Statistics
in Medicine

Research Article

Received 21 June 2012, Accepted 7 October 2013 Published online 22 October 2013 in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.6030

On the joint use of propensity and prognostic scores in estimation of the average treatment effect on the treated: a simulation study

Finbarr P. Leacy[a*†] and Elizabeth A. Stuart[b]

Propensity and prognostic score methods seek to improve the quality of causal inference in non-randomized or observational studies by replicating the conditions found in a controlled experiment, at least with respect to observed characteristics. Propensity scores model receipt of the treatment of interest; prognostic scores model the potential outcome under a single treatment condition. While the popularity of propensity score methods continues to grow, prognostic score methods and methods combining propensity and prognostic scores have thus far received little attention. To this end, we performed a simulation study that compared subclassification and full matching on a single estimated propensity or prognostic score with three approaches combining the estimated propensity and prognostic scores: full matching on a Mahalanobis distance combining the estimated propensity and prognostic scores (FULL–MAHAL); full matching on the estimated prognostic propensity score within propensity score calipers (FULL–PGPPTY); and subclassification on an estimated propensity and prognostic score grid with 5×5 subclasses (SUBCLASS(5*5)). We considered settings in which one, both, or neither score model was misspecified. The data generating mechanisms varied in the degree of linearity and additivity in the true treatment assignment and outcome models. FULL–MAHAL and FULL–PGPPTY exhibited strong to superior performance in root mean square error terms across all simulation settings and scenarios. Methods combining propensity and prognostic scores were no less robust to model misspecification than single-score methods even when both score models were incorrectly specified. Our findings support the joint use of propensity and prognostic scores in estimation of the average treatment effect on the treated. Copyright © 2013 John Wiley & Sons, Ltd.

Keywords: observational data; confounding; bias reduction; subclassification; full matching; Mahalanobis metric

1. Introduction

Propensity and prognostic score methods seek to improve the quality of causal inference in non-randomized or observational studies by replicating the conditions found in a controlled experiment, at least with respect to observed characteristics. In particular, these approaches attempt to induce between the observed covariate distributions of treatment groups a comparability similar to that obtained under the random assignment of treatment conditions. Importantly, propensity and prognostic score approaches cannot ensure balance on unobserved background characteristics and consequently causal inferences based on them carry an assumption of no unobserved confounding.

The propensity score is defined as the probability of treatment receipt conditional on the observed covariates [1]. It is a balancing score; the observed covariate distributions of treatment groups are identical at each value of the score. Under suitable unconfoundedness and overlap conditions, discussed further in section two, conditioning on the propensity score allows identification of both the average

[a]*MRC Biostatistics Unit, Cambridge Institute of Public Health, Forvie Site, Robinson Way, Cambridge Biomedical Campus, Cambridge CB2 0SR, U.K.*
[b]*Departments of Mental Health and Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, U.S.A.*
**Correspondence to: Finbarr P. Leacy, MRC Biostatistics Unit, Cambridge Institute of Public Health, Forvie Site Robinson Way, Cambridge Biomedical Campus, Cambridge CB2 0SR, U.K.*
[†]*E-mail: finbarr.leacy@mrc-bsu.cam.ac.uk*

treatment effect on the treated (ATT), that is, the difference in potential outcomes amongst those receiving the treatment, and the average treatment affect (ATE), the difference in potential outcomes amongst those receiving treatment or control [2, 3]. Such conditioning may take the form of matching, subclassification or weighting [2]. Since their introduction in the early 1980s, propensity score methods have enjoyed substantial attention in the causal inference literature and wide application across a wide number of research fields including education, sociology, psychology, economics, and public health [4–15].

Hansen's prognostic score [16] formalizes a second type of experimental control not exploited by the propensity score and thus offers an alternative approach to causal inference in observational data. This second type of experimental control, Hansen notes, seeks to standardize the outcome generation process, inducing comparability between the observed covariate distributions of trial participants with differing potential outcomes by minimizing or eliminating associations between the covariates and trial outcomes [16]. A prognostic score is also a balancing score and is defined as any scalar or vector-valued function of the covariates that when conditioned on induces independence between the potential outcome under the control condition and the unreduced covariates. For example, if the potential outcome under control follows a generalized linear model, one possible prognostic score is given by the conditional expectation of the outcome under the control condition given the observed covariates. Similar to the propensity score, in practice, the true prognostic score is unlikely to be known and so must be estimated.

Examples of prognostic score use in the biomedical and statistical literatures have been limited almost exclusively to the binary outcome case. In this setting, the score has commonly been referred to as a disease risk score. Disease risk scores have been estimated in two ways: in the nomenclature of Arbogast and Ray [17], 'the unexposed-only' disease risk score is estimated in the unexposed group only, while the 'full-cohort' disease risk score is estimated in the entire sample. The unexposed-only disease risk score is equivalent to a prognostic score while the full-cohort disease risk score is equivalent to the binary outcome form of Miettinen's multivariate confounder score [18]. Arbogast and Ray have undertaken a critical review of disease risk score use in pharmacoepidemiologic studies [19] and have explored its performance relative to traditional regression adjustment in a simulation study [17]. They have also utilized the score in a number of studies of cardiovascular risk [20–26]. Arbogast and Ray motivate the disease risk score primarily as a dimension reduction technique and advocate adjusting for it directly in regression models relating the outcome and treatment status. While Sturmer *et al.* [27] and Caderette *et al.* [28] also advocate use of the disease risk score as a summary measure, using the prognostic score in this manner may be suboptimal; it does not take advantage of the balancing property of the prognostic score and renders the degree of model extrapolation between treatment and control groups unchanged [2]. Finally, Glynn *et al.* provide an excellent discussion on the relative merits of propensity and disease risk score use in comparative effectiveness research for emerging therapies in the pharmacoepidemiologic context [29]. To our knowledge, very few, if any, studies published to date have performed analyses using subclassification or matching on the prognostic score alone or examined the performance of these methods in simulation studies.

Propensity and prognostic score methods represent complementary approaches to causal inference in observational studies. Matching on the propensity score tends to produce balance in, and consequently removes bias due to, covariates highly predictive of treatment assignment. Matching on the prognostic score tends to produce balance in covariates highly predictive of the outcome. While use of the outcome data is intrinsic to prognostic score methodology, propensity score methods purposefully disregard the association between the covariates and the outcome. While this separation of design and analysis has been advocated as one of the propensity score's greatest strengths, it is not without disadvantages: recent work has suggested that failure to include in the propensity score model a variable that is highly predictive of the outcome but not associated with treatment status can lead to increased bias and decreased precision in treatment effect estimates in some settings [30]. Further work has indicated the existence of a class of variables that when conditioned on tend to amplify rather than reduce bias: this includes variables strongly associated with treatment status but weakly associated with outcome [31, 32]. Classical implementations of propensity score methods cannot ensure balance on prognostically relevant variables unless these variables are also associated with treatment assignment. While a number of ad hoc adjustments to overcome this drawback have been proposed, including, for example, Mahalanobis metric matching on a subset of prognostically important covariates within propensity score calipers [33], these rely heavily on substantive knowledge of potential confounders. On the other hand, while a particular study will typically have one treatment or exposure of interest and consequently one propensity score, there may be multiple outcomes of interest and thus multiple prognostic scores. A single subclassification or matching procedure can be performed under a propensity score approach; under the prognostic

score approach, separate matching or subclassification procedures must be performed for each outcome of interest.

The complementary nature of propensity and prognostic score methods suggests, at least conceptually, that conditioning on both the propensity and prognostic scores may have the potential to reduce bias and/or improve the precision of treatment effect estimates beyond that achieved by single score methods. To this end, the simulation study presented in this article examines the performance of three methods that seek to estimate the ATT by conditioning on both the propensity and prognostic scores: full matching on a Mahalanobis distance combining the estimated propensity and prognostic scores; full matching on the estimated prognostic propensity score within propensity score calipers; and subclassification on an estimated propensity and prognostic score grid with 5×5 subclasses. It compares these approaches with a number of single score methods and with standard main effects linear regression. Method performance is examined across 25 data generating mechanisms that vary in the degree of linearity and additivity in the true treatment assignment and outcome models and in settings in which one, both, or neither of the propensity and prognostic score models is misspecified. Two of the three combination methods investigated have been proposed previously: full matching on a Mahalanobis distance combining the estimated propensity and prognostic scores and full matching on the estimated prognostic propensity score within propensity score calipers were first proposed by Hansen [34] but their comparative performance has not yet been investigated in a simulation study. Subclassification on a propensity and prognostic score grid with $k \times k$ subclasses has not been examined previously.

The remaining sections of this article are organized as follows: Section 2 provides a brief introduction to subclassification and full matching; Section 3 provides a brief summary of prognostic score theory and a detailed description of the three approaches combining propensity and prognostic scores examined in our simulation study; Section 4 provides detailed explanations of the simulation setup and the propensity and prognostic score estimation procedures employed; Section 5 presents the simulation results in detail. Section 6 discusses the simulation results and highlights directions for future research.

2. Estimation of treatment effects via subclassification or full matching

Every propensity and prognostic score method examined in this article first produces a partition of the data into a number of matched sets with similar values of a given distance measure, e.g., propensity score quintiles [35]. Each matched unit then receives a weight that depends on both the estimand of interest and on the relative number of treatment and control group units in the matched set to which it belongs. Finally, these weights are incorporated into the outcome analysis from which the treatment effect is estimated.

2.1. Production of matched sets

Subclassification and full matching represent two ways in which the matched sets can be formed. Subclassification forms matched sets by direct partition along the distance measures of the whole sample or of a single treatment group: in this article, we form subclasses with equal numbers of treated units. Full matching, introduced by Rosenbaum [36], partitions a sample of units into a collection of matched sets, each of which contains at least one treatment group member and at least one control group member [37]. Full matching is an optimal matching method; it minimizes the average of the distances between each treatment and control group member across each of the matched sets. It typically produces a much larger number of matched sets than subclassification and thus represents a finer partition of the data. The quality of matches can be further refined by imposing a caliper restriction on potential matches. Such a restriction prevents the matching of units whose estimated propensity and/or prognostic scores differ by more than a pre-specified value. Units are discarded if there exists no unit in the opposite treatment group falling within their caliper. Caliper restrictions will typically introduce a trade-off between bias and estimand consistency. Increasing caliper width to ensure that all treated units are matched or removing the caliper restriction will preserve the estimand but may be associated with increased bias; decreasing the caliper width may reduce bias but change the estimand from the ATT to the effect of the treatment on treated units with a certain propensity to receive the treatment.

2.2. Derivation of weights

When the estimand of interest is the ATT, each matched treated unit receives unit weight, while each matched control unit receives a weight proportional to the number of treated units divided by the

number of control units in the matched set to which it belongs. Unmatched units receive zero weight. The weights of the control group units are scaled such that their sum equals the total number of matched control units. These weights serve to weight the control group to resemble the treatment group.

2.3. Estimator form

Every subclassification and matching estimator examined in this simulation study can be written in the following form

$$\hat{\gamma} = \sum_{i=1}^{N} Z_i Y_i \Big/ \left(\sum_{i=1}^{N} Z_i\right) - \sum_{i=1}^{N} (1 - Z_i) M_{ij} \frac{N_{1j}}{N_{0j}} Y_i \Big/ \left(\sum_{i=1}^{N} (1 - Z_i) M_{ij} \frac{N_{1j}}{N_{0j}}\right)$$

where Z_i is an indicator function of treatment receipt, Y_i is the observed outcome, W_i is the assigned weight, M_{ij} indicates membership of subject i in subclass or matched set j, N_{1j} is the number of treated units in subclass or matched set j, and N_{0j} is the number of control units in subclass or matched set j. Each estimator depends on the propensity and/or prognostic score through its/their influence on matched set composition.

