# Landmark Models for Optimizing the Use of Repeated Measurements of Risk Factors in Electronic Health Records to Predict Future Disease Risk

Ellie Paige, Jessica Barrett, David Stevens, Ruth H Keogh, Michael J Sweeting, Irwin Nazareth, Irene Petersen, and Angela M Wood

Correspondence to Dr Angela Wood, Department of Public Health and Primary Care, University of Cambridge, Strangeways Research Laboratory, Cambridge, CB1 8RN, UK (e-mail: amw79@medschl.cam.ac.uk)

Editorial enquiries to Dr Ellie Paige, National Centre for Epidemiology and Population Health, Research School of Population Health, The Australian National University (e-mail: ellie.paige@anu.edu.au, phone: +61 2 6125 2852)

Author affiliations: Department of Public Health and Primary Care, University of Cambridge, Cambridge, United Kingdom (Ellie Paige, Jessica Barrett, David Stevens, Michael Sweeting, and Angela Wood); National Centre for Epidemiology and Population Health, Research School of Population, the Australian National University, Canberra, Australia (Ellie Paige); MRC Biostatistics Unit, University of Cambridge, Cambridge, United Kingdom (Jessica Barrett); Department of Medical Statistics, London School of Hygiene and Tropical Medicine, London, United Kingdom (Ruth H Keogh); Institute of Epidemiology & Health, Department of Primary Care and Population Health, University College London, London, United Kingdom (Irwin Nazareth, and Irene Petersen).

ORIGINAL UNEDITED MANUSCRIPT

© The Author(s) 2018. Published by Oxford University Press on behalf of the Johns Hopkins Bloomberg School of Public Health. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.



Downloaded from https://academic.oup.com/aje/advance-article-abstract/doi/10.1093/aje/kwy018/4952104
by London School of Hygiene & Tropical Medicine user
on 02 May 2018

This work was funded by the Medical Research council (MR/K014811/1). The study funders played no role in the design, analysis or interpretation of the study. Dr Jessica Barrett is funded by a Medical Research Council fellowship (G0902100) and Medical Research Council unit programme (MC_UU_00002/5). Dr Ruth H Keogh is funded by a Medical Research Council Methodology Fellowship (MR/M014827/1).

Conflicts of interest: none declared

Running head: Repeat Risk Factors in Electronic Health Records

ORIGINAL UNEDITED MANUSCRIPT



Downloaded from https://academic.oup.com/aje/advance-article-abstract/doi/10.1093/aje/kwy018/4952104
by London School of Hygiene & Tropical Medicine user
on 02 May 2018

**ABSTRACT**

The benefits of using electronic health records for disease risk screening and personalized heathcare decisions are becoming increasingly recognized. We present a computationally feasible statistical approach to address the methodological challenges in utilizing historical repeat measures of multiple risk factors recorded in electronic health records to systematically identify patients at high risk of future disease. The approach is principally based on a two-stage dynamic landmark model. The first stage estimates current risk factor values from all available historical repeat risk factor measurements by landmark-age-specific multivariate linear mixed-effects models with correlated random-intercepts, which account for sporadically recorded repeat measures, unobserved data and measurements errors. The second stage predicts future disease risk from a sex-stratified Cox proportional hazards model, with estimated current risk factor values from the first stage. Methods are exemplified by developing and validating a dynamic 10-year cardiovascular disease risk prediction model using electronic primary care records for age, diabetes status, hypertension treatment, smoking status, systolic blood pressure, total and high-density lipoprotein cholesterol from 41,373 individuals in 10 primary care practices in England and Wales contributing to The Health Improvement Network (1997-2016). Using cross-validation, the model was well-calibrated (Brier score=0.041 [95%CI: 0.039, 0.042]) and had good discrimination (C-index=0.768 [95%CI: 0.759, 0.777]).

**KEYWORDS**

Primacy care records, electronic health records; cardiovascular disease; dynamic risk prediction; landmarking; mixed-effects.

ORIGINAL UNEDITED MANUSCRIPT



Downloaded from https://academic.oup.com/aje/advance-article-abstract/doi/10.1093/aje/kwy018/4952104
by London School of Hygiene & Tropical Medicine user
on 02 May 2018

**ABBREVIATIONS**

CI=confidence interval; CVD=cardiovascular disease; Electronic Health Records=EHRs; HDL-C=high-density lipoprotein cholesterol

Using electronic health records (EHRs) to systematically identify individuals at high risk of developing future disease outcomes has the potential to improve cost-effective health care (1), however existing risk prediction models do not fully optimize available historical data. The development of computationally feasible statistical methods for predicting future disease risk from existing EHRs presents specific methodological challenges and opportunities.

First, risk prediction models are typically developed using traditional prospective designs which define a baseline-origin from which to predict future disease risk. However, EHRs are dynamic in nature, for example in primary care records an individuals' follow-up begins at registration with a general practice until they transfer out or die. Defining arbitrary time origins for model development without allowing for the in- and out-flow of study participants over time can introduce bias (2). Second, risk prediction models typically use single measures of error-prone risk factors (e.g., blood pressure and cholesterol), but EHRs often contain risk factors measured repeatedly over time which could be utilized both for model development and for predicting future disease risk. In particular, repeated measurements can be used to predict error-free 'estimated current values' of risk factors, which may increase their predictive ability (3). Third, most risk prediction models require complete risk factor data to predict future risk. An exception in cardiovascular disease (CVD) risk prediction is the QRISK2 model (4), which has a built-in-tool to substitute missing risk factors using age- and



ORIGINAL UNEDITED MANUSCRIPT

Downloaded from https://academic.oup.com/aje/advance-article-abstract/doi/10.1093/aje/kwy018/4952104
by London School of Hygiene & Tropical Medicine user
on 02 May 2018

sex-specific population average values. Noteworthy, this substitution approach is not compatible with the multiple imputation approach used for model development of QRISK2 and has not been formally validated (5). Since EHR systems are primarily designed for patient management and administrative purposes, there can be large amounts of unobserved information on risk factors that needs to be handled appropriately and compatibly in both model development and for predicting future disease risk.

While multiple methods exist for developing risk prediction models using EHRs, a previous systematic review found that only 8% of studies modelled repeated longitudinal measures, 54% accounted for missing data, 16% appropriately accounted for censoring and loss to follow-up, and none assessed informative observations (where the clinic visit itself provides meaningful information) (6). Our aim was to establish a computationally feasible generic statistical framework that accounts for these potential advantages and biases of EHRs in the development of dynamic risk prediction models that leverage repeated measurements and handle unobserved data on routinely recorded risk factors. Our approach combines two existing methods, landmark-age models and multivariate linear mixed-effects models (2,7). A landmark-age is a reference point (e.g., 40, 45, 50,…,85 years) at which we want to make risk predictions using risk factor information collected up to that age. A series of prediction models, which we call landmark-age models, are constructed with time origin at the landmark-age and past risk factor information from eligible individuals (e.g., in our setting these are individuals who are currently registered with a general practice and at future risk of disease at the landmark-age). As such, individuals may contribute to one or more prediction models depending on their eligibility at the landmark-age reference points. Typically, landmark-age models are constructed using Cox proportional hazards models with the last observed risk factor values. We propose an extension to this, whereby we replace the last



ORIGINAL UNEDITED MANUSCRIPT

Downloaded from https://academic.oup.com/aje/advance-article-abstract/doi/10.1093/aje/kwy018/4952104
by London School of Hygiene & Tropical Medicine user
on 02 May 2018

observed values with error-free risk factor values estimated from a multivariate linear mixed-effects model using all available repeated measures of past risk factor values for each landmark-age (8). Multivariate mixed-effects models intrinsically handle unobserved data and sporadically recorded repeat measures (9) and their measurement errors (10). The approach also provides flexibility to account for the number (or rate) of clinic visits as a proxy for illness severity or health anxiety. There is a strong body of statistical evidence showing the benefits and potential applications of modelling longitudinal data using mixed-effects linear regression models (3,11-14), but this method is not often employed in the development of risk prediction models using EHRs (6). Moreover, using landmarking to model data in EHRs has been previously proposed (15), and has been combined with univariate mixed-effects modelling (16,17) but not in the context of dynamic risk prediction models. In the current study, we explore how landmarking can be combined with multivariate mixed-effects linear regression models to leverage the advantages of each method to generate dynamic risk prediction models suitable for use in EHRs. We illustrate our approach through the estimation of 10-year CVD risk using EHRs from 10 general practices in England and Wales.

