Title: Using Propensity Scores to Estimate Effects of Treatment Initiation Decisions: State of the Science

Running Title: Propensity Scores for Treatment Decisions

Author list:

Michael Webster-Clark[1] (corresponding author)

Til Stürmer[1]

Tiansheng Wang[1]

Kenneth Man[2,3]

Danica Marinac-Dabic[4]

Kenneth J. Rothman[5,6]

Alan R. Ellis[7]

Mugdha Gokhale[1,8]

Mark Lunt[9]

Cynthia Girman[1,10]

Robert J. Glynn[11]

Affiliations:

1. Department of Epidemiology, UNC Chapel Hill, Chapel Hill, North Carolina, USA.
2. Research Department of Practice and Policy, UCL School of Pharmacy, London, UK.
3. Department of Pharmacology and Pharmacy, LKS Faculty of Medicine, University of Hong Kong, Hong Kong.
4. Office of Clinical Evidence and Analysis, Center for Devices and Radiological Health, FDA, USA.
5. RTI Health Solutions, Raleigh, NC, USA.
6. Department of Epidemiology, Boston University, Boston, Massachusetts, USA.
7. Department of Social Work, NC State University, Raleigh, North Carolina, USA.
8. Pharmacoepidemiology, Center for Observational & Real-World Evidence, Merck, West Point, Pennsylvania, USA.
9. The Arthritis Research UK Epidemiology Unit, University of Manchester, Manchester, UK.
10. CERobs Consulting, LLC; Chapel Hill, NC, USA.

11. Pharmacoepidemiology and Pharmacoeconomics, Brigham and Women's Hospital and Harvard Medical School, Boston, Massachusetts, USA.

Funding: This manuscript has been funded by the International Society for Pharmacoepidemiology (ISPE) and is going through ISPE's membership at-large review process. Author time was partly supported by R01 AG056479 from the National Institute on Aging.

Corresponding author contact information:
Michael Webster-Clark, PharmD, PhD
Department of Epidemiology, UNC Gillings School of Global Public Health
University of North Carolina at Chapel Hill
McGavran-Greenberg, CB #7435
Chapel Hill, NC 27599-7435
Phone: 1 919 966 7433
Fax: 1 919 966 2089

Keywords: Propensity scores; review; real world evidence; real world data; comparative effectiveness research
Word count: 7,514 words
Black and white figure count: 6
Table count: 1
Data availability statement: Literature review results are available upon request.

ABSTRACT (201/250 word limit)

Confounding can cause substantial bias in non-experimental studies that aim to estimate causal effects. Propensity score methods allow researchers to reduce bias from measured confounding by summarizing the distributions of many measured confounders in a single score based on the probability of receiving treatment. This score can then be used to mitigate imbalances in the distributions of these measured confounders between those who received the treatment of interest and those in the comparator population, resulting in less biased treatment effect estimates. This methodology was formalized by Rosenbaum and Rubin in 1983 and, since then, has been used increasingly often across a wide variety of scientific disciplines. In this review article, we provide an overview of propensity scores in the context of real world evidence generation with a focus on their use in the setting of single treatment decisions, i.e. choosing between two therapeutic options. We describe five aspects of propensity score analysis: alignment with the potential outcomes framework, implications for study design, estimation procedures, implementation options, and reporting. We add context to these concepts by highlighting how the types of comparator used, the implementation method, and balance assessment techniques have changed over time. Finally, we discuss evolving applications of propensity scores.

Generating relevant and reliable real world evidence on the comparative safety and effectiveness of medical treatments requires tools to reduce bias from confounding variables.[1,2] Both the availability of health data and the sophistication of analytic methods have increased over time due to innovations in statistics, epidemiology, digital health and computing. In the US, the 21st Century Cures Act and its paradigm-changing focus on Real World Evidence (RWE)[3] have amplified the demand for studies using routinely collected data to accelerate medical product innovation, and similar efforts are underway internationally.[4] The complexity of available data has also increased, especially with the ability to link across many data sources. While investigators in the 1950s trying to understand the causal relationship between smoking and lung cancer had access to data on only a handful of potentially confounding variables,[5] today's researchers have access to data on dozens if not hundreds of variables for users of a given drug, device or surgical therapy (though these variables are often measured with error).[6,7] The increasing number of measured confounders and the focus on marginal, rather than conditional, causal effects has rendered legacy techniques like full joint stratification increasingly unappealing.

Estimating and utilizing **propensity scores**, formally defined in 1983 as "the conditional probability of assignment to a particular treatment given a vector of observed covariates," is one way for modern researchers to make use of this rich data to reduce confounding in treatment effect estimates.[8,9] While creation of and stratification by forms of multivariable confounder scores predated this work,[10-12] some previous scores led to biased effect estimates while others were found to exaggerate precision.[13-17] Rosenbaum and Rubin's 1983 paper sharpened the focus to prediction of treatments in the entire study population and laid out a clear theoretical framework for the scores as well as three distinct ways to utilize them. Since then, propensity scores have been widely adopted as a tool to aid in estimating causal effects in applied research, and numerous excellent tutorials and orientations to aspects of propensity score analyses have been published across a variety of disciplines.[18-20]

The primary goal of this manuscript is to add to this body of work by providing an overview of the role that propensity scores currently play in generating real world evidence on treatment effects. We also highlight trends in implementation (use of active comparators, matching, and strategies to evaluate covariate balance) and recent methodological developments. To do so, we describe the theoretical framework of the propensity score in the context of

treatment decisions; study design considerations and recommendations when using propensity scores; methods for propensity score estimation and implementation; recommendations for propensity score reporting to facilitate the evaluation of the research as real world evidence; and evolving applications of propensity scores outside the context of estimating effects of a single treatment decision. To provide a foundation for this description, we conducted a literature review to quantify the increasing use of propensity score methods and examine changes in that use over time.

LITERATURE REVIEW

From 2004 to 2018, the period since the most recent systematic review by epidemiologists,[21] some 48,170 unique manuscripts were published with the phrase "propensity score" in their main text and indexed in Pubmed, Embase, Web of Science, or Scopus. From these 48,170 papers, we randomly sampled 300 articles that applied propensity scores to research questions. In the process of identifying these 300 articles, we excluded 258 articles that discussed methodological advancements, were referencing past studies, or otherwise did not apply propensity score methods. We reviewed all 300 articles to identify the type of comparator used, the type of implementation method, and any types of balance assessment.

As the 48,170 articles were not evenly distributed across calendar time and we wanted similarly precise estimates across this range of time, our set of 300 articles was made up of 75 articles randomly sampled from each of four calendar years (2004, 2009, 2014, and 2019). This strategy also allowed us to estimate the share of articles applying propensity scores for research purposes as percentages. We then applied said percentages to the raw number of articles identified in the databases to estimate the number of papers published in each calendar that used propensity scores for research. The full protocol for this literature search appears in **Appendix A.**

Figure 1 shows the 28-fold increase in the estimated number of papers applying propensity score methods per year from 220 in 2004 to 6,208 papers in 2018. For comparison, the total number of papers indexed by Medline doubled from 2004 to 2018. This widespread proliferation of propensity score methods highlights the increasing importance of understanding how to apply them.

PROPENSITY SCORE THEORY

The potential outcomes framework provides the theoretical basis for using propensity scores to control for measured confounding of treatment effects in non-randomized experiments.[8,22,23] When medical providers face a treatment decision (for example, whether to prescribe statin A or statin B to a patient), we can posit two contrasting potential outcomes for that patient: the outcome if statin A is prescribed (denoted by Y_A) and the outcome if statin B is prescribed instead (denoted by Y_B).[23] Although each of the subjects receives only one treatment, the absolute effect (i.e. the risk difference) of statin A versus statin B on each patient can then be described as $E(Y_A - Y_B)$, and the relative effect (i.e. the risk ratio) can be described as $E(Y_A/Y_B)$.

Typically, a given person or group will experience only one of their potential outcomes, $Y_{T=A}$ or $Y_{T=B}$, while the others are rendered unobservable (i.e. counterfactual). The absence of this counterfactual data makes it nearly impossible to directly observe these causal effects. We only have access to the realized potential outcomes in the two disjoint populations, which we will call $\bar{Y}_{T=A}$, the average outcome in the group that received treatment A, and $\bar{Y}_{T=B}$, the average outcome in the group that received treatment B (with the bars denoting population averages).

One way we could use these realized potential outcomes is comparing $\bar{Y}_{T=A}$ with $\bar{Y}_{T=B}$. This approach is problematic, as there may be variables that influence both treatment assignment and the outcome,[24] commonly referred to as **confounders.** If confounders are present, the population risk difference $\bar{Y}_{T=A} - \bar{Y}_{T=B}$ will generally not equal $E(\bar{Y}_A)–E(\bar{Y}_B)$; the discrepancy is known as confounding bias.

Removing this bias requires two things. First, it requires **consistency** of the treatment effects; each individual's potential outcome under a given treatment must be the outcome we observe when the individual is assigned that treatment.[25] Second, it requires patients receiving A and patients receiving B to be **exchangeable,** (i.e. the two treatment groups must possess the same covariate patterns resulting in similar baseline outcome risks except for the effect of treatment, including **positivity** for covariate patterns in both groups); if these conditions are met, then $\bar{Y}_{T=A} - \bar{Y}_{T=B}$ is equal to $E(\bar{Y}_A) - E(\bar{Y}_B)$.[26,25,27] Simple or stratified randomization to treatment A or B has become the standard to achieve exchangeability since, on average, randomization renders treatment allocation independent of other factors, whether observed or unobserved, that contribute to the outcome.

While randomization is the only way to balance unmeasured variables, there are other ways to achieve **conditional exchangeability** with respect to observed variables.[28,29] Restriction and matching have been used successfully for this purpose in scientific research for a long time. Under the assumption that all confounding variables had been measured well, restriction and matching were sometimes seen as equivalent to randomization, especially in small samples.[29] The curse of high dimensionality and rapid decrease in efficiency with simultaneous matching on many factors led to an interest in matching on summary scores.[10,14,30] Rosenbaum and Rubin[8] defined a balancing score as a function of the measured covariates *(x)* such that those with T=A and T=B have equal distributions of *(x)* given the balancing score, which thereby balances covariates across treatments A and B. Exact full stratification or perfect matching on all measured *(x)* can be thought of as the finest balancing score. The coarsest version that still balances covariates is the **propensity score**, the conditional probability of T=A given *(x)*.[8] An estimate of this conditional probability from realized data achieves balancing properties similar to the "true" probability of treatment assignment and can be used to remove confounding by measured variables through matching, stratification, modeling, weighting, or doubly-robust estimation.[23] Each of these methods creates cohorts that are exchangeable on the measured variables, estimating a variety of treatment effects with reduced bias from these factors.