The robust model-based standard errors associated with these estimators are approximations to the true standard error (SE). This is because they fail to account for the fact that the true values of the propensity and prognostic scores from which the weights are derived are not known but instead are estimated in the first step. Recent work in the propensity score literature [38] suggests that when estimating the ATE via inverse probability of treatment weighting, some routinely used model-based estimators of the SE produce estimates that are almost always too large. In the subclassification case, such estimators will yield particularly conservative inferences when the propensity score model contains variables that are strongly related to the outcome but only weakly related to treatment assignment. Variance estimation in these contexts is the subject of ongoing research.

3. Propensity and prognostic score subclassification: theory and practice

This section first provides a brief overview of some important properties of the prognostic score. It then examines the relationship between prognostic scores and Miettinen's multivariate confounder score. It proceeds to present some theoretical arguments underpinning the joint use of propensity and prognostic scores in estimation of treatment effects before finally providing detailed descriptions of each of the three combination methods examined in the simulation study.

3.1. Prognostic score theory

Following Hansen's notation, in the discussion that follows let Y_c denote the potential outcome under control, Y_t the potential outcome under treatment, Y the observed outcome, Z the treatment or exposure of interest, and $\boldsymbol{X}$ the set of unreduced covariates. Assuming that Y_c given $\boldsymbol{X}$ follows a generalized linear model, let $\phi(\boldsymbol{X}) \equiv E(Z|\boldsymbol{X} = \mathbf{x})$ denote the true propensity score and $\Psi(\boldsymbol{X}) \equiv E(Y_c|\boldsymbol{X} = \mathbf{x})$ denote one of the possible true prognostic scores [16].

Similar to the propensity score, the prognostic score is, by definition, a balancing score, $Y_c \perp X|\Psi(\boldsymbol{X})$. It can be scalar or vector-valued. Conditioning on a prognostic score will tend to produce balance in those covariates most strongly associated with the outcome, removing systematic associations between these covariates and the outcome, and thus reducing bias in treatment effect estimates. The prognostic score can be defined for binary, continuous, categorical, and ordinal outcome variables and can be estimated using standard regression modeling techniques. If substantial non-linearity or non-additivity in the response surface is anticipated, machine learning methods such as CART, boosted CART, or random forests may be fruitfully employed. Hansen advocates estimating prognostic scores in a single treatment group and not the whole sample.

Under the assumption that $\boldsymbol{X}$ deconfounds the potential outcome under control and treatment assignment, $Y_c \perp Z|\boldsymbol{X}$, potential responses and treatment assignment are deconfounded by subclassification on the prognostic score, $Y_c \perp Z|\Psi(\boldsymbol{X})$. Prognostic score subclassification is sufficient for identification of the ATT by proposition 4 of Hansen [16]

$$E(Y_t - Y_c|Z = 1) = E[E\{Y|Z = 1, \Psi(\boldsymbol{X})\} - E\{Y|Z = 0, \Psi(\boldsymbol{X})\}|Z = 1]$$

Furthermore, prognostic score subclassification allows estimation of the ATT under a weaker condition than propensity score subclassification. Valid estimation of the ATT using subclassification on the propensity score requires weak overlap on the unreduced covariate: that is, there should be no set of covariate values associated with certain treatment receipt, $pr\left[pr\left\{Z=1|\boldsymbol{X}\right\}<1\right]=1$. Valid estimation of the ATT using subclassification on the prognostic score requires only weak overlap on the prognostic score: that is, there should be no prognostic score stratum in which the treatment is received with probability one, $pr\left[pr\left\{Z=1|\Psi\left(\boldsymbol{X}\right)\right\}<1\right]=1$.

Propensity score subclassification is sufficient for valid estimation of the ATE under weak unconfoundedness and strong overlap on the unreduced covariate, $pr\left[0<pr\left\{Z=1|\boldsymbol{X}\right\}<1\right]=1$. In contrast, proposition 5 of Hansen demonstrates that identification of the ATE under prognostic score subclassification requires weak unconfoundedness, strong overlap on the prognostic score, $pr\left[0<pr\left\{Z=1|\Psi\left(\boldsymbol{X}\right)\right\}<1\right]=1$ and additional subclassification on any effect modifiers, $m(\boldsymbol{X})$:

$$E(Y_t-Y_c)=E[E\left\{Y|Z=1,\Psi\left(\boldsymbol{X}\right),m\left(\boldsymbol{X}\right)\right\}-E\left\{Y|Z=0,\Psi\left(\boldsymbol{X}\right),m\left(\boldsymbol{X}\right)\right\}]$$

Effect modification is present if there exists at least one prognostic score for Y_c that is not also a prognostic score for Y_t. This is a broad definition of effect modification that encompasses any distributional differences such that $pr\left(Y_t|Y_c,X\right)\neq pr\left(Y_t|Y_c\right)$. $m\left(\boldsymbol{X}\right)$ is an effect modifier if both the condition for effect modification is satisfied and $Y_t\perp X|\Psi\left(\boldsymbol{X}\right),m\left(\boldsymbol{X}\right)$ for every prognostic score $\Psi\left(\boldsymbol{X}\right)$.

3.2. *Comparison with Miettinen's multivariate confounder score*

Although similar in character, Miettinen's multivariate confounder score [18] is less general than Hansen's prognostic score. Miettinen's score is defined as the conditional expectation of the outcome in the control group given the observed covariates, $E[Y|Z=0,\boldsymbol{X}]$. Miettinen's score will correspond to a prognostic score if and only if $Y_c\perp\boldsymbol{X}|E[Y|Z=0,\boldsymbol{X}]$. In particular, this implies that all higher order moments of Y_c depend on $\boldsymbol{X}$ only through $E[Y|Z=0,\boldsymbol{X}]$. For example, suppose Y_c follows a linear regression on $\boldsymbol{X}$. Miettinen's score will correspond to a prognostic score if the variance of Y_c is constant or depends on $\boldsymbol{X}$ only through $E[Y|Z=0,\boldsymbol{X}]$. If the variance of Y_c depends on $\boldsymbol{X}$ in some other way, then every possible prognostic score consists of some transformation of the mean and variance functions [16].

In contrast to the prognostic scores estimated using a single treatment group, Miettinen's multivariate confounder score is estimated using the full sample. This difference has important implications for partial ancillarity of the estimated scores and renders Miettinen's score less robust to model misspecification. Given that

$$p\left(Y_t,Y_c,Z,\boldsymbol{X}\right)=pr\left(Y_t|Y_c,Z,\boldsymbol{X}\right)p\left(Y_c,Z,\boldsymbol{X}\right)=pr\left(Y_t|Y_c,\boldsymbol{X}\right)p\left(Y_c,Z,\boldsymbol{X}\right)$$

under weak unconfoundedness, true prognostic scores and prognostic scores estimated using a correctly or incorrectly specified regression model in a single treatment group are appropriate statistics on which to condition because they are partial ancillaries for the parameter of interest $pr\left(Y_t|Y_c,\boldsymbol{X}\right)$ [16]. Miettinen's score, and prognostic scores estimated using the full sample are generally not ancillary in this sense. For example, suppose Y_t and Y_c follow linear regressions on $\mathbf{X}$. Miettinen's score is guaranteed to be a partial ancillary for $pr\left(Y_t|Y_c,\boldsymbol{X}\right)$ only when the true response surfaces for Y_t and Y_c are parallel, the error structure does not vary with $\boldsymbol{X}$, and the regression model for Y has been correctly specified [34]. In separate simulation studies, Pike *et al.* [39] and Cook and Goldman [40] found that subclassification on Miettinen's score exaggerated the statistical significance of treatment effect estimates when the conditions for partial ancillarity were violated.

3.3. *Combining propensity and prognostic scores: underlying theory*

Prognostic score theory supports conditioning on both the propensity and prognostic scores; it asserts that conditioning on a prognostic score and any other function of the covariates allows unbiased estimation of treatment effects whenever conditioning on the prognostic score alone does. In addition to bias reduction, subclassification or matching on both the propensity and prognostic scores may have other benefits. Hansen suggests that subclassification or matching on both scores may reduce extrapolation when the degree of covariate separation between treatment and control units is large [16]. In practice, conditioning on both scores may also be associated with increased precision in treatment effect estimates and with improved robustness to prognostic and/or propensity score model misspecification.

In a technical report [34], Hansen derives first-order approximations to the bias of direct adjustment when the sample data have first been partitioned into subclasses or matched sets according to one or both of the estimated propensity and prognostic scores or according to some other function of the covariates. In the subclassification case the bias is approximately given by the following expression [34, Equation 13]:

$$\sum_s \frac{n_s}{n_s - 1} \{(1 - \hat{p}_s)\, cov\,(\mu_{ts.}, \theta_{s.}) + \hat{p}_s cov\,(\mu_{cs.}, \theta_{s.})\}$$

where n_s is the number of units in subclass s, $\hat{p}_s$ is the proportion of treated units in subclass s, $\mu_{ts.}$ is the average true prognostic score for the potential outcome under treatment for units in subclass s, $\mu_{cs.}$ is the average true prognostic score for the potential outcome under control for units in subclass s, and $\theta_{s.}$ is the average true propensity score on the logit scale for units in subclass s. This expression suggests that bias will be minimized if there is little variation in the true propensity or prognostic scores within subclasses, and/or the propensity and prognostic scores are uncorrelated within subclasses [34].

In the matching case, Hansen derives the following first-order approximation to the bias of direct adjustment under the assumption of a constant treatment effect across study units [34, Equation 15]:

$$\sum_s \frac{n_s}{n_s - 1} cov\,(\mu_{cs.}, \theta_{s.})$$

This approximation suggests that, within matched sets, bias is more strongly determined by variation in the true propensities on the logit scale rather than on the probability scale and by the standard deviations of $\mu_{cs.}$ and $\theta_{s.}$. Hansen proposes minimizing the bias within matched sets by matching on a Mahalanobis distance combining the estimated propensity and prognostic scores. He considers performing the matching first with no additional caliper restrictions and then within propensity or prognostic score calipers. All three approaches performed well in a simulation study included in the technical report where they were associated with smaller SEs than a number of propensity score matching procedures while maintaining nominal coverage rates. Our paper builds on this previous work by comparing the performance of a number of methods combining the estimated propensity and prognostic scores with that of single score methods, across a large number of data generating mechanisms and under correct and incorrect specification of the score models.

3.4. Approaches combining propensity and prognostic scores

3.4.1. Full matching on a Mahalanobis distance combining the estimated propensity and prognostic scores. This method, first proposed by Hansen [34], builds on a large body of knowledge associated with matching on Mahalanobis distances formed from the unreduced covariates. For individuals i and j with estimated propensity and prognostic score information $\begin{pmatrix} \hat{\phi}_i \\ \hat{\psi}_i \end{pmatrix}$ and $\begin{pmatrix} \hat{\phi}_j \\ \hat{\psi}_j \end{pmatrix}$, the Mahalanobis distance D_{ij} is defined as

$$D_{ij} = \left(\begin{pmatrix} \hat{\phi}_i \\ \hat{\psi}_i \end{pmatrix} - \begin{pmatrix} \hat{\phi}_j \\ \hat{\psi}_j \end{pmatrix}\right)' \Sigma^{-1} \left(\begin{pmatrix} \hat{\phi}_i \\ \hat{\psi}_i \end{pmatrix} - \begin{pmatrix} \hat{\phi}_j \\ \hat{\psi}_j \end{pmatrix}\right)$$

where Σ is the variance-covariance matrix of $\begin{pmatrix} \hat{\phi} \\ \hat{\psi} \end{pmatrix}$ in the full control group [2]. Mahalanobis distance matching has been found to perform well when there are relatively few covariates and these covariates are approximately normally distributed [2]. For this reason, as for all the methods examined in this article, the estimated propensity score is transformed to the scale of the linear predictor prior to matching.

