Curr Epidemiol Rep (2017) 4:271–280
https://doi.org/10.1007/s40471-017-0131-y

Propensity Scores in Pharmacoepidemiology: Beyond the Horizon

John W. Jackson[1,2] · Ian Schmid[2] · Elizabeth A. Stuart[2,3,4]

Published online: 6 November 2017

Abstract

Purpose of Review Propensity score methods have become commonplace in pharmacoepidemiology over the past decade. Their adoption has confronted formidable obstacles that arise from pharmacoepidemiology's reliance on large healthcare databases of considerable heterogeneity and complexity. These include identifying clinically meaningful samples, defining treatment comparisons, and measuring covariates in ways that respect sound epidemiologic study design. Additional complexities involve correctly modeling treatment decisions in the face of variation in healthcare practice and dealing with missing information and unmeasured confounding. In this review, we examine the application of propensity score methods in pharmacoepidemiology with particular attention to these and other issues, with an eye towards standards of practice, recent methodological advances, and opportunities for future progress.

Recent Findings Propensity score methods have matured in ways that can advance comparative effectiveness and safety research in pharmacoepidemiology. These include natural extensions for categorical treatments, matching algorithms that can optimize sample size given design constraints, weighting estimators that asymptotically target matched and overlap samples, and the incorporation of machine learning to aid in covariate selection and model building.

Summary These recent and encouraging advances should be further evaluated through simulation and empirical studies, but nonetheless represent a bright path ahead for the observational study of treatment benefits and harms.

Keywords Pharmacoepidemiology · Comparative effectiveness · Study design · Non-experimental study · Propensity score · Causal inference

Introduction

Many pharmacoepidemiology studies compare safety and effectiveness across treatment options that have not been randomly assigned. Treatment groups may differ in terms of prognostic factors, and crude comparisons will often lack causal interpretation. Such bias arises because health professionals rightly use clinical parameters to recommend treatment when they anticipate potential benefit and withhold it when concerned about adverse events. This phenomenon and its variants are referred to as confounding by indication, channeling, or protopathic bias [1]. Propensity score methods were designed to confront such confounding. They do so by modeling how prognostic factors (henceforth covariates, X) guide treatment decisions and using this knowledge to construct treatment groups with similar covariate distributions. Given their reliance on a treatment model instead of (or in addition to) an outcome model, propensity scores may be particularly relevant for pharmacoepidemiology, where it is often difficult to have adequate outcome models given rare events and potentially large numbers of confounders.

This article is part of the Topical Collection on *Pharmacoepidemiology*

✉ Elizabeth A. Stuart
estuart@jhu.edu

[1] Department of Epidemiology, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD 21205, USA

[2] Department of Mental Health, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD 21205, USA

[3] Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD 21205, USA

[4] Department of Health Policy and Management, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD 21205, USA

Moreover, in medicine, treatment decisions are well understood and investigators can leverage clinical expertise and guidelines to build plausible models. As they draw on the familiar concepts of balance from clinical trials, propensity score analyses are very accessible to clinical, industry, and regulatory stakeholders.

In this review, we discuss critical aspects in the use of propensity scores in pharmacoepidemiologic research. We address study design, covariate choice, model selection, using the propensity score, and strategies for dealing with unmeasured bias. For each, we highlight current understanding, recent developments, and opportunities for progress.

Data and Design in Pharmacoepidemiology

Treatment Comparisons

Pharmacoepidemiology often employs routinely collected healthcare data that facilitates payment and reimbursement for medical services and products (e.g., insurance claims) and also medical care (e.g., electronic medical records) [2]. To estimate causal effects from these data, investigators must carve out a sample of patient experience that resembles a well-designed study that is conceivable in the real world, ethics and practicalities aside [3••]. Many design flaws at this stage can be avoided by anchoring study entry, treatment group assignment, and start of follow-up around a change in drug use.

The "new user" or "incident user" design anchors a study at the initial prescription dispensing (henceforth "use") of a drug after some period of non-use, perhaps comparing users with those initiating an active comparison or with non-users, with treatment group membership held fixed. More elaborate permutations anchor at treatment decisions, e.g., a dose intensification [4, 5, 6•]. These designs, which can be incorporated in distributed data network studies, are motivated to mitigate pernicious confounding, selection, and immortal-time biases and to answer clinically relevant questions [7–9]. Unless noted otherwise, our review will consider a comparison of incident use vs. non-use but extends to other treatment decisions.

Targets of Inference

Another key aspect of design is the choice of a target population, which hinges on the underlying clinical question. When interested in treatment effects among the study sample, the focus of this review, one may seek to estimate effects among the entire sample (average treatment effect (SATE)) or among a treatment group (average effect of treatment on the treated (SATT)) [10]. These estimands will differ when there are heterogeneous effects and the distributions of effect modifiers vary across treatment groups [11]. When the distributions are grossly dissimilar (an extreme being ubiquitously prescribed or withheld treatment for certain types of patients), lack of overlap precludes estimating the SATE and, in severe cases, the SATT [12, 13]. In a later section, we discuss possible strategies to estimate the SATT or effects among an "overlap" population that has shared common support [12, 13].

Causal Inference, Potential Outcomes, and the Propensity Score

Potential Outcomes

Propensity score methods draw on the potential outcomes framework that was developed for randomized trials [14, 15]. Therein, we consider potential outcomes $Y_i(A = a)$ under treatment $Y_i(1)$ and no treatment $Y_i(0)$, observing only one for a given individual i. Under some circumstances, we can expect individuals to share the same distribution of potential outcomes regardless of their actual treatment status, i.e., $Y(a) \coprod A$ for all a, allowing a contrast of outcomes among those actually treated vs. not $E[Y|A = 1] - E[Y|A = 0]$ to stand in for a causal contrast of outcomes had everyone been treated vs. not $E[Y(a = 1)] - E[Y(a = 0)]$. This is an unbiased estimate in a randomized trial, where we know that treatment assignment is independent of potential outcomes. When treatment is merely observed, that assumption is not guaranteed. We attempt to measure enough covariates X such that the potential outcomes are rendered conditionally independent of treatment, i.e., $Y(a) \coprod A \mid X$, and then estimate the conditional average causal effect $E[Y|A = 1, X] - E[Y|A = 0, X]$. Independence between potential outcomes and treatment is often referred to as "ignorability" or "exchangeability" [15, 16]. Moving from potential to observed outcomes often relies on the Stable-Unit-Treatment-Value-Assumption (SUTVA) which encodes that (i) the treatment of one individual does not affect the outcome of another ("no interference") and (ii) the outcome observed under actual treatment and a hypothetical intervention assigning treatment are equivalent ("treatment-version irrelevance" and consistency) [17, 18]. Some methods have been extended to relax SUTVA [19, 20].

