Comparison of propensity score methods and covariate adjustment: Evaluation in four cardiovascular studies

Markus C. Elze[1,2], John Gregson[1],Usman Baber[3], Elizabeth Williamson[1], Samantha Sartori[3], Roxana Mehran[3],Melissa Nichols[4,5], Gregg W Stone[4,5], Stuart Pocock[1*]

[1]Department of Medical Statistics, London School of Hygiene and Tropical Medicine, Keppel Street, London, WC1E 7HT, UK

[2]F. Hoffmann-La Roche AG, Grenzacherstrasse 124, Basel, 4070, Switzerland

[3]Icahn School of Medicine at Mount Sinai, Gustave L. Levy Place, New York, NY 10029-5674, US

[4]Columbia University Medical Center, New York-Presbyterian Hospital, West 168th St., New York, NY 10032, US

[5]Cardiovascular Research Foundation, East 59th Street, New York, NY 10022-1202, USA

[*]Corresponding author. Tel: +44 (0)20 7927 2413, Fax: +44 (0)20 7637 2853, Email: stuart.pocock@lshtm.ac.uk

MCE is employed by F. Hoffmann-La Roche AG and owns stocks in the company. This research was conducted while MCE was employed by the London School of Hygiene and Tropical Medicine and represents independent research. All other authors confirm no conflicts of interest to report.

Main body word count (including tables, legends, footnotes): 6,351 (5,000 max)

Abstract:

Propensity scores (PS) are an increasingly popular method to adjust for confounding in observational studies. PS methods have theoretical advantages over traditional covariate adjustment, but their relative performance in real-word scenarios is poorly characterized. We used datasets from four large-scale cardiovascular observational studies (PROMETHEUS, ADAPT-DES, THIN, and CHARM) to compare the performance of traditional covariate adjustment and four commonly used PS methods: matching, stratification, inverse probability weighting and use of propensity score as a covariate. We found that stratification performed poorly with few outcome events, and inverse probability weighting gave imprecise estimates of treatment effect and undue influence to a small number of observations when substantial confounding was present. Covariate adjustment and matching performed well in all of our examples, although matching tended to give less precise estimates in some cases. PS methods are not necessarily superior to traditional covariate adjustment, and care should be taken to select the most suitable method.

Keywords: Propensity score, covariate adjustment, observational studies, bias, comparison of methods

Abbreviations: BMI=body mass index; HPR=high platelet reactivity; IPW=Inverse probability weighting; MACE=major adverse cardiovascular event; PS=propensity score; RCTs=Randomized clinical trials;

Introduction

Evaluation of therapeutic interventions generally fall into two categories, observational studies and randomized controlled trials (RCTs). The choice of treatment in observational studies may be influenced by patient characteristics; e.g. higher risk patients may be more or less likely to receive the intervention. Some of these differences are collected in standard databases, while others are not (e.g. frailty). In contrast, when studying the effect of an intervention in RCTs confounding from both measured and unmeasured variables is avoided, and RCTs are thus generally considered the highest form of scientific investigation. Nonetheless, accurate treatment effect estimates from observational databases can provide complementary value to RCTs. This is particularly true when RCTs enrol highly selected patients (yielding results not generalizable to all real world scenarios), are small (because of their greater complexity and cost), or are not feasible to conduct [1].

The traditional method to adjust for baseline differences between treatment groups in observational databases is covariate adjustment, where all relevant patient characteristics are included in a regression model relating the outcome of interest to the alternative treatments. A commonly cited concern is that such models might be over-fitted when the number of covariates is large compared to the number of patients or outcome events – a rule of thumb is to have at least 10 events per covariate included in the model.[2] However, more recent opinions favor relaxing this rule of thumb[3].

Propensity score (PS) methods are increasingly being used in observational studies of cardiovascular interventions as an alternative to traditional covariate adjustment; many such examples can be found published in JACC.[4-7] A propensity score is defined as the probability of a patient being assigned to an intervention given a set of covariates[8]. As the PS summarizes all patient characteristics into a single covariate, they reduce (although do not eliminate[9]) the potential for overfitting.PS methods aim to achieve some of the characteristics of RCTs by compensating for the fact that different patients had different probabilities to be assigned to the exposures under investigation. Their aim is to thus attenuate problems of confounding of patient characteristics and assignment to an intervention typically found in observational studies.

Popular PS methods include stratification, matching, inverse probability weighting (IPW), and using the PS as a covariate in a conventional regression model.[10-12]] However there is lack of clear guidance on how to make a sensible choice between these various PS methods or traditional covariate adjustment for any given database. We therefore applied several propensity score methods to 4 large-scale observational cardiovascular datasets, to critically examine the specific advantages and pitfalls of the different methods, and to compare their results with classical covariate adjustment.

Methods

Datasets

We analyzed data from the CHARM program[13], the ADAPT-DES study[14], the THIN study[15], and the PROMETHEUS study[16]. For each dataset, we focused on one “treatment” comparison and one outcome of prime interest. The overall goal was to produce relevant PS models across a range of different settings, so for some cases these choices differed from the primary objectives of the original publications. The terms “treatment” and “control” are used throughout to simplify the language even though one study[14] performed comparisons for platelet reactivity. All outcomes studied were time-to-event with censoring occurring at the end of planned follow-up or at the time of patient withdrawal or lost to follow-up.

The **CHARM** program[13]randomised 7,599 patients with chronic heart failure to candesartan versus placebo, with a median follow-up of 3.1 years. We herein investigate the association of treatment with beta-blockers at baseline (3,396 untreated, 4,203 treated) and all-cause death (1,831 events). That is, we utilize the CHARM program as an observational database for inference about the association of beta-blocker use and mortality risk. Our propensity score model contains cardiovascular risk factors, (age, sex, BMI, smoking, diabetes) as well as prior cardiovascular events and hospitalisations (18 variables in all).

The **ADAPT-DES** study[14] investigated the relationship between high platelet reactivity (HPR) on clopidogrel (HPR 4,930 patients, not HPR 3,650 patients) to stent thrombosis and other cardiovascular events at 12 months follow-up in a prospective, multicentre registry of patients receiving drug-eluting stents. Herein we focus on stent thrombosis (56 events). The study authors reported an adjusted hazard ratio (HR) of 2.49 for HPR vs norther. Our propensity score model will contain information about age, sex, medication, diabetes, ethnicity, smoking, renal function and other cardiovascular risk factors (39 variables in all).

The **THIN** population-based cohort study[15]compared 30,811 statin users with 60,921 patients not using statins treated by the same general practitioners (total 91,732 patients) for several outcome events including all-cause mortality (17,296 events, HR 0.79). The inclusion criteria required at least 12 months of follow-up, thus the first year must be excluded due to the so-called immortality bias. Herein we investigate the effect of statin use on all-cause mortality. The study authors reported an adjusted hazard comparing statin-users to non-users of 0.78. Previous a large RCT[17] in a similar patient population found a HR of 0.87. Our propensity score model will contain cardiovascular risk factors, age, sex, BMI, smoking, drinking, other medications and other diseases (48 variables in all).