There are a few caveats, however. Achieving covariate balance can require iterative fitting of the propensity score model. Unless correctly specified (e.g., by including relevant interaction or higher order terms), the propensity score may not balance univariate or multivariate distributions of *(x)* between treatment groups; moreover, it is impossible to know how complex these interaction and higher order terms need to be. Additionally, problems arise when some covariate combinations are exclusive to one treatment group (generally referred to as **non-positivity**).[31] Finally, the propensity score will not balance covariates that were not included in the score, particularly those that were unmeasured, except to the extent that unmeasured variables are correlated with those that are part of the propensity score model.[32]

Propensity score techniques have some drawbacks relative to adjustment for covariate differences via g-computation or standardization via outcome models. Most notably, propensity score analyses will generally result in less precise estimates (because they cannot approach parametric efficiency bounds) and can be more complex to implement.[33]

These drawbacks must be weighed against the benefits of using propensity scores. First, propensity score methods are preferable when it is easier or more plausible to identify the model for treatment than for the outcome, particularly in settings with few outcomes. In those cases, regression models for the outcome may be overfit with only a small number of confounding variables.[34-36] Second, unlike an outcome model, the performance of a propensity score at balancing covariates can be empirically checked (and the model refined to improve balance) without examining treatment group outcomes. Third, it is straightforward to check for covariate positivity after implementing the propensity score and to identify (and potentially exclude) types of patients who are virtually guaranteed to receive one of the treatments.[37,38]

Fourth, understanding the propensity score distribution in the treated and untreated can help researchers gain insight into whether there is insufficient overlap (empirical equipoise) between treatment groups to allow for meaningful comparative safety and effectiveness research.[39,40] Finally, if researchers wish to move beyond population level treatment effects, they can compare estimated effects across propensity score strata to identify treatment effect heterogeneity (given that providers are likely to channel specific treatment options towards those who they believe are more likely to benefit, treatments may be more beneficial for those in more extreme propensity score strata).[41] While these strata are not themselves clinically relevant, they can signal potential variability in benefit or risk; that said, it can be difficult to uncover which clinically relevant covariates are creating heterogeneous treatment effects.[42]

PROPENSITY SCORES AND STUDY DESIGN CONSIDERATIONS

Suppose, then, that the advantages of using a propensity score persuade researchers to adopt the method to estimate a treatment effect. Before deciding how to estimate or use the score, there are several key study design considerations. While these considerations are important in any study of medical treatments, the decisions below were often ignored in non-experimental work before propensity scores encouraged a focus on treatment assignment mechanisms and the importance of understanding the indications for and barriers to the use of the treatment of interest;[40,43] moreover, propensity scores can help inform some of these study design decisions, particularly with respect to identifying relevant study populations with empirical equipoise.[44]

Comparator Choice: One of the most critical decisions with respect to the analytic question and potential confounding is the choice of the treatments to be compared; typically, one is a treatment of interest and one is a **comparator**. Choice of comparator shapes and is shaped by the causal question being examined: if researchers want to estimate the effects of a treatment compared with no intervention, they should design the study to compare treated individuals with a **non-user** or **inactive comparator** group with similar health conditions. While this may seem similar to the use of placebos in randomized trials, the fact that non-users are simply continuing to receive nothing (rather than an intervention with no effect) means that the surveillance and care they receive may differ fundamentally and systematically from care received by treated individuals. Further, this difference in treatment may stem from differences in factors that are difficult to measure, such as baseline disease severity, frailty, lifestyle choices and behaviors, and risk of the outcome. While measured covariates can be integrated into the propensity score, unmeasured variables like these often contribute to **confounding by indication** (i.e. disease severity) which can bias estimates of treatment effect and yield misleading results.[45,46]

On the other hand, if the goal of the study is to compare the benefit-to-harm balance of a new drug in a class with its predecessors or other marketed products, it is likely more appropriate to use patients receiving those treatments to form an **active comparator** group.[47,48] Unlike non-user comparators[45,46], active comparators in many cases implicitly condition on the indication for treatment (and the severity of disease warranting treatment), resulting in considerable reductions in confounding by indication as well as possibly increased balance in other baseline covariates and risk of the outcome. These comparators also generally have more similar surveillance

surrounding confounding factors and contraindications, reducing the potential for differential confounder measurement error.[49] Based on our literature review (**Figure 2)**, active comparators were used in less than half of studies in 2004, 2009, and 2014, though by 2019, 57% of studies used some form of active comparator.

Starting Follow-up: Choosing when to start follow-up is also vital, as a lack of a clear time zero can result in invalid estimates of treatment effects.[50] Since propensity score theory is centered around the idea of treatment decisions, it is often useful to focus on the choice of treatment at the time of initiation by restricting the study population to **new users** of drug therapy, excluding prevalent users.[51] If instead prevalent users are included, the propensity score represents both the probability of initiating and remaining on treatment - a much more complex quantity.[48,50,52] Whether restricting to new users or prevalent users for the treatment of interest, it is often difficult to identify the start of follow-up for any non-user comparators. Properly using such non-user data is possible, but often complicates the analysis.[53] Whether studying new users, prevalent users, or non-users, improperly setting time-zero can lead to **immortal time** bias.[54]

Handling Subsequent Treatment: Another critical design decision with consequences for the causal question being examined is the extent to which subsequent treatment affects follow-up after time zero. Initial treatment designs follow individuals until the end of the study period under their first observed treatment, regardless of any stopping or switching; this is analogous to the intention-to-treat designs from randomized clinical trials. Since in non-experimental research the literal intention of the prescriber is rarely captured, it is sometimes referred to as an "initial treatment" approach. Such an analysis estimates the effect of **treatment initiation** given the population's persistence, adherence, re-initiation, and switching rates under each treatment. In real world settings, as time passes, that treatment effect will generally diverge more and more from the effects of initiation and continuous use of treatment.[55,56]

With on-treatment follow-up (i.e. as-treated follow-up), subjects are followed from treatment initiation until they deviate from some treatment protocol, typically by stopping or switching a therapy, at which point they are censored. Such a design estimates the hypothetical effect of **treatment initiation and continued persistence and adherence to a given protocol**[56,57] and produces estimates that are not conditional on the study-specific factors that shape initial-treatment estimates. The price paid, however, is that as-treated designs may show high effectiveness of treatments even if real world patients have poor treatment adherence and

persistence. Additionally, these designs open up the potential for selection bias via differential drop-out unless time-varying confounding is addressed appropriately (which is difficult, as treatment changes are often a function of subtle and not routinely captured differences in effectiveness and side effects).[58]

Study Population: Finally, the design stage requires a decision about the study population in which a treatment effect will be estimated, as heterogeneity in treatment effect can result in a difference in study findings depending on target population. Generally, investigators start by choosing from the effect of the treatment in the total population studied (population average treatment effect, or PATE) or in one of the arms being studied (average treatment effect in the treated, ATT, or average treatment effect in the comparator/untreated, ATU) or some other population entirely.[42]

Propensity scores can play a pivotal role in further refining this initial target population to a population with better exchangeability and reduced non-positivity. Are the investigators concerned about strong confounding among those with high or low probabilities of initial assignment to treatment, and if so are they planning to remove (trim) some of those individuals from the study population?[39,59,60] Excluding all those at the extremes, or "tails," of the propensity score distributions can improve the precision of estimates; moreover, if these individuals are already contraindicated or strongly indicated to receive one treatment, their best course of treatment may not be of interest to researchers. If trimming is to be used, specifying multiple trimming rules (e.g. different percentile cutpoints, symmetrical vs asymmetrical) at the design stage can help researchers protect against accusations of fishing for results while giving some insight into how much confounding (or effect heterogeneity) exists in the tails. Researchers should also be sure to describe those who were trimmed, make it clear that they have limited evidence about effects in them, and (if one characteristic is strongly predictive of being trimmed) consider explicitly reframing the study question to exclude those individuals.

PROPENSITY SCORE ESTIMATION

After design choices (including whether the propensity score will be used to shape the final target population) are made and data are gathered, the next step in a study using propensity scores is propensity score estimation. The propensity score can be used to balance treatment groups with respect to measured covariates. But which covariates should be balanced? Once we've chosen the covariates, how do we use those covariates to estimate the conditional probability of treatment? Once we have estimated this probability, what, if anything, can be done to check whether the balancing has been successful?

Variable Selection: The goal of the propensity score model is to balance the distribution of risk factors for the outcome across the treatment groups, while preserving variability in treatment assignment that is independent of outcome risk. The choice of covariates is critical, as including variables that predict only treatment in the propensity score reduces study efficiency, and that cost has to be weighted against gains in validity.[61,62]

Variable Types: Consider **Figure 3,** a directed acyclic graph that depicts assumed causal relations in the form of arrows from one variable to another. These arrows form causal paths that result in expected associations between variables.[63] Baseline covariates with open causal paths to the exposure but not the outcome, like the **instrumental** variable in **Figure 3A,** should not be included in propensity score models. These variables reduce precision[39] and amplify the effect of any unmeasured confounding (bias amplification).[64,65] Unfortunately, distinguishing these variables from confounders is usually impossible, and the comparatively small bias from including a true instrument versus excluding a variable with a very weak path to the outcome (i.e. a near-instrument) means that even near-instruments are typically worth including in the propensity score model.[66] On the other hand, including baseline variables with open causal paths to the outcome but not treatment assignment (like the risk factor variable in **Figure 3A**) can increase precision when random sampling error has led to spurious associations with treatment in the study sample.[61]

The final type of covariate, baseline variables with open causal paths to both treatment assignment and the outcome, are generally good candidates for inclusion in propensity score models. C_1 in **Figure 3A** is the archetypal example of such a variable. However, some variables meeting this description result in bias when included in the propensity score model if researchers are unable to close the paths opened by their inclusion.[67] Such variables are termed **colliders**

because they have arrows pointing into them from at least two other variables on a causal graph. That said, when a variable is both a confounder and a collider, like the "collider" variable in **Figure 3B,** the confounding bias will generally outweigh collider bias except in extreme scenarios.[68]

Notably, none of the above considerations about variable selection is specific to propensity score models. However, propensity scores have helped clarify these issues, in part because they may be prone to the inclusion of instrumental variables if misunderstood as pure treatment prediction models.