The Propensity Score for Binary and Categorical Treatments

For a binary treatment A of use ($A = 1$) vs. non-use ($A = 0$), the propensity score $e(X)$ is the probability of use given the covariates X, i.e., $P(A = 1|X)$. Rosenbaum and Rubin proved that (i) among patients with the same propensity score $e(X)$, the covariates X will be balanced; (ii) if one can estimate causal treatment effects by adjustment for X (i.e., ignorability holds given X), then one can estimate causal treatment effects by adjusting for the propensity score (i.e., ignorability holds given $e(x)$) [15].

To compare treatments 1 ($A = 1$) and 2 ($A = 2$) with a common referent of non-use ($A = 3$), one needs to balance covariates across all three groups. This can be achieved by defining a propensity score function—the generalized propensity score—that describes how the distribution of treatment A depends on covariates X. This is a set of probabilities $P(A = a| X)$ that sum to one, in this case $P(A = 1| X)$, $P(A = 2| X)$, and $P(A = 3| X)$ [21, 22]. Among individuals with the same propensity score function, the covariate distributions of all three treatment groups are balanced. The generalized propensity score can be extended to continuous and ordinal treatments such as dose [22, 23].

Building the Propensity Score Model

In modeling the propensity score, the goal is not to perfectly predict treatment. (In fact, perfect prediction implies intractable confounding). Rather, it is to reduce confounding by (i) selecting enough covariates to render potential outcomes independent of treatment; (ii) producing estimates of the probability of treatment that accurately reflect how treatment is related to the covariates; and (iii) using those probabilities to create treatment and comparison groups with similar covariate distributions. In this sense, measures of discriminatory power of a predictive model such as the *c*-statistic should not guide covariate choice nor model specification [24, 25].

Covariate Definition and Selection

Choosing covariates to measure is challenging when the underlying causal structure is unknown. For a thorough overview of relevant issues, see the review by Sauer et al. [26]. A safe strategy is to include risk factors for the outcome (these will improve statistical precision) [27]. Covariates that predict treatment but not the outcome (true instruments) are best avoided as these can not only reduce precision [27] but also inflate any bias from any remaining confounding, a phenomenon described as Z-bias [28]. For all other covariates, simulation and theoretical results favor adjusting for the covariate to reduce potential confounding [28–30, 31•].

In databases where it is possible to measure hundreds of covariates, algorithms have been proposed to help investigators choose among them. The high-dimensional propensity score (hdPS) approach selects all covariates that are associated with treatment and outcome, in the hope that adjusting for a rich set of proxy covariates will protect against unmeasured confounding [32]. Recent work on hdPS warns against prescreening covariates by their prevalence and suggest that its performance in studies with few exposed outcomes can be enhanced by incorporating machine learning tools and Bayesian estimation of the covariate-outcome association [33–35]. A variety of other algorithms have been developed to improve the treatment effect estimate in high-dimensional settings while limiting the selected covariates to a set that suffices to control for confounding. In one, a backwards selection algorithm sequentially discards variables that are independent of the outcome given treatment and the remaining covariates [36]. In another, the "least absolute shrinkage and selection operator" (lasso) is applied to an outcome regression model to select covariates for inclusion in the propensity score [34]. Several methods seek to also optimize the mean square error of the treatment effect, including procedures that iteratively select variables for candidate outcome and propensity score models (Collaborative Targeted Maximum Likelihood (C-TMLE); and Bayesian Adjustment for Confounding (BAC)) and modified stepwise "change-in-estimate" selection strategies [37–41]. All of these strategies presume a single outcome. When several are of interest, simulation results suggest that a generic propensity score model based on their shared confounders performs nearly as well as separate models built for each outcome [42]. These algorithms appear promising for variable selection but have not been studied in depth.

In defining and measuring covariates, one must ensure that they are truly pre-treatment covariates. This follows naturally in a design that aligns cohort entry, the start of follow-up, and treatment definitions at a change in treatment, e.g., the index date. If covariates are not stable attributes over the study period, such as a measure of symptoms, care must be taken to ensure that such time-varying covariates are not effects of treatment. One should assess these before the index date and determine whether the immediate pre-treatment value or a richer summary of its history are most predictive of the outcome and the indexing treatment decision [43]. Assessments should also consider the sensitivity of defining covariates by drawing on all available data or restricting to a fixed window of some length [44]. Though some exploratory studies have examined these strategies empirically and through simulation, there is not unequivocal evidence regarding their merits and shortfalls [45]. An excellent review by Brookhart et al. describe additional concerns that arise when using medical billing and service codes to assess health status [46].

Modeling the Propensity Score

In observational studies, the true propensity score is unknown and must be estimated. For binary treatments, this is typically accomplished through a logistic regression model with at least main effects. The impact of iteratively adding interaction and higher terms can be evaluated by how well the resulting propensity score approach balances covariates (discussed later). Some alternatives seek to automate the process of covariate and model selection by leveraging machine learning tools, regularization, or loss-based estimation. For example, implementations of ensemble methods of bagged/boosted classification and regression trees and random forests appear to balance covariates better than logistic regression in simulation studies though the reverse has been seen with empirical data [25, 47]. Other approaches seek to optimize not prediction error but covariate selection and the propensity score's performance in constructing comparable treatment groups. These include C-TMLE mentioned above [39, 40]; the outcome-adaptive lasso, which uses shrinkage to deselect variables that predict exposure but not the outcome [48]; generalized boosted models which combine piecewise regression trees to capture interactions [49, 50]; and the covariate balancing propensity score which uses a generalized method of moments approach [51, 52]. Early simulations suggest that these methods perform well though their limits and tradeoffs have yet to be fully characterized [53].