## METHODS

### Data source

We used patient data from 10 randomly selected general practices that contributed data to The Health Improvement Network (18), a United Kingdom (UK) general practice database that derives data from routine administrative and clinical practice. During consultations with patients, family doctors enter data on medical symptoms and diagnoses using Read codes (19) (hierarchical classification system) while information on drug prescriptions is entered automatically into the EHRs. The Health Improvement Network captures information on: patient demographics, practice-level data, diagnoses and symptoms, specialist referrals,



ORIGINAL UNEDITED MANUSCRIPT

Downloaded from https://academic.oup.com/aje/advance-article-abstract/doi/10.1093/aje/kwy018/4952104
by London School of Hygiene & Tropical Medicine user
on 02 May 2018

laboratory testing, disease monitoring, prescribing, and death. For this study, we created code lists for the risk factors and outcomes using previously described methods (20). Code lists were reviewed by a clinician (I. Nazareth) and will be published on ClinicalCodes.org following publication.

The main outcome was newly recorded diagnoses of nonfatal or fatal CVD, where CVD was defined as with previous primary care risk scores (4) as: angina, myocardial infarction, stroke, transient ischaemic attack or major coronary surgery and revascularization. Cause of death was ascertained using Read codes.

Risk factors were selected based on those in the validated ACC/AHA Pooled Cohort Equations (21,22) and included: age, sex, diabetes status (binary, ascertained using Read codes (23)), smoking status (binary), systolic blood pressure (adjusted for hypertension treatment), total cholesterol and high density lipoprotein cholesterol (HDL-C). Once individuals had a diabetes diagnosis or a prescription for a blood pressure-lowering medication they were considered to have this condition/treatment during all follow-up. Values of systolic blood pressure, total cholesterol and HDL-C were standardized by centering on sex-specific means and dividing by the standard deviation.

**Study population**

Data was available from 1 January 1997 to 18 January 2016. Individuals entered the study from the latest of: (i) date of registration at general practice plus 6 months, (ii) date for acceptable computer usage (quality measurement defined as the year in which a general practice continuously used their computer system for recording of medical events and prescribing) (24), (iii) date for acceptable mortality reporting (date when mortality recording



ORIGINAL UNEDITED MANUSCRIPT

Downloaded from https://academic.oup.com/aje/advance-article-abstract/doi/10.1093/aje/kwy018/4952104
by London School of Hygiene & Tropical Medicine user
on 02 May 2018

reflected that of the UK general population) (25), (iv) 30th birth date, or (v) 1st January 1997. Individuals exited the study at the earliest of: (i) their first (i.e. 'incident') newly recorded CVD event; (ii) transfer out of the practice; (iii) their date of death, or (iv) 18 January 2016. The target population for whom we wanted to estimate CVD risk included individuals with genral practice records and without a history of CVD or statin prescriptions (Web Figure 1). We excluded participants with statin prescriptions as these individuals are already being treated for being at risk of developing CVD and as such would not need to be identified by a screening algorithm. In addition, the study sample excluded those with: unknown sex, study entry date after age 85, and no measurements of smoking status or systolic blood pressure or total cholesterol or HDL-C between study entry and study exit (Web Figure 1).

The following measurements were considered biologically implausible and were changed to missing for the analysis: systolic blood pressure <60 or >250 mm Hg (26); total cholesterol <1.75 or >20 mmol/liter (27); and HDL-C <0.3 or >3.1 mmol/liter (26) (n=12,352 measurements out of a total 1,675,241 were changed to missing).

The scheme for The Health Improvement Network to obtain and provide anonymous patient data was approved by the National Health Service South-East Multicenter Research Ethics Committee in 2002 and scientific approval for this study was obtained from Cegegim Strategic Data Medical Research's Scientific Review Committee (13-017). EP, AW, DS, JB and IP had full access to the data used to create the study population. This article follows RECORD reporting guidelines (Web Table 1) (28).

**Statistical analysis**

*Two-stage dynamic risk prediction model*



ORIGINAL UNEDITED MANUSCRIPT

Downloaded from https://academic.oup.com/aje/advance-article-abstract/doi/10.1093/aje/kwy018/4952104
by London School of Hygiene & Tropical Medicine user
on 02 May 2018

We used a two-stage approach to construct a dynamic risk prediction model, first modelling historical repeated risk factor measurements using multivariate mixed-effects linear models and then estimating 10-year CVD risk using Cox proportional hazards models (Figure 1). We briefly present the methods here and provide more detail in Web Appendix 1. In both stages, models were developed at landmark-ages 40, 45,…, 85 years for eligible participants defined as those (i) registered with a general practice at the landmark-age (ii) with no CVD diagnoses prior to the landmark-age and (iii) no statin prescription prior to the landmark-age. Treating each landmark-age as a time origin, past risk factor information was extracted from age 30 onwards and participants were followed up for 10 years until their first CVD event or study exit date (Figure 1). Crude incidence rates by age at study entry, sex, and statin prescription by calendar year were calculated.

*Estimation of error-free current risk factor values*

For each landmark-age and separately for males and females, we fitted multivariate mixed-effects linear regression models (9) on past repeat measurements for smoking status, systolic blood pressure, total cholesterol and HDL-C. Each model included fixed intercepts and slopes for each risk factor, a time-dependent covariate for initiation of blood pressure-lowering medications for systolic blood pressure, and correlated individual-specific random intercepts for all four risk factors. These models were estimable on individuals with at least one measurement of at least one risk factor. From each model we estimated the error-free *current risk factor values* (i.e., the predicted values at the landmark-age) using the best linear unbiased predictors from the empirical Bayes posterior distribution of the random intercepts, conditional on the past observed risk factor measurements.

*Estimating 10-year CVD risk*



ORIGINAL UNEDITED MANUSCRIPT

Downloaded from https://academic.oup.com/aje/advance-article-abstract/doi/10.1093/aje/kwy018/4952104
by London School of Hygiene & Tropical Medicine user
on 02 May 2018

Ten year CVD risk was estimated from a landmark-age Cox proportional hazards model, stratified by sex and with time since landmark-age as the underlying time variable. The model was adjusted for landmark-age and landmark-age squared, and included the risk factors: last observed diabetes status, last observed treatment for hypertension and estimated current risk factor values for smoking status, systolic blood pressure, total cholesterol and HDL-C. Participants were followed up for a maximum of ten years. Proportional hazards are therefore assumed only across a ten-year period. A 'super-landmark model' approach (7) was used with robust standard errors. A super-landmark model is a version of landmarking in which the datasets contributing to the landmark models across all landmark-ages are stacked and a single time-to-event model is fitted to the stacked dataset (Web Appendix 1).