Selection Strategies: Several approaches can be used to select covariates for balancing. First, one could use *a priori* specified directed acyclic graphs like **Figure 3,** that depict assumed relations among variables based on prior knowledge, to identify adjustment sets that would render treatment assignment and the outcome independent except through effects of treatment.[63,69] These adjustment sets can be pared into what are sometimes referred to as **minimally sufficient adjustment sets**, and the sets that researchers believe can be measured with the least error can then be used to estimate the score.[70] This approach requires the causal graph to be properly specified, an untestable and often unrealistic assumption.

Another potential approach is to include all known factors that might be associated with the outcome or treatment in the data. This approach (sometimes called the **kitchen sink approach**) is often seen as a less subjective method with fewer assumptions compared with creating directed acyclic graphs to identify minimally sufficient adjustment sets, but the kitchen sink approach can induce bias from including colliders and amplify unmeasured confounding if instrumental variables are included in the propensity score model.[65] A slightly more restrictive version of this approach (including all variables that are causes of the outcome or treatment in a directed acyclic graph) has also been proposed.[71]

Finally, one can attempt to identify variables associated with treatment assignment and the outcome from the data by applying a selection algorithm to a vast quantity of potential baseline covariates. One such approach, called **high dimensional propensity scores,** identifies things like diagnosis codes or healthcare events that are associated with treatment assignment as well as the outcome and ranks them as candidates for the propensity score model based on their estimated confounding potential (including their association with treatment assignment and the outcome).[72] Notably, establishing this ranking by simply estimating marginal associations

between variables and the outcome does not always eliminate instrumental variables if treatment affects the outcome, since the instruments will, in expectation, be associated with the outcome through treatment.

Score Estimation: After the covariates are selected, they are used to estimate each participant's probability of receiving the treatment of interest. By far the most common estimation choice has been multivariable regression of treatment on the set of covariates, with the propensity score being the predicted probability of treatment for each person given their covariates.[73] Propensity score estimation usually involves logistic regression but can rely on the multinomial logit model for more than two exposure groups[74] or linear regression or more complex models for continuous treatments.[75] Multivariable regression is straightforward but requires decisions about what interactions and functional forms to use in the final model, including whether to categorize continuous covariates.

To aid in these decisions, researchers often specify a starting model and implement their analytic method (be it matching, stratification, or weighting; see **PROPENSITY SCORE IMPLEMENTATION** below), then check balance by comparing the standardized absolute mean differences (SAMDs) between the treatment groups for the covariates included in the model.[76] Larger SAMDs correspond to larger imbalances in covariates; if SAMDs are too large, researchers may re-fit the model with additional interaction terms or more flexible functional forms. Multiple iterations may be required to achieve acceptable covariate balance that reduces bias from measured confounders; to increase confidence in the results, researchers should pre-specify each step of the iterative process and avoid examining effect estimates while adjusting the model.[77,78] While such methods are used quite frequently, surprisingly little theoretical work has been done on their impact on the accuracy of standard errors of treatment effect estimates.

Another option for estimating the propensity score is the use of more flexible tools than logistic regression, especially **machine learning** approaches.[73,79,80] These classification and prediction techniques target a parameter like average standardized absolute mean difference (ASAMD) or overall performance of the prediction model and iterate through potential models and probability estimations until they identify a model or set of predictions that optimizes the target parameter. The result is a predicted probability of treatment for each individual in the data set, conditional on covariates, i.e. a propensity score. Researchers should be sure, however, to use cross-validation techniques with these data-driven approaches and to use methods that yield

appropriate standard errors and confidence limit coverage for the point estimate when using propensity scores created in this manner to avoid complications from overfitting.[81]

Regardless of propensity score estimation method, choosing the target parameter for balance assessment is difficult. While a low ASAMD typically evinces adequate overall balance, this value does not take into account that the effect of variable imbalances on bias depends on how strongly the imbalanced variables affect the risk of the outcome. Perfect balance of a near-instrument or many weak confounders in the presence of a largely imbalanced strong risk factor for the outcome can result in strong overall bias despite a low ASAMD. Given the need to balance distributions across groups, it can also be useful to assess variance ratios.[76] Currently, new methods are being developed and refined that incorporate strength of association with the outcome when assessing balance.[82-84]

PROPENSITY SCORE IMPLEMENTATION

Once each individual has a propensity score, the next step is actually using those scores to estimate a treatment effect. Propensity scores can be used in several different ways to estimate internally valid effects of treatment. Three approaches (**matching, stratification, and regression covariate adjustment**) were described by Rosenbaum and Rubin in 1983.[8] A fourth strategy, **weighting,** arose later and was combined with outcome-based approaches in the early 2000s to create a fifth category: **doubly robust** estimation.[85,86] While each of these methods can estimate a treatment effect without bias, and each will yield the same effect estimate when the treatment effect is homogeneous, their estimates may differ substantially in the presence of non-uniform treatment effects.[87,88]

Matching: Matching on the propensity score was one of the first methods to be developed as a way to improve the efficiency of matching in the presence of many covariates. After assigning propensity scores to all study participants, one group (typically the smaller, treated, group, if the comparison is between treated and untreated) is taken as the target group. Those in the comparator group whose propensity scores are "similar" to those in the target group are identified to be included in the analytic sample. Investigators have to choose what constitutes similarity: **nearest-neighbor** matching randomly chooses a target participant and matches it to the comparator participant with the closest propensity score (and repeats this process for the whole target group), while **optimal** matching algorithmically minimizes the overall distance between matched pairs in the data set. To avoid large differences in propensity scores within matched pairs, **calipers** are used to restrict the search for the nearest match to within some distance (say, 0.01, or 0.10, or 10% of the standard deviation of the propensity score) from the propensity score of the target participant.[89,90] Using calipers effectively trims from the analytic sample any target participants that are at least a half caliper width outside the zone of propensity score overlap, sometimes termed the region of common support.

Matching can be 1:1 (finding one match, at random if multiple exist, for each target group member) or one to many (finding a fixed number of matches, e.g. 2:1, or **all** matches in the comparator group within the caliper for each member of the treated group).[91] Matching can also be **with replacement** (comparator group members can match with multiple target group members) or **without replacement** (each comparator participant matches with only one target group member).[92] One to many matching with replacement typically results in the most precise

treatment effect estimates and approximates the weighting approaches discussed later; for such matching a balanced matching strategy can reduce potential bias from "one-sided" matching.[91]

After matching, outcomes in the two groups can generally be compared directly since matching leads to exchangeability on measured variables and therefore removes (measured) confounding.[93] The final treatment effect estimated after matching is the **treatment effect in the target group for which matches were found.** In settings with little similarity between groups, propensity score matching will highlight issues related to non-overlap (i.e. non-positivity) insofar as the proportion of the target that can be matched with comparator observations will be low and the estimated treatment effect may be a bad approximation of the ATT, potentially requiring a redefinition of the study population.[39,94,95] Additionally, removing the matches with the most dissimilar propensity scores runs the risk of creating more chance imbalances in the progressively smaller data set (similar to problems with small randomized samples);[96] this does not appear to be problematic in most pharmacoepidemiologic applications of the propensity score, however, given the large study sizes and types of variables used.[97]

Stratification: One alternative to matching that may lead to more precise results at the cost of additional assumptions is stratification (or, as it was referred by Rosenbaum and Rubin in 1983, subclassification) by the propensity score.[98] Just as one can reduce confounding by age by estimating treatment effects within strata of age, one can estimate treatment effects within strata of the propensity score. The narrower the strata, the less potential for residual confounding from within-stratum differences. In addition to the resulting stratum-specific treatment effects, a variety of methods (some assuming uniform treatment effects and some not, such as weights)[18] can be used to combine the results into a summary estimate**.**

Based on Cochran's work with linear confounders, [99] Rosenbaum and Rubin suggested that five strata based on propensity score quintiles would likely suffice to remove most bias in a binary treatment effect (assuming the outcome is a monotone function of the propensity score); of course, some bias is likely to remain.[8,100] While this is true with roughly equal numbers of treated and comparator patients, if the strata are derived from the overall propensity score distribution and treatment is rare, the average propensity score will be small and information will be concentrated in the higher propensity score strata where there are more treated individuals. This can lead to considerable residual confounding because of poorer within-strata balance in the low propensity score strata. In such cases, it is preferable to use **fine stratification** where a large

number of strata are formed based on the propensity score distribution in the treated; if we use the finest possible strata of the propensity score in the treated, we effectively perform a one to many matched analysis.[101] In general, regions of non-overlap should be excluded before stratification to reduce the potential for residual confounding (and the covariate distributions in those non-overlap region described), and balance within strata should be checked.

Modeling: The final method Rosenbaum and Rubin discussed in 1983 was modeling, specifically including the propensity score alongside treatment in a linear regression model. If the association between propensity score and the outcome is modeled adequately, this approach will estimate the **propensity-score conditional** treatment effect (unless g-computation is performed after adjustment).[102] Specifying the functional form correctly can be difficult, however, as the propensity score is a composite of many variables with their own effects on the outcome. This approach is therefore "doubly **un**-robust" in the sense that it requires correctly specified propensity score **and** outcome models. Bias may also arise when the variance of the propensity score estimating function differs between the treated and comparator groups.[103] Additionally, this method is one of the few propensity-score based analytic methods where the extent to which covariates were successfully balanced between treated and comparator groups is difficult to investigate and impossible to demonstrate empirically. Modeling also generally assumes uniform treatment effects. However, propensity score modeling can be combined with propensity score stratification, matching, or weighting, and researchers can reduce residual confounding from measured variables in the score by including them in the multivariable regression.[9]

Weighting: Weighting by the propensity score can be used to create a variety of exchangeable treated and comparator pseudo-populations with balanced distributions of measured covariates. Just as surveys can up- or down-weight the responses of specific groups to obtain estimates for a population of interest,[104] one can up- and down-weight treated and comparator observations to resemble some target population (and each other) using the propensity score.[105,106] The most common target populations are the total population (inverse probability of treatment weighting), the treated population (ATT, odds, or standardized mortality ratio weighting),[107] the treated population that would have been identified in a 1:1 matched analysis without replacement (match weighting),[108] or the population with an emphasis on the region of overlap (overlap weighting; this particular population can be difficult to articulate).[109]

Weights can also be stabilized to make the weighted sample size equal the unweighted sample size, improving the precision of inverse probability of treatment weighted estimators.[110]

Despite the versatility of weighted analyses, there are many potential pitfalls. First, not all statistical software is readily suited to weighted data when it comes to producing point estimates.[111] Another major concern with weighting is how researchers should deal with the problem of extremely large weights. When treated individuals have low propensity scores or untreated individuals have high propensity scores, they may receive large weights, particularly when estimating the population average treatment effect. These people, sometimes referred to as those treated **contrary to prediction**, can have a large influence on results and add considerable variance to treatment effect estimates. If they have unmeasured compelling indications or contraindications, or incorrectly measured treatment, they may cause bias as well. As a result, investigators should be careful to specify before starting the analysis whether they plan to truncate or trim (symmetrically or asymmetrically) past a certain weight (or propensity score – separate for treated and untreated) cutoff, and should decide on the final weighting model before examining outcomes.[60,112,94]

Doubly Robust Methods: Propensity scores can also be used as a component of **doubly robust** estimators that specify both outcome and treatment models, yielding an unbiased estimate if at least one of the two models is correctly specified.[86] Doubly robust estimates are typically less precise than those from outcome models but more precise than weighted estimates. While this option is appealing, it is unclear whether (or when) these estimators perform worse than the alternatives when both models are incorrect.[33] A newer form of doubly robust estimation, targeted maximum likelihood estimation (TMLE), estimates an outcome model (typically via SuperLearner, an ensemble machine learning method), then leverages a treatment assignment model to "target" the parameter of interest-the treatment effect-and reduce confounding.[113,114] The treatment effect estimates from TMLE will always be less biased than estimates separately applying the outcome or treatment assignment models.

