2019

Exploration of structural and statistical biases in the application of propensity score matching to pharmacoepidemiologic data

BOSTON UNIVERSITY

SCHOOL OF PUBLIC HEALTH

Dissertation

EXPLORATION OF STRUCTURAL AND STATISTICAL BIASES IN THE APPLICATION OF PROPENSITY SCORE MATCHING TO PHARMACOEPIDEMIOLOGIC DATA

by

JOHN EDWARD RIPOLLONE

B.S., Messiah College, 2008
MPH, S.U.N.Y., Downstate Medical Center, School of Public Health, 2012

Submitted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

2019

Approved by

First Reader ____
Kenneth J. Rothman, DrPH
Professor of Epidemiology

Second Reader ____
Krista F. Huybrechts, PhD
Associate Professor of Medicine
Harvard Medical School

Adjunct Assistant Professor of Epidemiology
Boston University, School of Public Health

Third Reader ____
Jessica M. Franklin, PhD
Assistant Professor of Medicine
Harvard Medical School

Fourth Reader ____
Ryan E. Ferguson, DSc
Director, Cooperative Studies Program Coordinating Center
U.S. Department of Veterans Affairs

Research Assistant Professor of Medicine
Boston University, School of Medicine

Adjunct Clinical Assistant Professor of Epidemiology
Boston University, School of Public Health

DEDICATION

Dedicated to two groups.

First, to all family members (notably, Dad, Sarah, Nonna, Ashley and Marlene) and friends who were around during these past 5 years. They were tough years, but this group made it work for me. Also, to Stanley and Sullivan – two cats who kept me company throughout my work day.

Second, to my previous academic mentors, especially my advisors at Messiah College (notably, Dr. Marlin Eby, Dr. Michael Shin, and Dr. Lawrence Mylin) and at S.U.N.Y., Downstate Medical Center, School of Public Health (notably, Dr. Carl Rosenberg). My academic mentors provided the background and support, without which I could not have completed this dissertation, let alone begin my career.

ACKNOWLEDGMENTS

I thank the Boston University, School of Public Health for providing the scholarship that covered my tuition during my doctoral studies. I thank Dr. Kenneth Rothman for agreeing to act as my academic advisor as well as my dissertation committee chair. I thank Dr. Krista Huybrechts and Dr. Jessica Franklin for agreeing to work with me to generate the ideas that led to this dissertation. I thank my dissertation committee (Dr. Kenneth Rothman, Dr. Krista Huybrechts, Dr. Jessica Franklin and Dr. Ryan Ferguson) for supporting my dissertation work, as well as Dr. Matthew Fox for agreeing to act as the external reader of my dissertation. I thank Dr. Matthew Fox, Dr. Gheorghe Doros and Dr. Robert Lew for introducing me to Bayesian methods for epidemiology. Finally, I thank the Division of Pharmacoepidemiology and Pharmacoeconomics at Brigham and Women’s Hospital and Harvard Medical school for providing the data that were used to complete this dissertation as well as the funds that were necessary to publish a portion of this dissertation.

EXPLORATION OF STRUCTURAL AND STATISTICAL BIASES IN THE APPLICATION OF PROPENSITY SCORE MATCHING TO PHARMACOEPIDMEIOLOGIC DATA

JOHN EDWARD RIPOLLONE

Boston University School of Public Health, 2019

Major Professor: Kenneth J. Rothman, DrPH, Professor of Epidemiology

ABSTRACT

Certain pitfalls associated with propensity score matching have come to light, recently. The extent to which these pitfalls might threaten validity and precision in pharmacoepidemiologic research, for which propensity score matching often is used, is uncertain. We evaluated the "propensity score matching paradox" – the tendency for covariate imbalance to increase in a propensity score-matched dataset upon continuous pruning of matched sets – as well as the utility of coarsened exact matching, a technique that has been posed as a preferable alternative to propensity score matching, especially in light of the "propensity score matching paradox". We show that the "propensity score matching paradox" may not threaten causal inference that is based on propensity score matching in typical pharmacoepidemiologic settings to the extent predicted by previous research. Moreover, even though coarsened exact matching substantially improves covariate balance, it may not be optimal in typical pharmacoepidemiologic settings due to the extreme loss of study size (and resulting increase in bias and variance) that may be required to build the matched dataset. Finally, we explain variability in 1:1 propensity score matching without replacement as well as methods that were developed to account for this variability, with application of these methods to an example claims-based study.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER ONE: IMPLICATIONS OF THE PROPENSITY SCORE MATCHING PARADOX IN PHARMACOEPIDEMIOLOGY[1]

INTRODUCTION

Propensity score matching (PSM) is a popular method to control for differences in propensity score distributions in observational research [Pearl, 2010; Hade and Lu, 2014; Wu et al., 2015]. Other methods, notably stratification by propensity score, may be preferable with respect to overall efficiency, but PSM remains popular, perhaps owing to its reduction of the matching process to one dimension [Rosenbaum and Rubin, 1983; D'Agostino, 1998; Pearl, 2010; Desai et al., 2017]. With PSM, index units are matched to reference units with similar propensity score values, even though their underlying covariate profiles might be dissimilar. Even with this underlying dissimilarity, the distributions of observed covariates should be similar, on average, between index and reference units, conditional on the propensity score [Rosenbaum and Rubin, 1983; Austin, 2011]. From a practical perspective, PSM is easily understood among researchers and is easily implemented with available algorithms [Rassen et al., 2012].

King and Nielsen recently argued that PSM should be avoided because of the potential for the “PSM paradox” to degrade causal inference [King and Nielsen, 2016]. The paradox, in brief, is: for datasets that already are well-balanced on measured covariates, pruning of matched sets with the largest propensity score distances between the index and

[1] Ripollone JE, et al. (2018) – *American Journal of Epidemiology*

reference units may lead to increased imbalance in the underlying covariate distributions between exposure groups and, thus, to increased bias in the effect estimate.

Because King and Nielsen demonstrated the paradox in datasets with fewer covariates and with better initial covariate balance than what typically is encountered in pharmacoepidemiology, the practical effect of the paradox in pharmacoepidemiologic analyses is not clear.

Here, we present a description of the paradox and the results of an analysis of the impact of the paradox in pharmacoepidemiologic applications using insurance claims data. We used methods similar to those used by King and Nielsen in order to track levels of imbalance produced by progressive pruning of matched pairs from datasets in which, initially, all index units are matched. We varied a number of key parameters in the matching process, generating multiple matched datasets. Our intent was to evaluate the practical implications of the theoretical findings of King and Nielsen.

THE PSM PARADOX

The standard approach to 1:1 PSM for a dichotomous exposure is: (1) generate propensity scores corresponding to the estimated probability of receiving the index exposure, conditional on observed covariates, for every unit in a dataset (commonly via logistic regression); (2) match a reference unit to each index unit via some algorithm (e.g., nearest neighbor matching); (3) prune from the resulting dataset the matched pairs with the largest propensity score distances in order to eliminate poorly-matched units and to ensure balanced propensity score distributions (usually via application of a caliper as part of step 2); (4) compare (usually at the univariate level) pre- and post-matched covariate distributions to assess the improvement in covariate balance due to PSM; (5) estimate the effect parameter of interest in the matched dataset [Pan and Bai, 2015]. The key benefit of matching on the propensity score is the dimension reduction that allows for efficient matching on a scalar summary of a potentially large vector of covariates.

Let **X** be the vector of observed covariates that inform the propensity score model. PSM guarantees balance among the matched sets on the conditional probability of exposure, Pr(Exposure|**X**), but it guarantees balance on **X** only asymptotically [Mielke and Berry, 2007; Iacus et al., 2011]. With asymptotic balance, any pruning of matched sets from the resulting dataset is expected to be random with respect to underlying covariate balance. The reduction in study size resulting from random pruning could, by chance, increase the underlying **X** distance between matched units. Thus, although the intent of pruning propensity score-matched sets is to increase covariate balance, this process could have the opposite effect. By extension, with better covariate balance prior to any matching or pruning, it becomes more likely that balance will begin to deteriorate after only a few prunings. If the same procedure of pruning the worst-matched units is applied in the context of matching on the actual components of **X**, rather than on the scalar propensity score, an increase in imbalance is not expected because distances between the *original* covariate values inform the matching and pruning decisions [Greevy et al., 2004; King and L., 2007; Hill, 2008; Imai et al., 2008; Imai et al., 2009].

We present a simple example of this phenomenon using only two covariates in Table 1. In this population of 12, 4 are exposed to the index exposure and 8 are exposed to the reference exposure. The distributions of sex and race in this population are perfectly balanced between the two exposure groups. The propensity score for *every* unit is Pr(Index Exposure|Sex, Race) = 1/3. If 1:1 PSM without replacement is performed, there should be no algorithmic preference to match any reference unit to any index unit, since all 12 units have the same propensity score value. There are 70 possible selections of 4 reference units from the pool of 8 reference units to build the matched cohort consisting of 8 total units. Only 16 of those selections will retain perfect covariate balance in the sex-race distribution. Thus, we expect that 77% of the time, covariate balance will be *worse* after the initial pruning of units via PSM, compared with the balance in the pre-matched dataset. This phenomenon occurs even though the distribution of propensity scores will be perfectly balanced in any matched dataset. If either of these two covariates is related to outcome, we expect the covariate imbalance to correspond to bias in the treatment effect estimate.

Unlike our example dataset, the typical pharmacoepidemiologic claims dataset, which comprises a large number of patients and a large number of potential confounders of an association between a drug and health outcome (e.g., corresponding to concomitant medications and comorbidities), is not well-balanced on **X** before matching [Petri and

Urquhart, 1991; Patorno et al., 2013; Patorno et al., 2014]. Thus, we expected to observe a notable improvement in balance after PSM long before pruning could worsen balance.

METHODS

Description of Datasets

Two retrospective cohorts were used in these analyses. The first was a cohort of 49,919 low-income Medicare beneficiaries, at least 65 years of age, who were enrolled in the Pharmaceutical Assistance Contract for the Elderly (PACE) database in New Jersey over the years 1999-2002 and who initiated non-selective NSAIDs or selective COX-2 inhibitors [Brookhart et al., 2006; Schneeweiss et al., 2006]. The PACE cohort was generated to perform an analysis of the effect of selective COX-2 inhibitors, compared with non-selective NSAIDs, on the risk of gastrointestinal complications. Approximately 60% of patients represented in this cohort were selective COX-2 inhibitor initiators. Approximately 2,000 cases of gastrointestinal complication were observed in this cohort.

The second cohort comprised information on 886,996 completed pregnancies and was generated from the Medicaid Analytic eXtract (MAX) over the years 2000-2007 [Huybrechts et al., 2014; Bateman et al., 2015; Desai et al., 2017]. The MAX cohort was used to perform an analysis of the effect of statin use during the first trimester of pregnancy, compared with no use during the first trimester of pregnancy, on the risk of congenital malformation in the infant. Statin use was defined as the existence of at least one claim for a dispensed statin within the first trimester. Approximately 0.13% of women represented in this cohort filled a statin prescription during the first trimester. Approximately 30,000 congenital malformations were observed in this cohort.

Creation of Matched Datasets

We created multiple 1:1-matched datasets using propensity scores generated via logistic regression. In order to relax distributional assumptions for the propensity score models, all continuous variables were categorized. The propensity score models based on PACE predicted the probability of exposure to non-selective NSAIDs (since there were fewer non-selective NSAID initiators than selective COX-2 inhibitor initiators), while the propensity score models based on MAX predicted the probability of exposure to statins. Each matched dataset represented a different manipulation of (1) the richness of the covariate set informing the propensity score model, (2) the prevalence of index exposure in the pre-matched dataset and (3) the matching algorithm.

Covariate Set Richness

To assess whether increasing the number of covariates in the propensity score model decreases the number of prunings required for covariate imbalance to increase, we used three PACE-based covariate sets. The first covariate set, “Small”, comprised 19 covariates that were selected based on clinical importance. The second and third covariate sets (“Standard” and “Large”, respectively) comprised additional covariates (representing concomitant medications, comorbidities and other medical encounters) selected by a high-dimensional propensity score (HDPS) algorithm [Schneeweiss et al., 2009], in addition to the 19 pre-determined covariates. The 50 covariates with the highest bias-based HDPS ranks were included in the “Standard” covariate set, and the 100 covariates with the highest bias-based HDPS ranks were included in the “Large” covariate set. All models generated from MAX were based on one covariate set comprising 20 categorical covariates, which were selected based on clinical importance.

Prevalence of Index Exposure in the Pre-matched Dataset

To determine how the size of the fully-matched dataset affects covariate balance during matched set pruning, the index exposure prevalence values of PACE and MAX were varied, via simple random sampling with replacement, but the original dataset sizes were retained. Matched datasets were generated from PACE, separately for each of the three covariate set scenarios, using the original index exposure prevalence, 50% of the original index exposure prevalence and 20% of the original index exposure prevalence. Matched datasets were generated from MAX using the original index exposure prevalence, 400% of the original index exposure prevalence, and 700% of the original index exposure prevalence.

Matching Algorithm

Since the matching quality may depend on the matching algorithm, we used two 1:1 PSM algorithms that have been used in previous pharmacoepidemiologic analyses: a variation of nearest neighbor matching (NNM) and a variation of Parson’s digit-based greedy matching (DGM) [Rassen et al., 2012]. While the former algorithm attempts to minimize the overall propensity score distance among matched sets, the latter algorithm matches units on decreasing levels of precision, up to the fifth digit of the propensity score, without consideration of overall distance.

Because King and Nielsen referred to Mahalanobis distance matching (MDM) as a potentially better option than PSM for maintaining covariate balance after matching, we also implemented MDM [King et al., 2011; King and Nielsen, 2016]. Like the propensity score, the Mahalanobis distance is a scalar summary of the original covariate space. However, unlike the propensity score, it is a direct representation of distance between units in the actual covariate space, and has the following form:

$$\sqrt{}[(\mathbf{X}_i\text{-}\mathbf{X}_j)'\underline{\Sigma}^{-1}(\mathbf{X}_i\text{-}\mathbf{X}_j)],$$

where i indexes the exposed unit, j indexes the unexposed unit, **X** is the vector of covariates for a given unit and $\underline{\Sigma}$ is the sample covariance matrix of the original data [Iacus et al., 2011]. We selected a nearest neighbor matching algorithm to implement MDM given the popularity of this algorithm for MDM [Ho et al., 2007; Ho et al., 2011].

We constructed 12 unique datasets (9 PACE datasets, 3 MAX datasets) and 36 unique matching scenarios for our analysis. Our manipulation strategy is summarized in Figure 1.

Pruning and Assessment of Imbalance

For each fully-matched dataset, matched pairs were ranked in order of decreasing absolute propensity score distance or Mahalanobis distance and the matched pair with the largest distance was pruned from the dataset. Covariate balance was assessed for the remaining dataset, then the matched pair with the largest distance in the remaining dataset was pruned and covariate balance was assessed again. This process was repeated until only a single matched pair was left in the dataset.

We used two metrics to summarize covariate imbalance: the Mahalanobis balance and the c-statistic. The Mahalanobis balance is a type of Mahalanobis distance that represents the extent of covariate balance in the actual covariate space, and has the following form:

$$\sqrt{}[(\bar{\mathbf{X}}_{T1}-\bar{\mathbf{X}}_{T0})'\underline{\Sigma}^{-1}(\bar{\mathbf{X}}_{T1}-\bar{\mathbf{X}}_{T0})],$$

where $\bar{\mathbf{X}}_{Tk}$ is the vector of covariate means in exposure group k, and $\underline{\Sigma}$ is the sample covariance matrix of the original data [Gu and Rosenbaum, 1993; Franklin et al., 2014]. Higher Mahalanobis balance values indicate worse covariate balance. We used the c-statistic to determine changes in the discriminatory power of the logistic model predicting index exposure in the matched dataset [Harrell et al., 1982; Harrell et al., 1984]. Balance on the covariates in the matched dataset should lead to poor ability of the corresponding logistic model to determine which units are exposed (i.e., c-statistics near 0.5) [Franklin et al., 2014]. Thus, higher c-statistic values (greater than 0.5) indicate worse covariate balance.

The points in the pruning process at which three absolute propensity score distance calipers were achieved were marked both for the NNM and DGM scenarios. We selected our calipers from the common range, [0.01, 0.05] [Oakes and Kaufman, 2017]. We focused on a 0.05 caliper and then applied the more conservative calipers of 0.025 and 0.01 in order to determine whether the further loss of matched sets would correspond to increased covariate imbalance. Each caliper criterion was satisfied when the maximum propensity score distance between two units of a matched pair in a pruned dataset was less than the caliper value.

Tracking Changes in the Effect Estimate

We calculated and plotted a point estimate of effect after each pruning. For PACE, we calculated the RR estimate corresponding to the effect of non-selective NSAIDs, compared with COX-2 inhibitors, on the risk of gastrointestinal complications. For MAX, we calculated the RR estimate corresponding to the effect of statin use during the first trimester of pregnancy, compared with no use during the first trimester of pregnancy, on the risk of congenital malformation. Our goal in generating these graphs was to depict the pattern describing how the paradox might lead to bias in the effect estimate.

RESULTS

We display example covariate distributions for the pre-matched dataset and for the fully-matched datasets for PACE and MAX in Tables 2 and 3, respectively. These tables indicate that covariate balance in the pre-matched dataset was far worse for MAX than for PACE. In both datasets, covariate balance improved after the creation of the fully-matched dataset. For PACE, improvement was more marked for NNM and DGM than for

MDM (Table 2). For MAX, the opposite was true (Table 3). We also analyzed standardized differences and drew the same conclusions (Figures 2 and 3) [Austin, 2009].

We display all Mahalanobis balance metric trend graphs for PACE and MAX in Figures 4 and 5, respectively. The c-statistic metric trend graphs were similar and are displayed in Figures 6 and 7 for PACE and MAX, respectively. We also present zoomed-in versions of Figures 4 and 5 in Figures 8 and 9, respectively.

In each panel of Figures 4 and 5, the fully-matched datasets produced by NNM and by DGM had much better covariate balance than the corresponding pre-matched dataset, although this was not always the case for MDM – in one case, balance actually was worse for MDM in the fully-matched dataset (Figure 4, panel G). Moreover, the points at which the caliper criteria were met always were near the lowest regions of the NNM and DGM trend lines. These results indicate that if a typical caliper on the absolute propensity score scale in the range, [0.01, 0.05] had been required after NNM or DGM, before performing inference on these data, the covariate balance in the corresponding pruned dataset always would have been near optimal (at least, measured by the Mahalanobis balance). However, even though NNM and DGM always greatly improved covariate balance with respect to the pre-matched datasets after only a few prunings, covariate imbalance did eventually increase after further pruning in certain cases.

Covariate Set Richness

For the PACE NNM- and DGM-matched datasets, for a given index exposure prevalence, fewer prunings were required for covariate imbalance to increase as the number of covariates used to construct the corresponding propensity score model increased (Figure 4). This result is demonstrated by the fact that the imbalance trends increased more quickly during the pruning process as the number of covariates increased, or by the fact that the Mahalanobis balance value of the fully-matched dataset increased as the number of covariates increased, or both. A similar trend occurred for the PACE MDM-matched datasets. As the number of covariates used to perform MDM increased, the Mahalanobis balance value of the fully-matched dataset increased. Finally, increasing the number of covariates used to construct the propensity score model generally increased the number of prunings required to achieve the caliper criteria (Figure 8).

Prevalence of Index Exposure in the Pre-matched Dataset

No consistently strong trends in imbalance across index exposure prevalence levels were noted, although the largest index exposure prevalence scenarios for PACE and MAX always required more prunings to minimize imbalance. This relation was especially clear for PACE (Figure 4, panels A, D and G). Also, for a given covariate set size, lower index exposure prevalence values always corresponded to fewer prunings required to achieve the caliper criteria (Figures 8 and 9).

Matching Algorithm

The differences between the performances of NNM and DGM in reducing imbalance were not substantial in any scenario. For MAX, MDM performed better overall than NNM and DGM with respect to maintaining low covariate imbalance (Figure 5). However, for PACE, as the number of covariates used to build the propensity score model increased, MDM performance became increasingly worse, as evidenced by the elevated MDM trend lines (Figure 4). Finally, all MDM imbalance trends were effectively monotonic decreasing, whereas the paradox was visible in some cases for the NNM and DGM trends.

Tracking Changes in the Effect Estimate

The RR estimate trends for PACE and MAX are displayed in Figures 10 and 11, respectively. We found that, in general, the NNM and DGM trends were similar, especially at the left-most portion of each panel (i.e., in the caliper regions). For PACE, in the larger covariate set scenarios, the MDM trends indicated RR estimates further from the null than did the NNM and DGM trends, whereas in the Small PACE scenarios and in all MAX scenarios, all three algorithms produced similar RR estimates early in the pruning process. These findings corresponded to the findings regarding imbalance. Finally, in most cases, there was a clear difference between the pre-matched RR estimate and the RR estimates early in the pruning process. This difference also corresponded to the clear differences in imbalance among the datasets (e.g., compare Figure 11, panel A to Figure 5, panel A).

DISCUSSION

PSM greatly improved covariate balance compared with balance in the pre-matched dataset. The points at which our caliper criteria would have been met always were near the lowest points on the imbalance trends, indicating that matched datasets constructed from these data by many would have corresponded to excellent covariate balance. Although imbalance increased with further pruning when the propensity score model was based on a higher number of covariates, this phenomenon occurred only after pruning more matched sets than would have been required to achieve our caliper criteria. Moreover, although MDM led to near-monotonic decreasing imbalance trends, PSM achieved better covariate balance with fewer prunings and much larger matched dataset sizes for the larger covariate set scenarios.

The fact that the paradox was clearer in the larger covariate set scenarios was not surprising. When more covariates are used to build the underlying propensity score model, there is a greater probability that different individuals with similar propensity score values will have more dissimilar underlying covariate profiles, thus increasing the chance that balance will deteriorate after only a few prunings [King and Nielsen, 2016]. A similar logic applies to our finding that, in general, more prunings were required to achieve the caliper criteria when the underlying propensity score model was based on a larger vector of covariates. Even so, matching on the propensity score based on a larger vector of covariates always provided a great improvement in covariate balance in the caliper-matched dataset, compared to the pre-matched dataset – more so than MDM.

We found that manipulation of the index exposure prevalence affected the balancing of propensity score distributions more than the balancing of the underlying covariate distributions. For both NNM and DGM, the fact that the caliper criteria always were achieved with fewer prunings as the index exposure prevalence decreased was not surprising when considered from the perspective of balancing propensity score distributions. Lower index exposure prevalence equates to a higher probability of a single index unit finding a good reference unit match on the propensity score simply because, for a given study size, the pool of reference units is relatively larger when the index exposure prevalence is lower. However, it was difficult to perceive a clear effect on the underlying covariate balance, as evinced by the fact that the imbalance trend shapes did not change much as the index exposure prevalence was altered.

For our analyses, the PSM algorithm was not an important indicator of the appearance of the paradox, although previous studies comparing NNM with DGM have suggested a preference for NNM over DGM with respect to bias [Rassen et al., 2012].

The monotonicity of the MDM trends also was not surprising [King et al., 2011]. The failure of MDM to achieve adequate covariate balance *early* in the pruning process with more covariates may be attributed to known issues with MDM [Rubin, 1979; Gu and Rosenbaum, 1993; Zhao, 2004; Stuart, 2010]. It has been suggested that higher dimensions diminish the efficiency of MDM since, unlike logit-based PSM, MDM attempts to match units while regarding all interactions in the covariate space as equally important. Thus, having more covariates equates to having more complicated interactions to balance. This phenomenon may explain our finding that certain covariates were balanced differently after MDM compared with NNM and DGM and that the RR estimates were usually different for MDM, compared with NNM and DGM, with larger covariate sets (Figures 2 and 3; Figures 10 and 11). Thus, PSM may be the better option for the high-dimensional matching scenarios that are common to pharmacoepidemiologic research.

During matching, only covariate distribution imbalance and study size may be controlled directly, although the bias-variance trade-off for effect estimation certainly may be affected by the imbalance-study size tradeoff [King et al., 2011]. Thus, it is difficult to make strong statements regarding our effect estimate trends. Even so, in general there were no large differences between the RR estimates from NNM and DGM early in the pruning process, whereas MDM produced clearly different RR estimates when based on larger covariate set sizes.

We conclude that in our claims data, PSM in its conventional application would not have harmed covariate balance in the manner predicted based on King and Nielsen’s work. Although our findings conform to King and Nielsen’s description of the paradox, implementing either version of PSM in our datasets with any standard absolute propensity score distance caliper resulted in very good balance and preservation of sample size. Conversely, the utility of MDM depended on the pre-matched dataset and either resulted in excellent balance with few prunings or in excellent balance only after pruning a very large portion of the matched dataset.

Although we analyzed a limited set of conditions, we focused on data and techniques that are common in pharmacoepidemiology. Thus, our results bear important implications for applied researchers. Specifically, our results indicate that the paradox might not arise for situations in which the pre-matched dataset has high covariate imbalance and in which a reasonable absolute propensity score distance caliper is applied. We expect that the paradox only should be a practical concern when the pre-matched dataset has very low covariate imbalance, such that covariate balance worsens either after the full match, or after only a few prunings, as in our simple example; or in the unlikely scenario in which pruning is allowed to continue well beyond the point at which a reasonable absolute propensity score distance caliper would stop the pruning process, as in our example studies. We stress the importance of checking covariate balance after PSM in order to identify any increase in covariate imbalance – at the very least, via a univariate comparison of the pre- and post-matched covariate distributions. Finally, existing algorithms may be used to explore imbalance trends in order to identify disagreements between propensity score distribution balance and covariate balance [King et al., 2017].