3.4.2. Full matching on the prognostic propensity score within propensity score calipers. The prognostic propensity or 'focused' propensity score was proposed by Hansen as a means to improve propensity balance on prognostically relevant variables and reduce bias in estimated treatment effects [16]. The prognostic score for the outcome of interest is estimated first. This estimated prognostic score is then included as the sole predictor in the propensity score model or as an additional predictor in a propensity score model relating the treatment assignment variable to the observed covariates. The fitted values from this propensity score model are the prognostic propensity scores. Hansen notes that the degree of model

dependence inherent in the prognostic score estimation procedure is increased the greater the separation of treatment and control group units on the observed covariates. He proposes that such dependence may be counteracted by performing full matching on the prognostic propensity score within an ordinary propensity score caliper [16]. An application of this method to the SAT coaching study of Powers and Rock can be found in [41]. We utilize a caliper of 0.2 pooled standard deviations of the estimated propensity score in our simulation study. Across every data generating mechanism and propensity/prognostic score specification setting considered in our simulation study on average fewer than 5% of treated units were discarded using this caliper width.

3.4.3. Subclassification on a propensity and prognostic score grid with $k \times k$ subclasses. Subclassification on a propensity and prognostic score grid with k × k subclasses is proposed here as an alternative means of combining the propensity and prognostic scores. It is inspired by the propensity function of Imai and van Dyk [42]. Imai and van Dyk use this method to estimate the effects of smoking on medical expenditure [42], where they treat the duration and frequency of smoking as a bivariate treatment, estimating a separate propensity score for each.

The propensity and prognostic scores are first estimated. The data are then divided into an n × m grid of subclasses based on the quantiles of the estimated propensity and prognostic scores for pre-specified n and m values. In practice, these values could be chosen in an iterative process where the number of subclasses is varied until an acceptable level of propensity and/or prognostic score balance on each covariate is attained (while maintaining sufficient sample sizes in each subclass). Propensity score balance can be checked by comparing the distribution of each covariate across treatment groups. Prognostic score balance can be checked by testing for association between the outcome and each covariate in the control group along subclasses of the prognostic score. The ATT is calculated as a weighted average of the within-subclass estimates. Previous simulation work examining the performance of subclassification on the propensity function has suggested that for sufficiently large sample sizes increasing the total number of subclasses is associated with greater reductions in bias [42]. Preliminary simulation work, not reported here, suggests that this observation also holds for subclassification on a propensity and prognostic score grid with k × k subclasses.

4. Methods

4.1. Objective

The goal of this study was to investigate the comparative performance of a number of propensity and prognostic score methods used for estimating the ATT, $ATT = E[Y_t - Y_c|Z = 1]$, when one, both, or neither of the propensity and prognostic score models was incorrectly specified.

The three methods combining estimated propensity and prognostic scores considered here may differ in their sensitivity to misspecification of the propensity and/or prognostic score models. In particular, full matching on a Mahalanobis distance combining the estimated propensity and prognostic scores and subclassification on an estimated propensity and prognostic score grid with k × k subclasses may be less sensitive to misspecification of one of the score models than full matching on the prognostic propensity score within propensity score calipers: because because the estimated prognostic score is used as a covariate in estimating the prognostic propensity score correct specification of the propensity score model will not be sufficient if the prognostic score has been misspecified.

The following eight methods were considered:

- A main effects linear regression model relating the outcome to exposure status and the covariates with no propensity or prognostic score weighting. The correct functional form for the linear regression is used only in the setting where both the propensity and prognostic score models are correctly specified. Otherwise, a (misspecified) main effects model is used (REG)
- Subclassification on the estimated propensity score with five subclasses (SUBCLASS(5)–PROP)
- Full matching on the estimated propensity score (FULL–PROP)
- Subclassification on the estimated prognostic score with five subclasses (SUBCLASS(5) –PROG)
- Full matching on the estimated prognostic score (FULL–PROG)
- Subclassification on an estimated propensity and prognostic score grid with 5 × 5 subclasses (SUBCLASS(5*5))
- Full matching on a Mahalanobis distance combining the estimated propensity and estimated prognostic scores (FULL–MAHAL)

- Full matching on the estimated prognostic propensity score within propensity score calipers using a caliper of 0.2 pooled standard deviations of the estimated propensity score (FULL–PGPPTY)

The relative performance of methods combining the estimated propensity and prognostic scores and methods utilizing a single score is of particular interest in settings in which only one score model has been correctly specified and consequently forms the focus of this article's discussion.

4.2. *Simulation setup*

The structure of the simulation study was adapted from previous studies conducted by Setoguchi *et al.* [43] and Lee *et al.* [44]. Ten mean zero, unit variance normal covariates, $W_1, W_2, \ldots, W_{10}$ with covariance structure as illustrated in Figure 1 were initially generated. Five of these covariates $(W_1, W_3, W_5, W_6, W_8, W_9)$ were then dichotomized, resulting in slight attenuation of the reported correlations. Of these covariates, four $(W_1 - W_4)$ were confounders, three $(W_5 - W_7)$ were predictors of treatment assignment only, and three $(W_8 - W_{10})$ were predictors of outcome only.

Method performance was evaluated across 25 data generating mechanisms that varied in the degree of linearity and additivity in the true treatment assignment and outcome models. A binary treatment and a continuous outcome were considered. The binary treatment variable, Z, was generated according to a logistic model of the form $logit\{P[Z = 1|\boldsymbol{W}]\} = g(\boldsymbol{W})$. The baseline treatment prevalence was fixed at 0.2 for all scenarios. All continuous outcomes, Y, were generated according to a normal model subject to a constant additive treatment effect of $\gamma = -0.4$ and a residual standard deviation of $\sigma = 0.1$. All outcome models were thus of the following form: $Y = f(\mathbf{W}) + \gamma Z + \varepsilon$, $\varepsilon \sim N(0, 0.1^2)$

Cohort studies of size $n = 1000$, 2000, and 5000 were examined. One thousand datasets of each study size were generated for each of the five propensity score models. In each of these datasets, five outcome variables were then generated subject to the single treatment assignment variable. Thus, five sets of datasets of each study size were generated, each with one treatment assignment variable, ten covariates, and five outcome variables.

The five treatment assignment/propensity score and outcome models can be characterized as follows:

(a) additivity and linearity (main effects only)
(b) moderate non-additivity (ten two-way interaction terms)
(c) mild non-additivity and non-linearity (one quadratic term and three two-way interaction terms)
(d) moderate non-linearity (three quadratic terms)
(e) moderate non-additivity and non-linearity (three quadratic terms and ten two-way interaction terms.

The exact functional forms of these models can be found in Appendix A. Average R^2 values for the five outcome models across the 1000 simulated datasets generated under propensity score model C with $N = 2000$ were 0.92, 0.82, 0.61, 0.51, and 0.59, respectively. These values varied little across the five possible propensity score models.

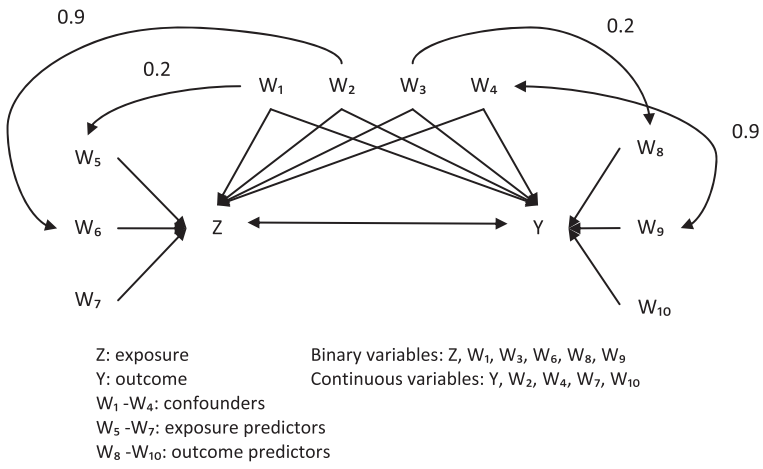

Figure 1. Simulation study data structure. (Reproduced with minor alterations from Lee *et al.* [44]).

4.3. Propensity and prognostic score estimation

Propensity scores were estimated using standard logistic regression. Incorrectly specified propensity score models consisted of main effect terms for each of the ten covariates; correctly specified propensity score models utilized the correct functional form. All estimated propensity scores were transformed to the scale of the linear predictor prior to subclassification or matching.

Prognostic scores were estimated using standard linear regression. Incorrectly specified prognostic score models consisted of main effect terms for each of the ten covariates; correctly specified prognostic score models utilized the correct functional form. All prognostic scores were estimated using the control group only, with predictions for the treated group obtained using the resulting fit.

We note that for true propensity score model A and for true prognostic score model A, linear regression models with main effects for each of the ten covariates are not in fact misspecified but overspecified.

4.4. Estimation of average treatment effect on the treated

Estimation of the ATT was a three-stage process. In the first stage, the data were partitioned into a collection of subclasses or matched sets according to the estimated propensity and/or prognostic score of each observation. In the second stage, weights for each observation in the simulated dataset were derived on the basis of matched set membership. In the final stage, estimates of the ATT and model-based approximations to its SE were obtained via weighted linear regression of the outcome on treatment assignment and use of the sandwich variance estimator.

All subclassification and matching procedures were conducted using the R statistical software program [45]; subclassification and full matching were implemented using the *MatchIt* [46,47] and *optmatch* [48] R packages respectively. Estimates of the ATT and model-based approximations to its SE were obtained using the R *survey* package [49].

4.5. Performance metrics

Method performance across each of the (5 treatment assignment models)*(5 outcome models)*(4 score model specification combinations)*(3 cohort sizes) = 300 simulation scenarios was evaluated according to the following standard criteria: mean percent bias of the effect estimate (BIAS), mean model-based standard error of the effect estimate (SE), root mean square error of the effect estimate (RMSE), and 95% confidence interval (CI) coverage rates (COVERAGE). We also considered the empirical standard error of the estimators, calculated as the standard deviation of the effect estimates across the 1000 simulation scenario replications (EMP-SE). All performance metrics were calculated with respect to the true ATT, $\gamma = -0.4$.

5. Results

Results are presented here for simulation settings with sample size $N = 2000$. While the output of simulations in which only one of the score models has been correctly specified forms the focus of attention, a brief overview of the results for simulation settings in which both score models are correctly or incorrectly specified is also provided. Plots of the percent bias, empirical SE, model-based SE, 95% CI coverage rates, and RMSE for all four specifications of the propensity and prognostic score models can be found in the online supplementary materials.