Implementation Trends: Based on our literature review, the proportion of applied papers performing matched analyses increased over time (**Figure 4).** While half of the sampled studies in 2004 and 2009 analyzed a propensity score matched dataset, three-quarters of studies in 2014 and seven-eighths of studies in 2019 used a matched approach, with a corresponding drop in the prevalence of stratified and modeled analyses. That said, our search method may have

underestimated the use of weighting to some extent, as papers describing weighted analyses may not use the term "propensity score."

Estimating Variance: Up to now, we have focused on using propensity scores to obtain point estimates in the study sample. This is only half the inferential problem as variance estimation is also critical. Propensity score methods have several notable features that affect variance estimation, and researchers should be sure to use a statistically sound method for estimating standard errors. Variance and confidence intervals should also be estimated using methods that take into account any machine learning done when selecting variables for the propensity score.[115]

Despite the prevalence of propensity score matching in research, obtaining appropriate standard errors in matched studies is not straightforward. The standard solution for identifying standard errors with limited assumptions, the non-parametric bootstrap, can yield overly narrow confidence intervals when matching with replacement (as multiple copies of an individual in the bootstrap will all match to the same individual).[116] Simpler approaches to analyzing matched data can also lead to incorrect estimates of the standard error, particularly in the setting of one-to-many-matching with variable matching ratios.[93,117] Fortunately, work has recently been done to derive statistically sound estimators of standard errors after matching for continuous outcomes, survival outcomes, and time-to-event outcomes.[118-120] Only 25 of the 202 matched studies in our literature review made any mention of incorporating the matched nature of the data when estimating standard errors; hopefully, these new methods for variance estimation will be adopted by the wider community applying propensity score matching.

When using other analytic approaches, it is worth noting that some statistical software packages take weighting into account in estimating point estimates but not standard errors, yielding too-small standard errors and overly narrow confidence intervals. Nonparametric methods like the sandwich estimator (now included in SAS and many R packages) or the bootstrap (which typically has slightly more accurate confidence intervals for weighted and stratified estimates) are sometimes required to achieve appropriate confidence interval coverage.[121,122]

REPORTING RESULTS OF PROPENSITY SCORE ANALYSES

Here we provide suggestions on presentation and describe tools available to researchers for reporting and interpreting the results of studies using propensity scores. We do not intend these suggestions as a rigid checklist on what is required for scientific manuscripts, but rather as a guide to information that will help readers evaluate propensity score analyses.

Implementation: One useful principle for presenting analytic methods used in a study is providing enough detail to allow readers to repeat the study themselves.[123,124] This detail includes the variables included in the propensity score; how those variables were chosen and measured; how the propensity score was estimated; how missing and misclassified data on covariates, treatment, and the outcome were handled; whether and how the propensity score estimation was iterated to improve covariate balance based on some diagnostic; and, once the propensity score was estimated, how it was used to estimate the treatment effect (including details like caliper width and whether matching allowed for replacement in a matched study). The extent of loss to follow-up and administrative censoring should be described, as should the methods used to account for potential selection bias from these processes. When few observations are affected, bounds-based analyses can be used to explore the potential impact of restricting to observed individuals.[125] Censoring that affects many observations can be addressed using methods similar to those used to account for bias when conditioning on treatment continuation. If nothing has been done, as we frequently saw in our review, a clear reasoning behind this decision should be provided by the authors.

Because unmeasured confounding is one of the largest concerns in the non-experimental studies where propensity scores are generally used (and because the propensity score does not reduce unmeasured confounding), informing readers of key missing variables and how they may affect the final results is good practice. Researchers should also be clear about the treatment effect they are estimating with their analysis to ensure readers can understand to whom it applies and to what other populations the inference could extend. Presenting the results of pre-specified sensitivity analyses with differing propensity score models and implementation approaches can also help readers understand the effect of the specific implementation structure chosen.

Imbalance: It is important to communicate how much imbalance existed in study characteristics in the crude and, if possible, in the final analytic groups.[126] The "typical" **Table 1** includes group size, the choice of risk factors considered, and the amount of imbalance of these

factors between treatment groups at baseline, often including a metric like the standardized mean difference that shows the degree of difference between treatment groups. When combined with clear statements about the causal effect being estimated, a good **Table 1** helps readers to assess whether a given study is answering a question that is relevant to them in a study population they care about (or a study population similar to one they care about), whether the most important risk factors for the outcome have been measured, and how different the two groups were before and after propensity score implementation.

If matching, stratification, or weighting is being used, columns describing the groups after propensity score adjustment can illustrate the final covariate balance or the lack thereof, overall and within strata if applicable. These and other balance diagnostics help readers understand the degree to which the propensity score has established exchangeability on measured covariates. While SAMDs are imperfect, they are often used to examine the imbalance between treatment groups, with the aim of getting SAMDs as close to 0 as possible (with an often arbitrary cutoff for poor performance at 0.1).[76,82] In the example **Table 1,** we see that applying the SMR weights reduced the SAMDs and led to balance in all covariates presented, suggesting good performance of the propensity score. In matched or trimmed analyses, providing both crude and matched or post-trimming statistics in **Table 1** is useful for readers interested in the effect trimming or matching had on group composition. It is especially helpful to know the number, and covariate distributions of individuals in each treatment group that were excluded from the analysis due to their propensity score or a failure to find a match, as they may be important in interpreting results and defining future study populations.

From 2004 to 2019, the proportion of studies assessing group balance after some manner of confounder adjustment has increased (see **Figure 5),** with a rise in the use of both SMDs and P-values. The increase in the use of matching has likely facilitated the reporting of these balance statistics. However, it is concerning that many of the balance diagnostics rely on p-values, an inappropriate metric, as balance assessment does not involve inference about a larger population and because p-values are study-size-dependent.[127]

Population Overlap: Supplementing **Table 1** with a density plot or histogram of the propensity score distribution by treatment group (**Figure 6)** can give substantial insight into the prevalence of the treatment as well as the amount of overlap (and treatment equipoise) between treatment groups. A plot of the preference score (a transformation of the propensity score) can

help assess overlap independent of the overall prevalence of treatment.[39] While they do not describe the performance of the propensity score, the C-statistic and other measures of model discrimination can also help describe the degree of overlap between treatment groups for the variables in the propensity score model. The lower the C-statistic, the more similar the groups, and the greater the overlap; a high (>0.8) C-statistic raises concerns about positivity and equipoise between treatment groups.[94] Critically, if instruments are being included in the model, they will decrease overlap of compared groups even if the groups are perfectly balanced on risk factors for the outcome, leading to loss of precision and bias amplification as described above.[94,128,129]

Other Items to Report: As with most studies, providing crude estimates of the treatment effect (in the total population for matched studies) alongside the (propensity-score) adjusted effect estimates can contextualize the overall direction of the measured confounding, which can be compared with expectations. If time-to-event analyses are being conducted, including crude survival curves alongside weighted or matched versions is helpful, particularly in weighted analyses where observations with large weights may manifest as large vertical jumps in the curves. If stabilized weights were used, reporting mean weights (which should be close to 1) and extreme weights can provide a useful diagnostic for potential problems in the propensity score. Heavily weighted observations may signal insufficient equipoise or problems with the coding of covariates.[130]

EVOLVING USAGE OF PROPENSITY SCORES

This overview has focused on using propensity scores to assess the effect of a treatment decision between two alternatives at one fixed point in time, as originally envisioned by Rosenbaum and Rubin. However, a great deal of work has considered alternative settings.

For example, the propensity score can be helpful when there are multiple consecutive treatment decisions. The conditional probability of treatment (i.e. the propensity score) can be used to fit marginal structural models for time-varying exposures. These methods have advanced considerably over the past 20 years-including the settings of time-varying instrumental variable analysis and possible interference between study units.[105,131,132] Marginal structural models are particularly valuable for estimating alternative, more complex causal effects such as the effect of dynamic treatment regimens (i.e. treatment regimens where exposure depends on time-varying measurements and factors, like treating HIV patients when CD4 falls below a given level).[133,134] Assuming all variables that influence decisions to swap or change therapy are available in the data, these methods can estimate important treatment effects without bias. Similarly, time-conditional propensity scores have been proposed to estimate the effect of switching to novel therapies among prevalent users of older treatments and augmenting older treatment regimens with new therapeutic agents.[135]

The propensity score can also be used beyond the case of a binary treatment decision. A similar framework can be applied when treatment has three or more categories.[74,136] Matching, weights, and trimming approaches are currently being developed that take into account the issues unique to this context.[95,137,138] Work has also been done to extend propensity score methods to continuous exposures and treatments, sometimes referred to as the generalized propensity score.[43,139,140]

CONCLUSIONS

With the proliferation of real world data sources and statistical software allowing easy matched, stratified, and weighted analyses and greater reliance on routinely collected data for research, it seems likely that the coming years will see continued use and refinement of propensity score methods to generate real world evidence on comparative effectiveness and safety. Focusing on treatment assignment during study design and analysis, as suggested by Rosenbaum and Rubin in 1983, has yielded insights ranging from the best choice of comparator to what should be done in the presence of limited covariate positivity. Summarizing many potential confounders in a single statistic such as the propensity score, allows simplified presentation and easy assessment of control for measured confounders, and further, allows treatment effect estimation even for rare outcomes. Propensity scores have been and will continue to be valuable tools for non-experimental research.