Some complexities are worth mentioning. Measurement error in covariates often occurs in administrative healthcare data and can lead to residual confounding [2, 46]. Recent theoretical work on leveraging prior knowledge and external validation samples for correcting the propensity score could be explored as a solution [54–56]. Another complexity is that treatments may be administered differently across time, physicians, health institutions, and systems, reflecting gradients in practice or quality of care. Some argue against estimating propensity scores within clusters when they have no effect on outcomes (i.e., the clustering variables are instruments) [57]. Others point out that failing to reflect heterogeneous relationships between covariates and treatment can lead to a mis-specified propensity score model [58] and propose a model-fitting strategy to confront this [59]. These concerns can be evaluated empirically through checking whether the estimates from propensity scores that ignore clustering produce comparable treatment groups. Future empirical work on this issue should consider therapeutic examples where there is considerable clinical uncertainty and discretion (e.g., psychiatry), complex trade-offs between benefits and risks (e.g., anticoagulant therapy), and where treatment rules are less established (e.g., newly marketed medications).

Evaluating the Propensity Score Approach

The performance of the propensity score approach should be assessed in terms of how well it has balanced covariates across treatment groups. Technically speaking, perfect balance would imply the same multivariate covariate distributions across treatment groups, though this is likely impossible to achieve let alone diagnose. A less ambitious goal is to only balance covariates in ways that reflect their role in a hypothesized model for the outcome [60•]. For example, an additive outcome model would suggest that balance of marginal means is sufficient, whereas a non-additive model would suggest that relevant interactions also be balanced. Moreover, balance on covariate transformations (such as a log-transformation or higher-order terms) should also be achieved if these are related to the outcome [61]. Therefore, aggregate measures of balance such as overlap in treatment densities, the Mahalanobis distance, or average standardized mean difference, may be useful in detecting gross imbalance but could still mask important differences [62•, 63].

Balance should thus be assessed for each covariate. While hypothesis tests of equality of means tend to reject when residual imbalances threaten causal inference, they are generally avoided in propensity score analyses as balance is an "in-sample" property and hypothesis tests conflate substantive differences with sample size [10, 64]; in large healthcare datasets, even clinically irrelevant differences can manifest as statistically significant. The standardized mean difference across each covariate can be reported by dividing the difference in covariate means by the pooled standard deviation in the original population (e.g., unmatched/unweighted). In terms of benchmarks, absolute standardized mean differences of less than 0.25 or 0.1 have been put forth as a rule of thumb, but ideally imbalances should be minimized without limit [10, 60]. One can go beyond the means to diagnose differences in covariate distributions using ratios of variance, the Kolmogorov-Smirnov distance, or box plots. It is worth pointing out that covariate balance is merely a sufficient condition for comparability of treatment groups [65]. The predicted counterfactual outcome among the referent treatment group can be leveraged to assess "prognostic" balance that summarizes over multiple covariates [66]. With all balance measures, their assessment should align with how the propensity score is to be used. They should be applied in matched samples, after inverse probability weighting, or within levels of propensity score subclasses [67, 68•, 69]. Residual imbalances can be tackled by including the unbalanced covariates in the outcome model, which is a form of double robustness [70].

Estimating Treatment Effects in Observational Studies

Matching

Matching is the most popular use of the propensity score in pharmacoepidemiology and will generally estimate the SATT [63]. Here, the propensity score is used as a measure of distance between treated and comparison units. The simplest algorithm is to use the propensity score to find a "matched" comparison for each treated unit ("one-to-one matching"). Differences in the outcome between treated and comparison groups in the matched sample, achieved non-parametrically or otherwise, provide estimates of the treatment effect. A frequent modification to this nearest neighbor approach is to only select comparisons that fall within a certain distance of the propensity score (a caliper). Though this leads to better balance, it can cause some treated units to be discarded, such that the target reflects an "overlap" population rather than the full treated group [71]. Most implementations of matching use a "greedy" algorithm that can exhaust the best matches early in the process without regard for the overall similarity of the treated and comparison groups. An alternative, optimal matching, seeks to minimize the average distance across pairs. Another variation finds more than one match for each treated unit (1:*k* matching, variable ratio matching). While improving precision, this can increase bias due to the inclusion of more distant comparisons [72]. See Stuart for a discussion on these tradeoffs [67]. It may be possible to apply a weight of $1/k$ after variable ratio matching to decrease the influence of larger matched sets, but the implications and utility of this proposal should be studied further. Extending these matching approaches to the generalized propensity score requires choosing an appropriate distance measure targeting a particular SATT or a population with common support, but success will depend on the degree and nature of overlap [73]. Exciting new methods such as cardinality matching and others bypass estimation of the propensity score entirely while optimizing sample size given specified covariate balance constraints (even for categorical treatments) [60]. Future empirical research might compare their performance with existing approaches, especially when the propensity score and prognostic score are integrated as distance measures.

Subclassification

An alternative to matching is to divide the population into subclasses according to the propensity score distribution in the overall populations or a particular treatment group [74]. Subclass indicators and their interactions with treatment can be entered as covariates in a regression model, otherwise marginal estimates of the SATE or SATT can be obtained by averaging over the subclass-specific effects through weighting [75]. The subclasses may be based on percentiles, quantiles, or some other scheme. The optimal classification strategy and its granularity may depend on whether treatment and outcome are rare, e.g., as in the case of safety outcomes for newly marketed medications [76]. Nevertheless, the working assumption is that within chosen subclasses, the treated and comparisons have similar covariate distributions which can be confirmed empirically [66]. It has been shown that creating just five to ten subclasses can remove at least 90% of the bias attributable to the covariates used to construct the propensity score [77]. With very fine strata, subclassification is akin to full matching; at its limit, it implies inverse probability weighting [78]. Subclassification can also be used with the generalized propensity score [23].

