuk Erkilinc, Mehmet Gokcu, Mustafa Bagci, Orhan Temel, Tugba Karadeniz, and Muzaffer Sanci. Analysis of endometrial carcinoma in young women at a high-volume cancer center. *International Journal of Surgery*, 44:185–190, aug 2017.
18. Marko Lukic, Ildir Licaj, Eiliv Lund, Guri Skeie, Elisabete Weiderpass, and Tonje Braaten. Coffee consumption and the risk of cancer in the Norwegian Women and Cancer (NOWAC) Study. *European Journal of Epidemiology*, 31(9):905–916, sep 2016.
19. Andrea Rocca, Alberto Farolfi, Roberta Maltoni, Elisa Carretta, Elisabetta Melegari, Cristiano Ferrario, Lorenzo Ceconetto, Samanta Sarti, Alessio Schirone, Anna Fedeli, Daniele Andreis, Elisabetta Pietri, Toni Ibrahim, Erika Montalto, and Dino Amadori. Efficacy of endocrine therapy in relation to progesterone receptor and Ki67 expression in advanced breast cancer. *Breast cancer research and treatment*, 152(1):57–65, 2015.
20. Eggemann H., Ignatov T., Burger E., Kantelhardt E.J., Fettke F., Thomssen C., and Dan Costa S. Moderate HER2 expression as a prognostic factor in hormone receptor positive breast cancer. *Endocrine-Related Cancer*, 22(5):725–733, 2015.
21. Yong Jin Kang, Won Sik Jang, Jong Kyou Kwon, Cheol Yong Yoon, Joo Yong Lee, Won Sik Ham, and Young Deuk Choi. Intermediate PSA half-life after neoadjuvant hormone therapy predicts reduced risk of castration-resistant prostate cancer development after radical prostatectomy. *BMC Cancer*, 17(1):789, dec 2017.
22. Rafael Bitzur, Ronen Brenner, Elad Maor, Maayan Antebi, Tomer Ziv-Baran, Shlomo Segev, Yechezkel Sidi, and Shayee Kivity. Metabolic syndrome, obesity, and the risk of cancer development. *European journal of internal medicine*, 34:89–93, 2016.
23. Oladeru O.T., Miccio J.A., Yang J., Xue Y., and Ryu S. Conformal external beam radiation or selective internal radiation therapy—a comparison of treatment outcomes for hepatocellular carcinoma. *Journal of Gastrointestinal Oncology*, 7(3):433–440, 2016.
24. Sarah Krull Abe, Manami Inoue, Norie Sawada, Junko Ishihara, Motoki Iwasaki, Taiki Yamaji, Taichi Shimazu, Shizuka Sasazuki, and Shoichiro Tsugane. Glycemic index and glycemic load and risk of colorectal cancer: a population-based cohort study (JPHC Study). *Cancer causes & control : CCC*, 27(4):583–593, 2016.
25. Alfonso Rivera Duarte, Alejandra Armengol Alonso, Elena Sandoval Cartagena, and Elena Tuna Aguilar. Blastic Transformation in Mexican Population With Chronic

- Myelomonocytic Leukemia. *Clinical Lymphoma Myeloma and Leukemia*, 17(8):532–538, aug 2017.
26. Sun G.E.C., Wells B.J., Yip K., Zimmerman R., Raghavan D., and Kattan M.W. Gender-specific effects of oral hypoglycaemic agents on cancer risk in type 2 diabetes mellitus. *Diabetes, Obesity and Metabolism*, 16(3):276–283, 2014.
27. Lindsay A Renfro, Axel Grothey, Yuan Xue, Leonard B Saltz, Thierry Andre, Chris Twelves, Roberto Labianca, Carmen J Allegra, Steven R Alberts, Charles L Loprinzi, Greg Yothers, Daniel J Sargent, and Adjuvant Colon Cancer Endpoints (ACCENT) Group. ACCENT-based web calculators to predict recurrence and overall survival in stage III colon cancer. *Journal of the National Cancer Institute*, 106(12), 2014.
28. Dante Wan, Diego Villa, Ryan Woods, Rinat Yerushalmi, and Karen Gelmon. Breast Cancer Subtype Variation by Race and Ethnicity in a Diverse Population in British Columbia. *Clinical breast cancer*, 16(3):e49–55, 2016.
29. Shamseddine A.I., Mukherji D., Melki C., Elias E., Eloubeidi M., Dimassi H., Khalife M., Abou-Alfa G., and O’Reilly E. Lymph node ratio is an independent prognostic factor after resection of periampullary malignancies: Data from a tertiary referral center in the middle east. *American Journal of Clinical Oncology: Cancer Clinical Trials*, 37(1):13–18, 2014.
30. Strimpakos A., Pentheroudakis G., Kotoula V., De Roock W., Kouvatsos G., Papakostas P., Makatsoris T., Papamichael D., Andreadou A., Sgouros J., Zizi-Sermpetzoglou A., Kominea A., Televantou D., Razis E., Galani E., Pectasides D., Tejpar S., and Syrigos K. The prognostic role of ephrin A2 and endothelial growth factor receptor pathway mediators in patients with advanced colorectal cancer treated with cetuximab. *Clinical Colorectal Cancer*, 12(4):267, 2013.
31. Vivek Thumbigere-Math, Lam Tu, Sabrina Huckabay, Arkadiusz Z Dudek, Scott Lunos, David L Basi, Pamela J Hughes, Joseph W Leach, Karen K Swenson, Rajaram Gopalakrishnan, Thumbigere-Math V., Tu L., Huckabay S., Dudek A.Z., Lunos S., Basi D.L., Hughes P.J., Leach J.W., and Swenson K.K. A retrospective study evaluating frequency and risk factors of osteonecrosis of the jaw in 576 cancer patients receiving intravenous bisphosphonates. *American journal of clinical oncology*, 35(4):386–392, 2012.
32. Sjoblom B., Gronberg B.H., Wentzel-Larsen T., Baracos V.E., Hjermsstad M.J., Aass N., Bremnes R.M., Flotten O., and Bye A. Skeletal muscle radiodensity is prognostic for survival in patients with advanced non-small cell lung cancer. *Clinical Nutrition*, 35(6):1386–1393, 2016.