References

1. Greenland S, Morgenstern H. Confounding in health research. *Annual review of public health.* 2001;22:189-212.
2. Rothman KJ, Greenland S, Lash TL. *Modern epidemiology.* Lippincott Williams & Wilkins; 2008.
3. FDA. Real World Evidence. 2018; https://www.fda.gov/science-research/science-and-research-special-topics/real-world-evidence. Accessed 8/6, 2019.
4. Eichler H-G, Bloechl-Daum B, Broich K, et al. Data Rich, Information Poor: Can We Use Electronic Health Records to Create a Learning Healthcare System for Pharmaceuticals? *Clinical Pharmacology & Therapeutics.* 2019;105(4):912-922.
5. Cornfield J, Haenszel W, Hammond EC, Lilienfeld AM, Shimkin MB, Wynder EL. Smoking and lung cancer: recent evidence and a discussion of some questions. *Journal of the National Cancer Institute.* 1959;22(1):173-203.
6. Lee CH, Yoon HJ. Medical big data: promise and challenges. *Kidney research and clinical practice.* 2017;36(1):3-11.
7. Slobogean GP, Giannoudis PV, Frihagen F, Forte ML, Morshed S, Bhandari M. Bigger Data, Bigger Problems. *Journal of orthopaedic trauma.* 2015;29 Suppl 12:S43-46.
8. ROSENBAUM PR, RUBIN DB. The central role of the propensity score in observational studies for causal effects. *Biometrika.* 1983;70(1):41-55.
9. D'Agostino RB, Jr. Propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. *Statistics in medicine.* 1998;17(19):2265-2281.
10. Miettinen OS. Stratification by a multivariate confounder score. *American journal of epidemiology.* 1976;104(6):609-620.
11. Peters CC. A method of matching groups for experiment with no loss of population. *The Journal of Educational Research.* 1941;34(8):606-612.
12. WA B. A technique for studying the effects of a television broadcast. *Applied Stat.* 1956;5:195-202.
13. Pike M, Anderson J, Day N. Some insights into Miettinen's multivariate confounder score approach to case-control study analysis. *Journal of Epidemiology & Community Health.* 1979;33(1):104-106.
14. Greenland S. Confounder Summary Score. In: Gail MHB, Jacques, ed. *Encyclopedia of Epidemiologic Methods*. Sussex, England: John Wiley and Sons, Ltd.; 2015:1116-1118.
15. Hansen BB. The prognostic analogue of the propensity score. *Biometrika.* 2008;95(2):481-488.
16. Glynn RJ, Gagne JJ, Schneeweiss S. Role of disease risk scores in comparative effectiveness research with emerging therapies. *Pharmacoepidemiology and drug safety.* 2012;21:138-147.
17. 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(4):317-324.
18. Austin PC. An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies. *Multivariate behavioral research.* 2011;46(3):399-424.
19. Brookhart MA, Wyss R, Layton JB, Stürmer T. Propensity score methods for confounding control in nonexperimental research. *Circulation Cardiovascular quality and outcomes.* 2013;6(5):604-611.
20. Lee J, Little TD. A practical guide to propensity score analysis for applied clinical research. *Behaviour research and therapy.* 2017;98:76-90.
21. Sturmer T, Joshi M, Glynn RJ, Avorn J, Rothman KJ, Schneeweiss S. A review of the application of propensity score methods yielded increasing use, advantages in specific settings, but not

substantially different estimates compared with conventional multivariable methods. *Journal of clinical epidemiology.* 2006;59(5):437-447.
22. Little RJ, Rubin DB. Causal effects in clinical and epidemiological studies via potential outcomes: concepts and analytical approaches. *Annual review of public health.* 2000;21:121-145.
23. Rosenbaum PR. Propensity Score. In: Gail MHB, Jacques, ed. *Encyclopedia of Epidemiologic Methods*. West Sussex, England: John Wiley and Sons, Ltd.; 2015:4267-4272.
24. Maldonado G, Greenland S. Estimating causal effects. *International Journal of Epidemiology.* 2002;31(2):422-429.
25. Cole SR, Frangakis CE. The consistency statement in causal inference: a definition or an assumption? *Epidemiology (Cambridge, Mass).* 2009;20(1):3-5.
26. Hernán MA, Robins JM. Estimating causal effects from epidemiological data. *Journal of epidemiology and community health.* 2006;60(7):578-586.
27. Greenland S, Robins JM. Identifiability, exchangeability, and epidemiological confounding. *International journal of epidemiology.* 1986;15(3):413-419.
28. Saint-Mont U. Randomization does not help much, comparability does. *PloS one.* 2015;10(7):e0132102.
29. Greenland S, Robins JM. Identifiability, exchangeability and confounding revisited. *Epidemiologic Perspectives & Innovations.* 2009;6(1):4.
30. McKinlay SM. Pair-Matching -- A Reappraisal of a Popular Technique. *Biometrics.* 1977;33(4):725-735.
31. Westreich D, Cole SR. Invited commentary: positivity in practice. *American journal of epidemiology.* 2010;171(6):674-677; discussion 678-681.
32. Drake C. Effects of misspecification of the propensity score on estimators of treatment effect. *Biometrics.* 1993:1231-1236.
33. Funk MJ, Westreich D, Wiesen C, Stürmer T, Brookhart MA, Davidian M. Doubly Robust Estimation of Causal Effects. *American journal of epidemiology.* 2011;173(7):761-767.
34. Braitman LE, Rosenbaum PR. Rare outcomes, common treatments: analytic strategies using propensity scores. *Annals of internal medicine.* 2002;137(8):693-695.
35. Cepeda MS. The use of propensity scores in pharmacoepidemiologic research. *Pharmacoepidemiology and drug safety.* 2000;9(2):103-104.
36. Cepeda MS, Boston R, Farrar JT, Strom BL. Comparison of logistic regression versus propensity score when the number of events is low and there are multiple confounders. *American journal of epidemiology.* 2003;158(3):280-287.
37. Austin PC. Propensity-score matching in the cardiovascular surgery literature from 2004 to 2006: a systematic review and suggestions for improvement. *The Journal of thoracic and cardiovascular surgery.* 2007;134(5):1128-1135. e1123.
38. 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(12):841-853.
39. Walker A, Patrick, Lauer M, et al. A tool for assessing the feasibility of comparative effectiveness research. *Comparative Effectiveness Research.* 2013;2013:11.
40. Rubin DB. Estimating causal effects from large data sets using propensity scores. *Annals of internal medicine.* 1997;127(8_Part_2):757-763.
41. Glynn RJ, Schneeweiss S, Stürmer T. Indications for propensity scores and review of their use in pharmacoepidemiology. *Basic & clinical pharmacology & toxicology.* 2006;98(3):253-259.
42. Sturmer T, Rothman KJ, Glynn RJ. Insights into different results from different causal contrasts in the presence of effect-measure modification. *Pharmacoepidemiology and drug safety.* 2006;15(10):698-709.

43. Hirano K, Imbens GW. The propensity score with continuous treatments. *Applied Bayesian modeling and causal inference from incomplete-data perspectives.* 2004;226164:73-84.
44. Glynn RJ. Use of Propensity Scores To Design Observational Comparative Effectiveness Studies. *Journal of the National Cancer Institute.* 2017;109(8).
45. Glynn RJ, Knight EL, Levin R, Avorn J. Paradoxical relations of drug treatment with mortality in older persons. *Epidemiology (Cambridge, Mass).* 2001;12(6):682-689.
46. Eurich DT, Marrie TJ, Johnstone J, Majumdar SR. Mortality reduction with influenza vaccine in patients with pneumonia outside "flu" season: pleiotropic benefits or residual confounding? *American journal of respiratory and critical care medicine.* 2008;178(5):527-533.
47. Johnson ES BB, Briesacher BA, Fleming NS, Gerhard T, Kornegay CJ, Nourjah P, Sauer B, Schumock GT, Sedrakyan A, Stürmer T, West SL, Schneeweiss S. *The Incident User Design in Comparative Effectiveness Research.* Rockville, MD: Agency for Healthcare Research and Quality;2012.
48. Kramer MS, Lane DA, Hutchinson TA. Analgesic use, blood dyscrasias, and case-control pharmacoepidemiology: A critique of the International Agranulocytosis and Aplastic Anemia Study. *Journal of Chronic Diseases.* 1987;40(12):1073-1081.
49. Greenland S, Robins JM. Confounding and misclassification. *American journal of epidemiology.* 1985;122(3):495-506.
50. Edwards JK, Htoo PT, Stürmer T. Counterpoint: Keeping the Demons at Bay When Handling Time-Varying Exposures—Beyond Avoiding Immortal Person-Time. *American journal of epidemiology.* 2019;188(6):1016-1022.
51. Ray WA. Evaluating medication effects outside of clinical trials: new-user designs. *American journal of epidemiology.* 2003;158(9):915-920.
52. Lund JL, Richardson DB, Stürmer T. The active comparator, new user study design in pharmacoepidemiology: historical foundations and contemporary application. *Current epidemiology reports.* 2015;2(4):221-228.
53. Hernán MA, Robins JM. Using Big Data to Emulate a Target Trial When a Randomized Trial Is Not Available. *American journal of epidemiology.* 2016;183(8):758-764.
54. Lévesque LE, Hanley JA, Kezouh A, Suissa S. Problem of immortal time bias in cohort studies: example using statins for preventing progression of diabetes. *BMJ.* 2010;340:b5087.
55. Hernán MA, Alonso A, Logan R, et al. Observational studies analyzed like randomized experiments: an application to postmenopausal hormone therapy and coronary heart disease. *Epidemiology (Cambridge, Mass).* 2008;19(6):766-779.
56. Sheiner LB, Rubin DB. Intention-to-treat analysis and the goals of clinical trials. *Clinical Pharmacology & Therapeutics.* 1995;57(1):6-15.
57. Hernán MA, Robins JM. Per-Protocol Analyses of Pragmatic Trials. *New England Journal of Medicine.* 2017;377(14):1391-1398.
58. Hernán MA, Hernández-Díaz S. Beyond the intention-to-treat in comparative effectiveness research. *Clinical trials (London, England).* 2012;9(1):48-55.
59. Crump RK, Hotz VJ, Imbens GW, Mitnik OA. Dealing with limited overlap in estimation of average treatment effects. *Biometrika.* 2009;96(1):187-199.
60. 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. *American journal of epidemiology.* 2010;172(7):843-854.
61. Brookhart MA, Schneeweiss S, Rothman KJ, Glynn RJ, Avorn J, Stürmer T. Variable selection for propensity score models. *American journal of epidemiology.* 2006;163(12):1149-1156.
62. Schuster T, Lowe WK, Platt RW. Propensity score model overfitting led to inflated variance of estimated odds ratios. *Journal of clinical epidemiology.* 2016;80:97-106.