5.1. Propensity score model incorrectly specified, prognostic score model correctly specified

5.1.1. Effect estimates - percent bias and standard error. As illustrated in Figure 2, full matching on the estimated prognostic score (FULL–PROG) performed best across all 25 data generating mechanisms, exhibiting close to zero bias in all scenarios. Full matching on a Mahalanobis distance combining the estimated propensity and prognostic scores (FULL–MAHAL) and full matching on the prognostic propensity score (FULL–PGPPTY) also exhibited excellent performance. While subclassification on an estimated propensity and prognostic score grid with 5×5 subclasses (SUBCLASS(5*5)) outperformed subclassification approaches using a single estimated score in all scenarios, it did not match the performance of FULL–MAHAL or FULL–PGPPTY. For a fixed true propensity score model, percent bias for full matching on the estimated propensity score (FULL–PROP) and subclassification on the estimated propensity score (SUBCLASS(5)–PROP) increased with the degree of non-linearity in the true outcome

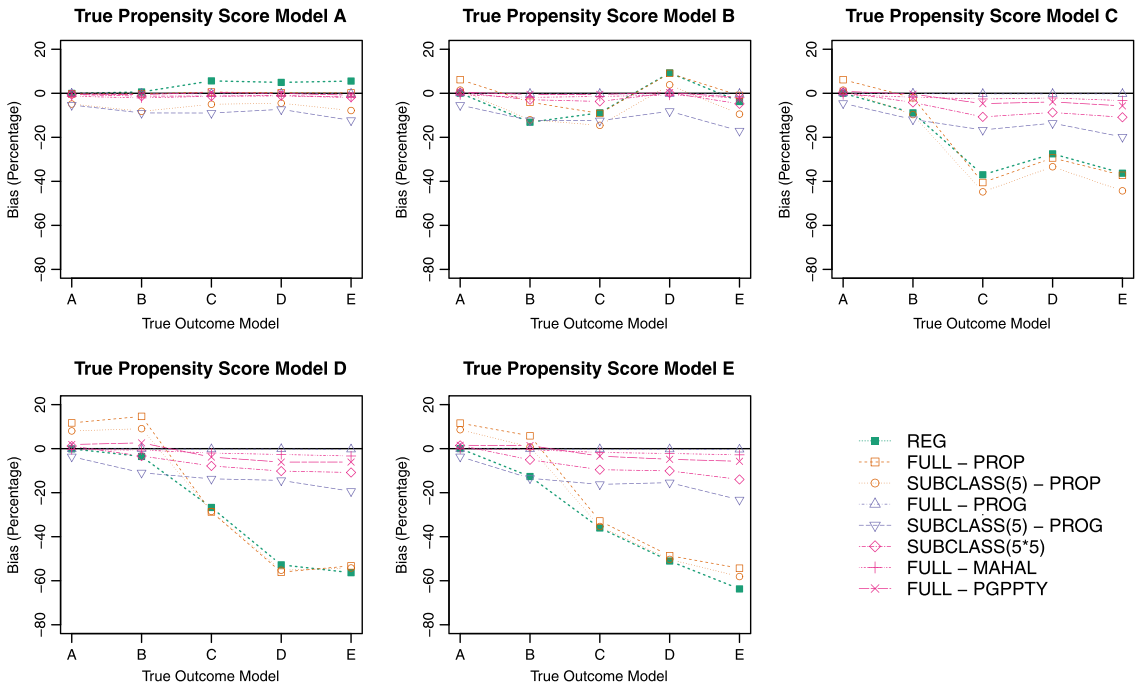

Figure 2. Percent bias in average treatment effect on the treated estimate by propensity/prognostic score method for an incorrectly specified propensity score model, $N = 2000$.

model. In the presence of non-linearity in the true outcome model, these methods performed similarly to a misspecified standard main effects linear regression (REG).

A consistent pattern in the model-based SE values was observed across the 25 data generating mechanisms; REG exhibited the smallest model-based SE in all cases, followed by SUBCLASS(5)–PROG, FULL–PROG, SUBCLASS(5)–PROP, and SUBCLASS(5*5). FULL–MAHAL, FULL–PGPPTY, and FULL–PROP exhibited the largest model-based SEs. For all methods, model-based SE values tended to increase with the degree of non-linearity and non-additivity in the true treatment assignment and outcome models.

As theoretically predicted [50], the empirical SEs for all methods other than REG were uniformly smaller than the corresponding model-based SEs. Combination methods exhibited the smallest empirical SE across all 25 data generating mechanisms. PROP–FULL exhibited the largest empirical SE. While the empirical SEs of the propensity score methods tended to increase with increasing non-linearity in the true outcome model, the empirical SEs of combination methods remained constant. The percentage relative error in SEs for all methods tended to decrease with increasing non-linearity in the true outcome model but was constant across treatment assignment models. Percentage relative error was smallest for propensity score methods but was extremely large for combination methods and for FULL–PROG.

5.1.2. 95% coverage. FULL–PROG, SUBCLASS(5)–PROG, FULL–MAHAL, FULL–PGPPTY, and SUBCLASS(5*5) achieved near perfect 95% CI coverage across all 25 data generating mechanisms (Figure 3). FULL–PROP and SUBCLASS(5)–PROP achieved nominal coverage for true outcome models A and B across all five true propensity score models but performed very poorly for true outcome models B, C, and E as the degree of non-linearity in the true propensity score model increased. REG exhibited the worst performance for true outcome models B–E in all cases.

5.1.3. Root mean square error. A consistent pattern was also observed in the RMSE values across the 25 data generating mechanisms (Figure 4). FULL–PROG demonstrated the smallest RMSE values in all cases, closely followed by FULL–MAHAL and then FULL–PGPPTY. RMSE values for these three methods tended to remain stable across the various data generating mechanisms; RMSE values for the other five methods tended to increase with the degree of non-linearity in the true propensity and outcome models. While SUBCLASS(5*5) and SUBCLASS(5)–PROG outperformed REG for true outcome models C–E, the latter method outperformed FULL–PROP and SUBCLASS(5)–PROP in all but a small minority of cases.

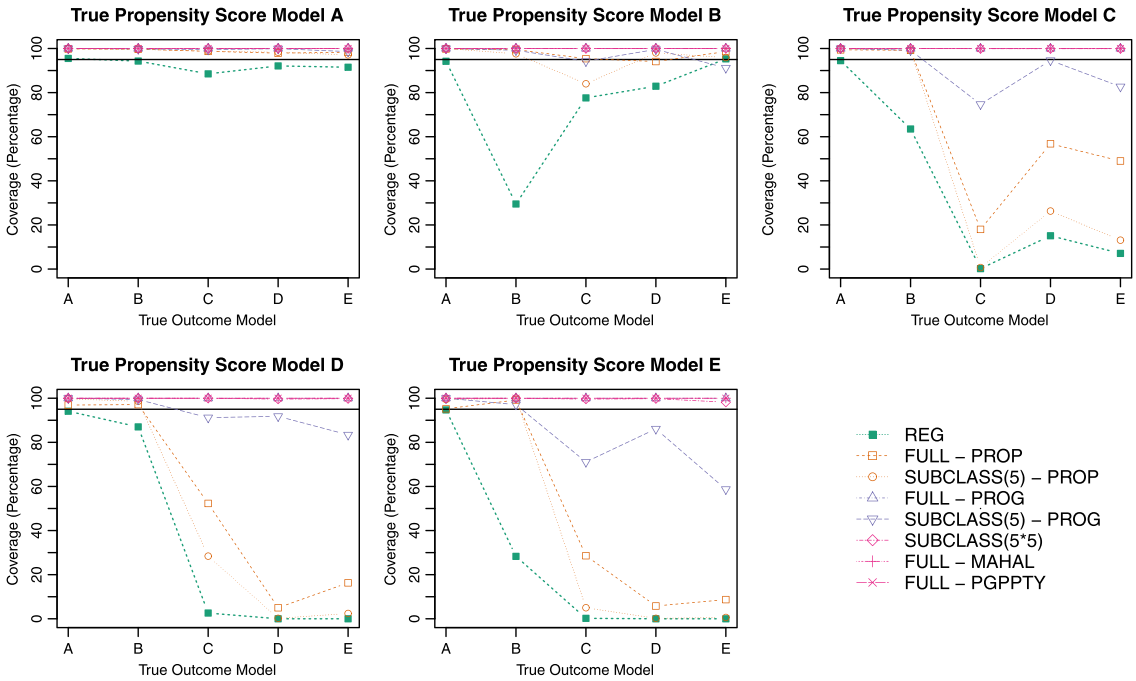

Figure 3. The 95% confidence interval coverage by propensity/prognostic score method for an incorrectly specified propensity score model, $N = 2000$.

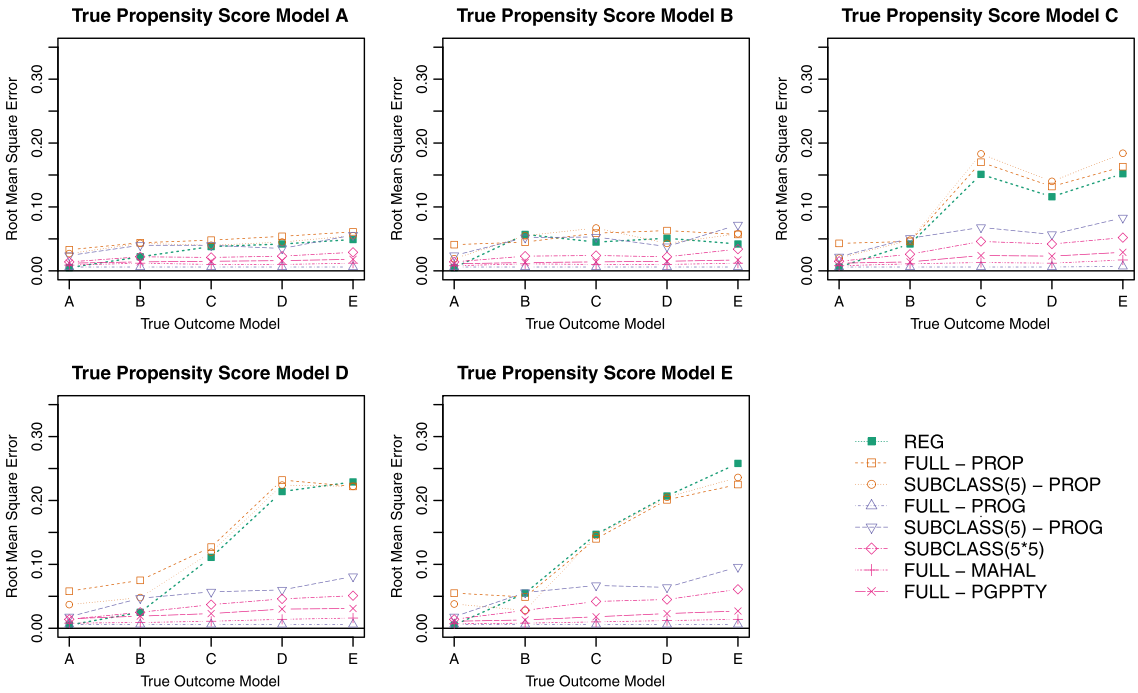

Figure 4. Root mean square error of average treatment effect on the treated estimate by propensity/prognostic score method for an incorrectly specified propensity score model, $N = 2000$.