CHAPTER 2: EVALUATING THE UTILITY OF COARSENED EXACT MATCHING FOR PHARMACOEPIDEMIOLOGY USING REAL AND SIMULATED CLAIMS DATA

INTRODUCTION

“Coarsened exact matching” (CEM) is a design strategy in cohort studies that has been shown to produce good covariate balance between exposure groups and, thus, to reduce the impact of confounding bias in observational causal inference [Iacus et al., 2011; Iacus et al., 2011]. The strategy is simply matching simultaneously by a set of multivariable values for potential confounders (“exact matching”). Coarsening refers to reducing the number of potential matching values for a given variable (e.g., by categorizing continuous variables) to increase the number of matches achieved. It has been demonstrated that CEM may outperform certain adjustment techniques that are common in pharmacoepidemiology with respect to covariate balance and effect bias [King et al., 2011; King and Nielsen, 2016]. For example, King, et al. [King et al., 2011] and King and Nielsen [King and Nielsen, 2016] demonstrated, using real and simulated data, that unlike CEM, propensity score matching (PSM) may increase covariate imbalance (although this may not apply to the typical pharmacoepidemiologic application of PSM [Ripollone et al., 2018]). Since CEM has, to our knowledge, not been implemented within the context of pharmacoepidemiologic analyses of claims data, and since CEM has been touted to have properties that may make it a preferable choice for causal inference [Iacus et al., 2011; Iacus et al., 2011], the utility of CEM for pharmacoepidemiology, compared with other standard techniques, should be explored.

Here, we compare CEM with 3 techniques for confounding control that have been used in pharmacoepidemiologic analyses [Rassen et al., 2012; Desai et al., 2017; Ripollone et al., 2018]: PSM, Mahalanobis distance matching (MDM) and fine stratification on the propensity score (FS). We present the results of a comparison of these four methods with respect to covariate balance, confounding control and effect estimate precision using real and simulated claims-based cohorts. We used typical pharmacoepidemiologic claims scenarios (i.e., large datasets with a large number of potential confounders of an association between a drug and health outcome [Petri and Urquhart, 1991; Patorno et al., 2013; Patorno et al., 2014]) to enhance the applicability of our results. To our knowledge, these techniques have not been compared, simultaneously, with respect to covariate balance, confounding control and effect estimate precision within the context of claims-based analyses, although some separate comparisons have been performed [Iacus et al., 2011; King et al., 2011; Fullerton et al., 2016; King and Nielsen, 2016; Desai et al., 2017; Ripollone et al., 2018].

METHODS

Here, we describe the mechanics of CEM, PSM, MDM and FS as well as our approach to comparing these methods.

Coarsened Exact Matching

Let **X** be the vector of observed covariates. CEM entails: (1) coarsening the covariates in **X** (i.e., categorizing continuous variables, or further collapsing categorical variables) so that similar units are assigned the same value for the coarsened covariate; (2) implementing exact matching with the coarsened data – index-exposed and reference-exposed units (i.e., units with and without the exposure of interest, respectively) that appear in the same bin of the multi-way array created by the coarsening strategy are considered "exactly-matched"; (3) eliminating units that appear in bins that do not contain units of opposite exposure status (i.e., eliminating unmatched units) – such bins represent regions of non-positivity that should not contribute to treatment effect estimates [Petersen et al., 2012]; (4) estimating the effect of interest in the matched dataset, with weights applied to individual units [Iacus et al., 2011; Iacus et al., 2011].

The coarsened boundaries in step 1 should be determined through substantive knowledge. However, empirical, "auto-" coarsening methods may be used when substantive knowledge is scarce [Iacus et al., 2011; Iacus et al., 2018]. The weighting scheme in step 4 is critical since unequal numbers of index-exposed and reference-exposed units may appear in a given bin, and across bins. Proper weighting is necessary to achieve the covariate balance between exposure groups. The scheme used for CEM applies a weight of 1 to each index-exposed unit and weights reference-exposed units in each matched set in proportion to the distribution of index-exposed units in the matched set. If a higher proportion of reference-exposed units, with respect to the number of reference-exposed units among *all* matched sets, appears in the matched set, compared with the equivalent proportion of index-exposed units, the reference-exposed units are down-weighted, and vice versa. Thus, reference-exposed units receive a weight characterized by the ratio of the proportion of total index-exposed units appearing in the matched set to the equivalent proportion of reference-exposed units [Iacus et al., 2011; Iacus et al., 2011]:

$$(N_{\text{Index-exposed in Matched Set}} \, / \, N_{\text{Total Index-exposed}}) \, / \, (N_{\text{Reference-exposed in Matched Set}} \, / \, N_{\text{Total Reference-exposed}}).$$

We present a complete derivation of this weight in the Appendix.

With CEM, covariate balance is never worse than the balance in the original dataset [Iacus et al., 2011; Iacus et al., 2011; King et al., 2011; King and Nielsen, 2016]. A coarsening strategy resulting in more strata will achieve better covariate balance. For scalar-based matching techniques, such as PSM and MDM, covariate balance for every variable is not necessarily guaranteed. It can be checked *after* matching, at which point it might be decided that the process should be performed again (e.g., using a different distance caliper in the matching algorithm) to improve covariate balance. Moreover, unlike other techniques, CEM guarantees balance for higher-order terms, such as interactions, between exposure groups [Iacus et al., 2011].

Propensity Score Matching

We focus on the case of 1:1 PSM without replacement, given its popularity in biomedical fields such as pharmacoepidemiology [Glynn et al., 2006; Austin, 2008; Austin, 2009; Austin and Small, 2014; Wu et al., 2015; Jackson et al., 2017]. PSM entails: (1) for each unit, estimating the propensity score: the probability of receiving the index exposure, conditional on **X**, for every unit in a dataset (commonly via logistic regression); (2) matching one reference-exposed unit to one index-exposed unit, without replacement, via some algorithm (e.g., nearest neighbor matching [Ho et al., 2018]); (3) pruning from the resulting dataset the matched pairs with the largest propensity score distances to eliminate poorly-matched units (usually via application of a caliper as part of step 2); (4) comparing (usually at the univariate level) pre- and post-matched **X** distributions to assess the improvement in covariate balance and re-running steps 1-3, if necessary; (5) estimating the effect of interest in the matched dataset [Austin, 2008; Austin, 2009; Austin and Small, 2014; Pan and Bai, 2015; Wu et al., 2015].

The key theoretical benefit of PSM is the ability to match on a scalar summary of **X**, which may involve a large number of variables in a typical claims study [Patorno et al., 2013; Patorno et al., 2014]. This benefit, along with other benefits that have been outlined extensively [Rosenbaum and Rubin, 1983; Austin, 2007; Austin et al., 2007; Austin, 2008; Austin, 2008; Hade and Lu, 2014; Austin and Schuster, 2016; Ripollone et al., 2018], may explain the popularity of PSM in pharmacoepidemiology [Rosenbaum and Rubin, 1983; Petri and Urquhart, 1991; Glynn et al., 2006; Patorno et al., 2013; Patorno et al., 2014; Jackson et al., 2017].

Mahalanobis Distance Matching

MDM operates similarly to PSM, except that it is based on the Mahalanobis distance, which, unlike the distance between propensity scores, is measured in the *actual* covariate space, and has the following form:

$$\sqrt{}[(\mathbf{X}_i-\mathbf{X}_j)'\underline{\Sigma}^{-1}(\mathbf{X}_i-\mathbf{X}_j)],$$

where i indexes the exposed unit, j indexes the unexposed unit and $\underline{\Sigma}$ is the sample covariance matrix of the original data [Iacus et al., 2011; Ripollone et al., 2018]. Similar to PSM, the key benefit of MDM is the dimension reduction reflected in a scalar summary of **X** [King et al., 2011; King and Nielsen, 2016; Ripollone et al., 2018].

Fine Stratification on the Propensity Score

FS is a modification of the normal approach to stratification on the propensity score, using a high number (e.g., 50) of propensity score strata [Cochran, 1968; Rosenbaum and Rubin, 1984; Desai et al., 2017]. Strata-specific estimates may be pooled, or the same weights described for CEM may be applied before effect estimation (i.e., to account for the unequal numbers of index-exposed and reference-exposed units within a given propensity score stratum, and across propensity score strata). A key benefit of FS, not shared by any of the matching techniques, is high retention of study subjects (leading to precise effect estimates) in the analytic dataset [Desai et al., 2017]. Only the nonoverlapping tails of the PS distribution are dropped from the analysis; within the range of overlap, every unit falls into a PS stratum and is counted in the analysis. FS overcomes the biggest drawback of matching, which is the exclusion of unmatched units.

Description of Real Datasets

Two claims cohorts were used. The first was a cohort of 49,919 low-income Medicare beneficiaries, at least 65 years of age, who were enrolled in the Pharmaceutical Assistance Contract for the Elderly database in New Jersey over the years 1999-2002 and who initiated non-selective NSAIDs or selective COX-2 inhibitors (hereafter, “NSAID cohort”) [Brookhart et al., 2006; Schneeweiss et al., 2006]. The NSAID cohort was previously generated to perform an analysis of the effect of selective COX-2 inhibitors, compared with non-selective NSAIDs, on the risk of gastrointestinal complications. Approximately 60% of patients represented in this cohort were selective COX-2 inhibitor initiators. Approximately 2,000 cases of gastrointestinal complication were observed in this cohort. Three covariate sets were used for the NSAID cohort analyses. The “small” set comprised 19 continuous and binary covariates that were selected based on clinical importance. The second and third covariate sets (“standard” and “large”, respectively) comprised binary covariates (representing concomitant medications, comorbidities and other medical encounters) selected by a high-dimensional propensity score algorithm [Schneeweiss et al., 2009], in addition to the 19 pre-determined covariates: the 50 covariates with the highest bias-based ranks were included in the standard covariate set, and the 100 covariates with the highest bias-based ranks were included in the large covariate set. The distribution of the small set of pre-matched covariates in the NSAID cohort set is shown in Table 2.

The second cohort comprised information on 886,996 completed pregnancies and was generated from the Medicaid Analytic eXtract over the years 2000-2007 (hereafter, “statin cohort”) [Huybrechts et al., 2014; Bateman et al., 2015; Desai et al., 2017]. The statin cohort was used to perform an analysis of the effect of statin use during the first trimester of pregnancy, compared with no use during the first trimester of pregnancy, on the risk of congenital malformation in the infant. Statin use was defined as the existence of at least one claim for a dispensed statin within the first trimester. Approximately 0.13% of women in this cohort filled a statin prescription during the first trimester. Approximately 30,000 congenital malformations were observed in this cohort. The statin cohort comprised 20 categorical covariates, which were selected based on clinical importance. The distribution of pre-matched covariates in the statin cohort is shown in Table 3.

Analysis of Real Datasets

For each of the 4 real datasets (3 NSAID cohort-based datasets plus the statin cohort), we performed CEM, PSM, MDM and FS. For CEM, we applied the R CEM package default auto-coarsening strategy, which attempts to divide the range of values for the numerical covariates in **X** into the number of bins required to approximate a normal density (Sturges' rule) [Iacus et al., 2011; Iacus et al., 2018]. For the NSAID cohort PSM and FS analyses, all continuous variables were categorized to relax distributional assumptions for the propensity score model. For PSM and MDM, we used a nearest neighbor matching algorithm. To emulate previous analyses of these data, we applied a 0.025 absolute propensity score distance caliper for PSM, but allowed all exposed units to be matched for MDM [Ripollone et al., 2018]. We performed MDM for all 3 NSAID cohort-based datasets for the sake of example, even though in practice, MDM is not warranted for high-dimensional scenarios, where the MDM algorithm is slow to implement and sub-optimal with respect to covariate balance [Rubin, 1979; Zhao, 2004; Stuart, 2010; King et al., 2011; Ripollone et al., 2018]). Thus, we expected to observe worse covariate balance from MDM in the larger NSAID cohort-based analyses. For FS, we trimmed regions of non-overlap between exposed and unexposed propensity score distributions and generated 50 strata based on quantiles of the exposed propensity score distribution.

We assessed covariate balance in the resulting analytic datasets using the Mahalanobis balance metric, which has been used in previous methodological assessments in pharmacoepidemiology [Franklin et al., 2014; Ripollone et al., 2018]. The Mahalanobis balance is a type of Mahalanobis distance that represents the extent of covariate balance in the actual covariate space, and has the following form:

$$\sqrt{}[(\bar{\mathbf{X}}_{T1}-\bar{\mathbf{X}}_{T0})'\underline{\Sigma}^{-1}(\bar{\mathbf{X}}_{T1}-\bar{\mathbf{X}}_{T0})],$$

where $\bar{\mathbf{X}}_{Tk}$ is the vector of covariate means in exposure group k, and $\underline{\Sigma}$ is the sample covariance matrix of the original data [Gu and Rosenbaum, 1993; Franklin et al., 2014]. Higher Mahalanobis balance values indicate worse covariate balance. For the CEM and FS scenarios, units were weighted before calculating the Mahalanobis balance.

We then estimated risk ratios and corresponding 95% Wald confidence intervals generated from log-binomial regression models. For the NSAID cohort, we estimated the risk ratio corresponding to the effect of non-selective NSAIDs, compared with COX-2 inhibitors, on the risk of gastrointestinal complications. For the statin cohort, we estimated the risk ratio corresponding to the effect of statin use during the first trimester of pregnancy, compared with no use during the first trimester of pregnancy, on the risk of congenital malformation. For the CEM and FS scenarios, units were weighted before calculating the risk ratios and corresponding standard errors.

Description of Simulated Datasets

A series of plasmode-simulated datasets were generated using the NSAID cohort. In plasmode simulation, the true effect of exposure on outcome in a real cohort is set to a known value, but the associations within the observed exposure-covariate data from that cohort are preserved and are allowed to confound this true effect [Franklin et al., 2014]. Plasmode simulation is particularly apt for methodologic research in claims data because it maintains observed complex data structures.

Plasmode simulation for a binary outcome scenario entails:

(1) Regressing outcome on exposure and on the set of desired covariates from the original cohort (using a generalized linear model approach) to obtain a set of model parameter estimates corresponding to exposure and to each covariate;

(2) Sampling, with replacement, exposed and unexposed units from the original cohort to obtain the desired study size and exposure prevalence, retaining the original exposure-covariate values for each unit in each sample;

(3) Altering the model parameter estimate for exposure and intercept from the model in step 1 to specify the desired exposure effect and the desired baseline prevalence of outcome, respectively, in the sample (the effect strengths of the other covariates from that model also may be altered).

(4) Applying the altered model from step 3 to each sample from step 2 to calculate the probability of outcome and, in turn, binary outcome status for each unit. Because of the specification of exposure effect, the true desired effect will be reflected in each sample [Vaughan et al., 2009; Franklin et al., 2014; Franklin et al., 2015].

Simulation scenarios were constructed by simulating outcome (gastrointestinal complications, 20% event rate in all scenarios), using all of the covariates included in a given scenario to predict outcome. The true risk ratio for each scenario was set at 1 (ln[1] = 0). Each scenario comprised 1,000 simulated cohorts of 25,000 units and represented a variation of index exposure prevalence and covariate set size. The index exposure prevalence values were 5%, 10%, 20%, 30% and 40% and the covariate set sizes were small, standard and large. Two additional small covariate scenarios included a product term representing the interaction between continuous age and continuous Charlson comorbidity score in the outcome generation model. In one scenario, the coefficient on the product term maintained its original estimated value from the real data ("default"). In the other scenario, the strength of the product term was increased by 200% ("exaggerated"). For both product term scenarios, index exposure prevalence was set at 20%. We generated product term scenarios because CEM guarantees balance on such terms (within the limits of the coarsening strategy), while PSM, FS and MDM do not guarantee balance on such terms [Stuart, 2010; Iacus et al., 2011]. We summarize our simulation scenarios in Table 4.

Analysis of Simulated Datasets

We applied the same methods that were used for the analysis of the real datasets. For the scenarios that included a product term between age and Charlson comorbidity score, we performed CEM using a manual coarsening strategy for the age and Charlson comorbidity score variables to ensure that lack of balance on those variables was not due to use of inappropriate coarsening boundaries. Specifically, we coarsened the age variable into groups of 5 years, and we coarsened the Charlson comorbidity score variable into the following groups: 0, 1, 2, 3, ≥4. We only performed MDM for the small covariate set scenarios, for reasons explained above.

We compared the following measures among the methods [Burton et al., 2006]:

(1) Average proportional decrease in Mahalanobis balance, from the original Mahalanobis balance;

(2) Bias = [average adjusted ln(risk ratio) value] - [true ln(risk ratio)];

(3) Variance of the adjusted ln(risk ratio) values;

(4) Square root of mean squared error (rMSE) = $\sqrt{[\text{bias}^2 + \text{variance}]}$.

RESULTS

Analysis of Real Datasets

We present the results of the analysis of real datasets in Table 5. CEM always produced essentially perfect covariate balance (Mahalanobis balance values never greater than 0.020), although PSM and FS still demonstrated notable improvement in covariate balance, compared with crude balance. MDM was worst with respect to covariate balance in each NSAID cohort analysis – with Mahalanobis balance values increasing from 0.207 to 0.681 (which was *worse* than the corresponding crude Mahalanobis balance) as covariate set size increased. However, for the statin cohort analysis, MDM performed better with respect to covariate balance compared with PSM and FS (adjusted Mahalanobis balance values: 0.244, 1.632, 0.586, respectively).

CEM always produced the least precise effect estimate (highest 95% confidence interval width in each case – even up to 30.58 for the large NSAID cohort analysis). Conversely, FS always was optimal with respect to precision (lowest 95% confidence interval width in each case). PSM and MDM produced effect estimates with similar levels of precision.

Analysis of Simulated Datasets

Non Interaction Scenarios

CEM and FS maintained the highest average proportional decrease in Mahalanobis balance among the 4 methods (Figure 12). CEM only performed worse than FS with respect to balance improvement in the 5% and 10% index exposure prevalence standard and large covariate set scenarios. Generally, PSM performed worst with respect to balance improvement in the lowest index exposure prevalence scenarios. For the small covariate set scenarios, MDM generally performed worst among the methods with respect to balance improvement and produced a consistently decreasing trend in balance improvement with increasing index exposure prevalence. Finally, balance improvement for CEM, PSM and FS became slightly worse, for a given index exposure prevalence, as covariate set size increased.

Perhaps the key finding is that CEM always produced the highest rMSE among the 4 methods, with the highest values seen in the standard and large covariate set scenarios (Figure 13, panels B and C, respectively). In the small covariate set scenarios, the rMSE from CEM was highest with 5% index exposure prevalence and generally declined as index exposure prevalence increased (Figure 13, panel A). For PSM, MDM and FS, rMSE generally decreased as index exposure prevalence increased (Figure 14). For a given index exposure prevalence, there was a slight upward trend in rMSE as covariate set size increased for all 3 methods. In most scenarios, FS produced the lowest rMSE. PSM and FS always produced similar rMSE values for the higher index exposure prevalence scenarios, but FS always produced lower rMSE values, compared with PSM, in the lower index exposure prevalence scenarios. For the small covariate set scenario, MDM always produced the highest rMSE.

It was clear that variance drove the high rMSE values for CEM, since the CEM variance trends (Figure 17) were similar to the CEM rMSE trends (Figure 13). The strong influence of variance on the rMSE trends also was seen for PSM, FS and MDM, among which the FS variance trends were lowest (Figure 18). The CEM bias trends were much higher, overall, compared with the PSM, FS and MDM bias trends (Figure 15 –

especially panels B and C). The latter 3 bias trends were relatively similar across all scenarios, with PSM and FS yielding the lowest bias values (Figure 16).

Interaction Scenarios

We display all interaction scenario results in Table 6. The trends among all measures for the default and exaggerated scenarios were the same as those seen for the non interaction scenarios. There were no substantial differences among the measures comparing the default scenario with the exaggerated scenario.

We demonstrate the extent to which CEM improved covariate balance between the index-exposed and reference-exposed groups within the context of the interaction between age and Charlson comorbidity score in Table 7. This table shows the absolute differences between the exposure groups with respect to the average of the average age (or weighted average age for CEM and FS) within each coarsened category of Charlson comorbidity score, and vice versa, across plasmode simulations (default scenario only). CEM yielded the lowest difference values among the 4 methods. Unlike the other 3 methods, CEM never produced a difference value that was higher than the corresponding difference value in the original simulated cohort. Thus, as expected, CEM led to much better covariate balance among the coarsened strata of the covariates associated with the product term compared with the other 3 methods.

DISCUSSION

Overall, the analyses of real and simulated datasets led to the same conclusions. CEM was optimal with respect to covariate balance and FS was optimal with respect to bias and precision (and still maintained excellent covariate balance). PSM tended to perform almost as well as FS with respect to all simulation metrics, especially for higher exposure prevalence scenarios. The performance of MDM generally never surpassed that of FS and PSM.

The optimal performance of CEM with respect to covariate balance effectively was guaranteed by the high number of binary covariates in our data (thus, CEM amounted to exact matching) [Iacus et al., 2011; Iacus et al., 2011; King et al., 2011; Fullerton et al., 2016; King and Nielsen, 2016]. FS performed almost as well as CEM, and better than PSM and MDM, with respect to covariate balance. Since 50 strata were used, the maximum distance between index-exposed and reference-exposed units within a given stratum usually was very low – even lower than the PSM absolute propensity score distance caliper of 0.025. The low "implied calipers" associated with FS corresponded to high covariate balance overall [5]. Moreover, since it already has been shown that FS tends to outperform PSM with rare index exposure prevalence, the differences between FS and PSM with respect to covariate balance improvement in the lowest index exposure prevalence scenarios were not surprising [Desai et al., 2017]. The fact that PSM, CEM and FS generally performed worse with respect to covariate balance improvement, for a given index exposure prevalence, as covariate set size increased, is attributable to the difficulties of achieving covariate balance in higher dimensions [King et al., 2011; Ripollone et al., 2018].

In the analysis of simulated datasets, the very high rMSE values associated with CEM were due to the extreme loss of study size, and the corresponding decrease in the number of outcomes, that occurred during creation of the matched datasets. This extreme loss of study size may explain the discrepancy between the CEM average proportional decrease in Mahalanobis balance trends and the CEM bias trends, which would be expected to coincide (i.e., improvement in covariate balance for true confounders should be complemented by low bias in the effect estimate). In other words, the decrease in effective study size and number of outcomes across simulations was so consequential that the resulting sparse data led to elevated bias trends [Greenland et al., 2016]. This extreme loss of study size also was clear in the analysis of the real NSAID cohort: in the small scenario, the matched dataset produced by CEM comprised 16,139 units and 106 outcomes, representing a decrease in study size and number of outcomes of approximately 70% and 80%, respectively (Table 5). These numbers decreased dramatically as covariate set size increased.

The decrease in study size associated with CEM is intuitive since CEM effectively was exact matching in our scenarios. This phenomenon also explains the finding that CEM performed best with respect to rMSE in the small covariate set scenarios, with higher index exposure prevalence: matching exactly on a small vector of covariates with many exposed units led to better retention of outcomes and, thus, to lower rMSE. Conversely, the large analytic cohorts resulting from FS (leading to low variance) and the consistently low bias values associated with FS were responsible for the low rMSE values observed for FS. Thus, overall, FS was optimal among the 4 methods with respect to rMSE. Notably, PSM performed almost as well as FS with respect to rMSE, increasingly so as index exposure prevalence increased – a result also seen in previous work [Desai et al., 2017].

The overall suboptimal performance of MDM, especially with respect to covariate balance, may be attributed to known issues with MDM [Rubin, 1979; Gu and Rosenbaum, 1993; Zhao, 2004; Stuart, 2010; Ripollone et al., 2018]. The fact that covariate balance for MDM decreased with higher index exposure prevalence was not surprising since no matched set pruning was performed. Thus, overall, with increasing index exposure prevalence, the matched dataset's Mahalanobis balance value approached the original dataset's Mahalanobis balance value. A similar logic applies to the decreasing bias trend for MDM: overall, since bias already was relatively low in the original dataset, the bias from MDM approached the bias from the original dataset as index exposure prevalence increased (Web Figure 2, panel A). It is worth noting that MDM performed almost as well as PSM with respect to variance, mainly because of the lack of matched set pruning for MDM.

Although in our analyses CEM always was optimal with respect to covariate balance, the ultimate objective is to obtain a valid and precise effect estimate. The high levels of balance achieved by CEM in our study were not complemented by low rMSE values because CEM produced heavy losses in study size (and numbers of outcomes) to achieve this balance. If not for this problem, there would be less motivation to pursue a dimension reduction technique, such as a propensity score-based method. Therefore, in these types of pharmacoepidemiologic analyses, CEM may not be the optimal choice, especially if the vector of important confounders is large. Instead, FS may be optimal with respect to confounding control and effect estimate precision. PSM may perform similarly to FS, especially if index exposure prevalence is high.

Our simulation study had some noteworthy limitations. Although we covered a wide range of scenarios (by varying index exposure prevalence and covariate set size), the simulated data were based on only one real cohort, exemplifying only one type of complex pharmacoepidemiologic claims exposure-covariate structure. The statin cohort, for example, also could have been used (although we note that the NSAID cohort allowed us more flexibility in terms of covariate set size variation). Also, we only implemented the 4 methods in the common manner (e.g., auto-coarsening strategy for CEM, use of a 0.025 absolute propensity score distance caliper for PSM, etc.), not necessarily in an optimal manner. Future work may be warranted to fill the gaps left by these limitations.

CHAPTER 3: ACCOUNTING FOR SAMPLING VARIABILITY IN EFFECT ESTIMATION AFTER 1:1 PROPENSITY SCORE MATCHING WITHOUT REPLACEMENT: A REVIEW OF THEORY AND METHODS

INTRODUCTION

The conventional approach to estimating the standard error of the effect estimate in propensity score analysis does not specifically account for the sampling variability associated with propensity score estimation [McCandless et al., 2009; Alvarez and Levin, 2014; Austin and Small, 2014; Pan and Bai, 2015]. This approach is common for propensity score matching (PSM), the most popular propensity score technique [Morgan and Winship, 2007; Pearl, 2010; Pan and Bai, 2015]. Generally, in PSM, only the variability directly associated with effect estimation is considered in the standard error of the effect estimate, leaving the impact of sampling variability on the actual propensity score estimation process unaccounted. This practice may lead to inaccurate estimation of the standard error of the effect estimate [Alvarez and Levin, 2014; Austin and Small, 2014; Abadie and Imbens, 2016].