33. Kyle A Richards, Joshua A Cohn, Michael C Large, Gregory T Bales, Norm D Smith, and Gary D Steinberg. The effect of length of ureteral resection on benign ureterointestinal stricture rate in ileal conduit or ileal neobladder urinary diversion following radical cystectomy. *Urologic oncology*, 33(2):65.e1–8, 2015.
34. Daniel A Morgenstern, Wendy B London, Derek Stephens, Samuel L Volchenbom, Barbara Hero, Andrea Di Cataldo, Akira Nakagawara, Hiroyuki Shimada, Peter F Ambros, Katherine K Matthay, Susan L Cohn, Andrew D J Pearson, and Meredith S Irwin. Metastatic neuroblastoma confined to distant lymph nodes (stage 4N) predicts outcome in patients with stage 4 disease: A study from the International Neuroblastoma Risk Group Database. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology*, 32(12):1228–1235, 2014.
35. Michael A Liss, Martha White, Loki Natarajan, and J Kellogg Parsons. Exercise Decreases and Smoking Increases Bladder Cancer Mortality. *Clinical genitourinary cancer*, 15(3):391–395, jun 2017.
36. Laura J Rasmussen-Torvik, Christina M Shay, Judith G Abramson, Christopher A Friedrich, Jennifer A Nettleton, Anna E Prizment, and Aaron R Folsom. Ideal cardiovascular health is inversely associated with incident cancer: the Atherosclerosis Risk In Communities study. *Circulation*, 127(12):1270–1275, 2013.
37. Katharine Bailey, Andy Ryan, Sophia Apostolidou, Evangelia Fourkala, Matthew Burnell, Aleksandra Gentry-Maharaj, Jatinderpal Kalsi, Max Parmar, Ian Jacobs, Hynek Pikhart, and Usha Menon. Socioeconomic indicators of health inequalities and female mortality: a nested cohort study within the United Kingdom Collaborative Trial of Ovarian Cancer Screening (UKCTOCS). *BMC public health*, 15:253, 2015.
38. Min J.-Y. Blood trihalomethane levels and the risk of total cancer mortality in US adults. *Environmental Pollution*, 212:90–96, 2016.
39. A M Behie and M H O'Donnell. Prenatal smoking and age at menarche: influence of the prenatal environment on the timing of puberty. *Human reproduction (Oxford, England)*, 30(4):957–962, 2015.
40. Mehta R., Gillan A.S., Ming Z.Y., Rai B.P., and Byrne D. Socio-economic deprivation and outcomes following radical nephroureterectomy for clinically localized upper tract transitional cell carcinoma. *World journal of urology*, 33(1):41–49, 2015.
41. Ohri N., Duan F., MacHtay M., Gorelick J.J., Snyder B.S., Alavi A., Siegel B.A., Johnson D.W., Bradley J.D., and Denittis A. Pretreatment FDG-PET metrics in stage III non-small cell lung cancer: ACRIN 6668/RTOG 0235. *Journal of the National Cancer Institute*, 107(4), 2015.

42. Lino-Silva L.S., Dominguez-Rodriguez J.A., Aguilar-Romero J.M., Martinez-Said H., Salcedo-Hernandez R.A., Garcia-Perez L., and Herrera-Gomez A. Melanoma in Mexico: Clinicopathologic Features in a Population with Predominance of Acral Lentiginous Subtype. *Annals of Surgical Oncology*, 23(13):4189–4194, 2016.
43. Anna Vogiatzoglou, Angela A Mulligan, Amit Bhaniani, Marleen A H Lentjes, Alison McTaggart, Robert N Luben, Christian Heiss, Malte Kelm, Marc W Merx, Jeremy P E Spencer, Hagen Schroeter, Kay-Tee Khaw, and Gunter G C Kuhnle. Associations between flavan-3-ol intake and CVD risk in the Norfolk cohort of the European Prospective Investigation into Cancer (EPIC-Norfolk). *Free radical biology & medicine*, 84:1–10, 2015.
44. Kelly M Cordoro, Deepti Gupta, Ilona J Frieden, Timothy McCalmont, and Mohammed Kashani-Sabet. Pediatric melanoma: results of a large cohort study and proposal for modified ABCD detection criteria for children. *Journal of the American Academy of Dermatology*, 68(6):913–925, 2013.
45. Zaragoza J., Kervarrec T., Touze A., Avenel-Audran M., Beneton N., Esteve E., Wierzbicka Hainaut E., Aubin F., Machet L., Julia Zaragoza, Thibault Kervarrec, Antoine Touze, Martine Avenel-Audran, Nathalie Beneton, Eric Esteve, Ewa Wierzbicka Hainaut, Francois Aubin, Laurent Machet, and Mahtab Samimi. A high neutrophil-to-lymphocyte ratio as a potential marker of mortality in patients with Merkel cell carcinoma: A retrospective study. *Journal of the American Academy of Dermatology*, 75(4):712–721.e1, 2016.
46. Marit Busund, Nora S. Bugge, Tonje Braaten, Marit Waaseth, Charlotta Rylander, and Eiliv Lund. Progestin-only and combined oral contraceptives and receptor-defined premenopausal breast cancer risk: The Norwegian Women and Cancer Study. *International Journal of Cancer*, feb 2018.
47. Lin-Hui Su, Li-Sheng Chen, Sheng-Che Lin, and Hsiu-Hsi Chen. Association of androgenetic alopecia with mortality from diabetes mellitus and heart disease. *JAMA dermatology*, 149(5):601–606, 2013.
48. N Saade, C Sadler, and M Goldfarb. Impact of Regional Lymph Node Dissection on Disease Specific Survival in Adrenal Cortical Carcinoma. *Hormone and metabolic research = Hormon- und Stoffwechselforschung = Hormones et metabolisme*, 47(11):820–825, 2015.
49. Daniel Orbach, Bernadette Brennan, Gianni Bisogno, Max Van Noesel, Veronique Minard-Colin, Julia Daragjati, Michela Casanova, Nadege Corradini, Ilaria Zanetti, Gian Luca De Salvo, Anne Sophie Defachelles, Anna Kelsey, Myriam Ben Arush, Nadine Francotte, and Andrea Ferrari. The EpSSG NRSTS 2005 treatment protocol

for desmoid-type fibromatosis in children: an international prospective case series. *The Lancet Child & Adolescent Health*, 1(4):284–292, dec 2017.

50. Zeljka Jutric, Jan Grendar, Helena M. Hoen, Sung W. Cho, Maria A. Cassera, Pippa H. Newell, Chet W. Hammill, Paul D. Hansen, and Ronald F. Wolf. Regional Metastatic Behavior of Nonfunctional Pancreatic Neuroendocrine Tumors. *Pancreas*, 46(7):898–903, aug 2017.
51. Harding J.L., Shaw J.E., Anstey K.J., Adams R., Balkau B., Brennan-Olsen S.L., Briffa T., Davis T.M., Davis W.A., Dobson A., Flicker L., Giles G., Grant J., Huxley R., Knuiman M., Luszcz M., Macinnis R.J., Mitchell P., Pasco J.A., Reid C., Simmons D., Simons L., Tonkin A., Woodward M., and Peeters A. Comparison of anthropometric measures as predictors of cancer incidence: A pooled collaborative analysis of 11 Australian cohorts. *International Journal of Cancer*, 2015.
52. Christopher J Keto, William J Aronson, Martha K Terris, Joseph C Presti, Christopher J Kane, Christopher L Amling, and Stephen J Freedland. Detectable prostate-specific antigen Nadir during androgen-deprivation therapy predicts adverse prostate cancer-specific outcomes: results from the SEARCH database. *European urology*, 65(3):620–627, 2014.
53. Fornaro L., Cereda S., Aprile G., Di Girolamo S., Santini D., Silvestris N., Lonardi S., Leone F., Milella M., Vivaldi C., Belli C., Bergamo F., Lutrino S.E., Filippi R., Russano M., Vaccaro V., Brunetti A.E., Rotella V., Falcone A., Barbera M.A., Corbelli J., Fasola G., Aglietta M., Zagonel V., Reni M., and Vasile E. Multivariate prognostic factors analysis for second-line chemotherapy in advanced biliary tract cancer. *British Journal of Cancer*, 110(9):2165–2169, 2014.
54. Hai-Xia Liu, Na Li, Li Wei, Fu-Xing Zhou, Rui Ma, Feng Xiao, Wei Zhang, Ying Zhang, Yan-Ping Hui, Hui Song, and Bi-Liang Chen. High expression of Kruppel-like factor 4 as a predictor of poor prognosis for cervical cancer patient response to radiotherapy. *Tumour biology : the journal of the International Society for Oncodevelopmental Biology and Medicine*, 39(6):1010428317710225, 2017.
55. Rapat Pittayanon, Rungsun Rerknimitr, and Alan Barkun. Prognostic factors affecting outcomes in patients with malignant GI bleeding treated with a novel endoscopically delivered hemostatic powder. *Gastrointestinal Endoscopy*, 87(4):994–1002, apr 2018.
56. Ahmed Elshafei, Kae Jack Tay, Onder Kara, Ercan Malkoc, Yaw Nyame, Hans Arora, Asmaa Hatem, Sahil A. Patel, Franco Lugnani, Thomas J. Polascik, and J. Stephen Jones. Associations Between Prostate Volume and Oncologic Outcomes in Men Undergoing Focal Cryoablation of the Prostate. *Clinical Genitourinary Cancer*, 16(2):e477–e482, apr 2018.