Weighting

In both matching and subclassification, the values of the propensity score are used to create sets where the treated and control have similar propensity scores (though not exactly so) and thus similar covariate distributions. In contrast, weighting uses the propensity score values directly. A weighting approach that estimates the SATE defines weights as the inverse of the probability of treatment received $\frac{1}{P[A=a|X]}$, mirroring weights in survey sampling [79]. The weights can then be used in non-parametric or parametric analyses of the SATE. Hypothesis tests and 95% confidence intervals can be constructed with a robust sandwich variance estimator [80]. To achieve more precision, the weights can be standardized by including the unconditional probability of treatment received in the numerator $\frac{P[A=a]}{P[A=a|X]}$. Assessments of effect heterogeneity across a moderator V should include it within the conditioning event of the numerator and denominator, i.e., $\frac{P[A=a|V]}{P[A=a|X,V]}$ [81]. This formulation can also be extended to continuous treatments [16]. If interest lies in the effect among a particular treatment group (i.e., the SATT) then a weighting strategy sometimes coined as "weighting by the odds" or "SMR weighting" uses that treatment's conditional probability as the numerator, i.e., $\frac{P[A=a'|X]}{P[A=a|X]}$, where a' is the reference treatment level of interest and a denotes the unit's actual treatment condition [82]. By definition, the weights for individuals in the reference treatment group of interest reduce to 1.

A limitation of weighting is that areas of limited overlap can result in low treatment probabilities and extremely large weights for certain individuals. These result in wider 95% confidence intervals, though if the propensity score model and weight specifications are correct, this level of influence and uncertainty are appropriate. In practice, though, it is common to sacrifice a little validity for precision by truncating the weights to the 99th or other percentile [83]. The impact of this procedure can be assessed by inspecting covariate balance [68•, 84••].

To some extent, areas of non-overlap can be addressed by "trimming" the sample to remove treated observations in the tails of the propensity score distribution that may lack comparison or treated units (as with caliper matching) [12, 13]. This can be particularly relevant, and important, in pharmacoepidemiology contexts where there is a group of people who would never be prescribed one of the drugs (treatments) of interest, because of contraindications or some other reason. Restricting attention to individuals in the area of common support can be thought of as focusing on the individuals for whom there is some clinical equipoise in terms of which treatment to provide [85]. An exciting advance is the development of weighting strategies that target not the SATE or SATT but the treatment effect among the population with common support. These include the "matching weight" which asymptotically targets a one-to-one matched sample by replacing the numerator with the minimum of the conditional treatment probabilities: e.g., $\frac{\min(P[A=0|X],P[A=1|X])}{P[A=a|X]}$ in the binary case [86]. An alternative is the "overlap weight" which targets the entire population of common support by replacing the numerator with the product of the conditional treatment probabilities: e.g., $\frac{P[A=0|X]\times P[A=1|X]}{P[A=a|X]}$ [87]. Like standard inverse probability weighting, they can accommodate categorical treatments but have the added benefit of being bounded between 0 and 1 [87, 88]. For categorical treatments, they do, however, presuppose sufficient overlap across all (and not just pairwise) comparisons. Early studies suggest that, in the context of up to three unequally sized treatment groups and rare binary outcomes, matching weights exhibit better balance and lower bias and mean square error compared to standard inverse probability weighting and matching [88, 89]. But their performance in simulated or empirical cases with more than three treatment groups (a realistic setting in pharmacoepidemiology) has yet to be evaluated.

Regression

A final approach is to regress the outcome on treatment and the estimated propensity score. In pharmacoepidemiology where outcomes are often non-linear and treatment effects are heterogeneous, estimating unbiased treatment effects will require that such models are correctly specified (which might involve higher order or spline terms for the propensity score and its interaction with treatment) [90]. For this reason, other uses of the propensity score are generally favored. Nonetheless, in recent years, some recent theoretical insights have been developed [90]. If the propensity score model is correct, in large samples, one can obtain valid tests of the null hypothesis of no treatment effect (on the difference or ratio scale) provided a robust variance estimator is used [90]. It remains unclear how to diagnose the balancing property of the estimated propensity score when it is to be used as a regressor.

Joint and Time-Varying Treatments

Our review has focused on causal inference for treatment decisions at a common (single) decision point. But pharmacoepidemiology often explores effects of discontinuing therapy or of adhering to treatment plans. Moreover, treatment decisions involving the addition or withdrawal of a second drug to increase effectiveness, mitigate side-effects, or to treat an emerging comorbidity, are vitally important. With each of these questions, including that of drug-drug interaction, naïve regression on time-varying versions of propensity scores can be inappropriate and induce bias because time-varying covariates (and propensity score summaries of them) may be affected by treatment [91, 92]. Special methods such as marginal structural models, structural nested models, or adaptations of the g-formula are required to adequately adjust for confounding in the presence of such feedback loops [80, 81, 93–96]. The spirit of design in propensity score approaches can be retained in these analyses by empirically assessing feedback and adapting balance measures to account for treatment history [84••].

Bias Correction for Unobserved Confounding

So far our discussion has assumed that there is no unmeasured confounding. When this assumption fails, the treatment effect estimate will be biased. Techniques are available to help quantify the potential bias. If the unmeasured confounder(s) are measured in a subset or external sample, a technique known as propensity score calibration has been proposed along with a two-stage approach that avoids its surrogacy assumption [97–99]. An alternative strategy is to carry out a sensitivity analysis (subsumed within "bias analysis") that models potential bias from unmeasured confounding as a function of how strongly the treatment and outcome are associated with an unmeasured confounder [100–102]. Practical implementations of these tools rely on certain no-interaction assumptions, but recent work has provided new bounding formulas that do not require them and are much easier to use [103••]. One can also pursue a sensitivity analysis by defining a bias parameter as the difference in potential outcomes across treatment groups given the propensity score [104]. Uncertainty in the parameters can be incorporated using frequentist or Bayesian frameworks, as in probabilistic bias analysis [105, 106]. The concept of design sensitivity extends the logic of sensitivity analysis to compare implications of potential unmeasured bias across a range of study design features and can be suited to evaluate effects that are substantial but rare [107••]. Concerning longitudinal treatment effects, many important time-varying confounders are often not available in electronic healthcare databases. Though a few correction and sensitivity-analysis techniques exist, the easier-to-use tools we have described have yet to be translated to these settings [108, 109].