*Assessment of predictive ability*

Performance of the 10-year CVD risk predictions were assessed with measures of calibration (i.e., calibration plots by decile of predicted risk), predictive accuracy (i.e., brier scores; an average of the squared difference between the observed outcome and predicted risk, where lower scores indicate better predictive accuracy and zero means perfect calibration) and discrimination (i.e., C-index; a measure of how well the model discriminates between those with and without CVD (29,30)). We estimated the C-index over all individuals (calculated over pairs of different individuals) and also separately at each landmark-age. The latter is estimated on subsets of individuals of the same age, thus we call this an age-adjusted C-index which naturally will have lower values to reflect poorer discrimination (31). We used ten-fold cross-validation, splitting the data by general practice, to account for over-optimism.

The above 10-year CVD risk predictions were compared against predictions from (i) a 'basic' landmark-age model, which included sex, age, last observed diabetes status and last observed



ORIGINAL UNEDITED MANUSCRIPT

Downloaded from https://academic.oup.com/aje/advance-article-abstract/doi/10.1093/aje/kwy018/4952104
by London School of Hygiene & Tropical Medicine user
on 02 May 2018

treatment for hypertension; (ii) a dynamic landmark-age model with landmark-age interactions with each covariate; (iii) a dynamic landmark-age model with last observed measurements of all risk factors instead of estimated current risk factor values; and (iv) a dynamic landmark-age model using cumulative means of all historical measurements recorded *before* each landmark-age, of smoking status, systolic blood pressure, total cholesterol and HDL-C. Predictions from (iii) and (iv) were only estimable for individuals with one or more measurements on *all* risk factors, which we call the restricted sample.

*Sensitivity analyses*

We conducted four sensitivity analyses. First instead of using all available historical repeat measurements of risk factors, we restricted the data to be within ten years before each landmark-age. Second, we adjusted the multivariate mixed-effects models by the annual rate of repeated measurements in the five years before each landmark-age (as a proxy to account for bias due to sicker or more health conscious individuals having more repeats (32)). Third, instead of estimating current risk factor values from only past information we estimated the future 10-year average risk factor levels from a multivariate mixed-effects model derived from both past and future risk factor information within the 10-year future horizon (Web Figure 2). Importantly, only past observed risk factors were subsequently used in the prediction of the future 10-year average risk factor levels for the Cox model. Fourth, since it might be useful to identify patients who are still at high absolute risk even after treatment with statins, we re-ran the main analyses including statin users in the models. The mixed-effects model including a time-dependent covariate for statin therapy initiation for total cholesterol and statin therapy at landmark-age was included as a risk factor in the Cox model.

ORIGINAL UNEDITED MANUSCRIPT



Downloaded from https://academic.oup.com/aje/advance-article-abstract/doi/10.1093/aje/kwy018/4952104
by London School of Hygiene & Tropical Medicine user
on 02 May 2018

All analyses were performed using Stata 14.2 (StataCorp) and 95% confidence intervals (95% CIs) were generated for all measures of association.

## RESULTS

*Description of the study sample*

The target population included 41,373 individuals with general practice records and without a history of CVD or statin use at study entry. Of these, 32,328 (78%) individuals had at least one measurement of smoking status, systolic blood pressure, total cholesterol or HDL-C recorded before first CVD event or statin (Web Figure 1). Mean age at study entry was 47.9 (standard deviation=13.6) years, 17,592 (54%) were men and 5,617 (17%) were prescribed statins after study entry (Table 1). Individuals generally had more repeat measures of systolic blood pressure than HDL-C (Table 1). On average, there were 1.1 years between repeated measurements of smoking status, 0.5 years between repeated measurements of systolic blood pressure, 1.1 years between repeated measurements of total cholesterol, and 1.2 years between repeated measurements of HDL-C. Overall, 2,861 participants (7%) had a newly recorded CVD event over a mean 10.4 (standard deviation=5.6) years of follow-up. Crude CVD incidence rates per 1,000 person-years increased from 2.9 for 40-44 year olds to 35.2 for 80-84 year olds, were higher in men than women, and decreased among statin users by increasing calendar year (Table 2). Participants in the study sample and restricted sample (n=12,292 (30% of target population); Web Figure 1) were similar in terms of age at study entry, sex, systolic blood pressure, and total and HDL-C levels but those in the restricted sample were more likely to have diabetes (Table 1). The study sample had more males compared to the target population but was otherwise similar (Web Table 2).

*Estimates from the landmark models*



ORIGINAL UNEDITED MANUSCRIPT

Downloaded from https://academic.oup.com/aje/advance-article-abstract/doi/10.1093/aje/kwy018/4952104
by London School of Hygiene & Tropical Medicine user
on 02 May 2018

Regression coefficients from the age- and sex-specific multivariate linear mixed-effects models and hazard ratios for the Cox models, without 10-fold cross-validation, are provided in Web Tables 3-6. Overall the values of the fixed intercepts from the multivariate mixed-effects linear models show that systolic blood pressure and total cholesterol increased over the landmark-ages, whereas HDL-C and smoking status decreased (Web Table 3). In addition, hazard ratios were generally stronger for the model using estimated current risk factor values compared to using last observed values or cumulative means (Web Table 6).

*Assessment of 10-year CVD risk*

28% of individuals had an estimated 10-year CVD risk of ≥10% and 10% had an estimated risk of ≥20% from the landmark model with estimated current risk factor values. The model appeared well-calibrated (Web Figure 3b), had Brier score of 0.041 (0.030, 0.042) and overall C-index of 0.768 (0.759, 0.777) (Figure 2b). Discrimination was better at younger ages (Figure 3). Additional age interactions did not further improve calibration or risk discrimination (Figure 2b and Web Figure 3c). The basic model (including only age, diabetes status and treatment for hypertension) also appeared well calibrated (Web Figure 3a), had Brier score of 0.041 (95% CI: 0.040, 0.043), and a lower overall C-index of 0.752 (95% CI: 0.742, 0.761) (Figure 2a). Similar to the main model, the basic model also discriminated risk better at younger compared to older ages (Web Figure 4).

Estimated 10-year CVD risk appeared slightly higher in models using last observed and cumulative mean risk factor values compared to estimated current values (Web Figure 5). Calibration, Brier scores and C-indices were similar across the landmark models with last observed, cumulative mean or estimated current risk factor values (Web Figures 6 and 7).



ORIGINAL UNEDITED MANUSCRIPT

Downloaded from https://academic.oup.com/aje/advance-article-abstract/doi/10.1093/aje/kwy018/4952104
by London School of Hygiene & Tropical Medicine user
on 02 May 2018

Risk discrimination was better at younger ages than older ages across all models (Web Figure 8).

*Sensitivity analyses*

There was no difference in risk discrimination when the model was restricted to using historical repeated measures data up to 10 years before landmark-age (C-index=0.768 [95% CI: 0.758, 0.777]) or when the estimated current risk factor values were adjusted for the rate of clinic visits (C-index=0.766 [95% CI: 0.756, 0.775]). However, we observed an increase in risk discrimination using estimated future 10-year average risk factor levels (C-index=0.774 [95% CI: 0.765, 0.783]) instead of estimated current risk factor values. C-indices were lower when statin users were included in the analysis but the patterns of risk discrimination and calibration remained the same as in the main analysis (Web Tables 7 and 8).

## DISCUSSION

We have presented a computationally feasible statistical framework for developing dynamic risk prediction models for use on EHRs with historical repeated measures of risk factors. The two-stage landmark approach combines Cox proportional hazards regression and age-specific multivariate linear mixed-effects models, which account for sporadically recorded repeat measures, unobserved data and measurements errors. We illustrated the framework for the derivation and validation of a primary care dynamic risk prediction model for 10-year CVD risk, but it has potential for wider application to other diseases and conditions and for use on other electronic patient records where repeated measurements are recorded, such as those collected in secondary care.