57. Akbar Fazel-Tabar Malekshah, Marsa Zaroudi, Arash Etemadi, Farhad Islami, Sadaf Sepanlou, Maryam Sharafkhah, Abbas-Ali Keshtkar, Hooman Khademi, Hossein Poustchi, Azita Hekmatdoost, Akram Pourshams, Akbar Feiz Sani, Elham Jafari, Farin Kamangar, Sanford M Dawsey, Christian C Abnet, Paul D Pharoah, Paul J Berenman, Paolo Boffetta, Ahmad Esmailzadeh, and Reza Malekzadeh. The Combined Effects of Healthy Lifestyle Behaviors on All-Cause Mortality: The Golestan Cohort Study. *Archives of Iranian medicine*, 19(11):752–761, 2016.
58. Cheng Yuan, Ning Li, Xiaoyong Mao, Zui Liu, Wei Ou, and Si-Yu Wang. Elevated pretreatment neutrophil/white blood cell ratio and monocyte/lymphocyte ratio predict poor survival in patients with curatively resected non-small cell lung cancer: Results from a large cohort. *Thoracic cancer*, 8(4):350–358, jul 2017.
59. Valerie A Smith, Roy B Sessions, and Eric J Lentsch. Cervical lymph node metastasis and papillary thyroid carcinoma: does the compartment involved affect survival? Experience from the SEER database. *Journal of surgical oncology*, 106(4):357–362, 2012.
60. Sarah Kawaguchi Jaimin R. Bhatt, Michael A. S. Jewett, Patrick O. Richard and Antonio Finelli Narhari Timilshina, Andrew Evans, Shabbir Alibhai. Multilocular cystic renal cell carcinoma: pathological t staging makes no difference to favorable outcomes and should be reclassified. *The Journal of urology*, 196:1350–1355, 2016.
61. A Necchi, R Miceli, M Bregni, C Bokemeyer, L A Berger, K Oechsle, K Schumacher, E Kanfer, J H Bourhis, C Massard, D Laszlo, J Montoro, A Flechon, F Arpaci, S Secondino, P Wuchter, P Dreger, M Crysandt, N Worel, W Kruger, M Ringhoffer, A Unal, A Nagler, A Campos, A Wahlin, M Michieli, G Sucak, I Donnini, R Schots, N Ifrah, M Badoglio, M Martino, D Raggi, P Giannatempo, G Rosti, P Pedrazzoli, and F Lanza. Prognostic impact of progression to induction chemotherapy and prior paclitaxel therapy in patients with germ cell tumors receiving salvage high-dose chemotherapy in the last 10 years: a study of the European Society for Blood and Marrow Transplantation S. *Bone marrow transplantation*, 51(3):384–390, 2016.
62. E Susan Amirian, Terri S Armstrong, Kenneth D Aldape, Mark R Gilbert, and Michael E Scheurer. Predictors of survival among pediatric and adult ependymoma cases: a study using Surveillance, Epidemiology, and End Results data from 1973 to 2007. *Neuroepidemiology*, 39(2):116–124, 2012.
63. McCabe E.L., Larson M.G., Lunetta K.L., Newman A.B., and Cheng S. Association of an Index of Healthy Aging With Incident Cardiovascular Disease and Mortality in a Community-Based Sample of Older Adults. *The journals of gerontology. Series A, Biological sciences and medical sciences*, 71(12):1695–1701, 2016.

64. Alicja Puszkiel, Melanie White-Koning, Nicolas Dupin, Nora Kramkimel, Audrey Thomas-Schoemann, Gaelle Noe, Nicolas Chapuis, Michel Vidal, Francois Goldwasser, Etienne Chatelut, and Benoit Blanchet. Plasma vemurafenib exposure and pre-treatment hepatocyte growth factor level are two factors contributing to the early peripheral lymphocytes depletion in BRAF-mutated melanoma patients. *Pharmacological research*, 113(Pt A):709–718, 2016.
65. Piet A van den Brandt and Maya Schulpen. Mediterranean diet adherence and risk of postmenopausal breast cancer: results of a cohort study and meta-analysis. *International journal of cancer*, 140(10):2220–2231, may 2017.
66. Ferguson M.K., Watson S., and Johnson E. Predicted postoperative lung function is associated with all-cause long-term mortality after major lung resection for cancer. *European Journal of Cardio-thoracic Surgery*, 45(4):660–664, 2014.
67. M Yi, L Huo, K B Koenig, E A Mittendorf, F Meric-Bernstam, H M Kuerer, I Bedrosian, A U Buzdar, W F Symmans, J R Crow, M Bender, R R Shah, G N Hortobagyi, and K K Hunt. Which threshold for ER positivity? a retrospective study based on 9639 patients. *Annals of oncology : official journal of the European Society for Medical Oncology*, 25(5):1004–1011, 2014.
68. Lv J.-W., Chen Y.-P., Zhou G.-Q., Tang L.-L., Mao Y.-P., Li W.-F., Guo R., Lin A.-H., and Ma J. Cigarette smoking complements the prognostic value of baseline plasma Epstein-Barr virus deoxyribonucleic acid in patients with nasopharyngeal carcinoma undergoing intensity-modulated radiation therapy: A large-scale retrospective cohort study. *Oncotarget*, 7(13):16806–16817, 2016.
69. John J Coen, Jonathan J Paly, Andrzej Niemierko, Donald S Kaufman, Niall M Heney, Daphne Y Spiegel, Jason A Efsthathiou, Anthony L Zietman, and William U Shipley. Nomograms predicting response to therapy and outcomes after bladder-preserving trimodality therapy for muscle-invasive bladder cancer. *International journal of radiation oncology, biology, physics*, 86(2):311–316, 2013.
70. Takahashi M., Komine K., Yamada H., Kasahara Y., Chikamatsu S., Okita A., Ito S., Ouchi K., Okada Y., Imai H., Saijo K., Shimodaira Shirota H., Takahashi S., Mori T., Shimodaira Shirota H., Masahiro Masanobu Takahashi, Masahiro Masanobu Takahashi, Keigo Komine, Hideharu Yamada, Yuki Kasahara, Sonoko Chikamatsu, Akira Okita, Shukuei Ito, Kota Ouchi, Yoshinari Okada, Hiroo Imai, Ken Saijo, Hidekazu Shirota, Shin Takahashi, Takahiro Mori, Hideki Shimodaira, and Chikashi Ishioka. The G8 screening tool enhances prognostic value to ECOG performance status in elderly cancer patients: A retrospective, single institutional study. *PLoS ONE*, 12(6):e0179694, 2017.