Conclusions

Since their adoption into pharmacoepidemiology, propensity score methods have expanded in ways that are important for studies of comparative effectiveness and drug safety.

Extensions include ways to simultaneously compare multiple drugs, both after a fixed decision point and over time. Novel matching algorithms have emerged that provide greater control over balance and sample size. Weighting estimators have evolved to target matched and overlap estimands and in doing so avoid potentially harmful extrapolation. Sensitivity and bias analysis techniques have emerged that can leverage validation data or examine robustness to unobserved confounding. And although not covered in detail in this review, extensions for time-varying exposures have also progressed, including our ability to approach such questions in a design-based paradigm. The field is still wrestling with issues of covariate and model selection, as well as variance estimation, and machine-learning algorithms are being re-tuned away from minimizing prediction error towards improving the quality of the treatment effect estimate [110, 111]. Many of these new and exciting developments will need to be explored through further simulation and empirical examples, but together they represent a bright path ahead in overcoming many of the design and analytic challenges in the study of therapeutic effects and harms.

Funding Information Ian Schmid reports grants from the US Department of Education Institute of Education Sciences during the conduct of the study. Elizabeth A. Stuart reports grants from the National Institute of Mental Health during the conduct of the study, grants from the Patient Centered Outcomes Research Institute, and grants from the US Department of Education, Institute of Education Sciences, outside the submitted work.

Compliance with Ethical Standards

Conflict of Interest The authors declare that they have no conflict of interest.

Human and Animal Rights and Informed Consent This article contains no studies with human or animal subjects performed by any of the authors.

References

Papers of particular interest, published recently, have been highlighted as:
- • Of importance
- •• Of major importance

1. Walker AM. Confounding by indication. Epidemiology. 1996;7:335–6.
2. Schneeweiss S, Avorn J. A review of uses of health care utilization databases for epidemiologic research on therapeutics. J Clin Epidemiol. 2005;58:323–37.
3.•• Hernán MA, Robins JM. Using big data to emulate a target trial when a randomized trial is not available. Am J Epidemiol. 2016;183:758–64. **Elaborates how to construct an observational study in the image of a target randomized trial from large observational datasets, i.e. "big data". Considers how to adjust the formulation of the target trial according to the quality of the observational data, how to manipulate these data to emulate each of the main components of the target trial, and how to address potential methodological challenges posed by the observational nature of these data**
4. Ray WA. Evaluating medication effects outside of clinical trials: new-user designs. Am J Epidemiol. 2003;158:915–20.
5. Johnson ES, Bartman BA, Briesacher BA, et al. The incident user design in comparative effectiveness research. Pharmacoepidemiol Drug Saf. 2013;22:1–6.
6.• Brookhart MA. Counterpoint: the treatment decision design. Am J Epidemiol. 2015;182:840–5. **The treatment decision design extends the new-user design to address pharmacoepidemiological problems beyond those in which patients are observed from the start of exposure without compromising its ability to establish temporal ordering among study variables and yield causal estimates for clinically relevant comparisons**
7. Seeger JD, Walker AM, Williams PL, Saperia GM, Sacks FM. A propensity score-matched cohort study of the effect of statins, mainly fluvastatin, on the occurrence of acute myocardial infarction. Am J Cardiol. 2003;92:1447–51.
8. Toh S, Gagne JJ, Rassen JA, Fireman BH, Kulldorff M, Brown JS. Confounding adjustment in comparative effectiveness research conducted within distributed research networks. Med Care. 2013;51:S4–10.
9. Hernán MA, Sauer BC, Hernández-Díaz S, Platt R, Shrier I. Specifying a target trial prevents immortal time bias and other self-inflicted injuries in observational analyses. J Clin Epidemiol. 2016;79:70–5.
10. Imai K, King G, Stuart EA. Misunderstandings between experimentalists and observationalists about causal inference. J R Stat Soc Ser A. 2008;171:481–502.
11. Kurth T, Walker AM, Glynn RJ, Chan KA, Gaziano JM, Berger K, et al. Results of multivariable logistic regression, propensity matching, propensity adjustment, and propensity-based weighting under conditions of nonuniform effect. Am J Epidemiol. 2006;163:262–70.
12. Crump RK, Hotz VJ, Imbens GW, Mitnik OA. Dealing with limited overlap in estimation of average treatment effects. Biometrika. 2009;96:187–99.
13. Stürmer T, Rothman KJ, Avorn J, Glynn RJ. Treatment effects in the presence of unmeasured confounding: dealing with observations in the tails of the propensity score distribution-a simulation study. Am J Epidemiol. 2010;172:843–54.
14. Neyman J. On the application of probability theory to agricultural experiments. Essay on principles. Section 9. Stat Sci. 5:465–72.
15. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. Biometrika. 1983;70:41–55.
16. Hernán MA, Robins JM. Estimating causal effects from epidemiological data. J Epidemiol Community Health. 2006;60:578–86.
17. Rubin DB. Comment on: randomization analysis of experimental data: the Fisher randomization test by D. Basu J Am Stat Assoc. 1980;75:575–82.
18. VanderWeele TJ. Concerning the consistency assumption in causal inference. Epidemiology. 2009;20:880–3.
19. Halloran ME, Hudgens MG. Dependent happenings: a recent methodological review. Curr Epidemiol Reports. 2016;3:297–305.
20. Vanderweele TJ, Hernán MA. Causal inference under multiple versions of treatment. J Causal Inference. 2014;1:1–20.
21. Joffe MM, Rosenbaum PR. Invited commentary: propensity scores. Am J Epidemiol. 1999;150:327–33.
22. Imbens GW. The role of the propensity score in estimating dose-response functions. Biometrika. 2000;87:706–10.

23. Imai K, van Dyk D (2004) Causal inference with general treatment regimes: generalizing the propensity score. J Am Stat Assoc 99: 854–866.

24. Westreich D, Cole SR, Funk MJ, Brookhart MA, Stürmer T. The role of the c-statistic in variable selection for propensity score models. Pharmacoepidemiol Drug Saf. 2011;20:317–20.

25. Moodie EM, Stephens DA. Treatment prediction, balance and propensity score adjustment. Epidemiology. 2017; https://doi.org/10.1097/EDE.0000000000000657.