Since Rubin and Thomas first highlighted unique characteristics of variance in PSM [Rubin and Thomas, 1992; Rubin and Thomas, 1996], the pool of literature on handling sampling variability in PSM has grown. However, it is difficult to find a straightforward depiction of how the sampling variability associated with propensity score estimation manifests in PSM. In light of the popularity of PSM, we sought to provide this depiction as well as an explanation of methods that may account for this sampling variability better than the conventional approach to PSM.

We focus on the case of 1:1 PSM *without* replacement (hereafter, “1:1 PSM”) since it is a popular PSM approach in biomedical fields, such as pharmacoepidemiology [Glynn et al., 2006; Austin, 2008; Austin, 2008; Austin, 2009; Austin and Small, 2014; Wu et al., 2015; Jackson et al., 2017]. For our explanation of sampling variability in 1:1 PSM, we assume use of a deterministic matching algorithm (i.e., one that always produces the same matches, for the same pre-matched sample, based on closest propensity score distance between an exposed unit and an unexposed unit) for the sake of simplicity [Rassen et al., 2012]. We summarize the main facets of bootstrap and Bayesian methods that attempt to account for this sampling variability and we illustrate the use of these methods using a real pharmacoepidemiologic claims-based study [Tu and Zhou, 2002; McCandless et al., 2009; Kaplan and Chen, 2012; Austin and Small, 2014; Pan and Bai, 2015].

Although not the focus of our review, the variability associated with propensity score estimation also affects standard error estimation in other types of propensity score analysis (e.g., stratification by propensity score, inverse weighting by propensity score) and that methods to address this issue for these other types of analyses exist as well [Hirano et al., 2003; Lunceford and Davidian, 2004; Li and Greene, 2013; Li et al., 2017]. These other types of propensity score analysis (notably, stratification by propensity score) may be preferable to 1:1 PSM with respect to overall efficiency. We focused on 1:1 PSM due to its frequent use in pharmacoepidemiology [Rosenbaum and Rubin, 1983; D'Agostino, 1998; Pearl, 2010; Desai et al., 2017].

SAMPLING VARIABILITY IN 1:1 PSM

Figure 19 demonstrates the conventional application of 1:1 PSM using 10 units (5 exposed, 5 unexposed). We use 10 units only for demonstration purposes, and we can assume that these units come from a larger population. From the 10-unit population, we demonstrate the derivation of 6 matched sets. In the 10-unit population, each numbered exposed unit has exactly the same true propensity score as its corresponding unexposed unit (i.e., exposed unit 1 and unexposed unit 1 have exactly the same true propensity score, etc.). Moreover, each of the 5 exposed-unexposed pairs has a unique, true propensity score.

By calculating the conventional standard error (e.g., the standard error of a risk ratio from a log binomial regression model), the analyst implicitly assumes that the effect estimate is generated from a direct “random” sample of the population. Thus, the matched sets are considered direct “random” samples of the population, as shown in Figure 19. However, the *pre-matched* samples (i.e., the original dataset for a given study), not the matched sets, are directly derived from the population. Therefore, sampling variability in 1:1 PSM first affects the selection of pre-matched samples and influences not only effect estimation but also the intermediate step of propensity score estimation.

In addition to the population and matched sets, Figure 20 displays the corresponding pre-matched samples, each of which is unique and, thus, is expected to result in *different* estimated propensity score models. The impact of sampling variability on propensity score estimation in 1:1 PSM can be demonstrated using exposed unit 1, which appears in all 6 pre-matched samples and corresponding matched sets.

Pre-matched Sample 1: Exposed unit 1 and unexposed unit 1 receive their true propensity scores as their propensity score estimates and no other unexposed unit receives this propensity score estimate. Consequently, unexposed unit 1 is guaranteed to be matched to exposed unit 1.

Pre-matched Sample 2: Exposed unit 1 receives its true propensity score as its propensity score estimate, but unexposed unit 1 does *not* receive its true propensity score as its propensity score estimate. Unexposed unit 2 is matched to exposed unit 1 because its propensity score estimate is closest to exposed unit 1's propensity score estimate.

It is possible for exposed unit 1 to receive its true propensity score as its propensity score estimate, but for unexposed unit 1 *not* to receive its true propensity score as its propensity score estimate if, for example, exposed unit 1 and unexposed unit 1 have different covariate values underlying their shared, true propensity score value (i.e., even though unexposed unit 1 and exposed unit 1 have the same true propensity score, the underlying covariate profiles still may differ between the two units [Rosenbaum and Rubin, 1983]). During propensity score estimation, a binary covariate may receive an inaccurate coefficient estimate, which may cause units that have a specific value for this binary covariate (e.g., "1" for unexposed unit 1) to receive inaccurate propensity score estimates.

Pre-matched Sample 3: Exposed unit 1 does *not* receive its true propensity score as its propensity score estimate, but unexposed unit 1 *does* receive its true propensity score as its propensity score estimate (this is the opposite of what was seen in pre-matched sample 2). Unexposed unit 5 is matched to exposed unit 1 because its propensity score estimate is closest to that of exposed unit 1's propensity score estimate.

Pre-matched Sample 4: Unexposed unit 1 is not in this pre-matched sample. Even if exposed unit 1 receives its true propensity score as its propensity score estimate, it cannot be matched to unexposed unit 1.

Pre-matched Sample 5: As in the first matched set, exposed unit 1 and unexposed unit 1 receive their true propensity scores as their propensity score estimates and are matched. However, unlike pre-matched sample 1, pre-matched sample 5 comprises exposed unit 5 and unexposed unit 5, making the resulting matched set distinct from matched set 1.

Pre-matched Sample 6: Although the sixth matched set is exactly the same as the first matched set, pre-matched sample 6 is not the same as the pre-matched sample 1. Sampling variability still is evident in the propensity score estimates (i.e., the propensity score estimates are slightly different because pre-matched sample 6 is different from pre-matched sample 1), even though the matching decisions are not different from those seen in the first matched set.

Thus, different pre-matched samples may result in different propensity score estimates for the same units and, consequently, in different matching decisions for those units. Moreover, the matched set may not include certain units, either because they were not matched or because they were not in the pre-matched sample. The key point is that the exposure model generated in each pre-matched sample is an *estimate* of the true population exposure model and, thus, is a manifestation of sampling variability. This manifestation of sampling variability is *not* necessarily represented in the standard error of the effect estimate from a conventional 1:1 PSM analysis.

Our depiction of sampling variability in 1:1 PSM is relevant for other types of propensity score analysis as well. For example, if stratification by propensity score was applied in our depiction instead, the strata in which exposed unit 1 would appear (i.e., in the final analytic dataset) may have comprised different exposed and unexposed units, depending on the composition of the corresponding original dataset (i.e., the dataset that was sampled from the population, before any stratification).

METHODS TO ACCOUNT FOR SAMPLING VARIABILITY IN 1:1 PSM

Bootstrap and Bayesian techniques have been shown to accurately estimate the standard error of an effect estimate from 1:1 PSM without replacement. We describe the methods that, to our knowledge, are the only established methods for accurately estimating the standard error of the effect estimate for the specific case of 1:1 PSM without replacement.

Bootstrap 1:1 PSM

Use of the bootstrap for standard error estimation (especially when the observed data are not associated with a known probability distribution) is an established statistical practice [Efron, 1979; Efron and Tibshirani, 1986]. It has been suggested that bootstrap procedures are ideal for propensity score analysis since these procedures provide non-parametric, robust statistics for complex distributions, such as the distribution of propensity scores [Guo and Fraser, 2010; Bai, 2013; Austin and Small, 2014]. To this end, Austin and Small [Austin and Small, 2014] described and evaluated the “simple bootstrap” and the “complex bootstrap”. It is worth noting that other bootstrap methods for propensity score matching, such as the “wild bootstrap”, have been developed, but that these methods are ideal only for the case of PSM *with* replacement and, thus, were not addressed here [Otsu and Rai, 2017; Bodory et al., 2018].

Simple Bootstrap 1:1 PSM

Simple bootstrap 1:1 PSM is performed by applying the standard bootstrap to the propensity score-matched set [Austin and Small, 2014]. The bootstrap sample is generated by sampling *matched pairs* (as opposed to individual units) from the propensity score-matched set, with replacement, so that the size of the bootstrap sample is the same as the size of the propensity score-matched set. Thus, if the propensity score-matched set comprises N matched pairs, the bootstrap sample also will comprise N matched pairs. M such bootstrap samples are drawn and the relevant effect is estimated using each of the M bootstrap samples. The standard deviation of the estimated effects across the M samples is used as an estimate of the standard error of the effect estimate derived from the original propensity score-matched set. We depict the mechanics of simple bootstrap 1:1 PSM in Figure 21.

Complex Bootstrap 1:1 PSM

In complex bootstrap 1:1 PSM, M standard bootstrap samples are generated by sampling *individual units* from the pre-matched sample, with replacement. In each of the M samples, 1:1 PSM is performed and the relevant effect is estimated. The standard deviation of the estimated effects across the M samples is used as an estimate of the standard error of the effect estimate derived from the original matched sample. Since the bootstrap procedure is based on the pre-matched sample (not on matched pairs from the matched set), the M samples may have varying numbers of matched pairs. Although, in their simulation study, Austin and Small [Austin and Small, 2014] demonstrated a slight advantage for standard error estimation by simple bootstrap 1:1 PSM over complex bootstrap 1:1 PSM, we included complex bootstrap 1:1 PSM in our review because, unlike for simple bootstrap 1:1 PSM, its sampling scheme *directly* accounts for the potential variation due to propensity score matching (i.e., by repeatedly performing 1:1 PSM). We depict the mechanics of complex bootstrap 1:1 PSM in Figure 22.

Bayesian 1:1 PSM

A key benefit of Bayesian methodology is the ability to incorporate prior information (“beliefs”) regarding a parameter of interest (usually a measure of effect, such as the risk ratio, in epidemiology) into the analysis of the data. The incorporation of prior information into the analysis of the data leads to a “posterior” distribution on which final estimates of the parameter of interest are based. We direct the reader who is unfamiliar with Bayesian methodology to the introductory literature regarding Bayesian statistics for epidemiology, especially Spiegelhalter, et al. [Spiegelhalter et al., 2004], Greenland [Greenland, 2006; Greenland, 2007; Greenland, 2009] and MacLehose [MacLehose, 2014].

Although propensity score analysis generally is performed within the context of frequentist statistics, there is a growing literature on Bayesian applications in propensity score analysis [Zigler, 2016]. Much of this work has addressed incorporation of the sampling variability associated with propensity score estimation into the final effect estimate using prior information to generate posterior distributions for the propensity score model parameters [Hoshino, 2008; McCandless et al., 2009; An, 2010; Kaplan and Chen, 2012; Kaplan and Chen, 2014]. This Bayesian approach (hereafter, “BPSM”; described by An [An, 2010] and Kaplan and Chen [Kaplan and Chen, 2012; Alvarez and Levin, 2014]) can produce accurate standard error estimates for the case of 1:1 PSM. BPSM proceeds as follows.

(1) Estimate the propensity score model using a Markov-chain Monte Carlo (MCMC) methodology. MCMC methods are simulation techniques, commonly used in Bayesian analyses, that can generate a sample from the posterior distribution without a specific algebraic form for that distribution [Spiegelhalter et al., 2004]. The posterior distribution summarizes the remaining uncertainty in the parameter, after accounting for the prior information and the data. For BPSM, a MCMC-based logistic regression modeling procedure may be used to generate a sample from the joint posterior distribution for the propensity score model parameters (the intercept and vector of slopes). The size of the posterior distribution sample for these parameters corresponds to the number of simulations saved from the MCMC process [Kaplan and Chen, 2012]. Thus, if the MCMC process contributes N simulations to the posterior distribution, there will be N different propensity score models (N sets of intercept and slope values).

(2) Apply each of the N different MCMC-based propensity score models from step 1 to the original data to create N different sets of propensity score estimates. This step effectively creates a posterior distribution for the propensity score estimates for each unit that accounts for the uncertainty in estimation of the propensity score model.

(3) Perform 1:1 PSM on the original data using each of the N sets of propensity score estimates from step 2, resulting in N different matched sets. The N matched sets are created independently, so they may have varying numbers of matched pairs.

(4) Generate effect and standard error estimates for each of the N matched datasets from step 3 using a standard *frequentist* approach (e.g., a log binomial outcome model for the risk ratio).

(5) Use the average of the N effect estimates as the final effect estimate.

(6) Apply a formula for estimating the standard error of the effect estimate that is based on the law of total variance, as described by Kaplan and Chen [Kaplan and Chen, 2012]. We depict the mechanics of complex bootstrap 1:1 PSM in Figure 23.

A “full Bayesian” approach, in which the outcome model also is estimated from the matched dataset via an MCMC method (incorporating a prior distribution for the outcome model in step 4), also could be used [Kaplan and Chen, 2012]. However, Kaplan and Chen [Kaplan and Chen, 2012] indicate that such an approach may lead to standard errors that are *less* accurate than the standard errors from the BPSM approach.

Empirical Example

Description of Dataset

We used a cohort of 49,919 low-income Medicare beneficiaries, at least 65 years of age, who were enrolled in the Pharmaceutical Assistance Contract for the Elderly database in New Jersey over the years 1999-2002 and who initiated non-selective NSAIDs or selective COX-2 inhibitors [Brookhart et al., 2006; Schneeweiss et al., 2006]. This cohort was generated to perform an analysis of the effect of selective COX-2 inhibitors, compared with non-selective NSAIDs, on the risk of gastrointestinal complications. Approximately 60% of patients represented in this cohort were selective COX-2 inhibitor initiators. Approximately 2,000 cases of gastrointestinal complication were observed. This cohort comprised 19 continuous and binary covariates that were selected based on clinical importance as well as 50 binary covariates (representing concomitant medications, comorbidities and other medical encounters) selected by a high-dimensional propensity score (HDPS) algorithm [Schneeweiss et al., 2009]. The distribution of the pre-matched non-HDPS covariates is shown in Table 2.

Analyses

We applied the following 1:1 PSM techniques, generating risk ratio estimates (in this case, corresponding to the effect of non-selective NSAIDs, compared with COX-2 inhibitors, on the risk of gastrointestinal complications) and corresponding standard errors (and Wald 95% confidence intervals). To emulate previous analyses of these data, 1:1 PSM was performed with a nearest neighbor matching algorithm and a 0.025 absolute propensity score distance caliper [Ripollone et al., 2018]. Risk ratios were estimated using log binomial regression, unless otherwise noted.

(1) *Conventional 1:1 PSM*. Two models were generated. One model was based on maximum likelihood estimation, yielding the conventional standard error estimate. The other model was based on generalized estimating equations (GEE), yielding a robust standard error estimate designed to account for matched set correlation. We generated the second model since it has been demonstrated that this approach may lead to a better approximation of the sampling distribution of the effect estimate from a propensity score matched dataset compared with the approach that ignores matched set correlation [Austin, 2009; Austin, 2011].

Additionally, we calculated a standard error estimate using a simple contingency table-based formula that accounts for 1:1 matching [Rothman, 1986] so that we could compare the results of the model-based approaches to the results of the simplest possible approach for a matched cohort analysis. Letting the number of matched pairs for which the exposed unit and the unexposed unit experienced the outcome, for which only the exposed unit experienced the outcome, for which only the unexposed unit experienced the outcome and for which neither unit experienced the outcome be f_{11}, f_{10}, f_{01} and f_{00}, respectively, the formula for the standard error of the risk ratio is:

$$\sqrt{}[(f_{10} + f_{01}) / ((f_{11} + f_{10})*(f_{11} + f_{01}))].$$

For this simple approach, the risk ratio was generated via the corresponding contingency table-based formula [Rothman, 1986]:

$$(f_{11} + f_{10}) / (f_{11} + f_{01})$$

(2) *Simple bootstrap 1:1 PSM.* 1,000 bootstrap samples were generated.

(3) *Complex bootstrap 1:1 PSM*. 1,000 bootstrap samples were generated.

(4) *BPSM*. To emulate the approach taken by Kaplan and Chen [Kaplan and Chen, 2012], we generated a MCMC-based logistic regression (using the “MCMClogit” function in the R package, “MCMCpack”) model to estimate propensity scores. We used a non-informative uniform prior for the intercept parameter (i.e., a prior that has minimal influence on the estimation of the intercept parameter in the propensity score model) and the same independent normal prior for each slope (i.e., one corresponding to each covariate in dataset). The independent normal prior always had a mean of zero, but its variance was altered to determine the impact on the resulting standard error estimate.

Specifically, the variance was set at 0, 1, 1/10 and 1/100 (with lower values indicating more precise prior distributions). The MCMC procedure generated a total of 10,000 simulations (after 1,000 burn-in iterations). A thinning interval of 10 was applied after the burn-in to reduce the potential impact of auto-correlation among the simulations. After thinning, 1,000 simulations contributed to the propensity score model posterior distribution.

Results

We display the results of the empirical example analyses in Table 8. It was clear that, overall, the standard error estimates were similar among all methods. The non-Bayesian techniques always produced the largest standard error values (with the highest value seen in the simple bootstrap analysis: 0.111). BPSM produced very similar standard error estimates (each approximately 0.104). Thus, the different prior variance values for the MCMC-based propensity score model did not noticeably impact the precision of the final effect estimate. It did, however, impact the exposure effect estimate, with more precision in the prior distribution corresponding to estimated risk ratios closer to null. All risk ratio estimates from BPSM were closer to the null than were the adjusted risk ratio estimates from the non-Bayesian techniques (which were effectively the same).

Discussion

The results of our empirical example indicate that for our dataset, standard error estimates from all techniques were similar. Thus, for these data, accounting for the variability associated with propensity score estimation did not result in standard error estimates that were much different from the standard error estimate from the conventional application of 1:1 PSM.

The fact that the standard error estimates from the bootstrap techniques were similar to the standard error estimates from the conventional techniques is not surprising in light of the results of the simulation study conducted by Austin and Small [Austin and Small, 2014], and the results of an ongoing simulation study of the comparative performance of bootstrap procedures for 1:1 PSM similar to those demonstrated here using datasets similar to our example dataset with respect to size and number of covariates [Desai et al., 2019].

In their simulation study, Austin and Small [Austin and Small, 2014] demonstrated that the mean ratio of estimated standard error to the standard deviation of simulated exposure effects tended to be smaller for simple bootstrap 1:1 PSM, and closer to the value from the scenario accounting for matched pair correlation (as in our example), than for complex bootstrap 1:1 PSM. The authors noted that this finding might make simple bootstrap 1:1 PSM preferable to complex bootstrap 1:1 PSM in practice. They concluded that a method for estimating the standard error of the effect estimate that accounts for matched pair correlation may be best in most applications (even though such a method also treats the propensity score as a fixed quantity, rather than an estimate that also is subject to sampling variability), but that the simple bootstrap should be considered for the rare case in which a parametric estimator for the standard error is unavailable (i.e., in a scenario that relies on a more complex PSM approach such as “double propensity score matching” [Austin, 2017]) [Austin and Small, 2014]. However, the preference for a method that accounts for matched pair correlation would seem to have made no difference in our data, given the similarity of the results between the robust log binomial model and the maximum likelihood-based log binomial.

The variance of the prior distribution for the propensity score model in our BPSM analyses had no notable influence on the standard error estimate. The fact that the standard error estimates from BPSM always were slightly smaller than the standard error estimates from the non-Bayesian approaches was surprising in light of the results of the simulation analysis performed by Kaplan and Chen [Kaplan and Chen, 2012], which indicated that BPSM tended to yield larger and more accurate standard error estimates compared with the standard error estimates from the conventional PSM approach. However, the simulated datasets in the Kaplan and Chen study were much smaller and comprised far fewer covariates (with scenarios using no more than 300 units and 3 covariates) than what would typically be encountered in pharmacoepidemiology [Patorno et al., 2013; Patorno et al., 2014]. Moreover, Kaplan and Chen note that BPSM may be preferable for such small-data scenarios. The preference for small-data scenarios might also explain why the risk ratio estimates from BPSM in our example were so different from the risk ratio estimates from the non-Bayesian results (although we note again that, unlike for the non-Bayesian techniques, the approach to effect estimation for BPSM involved averaging over effect estimates based on the MCMC posterior distribution). Thus, it is unclear how applicable the previous simulation findings are to typical claims scenarios.

A key point to consider is that it is difficult to predict whether incorporation of the variability in propensity score estimation into the standard error estimate from 1:1 PSM will lead to a higher (which may be the intuitive prediction, and which was seen in the BPSM analysis) or lower standard error estimates. This is the case because of the general complexity of the sampling distribution of effect estimates in PSM and because matching on the *estimated* propensity score may be more statistically efficient (leading to lower standard errors) than matching on the *true* propensity score [Austin and Small, 2014; Abadie and Imbens, 2016]. Since one cannot know whether the estimated propensity scores for the pre-matched sample actually are the true propensity scores for the pre-matched sample, one cannot predict how large the standard error estimates will be, even with methods that incorporate the uncertainty due to propensity score estimation in 1:1 PSM.

Our review provides an accessible depiction of sampling variability for the specific case of 1:1 PSM without replacement as well as the methods that attempt to account for this sampling variability. Our example scenario indicates that these methods may not produce appreciably different standard errors, compared with conventional 1:1 PSM methods, for claims-based analyses similar to ours. This could imply that the simplest approach to assessing the standard error estimate (e.g., use of the simple, contingency table-based formula) might be the best approach in these cases. However, additional simulation studies might be warranted to confirm these findings. Specifically, even though our results conform with the findings of previous simulation studies of the bootstrap techniques, more work may be required to determine the utility of BPSM for the typical claims-based study, compared with the other approaches.

TABLES

Table 1. Simple Example of the Propensity Score Matching Paradox

Sex and Race[a]	Index Exposure (n)	Reference Exposure (n)	Total	Stratum PS
Male				
White	1	2	3	0.3
Not white	1	2	3	0.3
Female				
White	1	2	3	0.3
Not white	1	2	3	0.3

Abbreviations: PS = Propensity Score

[a] The population represented in this table contains index and reference exposure groups that are perfectly balanced on sex and race. The propensity score values for all 12 units are equal. 1:1 propensity score matching without replacement would be expected to increase the underlying covariate imbalance in the matched dataset, compared to the pre-matched dataset.