71. Jens Kohler, Martin Schuler, Thomas Christoph Gauler, Stefanie Nopel-Dunnebacke, Maike Ahrens, Andreas-Claudius Hoffmann, Stefan Kasper, Felix Nensa, Benedikt Gomez, Maria Hahnemann, Frank Breitenbuecher, Danjouma Cheufou, Filiz Ozkan, Kaid Darwiche, Mathias Hoicznyk, Henning Reis, Stefan Welter, Wilfried Ernst Erich Eberhardt, Martin Eisenacher, Helmut Teschler, Georgios Stamatis, Wolff Schmiegel, Stephan Albrecht Hahn, and Alexander Baraniskin. Circulating U2 small nuclear RNA fragments as a diagnostic and prognostic biomarker in lung cancer patients. *Journal of cancer research and clinical oncology*, 142(4):795–805, 2016.
72. Paly J.J., Hallemeier C.L., Biggs P.J., Niemierko A., Roeder F., Martinez-Monge R., Whitson J., Calvo F.A., Fastner G., Sedlmayer F., Wong W.W., Ellis R.J., Haddock M.G., Choo R., Shipley W.U., and Zietman A.L. Outcomes in a multi-institutional cohort of patients treated with intraoperative radiation therapy for advanced or recurrent renal cell carcinoma. *International Journal of Radiation Oncology Biology Physics*, 88(3):618–623, 2014.
73. Gabriel E., Attwood K., Thirunavukarasu P., Al-Sukhni E., Boland P., Emmanuel Gabriel, Kristopher Attwood, Pragatheeshwar Thirunavukarasu, Eisar Al-Sukhni, Patrick Boland, and Steven Nurkin. Predicting Individualized Postoperative Survival for Stage II/III Colon Cancer Using a Mobile Application Derived from the National Cancer Data Base. *Journal of the American College of Surgeons*, 222(3):232–244, 2016.
74. Marko Lukic, Mie Jareid, Elisabete Weiderpass, and Tonje Braaten. Coffee consumption and the risk of malignant melanoma in the Norwegian Women and Cancer (NOWAC) Study. *BMC cancer*, 16:562, 2016.
75. G Gandaglia, G Lista, N Fossati, N Suardi, A Gallina, M Moschini, L Bianchi, M S Rossi, R Schiavina, S F Shariat, A Salonia, F Montorsi, and A Briganti. Non-surgically related causes of erectile dysfunction after bilateral nerve-sparing radical prostatectomy. *Prostate cancer and prostatic diseases*, 19(2):185–190, 2016.
76. Guru Sonpavde, Gregory R Pond, Andrew J Armstrong, Stephen J Clarke, Janette L Vardy, Arnoud J Templeton, Shaw-Ling Wang, Jolanda Paolini, Isan Chen, Edna Chow-Maneval, Mariajose Lechuga, Matthew R Smith, and M Dror Michaelson. Prognostic impact of the neutrophil-to-lymphocyte ratio in men with metastatic castration-resistant prostate cancer. *Clinical genitourinary cancer*, 12(5):317–324, 2014.
77. Geoffrey C Kabat, Charles E Matthews, Victor Kamensky, Albert R Hollenbeck, and Thomas E Rohan. Adherence to cancer prevention guidelines and cancer incidence, cancer mortality, and total mortality: a prospective cohort study. *The American journal of clinical nutrition*, 101(3):558–569, 2015.

78. Li J., Eriksson M., Czene K., and Hall P. Common diseases as determinants of menopausal age. *Human Reproduction*, 31(12):2856–2864, 2016.
79. B E Shaw, N P Mayor, R M Szydlo, W P Bultitude, C Anthias, K Kirkland, J Perry, A Clark, S Mackinnon, D I Marks, A Pagliuca, M N Potter, N H Russell, K Thomson, J A Madrigal, and S G E Marsh. Recipient/donor HLA and CMV matching in recipients of T-cell-depleted unrelated donor haematopoietic cell transplants. *Bone marrow transplantation*, 52(5):717–725, 2017.
80. Melina Arnold, Luohua Jiang, Marcia L Stefanick, Karen C Johnson, Dorothy S Lane, Erin S LeBlanc, Ross Prentice, Thomas E Rohan, Beverly M Snively, Mara Vitolins, Oleg Zaslavsky, Isabelle Soerjomataram, Hoda Anton-Culver, Arnold M., Jiang L., Stefanick M.L., Johnson K.C., Lane D.S., LeBlanc E.S., Prentice R., Rohan T.E., Snively B.M., Vitolins M., Zaslavsky O., Soerjomataram I., and Melina; ORCID: <http://orcid.org/0000-0003-1700-6831> A O Jiang Anton-Culver H. AO - Arnold Luohua; ORCID: <http://orcid.org/0000-0002-2281-7260>. Duration of Adulthood Overweight, Obesity, and Cancer Risk in the Women’s Health Initiative: A Longitudinal Study from the United States. *PLoS medicine*, 13(8):e1002081, 2016.
81. Atsumu Yuki, Rei Otsuka, Chikako Tange, Yukiko Nishita, Makiko Tomida, Fujiko Ando, and Hiroshi Shimokata. Physical frailty and mortality risk in Japanese older adults. *Geriatrics & Gerontology International*, apr 2018.
82. Brandon-Luke L. Seagle, Amy L. Alexander, Taliya Lantsman, and Shohreh Shahabi. Prognosis and treatment of positive peritoneal cytology in early endometrial cancer: matched cohort analyses from the National Cancer Database. *American Journal of Obstetrics and Gynecology*, 218(3):329.e1–329.e15, mar 2018.
83. Elizabeth Kersten, Patricia Scanlan, Steven G Dubois, and Katherine K Matthay. Current treatment and outcome for childhood acute leukemia in Tanzania. *Pediatric blood & cancer*, 60(12):2047–2053, 2013.
84. Masahiro Yanagiya, Jun-ichi Nitadori, Kazuhiro Nagayama, Masaki Anraku, Masaaki Sato, and Jun Nakajima. Prognostic significance of the preoperative neutrophil-to-lymphocyte ratio for complete resection of thymoma. *Surgery Today*, 48(4):422–430, apr 2018.
85. Miller R.E., Markt S.C., O’Donnell E., Bernard B., Albiges L.K., and Beard C. Age ≥ 40 Years Is Associated with Adverse Outcome in Metastatic Germ Cell Cancer Despite Appropriate Intended Chemotherapy. *European Urology Focus*, 2016.
86. Sara Hallden, Marketa Sjogren, Bo Hedblad, Gunnar Engstrom, Krzysztof Narkiewicz, Michal Hoffmann, Bjorn Wahlstrand, Thomas Hedner, and Olle Melander. Smok-

ing and obesity associated BDNF gene variance predicts total and cardiovascular mortality in smokers. *Heart (British Cardiac Society)*, 99(13):949–953, 2013.