26. Sauer BC, Brookhart MA, Roy J, VanderWeele TJ. A review of covariate selection for non-experimental comparative effectiveness research. Pharmacoepidemiol Drug Saf. 2013;22:1139–45.

27. Brookhart MA, Schneeweiss S, Rothman KJ, Glynn RJ, Avorn J, Stürmer T. Variable selection for propensity score models. Am J Epidemiol. 2006;163:1149–56.

28. Ding P, VanderWeele TJ, Robins JM. Instrumental variables as bias amplifiers with general outcome and confounding. Biometrika. 2017;104:291–302.

29. Myers JA, Rassen JA, Gagne JJ, Huybrechts KF, Schneeweiss S, Rothman KJ, et al. Effects of adjusting for instrumental variables on bias and precision of effect estimates. Am J Epidemiol. 2011;174:1213–22.

30. Zhu Y, Schonbach M, Coffman DL, Williams JS. Variable selection for propensity score estimation via balancing covariates. Epidemiology. 2015;26:e14–5.

31.• Ding P, Miratrix L. To adjust or not to adjust? Sensitivity analysis of M-bias and butterfly-bias. J Causal Inference. 2014;3:41–57. **Addresses the debate as to whether one should adjust in M-structures in which a pretreatment covariate M is a collider for two latent factors. Presents theoretical results comparing the bias between adjusting and not adjusting for M in various scenarios of linear structural equation models, including independent latent factors, correlated latent factors, and when M is also a confounder. Advises for adjusting for M in general except for in certain situations, e.g., when the system is close to deterministic**

32. Schneeweiss S, Rassen JA, Glynn RJ, Avorn J, Mogun H, Brookhart MA. High-dimensional propensity score adjustment in studies of treatment effects using health care claims data. Epidemiology. 2009;20:512–22.

33. Schuster T, Pang M, Platt RW. On the role of marginal confounder prevalence—implications for the high-dimensional propensity score algorithm. Pharmacoepidemiol Drug Saf. 2015;24:1004–7.

34. Franklin JM, Eddings W, Glynn RJ, Schneeweiss S. Regularized regression versus the high-dimensional propensity score for confounding adjustment in secondary database analyses. Am J Epidemiol. 2015;182:651–9.

35. Schneeweiss S, Eddings W, Glynn RJ, Patorno E, Rassen J, Franklin JM. Variable selection for confounding adjustment in high-dimensional covariate spaces when analyzing healthcare databases. Epidemiology. 2017;28:237–48.

36. Vanderweele TJ, Shpitser I. A new criterion for confounder selection. Biometrics. 2011;67:1406–13.

37. Vansteelandt S, Bekaert M, Claeskens G. On model selection and model misspecification in causal inference. Stat Methods Med Res. 2012;21:7–30.

38. Wang C, Parmigiani G, Dominici F. Bayesian effect estimation accounting for adjustment uncertainty. Biometrics. 2012;68:661–71.

39. Gruber S, van der Laan MJ (2015) Consistent causal effect estimation under dual misspecification and implications for confounder selection procedures. Stat Methods Med Res 24:1003–1008.

40. Schnitzer ME, Lok JJ, Gruber S. Variable selection for confounder control, flexible modeling and collaborative targeted minimum loss-based estimation in causal inference. Int J Biostat. 2016;12: 97–115.

41. Greenland S, Daniel R, Pearce N. Outcome modelling strategies in epidemiology: traditional methods and basic alternatives. Int J Epidemiol. 2016;45:565–75.

42. Wyss R, Girman CJ, LoCasale RJ, Brookhart AM, Stürmer T. Variable selection for propensity score models when estimating treatment effects on multiple outcomes: a simulation study. Pharmacoepidemiol Drug Saf. 2013;22:77–85.

43. Gilbertson DT, Bradbury BD, Wetmore JB, et al. Controlling confounding of treatment effects in administrative data in the presence of time-varying baseline confounders. Pharmacoepidemiol Drug Saf. 2016;25:269–77.

44. Brunelli SM, Gagne JJ, Huybrechts KF, Wang SV, Patrick AR, Rothman KJ, et al. Estimation using all available covariate information versus a fixed look-back window for dichotomous covariates. Pharmacoepidemiol Drug Saf. 2013;22:542–50.

45. Nakasian SS, Rassen JA, Franklin JM. Effects of expanding the look-back period to all available data in the assessment of covariates. Pharmacoepidemiol Drug Saf. 2017; https://doi.org/10.1002/pds.4210.

46. Brookhart MA, Sturmer T, Glynn RJ, Rassen JA, Schneeweiss S. Confounding control in healthcare database research: challenges and potential approaches. Med Care. 2010;48:S114–20.

47. Lee BK, Lessler J, Stuart EA. Improving propensity score weighting using machine learning. Stat Med. 2010;29:337–46.

48. Shortreed SM, Ertefaie A. Outcome-adaptive lasso: variable selection for causal inference. Biometrics. 2017; https://doi.org/10.1111/biom.12679.

49. McCaffrey DF, Ridgeway G, Morral AR. Propensity score estimation with boosted regression for evaluating causal effects in observational studies. Psychol Methods. 2004;9:403–25.

50. Mccaffrey DF, Griffin BA, Almirall D, Slaughter ME, Ramchand R, Burgette LF. A tutorial on propensity score estimation for multiple treatments using generalized boosted models. Stat Med. 2013;32:3388–414.

51. Imai K, Ratkovic M. Covariate balancing propensity score. J. R. Statist. Soc. B. 2014;76:243–63.

52. Ning Y, Peng S, Imai K (2017) High dimensional propensity score estimation via covariate balancing. Available at http://imai.princeton.edu/research/hdCBPS.html. Accessed 30 June 2017.

53. Wyss R, Ellis AR, Brookhart MA, Girman CJ, Funk MJ, LoCasale R, et al. The role of prediction modeling in propensity score estimation: an evaluation of logistic regression, bcart, and the covariate-balancing propensity score. Am J Epidemiol. 2014;180:645–55.

54. McCaffrey DF, Lockwood JR, Setodji CM. Inverse probability weighting with error-prone covariates. Biometrika. 2013;100: 671–80.