ORIGINAL UNEDITED MANUSCRIPT



Downloaded from https://academic.oup.com/aje/advance-article-abstract/doi/10.1093/aje/kwy018/4952104
by London School of Hygiene & Tropical Medicine user
on 02 May 2018

Our motivation was based on optimising electronic primary care data for automatically identifying high-risk individuals for full formal disease risk assessment, rather like a pre-screening tool with the potential to improve cost-effective health care. For example, several international guidelines for CVD risk assessment and management (21,33-35) recommend using a systematic strategy for prioristising people for full formal risk assessment on the basis of an estimate of their CVD risk using risk factors already recorded in EHRs. CVD risk assessment tools, such as the Framingham risk model (36) and QRISK2 (4), are now integrated into electronic primary care record systems, but are not purposefully designed for pre-screening use. QRISK2 estimates CVD risk using last observed values for the numerous risk factors, and when missing, imputes using age- and sex-specific population averages for continuous risk factors or assumes no adverse clinical indicators. Our proposed framework optimizes all available historical risk factor values, handling potential bias from spurious one-off measurements, and when missing, intrinsically imputes using all other risk factor information. Future work should formally compare such models for pre-screening use and assess their cost-effectiveness.

For illustration, we compared a basic CVD risk model using sex, age, diabetes status and treatment for hypertension against extended risk models with additional risk factors incorporated as cumulative means, last observed values or estimated current risk factor values for smoking status, systolic blood pressure, and HDL-C. Our findings showed a modest improvement in risk discrimination when including estimated current values of additional risk factors, but no difference in risk discrimination in the restricted dataset when comparing additional risk factors incorporated as last observed, cumulative means, or estimated current risk factor values. Cumulative mean risk values handle sporadically recorded repeat measurements and account for measurement errors, but they are only estimable for



ORIGINAL UNEDITED MANUSCRIPT

Downloaded from https://academic.oup.com/aje/advance-article-abstract/doi/10.1093/aje/kwy018/4952104
by London School of Hygiene & Tropical Medicine user
on 02 May 2018

individuals with at least one historical measurement on all risk factors and thus not suitable for population-wide screening. A major strength of the landmark model with estimated current values of risk factors is that it is estimable for individuals with at least one measure in any of the risk factors included in the multivariate mixed model (in our illustration this was approximately 80% of individuals).

Another strength of our landmark framework is that it is developed and internally validated using data that reflects the complexity and messiness of the EHRs that would be used to estimate disease risk for future individuals, unlike risk prediction models developed using purpose-designed cohort studies. Importantly, the assumptions made about the dynamic nature of the historical repeat measures data, unobserved risk factors and measurement errors in the model development are compatible with the assumptions required for making a risk prediction for a new individual using data from EHRs.

In our sensitivity analysis we investigated using predicted future 10-year average risk factor levels instead of estimated current values, and observed a modest improvement in risk discrimination. This suggests that future risk factor values of smoking, systolic blood pressure, total cholesterol and HDL-C are more predictive of future 10-year CVD risk than current values. A considerable limitation in this analysis is that it ignores informative censoring of individuals due to death or CVD event in the multivariate mixed-effects model, although evidence from empirical and simulation studies (11,14) suggest there is often little to be gained from more complex modelling (e.g., joint models (37)).

Other methods with which to develop risk prediction models for use on EHRs exist, including machine learning approaches such as neural networks (14,38,39), and statistical approaches



ORIGINAL UNEDITED MANUSCRIPT

Downloaded from https://academic.oup.com/aje/advance-article-abstract/doi/10.1093/aje/kwy018/4952104
by London School of Hygiene & Tropical Medicine user
on 02 May 2018

such as joint models (14). Prediction models developed using landmark and joint models for single risk factors have been previously compared (40) but not in the setting using multivariate risk factors. Joint models are more computationally burdensome than landmark models, and further development is required before they are computationally feasible for application to large EHR datasets. However, landmark models can be developed using any standard statistical software with multivariate mixed-effects models and Cox regression. The landmark-age- and sex-specific multivariate mixed-effects models can be run in parallel since the most computationally burdensome part is extracting the out-of-sample individual-specific random intercepts for estimating the current risk factor values.

Certain limitations of our proposed method remain. First, our approach assumes a multivariate normal distribution for estimated current values of continuous and binary risk factors. Such an assumption is not uncommon in statistical methodology for epidemiology (e.g., in regression calibration (10) and multiple imputation (41)), however, it would be possible to replace with a mixture of regression models with correlated latent variables (42). Second, the added distributional assumptions on the risk factors may limit transferability and implicate recalibration methods for use of the model to other populations, especially in comparison to conventional CVD prediction models. Investigating the impact of model misspecification is on our future research agenda. Third, uncertainties in the estimated current risk factor values are not accounted for in the Cox model. However, our previous work suggest that such uncertainties are often negligible relative to the estimated standard errors of the beta-coefficients in the Cox model (10). Fourth, individuals with more frequent EHRs are more likely to have health conditions or health anxiety. We attempted to account for this by adjusting the estimated current risk factor values by the annual rate of repeated measurements, although it may be plausible to additionally include this as a risk factor in the Cox model.



ORIGINAL UNEDITED MANUSCRIPT

Downloaded from https://academic.oup.com/aje/advance-article-abstract/doi/10.1093/aje/kwy018/4952104
by London School of Hygiene & Tropical Medicine user
on 02 May 2018

Fifth, for our illustration we assumed a lack of specific Read or drug codes to indicate no diagnosis or medication use and cause of death was only available for 13% of those who died, meaning the outcome of CVD is underestimated in this study. Sixth, we used the same definition of CVD events as used in CVD risk prediction models used in practice, such as QRISK2, which include 'soft' outcomes such as angina. However, while angina can be a symptom of coronary heart disease, it is not a disease itself, and the appropriateness of including it in the outcome definition of CVD risk prediction models will depend on the clinical context. Finally, despite using contemporary data, CVD screening and treatment practices have changed over time and are not accounted for in the models. These limitations are unlikely to affect our between model comparisons.

The benefits of optimizing EHRs for disease risk screening and personalized heath care decisions are becoming increasingly recognized. There is a growing need for suitable statistical methods, data analytics and machine learning approaches to address the computational and methodological challenges for the analysis of such "big data". The framework presented in this paper provides a practical, transparent and flexible solution for the development of dynamic risk prediction models for use on EHRs.

ORIGINAL UNEDITED MANUSCRIPT



Downloaded from https://academic.oup.com/aje/advance-article-abstract/doi/10.1093/aje/kwy018/4952104
by London School of Hygiene & Tropical Medicine user
on 02 May 2018

## FIGURE LEGENDS

**Figure 1: Schematic showing the landmark age approach**

The dotted line indicates historical repeat measures of smoking status, systolic blood pressure, total cholesterol and HDL cholesterol, modelled by landmark-age specific multivariate linear mixed-effects models. The diamonds show the landmark age (time of risk prediction). The arrows indicate the 10-year follow-up until CVD event or censoring, modelled by Landmark Cox model.

**Figure 2a: Calibration statistics for each risk prediction model in the study sample (n=32,328), data from The Health Improvement Network (United Kingdom, 1997-2016)**

Brier score and 95% confidence intervals (CI) for each model. Lower Brier score is interpreted as better calibration. The basic model includes: age, sex + last observed measures for diabetes status and hypertension treatment. The model with estimated current values of risk factors includes: basic model + predicted current values for smoking status, systolic blood pressure, total cholesterol and HDL. The model with age interactions includes: basic model + predicted current values for smoking status, systolic blood pressure, total cholesterol and HDL + age interactions with all risk factors.