87. Katherine A Janeway, Donald A Barkauskas, Mark D Krailo, Paul A Meyers, Cindy L Schwartz, David H Ebb, Nita L Seibel, Holcombe E Grier, Richard Gorlick, and Neyssa Marina. Outcome for adolescent and young adult patients with osteosarcoma: a report from the Children’s Oncology Group. *Cancer*, 118(18):4597–4605, 2012.
88. David B Stewart, Christopher Hollenbeak, Susan Desharnais, Fabian Camacho, Patricia Gladowski, Vickie L Goff, and Li Wang. Rectal cancer and teaching hospitals: hospital teaching status affects use of neoadjuvant radiation and survival for rectal cancer patients. *Annals of surgical oncology*, 20(4):1156–1163, 2013.
89. Maria Iachina, Erik Jakobsen, Anne Kudsk Fallesen, and Anders Green. Transfer between hospitals as a predictor of delay in diagnosis and treatment of patients with Non-Small Cell Lung Cancer - a register based cohort-study. *BMC health services research*, 17(1):267, 2017.
90. Wu S.-G., Zhang Z.-Q., Liu W.-M., He Z.-Y., Li F.-Y., Lin H.-X., Sun J.-Y., and Lin H. Impact of the number of resected lymph nodes on survival after preoperative radiotherapy for esophageal cancer. *Oncotarget*, 7(16):22497–22507, 2016.
91. Di Lorenzo G., Buonerba C., Bellelli T., Romano C., Montanaro V., Ferro M., Benincasa A., Ribera D., Lucarelli G., De Cobelli O., Sonpavde G., Giuseppe Di Lorenzo, Carlo Buonerba, Teresa Bellelli, Concetta Romano, Vittorino Montanaro, Matteo Ferro, Alfonso Benincasa, Dario Ribera, Giuseppe Lucarelli, Ottavio De Cobelli, Guru Sonpavde, Sabino De Placido, Di Lorenzo G., Buonerba C., Bellelli T., Romano C., Montanaro V., Ferro M., Benincasa A., Ribera D., Lucarelli G., De Cobelli O., Sonpavde G., and De Placido S. Third-Line chemotherapy for metastatic Urothelial Cancer: A retrospective observational study. *Medicine*, 94(51):e2297, 2015.
92. Amanda I Phipps, Paul J Limburg, John A Baron, Andrea N Burnett-Hartman, Daniel J Weisenberger, Peter W Laird, Frank A Sinicrope, Christophe Rosty, Daniel D Buchanan, John D Potter, and Polly A Newcomb. Association between molecular subtypes of colorectal cancer and patient survival. *Gastroenterology*, 148(1):77–87.e2, 2015.
93. Marco Carbone, Alessandra Nardi, Tania Marianelli, Kate Martin, Alex Hudson, David Collett, Renato Romagnoli, Antonio Pinna, Alexander Gimson, James M Neuberger, Mario Angelico, , Liver Match Investigators for Italian Association for the Study of the Liver National Transplant Centre Transplant, and the National

Health System Blood. International comparison of liver transplant programmes: differences in indications, donor and recipient selection and outcome between Italy and UK. *Liver international : official journal of the International Association for the Study of the Liver*, 36(10):1481–1489, 2016.

94. Thompson E.M., Hielscher T., Bouffet E., Remke M., Luu B., Gururangan S., McLendon R.E., Bigner D.D., Lipp E.S., Perreault S., Cho Y.-J., Grant G., Kim S.-K., Lee J.Y., Rao A.A.N., Giannini C., Li K.K.W., Ng H.-K., Yao Y., Kumabe T., Tominaga T., Grajkowska W.A., Perek-Polnik M., Low D.C.Y., Seow W.T., Chang K.T.E., Mora J., Pollack I.F., Hamilton R.L., Leary S., Moore A.S., Ingram W.J., Hallahan A.R., Jouvett A., Fevre-Montange M., Vasiljevic A., Faure-Contier C., Shofuda T., Kagawa N., Hashimoto N., Jabado N., Weil A.G., Gayden T., Wataya T., Shalaby T., Grotzer M., Zitterbart K., Sterba J., Kren L., Hortobagyi T., Klekner A., Laszlo B., Pocza T., Hauser P., Schuller U., Jung S., Jang W.-Y., French P.J., Kros J.M., van Veelen M.-L.C., Massimi L., Leonard J.R., Rubin J.B., Vibhakkar R., Chambless L.B., Cooper M.K., Thompson R.C., Faria C.C., Carvalho A., Nunes S., Pimentel J., Fan X., Muraszko K.M., Lopez-Aguilar E., Lyden D., Garzia L., Shih D.J.H., Kijima N., Schneider C., Adamski J., Northcott P.A., Kool M., Jones D.T.W., Chan J.A., Nikolic A., Garre M.L., Van Meir E.G., Osuka S., Olson J.J., Jahangiri A., Castro B.A., Gupta N., Weiss W.A., Moxon-Emre I., Mabbott D.J., Lassaletta A., Hawkins C.E., Tabori U., Drake J., Kulkarni A., Dirks P., Rutka J.T., Korshunov A., Pfister S.M., Packer R.J., and Ramaswamy V. Prognostic value of medulloblastoma extent of resection after accounting for molecular subgroup: a retrospective integrated clinical and molecular analysis. *The Lancet Oncology*, 17(4):484–495, 2016.
95. Farkhad Manapov, Maximilian Niyazi, Sabine Gerum, Olarn Roengvoraphoj, Chukwuka Eze, Minglun Li, Guido Hildebrandt, Rainer Fietkau, Gunther Klautke, and Claus Belka. Evaluation of the role of remission status in a heterogeneous limited disease small-cell lung cancer patient cohort treated with definitive chemoradiotherapy. *BMC cancer*, 16:216, 2016.
96. A. Ciarrocchi, R. Pietroletti, F. Carlei, and G. Amicucci. Extensive surgery and lymphadenectomy do not improve survival in primary melanoma of the anorectum: results from analysis of a large database (SEER). *Colorectal Disease*, 19(2):158–164, feb 2017.
97. Sughosh Dhakal, James E Bates, Carla Casulo, Jonathan W Friedberg, Michael W Becker, Jane L Liesveld, and Louis S Constine. Patterns and Timing of Failure for Diffuse Large B-Cell Lymphoma After Initial Therapy in a Cohort Who Underwent Autologous Bone Marrow Transplantation for Relapse. *International journal of radiation oncology, biology, physics*, 96(2):372–378, 2016.