The **PROMETHEUS** cohort study[16]compared prasugrel ("treatment") with clopidogrel ("control") for MACE outcomes (death, myocardial infarction, stroke, or unplanned revascularization) at 90 days (1,580 events) in 19,914 patients (4,017 prasugrel, 15,587 clopidogrel) using databases from 8 US hospitals. The authors reported an unadjusted HR of 0.58 and an adjusted HR using a propensity score model of 0.89. Our propensity score model will contain cardiovascular risk factors, age, sex, BMI, smoking, prior cardiovascular events, as well as details about the implanted stent and an indicator for study centre (35 variables in all).

Propensity scores – a brief overview

The PS for an individual is defined as the probability of being assigned to "treatment" given all relevant covariates[8]. The PS is typically estimated using a logistic regression model that incorporates all variables that may be related to the outcome and/or the treatment decision. All such variables should be included in the logistic model irrespective of their statistical significance or collinearity with other variables in the model. However, variables that are exclusively associated with the treatment decision, but not the outcome, should not be incorporated[18]. As in any predictive regression model, any variable collected after the treatment decision should not be used. As far as possible, covariates identified as relevant in the four original studies will be incorporated in the PS models used here. Note that all relevant variables remain in the model regardless of their statistical significance.

For each covariate, individuals with the same PS should have, on average, the same distribution of that covariate irrespective of treatment decision (“covariate balance”). This can be checked using plots of the covariate balance or several diagnostic tests.

After the PS has been calculated, there are several options for how to use them to estimate “treatment” effects. Note throughout that while PS methods strive to estimate the true “treatment effect”, the usual caveats for observational studies apply, such as the inability to include all relevant confounders (especially those unmeasured). As described below, popular PS methods include matching or stratifying observations based on the PS, inverse probability weighting (IPW) applied to each observation, or simply including the PS as an additional variable in a regression model. The more traditional covariate adjustment offers an alternative to PS techniques by simply incorporating all relevant covariates into the final model[19].

Four propensity score methods

For each dataset the goal was to estimate the “treatment” effect on a time-to-event outcome using Cox proportional hazards models. After creating the PS for each individual, there are several ways to adjust for confounding.

PS stratification splits the dataset into several strata based on the individual’s PS alone, without reference to their treatment (exposure) group. A treatment effect is then estimated within each stratum, and an overall estimated treatment effect is calculated by taking a (weighted) average across strata. Here, 5 and 10 strata with an equal number of individuals in each stratum are used. An alternative is to split the range of possible PS into equal parts, which usually results in fewer individuals in the more extreme strata. Stratification has the additional advantage that effect estimates are available for each stratum, which may reveal potential heterogeneity of “treatment” effects across strata.

PS matching tries to find one (or more) individuals with similar PS in the “treatment” and “control groups”. There are various methods to match individuals, but here we use 1:1 nearest-neighbour matching with an added constraint that the difference between the PS (“caliper width”) may be at most 0.1 to avoid pairing dissimilar individuals. We chose this method for its computational simplicity. Following matching, the treatment effect is calculated by applying either a conventional (unmatched) regression model or a matched pair analysis to the set of patients who are successfully matched.[20] We opt for an unpaired analysis here due to its greater simplicity, noting that in our examples a paired analyses gave almost identical results (Supplementary Table 1).

The matching process results in an analysis based upon only those patients who are successfully matched. Therefore, if the treatment effect varies according to patients’ characteristics and their likelihood of receiving treatment, the treatment effect estimated from this subset of patients may differ from the effect in the original study population. This issue is covered in greater detail in the discussion.

Inverse probability weighting (IPW) uses the whole dataset, but reweights individuals to increase the weights of those who received unexpected exposures. This procedure can be thought of as producing additional observations for those parts of the target population from which there were few observations. It effectively generates a pseudo-population with near perfect covariate balance between treatment groups.[12] IPW applies weights corresponding to 1/PS for patients in the “treated” cohort and 1/(1-PS) for those in the “control”cohort.PS close to 0 (for the “treated”) or 1 (for the “control”) may be problematic for IPW due to the large weight assigned to these observations. We discuss below methods to resolve this issue such as trimming or truncating large weights.

While these three PS methods aim to balance all covariates between the treatment and control groups, the more traditional **covariate adjustment** aims to control for covariate effects (confounding) using a prediction model for the outcome event (in our case a proportional hazards model for a time-to-event outcome). Care must be taken to specify the correct functional form for any covariates that may have non-linear effects. Covariate adjustment has its critics, but there is little practical evidence that it gives misleading results. For comparison, we provide the crude effect estimate as well as the covariate-adjusted effect estimate using all covariates from the PS models.

Including the propensity score as an additional covariate in the regression model represents the fourth PS method examined. Alternatively, one could have only the PS and treatment in a model of the outcome of interest.

Variations on propensity score methods

There are several variations on the four PS methods presented above. A lack of covariate balance can be compensated for by using a “doubly robust” approach. The dataset can be pre-processed by “trimming” away (removing) individuals with extreme PS. Alternatively, large IPW weights can be avoided by truncating the weights.

“Doubly robust” methods incorporate relevant covariates both in the propensity score model and the outcome regression model for the treatment effect: this can compensate for insufficient covariate balance.[21] As the name implies, this approach offers some robustness to model misspecification either in the PS or the outcome regression model. It is recommended[8, 22] when using the propensity score as a covariate to also include individual covariates in the outcome regression model. When the method is used in this way, as we do throughout this report, it is “doubly robust”. “Doubly robust” methods are also commonly used with IPW, but less frequently with matching or stratification, possibly due to the reduced sample size when using these methods. However, the doubly robust approach removes a key advantage of PS methods: having only one covariate in the final model.

Trimming can be performed after the calculation of the PS. This involves dropping the individuals with most extreme PS values in both the “treatment” and “control” groups as they may lack a match in the other group and can be predisposed to residual confounding. This can help avoid extreme weights in inverse probability weighting, improve comparability between the exposures, and remove unusual “outlying” patients for whom the expected treatment (or control) was not chosen. Typical trimming methods might remove the most extreme 1% or 5% of all observations.

Weight truncation reduces any “large” weight down to a maximum weight. There is no standard definition of a “large” weight for IPW. Here, we considered any weight above 10 to be “large” and reduced any weight greater than 10 down to this threshold. Removal of large weights is sometimes recommended for sensitivity analyses. However, complete removal of all individuals with weights larger than 10 may increase the imbalance between “treatment” and “control” groups.

Standard errors and p-values

[23]All standard errors (SE) reported here are given on the effect (i.e. log HR) scale and we used the usual sandwich variance estimator when using IPW.[23] The calculation of p-values can then be done in the usual fashion. A special case is stratification, where it is necessary to aggregate SEs and p-values from multiple models. This is done by calculating the overall variance for a particular parameter as the weighted average of the variances for that parameter from each stratum and dividing by the number of strata[24]. Assuming asymptotic normality on the overall effect, p-values can then be calculated.

Results

PSs for the CHARM, ADAPT-DES, THIN, and PROMETHEUS studies showed a range of different distributions (Fig. 1). Full PS models are given in Supplemental Tables 2-5 and for comparison, covariate-adjusted models are given in Supplemental Tables 6-9. Both CHARM and ADAPT-DES exhibited good overlap between the PS for the "treatment" and "control" groups. A single individual in the "control" group for ADAPT-DES had a PS close to zero and could be considered an outlier.