55. Hong H, Rudolph KE, Stuart EA. Bayesian approach for addressing differential covariate measurement error in propensity score methods. Psychometrika. 2016;1–19

56. Webb-Vargas Y, Rudolph KE, Lenis D, Murakami P, Stuart EA. An imputation-based solution to using mismeasured covariates in propensity score analysis. Stat Methods Med Res. 2015; https://doi.org/10.1177/0962280215588771.

57. Walker AM. Matching on provider is risky. J Clin Epidemiol. 2013;66:S65–8.

58. Dusetzina SB, Mack CD, Stürmer T. Propensity score estimation to address calendar time-specific channeling in comparative effectiveness research of second generation antipsychotics. PLoS One. 2013; https://doi.org/10.1371/journal.pone.0063973.

59. Wyss R, Ellis AR, Lunt M, Brookhart MA, Glynn RJ, Stürmer T. Model misspecification when excluding instrumental variables from PS models in settings where instruments modify the effects of covariates on treatment. Epidemiol Method. 2014;3:83–96.

60.• de los Angeles Resa M, Zubizarreta JR. Evaluation of subset matching methods and forms of covariate balance. Stat Med.

2016;35:4961–79. **Finds through simulation studies that optimal matching methods such as cardinality matching and optimal subset matching outperform nearest neighbor matching with respect to balancing covariates, maximizing size of the matched samples, minimizing covariate distances between matched pairs, and estimating the treatment effect. Advises matching with fine balance on nominal covariates (i.e. forcing marginal distributions to be identical between treated and comparison groups) and with stronger balance than the heuristic of limiting standardized mean differences to under 0.1**

61. Belitser S V., Martens EP, Pestman WR, Groenwold RHH, de Boer A, Klungel OH (2011) Measuring balance and model selection in propensity score methods. Pharmacoepidemiol Drug Saf 20:1115–1129.

62.• Franklin JM, Rassen JA, Ackermann D, Bartels DB, Schneeweiss S. Metrics for covariate balance in cohort studies of causal effects. Stat Med. 2014;33:1685–99. **Compares ten single overall measures with respect to their association with bias in the estimation of a treatment effect across seven simulation scenarios with varying specifications of the covariate-exposure associations, covariate-outcome associations, and sample size. Concludes that the standardized difference, post-matching C-statistic, and general weighted difference performed the best overall**

63. Ali MS, Groenwold RHH, Belitser S V., Pestman WR, Hoes AW, Roes KCB, Boer A De, Klungel OH (2015) Reporting of covariate selection and balance assessment in propensity score analysis is suboptimal: a systematic review. J Clin Epidemiol 68:112–121.

64. Hansen BB. The essential role of balance tests in propensity-matched observational studies: comments on “a critical appraisal of propensity-score matching in the medical literature between 1996 and 2003” by Peter Austin, statistics in medicine. Stat Med. 2008;27:2050–4.

65. Hansen BB. The prognostic analogue of the propensity score. Biometrika. 2008;95:481–8.

66. Leacy FP, Stuart EA. On the joint use of propensity and prognostic scores in estimation of the average treatment effect on the treated: a simulation study. Stat Med. 2014;33:3488–508.

67. Stuart EA. Matching methods for causal inference: a review and a look forward. Stat Sci. 2010;25:1–21.

68.• Austin PC, Stuart EA. Moving towards best practice when using inverse probability of treatment weighting (IPTW) using the propensity score to estimate causal treatment effects in observational studies. Stat Med. 2015;34:3661–79. **Notes that most applications of IPTW using the propensity score to estimate effects do not include essential balance diagnostics. Describes the importance of measuring balance in the weighted sample. Advises using weighted standardized differences to compare means, higher-order moments, and interactions as well as cumulative distribution functions, side-by-side boxplots, and the Kolmogorov-Smirnov test statistic to compare both qualitatively and quantitatively the distributions of continuous variables between treatment groups in the weighted sample**

69. Rosenbaum PR, Rubin DB. Reducing bias in observational studies using subclassification on the propensity score. J Am Stat Assoc. 1984;79:516–24.

70. Kang JDY, Schafer JL. Demystifying double robustness: a comparison of alternative strategies for estimating a population mean from incomplete data. Stat Sci. 2007;22:523–39.

71. Lunt M. Selecting an appropriate caliper can be essential for achieving good balance with propensity score matching. Am J Epidemiol. 2014;179:226–35.

72. Wang SV, Schneeweiss S, Rassen JA. Optimal matching ratios in drug safety surveillance. Epidemiology. 2014;25:772–3.

73. Rassen J, Shelat A, Franklin JM, Glynn RJ, Solomon DH, Schneeweiss S. Matching by propensity score in cohort studies with three treatment groups. Epidemiology. 2013;24:401–9.

74. D’agostino RB. Tutorial in biostatistics propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. Stat Med Stat Med. 1998;17:2265–81.

75. Rudolph KE, Colson KE, Stuart EA, Ahern J. Optimally combining propensity score subclasses. Stat Med. 2016;35:4937–47.

76. Desai RJ, Rothman KJ, Bateman BT, Hernandez-Diaz S, Huybrechts KF. A propensity score based fine stratification approach for confounding adjustment when exposure is infrequent. Epidemiology. 2016;28:249–57.

77. Rosenbaum PR, Rubin DB. Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. Am Stat. 1985;39:33–8.

78. Hansen BB. Full matching in an oservational study of coaching for the SAT. J Am Stat Assoc. 2004;99:609–18.

79. Horvitz DG, Thompson D. A generalization of sampling without replacement from a finite universe. J Am Stat Assoc. 1952;44: 663–85. Available from: http://www.jstor.org/stable/2280784

80. Robins JM, Hernán MA, Brumback B. Marginal structural models and causal inference in epidemiology. Epidemiology. 2000;11: 550–60.

81. VanderWeele TJ. On the distinction between interaction and effect modification. Epidemiology. 2009;20:863–71.

82. Sato T, Matsuyama Y. Marginal structural models as a tool for standardization. Epidemiology. 2003;14:680–6.

83. Cole SR, Hernán MA. Constructing inverse probability weights for marginal structural models. Am J Epidemiol. 2008;168:656–64.