98. Ali H.R., Dawson S.-J., Blows F.M., Provenzano E., Leung S., Nielsen T., and Pharoah P.D. A Ki67/BCL2 index based on immunohistochemistry is highly prognostic in ER-positive breast cancer. *Journal of Pathology*, 226(1):97–107, 2012.
99. Seok Jin Kim, Dok Hyun Yoon, Arnaud Jaccard, Wee Joo Chng, Soon Thye Lim, Huangming Hong, Yong Park, Kian Meng Chang, Yoshinobu Maeda, Fumihiko Ishida, Dong-Yeop Shin, Jin Seok Kim, Seong Hyun Jeong, Deok-Hwan Yang, Jae-Cheol Jo, Gyeong-Won Lee, Chul Won Choi, Won-Sik Lee, Tsai-Yun Chen, Kiyeun Kim, Sin-Ho Jung, Tohru Murayama, Yasuhiro Oki, Ranjana Advani, Francesco D'Amore, Norbert Schmitz, Cheolwon Suh, Ritsuro Suzuki, Yok Lam Kwong, Tong-Yu Lin, and Won Seog Kim. A prognostic index for natural killer cell lymphoma after non-anthracycline-based treatment: a multicentre, retrospective analysis. *The Lancet. Oncology*, 17(3):389–400, 2016.
100. Iero Meattini, Calogero Saieva, Paolo Bastiani, Francesca Martella, Giulio Francolini, Monica Lo Russo, Lisa Paoletti, Morena Doria, Isacco Desideri, Francesca Terziani, Carla De Luca Cardillo, Benedetta Bendinelli, Cinzia Ciabatti, Cristina Muntoni, Galliano Tinacci, Jacopo Nori, Herd Smith, Beniamino Brancato, Lorenzo Galli, Luis Jose Sanchez, Donato Casella, Marco Bernini, Lorenzo Orzalesi, Giulio Alberto Carta, Simonetta Bianchi, Francesca Rossi, and Lorenzo Livi. Impact of hormonal status on outcome of ductal carcinoma in situ treated with breast-conserving surgery plus radiotherapy: Long-term experience from two large-institutional series. *Breast (Edinburgh, Scotland)*, 33:139–144, 2017.
101. Piet A van den Brandt and Leo J Schouten. Relationship of tree nut, peanut and peanut butter intake with total and cause-specific mortality: a cohort study and meta-analysis. *International journal of epidemiology*, 44(3):1038–1049, 2015.
102. Ali H.R., Provenzano E., Dawson S.-J., Blows F.M., Liu B., Shah M., Earl H.M., Poole C.J., Hiller L., Dunn J.A., Bowden S.J., Twelves C., Bartlett J.M.S., Mahmoud S.M.A., Rakha E., Ellis I.O., Liu S., Gao D., Nielsen T.O., and Pharoah P.D.P. Association between CD8+ T-cell infiltration and breast cancer survival in 12 439 patients. *Annals of Oncology*, 25(8):1536–1543, 2014.
103. Dianed Zheng D., Christ S.L., Lam B.L., Arheart K.L., and Galor A. Increased mortality risk among the visually impaired: The roles of mental well-being and preventive care practices. *Investigative Ophthalmology and Visual Science*, 53(6):2685–2692, 2012.
104. Elias Jabbour, Guillermo Garcia-Manero, A Megan Cornelison, Jorge E Cortes, Farhad Ravandi, Naval Daver, Tapan Kadia, Angela Teng, and Hagop Kantarjian. The effect of decitabine dose modification and myelosuppression on response

- and survival in patients with myelodysplastic syndromes. *Leukemia & lymphoma*, 56(2):390–394, 2015.
105. N A Quraishi, S R Manoharan, G Arealis, A Khurana, S Elsayed, K L Edwards, and B M Boszczyk. Accuracy of the revised Tokuhashi score in predicting survival in patients with metastatic spinal cord compression (MSCC). *European spine journal: official publication of the European Spine Society, the European Spinal Deformity Society, and the European Section of the Cervical Spine Research Society*, 22 Suppl 1:S21–6, 2013.
106. Rance J.T. Fujiwara, Barbara Burtness, Zain A. Husain, Benjamin L. Judson, Aarti Bhatia, Clarence T. Sasaki, Wendell G. Yarbrough, and Saral Mehra. Treatment guidelines and patterns of care in oral cavity squamous cell carcinoma: Primary surgical resection vs. nonsurgical treatment. *Oral Oncology*, 71:129–137, aug 2017.
107. Daniel Carlzon, Johan Svensson, Max Petzold, Magnus K Karlsson, Osten Ljunggren, Mohammad-Ali Haghsheno, Jan-Erik Damber, Dan Mellstrom, and Claes Ohlsson. Insulin-like growth factor I and risk of incident cancer in elderly men - results from MrOS (Osteoporotic Fractures in Men) in Sweden. *Clinical endocrinology*, 84(5):764–770, 2016.
108. Prinelli F., Yannakoulia M., Anastasiou C.A., Adorni F., Di Santo S.G., Musicco M., and Scarmeas N. Mediterranean diet and other lifestyle factors in relation to 20-year all-cause mortality: A cohort study in an Italian population. *British Journal of Nutrition*, 113(6):1003–1011, 2015.
109. Eugene P Ceppa, Alexandra M Roch, Jessica L Cioffi, Neil Sharma, Jeffrey J Easler, John M DeWitt, Michael G House, Nicholas J Zyromski, Attila Nakeeb, and C Max Schmidt. Invasive, mixed-type intraductal papillary mucinous neoplasm: superior prognosis compared to invasive main-duct intraductal papillary mucinous neoplasm. *Surgery*, 158(4):935–937, 2015.
110. Naamit K Gerber, Yoshiya Yamada, Andreas Rimmer, Weiji Shi, Gregory J Riely, Kathryn Beal, Helena A Yu, Timothy A Chan, Zhigang Zhang, and Abraham J Wu. Erlotinib versus radiation therapy for brain metastases in patients with EGFR-mutant lung adenocarcinoma. *International journal of radiation oncology, biology, physics*, 89(2):322–329, 2014.
111. Danielle Rodin, Michael Drumm, Rebecca Clayman, Daniela L. Buscariollo, Sigolene Galland-Girodet, Alec Eidelman, Adam S. Feldman, Douglas M. Dahl, Francis J. McGovern, Aria F. Olumi, Andrzej Niemierko, William U. Shipley, Anthony L. Zietman, and Jason A. Efstathiou. Risk Factors for Disease Progression After Post-prostatectomy Salvage Radiation: Long-term Results of a Single-institution Experience. *Clinical Genitourinary Cancer*, 16(1):21–27.e1, feb 2018.

112. Kent M.S., Mandrekar S.J., Landreneau R., Nichols F., Foster N.R., Dipetrillo T.A., Meyers B., Heron D.E., Jones D.R., Tan A.D., Starnes S., and Putnam J.B. A Nomogram to Predict Recurrence and Survival of High-Risk Patients Undergoing Sublobar Resection for Lung Cancer: An Analysis of a Multicenter Prospective Study (ACOSOG Z4032). *Annals of Thoracic Surgery*, 102(1):239–246, 2016.
113. Anthony N Karnezis, Samuel Leung, Jamie Magrill, Melissa K McConechy, Winnie Yang, Christine Chow, Martin Kobel, Cheng-Han Lee, David G Huntsman, Aline Talhouk, Friederich Kommoss, C Blake Gilks, and Jessica N McAlpine. Evaluation of endometrial carcinoma prognostic immunohistochemistry markers in the context of molecular classification. *The Journal of Pathology: Clinical Research*, 3(4):279–293, oct 2017.
114. Akihiro Naito, Satoru Taguchi, Tohru Nakagawa, Akihiko Matsumoto, Yasushi Nagase, Mariko Tabata, Jimpei Miyakawa, Motofumi Suzuki, Hiroaki Nishimatsu, Yutaka Enomoto, Shintaro Takahashi, Toshikazu Okaneya, Daisuke Yamada, Takamitsu Tachikawa, Shigeru Minowada, Tetsuya Fujimura, Hiroshi Fukuhara, Haruki Kume, and Yukio Homma. Prognostic significance of serum neuron-specific enolase in small cell carcinoma of the urinary bladder. *World journal of urology*, 35(1):97–103, 2017.
115. Satoru Taguchi, Tohru Nakagawa, Akihiko Matsumoto, Yasushi Nagase, Taketo Kawai, Yoshinori Tanaka, Kanae Yoshida, Sachi Yamamoto, Yutaka Enomoto, Yorito Nose, Toshikazu Sato, Akira Ishikawa, Yukari Uemura, Tetsuya Fujimura, Hiroshi Fukuhara, Haruki Kume, and Yukio Homma. Pretreatment neutrophil-to-lymphocyte ratio as an independent predictor of survival in patients with metastatic urothelial carcinoma: A multi-institutional study. *International journal of urology : official journal of the Japanese Urological Association*, 22(7):638–643, 2015.
116. Lars Barregard, Gerd Sallsten, Bjorn Fagerberg, Yan Borne, Margaretha Persson, Bo Hedblad, and Gunnar Engstrom. Blood Cadmium Levels and Incident Cardiovascular Events during Follow-up in a Population-Based Cohort of Swedish Adults: The Malmo Diet and Cancer Study. *Environmental health perspectives*, 124(5):594–600, 2016.
117. Chen S., Huang L., Liu Y., Chen C.M., and Wu J. The predictive and prognostic significance of pre- and post-treatment topoisomerase IIalpha in anthracycline-based neoadjuvant chemotherapy for local advanced breast cancer. *European Journal of Surgical Oncology*, 39(6):619–626, 2013.
118. Mantripragada K.C., Hamid F., Shafqat H., Kalyan C Mantripragada, Fatima Hamid, Hammad Shafqat, and Adam J Olszewski. Adjuvant Therapy for Resected Gallbladder Cancer: Analysis of the National Cancer Data Base. *Journal of the National Cancer Institute*, 109(2), 2017.