5.2. *Propensity score model correctly specified, prognostic score model incorrectly specified*

5.2.1. Effect estimates - percent bias and standard error. FULL-PROP performed best across all 25 data generating mechanisms, exhibiting close to zero bias in all cases with FULL-PGPPTY exhibiting near identical performance (Figure 5). FULL–MAHAL also exhibited excellent performance with percent bias values tending to remain stable even as the degree of non-linearity and non-additivity in

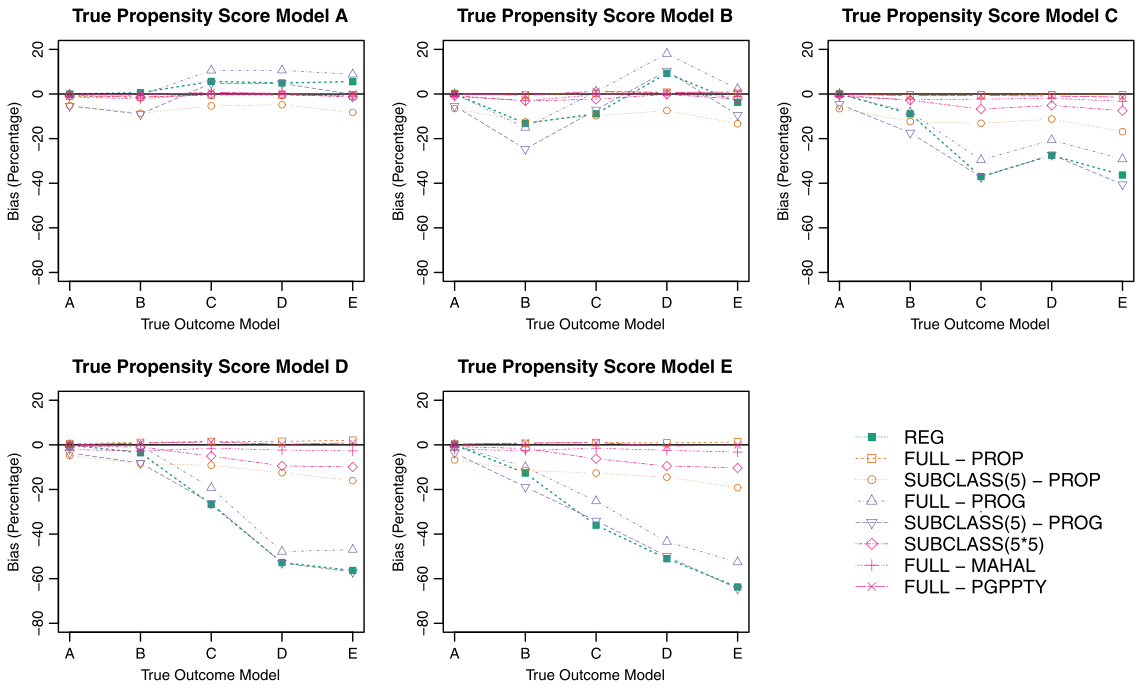

Figure 5. Percent bias in average treatment effect on the treated estimate by propensity/prognostic score method for an incorrectly specified prognostic score model, $N = 2000$.

the true outcome models increased. While SUBCLASS(5*5) demonstrated similar performance for true propensity score models A and B, its performance relative to FULL–PGPPTY and FULL–MAHAL worsened as the degree of non-additivity and non-linearity in the true treatment assignment and outcome models increased. SUBCLASS(5)–PROG did not perform well in this setting and was associated with biases at least as great as those for REG across all 25 data generating mechanisms. While SUBCLASS(5)-PROG performed similarly under prognostic score model misspecification as SUBCLASS(5)-PROP performed under propensity score model misspecification, FULL-PROG performed less poorly than FULL-PROP.

Model-based SEs, while generally larger in magnitude than when the propensity score was incorrectly specified, exhibited an identical pattern to that observed under propensity score misspecification: REG exhibited the smallest model-based SE in all cases, followed by SUBCLASS(5)–PROG, FULL–PROG, SUBCLASS(5)–PROP, and SUBCLASS(5*5). FULL–PROP was associated with the largest SE for every data generating mechanism. For all methods, model-based SE values tended to increase with the degree of non-linearity and non-additivity in the true treatment assignment and outcome models with clearer separation of methods than under propensity score misspecification.

The empirical SEs for all methods other than REG were uniformly smaller than the corresponding model-based SEs. Empirical SEs were greater under prognostic score misspecification than under propensity score misspecification. REG exhibited the smallest empirical SEs, closely followed by FULL–PROG, combination methods, and SUBCLASS(5)-PROG. PROP–FULL and SUBCLASS(5)–PROP exhibited substantially larger empirical SEs. The empirical SE of single score methods tended to increase with increasing non-linearity in the true outcome model, while the empirical SEs of combination methods remained constant. The percentage relative error in SE tended to decrease with increasing non-linearity in the true outcome model but was constant across treatment assignment models. Percentage relative error was smallest for propensity score methods but was extremely large for combination methods and for FULL–PROG.

5.2.2. 95% coverage. Coverage rates exhibited a pattern almost identical to that observed under propensity score model misspecification (Figure 6). FULL-MAHAL, FULL–PGPPTY, and SUBCLASS(5*5) achieved near perfect 95% CI coverage across all 25 data generating mechanisms. FULL–PROP tended to achieve CI coverage at close to the nominal rate. FULL–PROG achieved nominal coverage for true outcome models A and B across all five true propensity score models but performed poorly for true

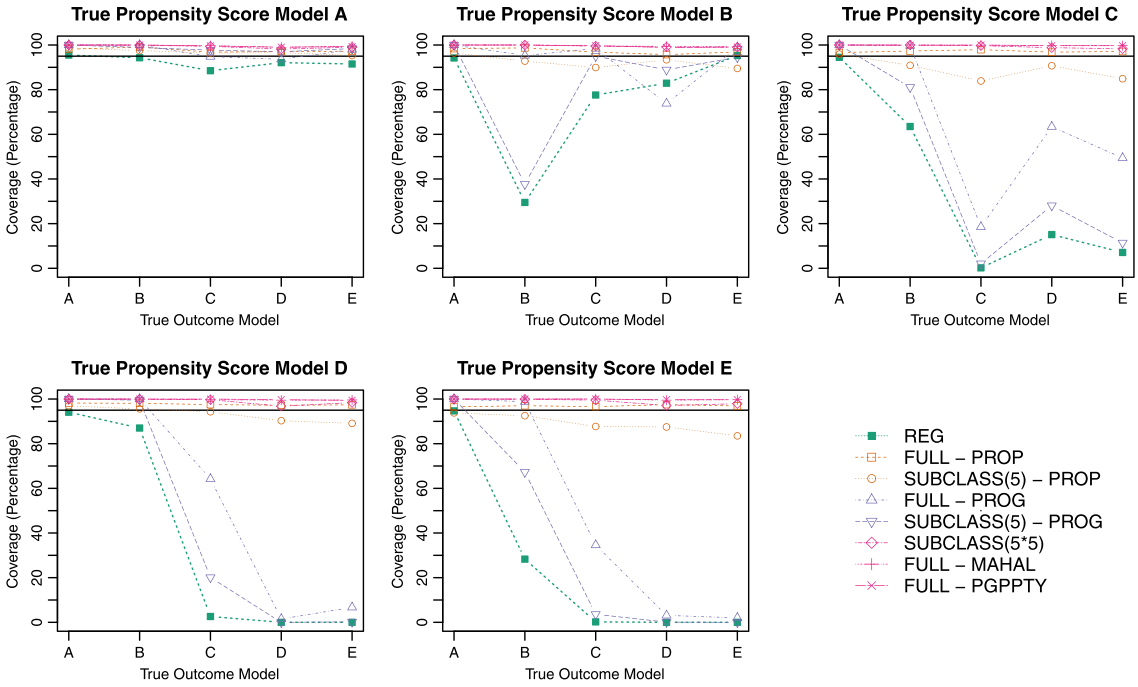

Figure 6. 95% confidence interval coverage by propensity/prognostic score method for an incorrectly specified prognostic score model, $N = 2000$.

outcome models B, C, and E. SUBCLASS(5) –PROG exhibited very poor performance across all data generating mechanisms. REG exhibited the worst performance for true outcome models B–E across the five true propensity score models.

5.2.3. Root mean square error. FULL–PGPPTY, FULL–MAHAL, and SUBCLASS(5*5) exhibited significantly smaller RMSE values than FULL–PROP and SUBCLASS(5)–PROP in all scenarios (Figure 7). These methods generally performed best across the 25 data generating mechanisms and significantly outperformed other methods in the presence of moderate non-linearity in the true propensity and/ or outcome models (C–E). For a given true propensity score model, FULL–PROP and SUBCLASS(5)–PROP were associated with the largest RMSE values for true outcome models A and B. However, these values tended to be stable across the five outcome models. SUBCLASS(5)–PROG and REG performed almost identically across the 25 data generating mechanisms. These methods performed strongly in the absence of non-linearity in the true propensity and/or prognostic score models but otherwise performed the most poorly of all the methods examined.

5.3. Both score models incorrectly specified

When both score models were incorrectly specified, methods combining the estimated propensity and/or prognostic scores exhibited similar performance to a misspecified standard main effects linear regression across all performance metrics. In the presence of non-linearity in the true treatment assignment and/or outcome models (C–E), FULL–PGPPTY, FULL–MAHAL, and SUBCLASS(5*5) tended to outperform REG with respect to percent bias but performed no better than REG otherwise. Similarly, while methods combining the estimated propensity and prognostic scores exhibited superior 95% CI coverage across all 25 data generating mechanisms, they achieved nominal coverage only in the absence of non-linearity. Model-based SEs followed an identical pattern to that reported previously. Empirical SEs were again uniformly smaller than the corresponding model-based SEs and were only slightly larger than the empirical SE for REG. The empirical SE increased with non-linearity in the true treatment assignment and outcome models. The percentage relative error in SE followed a very similar pattern to that reported previously. With respect to RMSE (Figure 8), FULL–PGPPTY, FULL–MAHAL, and SUBCLASS(5*5) tended to outperform FULL–PROP, SUBCLASS(5)–PROP, and SUBCLASS(5)–PROG under most data generating mechanisms but failed to match the performance of FULL–PROG for true propensity score models C and D.

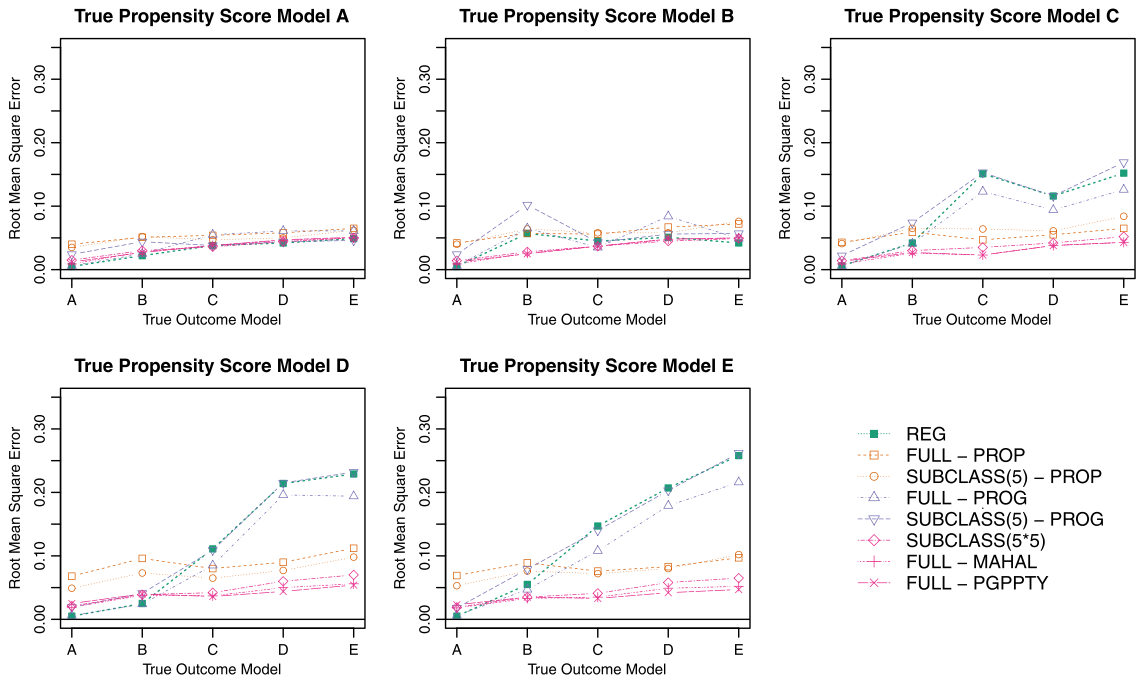

Figure 7. Root mean square error of average treatment effect on the treated estimate by propensity/prognostic score method for an incorrectly specified prognostic score model, $N = 2000$.