Table 2. Example Distributions of the Non-High Dimensional Propensity Score Covariates in the Pre-matched Pharmaceutical Assistance Contract for the Elderly (1999-2002), Standard Covariate Set, Original Index Exposure Prevalence Dataset and in the Three Corresponding Fully-matched Datasets

	Pre-matched (n = 49,653)				Full, NNM		Full, DGM		Full, MDM	
Covariate	Non-selective NSAIDs (n = 17,611)[a]		Selective COX-2 Inhibitors (n = 32,042)		Selective COX-2 Inhibitors (n = 17,611)		Selective COX-2 Inhibitors (n = 17,611)		Selective COX-2 Inhibitors (n = 17,611)	
	Mean (SD)	%	Mean (SD)	%	Mean (SD)	%	Mean (SD)	%	Mean (SD)	%
Age	77.79 (7.30)		79.76 (7.24)		78.15 (7.24)		78.16 (7.23)		78.95 (7.06)	
Generics	7.43 (5.02)		8.41 (5.25)		7.56 (5.02)		7.60 (5.03)		6.75 (4.17)	
Any Medical Visit	7.74 (6.61)		8.60 (6.67)		7.86 (6.53)		7.90 (6.59)		6.96 (5.32)	
Charlson Comorbidity	1.85 (1.97)		2.05 (2.01)		1.85 (1.95)		1.87 (1.96)		1.47 (1.58)	
Male		18.84		14.09		17.47		17.50		13.43
Race										
White		89.76		95.45		92.94		92.86		94.61
Black		8.97		3.54		5.91		5.96		4.15
Other		1.27		1.02		1.15		1.19		1.24
Comorbidities										
Bleeding		1.11		1.72		1.15		1.25		1.08
CHF		24.58		30.36		24.76		25.17		18.80
Coronary Disease		14.78		16.43		14.89		14.87		9.60
Hypertension		70.20		72.82		70.18		70.29		70.82
Rheumatoid Arthritis		2.70		5.00		3.02		2.84		2.54
Osteoarthritis		33.49		48.53		35.16		35.01		41.23
Ulcer		2.42		3.71		2.58		2.58		2.14
Hospitalization in Prior Year		26.07		30.60		26.47		26.90		17.86
Nursing Home Resident		5.66		8.34		6.18		6.23		3.64
Other Medications										
Corticosteroid		7.80		8.74		8.08		8.17		5.48

Covariate	Pre-matched (n = 49,653)				Full, NNM		Full, DGM		Full, MDM	
	Non-selective NSAIDs (n = 17,611)[a]		Selective COX-2 Inhibitors (n = 32,042)		Selective COX-2 Inhibitors (n = 17,611)		Selective COX-2 Inhibitors (n = 17,611)		Selective COX-2 Inhibitors (n = 17,611)	
	Mean (SD)	%	Mean (SD)	%	Mean (SD)	%	Mean (SD)	%	Mean (SD)	%
Other Gastrointestinal Medication		20.44		27.42		21.70		21.75		20.28
Warfarin		6.55		13.27		7.00		7.02		5.95
Year of Exposure Initiation										
1999		48.79		41.68		47.09		47.11		43.21
2000		23.91		29.94		24.90		24.79		29.10
2001		20.00		21.28		20.49		20.73		21.08
2002		7.30		7.09		7.52		7.38		6.62

Abbreviations: DGM = Propensity score digit-based greedy Matching; MDM = Mahalanobis distance matching; NNM = Propensity score nearest neighbor matching

[a] The Non-selective NSAIDs covariate distribution is shown only once, since this distribution was the same in each dataset

Table 3. Example Distributions (%) of all Covariates in the Pre-matched Medicaid Analytic eXtract (United States, 2000-2007), Original Index Exposure Prevalence Dataset and in the Three Corresponding Fully-matched Datasets

Covariate	Pre-matched (n = 886,996)		Full, NNM	Full, DGM	Full, MDM
	Statins (n = 1,152)[a]	No Statins (n = 885,844)	No Statins (n = 1,152)	No Statins (n = 1,152)	No Statins (n = 1,152)
Age Categories					
≤ 19	5.56	29.43	5.21	4.25	5.21
20–24	14.06	35.6	12.76	14.41	14.24
25–29	21.09	20.41	21.96	22.31	22.74
30–34	28.13	9.48	28.91	28.65	27.34
35–39	22.22	4.17	21.96	21.61	21.53
≥ 40	8.94	0.91	9.20	8.77	8.94
Race					
Asian/Other Pacific Islander	6.51	3.42	6.42	5.90	5.38
Black/African American	25.69	34.09	22.92	24.31	27.95
Hispanic/Latino	17.10	15.08	21.09	17.88	17.88
Other	5.73	4.74	6.08	7.47	4.86
Unknown	2.95	2.01	3.39	3.21	2.78
White	42.01	40.67	40.10	41.23	41.15
U.S. Region					
Midwest	23.18	32.02	22.48	20.92	24.39
Northeast	21.27	14.97	20.57	22.83	18.75
South	26.04	26.07	24.13	26.13	26.48
West	29.51	26.94	32.81	30.12	30.38
Number of Non-antihypertensive Generics Used					
None	8.33	46.45	6.25	7.29	10.76
1–3	27.00	36.64	30.30	28.39	28.21
> 3	64.67	16.91	63.45	64.32	61.02

Covariate	Pre-matched (n = 886,996)		Full, NNM	Full, DGM	Full, MDM
	Statins (n = 1,152)[a]	No Statins (n = 885,844)	No Statins (n = 1,152)	No Statins (n = 1,152)	No Statins (n = 1,152)
Number of Physician Visits During the Pre-index Period					
None	27.08	52.07	25.78	25.52	25.87
1–3	49.91	39.52	51.82	51.48	53.39
> 3	23.00	8.41	22.40	23.00	20.75
Year of Delivery					
2000	0.00	0.14	0.00	0.00	0.00
2001	4.17	9.65	4.51	4.25	3.39
2002	5.56	11.04	6.34	4.77	6.34
2003	10.42	14.59	10.33	9.72	10.33
2004	19.10	17.61	18.23	18.92	17.36
2005	20.14	16.88	20.23	20.23	20.31
2006	23.78	17.49	21.18	24.05	24.74
2007	16.84	12.60	19.18	18.06	17.53
Comorbidities					
Hypertension	40.63	5.00	39.76	40.97	40.02
Diabetes	45.14	3.06	40.71	41.75	45.14
Renal Disease	4.17	0.46	3.91	3.82	4.17
Obesity	23.35	5.31	23.87	25.26	23.35
Tobacco Use	11.02	7.77	10.16	11.11	8.85
Alcohol Abuse	3.99	2.61	4.60	4.69	3.13
Illicit Drug Use	6.42	5.33	6.60	6.68	5.38
Dyslipidemia	67.10	3.14	71.09	71.53	66.58
Multiple Gestation	6.60	3.55	6.16	7.03	5.64
Multipara	88.80	75.69	88.02	88.54	92.01
Other Medications					
Insulin	30.47	1.24	26.30	25.95	30.47

Covariate	Pre-matched (n = 886,996)		Full, NNM	Full, DGM	Full, MDM
	Statins (n = 1,152)[a]	No Statins (n = 885,844)	No Statins (n = 1,152)	No Statins (n = 1,152)	No Statins (n = 1,152)
Antidiabetic	38.80	1.27	33.94	34.29	38.80
Hypertension Medication	53.73	6.65	52.52	50.95	52.78
Potentially Teratogenic Medication	31.68	3.63	29.08	28.47	30.30

Abbreviations: DGM = Propensity score digit-based greedy Matching; MDM = Mahalanobis distance matching; NNM = Propensity score nearest neighbor matching

[a] The Statins covariate distribution is shown only once, since this distribution was the same in each dataset

Table 4. Summary of Plasmode Simulation Scenarios

Exposure Prevalence[a]	Covariate Set[b]	Product Term[c] Estimate Strength
0.05	Small	—
	Standard	—
	Large	—
0.10	Small	—
	Standard	—
	Large	—
0.20	Small	—
		Default
		Exaggerated
	Standard	—
	Large	—
0.30	Small	—
	Standard	—
	Large	—
0.40	Small	—
	Standard	—
	Large	—

[a] All plasmode scenarios were based on the NSAID cohort.

[b] The "Small" set comprised 19 pre-determined covariates; the "Standard" and "Large" sets comprised an additional 50 and 100 covariates, respectively, selected from a high-dimensional propensity score algorithm.

[c] The product term represented the interaction between age and Charlson comorbidity score. The "default" scenario maintained the original product term and the "exaggerated" scenario was based on a product term that was 200% greater than the default product term.

Table 5. Real Dataset Analysis Results

Original Dataset	Method	Number of Units Analyzed	Number of Outcomes Analyzed	RR	95% CI	95% CI Width[a]	MB
NSAID, Small	Crude	49,653	552	0.92	—	—	0.558
	CEM	16,139	106	1.68	[1.09, 2.58]	2.36	0.017
	PSM	34,150	355	1.05	[0.86, 1.29]	1.51	0.089
	MDM	35,222	361	1.05	[0.86, 1.29]	1.51	0.207
	FS	49,634	552	1.08	[0.90, 1.31]	1.45	0.026
NSAID, Standard	Crude	49,653	552	0.92	—	—	0.641
	CEM	3,226	10	2.55	[0.64, 10.09]	15.73	0.014
	PSM	33,368	339	1.12	[0.90, 1.38]	1.53	0.087
	MDM	35,222	318	1.39	[1.11, 1.74]	1.56	0.541
	FS	49,626	552	1.12	[0.93, 1.36]	1.47	0.051
NSAID, Large	Crude	49,653	552	0.92	—	—	0.654
	CEM	1,763	6	1.71	[0.31, 9.48]	30.58	0.020
	PSM	33,174	340	1.09	[0.88, 1.34]	1.53	0.089
	MDM	35,222	309	1.49	[1.19, 1.87]	1.57	0.681
	FS	49,626	552	1.12	[0.92, 1.37]	1.48	0.057
Statin	Crude	886,996	31,489	1.79	—	—	5.127
	CEM	11,321	307	1.13	[0.54, 2.36]	4.35	0.000
	PSM	2,302	144	1.03	[0.75, 1.41]	1.88	1.632
	MDM	2,304	147	0.99	[0.72, 1.35]	1.87	0.244
	FS	809,732	29,072	1.03	[0.82, 1.31]	1.60	0.586

Abbreviations: CEM = Coarsened exact matching; CI = Confidence interval; FS = Fine stratification on the propensity score; MB = Mahalanobis balance; MDM = Mahalanobis distance matching; PSM = Propensity score matching; RR = Risk ratio.

[a] The 95% CI width was calculated by dividing the upper 95% CI endpoint by the lower 95% CI endpoint (using all available digits).

Table 6. Plasmode Analysis Results, Small Covariate Set, 20% Index Exposure Prevalence Interaction Scenarios – All Simulation Metrics

Scenario[a,b]	Method	Bias	Variance	Square Root of MSE	AMB
Default	Crude	-0.103	—	—	—
	CEM	0.327	0.226	0.577	0.967
	PSM	0.067	0.040	0.210	0.886
	MDM	0.131	0.041	0.242	0.792
	FS	0.070	0.027	0.178	0.946
Exaggerated	Crude	-0.091	—	—	—
	CEM	0.341	0.220	0.580	0.967
	PSM	0.079	0.040	0.214	0.886
	MDM	0.143	0.038	0.242	0.792
	FS	0.080	0.023	0.172	0.946

Abbreviations: AMB = Average proportion decrease in Mahalanobis balance; CEM = Coarsened exact matching; CI = Confidence interval; FS = Fine stratification on the propensity score; MDM = Mahalanobis distance matching; PSM = Propensity score matching.

[a] The product term represented the interaction between age and Charlson comorbidity score.

[b] The “default” scenario maintained the original product term and the “exaggerated” scenario was based on a product term that was 200% greater than the default product term.

Table 7. Plasmode Analysis Results, Small Covariate Set, 20% Index Exposure Prevalence Interaction Scenarios – Absolute Differences between Index-exposed and Reference-exposed Groups with Respect to the Average of the Average Age, Within Each Coarsened Category of Charlson Comorbidity Score, and Vice Versa, Across the Plasmode Simulations; Default Scenario Only

Difference in Average	Original	CEM[a]	PSM	MDM	FS[a]
Average Age					
Within Score 0	2.09	0.05	0.38	0.47	0.28
Within Score 1	2.03	0.11	0.25	0.42	0.15
Within Score 2	1.66	0.11	0.06	0.45	0.14
Within Score 3	1.93	0.01	0.16	0.94	0.07
Within Score ≥ 4	1.43	0.02	0.25	0.99	0.31
Average Score					
Within Age < 70	0.16	0.00	0.06	0.28	0.04
Within Age 70-74	0.23	0.00	0.07	0.22	0.06
Within Age 75-79	0.15	0.00	0.01	0.15	0.01
Within Age 80-84	0.20	0.01	0.05	0.01	0.05
Within Age 85-89	0.06	0.01	0.12	0.15	0.12
Within Age 90-94	0.01	0.01	0.12	0.10	0.12
Within Age ≥ 95	0.18	0.00	0.00	0.04	0.00

Abbreviations: CEM = Coarsened exact matching; FS = Fine stratification on the propensity score; MDM = Mahalanobis distance matching; PSM = Propensity score matching; Score = Charlson Comorbidity Score.

[a] The average age and average score values were weighted (at the unit level) for the CEM and FS scenarios.

Table 8. Results of the Example 1:1 Propensity Score Matching Without Replacement Analyses Using the Pharmaceutical Assistance Contract for the Elderly (1999-2002) Dataset

1:1 PSM Method	Adjusted RR[a,b]	SE (ln[RR])	95% CI	95% CI Width[c]
Non-Bayesian				
Conventional, Simple SE Formula	1.13	0.109	[0.92, 1.40]	1.53
Conventional, Log Binomial	1.13	0.109	[0.92, 1.40]	1.53
Robust SE Log Binomial	1.13	0.109	[0.92, 1.40]	1.53
Simple Bootstrap	1.13	0.111	[0.91, 1.41]	1.55
Complex Bootstrap	1.13	0.106	[0.92, 1.40]	1.52
Bayesian (BPSM)				
$\boldsymbol{\beta}_{\text{prior}}$ ~ N(0, 0)	1.04	0.104	[0.84, 1.27]	1.51
$\boldsymbol{\beta}_{\text{prior}}$ ~ N(0, 1)	1.03	0.104	[0.84, 1.27]	1.51
$\boldsymbol{\beta}_{\text{prior}}$ ~ N(0, 1/10)	1.03	0.104	[0.84, 1.26]	1.50
$\boldsymbol{\beta}_{\text{prior}}$ ~ N(0, 1/100)	1.01	0.104	[0.82, 1.23]	1.50

Abbreviations: BPSM = Bayesian 1:1 propensity score matching without replacement; CI = Confidence interval; 1:1 PSM = 1:1 propensity score matching without replacement; SE = Standard error; RR = Risk ratio.

[a] The crude RR was 0.92.

[b] The adjusted RR for the Conventional, Simple SE Formula was based on the corresponding simple contingency table-based formula. The same maximum likelihood-based adjusted RR estimate was used for the Conventional, Log Binomial, Simple Bootstrap and Complex Bootstrap scenarios. The adjusted RR for the Robust SE Log Binomial scenario was based on a generalized estimating equations model. The adjusted RR for the Bayesian scenarios was the exponent of the average value adjusted estimate across the matched datasets (generated from the MCMC-based logistic propensity score distribution).

[c] The 95% CI width was calculated by dividing the upper 95% CI endpoint by the lower 95% CI endpoint (using all available digits).

FIGURES

Figure 1. Design Flowchart for the 36 Dataset Scenarios

The Matching/Distance Metric and Matching Algorithm branches apply to each of the Index Exposure Prevalence branches. The former branches were collapsed for the sake of efficiency.

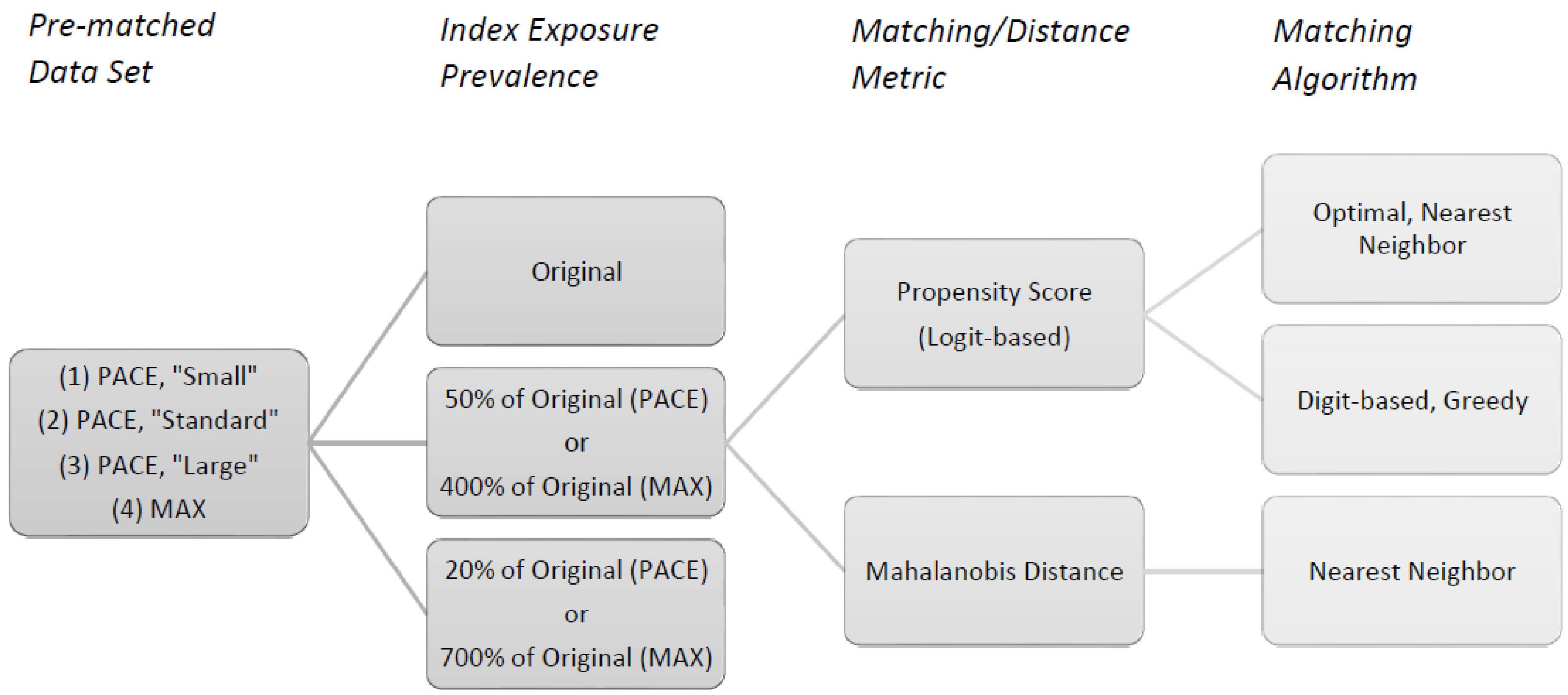


Figure 2. Forest Plots of Standardized Differences Among All Covariates for the PACE, Standard Covariate Set, Original Index Exposure Prevalence Dataset

A) Fully-matched datasets. B) Matched datasets pruned to the sample size of the pruned dataset that first met the propensity score nearest neighbor matching 0.025 absolute propensity score distance caliper. The black, red, green and blue markers correspond to the covariates from the original/pre-matched dataset, from the propensity score nearest neighbor-matched dataset, from the propensity score digit-based greedy-matched dataset and from the Mahalanobis distance-matched dataset, respectively.

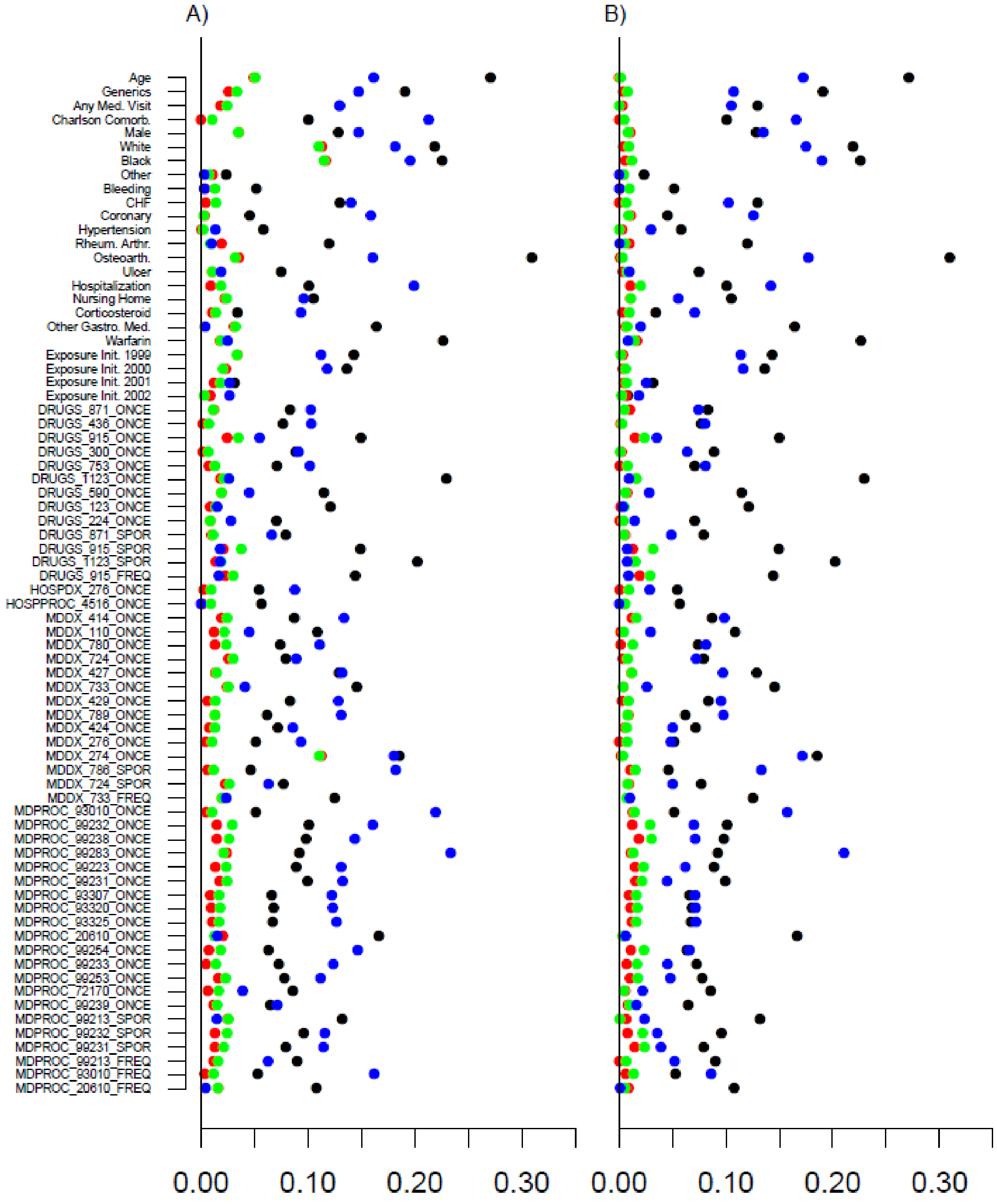

A)
B)
Age
Generics
Any Med. Visit
Charlson Comorb.
Male
White
Black
Other
Bleeding
CHF
Coronary
Hypertension
Rheum. Arthr.
Osteoarth.
Ulcer
Hospitalization
Nursing Home
Corticosteroid
Other Gastro. Med.
Warfarin
Exposure Init. 1999
Exposure Init. 2000
Exposure Init. 2001
Exposure Init. 2002
DRUGS_871_ONCE
DRUGS_436_ONCE
DRUGS_915_ONCE
DRUGS_300_ONCE
DRUGS_753_ONCE
DRUGS_T123_ONCE
DRUGS_590_ONCE
DRUGS_123_ONCE
DRUGS_224_ONCE
DRUGS_871_SPOR
DRUGS_915_SPOR
DRUGS_T123_SPOR
DRUGS_915_FREQ
HOSPDX_276_ONCE
HOSPPROC_4516_ONCE
MDDX_414_ONCE
MDDX_110_ONCE
MDDX_780_ONCE
MDDX_724_ONCE
MDDX_427_ONCE
MDDX_733_ONCE
MDDX_429_ONCE
MDDX_789_ONCE
MDDX_424_ONCE
MDDX_276_ONCE
MDDX_274_ONCE
MDDX_786_SPOR
MDDX_724_SPOR
MDDX_733_FREQ
MDPROC_93010_ONCE
MDPROC_99232_ONCE
MDPROC_99238_ONCE
MDPROC_99283_ONCE
MDPROC_99223_ONCE
MDPROC_99231_ONCE
MDPROC_93307_ONCE
MDPROC_93320_ONCE
MDPROC_93325_ONCE
MDPROC_20610_ONCE
MDPROC_99254_ONCE
MDPROC_99233_ONCE
MDPROC_99253_ONCE
MDPROC_72170_ONCE
MDPROC_99239_ONCE
MDPROC_99213_SPOR
MDPROC_99232_SPOR
MDPROC_99231_SPOR
MDPROC_99213_FREQ
MDPROC_93010_FREQ
MDPROC_20610_FREQ
0.00
0.10
0.20
0.30
0.00
0.10
0.20
0.30

Figure 3. Forest Plots of Standardized Differences Among All Covariates for the MAX, Original Index Exposure Prevalence Dataset

A) Fully-matched datasets. B) Matched datasets pruned to the sample size of the pruned dataset that first met the propensity score nearest neighbor matching 0.025 absolute propensity score distance caliper. The black, red, green and blue markers correspond to the covariates from the original/pre-matched dataset, from the propensity score nearest neighbor-matched dataset, from the propensity score digit-based greedy-matched dataset and from the Mahalanobis distance-matched dataset, respectively.

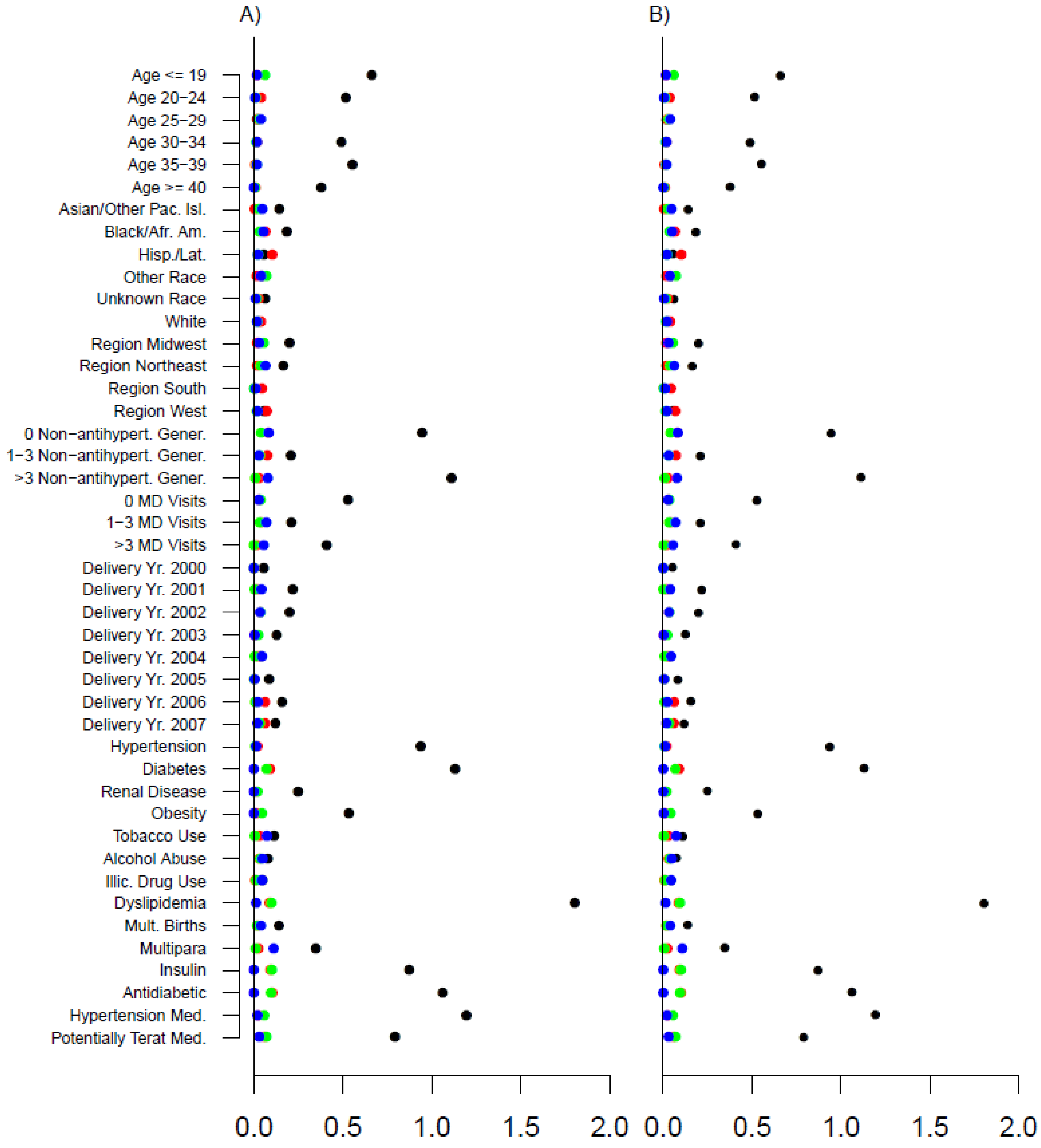

A)
B)
Age <= 19
Age 20−24
Age 25−29
Age 30−34
Age 35−39
Age >= 40
Asian/Other Pac. Isl.
Black/Afr. Am.
Hisp./Lat.
Other Race
Unknown Race
White
Region Midwest
Region Northeast
Region South
Region West
0 Non−antihypert. Gener.
1−3 Non−antihypert. Gener.
>3 Non−antihypert. Gener.
0 MD Visits
1−3 MD Visits
>3 MD Visits
Delivery Yr. 2000
Delivery Yr. 2001
Delivery Yr. 2002
Delivery Yr. 2003
Delivery Yr. 2004
Delivery Yr. 2005
Delivery Yr. 2006
Delivery Yr. 2007
Hypertension
Diabetes
Renal Disease
Obesity
Tobacco Use
Alcohol Abuse
Illic. Drug Use
Dyslipidemia
Mult. Births
Multipara
Insulin
Antidiabetic
Hypertension Med.
Potentially Terat Med.
0.0
0.5
1.0
1.5
2.0
0.0
0.5
1.0
1.5
2.0

Figure 4. Mahalanobis Balance Metric Trends for the 9 PACE Datasets

A) Small covariate set, original index exposure prevalence (IEP). B) Small covariate set, 50% of IEP. C) Small covariate set, 20% of IEP. D) Standard covariate set, IEP. E) Standard covariate set, 50% of IEP. F) Standard covariate set, 20% of IEP. G) Large covariate set, IEP. H) Large covariate set, 50% of IEP. I) Large covariate set, 20% of IEP. The black dots indicate the Mahalanobis balance values of the pre-matched datasets. Red lines indicate propensity score nearest neighbor matching trends, green lines indicate propensity score digit-based greedy matching trends and blue lines indicate Mahalanobis distance matching trends. The dotted and dashed vertical lines (for propensity score nearest neighbor matching and propensity score digit-based greedy matching, respectively) mark the 6 points at which the propensity score matching trends first met the 0.05, 0.025 and 0.01 absolute propensity score distance caliper criteria (vertical line colors correspond to trend colors). The caliper criteria always were met in the order 0.05, 0.025, 0.01 during the pruning process.