In contrast, the THIN and PROMETHEUS studies showed markedly different PS distributions for the treatment and control groups. This indicates that it may be difficult to provide valid comparisons between the two groups. THIN had a substantial number of PS close to zero or one. There were 1,134 "treated" patients (4% of all "treated") and 15,514 "control" patients (25% of all "control") with a PS less than 0.1. Conversely, there were 2,235 "treated" individuals (7% of all "treated") and 173 "control" patients (0.3% of all "control") above a PS of 0.9. Clearly, there are key variables in the PS that played an important role in who did (and did not) receive a statin. Patients where PS and chosen exposure strongly disagreed (high PS but received “control”, low PS but received “treatment”) may be atypical, but received large IPW weights.

PROMETHEUS had a very large number of PS close to zero, especially in the “control” group receiving clopidogrel (7282 “control” individuals below PS 0.1, 47% of the “control” group). This indicates that key variables in the PS had a marked influence on physician choice of clopidogrel rather than prasugrel. There were also a considerable number of patients in the “treatment” group receiving prasugrel, with a PS close to zero (330 “treated” individuals below PS 0.1, 8% of the “treated” group). These individuals may be unusual and may not offer a representative comparison with the other group.

CHARM

Results for CHARM, a non-randomised comparison of the effect of beta-blocker use versus control on all-cause death, showed excellent agreement across all PS methods and covariate adjustment (Fig. 2a). As expected, the crude estimate (first row) was different from the covariate adjusted estimate (second row) or estimates provided by the different PS methods (other rows). The adjusted HRs were all ~0.73, with 95% CIs ~0.65 to 0.81. SEs were very similar across all methods, and p-values were highly significant for all methods.

ADAPT-DES

The ADAPT-DES study, which investigated the relationship between HPR and the risk of stent thrombosis, produced similar HRs for most methods (Fig 2b). Covariate adjustment, matching, IPW and using the PS as a covariate all arrived at a HR of ~2.2 comparing HPR to not HPR. A notable exception is stratification, which showed an unstable result when using 10 strata and a wider CI with 5 strata. Otherwise, SEs and p-values were comparable for all methods, although matching had slightly poorer precision.

An investigation of the strata for ADAPT-DES reveals that the relatively low number of 56 events in the dataset was divided unevenly in the 10 strata (Supplemental Table 10). Two strata received only a single event, making precise estimation of the treatment effect within those strata impossible. These findings strongly suggest that stratification with this many strata should not be used when the number of events is sparse.

THIN

The different PS methods and covariate adjustment mostly produced similar results for the THIN study, arriving at HRs ~0.85 and a highly significant mortality reduction for those taking a statin (Fig. 2c). The exception was IPW, which estimated a smaller treatment effect with a wider CI. Trimming individuals with extreme PS from IPW gave similar results while truncating large weights in IPW brought the HR in line with the other methods. Similarly, a “doubly robust” approach of including all covariates in the final regression model also brought the HR in agreement with the other approaches. Additionally, a strong influence of confounders in this database was noted: the crude HR of 0.55 greatly exaggerated the treatment effect, due to the fact that individuals on statins tended to be at lower mortality risk. Note that from RCTs, a HR of approximately 0.87 is expected.

A plot of the IPW weights revealed very large weights for some individuals (Fig. 3a), which may be why IPW produced different results from other methods. For 1,307 patients, weights exceeded 10. These 1.4% of patients had the same total weight as the 22% of patients with the lowest weights. This may have given undue influence to very few observations, which is especially problematic considering that those large weights were given to the most unusual individuals. The majority of the large weights were given to patients on statin treatment who the PS model strongly predicted would be controls (i.e. not taking a statin).

PROMETHEUS

Results for the PROMETHEUS study (Fig. 2d), comparing prasugrel versus clopidogrel for risk of MACE events, showed substantial disagreement between the methods, although the results were non-significant for almost all methods. Covariate adjustment, stratification, IPW, and using the PS as a covariate all produced HRs of ~0.94. Matching showed a lower HR of ~0.85. IPW without any modification had a much higher SE than other methods. Investigating the IPW weight distribution revealed very large weights for 8% (330 individuals) of the “treatment” group (Fig. 3b), which may explain the stark change in HR seen when truncating large weights. The crude estimate of treatment (HR 0.59) is attributable to the marked confounding present; i.e. patients chosen to receive prasugrel tended to be at lower risk of MACE events.

Further examination showed that covariate balance is insufficient for some methods. Figure 4 compares the covariate balance for matching, stratification, and IPW using the absolute standardised difference between the “treatment” and “control” groups. Without use of PS methods, covariate balance was insufficient for almost all variables. Matching produced excellent balance for all variables. Stratification mostly achieved satisfactory covariate balance except for previous PCI, with age, hypertension, and prior CHF as borderline cases. IPW showed very poor covariate balance for previous PCI and poor or borderline balance for hypertension, previous MI, and prior PAD. Due to the lack of covariate balance, the results for stratification and IPW may be considered unreliable.

Effect of trimming and truncation of IPW weights

We used PS trimming in the THIN and PROMETHEUS studies to attempt to reduce the impact of large weights in IPW. However, for both studies even 5% trimming was not sufficient to fully remove all large IPW weights from the datasets. Consequently, the standard errors for the estimated treatment effect remained large relative to other methods after trimming, particularly in PROMETHEUS (Fig 3d). We additionally applied 1% and 5% trimming and compared findings for each PS method and covariate adjustment on the trimmed datasets. However, trimming did little to reconcile differences in the estimates produced (Supplemental Figures S1-S4). Finally, we truncated large weights in the THIN and PROMETHEUS studies to a maximum of 10. In both examples, this resulted in a large reduction in the standard error and an estimated HR much closer to the crude estimate (Figs. 3c-d). However, this brought the estimates closer to the other methods in THIN (Fig. 3c) whereas it took estimates further from other methods in PROMETHEUS (Fig. 3d).

Discussion

For observational cohort studies that compare alternative treatments (and other exposures), it has become standard practice to use propensity score (PS) methods to correct for selection biases and potential confounding when examining the relative risks (hazards) of event outcomes. While the principles of PS methods are clear, there exists a diversity of alternative approaches (e.g. propensity matching, stratification, inverse probability weighting (IPW) with or without trimming) alongside the more traditional method of covariate adjustment. Although there is a substantial methodologic literature on PS approaches with some limited guidance on which specific methods may be preferable[10, 11], there is no general agreement as to the choice of PS method that is best suited to any particular scenario. Thus, researchers may choose a suboptimal method that preserves more bias and/or imprecision than is necessity.

To provide insight to this common problem, we have here undertaken an in-depth assessment of many of the available PS and covariate-adjustment approaches as applied to four large-scale cardiovascular studies. The present analysis illustrates the challenges faced in determining which methods actually produce the most valid results in different settings.

Our first example, the CHARM study examining the impact of beta-blocker use at baseline on mortality in heart failure patients, is the most straightforward. The PS distributions for the 17 chosen baseline variables showed considerable overlap between the two groups with no extreme values. In addition the study was large, with the two groups being of similar size. The results showed a consistency across all PS methods and also covariate adjustment. Note the crude estimate produced an exaggerated treatment effect, indicating the importance of taking confounding into account by using any of these methods. However, the extent of confounding is less extreme than in several of the other studies.