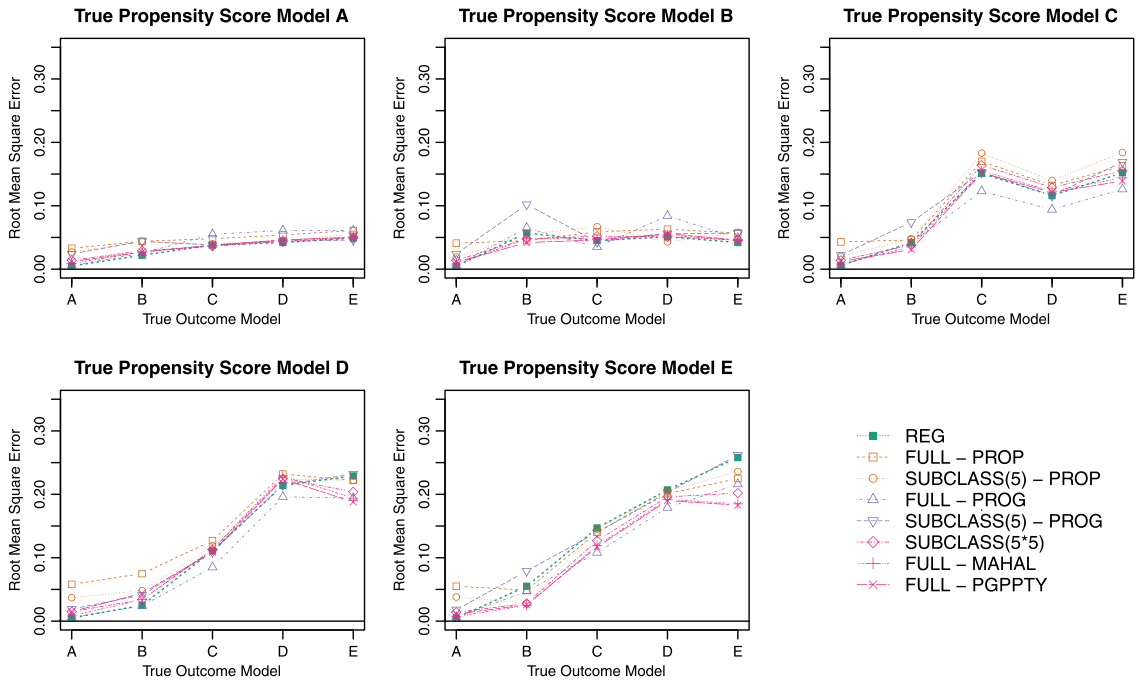

Figure 8. Root mean square error of average treatment effect on the treated estimate by propensity/prognostic score method for incorrectly specified propensity and prognostic score models, $N = 2000$.

5.4. Both score models correctly specified

REG and FULL–PROG exhibited the strongest performance across all performance metrics in this setting. These methods exhibited near zero bias across all 25 data generating mechanisms. They achieved 95% CI coverage at the nominal rate while having the smallest SEs of any method examined. Empirical SEs were again uniformly smaller than the corresponding model-based SEs. The empirical SE for FULL-PROG was almost identical to that of REG, while the empirical SEs for FULL-MAHAL,

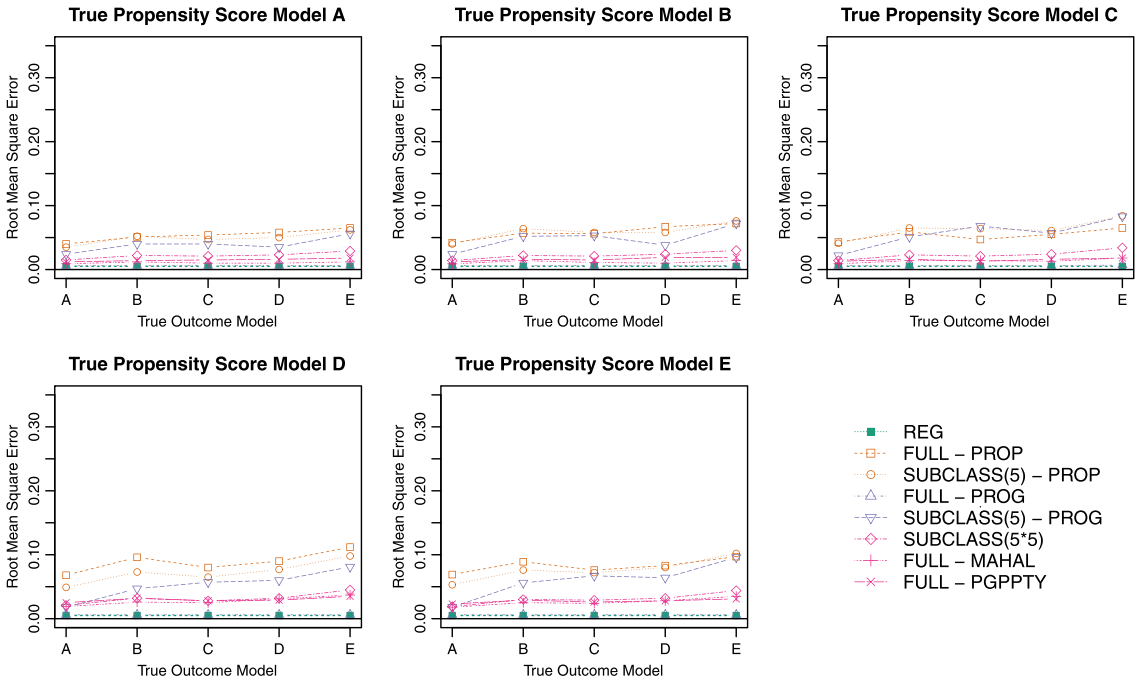

Figure 9. Root mean square error of average treatment effect on the treated estimate by propensity/prognostic score method for correctly specified propensity and prognostic score models, $N = 2000$.

FULL-PGPPTY, and SUBCLASS(5*5) were only slightly larger. Model-based estimators grossly overestimated the SE for combination methods in this setting but only marginally overestimated the SE for propensity score methods. Figure 9 illustrates that methods combining the estimated propensity and/or prognostic score performed better in RMSE terms than methods using the propensity score alone. FULL–PGPPTY, FULL–MAHAL, and SUBCLASS(5*5) exhibited near zero bias, near perfect 95% CI coverage, and almost identical RMSE values. These methods became less competitive relative to REG and FULL–PROG as the degree of non-linearity in the true propensity score model increased. The relatively poor performance of FULL–PROP may be attributed to a large SE, which increased with the degree of non-linearity in the true propensity score model. Similarly, SUBCLASS(5)–PROP and SUBCLASS(5)–PROG were the only methods associated with significant bias in this setting. These methods exhibited negative biases of approximately 10–15%.

5.5. *Simulations with* $N = 1000$ *and* $N = 5000$

Changes in sample size were not associated with any qualitative differences in relative method performance under any of the simulation settings or data generating mechanisms considered. While larger sample sizes were associated with significant increases in bias for SUBCLASS(5)–PROP and SUBCLASS(5)–PROG, increasing the sample size was associated with significant improvements in bias reduction for FULL–PGPPTY and FULL–MAHAL. While larger sample sizes were associated with reductions in the SEs for all methods, these did not unduly influence the attainment of nominal CI coverage.

5.6. *Summary*

In summary, the results of the present simulation study support the joint use of propensity and prognostic scores in estimation of the ATT. In particular, they support the use of full matching on a Mahalanobis distance combining the estimated propensity and prognostic scores (FULL–MAHAL) and full matching on the prognostic propensity score within propensity score calipers (FULL–PGPPTY). These two methods exhibited superior performance across all simulation settings and scenarios. Even when both score models were incorrectly specified, these methods outperformed full matching and subclassification on the estimated propensity score and performed no worse than an incorrectly specified main effects linear regression. Furthermore, when both score models were correctly specified, methods combining

the propensity and prognostic scores tended to estimate the effect more precisely than corresponding propensity score approaches while maintaining nominal CI coverage. These methods performed only marginally less well than a correctly specified linear regression model. When only one of the score models was correctly specified, methods combining the estimated propensity and prognostic scores were as robust to model misspecification as single score methods and proved significantly more robust than these and regression based approaches in the presence of non-linearity in the true treatment assignment and/or outcome models. FULL–MAHAL and FULL–PGPPTY significantly outperformed all other approaches under prognostic score misspecification, and all except full matching on the prognostic score (FULL–PROG) when the propensity score model was incorrectly specified. The superiority of these methods became more pronounced with increasing non-linearity in the true treatment assignment and/or outcome models. Importantly, incorrectly specifying the prognostic score model was associated with larger RMSE values for FULL–MAHAL and FULL–PGPPTY than incorrectly specifying the propensity score model.

6. Discussion

Our simulation study suggests that full matching on a Mahalanobis distance combining the estimated propensity and prognostic scores (FULL–MAHAL) represents the most promising approach by which these scores may be combined to estimate the average treatment effect on the treated. This method performed consistently well across all simulation settings considered and proved robust to the presence of non-linearity and non-additivity in the true treatment assignment and/or outcome models. While full matching on the prognostic propensity score within a propensity score caliper (FULL–PGPPTY) and FULL–MAHAL exhibited comparable performance in most cases, FULL–MAHAL exhibited superior performance when only the prognostic score was correctly specified. Additionally, as FULL–MAHAL does not impose any caliper restrictions on potential matches, it is not subject to the changes in estimand and associated complications in interpretation that such restrictions can induce. Nevertheless, this method could also be easily adapted to incorporate a caliper on the estimated propensity and/or prognostic score. Finally, the FULL–MAHAL method can easily incorporate multiple propensity and/or prognostic scores in the matching process. While this remains a rarely performed and underexplored practice, it distinguishes FULL–MAHAL as the most versatile of the methods considered here.

Our simulation study found full matching on the prognostic propensity score within propensity score calipers to be quite robust to misspecification of either one of the propensity or prognostic score models. It is possible that any bias induced by misspecification is outweighed by gains associated with providing greater weight in the propensity score model to covariates more strongly related to outcome than treatment. This is in line with the work by Brookhart *et al.* suggesting that including variables strongly related to treatment assignment only in the propensity score model may amplify bias in treatment effect estimates [30].

The strong RMSE performance of full matching on the estimated prognostic score (FULL–PROG) represents an ancillary finding of this simulation work. While full matching and subclassification on the estimated propensity score (FULL-PROP, SUBCLASS(5)–PROP) and subclassification on the estimated prognostic score (SUBCLASS(5)–PROG) exhibited inferior performance to methods combining these scores across all simulation settings and data generating mechanisms, this was not the case for FULL–PROG. FULL–PROG performed almost identically to a correctly specified linear regression model and outperformed all other methods when either the prognostic score or the propensity score was correctly specified. When both score models were incorrectly specified, FULL–PROG performed no worse than an incorrectly specified linear regression model and exhibited the best performance of any method in the presence of non-linearity in the true treatment assignment and outcome models. However, FULL–PROG performed significantly less well than methods combining the estimated propensity and prognostic scores when only the propensity score was correctly specified. These results suggest that methods combining the estimated propensity and prognostic scores should be preferred to methods utilizing the propensity score alone, but that full matching on a prognostic score may be preferred if the researcher is confident that the prognostic score model has been correctly specified.

Our subclassification and matching estimators were associated with robust model-based estimates of the SE that systematically exceeded the empirical SE. One possible explanation for this discrepancy is that our variance estimators did not account for the fact that the propensity and prognostic scores were estimated from the data. Recent work by Williamson and colleagues [38] in the propensity score context has demonstrated that routinely used variance estimators for the subclassification estimator of the ATE tend to be associated with confidence intervals that are too conservative when the propensity

score model contains variables that are strongly related to the outcome but only weakly related to the treatment. Williamson *et al.* further show that routinely used robust variance estimators for the inverse probability of treatment estimator, similar in nature to those employed in this simulation study, are associated with confidence intervals that are almost always too conservative (i.e., too wide, as we find here). While the authors derive the asymptotic marginal variance of one type of subclassification estimator, they advocate that bootstrapping techniques be used in practice. Future work could examine the performance of bootstrap estimators of the variance for the subclassification and matching estimators examined in this study.