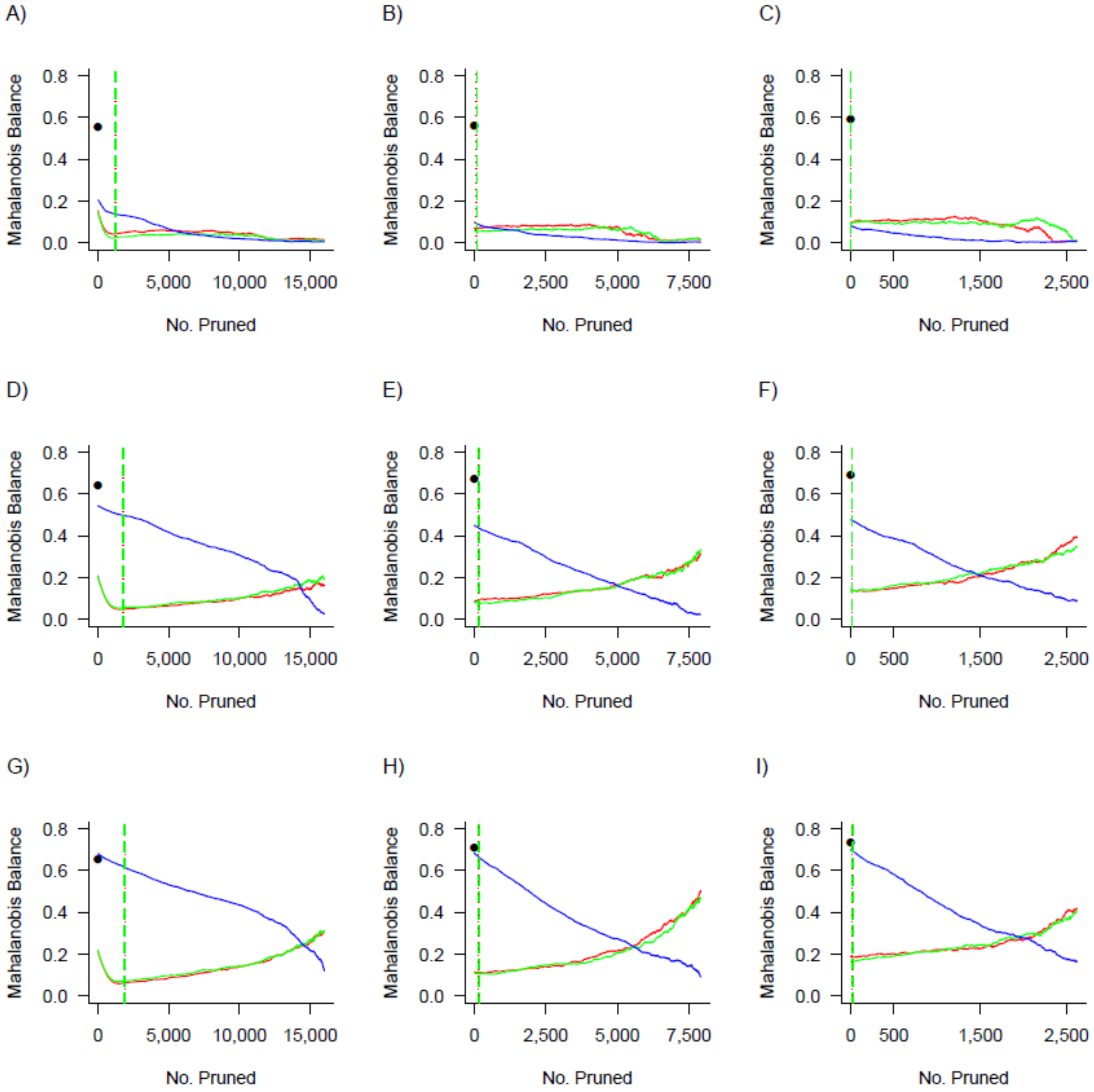

A)
B)
C)
D)
E)
F)
G)
H)
I)
Mahalanobis Balance
0.0
0.2
0.4
0.6
0.8
No. Pruned
0
5,000
10,000
15,000
2,500
7,500
500
1,500

Figure 5. Mahalanobis Balance Metric Trends for the 3 MAX Datasets

A) Original index exposure prevalence (IEP). B) 400% of IEP. C) 700% of IEP. The black dots indicate the Mahalanobis balance values of the pre-matched datasets. Red lines indicate propensity score nearest neighbor matching trends, green lines indicate propensity score digit-based greedy matching trends and blue lines indicate Mahalanobis distance matching trends. The dotted and dashed vertical lines (for propensity score nearest neighbor matching and propensity score digit-based greedy matching, respectively) mark the 6 points at which the propensity score matching trends first met the 0.05, 0.025 and 0.01 absolute propensity score distance caliper criteria (vertical line colors correspond to trend colors). The caliper criteria always were met in the order 0.05, 0.025, 0.01 during the pruning process.

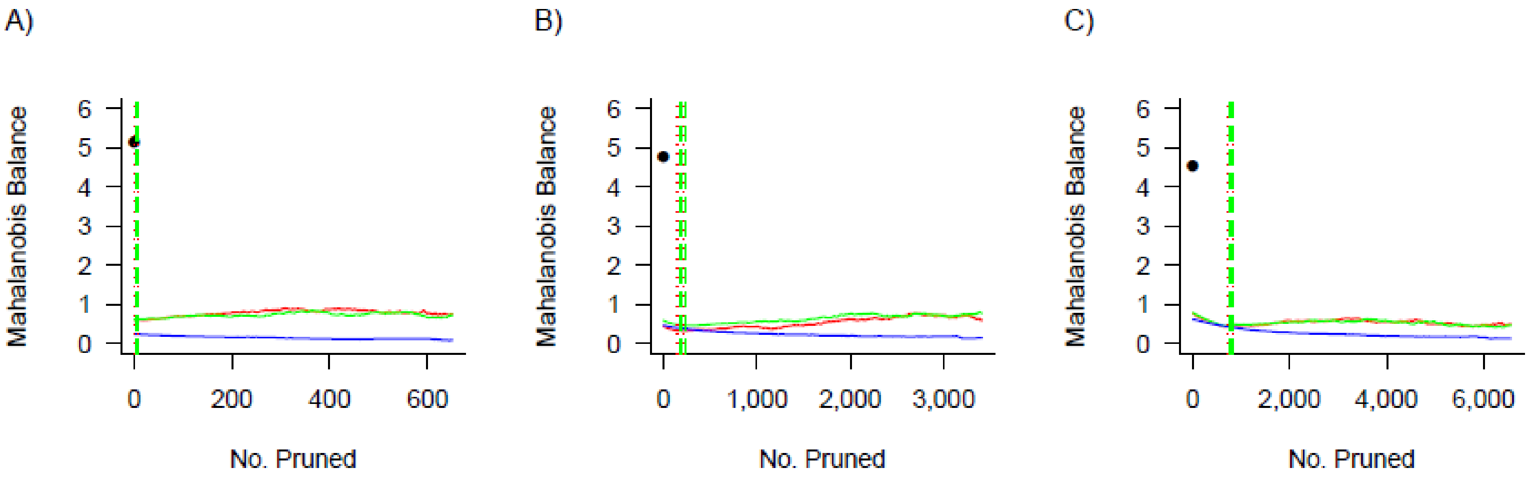


Figure 6. C-statistic Metric Trends for the 9 PACE Datasets

A) Small covariate set, original index exposure prevalence (IEP). B) Small covariate set, 50% of IEP. C) Small covariate set, 20% of IEP. D) Standard covariate set, IEP. E) Standard covariate set, 50% of IEP. F) Standard covariate set, 20% of IEP. G) Large covariate set, IEP. H) Large covariate set, 50% of IEP. I) Large covariate set, 20% of IEP. The black dots indicate the Mahalanobis balance values of the pre-matched datasets. Red lines indicate propensity score nearest neighbor matching trends, green lines indicate propensity score digit-based greedy matching trends and blue lines indicate Mahalanobis distance matching trends. The dotted and dashed vertical lines (for propensity score nearest neighbor matching and propensity score digit-based greedy matching, respectively) mark the 6 points at which the propensity score matching trends first met the 0.05, 0.025 and 0.01 absolute propensity score distance caliper criteria (vertical line colors correspond to trend colors). The caliper criteria always were met in the order 0.05, 0.025, 0.01 during the pruning process.

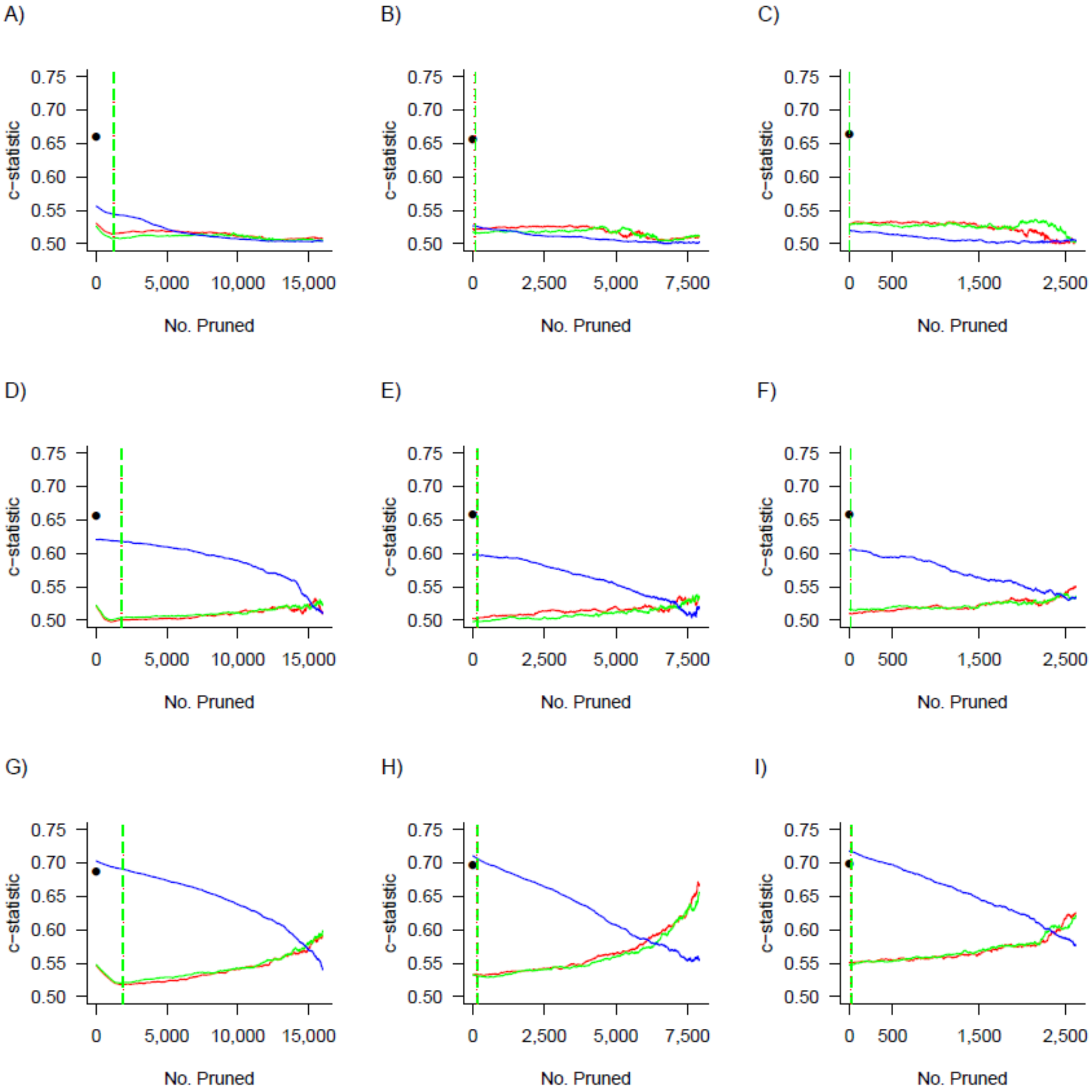

A)
B)
C)
D)
E)
F)
G)
H)
I)
c−statistic
0.75
0.70
0.65
0.60
0.55
0.50
No. Pruned
0
5,000
10,000
15,000
2,500
7,500
500
1,500

Figure 7. C-statistic Metric Trends for the 3 MAX Datasets

A) Original index exposure prevalence (IEP). B) 400% of IEP. C) 700% of IEP. The black dots indicate the Mahalanobis balance values of the pre-matched datasets. Red lines indicate propensity score nearest neighbor matching trends, green lines indicate propensity score digit-based greedy matching trends and blue lines indicate Mahalanobis distance matching trends. The dotted and dashed vertical lines (for propensity score nearest neighbor matching and propensity score digit-based greedy matching, respectively) mark the 6 points at which the propensity score matching trends first met the 0.05, 0.025 and 0.01 absolute propensity score distance caliper criteria (vertical line colors correspond to trend colors). The caliper criteria always were met in the order 0.05, 0.025, 0.01 during the pruning process.

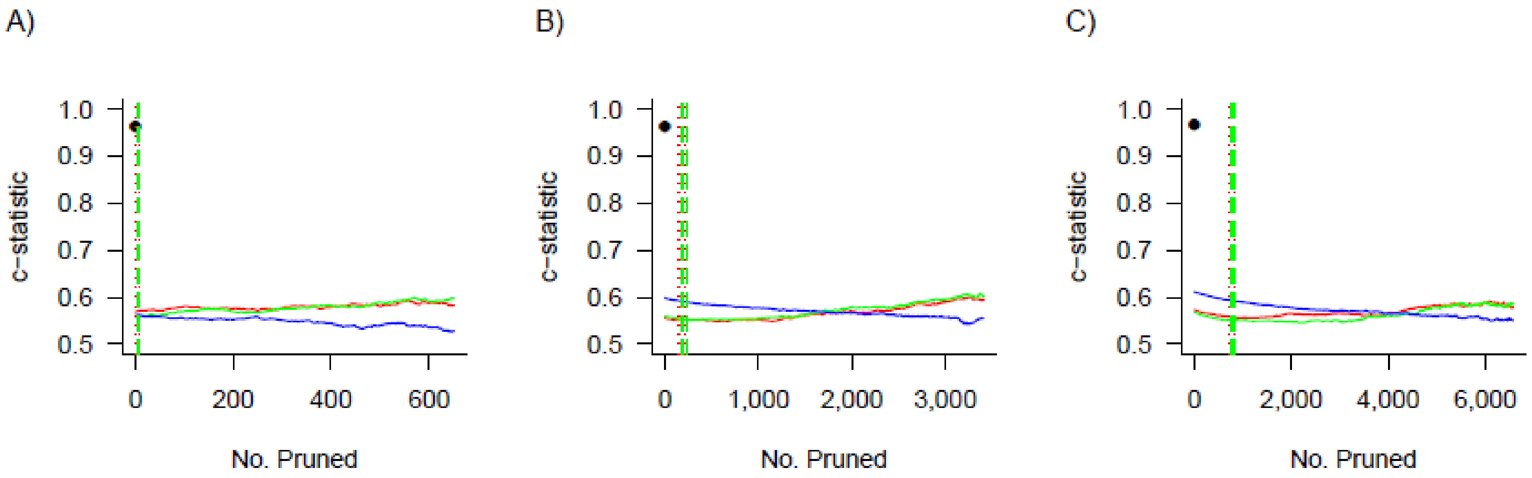


Figure 8. Zoomed-in Version of Figure 4

A) Small covariate set, original index exposure prevalence (IEP). B) Small covariate set, 50% of IEP. C) Small covariate set, 20% of IEP. D) Standard covariate set, IEP. E) Standard covariate set, 50% of IEP. F) Standard covariate set, 20% of IEP. G) Large covariate set, IEP. H) Large covariate set, 50% of IEP. I) Large covariate set, 20% of IEP. The ranges of the “No. Pruned” axes in these panels are much smaller than the ranges of the corresponding panels in Figure 1. The black dots indicate the Mahalanobis balance values of the pre-matched datasets. Red lines indicate propensity score nearest neighbor matching trends, green lines indicate propensity score digit-based greedy matching trends and blue lines indicate Mahalanobis distance matching trends. The dotted and dashed vertical lines (for propensity score nearest neighbor matching and propensity score digit-based greedy matching, respectively) mark the 6 points at which the propensity score matching trends first met the 0.05, 0.025 and 0.01 absolute propensity score distance caliper criteria (vertical line colors correspond to trend colors). The caliper criteria always were met in the order 0.05, 0.025, 0.01 during the pruning process.

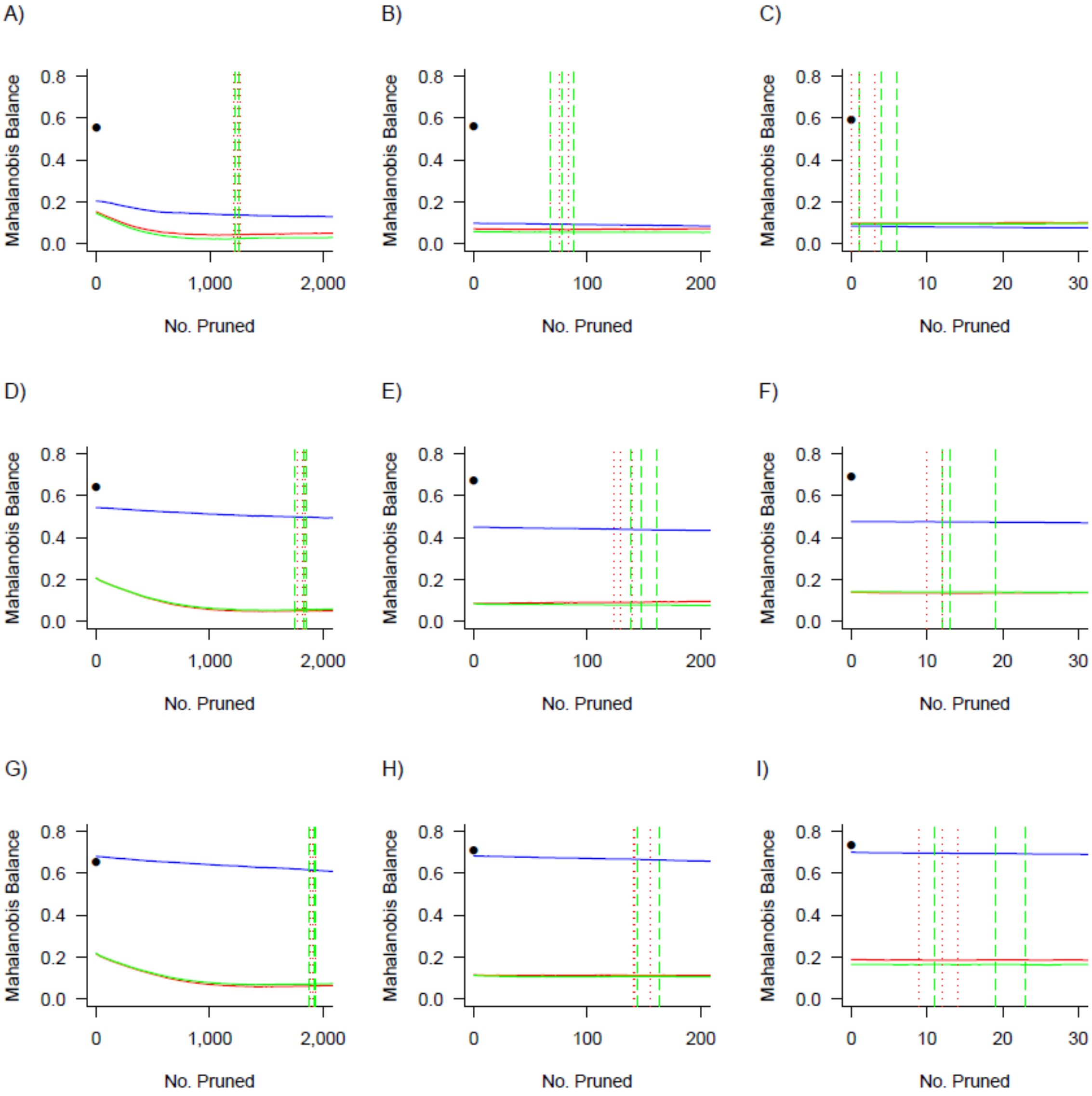

A)
B)
C)
D)
E)
F)
G)
H)
I)
Mahalanobis Balance
0.8
0.6
0.4
0.2
0.0
0
1,000
2,000
100
200
10
20
30
No. Pruned

Figure 9. Zoomed-in Version of Figure 5

A) Original index exposure prevalence (IEP). B) 400% of IEP. C) 700% of IEP. The ranges of the “No. Pruned” axes in these panels are much smaller than the ranges of the corresponding panels in Figure 1. The black dots indicate the Mahalanobis balance values of the pre-matched datasets. Red lines indicate propensity score nearest neighbor matching trends, green lines indicate propensity score digit-based greedy matching trends and blue lines indicate Mahalanobis distance matching trends. The dotted and dashed vertical lines (for propensity score nearest neighbor matching and propensity score digit-based greedy matching, respectively) mark the 6 points at which the propensity score matching trends first met the 0.05, 0.025 and 0.01 absolute propensity score distance caliper criteria (vertical line colors correspond to trend colors). The caliper criteria always were met in the order 0.05, 0.025, 0.01 during the pruning process.

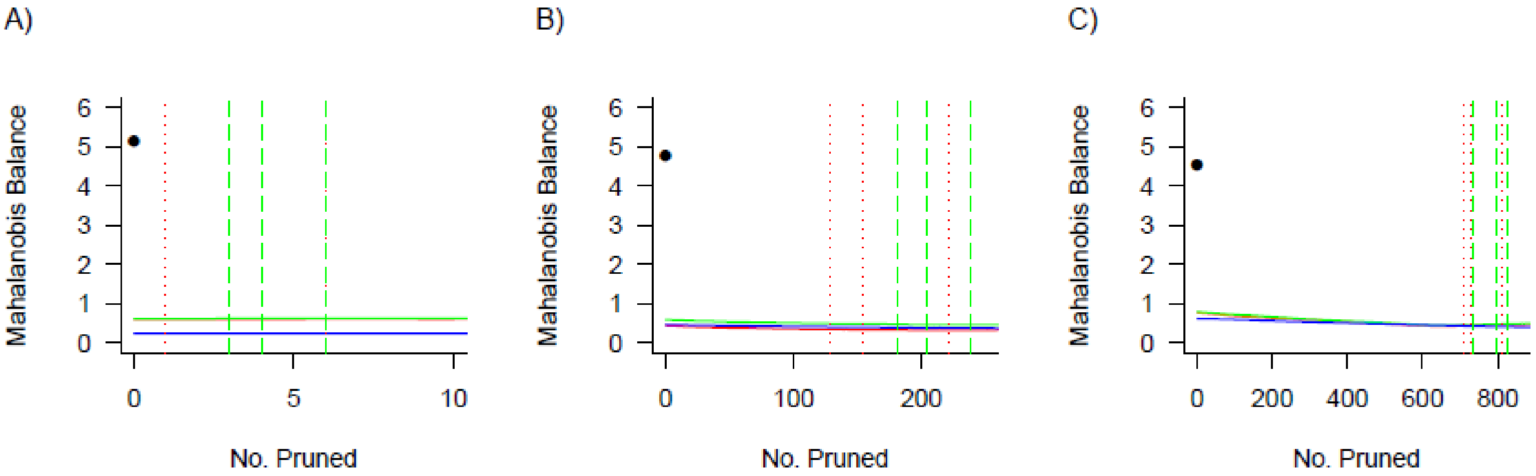


Figure 10. Relative Risk Estimate Trends for the 9 PACE Datasets

A) Small covariate set, original index exposure prevalence (IEP). B) Small covariate set, 50% of IEP. C) Small covariate set, 20% of IEP. D) Standard covariate set, IEP. E) Standard covariate set, 50% of IEP. F) Standard covariate set, 20% of IEP. G) Large covariate set, IEP. H) Large covariate set, 50% of IEP. I) Large covariate set, 20% of IEP. A dashed horizontal black line at the relative risk estimate value of 1.00 is included for reference. The black dots indicate the relative risk estimates of the pre-matched datasets. Red lines indicate propensity score nearest neighbor matching trends, green lines indicate propensity score digit-based greedy matching trends and blue lines indicate Mahalanobis distance matching trends. The dotted and dashed vertical lines (for propensity score nearest neighbor matching and propensity score digit-based greedy matching, respectively) mark the 6 points at which the propensity score matching trends first met the 0.05, 0.025 and 0.01 absolute propensity score distance caliper criteria (vertical line colors correspond to trend colors). The caliper criteria always were met in the order 0.05, 0.025, 0.01 during the pruning process.