63. Pearl J. Causal diagrams for empirical research. *Biometrika.* 1995;82(4):669-688.
64. Pearl J. Invited commentary: understanding bias amplification. *American journal of epidemiology.* 2011;174(11):1223-1227; discussion pg 1228-1229.
65. Patrick AR, Schneeweiss S, Brookhart MA, et al. The implications of propensity score variable selection strategies in pharmacoepidemiology: an empirical illustration. *Pharmacoepidemiology and drug safety.* 2011;20(6):551-559.
66. Myers JA, Rassen JA, Gagne JJ, et al. Effects of adjusting for instrumental variables on bias and precision of effect estimates. *American journal of epidemiology.* 2011;174(11):1213-1222.
67. Cole SR, Platt RW, Schisterman EF, et al. Illustrating bias due to conditioning on a collider. *Int J Epidemiol.* 2010;39(2):417-420.
68. Liu W, Brookhart MA, Schneeweiss S, Mi X, Setoguchi S. Implications of M bias in epidemiologic studies: a simulation study. *American journal of epidemiology.* 2012;176(10):938-948.
69. Greenland S, Pearl J, Robins JM. Causal diagrams for epidemiologic research. *Epidemiology (Cambridge, Mass).* 1999;10(1):37-48.
70. Pearl J, Paz A. Confounding equivalence in causal inference. *Journal of Causal Inference J Causal Infer.* 2014;2(1):75-93.
71. VanderWeele TJ, Shpitser I. A new criterion for confounder selection. *Biometrics.* 2011;67(4):1406-1413.
72. 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 (Cambridge, Mass).* 2009;20(4):512-522.
73. Westreich D, Lessler J, Funk MJ. Propensity score estimation: neural networks, support vector machines, decision trees (CART), and meta-classifiers as alternatives to logistic regression. *Journal of clinical epidemiology.* 2010;63(8):826-833.
74. 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. *Statistics in medicine.* 2013;32(19):3388-3414.
75. Fong C, Hazlett C, Imai K. Covariate balancing propensity score for a continuous treatment: Application to the efficacy of political advertisements. *The Annals of Applied Statistics.* 2018;12(1):156-177.
76. Rubin DB. Using propensity scores to help design observational studies: application to the tobacco litigation. *Health Services and Outcomes Research Methodology.* 2001;2(3-4):169-188.
77. Rubin DB. The design versus the analysis of observational studies for causal effects: parallels with the design of randomized trials. *Statistics in medicine.* 2007;26(1):20-36.
78. Maislin G, Rubin D. Design of non-randomized medical device trials based on sub-classification using propensity score quintiles, topic contributed session on medical devices. Paper presented at: Proceedings of the Joint Statistical Meetings2010.
79. Low YS, Gallego B, Shah NH. Comparing high-dimensional confounder control methods for rapid cohort studies from electronic health records. *Journal of comparative effectiveness research.* 2016;5(2):179-192.
80. 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.
81. Lee BK, Lessler J, Stuart EA. Improving propensity score weighting using machine learning. *Statistics in medicine.* 2010;29(3):337-346.
82. Stuart EA, Lee BK, Leacy FP. Prognostic score-based balance measures can be a useful diagnostic for propensity score methods in comparative effectiveness research. *Journal of clinical epidemiology.* 2013;66(8 Suppl):S84-S90.e81.

83. Colson KE, Rudolph KE, Zimmerman SC, et al. Optimizing matching and analysis combinations for estimating causal effects. *Scientific reports.* 2016;6:23222.
84. Wyss R, Lunt M, Brookhart MA, Glynn RJ, Sturmer T. Reducing Bias Amplification in the Presence of Unmeasured Confounding Through Out-of-Sample Estimation Strategies for the Disease Risk Score. *Journal of causal inference.* 2014;2(2):131-146.
85. Rosenbaum PR. Model-based direct adjustment. *Journal of the American Statistical Association.* 1987;82(398):387-394.
86. Bang H, Robins JM. Doubly robust estimation in missing data and causal inference models. *Biometrics.* 2005;61(4):962-973.
87. Sturmer T, Rothman KJ, Glynn RJ. Insights into different results from different causal contrasts in the presence of effect-measure modification. *Pharmacoepidemiology and Drug Safety.* 2006;15(10):698-709.
88. Kurth T, Walker AM, Glynn RJ, et al. Results of multivariable logistic regression, propensity matching, propensity adjustment, and propensity-based weighting under conditions of nonuniform effect. *American journal of epidemiology.* 2006;163(3):262-270.
89. Austin PC. A comparison of 12 algorithms for matching on the propensity score. *Statistics in medicine.* 2014;33(6):1057-1069.
90. Lunt M. Selecting an Appropriate Caliper Can Be Essential for Achieving Good Balance With Propensity Score Matching. *American journal of epidemiology.* 2013;179(2):226-235.
91. Rassen JA, Shelat AA, Myers J, Glynn RJ, Rothman KJ, Schneeweiss S. One-to-many propensity score matching in cohort studies. *Pharmacoepidemiology and drug safety.* 2012;21 Suppl 2:69-80.
92. Stuart EA. Matching methods for causal inference: A review and a look forward. *Statistical science : a review journal of the Institute of Mathematical Statistics.* 2010;25(1):1-21.
93. Stuart EA. Developing practical recommendations for the use of propensity scores: discussion of 'A critical appraisal of propensity score matching in the medical literature between 1996 and 2003' by Peter Austin, Statistics in Medicine. *Statistics in medicine.* 2008;27(12):2062-2065; discussion 2066-2069.
94. Glynn RJ, Lunt M, Rothman KJ, Poole C, Schneeweiss S, Stürmer T. Comparison of alternative approaches to trim subjects in the tails of the propensity score distribution. *Pharmacoepidemiology and drug safety.* 2019;28(10):1290-1298.
95. Yoshida K, Solomon DH, Haneuse S, et al. A tool for empirical equipoise assessment in multigroup comparative effectiveness research. *Pharmacoepidemiology and drug safety.* 2019;28(7):934-941.
96. King G, Nielsen R. Why Propensity Scores Should Not Be Used for Matching. *Political Analysis.* 2019;27(4).
97. Ripollone JE, Huybrechts KF, Rothman KJ, Ferguson RE, Franklin JM. Implications of the Propensity Score Matching Paradox in Pharmacoepidemiology. *American journal of epidemiology.* 2018;187(9):1951-1961.
98. Wang Z. Propensity score methods to adjust for confounding in assessing treatment effects: Bias and precision. *The Internet Journal of Epidemiology.* 2009;7(2):x-x.
99. Cochran WG. The effectiveness of adjustment by subclassification in removing bias in observational studies. *Biometrics.* 1968:295-313.
100. 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(19):2937-2960.
101. 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 (Cambridge, Mass).* 2017;28(2):249-257.

102. Austin PC. The relative ability of different propensity score methods to balance measured covariates between treated and untreated subjects in observational studies. *Medical decision making : an international journal of the Society for Medical Decision Making.* 2009;29(6):661-677.
103. Hade EM, Lu B. Bias associated with using the estimated propensity score as a regression covariate. *Statistics in medicine.* 2014;33(1):74-87.
104. Pfeffermann D. The use of sampling weights for survey data analysis. *Statistical methods in medical research.* 1996;5(3):239-261.
105. Robins JM, Hernan MA, Brumback B. Marginal structural models and causal inference in epidemiology. *Epidemiology (Cambridge, Mass).* 2000;11(5):550-560.
106. Hirano K, Imbens GW. Estimation of causal effects using propensity score weighting: An application to data on right heart catheterization. *Health Services and Outcomes research methodology.* 2001;2(3-4):259-278.
107. Sato T, Matsuyama Y. Marginal structural models as a tool for standardization. *Epidemiology (Cambridge, Mass).* 2003;14(6):680-686.
108. Li L, Greene T. A weighting analogue to pair matching in propensity score analysis. *The international journal of biostatistics.* 2013;9(2):215-234.
109. Li F, Thomas LE, Li F. Addressing Extreme Propensity Scores via the Overlap Weights. *American journal of epidemiology.* 2019;188(1):250-257.
110. Xu S, Ross C, Raebel MA, Shetterly S, Blanchette C, Smith D. Use of stabilized inverse propensity scores as weights to directly estimate relative risk and its confidence intervals. *Value in health : the journal of the International Society for Pharmacoeconomics and Outcomes Research.* 2010;13(2):273-277.
111. 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(19):2937-2960.
112. Lee BK, Lessler J, Stuart EA. Weight trimming and propensity score weighting. *PLoS One.* 2011;6(3):e18174.
113. Van Der Laan MJ, Rubin D. Targeted maximum likelihood learning. *The international journal of biostatistics.* 2006;2(1).
114. van der Laan MJ, Gruber S. Collaborative double robust targeted maximum likelihood estimation. *The international journal of biostatistics.* 2010;6(1).
115. Dukes O, Vansteelandt S. How to obtain valid tests and confidence intervals after propensity score variable selection? *Statistical methods in medical research.* 2020;29(3):677-694.
116. Abadie A, Imbens GW. On the failure of the bootstrap for matching estimators. *Econometrica.* 2008;76(6):1537-1557.
117. Austin PC. Comparing paired vs non-paired statistical methods of analyses when making inferences about absolute risk reductions in propensity-score matched samples. *Statistics in medicine.* 2011;30(11):1292-1301.
118. Austin PC, Cafri G. Variance estimation when using propensity-score matching with replacement with survival or time-to-event outcomes. *Statistics in medicine.* 2020;39(11):1623-1640.
119. Abadie A, Imbens GW. Large sample properties of matching estimators for average treatment effects. *econometrica.* 2006;74(1):235-267.
120. Abadie A, Imbens GW. A martingale representation for matching estimators. *Journal of the American Statistical Association.* 2012;107(498):833-843.
121. Austin PC. Variance estimation when using inverse probability of treatment weighting (IPTW) with survival analysis. *Statistics in medicine.* 2016;35(30):5642-5655.
122. Williamson EJ, Morley R, Lucas A, Carpenter JR. Variance estimation for stratified propensity score estimators. *Statistics in medicine.* 2012;31(15):1617-1632.