The next example, ADAPT-DES comparing the risk of stent thrombosis in ACS patients with and without HPR, has some methodologic similarities to CHARM, but also the complication of having fewer outcome events (only 56 stent thromboses). Here, PS stratification performed badly, with too few events per stratum. PS matching and IPW showed good agreement, although the former had less precise estimates due to not using all patients in the matched analysis. Surprisingly, covariate adjustment with 39 covariates and only 56 events held up well, producing very similar estimates to IPW.

The last two examples, the THIN study comparing the mortality of individuals on and off statins, and the PROMETHEUS study comparing prasugrel versus clopidogrel for the risk of MACE events in ACS patients, both presented more of a challenge in choosing a robust PS method. The reason in both cases was the marked separation of PS probability distributions between the two groups: statin versus no statin and prasugrel versus clopidogrel respectively. In particular, there were a substantial number of PSs close to 0 and 1 in THIN and close to 0 in PROMETHEUS. As a consequence, IPW included more than a few very influential individuals with very large weights in the IPW analysis. This in turn led to imprecise estimates of treatment effect and a worrying lack of covariate balance for some potentially important confounders. Additionally, in both these examples IPW analyses estimated HRs closer to the null.

In both examples, the use of IPW in a “doubly robust” fashion (i.e. also including all covariates in the final analysis) induced compatibility with other methods, but did not reduce the SE thereby leaving the 95% CIs unduly wide. The use of trimming (e.g. removing the 5% of individuals with the most extreme PS) was somewhat helpful, but the imprecision of estimates remained greater than for other methods. The use of PS stratification also has its problems when there is marked selection bias, as seen in THIN and PROMETHEUS. This is because using only 5 (or 10) strata does not wholly correct for covariate imbalance.

What can we learn from these experiences in order to make recommendations for the future use of PS methods and covariate adjustment? As in all studies, the primary analysis strategy should be pre-specified in advance. Post hoc selection of a preferred method after data exploration introduces bias and should only be considered for exploratory or sensitivity analysis.

One useful approach is to examine the baseline covariates prior to accessing any outcome data in order to determine which propensity score method (or covariate adjustment) may be most suitable given the characteristics of the PS, such as the degree of overlap in PS between treatment and control groups. Even so, relying on one method of analysis (which may have its flaws) may be too restrictive and it is wise to pre-define a number of secondary sensitivity analyses using alternative approaches. This enables one to determine if there is a consistency of the findings regarding the estimated treatment "effect", which if present instills confidence in the primary results.

But for any specific study what should be chosen as the primary analysis method? We see no single "right answer" to meet all circumstances, but the following insights should help in making the choice:

1) PS matching appears to be a reliable method in that it provides excellent covariate balance in most circumstances. It has the advantage of being simple to analyse, present and interpret. Its main disadvantage is that some individuals end up not matched and hence excluded from the analysis, resulting in a loss of both precision and generalizability. In our examples, up to 60% of patients were excluded by matching, although it should be noted that some of these patients could have been successfully matched by using more sophisticated matching algorithms. Finally, whatever the choice of matching algorithm, it is important to pre-define the precise algorithm to be used.

2) PS stratification tends to work well when covariate imbalance is not very marked. It has the merit of keeping all individuals in the analysis and also provides the opportunity to explore potential interactions between treatment and the PS on outcome risk. Stratification

tends to perform less well in datasets with few outcomes, particularly when the number of strata is large. When choosing the number of strata, one needs to trade-off the need for accurate control of confounding with the requirement of having a sufficient number of events in each stratum. Previous research shows that 5 strata may reduce confounding bias by up to 90%, so a modest number of strata should suffice in studies with few outcomes and/or only moderate confounding bias.[25] However, in studies with many outcome events using up to 10 strata will further reduce confounding bias, which may be important if covariate imbalance is marked.[26, 27] Beyond these recommendations, further research is needed to determine the best strategy to define the number and size distribution of strata. Are equal size strata preferred, or is it better to have larger numbers in the middle of the PS distribution, therefore enabling a more detailed exploration of the tails?

3) Inverse probability weighting (IPW) offers a conceptually simple method that is easy to implement in practice and retains all study participants. Some have advocated it as a preferred method[28, 29] However, when there is marked covariate imbalance, PS scores close to the extreme probabilities of 0 and 1occur, with some individuals ending up with very large weights. This seems intuitively inappropriate since these influential data points occur in individuals who represent a small proportion in their chosen treatment group. In our two examples with such extreme weights, THIN and PROMETHEUS, use of IPW produced less precise estimates than other methods and notable covariate imbalance.

Trimming has been recommended as an appropriate way of limiting the influence of heavily-weighted individuals. The difficulty here is to define in advance what level of trimming is desirable: should we exclude just 1% or up to 5% of extreme weights? In our examples with major covariate imbalance, trimming increased the precision of our estimates but did not alter the estimated associations. Truncating large weights resulted in more precise estimates and had large effects on the estimated associations. In both our examples, the estimated associations moved closer to the crude HR following truncation, perhaps suggesting that this method may lead to inadequate adjustment for covariate imbalance. Given the difficulty of limiting the influence of heavily weighted individuals, IPW may be best confined to datasets for which extremes of the PS distribution do not occur, such as in CHARM and ADAPT-DES, although this will generally not be known in advance of examining the data distribution.

Using a “doubly robust” IPW approach, where covariates are also included in the outcome regression model, appeared to produce results similar to traditional coverage adjustment, but with notably wider CIs. They also add a level of complexity to the analysis, and the inclusion of covariates in the outcome regression model removes a key advantage of the propensity score methods. They may therefore be unattractive as a primary method of analysis, and be best reserved for sensitivity explorations.

4) Covariate adjustment is the traditional method for correcting for covariate imbalance, selection bias and potential confounding, and existed long before PS methods were developed. Recently, some have argued that PS methods may be more robust or offer more complete adjustment for confounders.[28] However, whilst there are theoretical grounds on which to favor propensity score adjustment, we see little practical evidence to justify such negative claims[30]and in our examples, covariate adjustment provided reliable and statistically efficient estimates. One issue for datasets with few event outcomes is that the number of covariates considered for inclusion in the model may be limited, whereas many

more covariates may be included in a PS model without raising concerns of over-fitting or lack of model convergence. However, in ADAPT-DES including 39 covariates with only 56 events still produced reliable results. This example demonstrates that having fewer than 10 events per covariate does not necessary preclude using covariate adjustment[7], although it does not allay concerns of over-fitting in all similar scenarios. A further advantage of covariate adjustment is that it provides a predictive model (including treatment) for the risk (hazard) of the event outcome, which gives insight as to which covariates have the strongest influence on risk. Perhaps it is time that old-fashioned covariate adjustment deserves a revival in its use. Finally, adding the propensity score as an additional covariate produced results very similar to covariate adjustment, with similar estimates and standard errors across all examples.