84.•• Jackson JW. Diagnostics for confounding of time-varying and other joint exposures. Epidemiology. 2016;27:859–69. **Provides a framework to assess confounding with respect to joint or time-varying exposures that are common in pharmacoepidemiology. This includes one diagnostic that assesses time-varying confounding in the study population, another that identifies exposure-covariate feedback that indicates the use of g-methods, and a third that assesses time-varying confounding in weighted or stratified populations following the use of g-methods. Further guidance is given regarding how to estimate these diagnostics, present them graphically, and adapt them to settings of right censoring**

85. Schneeweiss S. Developments in post-marketing comparative effectiveness research. Clin Pharmacol Ther. 2007;82:143–56.

86. Li L, Greene T. A weighting analogue to pair matching in propensity score analysis. Int J Biostat. 2013;9:215–34.

87. Li F, Morgan KL, Zaslavsky AM. Balancing covariates via propensity score weighting. J Am Stat Assoc. 2016; https://doi.org/10.1080/01621459.2016.1260466.

88. Yoshida K, Hernández-Díaz S, Solomon DH, Jackson JW, Gagne JJ, Glynn RJ, et al. Matching weights to simultaneously compare three treatment groups. Epidemiology. 2017;28:387–95.

89. Franklin JM, Eddings W, Austin PC, Stuart EA, Schneeweiss S. Comparing the performance of propensity score methods in healthcare database studies with rare outcomes. Stat Med. 2017; https://doi.org/10.1002/sim.7250.

90. Vansteelandt S, Daniel RM. On regression adjustment for the propensity score. Stat Med. 2014;33:4053–72.

91. Ray WA, Liu Q, Shepherd BE. Performance of time-dependent propensity scores: a pharmacoepidemiology case study. Pharmacoepidemiol Drug Saf. 2015;24:98–106.

92. Hernán MA, Robins JM. Longitudinal causal inference. In: International encyclopedia of the social behavioral sciences. 2nd ed. Oxford, England: Elsevier; 2015. p. 340–4.

93. Westreich D, Cole SR, Young JG, Palella F, Tien PC, Kingsley L, et al. The parametric g-formula to estimate the effect of highly active antiretroviral therapy on incident AIDS or death. Stat Med. 2012;31:2000–9.

94. Vansteelandt S, Joffe M. Structural nested models and G-estimation: the partially realized promise. Stat Sci. 2014;29:707–31.

95. Shinohara RT, Narayan AK, Hong K, Kim HS, Coresh J, Streiff MB, et al. Estimating parsimonious models of longitudinal causal effects using regressions on propensity scores. Stat Med. 2013;32: 3829–37.

96. VanderWeele TJ, Jackson JW, Li S. Causal inference and longitudinal data: a case study of religion and mental health. Soc Psychiatry Psychiatr Epidemiol. 2016;51:1457–66.

97. Stürmer T, Schneeweiss S, Avorn J, Glynn RJ. Adjusting effect estimates for unmeasured confounding with validation data using propensity score calibration. Am J Epidemiol. 2005;162:279–89.

98. Stürmer T, Schneeweiss S, Rothman KJ, Avorn J, Glynn RJ. Performance of propensity score calibration–a simulation study. Am J Epidemiol. 2007;165:1110–8.

99. Lin HW, Chen YH. Adjustment for missing confounders in studies based on observational databases: 2-stage calibration combining propensity scores from primary and validation data. Am J Epidemiol. 2014;180:308–17.

100. Schneeweiss S. Sensitivity analysis and external adjustment for unmeasured confounders in epidemiologic database studies of therapeutics. Pharmacoepidemiol Drug Saf. 2006;15:291–303.

101. VanderWeele TJ, Arah OA. Bias formulas for sensitivity analysis of unmeasured confounding for general outcomes, treatments, and confounders. Epidemiology. 2011;22:42–52.

102. Arah OA. Bias analysis for uncontrolled confounding in the health sciences. Annu Rev Public Health. 2017;38:23–38.

103.•• Ding P, Vander Weele TJ. Sensitivity analysis without assumptions. Epidemiology. 2016;27:368–77. **Introduces a bounding factor for analyses of sensitivity to unobserved confounding in observational studies that does not require the investigator to assume that there is only a single binary confounder or that there is not exposure-confounder interaction of effects on the outcome. By specifying two sensitivity parameters in the bounding factor, representing the strength of confounding between the exposure and outcome induced by unmeasured confounding, the investigator can determine the degree to which unmeasured confounding could explain the observed effect estimate**

104. Li L, Shen C, Wu AC, Li X. Propensity score-based sensitivity analysis method for uncontrolled confounding. Am J Epidemiol. 2011;174:345–53.

105. McCandless LC, Gustafson P, Levy AR. A sensitivity analysis using information about measured confounders yielded improved uncertainty assessments for unmeasured confounding. J Clin Epidemiol. 2008;61:247–55.

106. Lash TL, Fox MP, Fink AK. Applying quantitative bias analysis to epidemiologic data. New York: Springer-Verlag; 2009.

107.•• Zubizarreta JR, Cerdá M, Rosenbaum PR. Effect of the 2010 Chilean earthquake on posttraumatic stress. Epidemiology. 2013;24:79–87. **Uses recently developed propensity score methods as part of a greater effort to use design and analytical elements that are tailored to detect the specific patterns of effect they have hypothesized. Promotes attention on how to formally mitigate sensitivity to unobserved confounding a priori**

108. Brumback B, Hernán MA, Haneuse SJ, Robins JM. Sensitivity analyses for unmeasured confounding assuming a marginal structural model for repeated measures. Stat Med. 2004;23:749–67.

109. Burne R, Abrahamowicz M. Martingale residual-based method to control for confounders measured only in a validation sample in time-to-event analysis. Stat Med. 2016;35:4588–606.

110. Zou B, Zou F, Shuster JJ, Tighe PJ, Koch GG, Zhou H. On variance estimate for covariate adjustment by propensity score analysis. Stat Med. 2016;35:3537–48.

111. Cefalu M, Dominici F, Arvold N, Parmigiani G. Model averaged double robust estimation. Biometrics. 2017;73:410–21.