- 119.Randy C Miles, Rachel E Gullerud, Christine M Lohse, James W Jakub, Amy C Degnim, and Judy C Boughey. Local recurrence after breast-conserving surgery: multivariable analysis of risk factors and the impact of young age. *Annals of surgical oncology*, 19(4):1153–1159, 2012.
- 120.Leu S., Von Felten S., Frank S., Vassella E., Vajtai I., Taylor E., Schulz M., Hutter G., Hench J., Schucht P., Boulay J.-L., Severina Leu, Stefanie von Felten, Stephan Frank, Erik Vassella, Istvan Vajtai, Elisabeth Taylor, Marianne Schulz, Gregor Hutter, Jurgen Hench, Philippe Schucht, Jean-Louis Boulay, and Luigi Mariani. IDH/MGMT-driven molecular classification of low-grade glioma is a strong predictor for long-term survival. *Neuro-Oncology*, 15(4):469–479, 2013.
- 121.Joost C. de Vries, Berdien Oortgiesen, Marc H. Hemmeler, Eric van Roon, Robby E. Kibbelaar, Nic Veeger, and Mels Hoogendoorn. Restoration of renal function in patients with newly diagnosed multiple myeloma is not associated with improved survival: a population-based study. *Leukemia & Lymphoma*, 58(9):2101–2109, sep 2017.
- 122.Fradet V., Mauermann J., Kassouf W., Rendon R., Jacobsen N., Fairey A., Izawa J., Kapoor A., Black P., Tanguay S., Chin J., So A., Lattouf J.-B., Bell D., Saad F., Sheyegan B., Drachenberg D., and Cagiannos I. Risk factors for bladder cancer recurrence after nephroureterectomy for upper tract urothelial tumors: Results from the Canadian Upper Tract Collaboration. *Urologic Oncology: Seminars and Original Investigations*, 32(6):839–845, 2014.
- 123.G David Batty, Tom C Russ, Emmanuel Stamatakis, and Mika Kivimaki. Psychological distress in relation to site specific cancer mortality: Pooling of unpublished data from 16 prospective cohort studies. *BMJ (Online)*, 356:j108, 2017.
- 124.Waqas R Shaikh, Stephen W Dusza, Martin A Weinstock, Susan A Oliveria, Alan C Geller, and Allan C Halpern. Melanoma Thickness and Survival Trends in the United States, 1989 to 2009. *Journal of the National Cancer Institute*, 108(1), jan 2016.
- 125.Schmidt N., Hess V., Zumbrunn T., Rothermundt C., and Bongartz G. Choi response criteria for prediction of survival in patients with metastatic renal cell carcinoma treated with anti-angiogenic therapies. *European Radiology*, 23(3):632–639, 2013.
- 126.Sara Alonso-Alvarez, Laura Magnano, Miguel Alcoceba, Marcio Andrade-Campos, Natalia Espinosa-Lara, Guillermo Rodriguez, Santiago Mercadal, Itziar Carro, Juan M Sancho, Miriam Moreno, Antonio Salar, Francesc Garcia-Pallarols, Reyes Arranz, Jimena Cannata, Maria Jose Terol, Ana I Teruel, Antonia Rodriguez, Ana Jimenez-Ubieto, Sonia Gonzalez de Villambrosia, Jose L Bello, Lourdes Lopez,

- Silvia Monsalvo, Silvana Novelli, Erik de Cabo, Maria S Infante, Emilia Pardal, Maria Garcia-Alvarez, Julio Delgado, Marcos Gonzalez, Alejandro Martin, Armando Lopez-Guillermo, and Maria D Caballero. Risk of, and survival following, histological transformation in follicular lymphoma in the rituximab era. A retrospective multicentre study by the Spanish GELTAMO group. *British journal of haematology*, 178(5):699–708, 2017.
127. Palak J Trivedi, Willem J Lammers, Henk R van Buuren, Albert Pares, Annarosa Floreani, Harry L A Janssen, Pietro Invernizzi, Pier Maria Battezzati, Cyriel Y Ponsioen, Christophe Corpechot, Raoul Poupon, Marlyn J Mayo, Andrew K Burroughs, Frederik Nevens, Andrew L Mason, Kris V Kowdley, Ana Lleo, Llorenç Caballeria, Keith D Lindor, Bettina E Hansen, Gideon M Hirschfield, and Global P B C Study Group. Stratification of hepatocellular carcinoma risk in primary biliary cirrhosis: a multicentre international study. *Gut*, 65(2):321–329, 2016.
128. Malte W Vetterlein, Philipp Gild, Luis A Kluth, Thomas Seisen, Michael Gierth, Hans-Martin Fritsche, Maximilian Burger, Chris Protzel, Oliver W Hakenberg, Nicolas von Landenberg, Florian Roghmann, Joachim Noldus, Philipp Nuhn, Armin Pycha, Michael Rink, Felix K-H Chun, Matthias May, Margit Fisch, Atiqullah Aziz, and PROMETRICS 2011 Study Group. Peri-operative allogeneic blood transfusion does not adversely affect oncological outcomes after radical cystectomy for urinary bladder cancer: a propensity score-weighted European multicentre study. *BJU international*, 121(1):101–110, 2018.
129. Peters M., van der Voort van Zyp J.R.N., Moerland M.A., Hoekstra C.J., van de Pol S., Westendorp H., Maenhout M., Kattevilder R., Verkooijen H.M., van Rossum P.S.N., Ahmed H.U., Shah T.T., Emberton M., and van Vulpen M A O Peters M.; ORCID: <http://orcid.org/0000-0002-7981-5768> A O Westendorp H.; ORCID: <http://orcid.org/0000-0001-9549-2391> A O Shah T.T.; ORCID: <http://orcid.org/0000-0003-1642-1208>. Development and internal validation of a multivariable prediction model for biochemical failure after whole-gland salvage iodine-125 prostate brachytherapy for recurrent prostate cancer. *Brachytherapy*, 15(3):296–305, 2016.
130. Mette Calundann Noer, Pia Leandersson, Torbjørn Paulsen, Susanne Rosthøj, Sofie Leisby Antonsen, Christer Borgfeldt, and Claus Høgdall. Confounders other than comorbidity explain survival differences in Danish and Swedish ovarian cancer patients – a comparative cohort study. *Acta Oncologica*, pages 1–9, feb 2018.
131. Nathan Papa, Nathan Lawrentschuk, David Muller, Robert MacInnis, Anthony Ta, Gianluca Severi, Jeremy Millar, Rodney Syme, Graham Giles, and Damien Bolton. Rural residency and prostate cancer specific mortality: results from the Victorian Radical Prostatectomy Register. *Australian and New Zealand journal of public health*, 38(5):449–454, 2014.