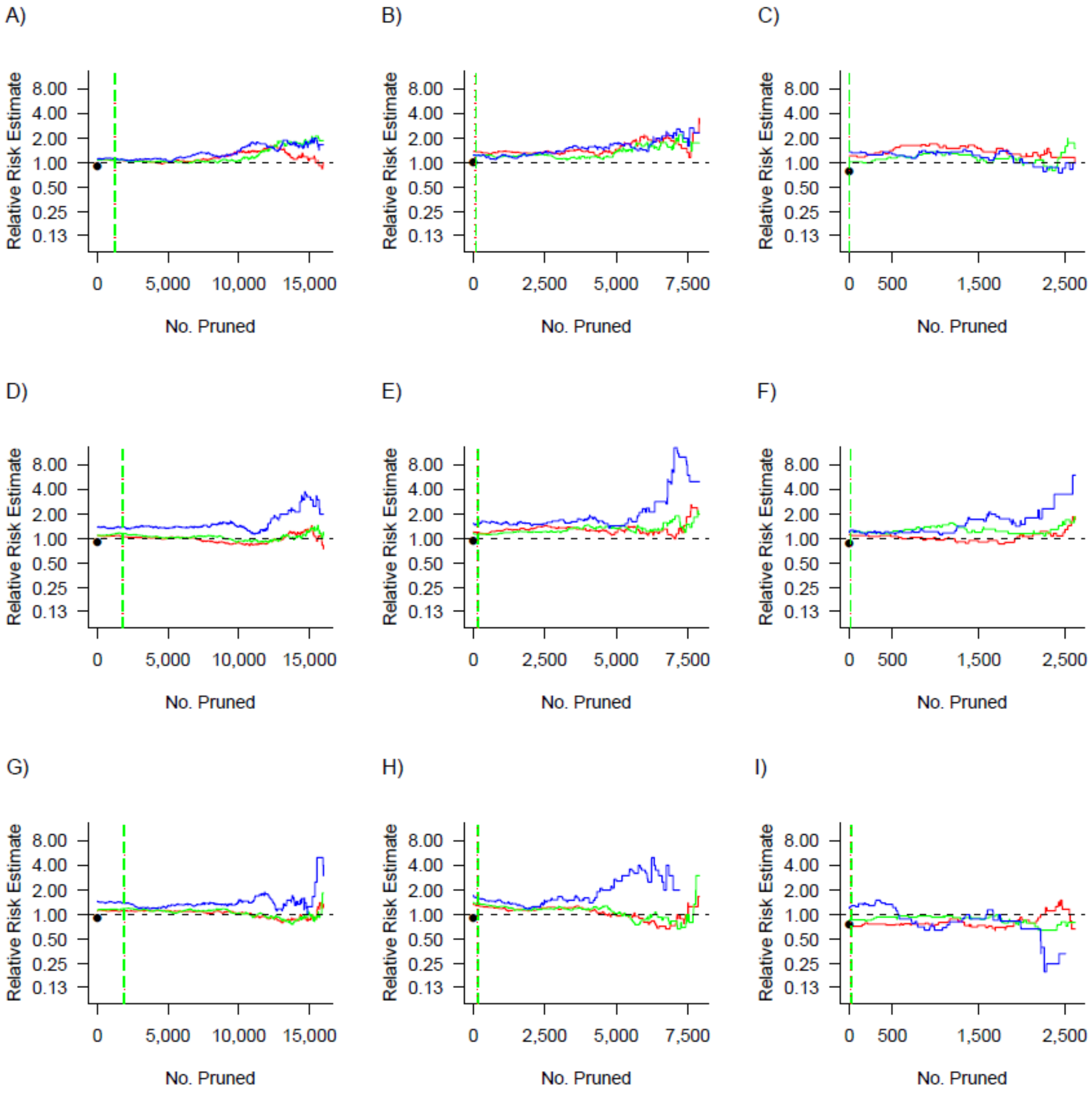

A)
B)
C)
D)
E)
F)
G)
H)
I)
Relative Risk Estimate
8.00
4.00
2.00
1.00
0.50
0.25
0.13
0
5,000
10,000
15,000
2,500
7,500
500
1,500
No. Pruned

Figure 11. Relative Risk Estimate Trends for the 3 MAX Datasets

A) Original index exposure prevalence (IEP). B) 400% of IEP. C) 700% of IEP. A dashed horizontal black line at the relative risk estimate value of 1.00 is included for reference. The black dots indicate the relative risk estimates of the pre-matched datasets. Red lines indicate propensity score nearest neighbor matching trends, green lines indicate propensity score digit-based greedy matching trends and blue lines indicate Mahalanobis distance matching trends. The dotted and dashed vertical lines (for propensity score nearest neighbor matching and propensity score digit-based greedy matching, respectively) mark the 6 points at which the propensity score matching trends first met the 0.05, 0.025 and 0.01 absolute propensity score distance caliper criteria (vertical line colors correspond to trend colors). The caliper criteria always were met in the order 0.05, 0.025, 0.01 during the pruning process.

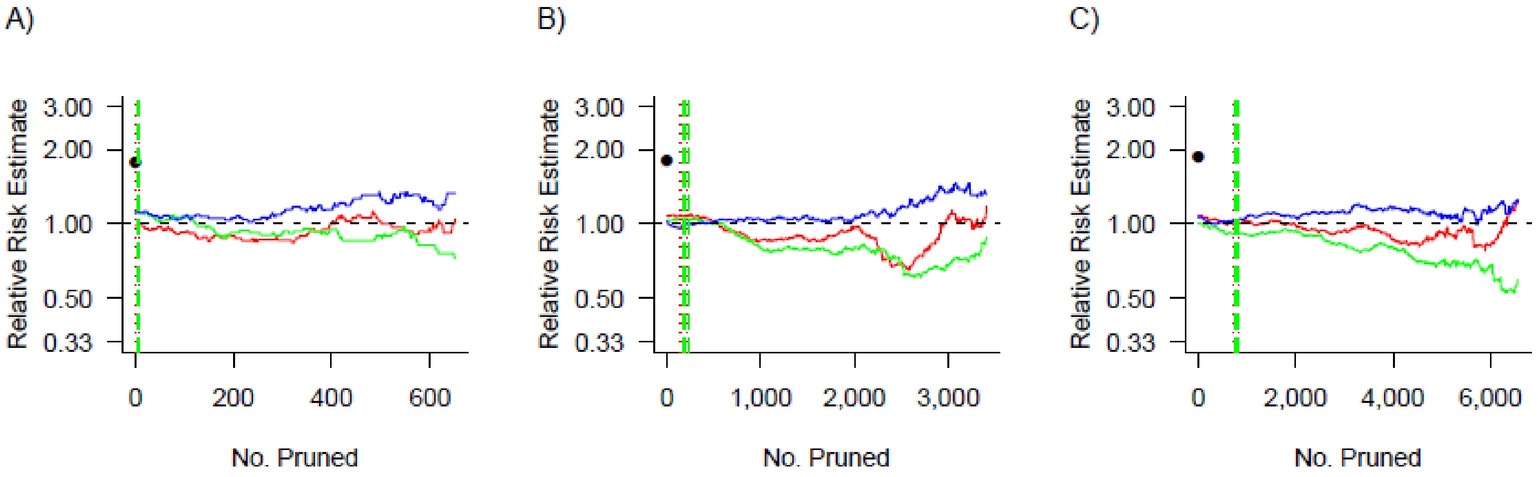


Figure 12. Plasmode Analysis Results, Non Interaction Scenarios – Average Proportional Decrease in Mahalanobis Balance

A) Small covariate set scenarios. B) Standard covariate set scenarios. C) Large covariate set scenarios. Blue lines indicate coarsened exact matching trends, green lines indicate propensity score matching trends, purple line indicates Mahalanobis distance matching trend and red lines indicate fine stratification on the propensity score trends. IEP = Index exposure prevalence, MB = Mahalanobis balance.

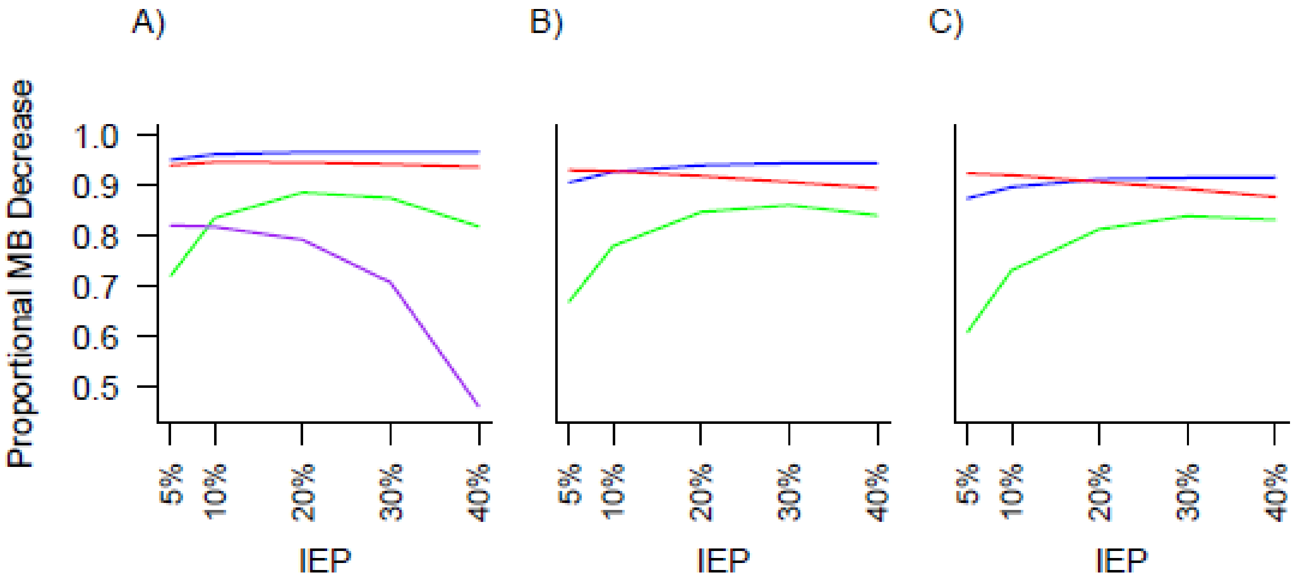


Figure 13. Plasmode Analysis Results, Non Interaction Scenarios – Square Root of Mean Squared Error, Including Coarsened Exact Matching Results

A) Small covariate set scenarios. B) Standard covariate set scenarios. C) Large covariate set scenarios. Blue lines indicate coarsened exact matching trends, green lines indicate propensity score matching trends, purple line indicates Mahalanobis distance matching trend and red lines indicate fine stratification on the propensity score trends. IEP = Index exposure prevalence, MSE = Mean squared error.

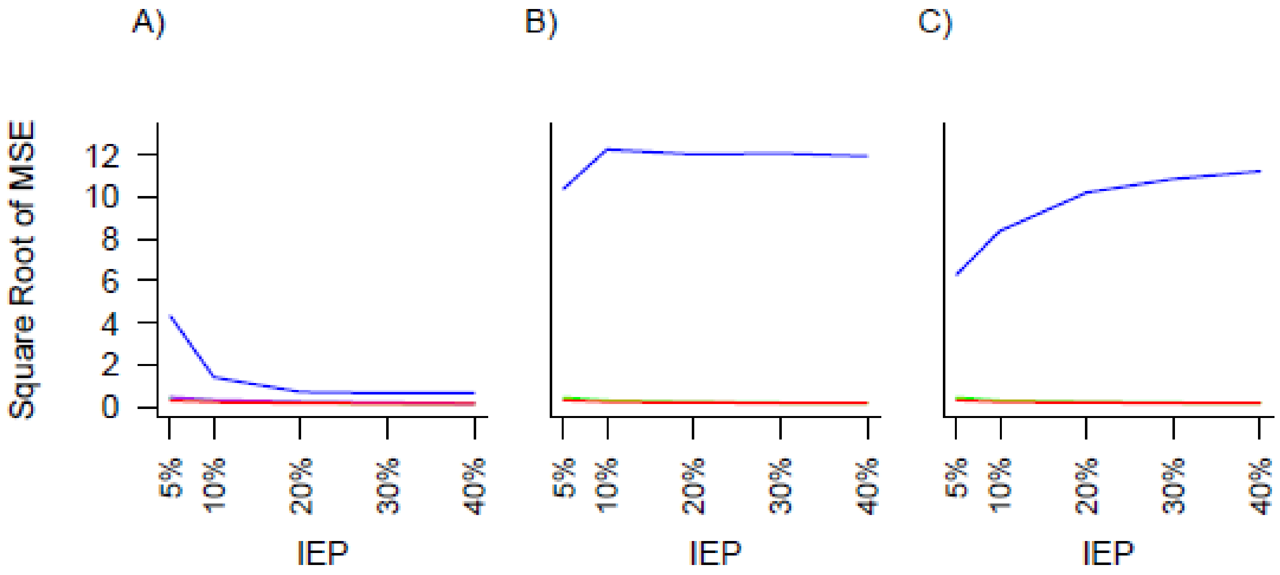


Figure 14. Plasmode Analysis Results, Non Interaction Scenarios – Square Root of Mean Squared Error, Excluding Coarsened Exact Matching Results

A) Small covariate set scenarios. B) Standard covariate set scenarios. C) Large covariate set scenarios. Green lines indicate propensity score matching trends, purple line indicates Mahalanobis distance matching trend and red lines indicate fine stratification on the propensity score trends. IEP = Index exposure prevalence, MSE = Mean squared error.

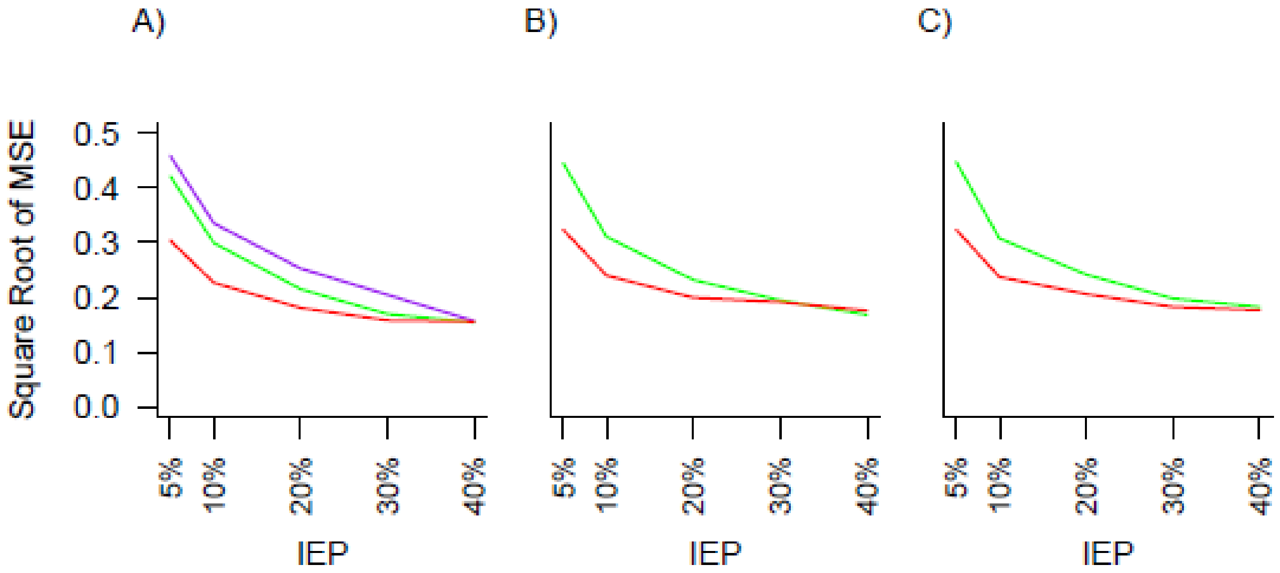


Figure 15. Plasmode Analysis Results, Non Interaction Scenarios – Bias, Including Coarsened Exact Matching Results

A) Small covariate set scenarios. B) Standard covariate set scenarios. C) Large covariate set scenarios. Black dots indicate the bias corresponding to the crude log risk ratio. Blue lines indicate coarsened exact matching trends, green lines indicate propensity score matching trends, purple line indicates Mahalanobis distance matching trend and red lines indicate fine stratification on the propensity score trends. IEP = Index exposure prevalence.

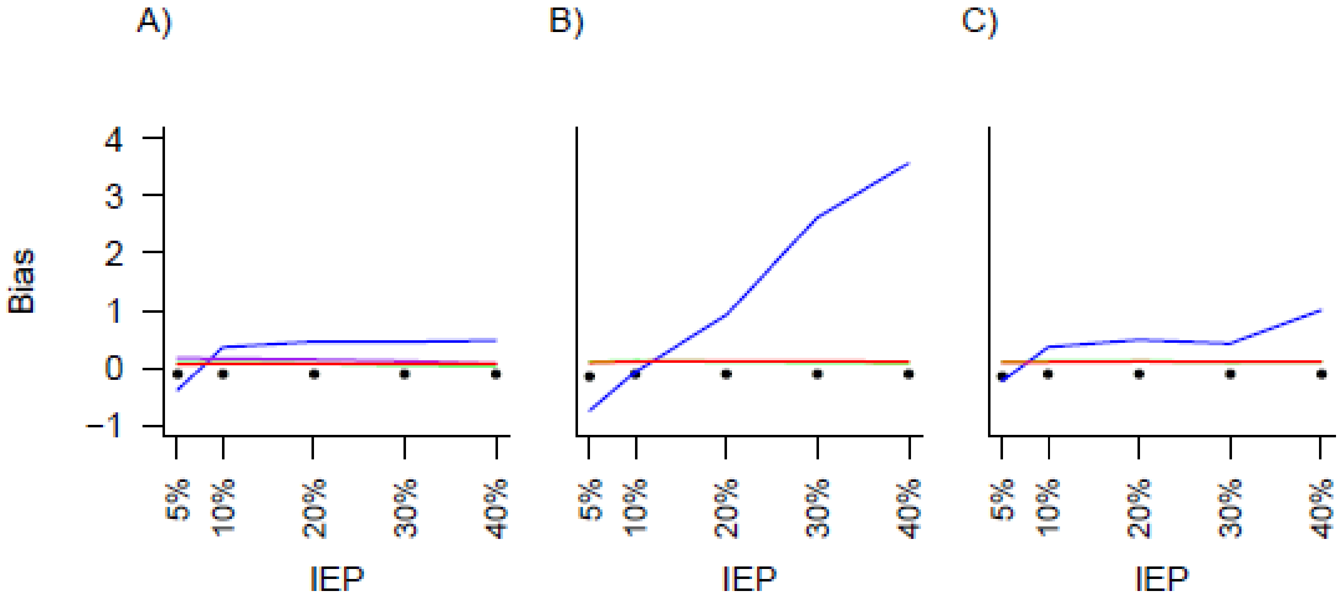


Figure 16. Plasmode Analysis Results, Non Interaction Scenarios – Bias, Excluding Coarsened Exact Matching Results

A) Small covariate set scenarios. B) Standard covariate set scenarios. C) Large covariate set scenarios. Black dots indicate the bias corresponding to the crude log risk ratio. Green lines indicate propensity score matching trends, purple line indicates Mahalanobis distance matching trend and red lines indicate fine stratification on the propensity score trends. IEP = Index exposure prevalence.

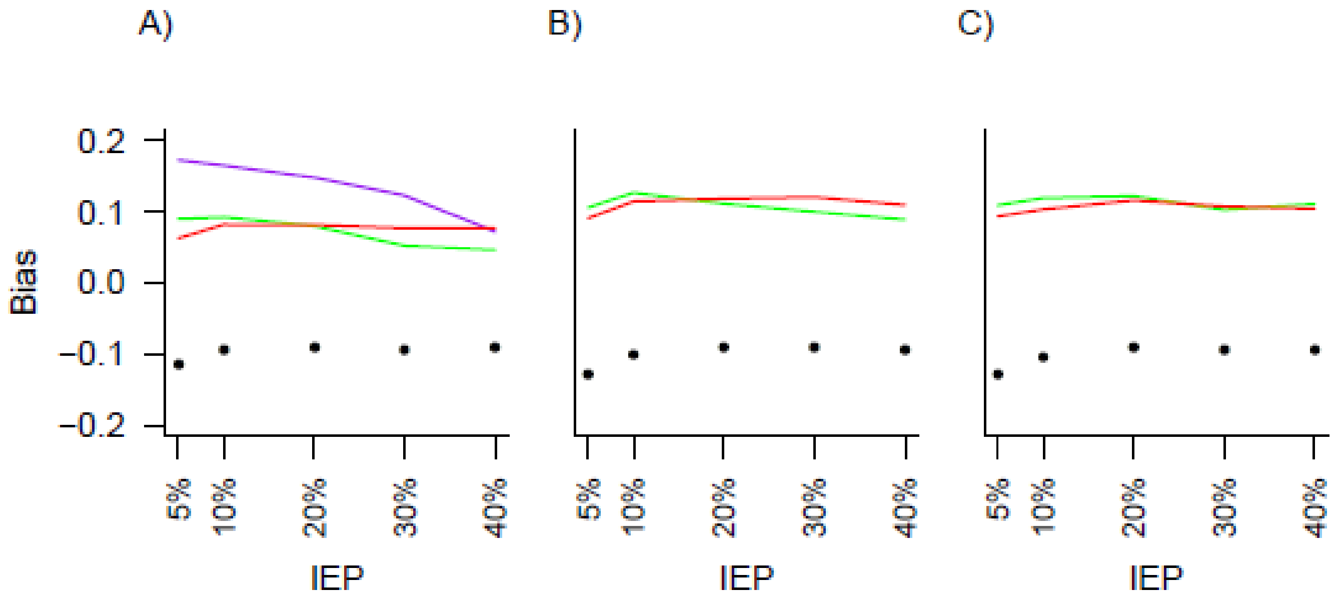


Figure 17. Plasmode Analysis Results, Non Interaction Scenarios – Variance, Including Coarsened Exact Matching Results

A) Small covariate set scenarios. B) Standard covariate set scenarios. C) Large covariate set scenarios. Blue lines indicate coarsened exact matching trends, green lines indicate propensity score matching trends, purple line indicates Mahalanobis distance matching trend and red lines indicate fine stratification on the propensity score trends. IEP = Index exposure prevalence.

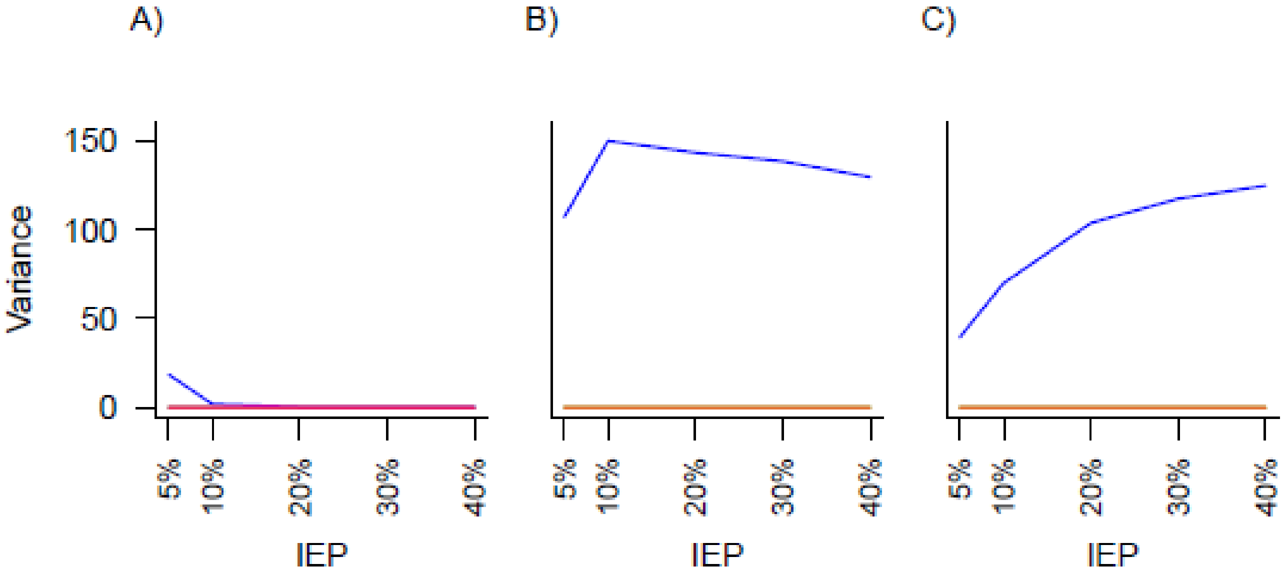


Figure 18. Plasmode Analysis Results, Non Interaction Scenarios – Variance, Excluding Coarsened Exact Matching Results

A) Small covariate set scenarios. B) Standard covariate set scenarios. C) Large covariate set scenarios. Green lines indicate propensity score matching trends, purple line indicates Mahalanobis distance matching trend and red lines indicate fine stratification on the propensity score trends. IEP = Index exposure prevalence.