**Figure 2b: Risk discrimination statistics for each risk prediction model in the study sample (n=32,328), data from The Health Improvement Network (United Kingdom, 1997-2016)**

C-index and 95% CI for each model. Higher C-index is interpreted as better discrimination. The basic model includes: age, sex + last observed measures for diabetes status and hypertension treatment. The model with estimated current values of risk factors includes: basic model + predicted current values for smoking status, systolic blood pressure, total



ORIGINAL UNEDITED MANUSCRIPT

Downloaded from https://academic.oup.com/aje/advance-article-abstract/doi/10.1093/aje/kwy018/4952104
by London School of Hygiene & Tropical Medicine user
on 02 May 2018

cholesterol and HDL. The model with age interactions includes: basic model + predicted current values for smoking status, systolic blood pressure, total cholesterol and HDL + age interactions with all risk factors.

**Figure 2c: Change in risk discrimination for each risk prediction model in the study sample (n=32,328), data from The Health Improvement Network (United Kingdom, 1997-2016)**

Change in C-index between the models in relation to the basic model. The basic model includes: age, sex + last observed measures for diabetes status and hypertension treatment. The model with estimated current values of risk factors includes: basic model + predicted current values for smoking status, systolic blood pressure, total cholesterol and HDL. The model with age interactions includes: basic model + predicted current values for smoking status, systolic blood pressure, total cholesterol and HDL + age interactions with all risk factors.

**Figure 3: Overall and age-adjusted C-index, data from The Health Improvement Network (United Kingdom, 1997-2016)**

ORIGINAL UNEDITED MANUSCRIPT



Downloaded from https://academic.oup.com/aje/advance-article-abstract/doi/10.1093/aje/kwy018/4952104
by London School of Hygiene & Tropical Medicine user
on 02 May 2018

**Table 1. Sample Characteristics of Participants in the Study, data from The Health Improvement Network (United Kingdom, 1997-2016)**

| | Baseline characteristics | | | | | | No. measurements per year | |
|---|---|---|---|---|---|---|---|---|
| | Study sample n=32,328 | | | Restricted sample[a] n=12,292 | | | Study sample | Restricted sample[a] |
| **Characteristics** | **No. of Persons** | **%** | **Mean (SD)** | **No. of Persons** | **%** | **Mean (SD)** | **Mean (SD)** | **Mean (SD)** |
| **Age at study entry, years** | | | 47.9 (13.6) | | | 47.5 (12.3) | | |
| **Males** | 17,592 | 54 | | 6,819 | 55 | | | |
| **History of diabetes[b]** | 3,743 | 12 | | 2,175 | 18 | | | |
| **Blood pressure-lowering medication prescriptions[b]** | 9,935 | 31 | | 4,685 | 38 | | | |
| **Statin prescriptions[b]** | 5,617 | 17 | | 2,003 | 16 | | | |
| **Current smokers[b]** | 9,453 | 31 | | 3,358 | 27 | 135.3 (21.1) | 0.6 (0.4) | 0.6 (0.4) |
| **Systolic blood pressure, mm Hg[c]** | | | 134.8 (21.0) | | | 5.4 (1.0) | 1.4 (1.4) | 1.6 (1.4) |
| **Total cholesterol, mmol/liter[c]** | | | 5.5 (1.1) | | | 1.4 (0.4) | 0.4 (0.4) | 0.5 (0.4) |
| **HDL-C, mmol/liter[c]** | | | 1.4 (0.4) | | | 135.3 (21.1) | 0.3 (0.3) | 0.4 (0.3) |

**HDL-C=high-density lipoprotein cholesterol; SD=standard deviation.**

**[a]restricted sample contains only patients with at least one measurement of each smoking status, systolic blood pressure, total cholesterol and HDL-C**

**[b]number and % calculated across follow-up period (i.e. a diagnosis of diabetes at any point during follow-up is counted as a history of diabetes for that individual)**

**[c]based on first measurements after study entry**

ORIGINAL UNEDITED MANUSCRIPT



Downloaded from https://academic.oup.com/aje/advance-article-abstract/doi/10.1093/aje/kwy018/4952104
by London School of Hygiene & Tropical Medicine user
on 02 May 2018

**Table 2. Crude CVD Incidence Rate per 1,000 Person-years by Entry Age, Sex, and Statin Prescriptions by Calendar Year for the Study Sample, data from The Health Improvement Network (United Kingdom. 1997-2016)**

| Factors | No. incident CVD cases | Total person-years | Crude incidence rate per 1,000 person-years |
|---|---|---|---|
| **Age (years) at study entry** | | | |
| 40 – 44 | 167 | 57,754 | 2.9 |
| 45 – 49 | 239 | 53,056 | 4.5 |
| 50 – 54 | 307 | 49,903 | 6.2 |
| 55 – 59 | 356 | 37,132 | 9.6 |
| 60 – 64 | 382 | 29,552 | 12.9 |
| 65 – 69 | 396 | 22,417 | 17.7 |
| 70 – 74 | 386 | 15,626 | 24.7 |
| 75 – 79 | 299 | 10,575 | 28.3 |
| 80 – 84 | 187 | 5,317 | 35.2 |
| **Sex** | | | |
| Male | 1,520 | 198,797 | 7.6 |
| Female | 1,341 | 232,166 | 5.8 |
| **Statin initiation by calendar year[a]** | | | |
| 1997 – 2001 | 225 | 4,828 | 46.6 |
| 2002 – 2006 | 968 | 38,857 | 24.9 |
| 2007 – 2011 | 687 | 46,662 | 14.7 |
| 2012 – 2016 | 365 | 27,543 | 13.3 |

**CVD=cardiovascular disease**

**[a]Index statin prescription stratified by calendar year of prescribing date**

ORIGINAL UNEDITED MANUSCRIPT



Downloaded from https://academic.oup.com/aje/advance-article-abstract/doi/10.1093/aje/kwy018/4952104
by London School of Hygiene & Tropical Medicine user
on 02 May 2018

## REFERENCES


1. Bates DW, Saria S, Ohno-Machado L, Shah A, Escobar G. Big data in health care: using analytics to identify and manage high-risk and high-cost patients. *Health affairs (Project Hope).* 2014;33(7):1123-1131.
2. Dafni U. Landmark Analysis at the 25-Year Landmark Point. *Circulation: Cardiovascular Quality and Outcomes.* 2011;4(3):363-371.
3. Paige E, Barrett J, Pennells L, et al. Repeated measurements of blood pressure and cholesterol improves cardiovascular disease risk prediction: an individual-participant-data meta-analysis. *American journal of epidemiology.* 2017;186(8):899-907.
4. Hippisley-Cox J, Coupland C, Vinogradova Y, et al. Predicting cardiovascular risk in England and Wales: prospective derivation and validation of QRISK2. *BMJ : British Medical Journal.* 2008;336(7659):1475-1482.
5. Collins GS, Altman DG. An independent and external validation of QRISK2 cardiovascular disease risk score: a prospective open cohort study. *The BMJ.* 2010;340:c2442.
6. Goldstein BA, Navar AM, Pencina MJ, Ioannidis JP. Opportunities and challenges in developing risk prediction models with electronic health records data: a systematic review. *Journal of the American Medical Informatics Association : JAMIA.* 2017;24(1):198-208.
7. van Houwelingen HC, Putter H. *Dynamic prediction in clinical survival analysis.* Florida, US: CRC Press, Taylor & Francis Group; 2012.
8. Xanthakis V, Sullivan LM, Vasan RS. Multilevel modeling versus cross-sectional analysis for assessing the longitudinal tracking of cardiovascular risk factors over time. *Stat Med.* 2013;32(28):5028-5038.
9. Laird NM, Ware JH. Random-effects models for longitudinal data. *Biometrics.* 1982;38(4):963-974.
10. Fibrinogen Studies C. Correcting for multivariate measurement error by regression calibration in meta-analyses of epidemiological studies. *Statistics in medicine.* 2009;28(7):1067-1092.
11. Sweeting MJ, Barrett JK, Thompson SG, Wood AM. The use of repeated blood pressure measures for cardiovascular risk prediction. A comparison of statistical models in the ARIC study [available online ahead of print October 11 2016]. *Statistics in Medicine.* 2017;36(28):4514-4528.
12. Singh A, Nadkarni G, Gottesman O, Ellis SB, Bottinger EP, Guttag JV. Incorporating temporal EHR data in predictive models for risk stratification of renal function deterioration. *Journal of biomedical informatics.* 2015;53:220-228.
13. Akbarov A, Williams R, Brown B, et al. A Two-stage Dynamic Model to Enable Updating of Clinical Risk Prediction from Longitudinal Health Record Data: Illustrated with Kidney Function. *Studies in health technology and informatics.* 2015;216:696-700.
14. Goldstein BA, Pomann GM, Winkelmayer WC, Pencina MJ. A comparison of risk prediction methods using repeated observations: an application to electronic health records for hemodialysis. *Stat Med.* 2017;36(17):2750-2763.
15. Wells BJ, Chagin KM, Li L, Hu B, Yu C, Kattan MW. Using the landmark method for creating prediction models in large datasets derived from electronic health records. *Health care management science.* 2015;18(1):86-92.
16. Damman K, Jaarsma T, Voors AA, Navis G, Hillege HL, van Veldhuisen DJ. Both in- and out-hospital worsening of renal function predict outcome in patients with heart failure: results from the Coordinating Study Evaluating Outcome of Advising and