Our study has limitations. Firstly, with just four datasets explored in depth, caution is needed in drawing any generalizable conclusions. This is particularly the case for small studies which are not examined here, although it should be noted that ADAPT-DES is small in terms of the number of outcome events included. Despite these limitations, we feel that the diversity of our examples facilitate a practical debate based on real experiences, which is better than relying on purely theoretical arguments. Secondly, our study assumes throughout that the effect of treatment on outcome does not differ by the likelihood that a patient is treated. When treatment effects do differ, as can be detected by comparing estimated HRs across strata of the PS, some PS methods will produce results that are systematically different from covariate adjustment even when both methods provide adequate adjustment for confounders.[31] This is because certain PS methods estimate the treatment effects relating to certain sections of the study population, such as only the treated or only the control patients. In these scenarios, investigators need to select an appropriate method to estimate the treatment effect in the set of patients in whom they most want to understand the impact of treatment: usually this is either the treated patients, the control patients, or the entire study population.[11] In addition a further technical detail is that IPW and PS matching (using an unpaired analyses) both estimate a marginal treatment effect, whereas multivariate regression, stratification and doubly robust methods all estimate a conditional HR.

In conclusion, in the present detailed examination of alternative PS methods and covariate adjustment in several topical cardiovascular studies, covariate adjustment and matching performed well in all of our examples, although matching tended to give less precise estimates in some cases. PS methods are not necessarily superior to traditional covariate adjustment, and care should be taken to select the most suitable method. We hope these insights will guide others to make wise choices in their use of PS methods, and to rekindle interest in old-fashioned covariate adjustment, which may be viewed as a suitable primary analysis method in many cases.

References

[1] Rothwell PM. External validity of randomised controlled trials: "to whom do the results of this trial apply?". Lancet 2005;365:82–93.

[2] Harrell F Jr, Lee KL, Califf RM, Pryor DB, Rosati RA. Regression modelling strategies for improved prognostic prediction. Stat Med 1984;3:143–152.

[3] Vittinghoff E, McCulloch CE. Relaxing the rule of ten events per variable in logistic and Cox regression. Am J Epidemiol 2007;165:710–718.

[4] Park DW, Seung KB, Kim YH, et al. Long-term safety and efficacy of stenting versus coronary artery bypass grafting for unprotected left main coronary artery disease: 5-year results from the MAIN-COMPARE (Revascularization for Unprotected Left Main Coronary Artery Stenosis: Comparison of Percutaneous Coronary Angioplasty Versus Surgical Revascularization) registry. J Am Coll Cardiol 2010;56:117–124.

[5] Ramos R, Garcá-Gil M, Comas-Cuf M, et al. Statins for Prevention of Cardiovascular Events in a Low-Risk Population With Low Ánkle Brachial Index. J Am Coll Cardiol 2016;67:630–640.

[6] Tamburino C, Barbanti M, D’Errigo P, et al. 1-Year Outcomes After Transfemoral Transcatheter or Surgical Aortic Valve Replacement: Results From the Italian OBSERVANT Study. J Am Coll Cardiol 2015;66:804–812.

[7] Solomon MD, Go AS, Shilane D, et al. Comparative effectiveness of clopidogrel in medically managed patients with unstable angina and non-ST-segment elevation myocardial infarction. J Am Coll Cardiol 2014;63:2249–2257.

[8] Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. Biometrika 1983;70:41–55.

[9] Senn S, Graf E, Caputo A. Stratification for the propensity score compared with linear regression techniques to assess the effect of treatment or exposure. Stat Med 2007;26:5529–5544.

[10] Austin PC. An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies. Multivariate Behav Res 2011;46:399–424.

[11] Williamson E, Morley R, Lucas A, Carpenter J. Propensity scores: from naive enthusiasm to intuitive understanding. Stat Methods Med Res 2012;21:273–293.

[12] Heinze G, Jüni P. An overview of the objectives of and the approaches to propensity score analyses. Eur Heart J 2011;32:1704–1708.

[13] Swedberg K, Pfeffer M, Granger C, et al. Candesartan in heart failure–assessment of reduction in mortality and morbidity (CHARM): rationale and design. Charm-Programme Investigators. J Card Fail 1999;5:276–282.

[14] Stone GW, Witzenbichler B, Weisz G, et al. Platelet reactivity and clinical outcomes after coronary artery implantation of drug-eluting stents (ADAPT-DES): a prospective multicentre registry study. Lancet 2013;382:614–623.

[15] Smeeth L, Douglas I, Hall AJ, Hubbard R, Evans S. Effect of statins on a wide range of health outcomes: a cohort study validated by comparison with randomized trials. Br J Clin Pharmacol 2009;67:99–109.

[16] Wayangankar SA, Baber U, Poddar K, et al. Predictors of 1 year net adverse cardiovascular events (NACE) among ACS patients undergoing PCI with clopidogrel or prasugrel: analysis from the PROMETHEUS registry. Journal of the American College of Cardiology 2016;67:562–562.

[17] Group HPSC. MRC/BHF Heart Protection Study of cholesterol lowering with simvastatin in 20,536 high-risk individuals: a randomised placebo-controlled trial. Lancet 2002;360:7–22.

[18] Myers JA, Rassen JA, Gagne JJ, et al. Effects of adjusting for instrumental variables on bias and precision of effect estimates. Am J Epidemiol 2011;174:1213–1222.

[19] Fisher R. *The Design of Experiments. 9th ed.* London: Macmillan 1971.

[20] Stuart EA. Developing practical recommendations for the use of propensity scores: discussion of 'A critical appraisal of propensity score matching in the medical literature between 1996 and 2003' by Peter Austin, Statistics in Medicine. Stat Med 2008;27:2062–5; discussion 2066–9.

[21] Kang JD, Schafer JL. Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. Statistical science 2007;523–539.

[22] D'Agostino R Jr. Propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. Stat Med 1998;17:2265–2281.

[23] Joffe MM, Ten Have TR, Feldman HI, Kimmel SE. Model selection, confounder control, and marginal structural models. The American Statistician 2012;.58: 272-279

[24] Mosteller F TJ. *Data Analysis and Regression: A Second Course in Statistics. 1st ed.* Pearson 1977.

[25] Rosenbaum PR, Rubin DB. Reducing bias in observational studies using subclassification on the propensity score. Journal of the American statistical Association 1984;79:516–524.

[26] Cochran WG. The effectiveness of adjustment by subclassification in removing bias in observational studies. Biometrics 1968;24:295–313.

[27] Hullsiek KH, Louis TA. Propensity score modeling strategies for the causal analysis of observational data. Biostatistics 2002;3:179–193.

[28] Austin PC. The relative ability of different propensity score methods to balance measured covariates between treated and untreated subjects in observational studies. Med Decis Making 2009;29:661–677.

[29] Lunceford JK, Davidian M. Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. Stat Med 2004;23:2937–2960.

[30] Glynn RJ, Schneeweiss S, Stürmer T. Indications for propensity scores and review of their use in pharmacoepidemiology. Basic Clin Pharmacol Toxicol 2006;98:253–259.

[31] Sturmer T, Rothman KJ, Glynn RJ. Insights into different results from different causal contrasts in the presence of effect-measure modification. Pharmacoepidemiol Drug Saf 2006;15:698–709.

Figure Legends

Central Illustration. An overview of the pros and cons of covariate adjustment and various propensity score methods.

Figure 1. Overview of the propensity score distribution for the “control” (blue) and “treatment” (red) exposures for: a) CHARM; b) ADAPT-DES; c) THIN; and d) PROMETHEUS.