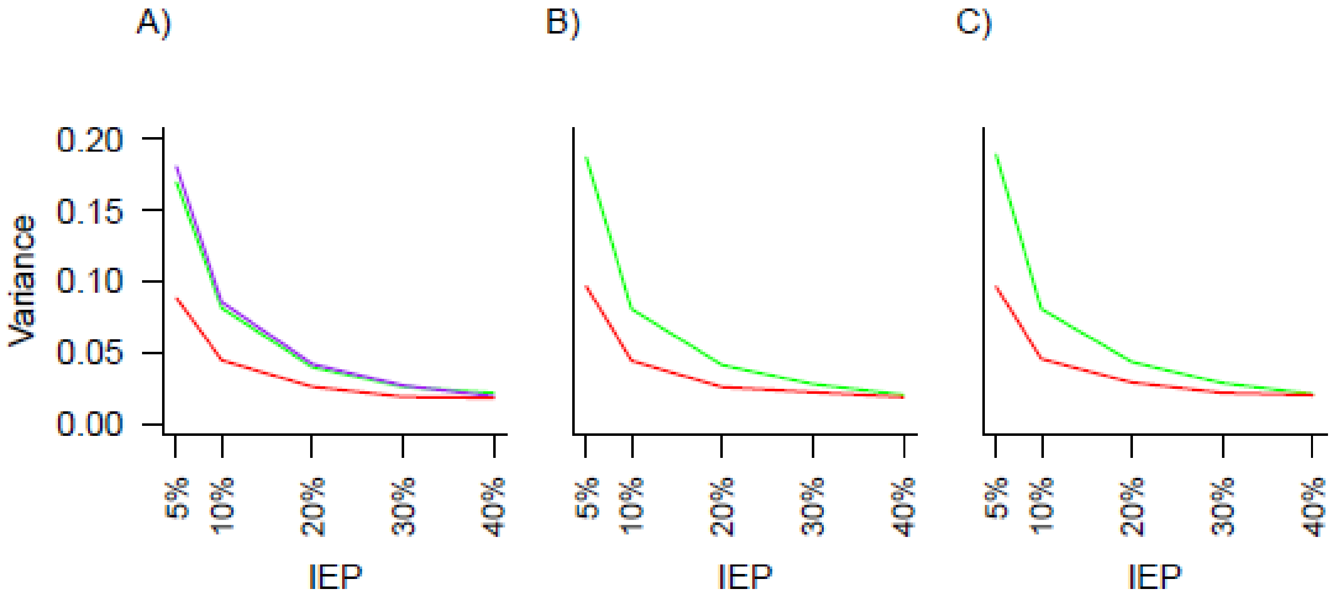


Figure 19. Depiction of the Usual Application of 1:1 Propensity Score Matching Without Replacement

1:1 PSM = 1:1 propensity score matching without replacement; E = Exposed; U = Unexposed; Y = Outcome variable; Z = Treatment status variable.

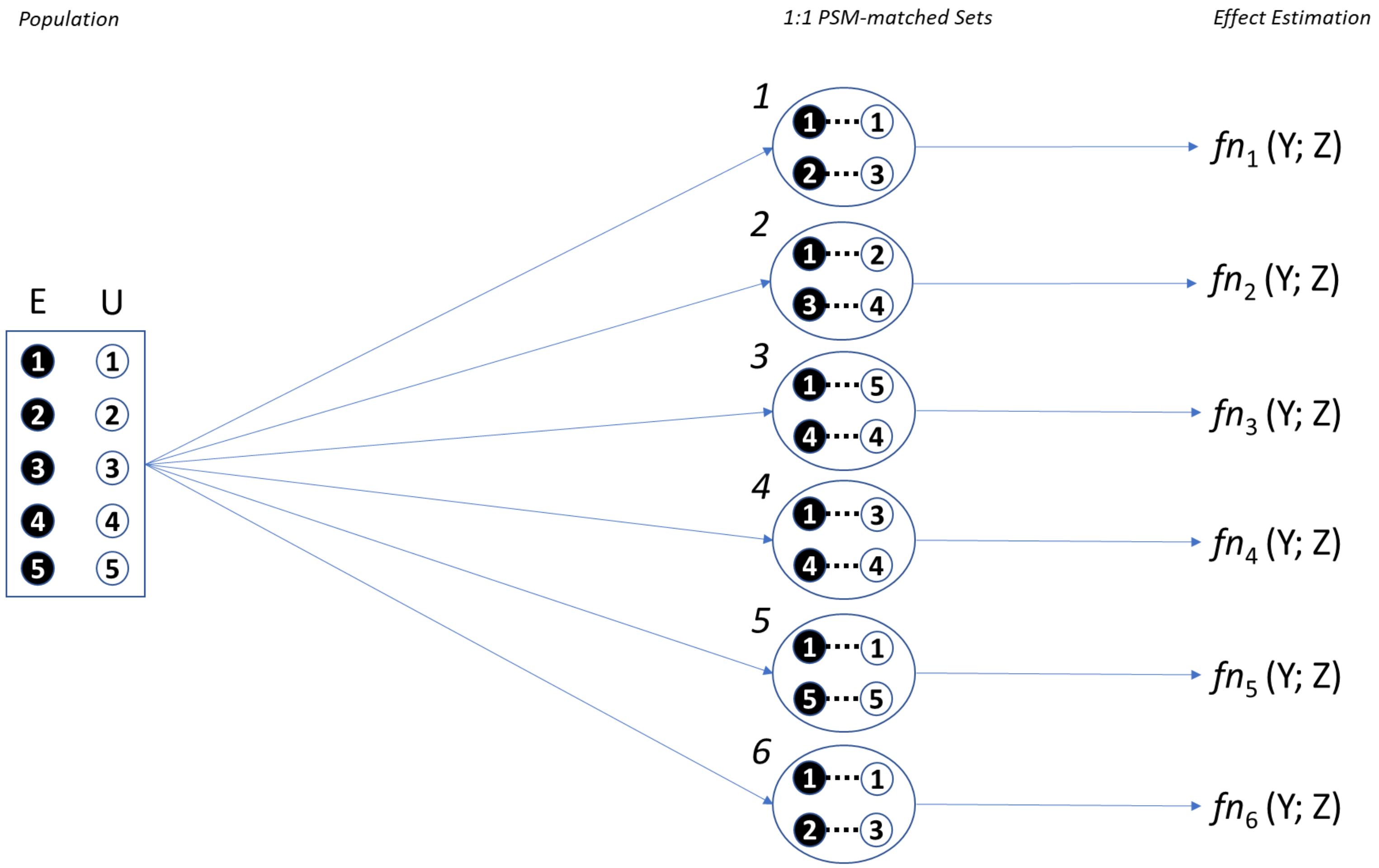


Figure 20. Depiction of the Usual Application of 1:1 Propensity Score Matching Without Replacement (Figure 1), Including Corresponding Pre-matched Samples

1:1 PSM = 1:1 propensity score matching without replacement; E = Exposed; PS = Propensity score; U = Unexposed; X = Covariate vector; Y = Outcome variable; Z = Treatment status variable.

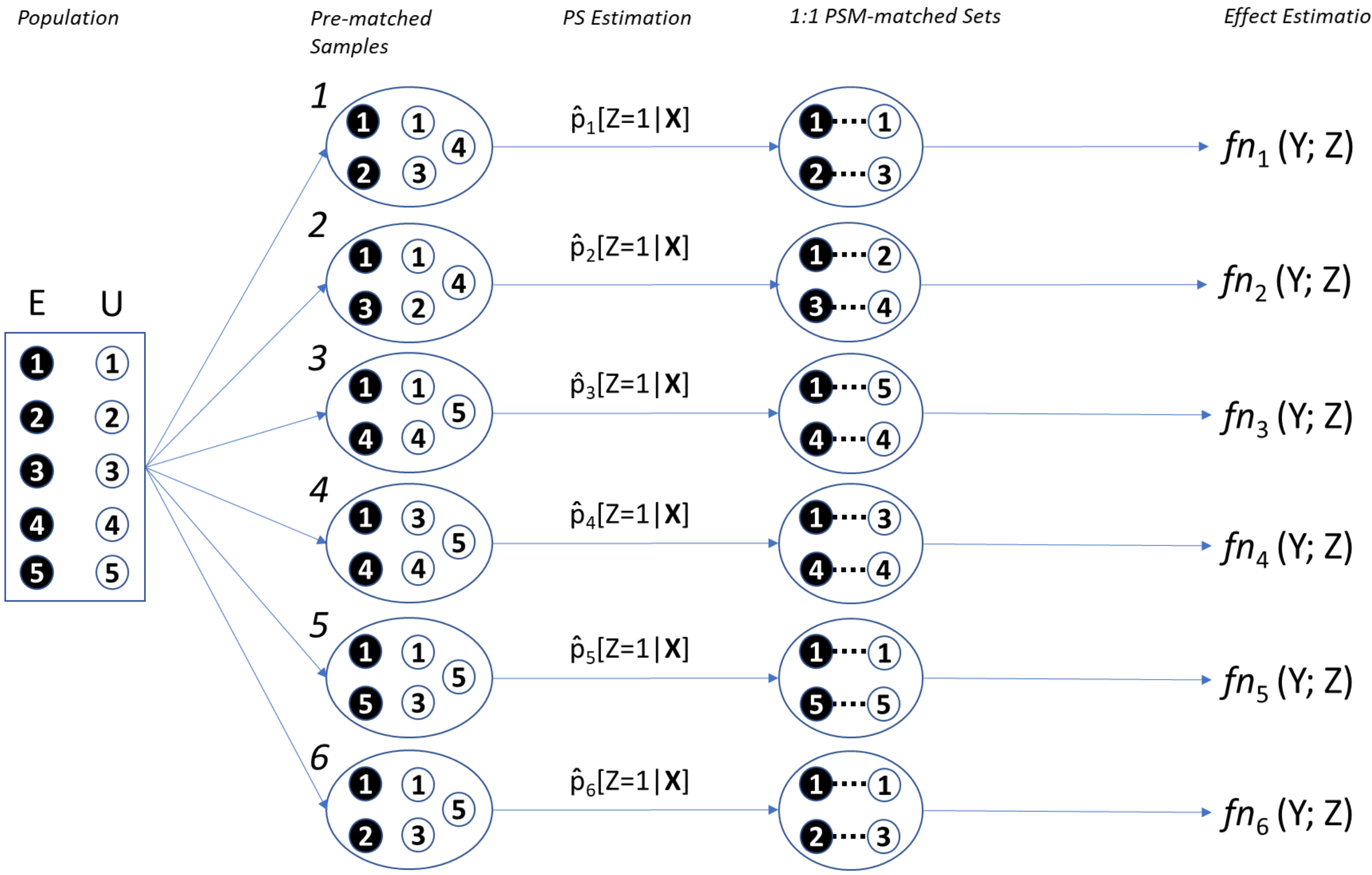


Figure 21. Depiction of the Mechanics of Simple Bootstrap 1:1 Propensity Score Matching for Standard Error Estimation

1:1 PSM = 1:1 propensity score matching without replacement; E = Exposed; PS = Propensity score; U = Unexposed; X = Covariate vector; Y = Outcome variable; Z = Treatment status variable.

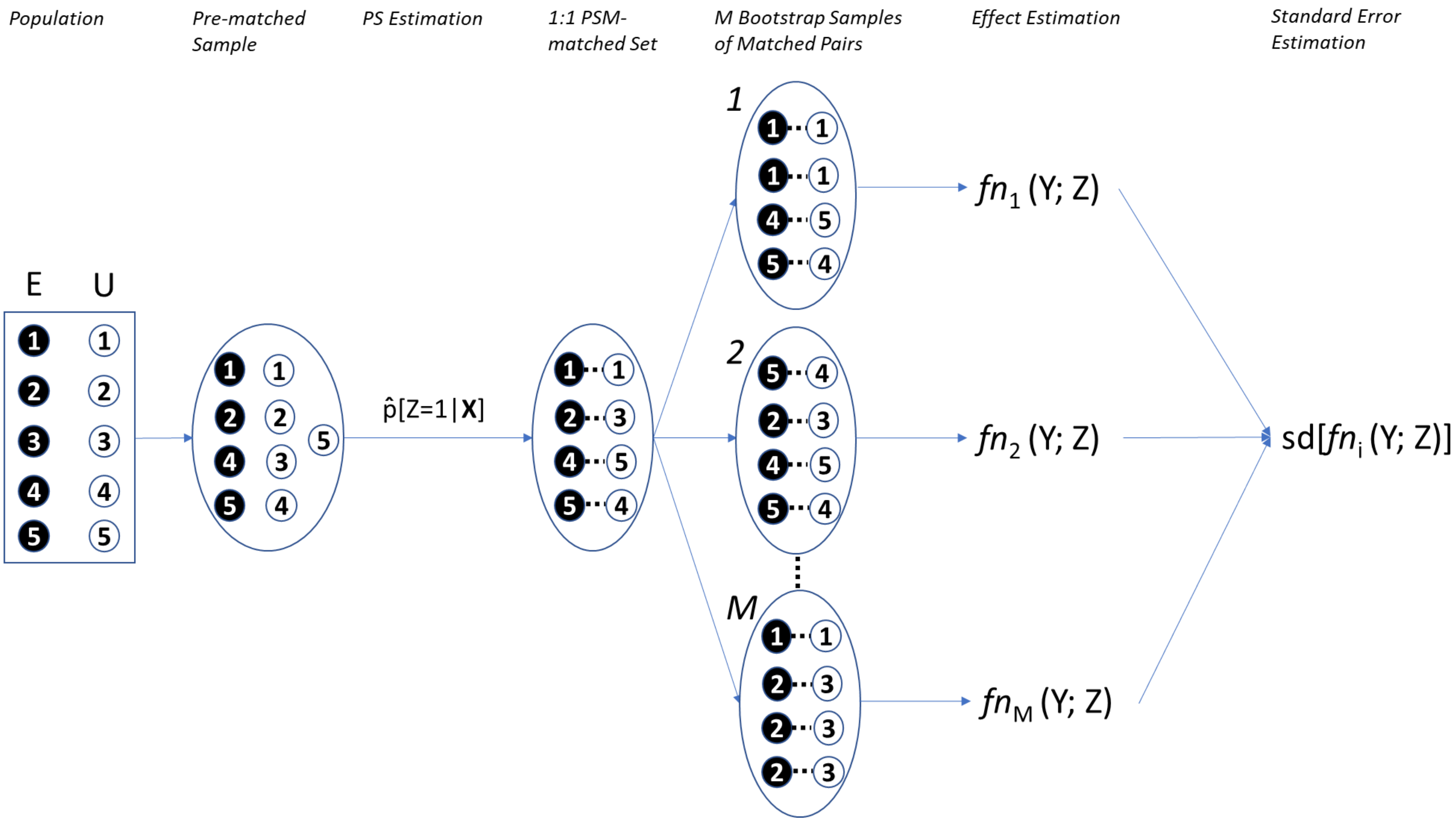


Figure 22. Depiction of the Mechanics of Complex Bootstrap 1:1 Propensity Score Matching for Standard Error Estimation

1:1 PSM = 1:1 propensity score matching without replacement; E = Exposed; PS = Propensity score; U = Unexposed; X = Covariate vector; Y = Outcome variable; Z = Treatment status variable.

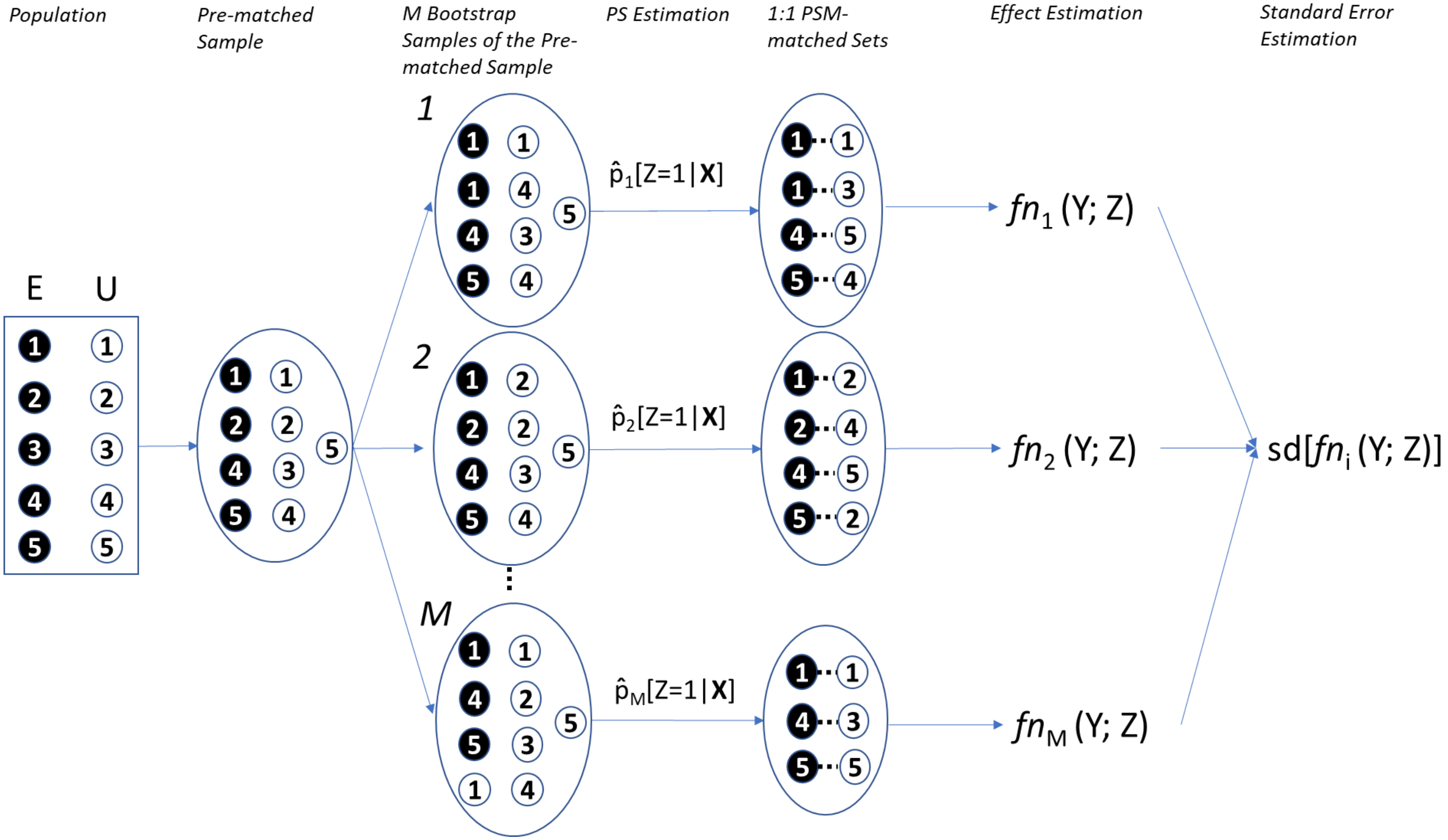
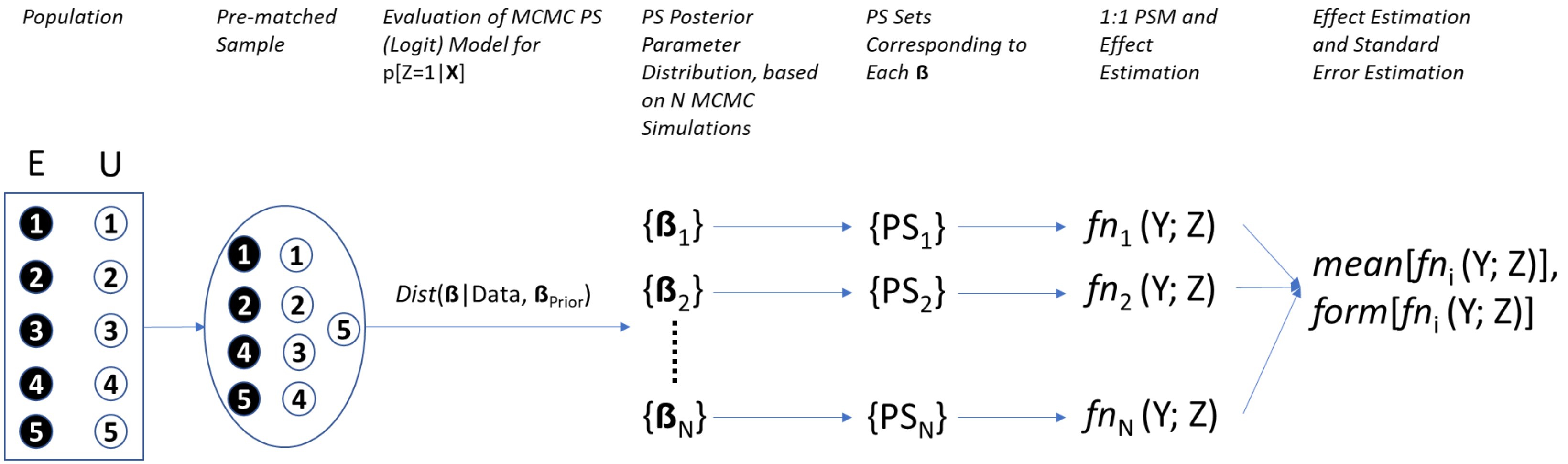


Figure 23. Depiction of the Mechanics of Bayesian 1:1 Propensity Score Matching for Standard Error Estimation

The "form" function represents the formula for the standard error used by Kaplan and Chen (2014). 1:1 PSM = 1:1 propensity score matching without replacement; E = Exposed; PS = Propensity score; U = Unexposed; X = Covariate vector; Y = Outcome variable; Z = Treatment status variable.


Population
Pre-matched Sample
Evaluation of MCMC PS (Logit) Model for p[Z=1|**X**]
PS Posterior Parameter Distribution, based on N MCMC Simulations
PS Sets Corresponding to Each **ß**
1:1 PSM and Effect Estimation
Effect Estimation and Standard Error Estimation
E
U
1
2
3
4
5
1
2
3
4
5
Dist(**ß**|Data, **ß**$_{Prior}$)
{**ß**$_1$}
{**ß**$_2$}
{**ß**$_N$}
{PS$_1$}
{PS$_2$}
{PS$_N$}
fn$_1$ (Y; Z)
fn$_2$ (Y; Z)
fn$_N$ (Y; Z)
mean[*fn*$_i$ (Y; Z)],
form[*fn*$_i$ (Y; Z)]

APPENDIX

DERIVATION OF THE WEIGHTING SCHEME FOR COARSENED EXACT MATCHING (CEM) AND FOR FINE STRATIFICATION ON THE PROPENSITY SCORE (FS) [Iacus et al., 2011; Iacus et al., 2011; Desai et al., 2017]

For this exposition, let "treatment" stand for the index exposure of interest and "control" stand for the reference exposure.

Assuming that the causal effect of interest is the average effect of treatment on the treated (ATT), the control group in the analytic dataset must be the counterfactual ideal for the treated group in the analytic dataset. From a statistical perspective, this means that the distribution of covariates in the entire control group must be the same as the distribution of covariates in the entire treated group (such equivalence approximates ignorability, which is required for recovery of the causal effect).

Both CEM and FS attempt to balance the distributions of covariates between treated and control units within the context of strata, which are determined by the respective method (for CEM, the method is exact matching within coarsened boundaries that define the strata; for FS, the method is grouping of control units with estimated propensity score values that are similar to the estimated propensity score values of treated units, with individual groups [strata] defined by quantiles of the estimated propensity score distribution of the treated units).

To maximize the number of units that appear in the analytic dataset (i.e., to maximize statistical efficiency), both methods allow for multiple treated units and multiple control units to appear in a stratum, in a *variable* ratio across strata (i.e., the same numbers of treated and control units might not appear across strata). Because of the variable placement of units across strata, the distributions of covariates between treated units and control units within a given stratum, and, by extension, among all strata, are not necessarily comparable [Rassen et al., 2012]. Consequently, placement of units into strata does *not* necessarily guarantee that the covariate balance that actually is achieved by the method will be perceivable. However, covariate balance may be perceivable, prior to any analysis, via an appropriate weighting scheme.

Let N_{iT} be the total number of treated units in stratum *i* and N_{iC} be the total number of control units in stratum *i*. Then N_T ($\sum_i N_{iT}$) is the total number of treated units in the analytic dataset (i.e., among all strata) and N_C ($\sum_i N_{iC}$) is the equivalent number of control units.

One way to observe the covariate balance produced by CEM and FS in the resulting analytic dataset (i.e., to make the distributions of covariates between the treated and control units comparable) is to weight control units in each stratum so that the proportion of treated units across all strata who are in stratum *i* is the same as the proportion of control units across all strata who are in stratum *i* (the variable ratio placement for CEM and FS does *not* guarantee that these proportion are equal across stratum). Thus, the requirement is:

$$\mathbf{1.}\ \{N_{iT} / N_T = N_{iC} / N_C\}\ \forall\, i$$

By rearranging the equation, it becomes clear that if this condition holds, the ratio of treated units to control units within each stratum is the same as the ratio of treated units to control units in the entire analytic dataset:

$$\mathbf{2.}\ \{N_{iT} / N_{iC} = N_T / N_C\}\ \forall\, i$$

Thus, the covariate distribution balance within each stratum will be correctly reflected in an analysis of the *entire* analytic dataset.

By rearranging the equation in **1** again, it is clear that the following also is true:

$$\mathbf{3.}\ \{(N_{iT} / N_T) / (N_{iC} / N_C) = 1\}\ \forall\, i$$

In other words, if the condition represented by **1** holds, then the ratio of the proportion of treated units across all strata who are in stratum *i* to the proportion of control units across all strata who are in stratum *i* must be unity.