Unlike propensity score methods, which purposefully disregard the association between covariates and the outcome, prognostic score methods have been criticized for compromising the separation of design and outcome analysis. While the use of outcome data is intrinsic to the prognostic score methodology, it is important to remember that only the outcomes of a single treatment group are used in the estimation of prognostic scores. As actual treatment effects remain unobserved, there is less potential for conscious or unconscious data dredging. This is in contrast to other approaches in which the observed exposure data, and outcomes for both treatment groups, are used directly in score estimation, as is the case with Miettinen's multivariate confounder score. Separation of design and analysis can be further safeguarded by estimating the prognostic score in a separate sample of controls where access to the outcome data of treated units is restricted. At worst, a prognostic score analysis conscientiously and faithfully implemented can uphold the spirit of the separation ideal.

The strong performance of methods combining the estimated propensity and prognostic scores when only one of the scores was correctly specified is a particularly noteworthy finding of this work. While there is no theory to suggest that any of the three methods examined here should yield consistent estimates of the ATT when only one of the score models is correctly specified, it is nevertheless interesting to note parallels with the doubly robust weighting estimators examined in Kang and Schafer [51]. In a similar manner to the methods examined in this article, doubly robust estimators leverage information regarding the association between the outcome and covariates to improve estimated precision. Doubly robust estimators produce consistent estimates of the ATE (or ATT) if at least one of the propensity score and outcome models are correctly specified. Similarly, Lunceford and Davidian [52] demonstrate that while subclassification on the estimated propensity score alone typically does not yield consistent estimates of the treatment effect, a modification including within-stratum regression adjustment may do so. The results of the present simulation study indicate that subclassification on both the propensity and prognostic scores can be associated with significant bias reduction and precision gain compared with subclassification on the propensity or prognostic score alone. These findings suggest that further investigation of the theoretical properties of methods combining propensity and prognostic score would be valuable. We have compared the performance of FULL-MAHAL and FULL-PGPPTY with a doubly robust weighted linear regression estimator of the ATT under a subset of the simulation scenarios presented here. Preliminary findings suggest that FULL-MAHAL and FULL-PGPPTY can be quite competitive with the doubly robust estimator when at least one of the score models is correctly specified and can outperform the doubly robust estimator in RMSE terms when both score models are misspecified.

The simulations conducted in this paper are limited to data generating mechanisms in which there is a constant treatment effect. Simulation studies examining method performance under effect modification are needed not just for the prognostic score based approaches examined here but also for basic propensity score methods whose performance in the presence of heterogeneous treatment effects has also received little attention. The simulation work presented here is nevertheless strengthened by the inclusion of simulation scenarios in which covariates can be confounders or predictors of exposure or outcome only; previous studies exploring prognostic score methods have assumed all covariates to be confounders. Mirroring current research in the propensity score literature, future work should seek to explore the sensitivity of prognostic score methods to omitted confounders and/or prognostically important variables and to examine the performance of methods incorporating multiple propensity and/or prognostic scores. A comprehensive study of the impact of variable selection would represent a substantial contribution to understanding how these methods might best be implemented in practice. The impact on method performance of including an estimated propensity score in the prognostic score model or of including an estimated prognostic score in the propensity score is likely to be of particular interest in this regard. Finally, noting that same-sample estimation of prognostic scores may be particularly vulnerable to overfitting even in comparison with propensity score estimation [16], application of machine learning penalized regression modeling techniques that can minimize such overfitting warrant further investigation.

The prognostic score methods considered in this article are only valid for estimation of the average treatment effect on the treated. As noted previously, estimation of the ATE using prognostic scores requires additional conditioning on all effect modifiers. Recent unpublished work by Waernbaum, following directly from her theory of covariate scores [53], demonstrates that the ATE can be validly estimated by conditioning on both a prognostic score for the potential outcome under treatment and a prognostic score for the potential outcome under control. The combination methods examined in this article can be directly applied to these two scores. Future work will investigate the performance of methods combining these two prognostic scores and the propensity score to estimate the ATE.

In conclusion, the simulation results presented here support the joint use of propensity and prognostic scores in the estimation of the ATT. In particular, the results support the use of full matching on a Mahalanobis distance combining the estimated propensity and prognostic scores and of full matching on the prognostic propensity score within propensity score calipers. These methods exhibited superior performance to corresponding propensity score methods when one or both scores were incorrectly specified. While method performance was generally comparable with that of propensity score methods when the propensity score model was correctly specified, approaches that combine the propensity and prognostic scores tended to estimate the effect more precisely while maintaining nominal CI coverage. These methods were, in general, as robust to model misspecification as those using the propensity or prognostic score alone and performed significantly better than single score methods in the presence of non-linearity in the true treatment assignment and/or outcome models when the single score model was incorrectly specified. The results suggest that methods combining the estimated propensity and prognostic scores should be preferred to methods utilizing the estimated propensity score alone but that full matching on a prognostic score may be preferred if the researcher is confident that the prognostic score model has been correctly specified.

Appendix A: data generation model specification

Correlation structure

$$corr\,(W_1, W_5) = corr\,(W_3, W_8) = 0.2$$

$$corr\,(W_2, W_6) = corr\,(W_3, W_9) = 0.9$$

All other pairwise correlations were set to zero.

True propensity score models

(a) Additivity and linearity

$$logit\{P[Z=1|\boldsymbol{W}]\} = \alpha_0 + \alpha_1 W_1 + \alpha_2 W_2 + \alpha_3 W_3 + \alpha_4 W_4 + \alpha_5 W_5 + \alpha_6 W_6 + \alpha_7 W_7$$

(b) Moderate non-additivity

$$\begin{aligned} logit\{P[Z=1|\boldsymbol{W}]\} = {} & \alpha_0 + \alpha_1 W_1 + \alpha_2 W_2 + \alpha_3 W_3 + \alpha_4 W_4 + \alpha_5 W_5 + \alpha_6 W_6 + \alpha_7 W_7 \\ & + 0.5 \times \alpha_1 W_1 W_3 + 0.7 \times \alpha_2 W_2 W_4 + 0.5 \times \alpha_3 W_3 W_5 \\ & + 0.7 \times \alpha_4 W_4 W_6 + 0.5 \times \alpha_5 W_5 W_7 + 0.5 \times \alpha_1 W_1 W_6 \\ & + 0.7 \times \alpha_2 W_2 W_3 + 0.5 \times \alpha_3 W_3 W_4 + 0.5 \times \alpha_4 W_4 W_5 \\ & + 0.5 \times \alpha_5 W_5 W_6 \end{aligned}$$

(c) Mild non-additivity and non-linearity

$$\begin{aligned} logit\{P[Z=1|\boldsymbol{W}]\} = {} & \alpha_0 + \alpha_1 W_1 + \alpha_2 W_2 + \alpha_3 W_3 + \alpha_4 W_4 + \alpha_5 W_5 + \alpha_6 W_6 + \alpha_7 W_7 \\ & + \alpha_2 W_2 W_2 + 0.5 \times \alpha_1 W_1 W_3 + 0.7 \times \alpha_2 W_2 W_4 + 0.5 \times \alpha_4 W_4 W_5 \\ & + 0.5 \times \alpha_5 W_5 W_6 \end{aligned}$$

(d) Moderate non-linearity

$$\begin{aligned} logit\{P[Z=1|\boldsymbol{W}]\} = {} & \alpha_0 + \alpha_1 W_1 + \alpha_2 W_2 + \alpha_3 W_3 + \alpha_4 W_4 + \alpha_5 W_5 + \alpha_6 W_6 + \alpha_7 W_7 \\ & + \alpha_2 W_2 W_2 + \alpha_4 W_4 W_4 + \alpha_7 W_7 W_7 \end{aligned}$$

(e) Moderate non-additivity and non-linearity

$$\begin{aligned} logit\{P[Z=1|\boldsymbol{W}]\} = {} & \alpha_0 + \alpha_1 W_1 + \alpha_2 W_2 + \alpha_3 W_3 + \alpha_4 W_4 + \alpha_5 W_5 + \alpha_6 W_6 + \alpha_7 W_7 \\ & + \alpha_2 W_2 W_2 + \alpha_4 W_4 W_4 + \alpha_7 W_7 W_7 + 0.5 \times \alpha_1 W_1 W_3 \\ & + 0.7 \times \alpha_2 W_2 W_4 + 0.5 \times \alpha_3 W_3 W_5 + 0.7 \times \alpha_4 W_4 W_6 \\ & + 0.5 \times \alpha_5 W_5 W_7 + 0.5 \times \alpha_1 W_1 W_6 + 0.7 \times \alpha_2 W_2 W_3 \\ & + 0.5 \times \alpha_3 W_3 W_4 + 0.5 \times \alpha_4 W_4 W_5 + 0.5 \times \alpha_5 W_5 W_6 \end{aligned}$$

True outcome models

All outcome models contain an error term, $\varepsilon \sim N(0,0.1^2)$

(a) Additivity and linearity

$$Y = \beta_0 + \beta_1 W_1 + \beta_2 W_2 + \beta_3 W_3 + \beta_4 W_4 + \beta_5 W_8 + \beta_6 W_9 + \beta_7 W_{10} + \gamma Z + \varepsilon$$

(b) Moderate non-additivity

$$\begin{aligned} Y = {} & \beta_0 + \beta_1 W_1 + \beta_2 W_2 + \beta_3 W_3 + \beta_4 W_4 + \beta_5 W_8 + \beta_6 W_9 + \beta_7 W_{10} + 0.5 \times \beta_1 W_1 W_3 \\ & + 0.7 \times \beta_2 W_2 W_4 + 0.5 \times \beta_3 W_3 W_8 + 0.7 \times \beta_4 W_4 W_9 + 0.5 \times \beta_5 W_8 W_{10} \\ & + 0.5 \times \beta_1 W_1 W_9 + 0.7 \times \beta_2 W_2 W_3 + 0.5 \times \beta_3 W_3 W_4 + 0.5 \times \beta_4 W_4 W_8 \\ & + 0.5 \times \beta_5 W_8 W_9 + \gamma Z + \varepsilon \end{aligned}$$

(c) Mild non-additivity and non-linearity

$$\begin{aligned} Y = {} & \beta_0 + \beta_1 W_1 + \beta_2 W_2 + \beta_3 W_3 + \beta_4 W_4 + \beta_5 W_8 + \beta_6 W_9 + \beta_7 W_{10} + \beta_2 W_2 W_2 \\ & + 0.5 \times \beta_1 W_1 W_3 + 0.7 \times \beta_2 W_2 W_4 + 0.5 \times \beta_4 W_4 W_8 + 0.5 \times \beta_5 W_8 W_9 + \gamma Z + \varepsilon \end{aligned}$$

(d) Moderate non-linearity

$$\begin{aligned} Y = {} & \beta_0 + \beta_1 W_1 + \beta_2 W_2 + \beta_3 W_3 + \beta_4 W_4 + \beta_5 W_8 + \beta_6 W_9 + \beta_7 W_{10} + \beta_2 W_2 W_2 \\ & + \beta_4 W_4 W_4 + \beta_7 W_{10} W_{10} + \gamma Z + \varepsilon \end{aligned}$$

(e) Moderate non-additivity and non-linearity

$$\begin{aligned} Y = {} & \beta_0 + \beta_1 W_1 + \beta_2 W_2 + \beta_3 W_3 + \beta_4 W_4 + \beta_5 W_8 + \beta_6 W_9 + \beta_7 W_{10} + \beta_2 W_2 W_2 \\ & + \beta_4 W_4 W_4 + \beta_7 W_{10} W_{10} + 0.5 \times \beta_1 W_1 W_3 + 0.7 \times \beta_2 W_2 W_4 + 0.5 \times \beta_3 W_3 W_8 \\ & + 0.7 \times \beta_4 W_4 W_9 + 0.5 \times \beta_5 W_8 W_{10} + 0.5 \times \beta_1 W_1 W_9 + 0.7 \times \beta_2 W_2 W_3 \\ & + 0.5 \times \beta_3 W_3 W_4 + 0.5 \times \beta_4 W_4 W_8 + 0.5 \times \beta_5 W_8 W_9 + \gamma Z + \varepsilon \end{aligned}$$

Model coefficient values:

$\alpha_0 = -1.897$	$\beta_0 = -1.386$
$\alpha_1 = 0.8$	$\beta_1 = 0.3$
$\alpha_2 = -0.25$	$\beta_2 = -0.36$
$\alpha_3 = 0.6$	$\beta_3 = -0.73$
$\alpha_4 = -0.4$	$\beta_4 = -0.2$
$\alpha_5 = -0.8$	$\beta_5 = 0.71$
$\alpha_6 = -0.5$	$\beta_6 = -0.19$
$\alpha_7 = 0.7$	$\beta_7 = 0.26$
	$\gamma = -0.4$

Acknowledgements

This work was supported by a Fulbright Foreign Student Award (Leacy) and by Award K25MH083846 from the National Institutes of Mental Health (Stuart). It was completed while the first author was a graduate student in the Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health. The content is solely the responsibility of the authors and does not necessarily represent the official views of the Fulbright Program, the National Institute of Mental Health, or the National Institutes of Health. The authors thank two anonymous reviewers for their helpful and insightful comments.