ORIGINAL UNEDITED MANUSCRIPT

Downloaded from https://academic.oup.com/aje/advance-article-abstract/doi/10.1093/aje/kwy018/4952104
by London School of Hygiene & Tropical Medicine user
on 02 May 2018

Counseling in Heart Failure (COACH). *European journal of heart failure.* 2009;11(9):847-854.
17. Maziarz M, Heagerty P, Cai T, Zheng Y. On longitudinal prediction with time-to-event outcome: Comparison of modeling options. *Biometrics.* 2017;73(1):83-93.
18. In Practice Systems Ltd. The Health Improvement Network (THIN). 2016; http://www.inps.co.uk/vision/health-improvement-network-thin. Accessed 5 July 2016.
19. Chisholm J. The Read clinical classification. *BMJ (Clinical research ed.).* 1990;300(6732):1092.
20. Dave S, Petersen I. Creating medical and drug code lists to identify cases in primary care databases. *Pharmacoepidemiology and drug safety.* 2009;18(8):704-707.
21. Goff DC, Jr., Lloyd-Jones DM, Bennett G, et al. 2013 ACC/AHA guideline on the assessment of cardiovascular risk: a report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines. *Journal of the American College of Cardiology.* 2014;63(25 Pt B):2935-2959.
22. Muntner P, Colantonio LD, Cushman M, et al. Validation of the atherosclerotic cardiovascular disease Pooled Cohort risk equations. *Jama.* 2014;311(14):1406-1415.
23. Sharma M, Petersen I, Nazareth I, Coton SJ. An algorithm for identification and classification of individuals with type 1 and type 2 diabetes mellitus in a large primary care database. *Clinical Epidemiology.* 2016;8:373-380.
24. Horsfall L, Walters K, Petersen I. Identifying periods of acceptable computer usage in primary care research databases. *Pharmacoepidemiology and drug safety.* 2013;22(1):64-69.
25. Maguire A, Blak BT, Thompson M. The importance of defining periods of complete mortality reporting for research using automated data from primary care. *Pharmacoepidemiology and drug safety.* 2009;18(1):76-83.
26. Littman AJ, Boyko EJ, McDonell MB, Fihn SD. Evaluation of a Weight Management Program for Veterans. *Preventing Chronic Disease.* 2012;9:E99.
27. Hajifathalian K, Ueda P, Lu Y, et al. A novel risk score to predict cardiovascular disease risk in national populations (Globorisk): a pooled analysis of prospective cohorts and health examination surveys. *The lancet. Diabetes & endocrinology.* 2015;3(5):339-355.
28. Benchimol EI, Smeeth L, Guttmann A, et al. The REporting of studies Conducted using Observational Routinely-collected health Data (RECORD) statement. *PLoS medicine.* 2015;12(10):e1001885.
29. Lloyd-Jones DM. Cardiovascular risk prediction: basic concepts, current status, and future directions. *Circulation.* 2010;121(15):1768-1777.
30. Harrell FE, Jr., Califf RM, Pryor DB, Lee KL, Rosati RA. Evaluating the yield of medical tests. *Jama.* 1982;247(18):2543-2546.
31. White IR, Rapsomaniki E. Covariate-adjusted measures of discrimination for survival data. *Biometrical journal. Biometrische Zeitschrift.* 2015;57(4):592-613.
32. Goldstein BA, Bhavsar NA, Phelan M, Pencina MJ. Controlling for Informed Presence Bias Due to the Number of Health Encounters in an Electronic Health Record. *American journal of epidemiology.* 2016;184(11):847-855.
33. National Institute for Health and Care Excellence. *Lipid modification: Cardiovascular risk assessment and the modification of blood lipids for the primary and secondary prevention of cardiovascular disease (clinical guideline CG181).* 2014.
34. New Zealand Ministry of Health. *Cardiovascular disease risk assessment: updated 2013.* 2013.

ORIGINAL UNEDITED MANUSCRIPT



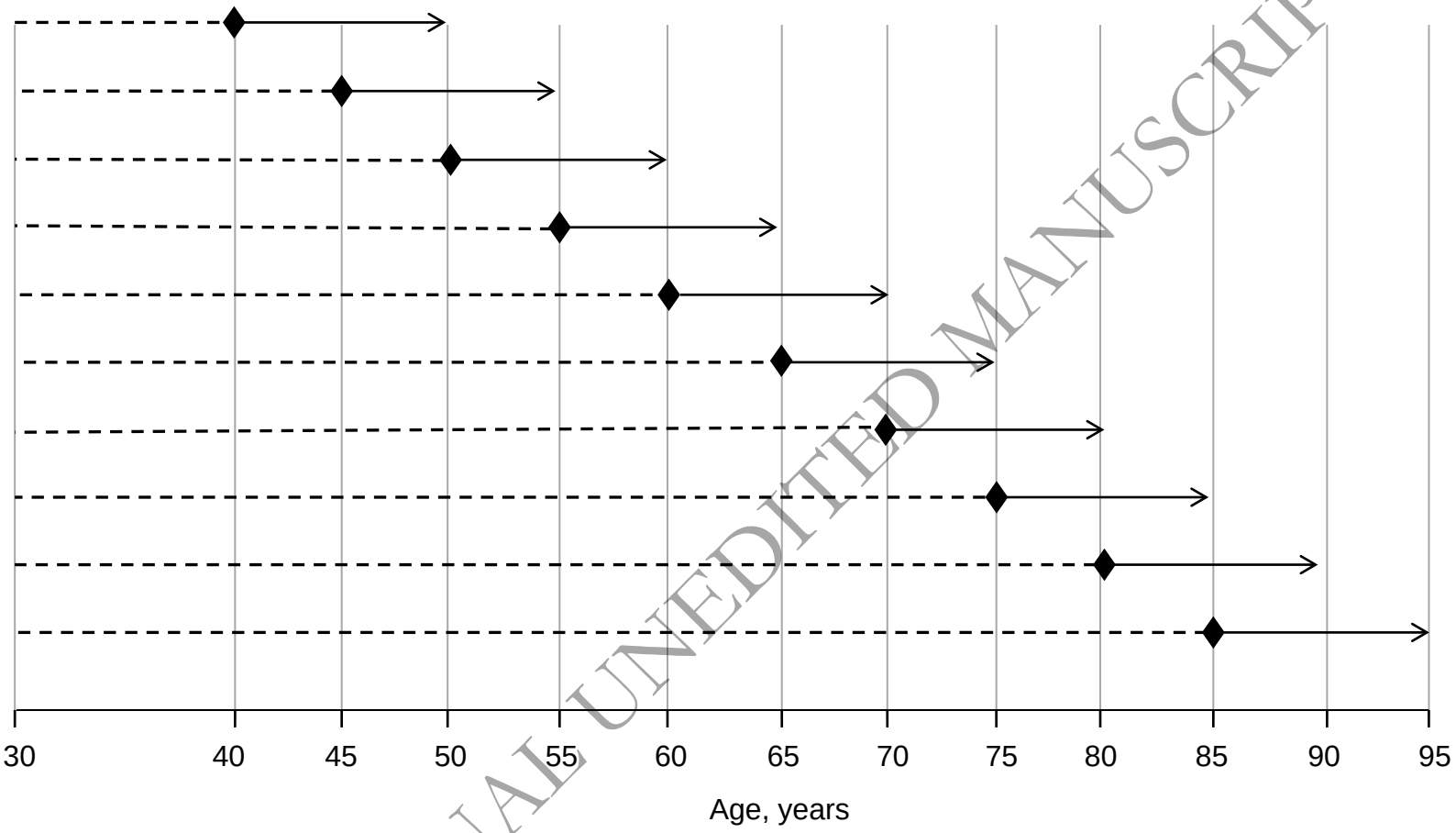
Downloaded from https://academic.oup.com/aje/advance-article-abstract/doi/10.1093/aje/kwy018/4952104
by London School of Hygiene & Tropical Medicine user
on 02 May 2018