123. Yao XI, Wang X, Speicher PJ, et al. Reporting and Guidelines in Propensity Score Analysis: A Systematic Review of Cancer and Cancer Surgical Studies. *Journal of the National Cancer Institute.* 2017;109(8).
124. Langan SM, Schmidt SA, Wing K, et al. The reporting of studies conducted using observational routinely collected health data statement for pharmacoepidemiology (RECORD-PE). *Bmj.* 2018;363:k3532.
125. Smith LH, VanderWeele TJ. Bounding Bias Due to Selection. *Epidemiology (Cambridge, Mass).* 2019;30(4):509-516.
126. Austin PC. Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples. *Statistics in medicine.* 2009;28(25):3083-3107.
127. Imai K, King G, Stuart EA. Misunderstandings between experimentalists and observationalists about causal inference. *Journal of the royal statistical society: series A (statistics in society).* 2008;171(2):481-502.
128. Austin PC, Grootendorst P, Anderson GM. A comparison of the ability of different propensity score models to balance measured variables between treated and untreated subjects: a Monte Carlo study. *Statistics in medicine.* 2007;26(4):734-753.
129. Westreich D, Cole SR, Funk MJ, Brookhart MA, Stürmer T. The role of the c-statistic in variable selection for propensity score models. *Pharmacoepidemiology and drug safety.* 2011;20(3):317-320.
130. Cole SR, Hernan MA. Constructing inverse probability weights for marginal structural models. *American journal of epidemiology.* 2008;168(6):656-664.
131. Tchetgen EJT, Michael H, Cui Y. Marginal Structural Models for Time-varying Endogenous Treatments: A Time-Varying Instrumental Variable Approach. *arXiv preprint arXiv:180905422.* 2018.
132. He J, Stephens-Shields A, Joffe M. Marginal structural models to estimate the effects of time-varying treatments on clustered outcomes in the presence of interference. *Statistical methods in medical research.* 2019;28(2):613-625.
133. Kang S, Lu W, Zhang J. ON ESTIMATION OF THE OPTIMAL TREATMENT REGIME WITH THE ADDITIVE HAZARDS MODEL. *Statistica Sinica.* 2018;28(3):1539-1560.
134. Neugebauer R, Fireman B, Roy JA, O'Connor PJ, Selby JV. Dynamic marginal structural modeling to evaluate the comparative effectiveness of more or less aggressive treatment intensification strategies in adults with type 2 diabetes. *Pharmacoepidemiology and drug safety.* 2012;21 Suppl 2:99-113.
135. Suissa S, Moodie EE, Dell'Aniello S. Prevalent new-user cohort designs for comparative drug effect studies by time-conditional propensity scores. *Pharmacoepidemiology and drug safety.* 2017;26(4):459-468.
136. Yang S, Imbens GW, Cui Z, Faries DE, Kadziola Z. Propensity score matching and subclassification in observational studies with multi-level treatments. *Biometrics.* 2016;72(4):1055-1065.
137. Yoshida K, Hernández-Díaz S, Solomon DH, et al. Matching Weights to Simultaneously Compare Three Treatment Groups: Comparison to Three-way Matching. *Epidemiology (Cambridge, Mass).* 2017;28(3):387-395.
138. Lopez MJ, Gutman R. Estimation of causal effects with multiple treatments: a review and new ideas. *Statistical Science.* 2017;32(3):432-454.
139. Austin PC. Assessing the performance of the generalized propensity score for estimating the effect of quantitative or continuous exposures on binary outcomes. *Statistics in medicine.* 2018;37(11):1874-1894.

140. Zhang Z, Zhou J, Cao W, Zhang J. Causal inference with a quantitative exposure. *Statistical methods in medical research.* 2016;25(1):315-335.

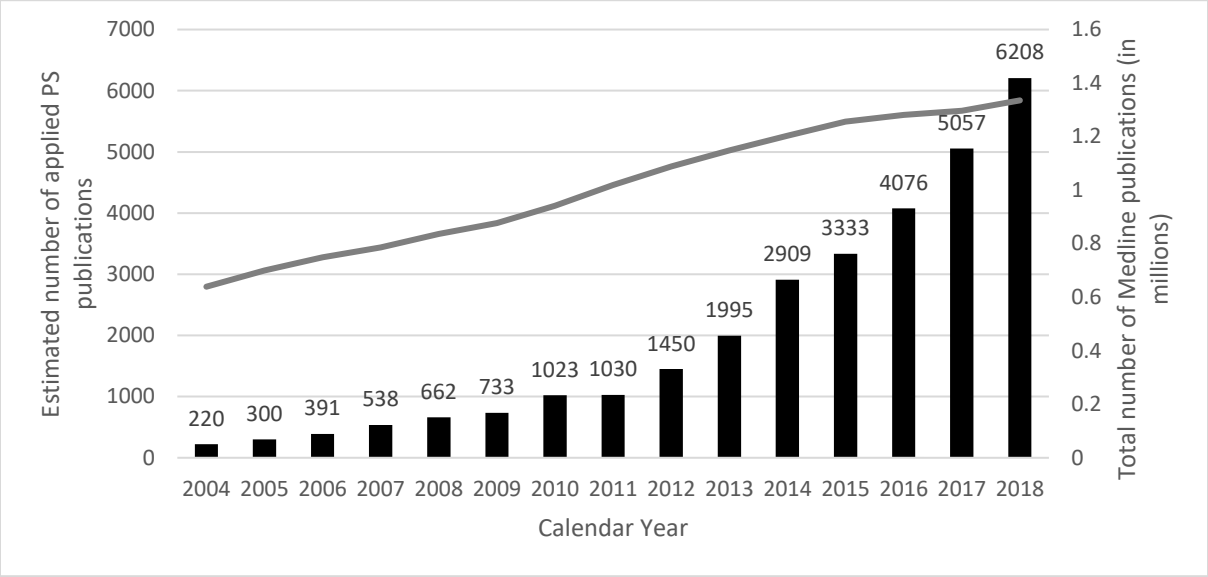

Figure 1: Black bars represent the number of published papers estimated to have applied propensity score methods to research questions from 2004 to 2018. The gray line denotes the total number of publications indexed by Medline in that calendar year in millions.

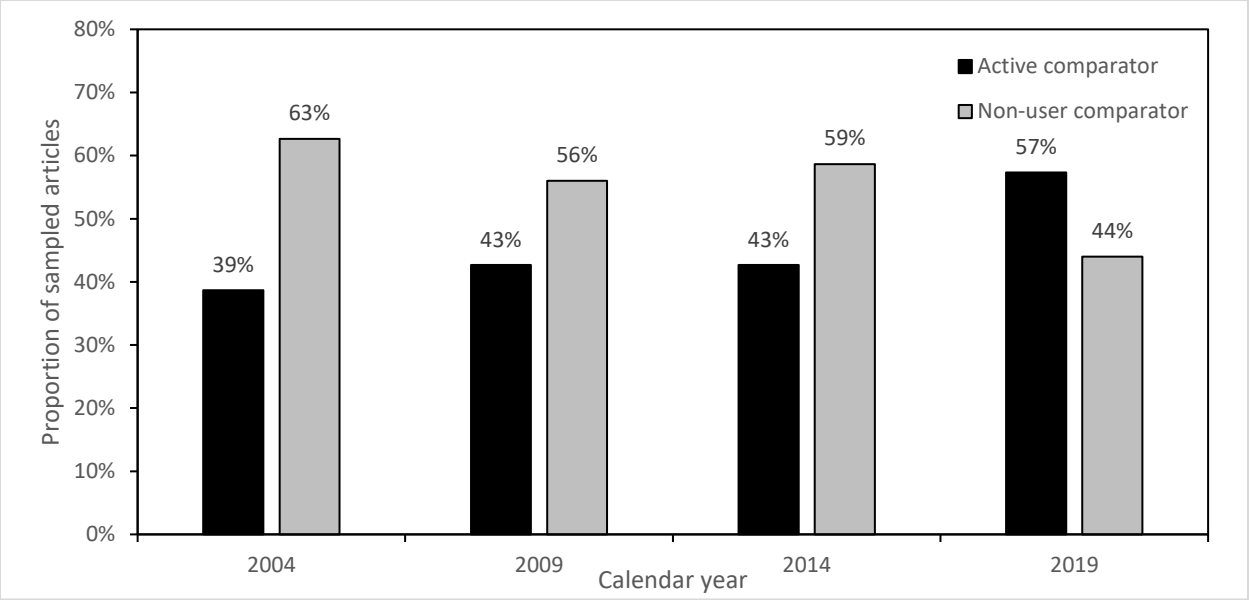

Figure 2: Proportions of sampled articles using active comparators (black) or non-user comparators (gray). Proportions could sum to more than 100% in a given calendar period when studies had more than one comparator group.

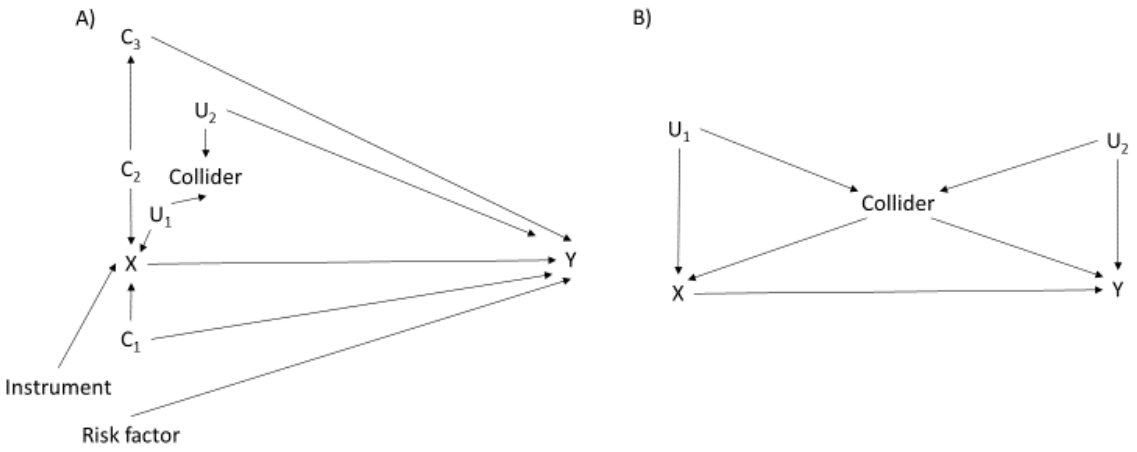

Figure 3: Directed acyclic graphs illustrating different types of variables that can be considered for inclusion in a propensity score model. An arrow is draw from one node to another if intervening on the first causes a change in the second for some portion of the population. C variables in panel A are measured variables on confounding paths. U variables in both panels are unmeasured variables. Other variable types are labeled with descriptive names.