If **3** does not hold in stratum *i*, then the following holds:

$$\mathbf{4.}\ (N_{iT} / N_T) / (N_{iC} / N_C) = \omega_i$$

(where $\omega_i \neq 1$). To ensure that **3** holds in stratum *i*, both sides of **4** must be divided by ω_i:

$$\mathbf{5.}\ (N_{iT} / N_T) / ([\omega_i{*}N_{iC}] / N_C) = \omega_i / \omega_i = 1$$

Therefore, ω_i, is the weight that should be applied *to each control unit* in stratum *i* to ensure that **3** holds. To show this, consider an example stratum *i* that comprises 3 control units and for which **4** holds. Each of these 3 control units contributes a weight of 1 after performing CEM or FS (i.e., without any further weighting). Therefore, the following is true:

$$\textbf{6.}\ N_{iC} = (1 + 1 + 1)$$

Thus, applying ω_i to N_{iC} as in **5** yields:

$$\textbf{7.}\ \omega_i{*}N_{iC} = \omega_i{*}(1 + 1 + 1) = (\omega_i{*}1 + \omega_i{*}1 + \omega_i{*}1)$$

Hence, each control unit in stratum *i* receives ω_i as its weight prior to any analysis in order to recover the condition represented by **3** and, thus, to reveal the extent of covariate balance achieved by CEM or FS. Using the weighted analytic dataset, the ATT may be estimated.

Of note, the risk ratio that is weighted using this scheme is equivalent to the common measure of association, the standardized morbidity ratio [Rothman et al., 2008].

Example

As an example, consider 2 strata from a hypothetical analytic dataset, resulting from an application of CEM. For this dataset, $N_T = 30$ and $N_{C =}$ 50. The 2 strata are composed as follows.

Stratum 1		Stratum 2	
# Treated Units	# Control Units	# Treated Units	# Control Units
2	3	4	1

Then, ω_1 and ω_2 are calculated as follows.

Stratum 1: $\omega_1 = (2/30) / (3/50) = 1.1111$

Stratum 2: $\omega_2 = (4/30) / (1/50) = 6.6667$

Thus, for stratum 1, each control unit should receive weight, 1.11 and for stratum 2, each control unit (the single control unit in this case) should receive weight, 6.67. For both strata, control units are *up-weighted* since, prior to weighting, the proportion of control units across all strata who are in the stratum is *less than* the proportion of treated units across all strata who are in the stratum (i.e., $\omega_i > 1$). To recover the condition represented by **3**, the weights are applied as follows (rounding error allowed here).

Stratum 1: $(2/30) / ([1.11*3]/50) = 1$

Stratum 2: $(4/30) / ([6.67*1]/50) = 1$

BIBLIOGRAPHY

Abadie, A. and Imbens, G. (2016). "Matching on the Estimated Propensity Score " Econometrica 84(2): 781-807 DOI: 10.3982/ECTA11293.

Alvarez, R. M. and Levin, I. (2014). "Uncertain Neighbors: Bayesian Propensity Score Matching for Causal Inference ". https://polisci.wustl.edu/files/polisci/imce/slammlevin.pdf.

An, W. (2010). "Bayesian Propensity Score Estimators: Incorporating Uncertainties in Propensity Scores into Causal Inference." Sociological Methodology 40: 151-189.

Austin, P. C. (2007). "Propensity-score matching in the cardiovascular surgery literature from 2004 to 2006: a systematic review and suggestions for improvement." Journal of Thoracic and Cardiovascular Surgery 134(5): 1128-1135 DOI: 10.1016/j.jtcvs.2007.07.021. https://www.ncbi.nlm.nih.gov/pubmed/17976439.

Austin, P. C. (2008). "A critical appraisal of propensity-score matching in the medical literature between 1996 and 2003." Statistics in Medicine 27(12): 2037-2049 DOI: 10.1002/sim.3150. https://www.ncbi.nlm.nih.gov/pubmed/18038446.

Austin, P. C. (2008). "The performance of different propensity-score methods for estimating relative risks." Journal of Clinical Epidemiology 61(6): 537-545 DOI: 10.1016/j.jclinepi.2007.07.011. https://www.ncbi.nlm.nih.gov/pubmed/18471657.

Austin, P. C. (2008). "Primer on statistical interpretation or methods report card on propensity-score matching in the cardiology literature from 2004 to 2006: a systematic review." Circulation: Cardiovascular Quality and Outcomes 1(1): 62-67 DOI: 10.1161/CIRCOUTCOMES.108.790634. https://www.ncbi.nlm.nih.gov/pubmed/20031790.

Austin, P. C. (2009). "Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples." Statistics in Medicine 28(25): 3083-3107 DOI: 10.1002/sim.3697. https://www.ncbi.nlm.nih.gov/pubmed/19757444.

Austin, P. C. (2009). "Type I error rates, coverage of confidence intervals, and variance estimation in propensity-score matched analyses." International Journal of Biostatistics 5(1): Article 13 DOI: 10.2202/1557-4679.1146. https://www.ncbi.nlm.nih.gov/pubmed/20949126.

Austin, P. C. (2011). "Comparing paired vs non-paired statistical methods of analyses when making inferences about absolute risk reductions in propensity-score matched samples." Statistics in Medicine 30(11): 1292-1301 DOI: 10.1002/sim.4200. https://www.ncbi.nlm.nih.gov/pubmed/21337595.

Austin, P. C. (2011). "An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies." Multivariate Behavioral Research 46(3): 399-424 DOI: 10.1080/00273171.2011.568786. https://www.ncbi.nlm.nih.gov/pubmed/21818162.

Austin, P. C. (2017). "Double propensity-score adjustment: A solution to design bias or bias due to incomplete matching." Statistical Methods in Medical Research 26(1): 201-222 DOI: 10.1177/0962280214543508. https://www.ncbi.nlm.nih.gov/pubmed/25038071.

Austin, P. C., Grootendorst, P. and Anderson, G. M. (2007). "A comparison of the ability of different propensity score models to balance measured variables between treated and untreated subjects: a Monte Carlo study." Statistics in Medicine 26(4): 734-753 DOI: 10.1002/sim.2580. https://www.ncbi.nlm.nih.gov/pubmed/16708349.

Austin, P. C. and Schuster, T. (2016). "The performance of different propensity score methods for estimating absolute effects of treatments on survival outcomes: A simulation study." Statistical Methods in Medical Research 25(5): 2214-2237 DOI: 10.1177/0962280213519716. https://www.ncbi.nlm.nih.gov/pubmed/24463885.

Austin, P. C. and Small, D. S. (2014). "The Use of Bootstrapping When Using Propensity-Score Matching Without Replacement: A Simulation Study." Statistics in Medicine 33: 4306-4319.

Bai, H. (2013). "A Bootstrap Procedure of Propensity Score Estimation." Journal of Experimental Education 81(2): 157-177 DOI: 10.1080/00220973.2012.700497.

Bateman, B. T., Hernandez-Diaz, S., Fischer, M. A., Seely, E. W., Ecker, J. L., Franklin, J. M., Desai, R. J., Allen-Coleman, C., Mogun, H., Avorn, J. and Huybrechts, K. F. (2015). "Statins and congenital malformations: cohort study." BMJ: British Medical Journal 350: h1035 DOI: 10.1136/bmj.h1035. https://www.ncbi.nlm.nih.gov/pubmed/25784688.

Bodory, H., Camponovo, L., Huber, M. and Lechner, M. (2018). "The Finite Sample Performance of Inference Methods for Propensity Score Matching and Weighting Estimators." Journal of Business & Economic Statistics DOI: 10.1080/07350015.2018.1476247.

Brookhart, M. A., Wang, P. S., Solomon, D. H. and Schneeweiss, S. (2006). "Evaluating short-term drug effects using a physician-specific prescribing preference as an instrumental variable." Epidemiology 17(3): 268-275 DOI:

10.1097/01.ede.0000193606.58671.c5. https://www.ncbi.nlm.nih.gov/pubmed/16617275.

Burton, A., Altman, D. G., Royston, P. and Holder, R. L. (2006). "The design of simulation studies in medical statistics." Statistics in Medicine 25(24): 4279-4292 DOI: 10.1002/sim.2673. https://www.ncbi.nlm.nih.gov/pubmed/16947139.

Cochran, W. G. (1968). "The effectiveness of adjustment by subclassification in removing bias in observational studies." Biometrics 24(2): 295-313. https://www.ncbi.nlm.nih.gov/pubmed/5683871.

D'Agostino, R. B., Jr. (1998). "Propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group." Statistics in Medicine 17(19): 2265-2281. https://www.ncbi.nlm.nih.gov/pubmed/9802183.

Desai, R. J., Rothman, K. J., Bateman, B. T., Hernandez-Diaz, S. and Huybrechts, K. F. (2017). "A Propensity-score-based Fine Stratification Approach for Confounding Adjustment When Exposure Is Infrequent." Epidemiology 28(2): 249-257 DOI: 10.1097/EDE.0000000000000595. https://www.ncbi.nlm.nih.gov/pubmed/27922533.

Desai, R. J., Wyss, R., Abdia, Y., Toh, S., Johnson, M., Lee, H., Karami, S., Major, J. M., Nguyen, M., Wang, S. V., Franklin, J. M. and Gagne, J. J. (2019). "*Under Review* - Evaluating the use of bootstrapping in cohort studies conducted with 1:1 propensity score matching - A plasmode simulation study."

Efron, B. (1979). "Bootstrap Methods: Another Look at the Jackknife." The Annals of Statistics 7(1): 1-26.

Efron, B. and Tibshirani, R. (1986). "Bootstrap Methods for Standard Errors, Confidence Intervals, and Other Measures of Statistical Accuracy." Statistical Science 1(1): 54-77.

Franklin, J. M., Eddings, W., Glynn, R. J. and Schneeweiss, S. (2015). "Regularized Regression Versus the High-Dimensional Propensity Score for Confounding Adjustment in Secondary Database Analyses." Am J Epidemiol 182(7): 651-659 DOI: 10.1093/aje/kwv108. https://www.ncbi.nlm.nih.gov/pubmed/26233956.

Franklin, J. M., Rassen, J. A., Ackermann, D., Bartels, D. B. and Schneeweiss, S. (2014). "Metrics for covariate balance in cohort studies of causal effects." Statistics in Medicine 33(10): 1685-1699 DOI: 10.1002/sim.6058. https://www.ncbi.nlm.nih.gov/pubmed/24323618.

Franklin, J. M., Schneeweiss, S., Polinski, J. M. and Rassen, J. A. (2014). "Plasmode simulation for the evaluation of pharmacoepidemiologic methods in complex healthcare databases." Computational Statistics and Data Analysis 72: 219-226 DOI: 10.1016/j.csda.2013.10.018. https://www.ncbi.nlm.nih.gov/pubmed/24587587.

Fullerton, B., Pohlmann, B., Krohn, R., Adams, J. L., Gerlach, F. M. and Erler, A. (2016). "The Comparison of Matching Methods Using Different Measures of Balance: Benefits and Risks Exemplified within a Study to Evaluate the Effects of German Disease Management Programs on Long-Term Outcomes of Patients with Type 2 Diabetes." Health Services Research 51(5): 1960-1980 DOI: 10.1111/1475-6773.12452. https://www.ncbi.nlm.nih.gov/pubmed/26841379.

Glynn, R. J., Schneeweiss, S. and Sturmer, T. (2006). "Indications for propensity scores and review of their use in pharmacoepidemiology." Basic and Clinical Pharmacology and Toxicology 98(3): 253-259 DOI: 10.1111/j.1742-7843.2006.pto_293.x. https://www.ncbi.nlm.nih.gov/pubmed/16611199.

Greenland, S. (2006). "Bayesian perspectives for epidemiological research: I. Foundations and basic methods." International Journal of Epidemiology 35(3): 765-775 DOI: 10.1093/ije/dyi312. https://www.ncbi.nlm.nih.gov/pubmed/16446352.

Greenland, S. (2007). "Bayesian perspectives for epidemiological research. II. Regression analysis." International Journal of Epidemiology 36(1): 195-202 DOI: 10.1093/ije/dyl289. https://www.ncbi.nlm.nih.gov/pubmed/17329317.

Greenland, S. (2009). "Bayesian perspectives for epidemiologic research: III. Bias analysis via missing-data methods." International Journal of Epidemiology 38(6): 1662-1673 DOI: 10.1093/ije/dyp278. https://www.ncbi.nlm.nih.gov/pubmed/19744933.

Greenland, S., Mansournia, M. A. and Altman, D. G. (2016). "Sparse data bias: a problem hiding in plain sight." BMJ: British Medical Journal 352: i1981 DOI: 10.1136/bmj.i1981. https://www.ncbi.nlm.nih.gov/pubmed/27121591.

Greevy, R., Lu, B., Silber, J. H. and Rosenbaum, P. (2004). "Optimal multivariate matching before randomization." Biostatistics 5(2): 263-275 DOI: 10.1093/biostatistics/5.2.263. https://www.ncbi.nlm.nih.gov/pubmed/15054030.

Gu, X. and Rosenbaum, P. R. (1993). "Comparison of multivariate matching methods: structures, distances and algorithms." Journal of Computational and Graphical Statistics 2: 405-420.

Guo, S. and Fraser, W. M. (2010). Propensity score analysis: Statistical methods and applications. Thousand Oaks, CA, Sage.

Hade, E. M. and Lu, B. (2014). "Bias associated with using the estimated propensity score as a regression covariate." Statistics in Medicine 33(1): 74-87 DOI: 10.1002/sim.5884. https://www.ncbi.nlm.nih.gov/pubmed/23787715.

Harrell, F. E., Jr., Califf, R. M., Pryor, D. B., Lee, K. L. and Rosati, R. A. (1982). "Evaluating the yield of medical tests." Journal of the American Statistical Association 247(18): 2543-2546. https://www.ncbi.nlm.nih.gov/pubmed/7069920.

Harrell, F. E., Lee, K. L., Califf, R. M., Pryor, D. B., Lee, K. L. and Rosati, R. A. (1984). "Regression modeling strategies for improved prognostic prediction." Statistics in Medicine 3: 143-152.

Hill, J. (2008). "Discussion of research using propensity-score matching: comments on 'A critical appraisal of propensity-score matching in the medical literature between 1996 and 2003' by Peter Austin, Statistics in Medicine." Statistics in Medicine 27(12): 2055-2061; discussion 2066-2059 DOI: 10.1002/sim.3245. https://www.ncbi.nlm.nih.gov/pubmed/18446836.

Hirano, K., Imbens, G. W. and G, R. (2003). "Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score." Econometrica 71(4): 1161-1189.

Ho, D., Imai, K., King, G. and Stuart, E. (2011). "Matchit: nonparametric preprocessing for parametric causal inference." Journal of Statistical Software 42(8): 1-28. http://gking.harvard.edu/matchit/.

Ho, D., Imai, K., King, G., Stuart, E. and Whitworth, A. (2018). "Package 'MatchIt'." https://cran.r-project.org/web/packages/MatchIt/MatchIt.pdf.

Ho, D., King, G. and Stuart, E. (2007). "Matching as nonparametric preprocessing for reducing model dependence in parameteric causal inference " Political Analysis 15(3): 199-236. http://gking.harvard.edu/files/abs/matchp-abs.shtml.

Hoshino, T. (2008). "A Bayesian propensity score adjustment for latent variable modeling and MCMC algorithm " Computational Statistics and Data Analysis 52: 1413-1429 DOI: 10.1016/j.csda.2007.03.024.

Huybrechts, K. F., Palmsten, K., Avorn, J., Cohen, L. S., Holmes, L. B., Franklin, J. M., Mogun, H., Levin, R., Kowal, M., Setoguchi, S. and Hernandez-Diaz, S. (2014). "Antidepressant use in pregnancy and the risk of cardiac defects." New England Journal of Medicine 370(25): 2397-2407 DOI: 10.1056/NEJMoa1312828. https://www.ncbi.nlm.nih.gov/pubmed/24941178.

Iacus, S. M., King, G. and Porro, G. (2011). "Causal inference without balance checking: coarsened exact matching." Political Analysis 20(1): 1-24 DOI: 10.1093/pan/mpr013.

Iacus, S. M., King, G. and Porro, G. (2011). "Multivariate matching methods that are monotonic imbalance bounding." Journal of the American Statistical Association 106: 345-361.

Iacus, S. M., King, G. and Porro, G. (2018). "Package 'cem'." https://cran.r-project.org/web/packages/cem/cem.pdf.

Imai, K., King, G. and Nall, C. (2009). "The essential role of pair matching in cluster-randomized experiments, with application to the Mexican Universal Health Insurance Evaluation." Statistical Science 24: 29-53.

Imai, K., King, G. and Stuart, E. (2008). "Misunderstandings among experimentalists and observationalists." Journal of the Royal Statistical Society, Series A 171: 481-502.

Jackson, J. W., Schmid, I. and Stuart, E. A. (2017). "Propensity Scores in Pharmacoepidemiology: Beyond the Horizon." Current Epidemiology Reports 4(4): 271-280 DOI: 10.1007/s40471-017-0131-y. https://www.ncbi.nlm.nih.gov/pubmed/29456922.

Kaplan, D. and Chen, J. (2012). "A Two-Step Bayesian Approach for Propensity Score Analysis: Simulations and Case Study." Psychometrika 77(3): 581-609 DOI: 10.1007/s11336-012-9262-8. https://www.ncbi.nlm.nih.gov/pubmed/27519782.

Kaplan, D. and Chen, J. (2014). "Bayesian Model Averaging for Propensity Score Analysis." Multivariate Behavioral Research 49(6): 505-517 DOI: 10.1080/00273171.2014.928492. https://www.ncbi.nlm.nih.gov/pubmed/26735355.

King, G. and L., Z. (2007). "When can history be our guide? The pitfalls of counterfactual inference." International Studies Quarterly 51: 183-210.

King, G., Lucas, C. and Nielsen, R. A. (2017). "The balance-sample size frontier in matching methods for causal inference." American Journal of Political Science 61(2): 473-489 DOI: 10.1111/ajps.12272.

King, G. and Nielsen, R. (2016). Why propensity scores should not be used for matching https://gking.harvard.edu/publications/why-propensity-scores-should-not-be-used-formatching.

King, G., Nielsen, R., Coberley, C., Pope, J. E. and Wells, A. (2011). Comparative effectiveness of matching methods for causal inference. http://gking.harvard.edu/publications/comparative-effectiveness-matching-methods-causal-inference.

Li, F., Morgan, K. L. and Zaslavsky, A. M. (2017). "Balancing Covariates via Propensity Score Weighting " Journal of the American Statistical Association 113(521): 390-400.

Li, L. and Greene, T. (2013). "A weighting analogue to pair matching in propensity score analysis." International Journal of Biostatistics 9(2): 215-234 DOI: 10.1515/ijb-2012-0030. https://www.ncbi.nlm.nih.gov/pubmed/23902694.

Lunceford, J. K. and Davidian, M. (2004). "Stratification and weighting via the propensity score in estimation of causal treatment effects: A comparative study." Statistics in Medicine 23(19): 2937-2960 DOI: 10.1002/sim.1903. https://www.ncbi.nlm.nih.gov/pubmed/15351954.

MacLehose, R. (2014). "Applications of Bayesian Methods to Epidemiologic Research." Current Epidemiology Reports 1(3): 103-109.

McCandless, L. C., Gustafson, P. and Austin, P. C. (2009). "Bayesian propensity score analysis for observational data." Statistics in Medicine 28(1): 94-112 DOI: 10.1002/sim.3460. https://www.ncbi.nlm.nih.gov/pubmed/19012268.

Mielke, P. W. and Berry, K. J. (2007). Permutation Methods: A Distance Function Approach. New York, New York, Springer.

Morgan, S. and Winship, C. (2007). Counterfactuals and Causal Inference: Methods and Principles for Social Research (Analytical Methods for Social Research). New York, Cambridge University Press.

Oakes, J. M. and Kaufman, J. S. (2017). Methods in Social Epidemiology. San Francisco, CA, John Wiley & Sons.

Otsu, T. and Rai, Y. (2017). "Bootstrap Inference of Matching Estimators for Average Treatment Effects." Journal of the American Statistical Association 112(520): 1720-1732 DOI: 10.1080/01621459.2016.1231613.

Pan, W. and Bai, H. (2015). Propensity Score Analysis. New York, New York, The Guilford Press.

Pan, W. and Bai, H. (2015). "Propensity score interval matching: using bootstrap confidence intervals for accommodating estimation errors of propensity scores." British Medical Journal; Medical Research Methodology 15: 53 DOI: 10.1186/s12874-015-0049-3. https://www.ncbi.nlm.nih.gov/pubmed/26215035.

Patorno, E., Glynn, R. J., Hernandez-Diaz, S., Liu, J. and Schneeweiss, S. (2014). "Studies with many covariates and few outcomes: selecting covariates and implementing propensity-score-based confounding adjustments." Epidemiology

25(2): 268-278 DOI: 10.1097/EDE.0000000000000069. https://www.ncbi.nlm.nih.gov/pubmed/24487209.

Patorno, E., Grotta, A., Bellocco, R. and Schneeweiss, S. (2013). "Propensity sore methodology for confounding control in health care utilization databases." Epidemiology Biostatistics and Public Health 10 (3): e89401-894016.

Pearl, J. (2010). "The foundations of causal inference " Sociological Methodology 40: 75-149.

Petersen, M. L., Porter, K. E., Gruber, S., Wang, Y. and van der Laan, M. J. (2012). "Diagnosing and responding to violations in the positivity assumption." Statistical Methods in Medical Research 21(1): 31-54 DOI: 10.1177/0962280210386207. https://www.ncbi.nlm.nih.gov/pubmed/21030422.

Petri, H. and Urquhart, J. (1991). "Channeling bias in the interpretation of drug effects." Statistics in Medicine 10(4): 577-581. https://www.ncbi.nlm.nih.gov/pubmed/2057656.

Rassen, J. A., Shelat, A. A., Myers, J., Glynn, R. J., Rothman, K. J. and Schneeweiss, S. (2012). "One-to-many propensity score matching in cohort studies." Pharmacoepidemiology and Drug Safety 21 Suppl 2: 69-80 DOI: 10.1002/pds.3263. https://www.ncbi.nlm.nih.gov/pubmed/22552982.

Ripollone, J. E., Huybrechts, K. F., Rothman, K. J., Ferguson, R. E. and Franklin, J. M. (2018). "Implications of the Propensity Score Matching Paradox in Pharmacoepidemiology." American Journal of Epidemiology DOI: 10.1093/aje/kwy078. https://www.ncbi.nlm.nih.gov/pubmed/29750409.

Rosenbaum, P. R. and Rubin, D. B. (1983). "The central role of the propensity score in observational studies for causal effects." Biometrika 70: 41-55.

Rosenbaum, P. R. and Rubin, D. B. (1984). "Reducing bias in observational studies using subclassification on the propensity score." Journal of the American Statistical Association 79: 516-524.

Rothman, K. J. (1986). Modern Epidemiology. Boston, MA, Little, Brown.

Rothman, K. J., Greenland, S. and Lash, T. L. (2008). Modern Epidemiology Philadelphia, PA, Lippincott Williams & Wilkins.

Rubin, D. B. (1979). "Using multivariate matched samplng and regression adjustment to control bias in observational studies." Journal of the American Statistical Association 74: 318-328.

Rubin, D. B. and Thomas, N. (1992). "Characterizing the Effect of Matching Using Linear Propensiy Score Methods with Normal Distributions." Biometrika 79: 797-809.

Rubin, D. B. and Thomas, N. (1996). "Matching using estimated propensity scores: relating theory to practice." Biometrics 52(1): 249-264. https://www.ncbi.nlm.nih.gov/pubmed/8934595.

Schneeweiss, S., Rassen, J. A., Glynn, R. J., Avorn, J., Mogun, H. and Brookhart, M. A. (2009). "High-dimensional propensity score adjustment in studies of treatment effects using health care claims data." Epidemiology 20(4): 512-522 DOI: 10.1097/EDE.0b013e3181a663cc. https://www.ncbi.nlm.nih.gov/pubmed/19487948.

Schneeweiss, S., Solomon, D. H., Wang, P. S., Rassen, J. and Brookhart, M. A. (2006). "Simultaneous assessment of short-term gastrointestinal benefits and cardiovascular risks of selective cyclooxygenase 2 inhibitors and nonselective nonsteroidal antiinflammatory drugs: an instrumental variable analysis." Arthritis and Rheumatology 54(11): 3390-3398 DOI: 10.1002/art.22219. https://www.ncbi.nlm.nih.gov/pubmed/17075817.

Spiegelhalter, D. J., Abrams, K. R. and Myles, J. P. (2004). Bayesian Approaches to Clinical Trials and Health-Care Evaluation. West Sussex, England, John Wiley & Sons Ltd.

Stuart, E. A. (2010). "Matching methods for causal inference: A review and a look forward." Statistical Science 25(1): 1-21 DOI: 10.1214/09-STS313. https://www.ncbi.nlm.nih.gov/pubmed/20871802.

Tu, W. and Zhou, X. (2002). "A Bootstrap Confidence Interval Procedure for the Treatment Effect Using Propensity Score Sublassification " Health Services & Outcomes Research Methodology 3: 135-147.

Vaughan, L. K., Divers, J., Padilla, M., Redden, D. T., Tiwari, H. K., Pomp, D. and Allison, D. B. (2009). "The use of plasmodes as a supplement to simulations: A simple example evaluating individual admixture estimation methodologies." Computational Statistics & Data Analysis 53(5): 1755-1766 DOI: 10.1016/j.csda.2008.02.032. https://www.ncbi.nlm.nih.gov/pubmed/20161321.

Wu, S., Ding, Y., Wu, F., Hou, J. and Mao, P. (2015). "Application of propensity-score matching in four leading medical journals." Epidemiology 26(2): e19-20 DOI: 10.1097/EDE.0000000000000249. https://www.ncbi.nlm.nih.gov/pubmed/25643113.

Zhao, Z. (2004). "Using matching to estimate treatment effects: data requirements, matching metrics, and Monte Carlo evidence." Review of Economics and Statistics 86(1): 91-107.

Zigler, C. M. (2016). "The Central Role of Bayes' Theroem for Join Estimation of Causal Effects and Propensity Scores." The American Statistician 70(1): 47-54.

CURRICULUM VITAE

[redacted]

[redacted]

[redacted]