35. Perk J, De Backer G, Gohlke H, et al. European Guidelines on cardiovascular disease prevention in clinical practice (version 2012). *European Heart Journal.* 2012;223(1):1-68.
36. D'Agostino RB, Sr., Vasan RS, Pencina MJ, et al. General cardiovascular risk profile for use in primary care: the Framingham Heart Study. *Circulation.* 2008;117(6):743-753.
37. Rizopoulos D. Dynamic predictions and prospective accuracy in joint models for longitudinal and time-to-event data. *Biometrics.* 2011;67(3):819-829.
38. Miotto R, Li L, Kidd BA, Dudley JT. Deep Patient: An Unsupervised Representation to Predict the Future of Patients from the Electronic Health Records. *Scientific reports.* 2016;6:26094.
39. Shameer K, Johnson KW, Yahi A, et al. Predictive Modeling of Hospital Readmission Rates Using Electronic Medical Record-Wide Machine Learning: A Case-Study Using Mount Sinai Heart Failure Cohort. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing.* 2016;22:276-287.
40. Suresh K, Taylor JMG, Spratt DE, Daignault S, Tsodikov A. Comparison of joint modeling and landmarking for dynamic prediction under an illness-death model. *Biometrical journal. Biometrische Zeitschrift.* 2017;59(6):1277-1300.
41. Schafer JL. *Analysis of incomplete multivariate data.* CRC press; 1997.
42. Fitzmaurice GM, Laird NM. Regression models for mixed discrete and continuous responses with potentially missing values. *Biometrics.* 1997;53(1):110-122.

ORIGINAL UNEDITED MANUSCRIPT



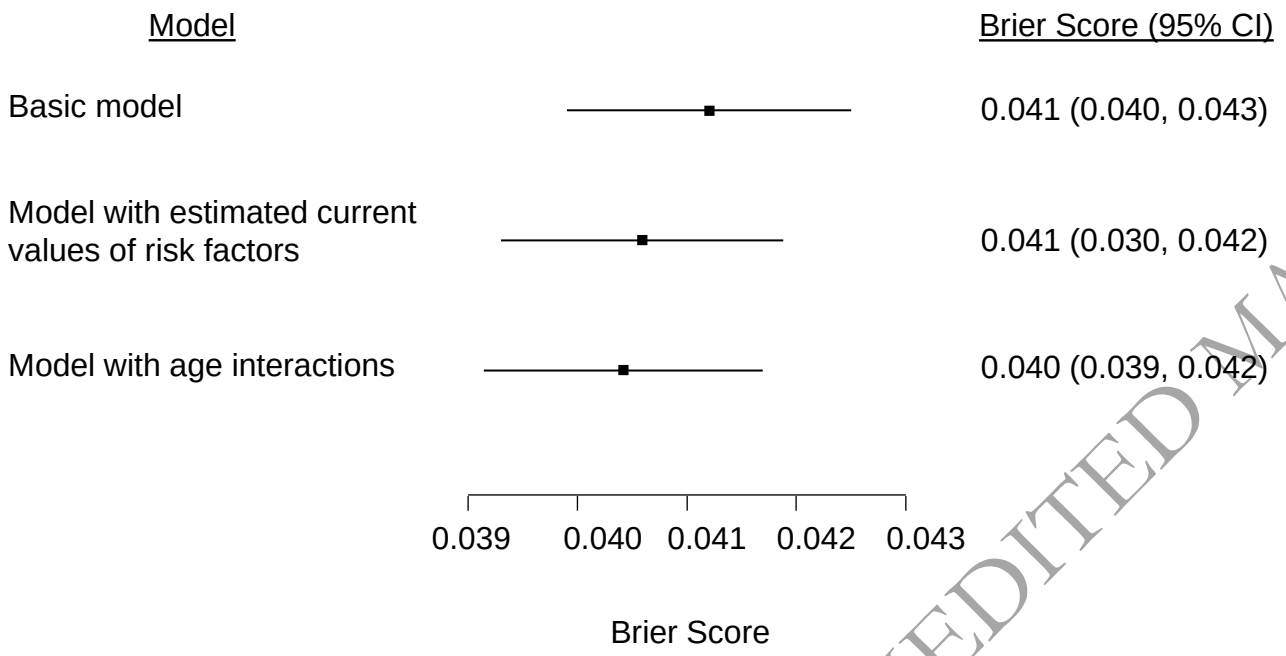
Downloaded from https://academic.oup.com/aje/advance-article-abstract/doi/10.1093/aje/kwy018/4952104
by London School of Hygiene & Tropical Medicine user
on 02 May 2018




ORIGINAL UNEDITED MANUSCRIPT

Downloaded from https://academic.oup.com/aje/advance-article-abstract/doi/10.1093/aje/kwy018/4952104
by London School of Hygiene & Tropical Medicine user
on 02 May 2018