Figure 2. **Comparison of hazard ratios from different PS methods and covariate adjustment for a) CHARM; b) ADAPT-DES; c) THIN; d) PROMETHEUS.**

In the plot, covariate adjustment is used as basis for the comparison (dashed line). Colors are used if results for the other methods differ by more than 5%.

Figure 3. Distribution of the weights for IPW in: a) THIN; b) PROMETHEUS.

To facilitate display, the vertical axis is on a logarithmic scale. There are no patients with extreme weights in the clopidogrel group, hence all patients treated with appear in the bar with the smallest inverse probability weight.

Figure 4.Comparison of the extent of covariate imbalance in PROMETHEUS using: a) crude comparisons; b) propensity matching c) propensity stratification; d) IPW. Graphs show the absolute standardized difference between treatment and control; values <0.1 are conventionally considered acceptable.

Central illustration

Covariate adjustment

+ provides a prognostic model for outcome of interest
– may not be suitable with many covariates in smaller studies

Propensity score methods

Stratification

+ provides effect estimates for every stratum
– may not fully account for strong confounding

Inverse probability weighting

+ creates a pseudo population with perfect covariate balance
– can be unstable when extreme weights occur

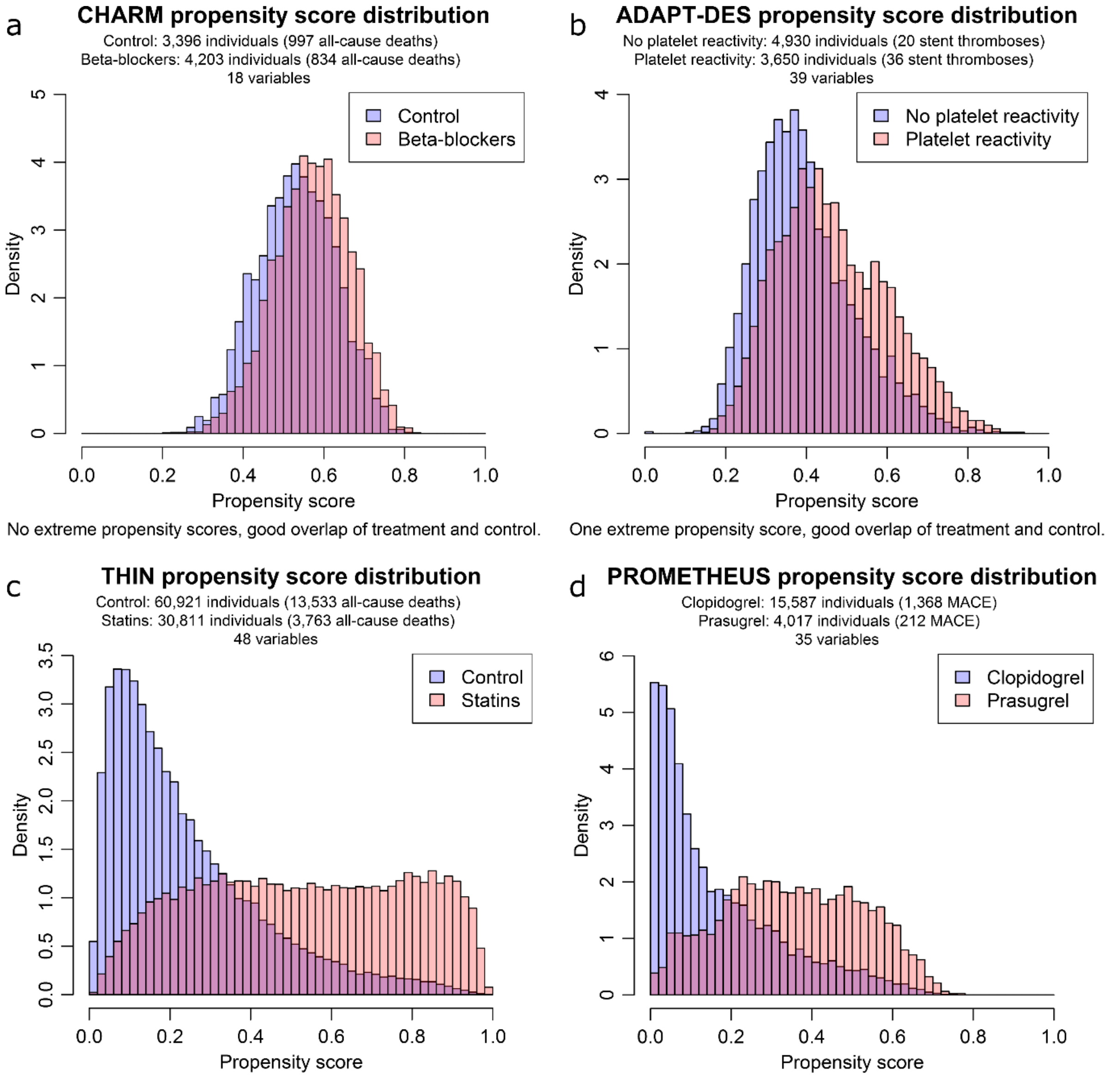
Matching

+ simple presentation of results
– loss of information for unmatched patients

Conclusions

- Covariate adjustment works well in many cases
- Sensitivity analyses can help show consistency of findings across alternative propensity score method
- Unmeasured confounding can still cause bias with any of these methods

Figure 1



a CHARM propensity score distribution

Control: 3,396 individuals (997 all-cause deaths)
Beta-blockers: 4,203 individuals (834 all-cause deaths)
18 variables

No extreme propensity scores, good overlap of treatment and control.

b ADAPT-DES propensity score distribution

No platelet reactivity: 4,930 individuals (20 stent thromboses)
Platelet reactivity: 3,650 individuals (36 stent thromboses)
39 variables

One extreme propensity score, good overlap of treatment and control.

c THIN propensity score distribution

Control: 60,921 individuals (13,533 all-cause deaths)
Statins: 30,811 individuals (3,763 all-cause deaths)
48 variables