- 132.Clive A.O., Kahan B.C., Hooper C.E., Bhatnagar R., Morley A.J., Zahan-Evans N., Bintcliffe O.J., Boshuizen R.C., Fysh E.T.H., Tobin C.L., Medford A.R.L., Harvey J.E., Van Den Heuvel M.M., and Lee Y.C.G. Predicting survival in malignant pleural effusion: Development and validation of the LENT prognostic score. *Thorax*, 69(12):1098–1104, 2014.
- 133.Shanna A. Arnold Egloff, Liping Du, Holli A. Loomans, Alina Starchenko, Pei-Fang Su, Tatiana Ketova, Paul B. Knoll, Jifeng Wang, Ahmed Q. Haddad, Oluwole Fadare, Justin M. Cates, Yair Lotan, Yu Shyr, Peter E. Clark, and Andries Zijlstra. Shed urinary ALCAM is an independent prognostic biomarker of three-year overall survival after cystectomy in patients with bladder cancer. *Oncotarget*, 8(1), jan 2017.
- 134.Riedel D.J., Cox E.R., and Stafford K.A. Clinical presentation and outcomes of prostate cancer in an urban cohort of predominantly African American, human immunodeficiency virus-infected patients. *Urology*, 85(2):415–421, 2015.
- 135.M C van Maaren, L de Munck, J J Jobsen, P Poortmans, G H de Bock, S Sieling, and L J A Strobbe. Breast-conserving therapy versus mastectomy in T1-2N2 stage breast cancer: a population-based study on 10-year overall, relative, and distant metastasis-free survival in 3071 patients. *Breast cancer research and treatment*, 160(3):511–521, 2016.
- 136.Paul W Sperduto, Norbert Kased, David Roberge, Samuel T Chao, Ryan Shanley, Xianghua Luo, Penny K Sneed, John Suh, Robert J Weil, Ashley W Jensen, Paul D Brown, Helen A Shih, John Kirkpatrick, Laurie E Gaspar, John B Fiveash, Veronica Chiang, Jonathan P S Knisely, Christina Maria Sperduto, Nancy Lin, and Minesh Mehta. The effect of tumor subtype on the time from primary diagnosis to development of brain metastases and survival in patients with breast cancer. *Journal of neuro-oncology*, 112(3):467–472, 2013.
- 137.Cata J.P., Jones J., Sepesi B., Mehran R.J., Rodriguez-Restrepo A., Lasala J., and Feng L. Lack of Association Between Dexamethasone and Long-Term Survival After Non-Small Cell Lung Cancer Surgery. *Journal of Cardiothoracic and Vascular Anesthesia*, 30(4):930–935, 2016.
- 138.Johann von Felden, Denise Heim, Kornelius Schulze, Till Krech, Florian Ewald, Bjorn Nashan, Ansgar W Lohse, and Henning Wege. High expression of micro RNA- 135A in hepatocellular carcinoma is associated with recurrence within 12 months after resection. *BMC cancer*, 17(1):60, 2017.
- 139.Malte W. Vetterlein, Julia Roschinski, Philipp Gild, Phillip Marks, Armin Soave, Ousman Doh, Hendrik Isbarn, Wolfgang H"oppner, Walter Wagner, Shahrokh F.

- Shariat, Maurizio Brausi, Franziska Bu"schek, Guido Sauter, Margit Fisch, and Michael Rink. Impact of the Ki-67 labeling index and p53 expression status on disease-free survival in pT1 urothelial carcinoma of the bladder. *Translational Andrology and Urology*, 6(6):1018–1026, dec 2017.
140. Nemelec R.M., Stadhouders A., and van Royen B.J. The outcome and survival of palliative surgery in thoraco-lumbar spinal metastases: contemporary retrospective cohort study. *European Spine Journal*, 2014.
141. Ishimaru M., Ono S., Suzuki S., Matsui H., Fushimi K., Miho; ORCID: <http://orcid.org/0000-0002-5269-5698> Yasunaga H. AO - Ishimaru, Miho Ishimaru, Sachiko Ono, Sayaka Suzuki, Hiroki Matsui, Kiyohide Fushimi, and Hideo Yasunaga. Risk Factors for Free Flap Failure in 2,846 Patients With Head and Neck Cancer: A National Database Study in Japan. *Journal of oral and maxillofacial surgery: official journal of the American Association of Oral and Maxillofacial Surgeons*, 74(6):1265–1270, 2016.
142. Hsia T.E.-C., Tu C.-Y., Chen H.-J., Chen S.-C., Liang J.I.-A.N., Chen C.-Y.I., and Wang Y.-C. A population-based study of primary chemoradiotherapy in clinical stage III non-small cell lung cancer: Intensity-modulated radiotherapy versus 3D conformal radiotherapy. *Anticancer Research*, 34(9):5175–5180, 2014.
143. Anna Ehinger, Per Malmstrom, Par-Ola Bendahl, Christopher W Elston, Anna-Karin Falck, Carina Forsare, Dorthe Grabau, Lisa Ryden, Olle Stal, Marten Ferno, and South and South-East Swedish Breast Cancer Groups. Histological grade provides significant prognostic information in addition to breast cancer subtypes defined according to St Gallen 2013. *Acta oncologica (Stockholm, Sweden)*, 56(1):68–74, jan 2017.
144. Wasil Jastaniah, Naglla Elimam, Razan S Alluhaibi, Alaa T Alharbi, Adil Ah Abbas, and Mohammed B Abrar. The prognostic significance of hypertension at diagnosis in children with wilms tumor. *Saudi medical journal*, 38(3):262–267, 2017.
145. Rachel S van der Post, Ingrid P Vogelaar, Peggy Manders, Lizet E van der Kolk, Annemieke Cats, Liselotte P van Hest, Rolf Sijmons, Cora M Aalfs, Margreet G E M Ausems, Encarna B Gomez Garcia, Anja Wagner, Frederik J Hes, Neeltje Arts, Arjen R Mensenkamp, J Han van Krieken, Nicoline Hoogerbrugge, and Marjolijn J L Ligtenberg. Accuracy of Hereditary Diffuse Gastric Cancer Testing Criteria and Outcomes in Patients With a Germline Mutation in CDH1. *Gastroenterology*, 149(4):897–906.e19, 2015.
146. Yu E., Stitt L., Vujovic O., Joseph K., Assouline A., Au J., Younus J., and Perera F. Prognostic factors for male breast cancer: Similarity to female counterparts. *Anticancer Research*, 33(5):2227–2232, 2013.

147. Rusthoven C.G., Koshy M., Sher D.J., Ney D.E., Gaspar L.E., Jones B.L., Karam S.D., Amini A., Ormond D.R., and Youssef A.S. Combined-modality therapy with radiation and chemotherapy for elderly patients with glioblastoma in the temozolomide era: A national cancer database analysis. *JAMA Neurology*, 73(7):821–828, 2016.
148. Laura E Hudson, Shishir K Maithel, Grant W Carlson, Monica Rizzo, Douglas R Murray, Andrea C Hestley, and Keith A Delman. 1 or 2 cm margins of excision for T2 melanomas: do they impact recurrence or survival?. *Annals of surgical oncology*, 20(1):346–351, 2013.