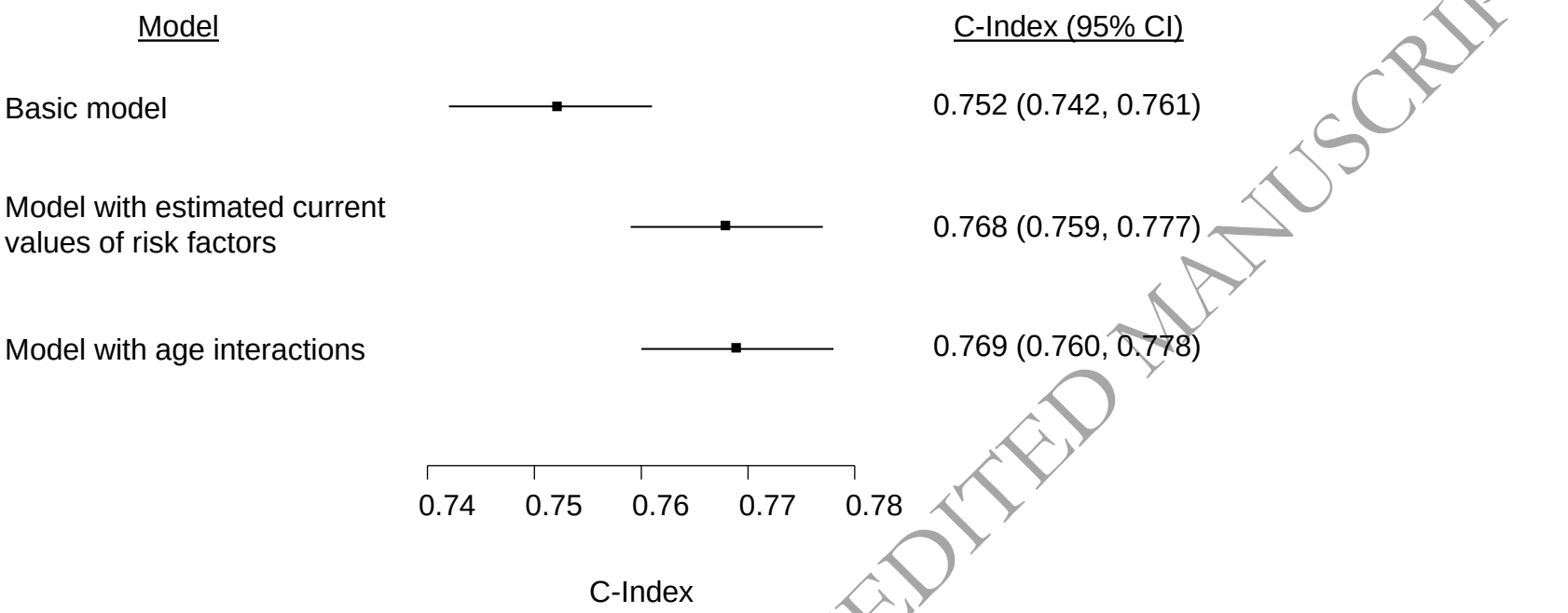
A)




ORIGINAL UNEDITED MANUSCRIPT

Downloaded from https://academic.oup.com/aje/advance-article-abstract/doi/10.1093/aje/kwy018/4952104
by London School of Hygiene & Tropical Medicine user
on 02 May 2018

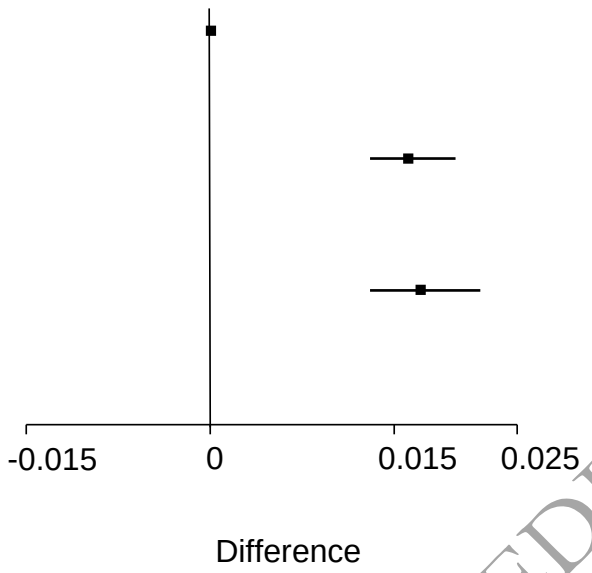
B)




ORIGINAL UNEDITED MANUSCRIPT

Downloaded from https://academic.oup.com/aje/advance-article-abstract/doi/10.1093/aje/kwy018/4952104
by London School of Hygiene & Tropical Medicine user
on 02 May 2018

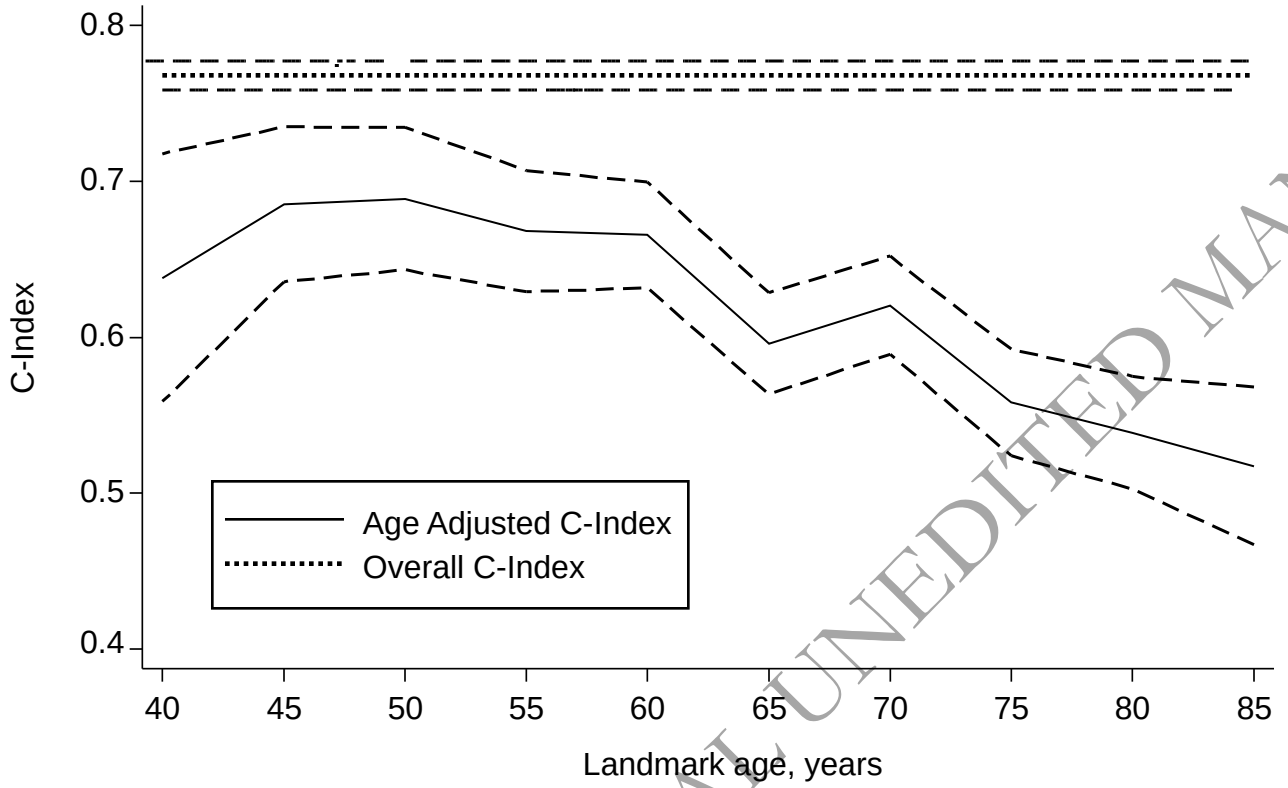
C)




| Model | Change in C-Index (95% CI) | p-value |
|---|---|---|
| Basic model | 0.000 (Ref) | |
| Model with estimated current values of risk factors | 0.016 (0.013, 0.020) | p<0.01 |
| Model with age interactions | 0.017 (0.013, 0.022) | p<0.01 |

ORIGINAL UNEDITED MANUSCRIPT

Downloaded from https://academic.oup.com/aje/advance-article-abstract/doi/10.1093/aje/kwy018/4952104
by London School of Hygiene & Tropical Medicine user
on 02 May 2018




ORIGINAL UNEDITED MANUSCRIPT

Downloaded from https://academic.oup.com/aje/advance-article-abstract/doi/10.1093/aje/kwy018/4952104
by London School of Hygiene & Tropical Medicine user
on 02 May 2018