References

1. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika* 1983; **70**:41–55.
2. Stuart EA. Matching methods for causal inference: a review and a look forward. *Statistical Science* 2010; **25**(1):1–21.
3. Imbens GW. Nonparametric estimation of average treatment effects under exogeneity: a review. *Review of Economics and Statistics* 2004; **86**:4–29.
4. Barnard J, Frangakis C, Hill J, Rubin DB. Principal stratification approach to broken randomized experiments: a case study of school choice vouchers in New York City (with discussion). *Journal of the American Statistical Association* 2003; **98**:299–323.
5. Behrman JR, Cheng Y, Todd PE. Evaluating preschool programs when length of exposure to the program varies: a nonparametric approach. *Review of Economics and Statistics* 2004; **86**(1):108–132.
6. Rosenbaum PR. Dropping out of high school in the United States: an observational study. *Journal of Educational Statistics* 1986; **11**:207–224.
7. DiPrete T, Gangl M. Assessing bias in the estimation of causal effects: Rosenbaum bounds on matching estimators and instrumental variables estimation with imperfect instruments. *Sociological Methodology* 2004; **34**:271–310.
8. Morgan SL, Harding DJ. Matching estimators of causal effects: prospects and pitfalls in theory and practice. *Sociological Methods and Research* 2006; **35**:3–60.
9. Harder VS, Morral AR, Arkes J. Marijuana use and depression among adults: testing for causal associations. *Addiction* 2006; **101**:1463–1472.
10. Hill J, Waldfogel J, Brooks-Gunn J, Han W. Maternal employment and child development: a fresh look using newer methods. *Developmental Psychology* 2005; **41**:833–850.
11. Dehejia RH, Wahba S. Causal effects in non-experimental studies: re-evaluating the evaluation of training programs. *Journal of the American Statistical Association* 1999; **94**:1053–1062.
12. Christakis NA, Iwashyna TI. The health impact of health care on families: a matched cohort study of hospice use by decedents and mortality outcomes in surviving, widowed spouses. *Social Science & Medicine* 2003; **57**:465–475.
13. Perkins SM, Tu W, Underhill MG, Zhou XH, Murray MD. The use of propensity scores in pharmacoepidemiological research. *Pharmacoepidemiology and Drug Safety* 2000; **9**:93–101.
14. Weitzen S, Lapane KL, Toledano AY, Hume AL, Mor V. Principles for modeling propensity scores in medical research: a systematic literature review. *Pharmacoepidemiology and Drug Safety* 2004; **13**:841–853.
15. Schneeweiss S, Rassen JA, Glynn RJ, Avorn J, Morgan H, Brookhart MA. High-dimensional propensity score adjustment in studies of treatment effects using health care claims data. *Epidemiology* 2009; **20**(4):512–522.
16. Hansen BB. The prognostic analogue of the propensity score. *Biometrika* 2008; **95**(2):481–488.
17. Arbogast PG, Ray WA. Performance of disease risk scores, propensity scores and traditional multivariable outcome regression in the presence of multiple confounders. *American Journal of Epidemiology* 2011; **174**(5):613–620.
18. Miettinen OS. Stratification by a multivariate confounder score. *American Journal of Epidemiology* 1976; **104**(6): 609–620.
19. Arbogast PG, Ray WA. Use of disease risk scores in pharmacoepidemiologic studies. *Statistical Methods in Medical Research* 2009; **18**(1):67–80.
20. Ray WA, Meredith S, Thapa PB, Meador KG, Hall K, Murray KT. Antipsychotics and the risk of sudden cardiac death. *Archives of General Psychiatry* 2001; **58**:1161–1167.
21. Ray WA, Stein CM, Daugherty JR, Hall K, Arbogast PG, Griffin MR. COX-2 selective non-steroidal anti-inflammatory drugs and risk of serious coronary heart disease. *Lancet* 2002; **360**:1071–1073.
22. Ray WA, Stein CM, Hall K, Daugherty JR, Griffin MR. Non-steroidal anti-inflammatory drugs and risk of serious coronary heart disease: an observational cohort study. *Lancet* 2002; **359**:118–123.
23. Ray WA, Meredith S, Thapa PB, Hall K, Murray KT. Cyclic antidepressants and the risk of sudden cardiac death. *Clinical Pharmacology & Therapeutics* 2004; **75**(3):234–241.
24. Ray WA, Murray KT, Meredith S, Narasimhulu SS, Hall K, Stein CM. Oral erythromycin and the risk of sudden death from cardiac causes. *New England Journal of Medicine* 2004; **351**:1089–1096.
25. Graham DJ, Campen D, Hui R, Spence M, Cheetham C, Levy G, Shoor S, Ray WA. Risk of acute myocardial infarction and sudden cardiac death in patients treated with cyclo-oxygenase 2 selective and non-selective non-steroidal anti-inflammatory drugs: nested case-control study. *Lancet* 2005; **365**:475–481.
26. Ray WA, Chung CP, Stein CM, Smalley WE, Hall K, Arbogast PG, Griffin MR. Risk of peptic ulcer hospitalizations in users of NSAIDS with gastroprotective cotherapy versus coxibs. *Gastroenterology* 2007; **133**(3):790–798.
27. Stürmer T, Schneeweiss S, Brookhart MA, Rothman KJ, Avorn J, Glynn RJ. Analytic Strategies to adjust confounding using exposure propensity scores and disease risk scores: nonsteroidal anti-inflammatory drugs and short-term mortality in the elderly. *American Journal of Epidemiology* 2005; **161**(9):891–898.
28. Cadarette SM, Gagne JJ, Solomon DH, Katz JN, Stürmer T. Confounder summary scores when comparing the effects of multiple drug exposures. *Pharmacoepidemiology and Drug Safety* 2010; **19**:2–9.
29. Glynn RJ, Gagne JJ, Schneeweiss S. Role of disease risk scores in comparative effectiveness research with emerging therapies. *Pharmacoepidemiology and Drug Safety* 2012; **21**(2):138–147.
30. Brookhart MA, Schneeweiss S, Rothman KJ, Glynn RJ, Avorn J, Sturmer T. Variable selection for propensity score models. *American Journal of Epidemiology* 2006; **163**(12):1149–1156.
31. Myers JA, Rassen JA, Gagne JJ, Huybrechts KF, Schneeweiss S, Rothman KJ, Joffe MM, Glynn RJ. Effects of adjusting for instrumental variables on bias and precision of effect estimates. *American Journal of Epidemiology* 2011; **174**(11):1223–1227.
32. Pearl J. Invited commentary: understanding bias amplification. *American Journal of Epidemiology* 2011; **174**(11):1213–1222.

33. Rubin DB, Thomas N. Combining propensity score matching with additional adjustments for prognostic covariates. *Journal of the American Statistical Association* 2000; **95**:573–585.
34. Hansen BB. Bias reduction in observational studies via prognosis scores. *Technical Report No. 441*, Statistics Department, University of Michigan, Ann Arbor, Michigan, June 2006. Available from: http://dept.stat.lsa.umich.edu/~bbh/rspaper2006-06.pdf [accessed 29 October 2010].
35. Rosenbaum PR, Rubin DB. Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association* 1984; **79**:516–524.
36. Rosenbaum PR. A characterization of optimal designs for observational studies. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)* 1991; **53**:597–610.
37. Hansen BB. Full matching in an observational study of coaching for the SAT. *Journal of the American Statistical Association* 2004; **99**:609–618.
38. Willamson EJ, Morley R, Lucas A, Carpenter JR. Variance estimation for stratified propensity score estimators. *Statistics in Medicine* 2012; **31**(15):1617–1632.
39. Pike MC, Anderson J, Day N. Some insights into Miettinen's multivariate confounder score approach to case-control study analysis. *Journal of Epidemiology and Community Health* 1979; **33**(1):104–106.
40. Cook EF, Goldman L. Performance of tests of significance based on stratification by a multivariate confounder score or by a propensity score. *Journal of Clinical Epidemiology* 1989; **42**:317–324.
41. Hansen BB. Propensity score matching to recover latent experiments from nonexperimental data: a case study. In *Looking Back: Proceedings of a Conference in Honor of Paul W. Holland*, Lecture Notes in Statistics 202. Springer: New York, 2011; 149-181.
42. Imai K, van Dyk DA. Causal inference with general treatment regimes: generalizing the propensity score. *Journal of the American Statistical Association* 2004; **99**:854–866.
43. Setoguchi S, Schneeweiss S, Brookhart MA, Glynn RJ, Cook EF. Evaluating uses of data mining techniques in propensity score estimation: a simulation study. *Pharmacoepidemiology and Drug Safety* 2008; **17**(6):546–555.
44. Lee B, Lessler J, Stuart EA. Improving propensity score weighting using machine learning. *Statistics in Medicine* 2009; **29**(3):337-346. PMCID PMC2943670.
45. R Development Core Team. R: A language and environment for statistical computing, R Foundation for Statistical Computing, Vienna, Austria, 2011. ISBN 3-900051-07-0, Available from: http://www.R-project.org/.
46. Ho DE, Imai K, King G, Stuart EA. Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Political Analysis* 2007; **15**(3):199–236.
47. Ho DE, Imai K, King G, Stuart EA. MatchIt: nonparametric preprocessing for parametric causal inference. *Journal of Statistical Software* 2011; **42**(8):1–28.
48. Hansen BB, Knopfler SO. Optimal full matching and related designs via network flows. *Journal of Computational and Graphical Statistics* 2006; **15**:609–627.
49. Lumley T. *Survey: analysis of complex survey samples*. R package version 3.26, 2011.
50. Rubin DB, Thomas N. Matching using estimated propensity scores: relating theory to practice. *Biometrics* 1996; **52**:249–264.
51. Kang JDY, Schafer JL. Demystifying double robustness: a comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical Science* 2007; **22**(4):523–539.
52. Lunceford JK, Davidian M. Stratification and weighting via the propensity score in estimation of causal treatment effects: A comparative study. *Statistics in Medicine* 2004; **23**:2937–2960.
53. Waernbaum I. Model misspecification and robustness in causal inference: comparing matching with double robustness. *Statistics in Medicine* 2012; **31**(15):1572–1581.

Supporting information

Additional supporting information may be found in the online version of this article at the publisher's web site.