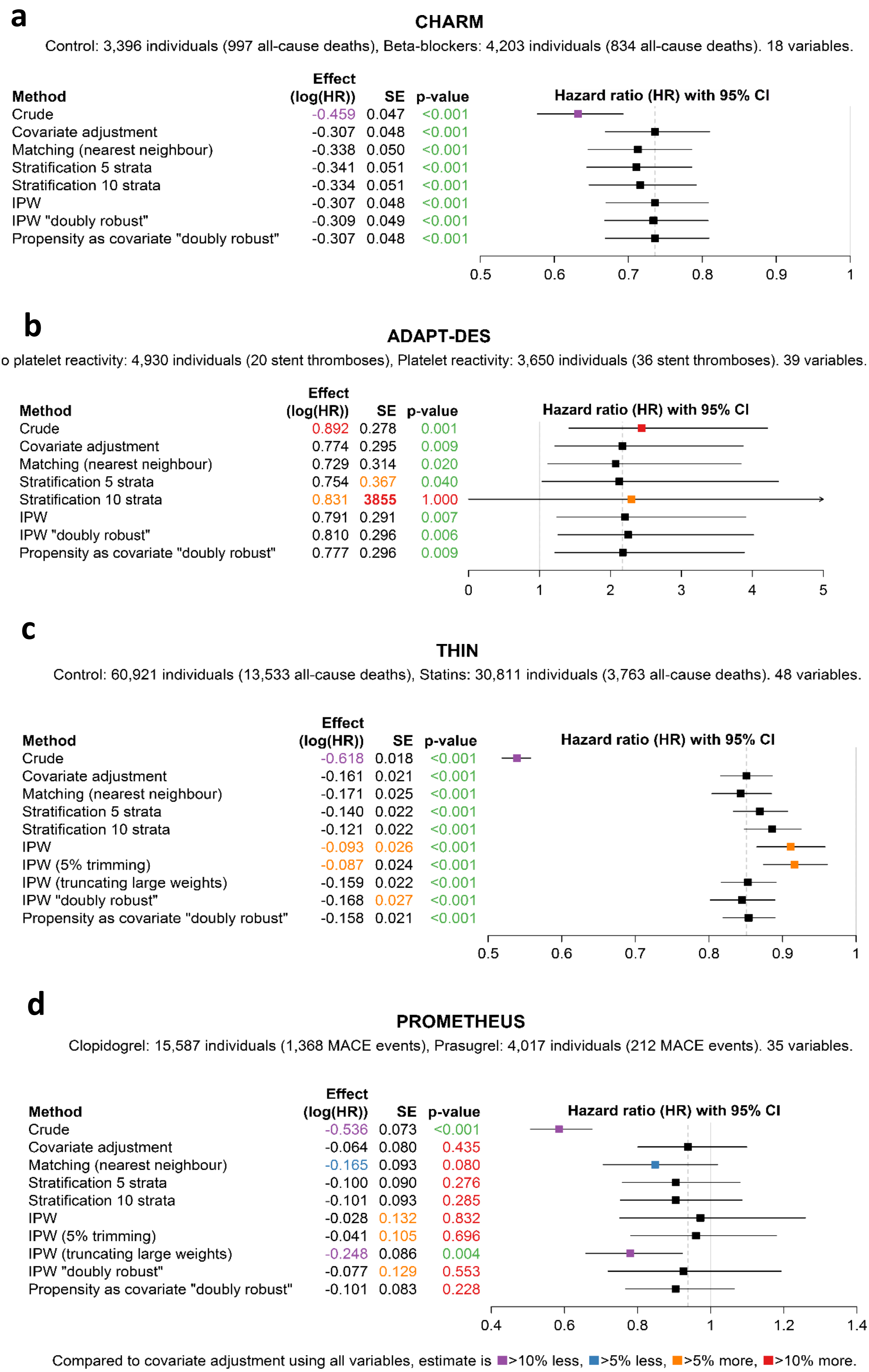
Some extreme propensity scores, poor overlap of treatment and control.

d PROMETHEUS propensity score distribution

Clopidogrel: 15,587 individuals (1,368 MACE)
Prasugrel: 4,017 individuals (212 MACE)
35 variables

Many extreme propensity scores, poor overlap of treatment and control.

Figure 2



a

CHARM

Control: 3,396 individuals (997 all-cause deaths), Beta-blockers: 4,203 individuals (834 all-cause deaths). 18 variables.

Method	Effect (log(HR))	SE	p-value
Crude	-0.459	0.047	<0.001
Covariate adjustment	-0.307	0.048	<0.001
Matching (nearest neighbour)	-0.338	0.050	<0.001
Stratification 5 strata	-0.341	0.051	<0.001
Stratification 10 strata	-0.334	0.051	<0.001
IPW	-0.307	0.048	<0.001
IPW "doubly robust"	-0.309	0.049	<0.001
Propensity as covariate "doubly robust"	-0.307	0.048	<0.001

b

ADAPT-DES

No platelet reactivity: 4,930 individuals (20 stent thromboses), Platelet reactivity: 3,650 individuals (36 stent thromboses). 39 variables.

Method	Effect (log(HR))	SE	p-value
Crude	0.892	0.278	0.001
Covariate adjustment	0.774	0.295	0.009
Matching (nearest neighbour)	0.729	0.314	0.020
Stratification 5 strata	0.754	0.367	0.040
Stratification 10 strata	0.831	**3855**	1.000
IPW	0.791	0.291	0.007
IPW "doubly robust"	0.810	0.296	0.006
Propensity as covariate "doubly robust"	0.777	0.296	0.009

c

THIN

Control: 60,921 individuals (13,533 all-cause deaths), Statins: 30,811 individuals (3,763 all-cause deaths). 48 variables.

Method	Effect (log(HR))	SE	p-value
Crude	-0.618	0.018	<0.001
Covariate adjustment	-0.161	0.021	<0.001
Matching (nearest neighbour)	-0.171	0.025	<0.001
Stratification 5 strata	-0.140	0.022	<0.001
Stratification 10 strata	-0.121	0.022	<0.001
IPW	-0.093	0.026	<0.001
IPW (5% trimming)	-0.087	0.024	<0.001
IPW (truncating large weights)	-0.159	0.022	<0.001
IPW "doubly robust"	-0.168	0.027	<0.001
Propensity as covariate "doubly robust"	-0.158	0.021	<0.001

d

PROMETHEUS

Clopidogrel: 15,587 individuals (1,368 MACE events), Prasugrel: 4,017 individuals (212 MACE events). 35 variables.

Method	Effect (log(HR))	SE	p-value
Crude	-0.536	0.073	<0.001
Covariate adjustment	-0.064	0.080	0.435
Matching (nearest neighbour)	-0.165	0.093	0.080
Stratification 5 strata	-0.100	0.090	0.276
Stratification 10 strata	-0.101	0.093	0.285
IPW	-0.028	0.132	0.832
IPW (5% trimming)	-0.041	0.105	0.696
IPW (truncating large weights)	-0.248	0.086	0.004
IPW "doubly robust"	-0.077	0.129	0.553
Propensity as covariate "doubly robust"	-0.101	0.083	0.228

Compared to covariate adjustment using all variables, estimate is 🟪>10% less, 🟦>5% less, 🟧>5% more, 🟥>10% more.

Figure 3

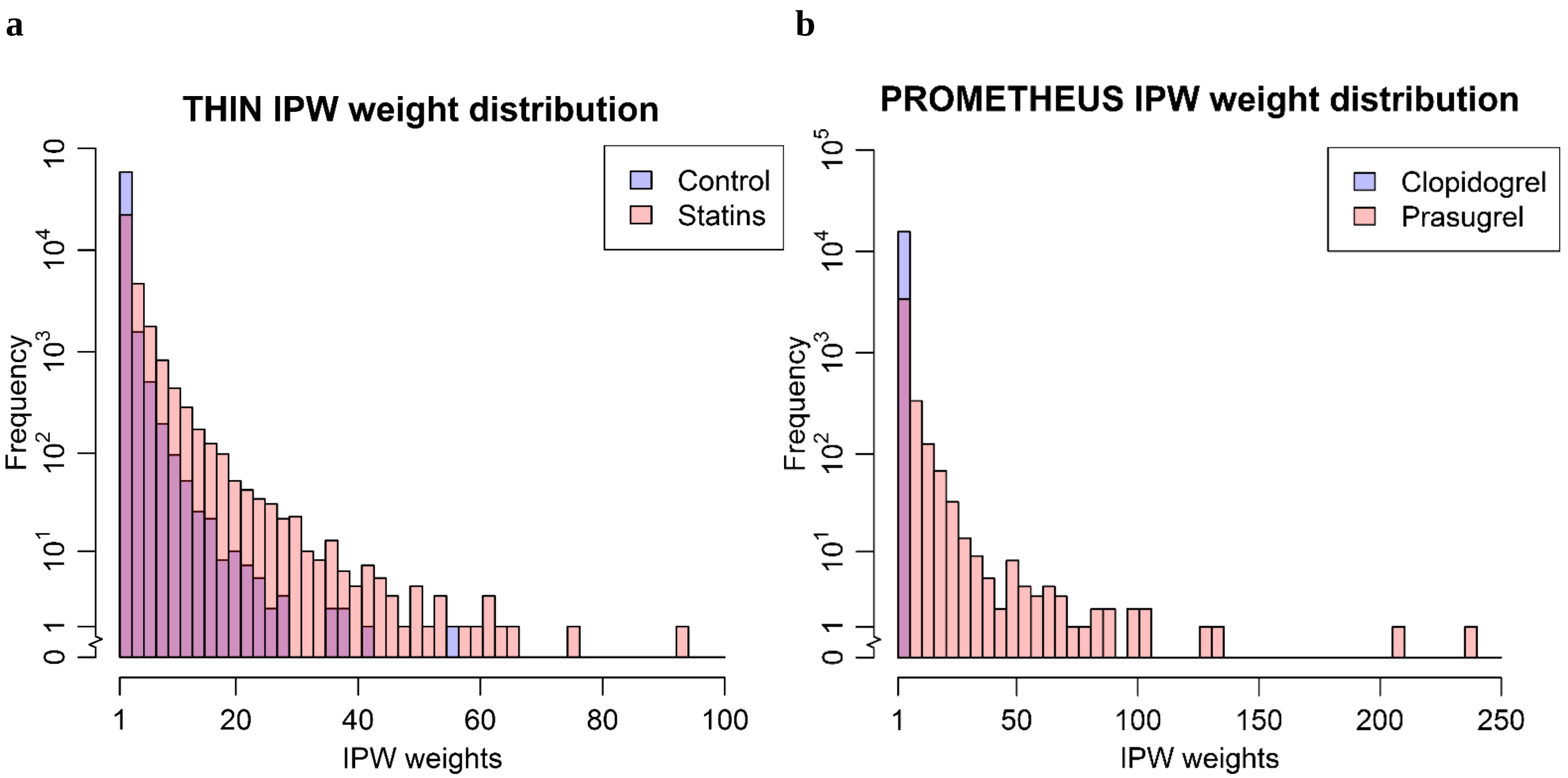
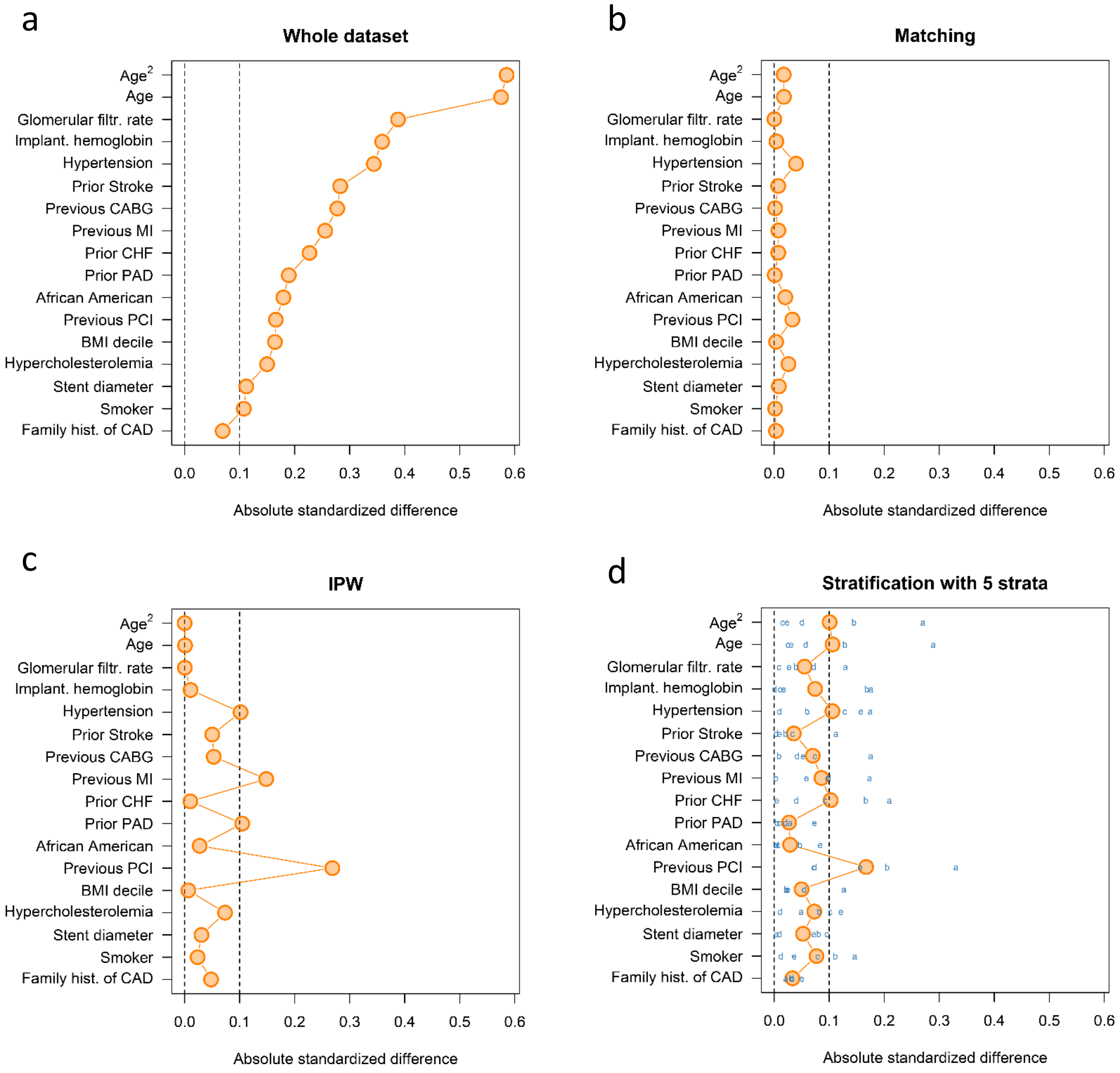

a
b
THIN IPW weight distribution
Control
Statins
Frequency
10
10^4
10^3
10^2
10^1
1
0
1
20
40
60
80
100
IPW weights
PROMETHEUS IPW weight distribution
Clopidogrel
Prasugrel
10^5
10^4
10^3
10^2
10^1
1
0
1
50
100
150
200
250
IPW weights

Figure 4



a
Whole dataset
b
Matching
c
IPW
d
Stratification with 5 strata
Age^2
Age
Glomerular filtr. rate
Implant. hemoglobin
Hypertension
Prior Stroke
Previous CABG
Previous MI
Prior CHF
Prior PAD
African American
Previous PCI
BMI decile
Hypercholesterolemia
Stent diameter
Smoker
Family hist. of CAD
0.0
0.1
0.2
0.3
0.4
0.5
0.6
Absolute standardized difference