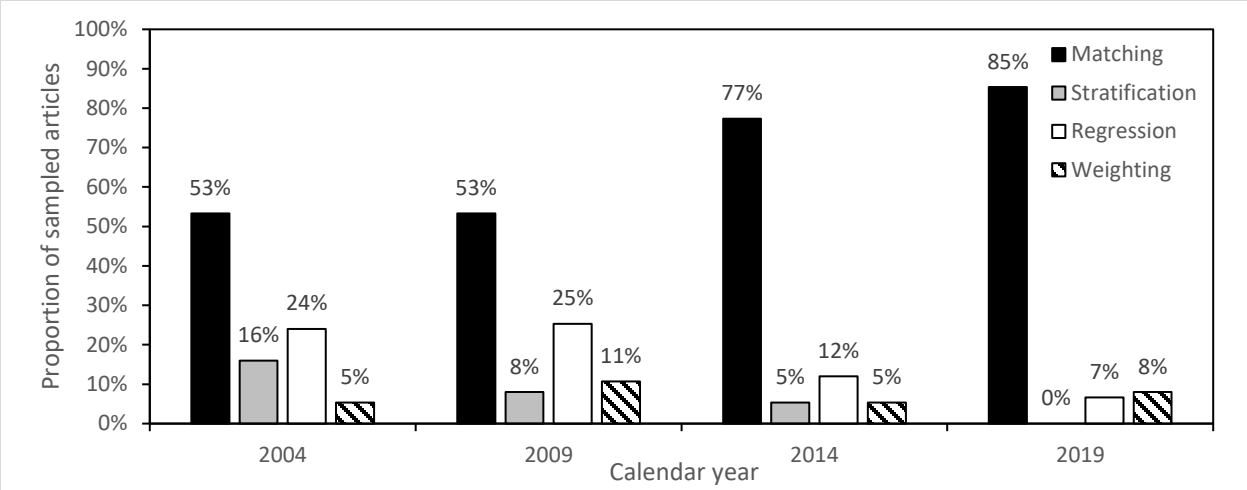

Figure 4: Proportions of sampled articles applying each type of method in the four calendar years we assessed. The solid black bar shows matched analyses, the gray bar shows stratified analyses, the white bar shows regression analyses, and the bar with diagonal lines shows weighted analyses. Proportions could sum to more than 100% in a given calendar period when multiple methods were used, and less than 100% if another method (typically invalid) was used.

Table 1: Example Table with Distributions of Covariates and Standardized Mean Differences (SMDs) Before and After Weighting Treatment 2 Patients to Resemble Treatment 1 Patients with Standardized Mortality Ratio (SMR) Weights

Covariate	Treatment 1 N=10,717	Treatment 2 N=74,8910	Crude SMDs	Treatment 2 (SMR weighted) N=10,717	SMR weighted SMDs
Male	5,316 (50%)	32,430 (43%)	0.127	5,206 (49%)	0.005
Hypertension	10,522 (98%)	73,340 (98%)	0.018	10,523 (98%)	-0.002
Diabetes	3,334 (31%)	24,329 (33%)	-0.029	3,352 (31%)	-0.007
Coronary Artery Disease	5,178 (48%)	37,389 (49%)	-0.032	5,203 (49%)	-0.010
Congestive Heart Failure	3,839 (36%)	30,404 (41%)	-0.098	3,861 (36%)	-0.008

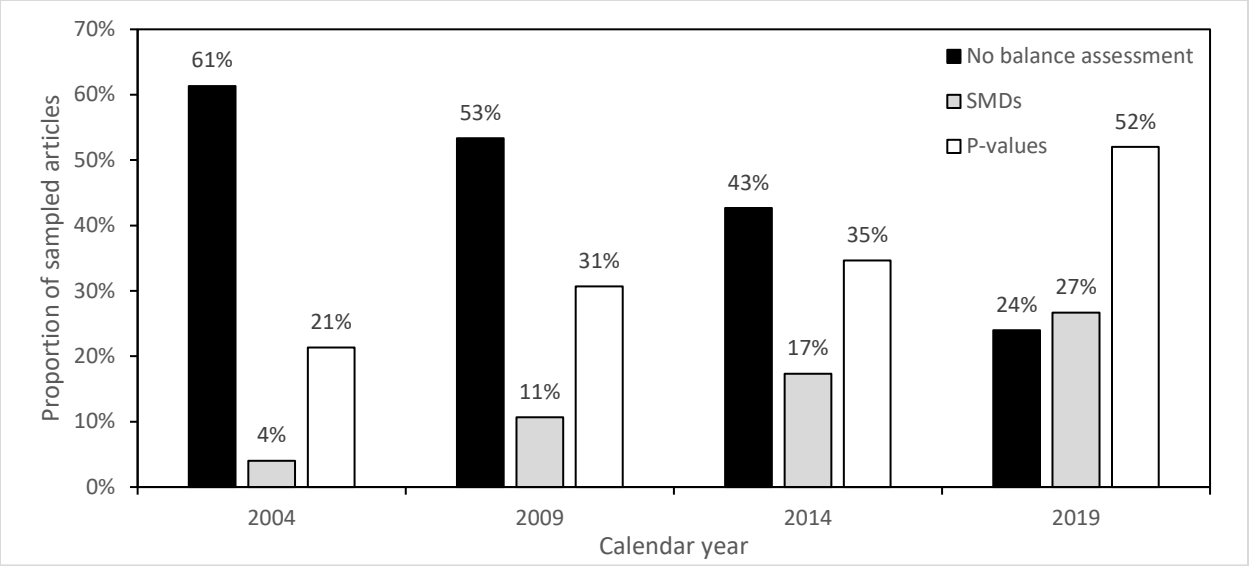

Figure 5: Calendar time trends in balance assessment. The solid black bar is the proportion of papers with no mention of balance assessment. The gray bar is the use of standardized mean differences for balance assessment, and the white bar is the use of P-values.

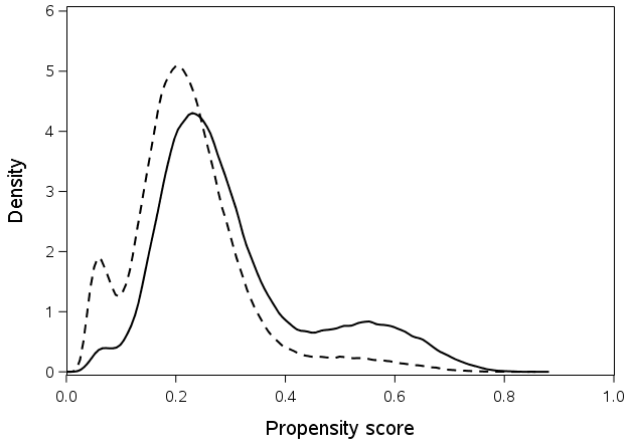

Figure 6: A sample density plot of propensity scores. The solid line is the treated population, while the dashed line is the comparator. The scores show fairly good overlap, though there might be some issues in the tails.

Appendix A: Literature Search Protocol

Final search and pull of articles took place on **October 4th, 2019.**

To assess overall trends in use:

- We searched **paper text (including abstract text)** from calendar years 2004 to 2018 for the term **"propensity score"** in the main text
- This search was done in **Pubmed, Embase, Web of Science,** and **Scopus (abstracts were excluded)**
- We then recorded the total number of articles after deduplication in each calendar year from 2004 to 2018

To correct these trends to account non-applied papers:

- We abstracted **75** articles at random from each of four years: **2004, 2009, 2014,** and **2019**
- If one of the 75 random articles from the year was not an application of propensity score methods to a substantive question, we recorded this and draw a replacement
    - "Non-substantive" articles included reviews, developmental methods papers, and commentaries
- The rejection rate in each of the subsampled years, and a weighted average of the prior and next rejection rate for the non-subsampled years, was then used to correct the raw results of the database search to exclude articles that were not applications of propensity score methods to a substantive question

To assess trends in implementation:

- Reviewers assessed the following components of the articles:
    - **Comparator** (active comparator vs non-user comparator)
    - **PS implementation** (matching 1:1 vs 1:n vs 1:all (subclass nearest neighbor vs calipers (with caliper values)) vs stratification (subclass fine stratification) vs regression covariate vs weighting (subclass IPTW vs SMR))
    - **Post-implementation covariate balance assessment** (SMDs, P values, both, neither, other; if provided with a cutoff note cutoff)

RESPONSE TO ISPE MEMBER COMMENTS

October 13th, 2020

Dear Dr. Davis and members of the ISPE Public Policy Committee,

We have revised our manuscript reviewing propensity scores in response to comments left by ISPE members during the open comment period, as well as comments from our reviewers at Statistics in Medicine on two revisions at the journal.

We have made several major additions to the manuscript, including two full new sections and a new subsection. We have also expanded our literature review and made changes to the main text itself in response to feedback from the membership; responses to their specific comments are detailed below.

In our opinion, the manuscript is now ideal for publication at Statistics in Medicine.

Regards,

Michael Webster-Clark

On behalf of my co-authors: Til Stürmer, Tiansheng Wang, Kenneth Man, Danica Marinac-Dabic, Kenneth J. Rothman, Alan R. Ellis, Mugdha Gokhale, Mark Lunt, Cynthia Girman, and Robert Glynn.

Comment. *The article seems to focus quite strongly on PS use and theory in pharmacoepidemiology (or at least disease treatment/intervention), with some aspects discussed in a quite specific way realted to this subject area, and it may be useful to reflect this in the title. Alternatively, or additionally, it may be useful to provide a broader setting for PS as described currently in the paper, for example by referring to the more general concept of exposure PS (sometimes abbreviated EPS) for balancing e.g. pollution exposed and not exposed, or even more generally to PS as a balancing score for balancing two groups differing on a characteristic (perhaps sex) assumed to be causally (or even noncausally for some risk indicator of interest) related to some other characteristic/outcome.*

This is a good point. Based on this and the feedback of the journal's reviewers, we have changed the title to "Using Propensity Scores to Estimate Effects of Treatment Initiation Decisions: State of the Science."

Comment. *suggest "similar initiatives have occurred" or "similar efforts have been made"*

That text now reads *"similar efforts are underway internationally."*

Comment. *This sounds a bit on the optimistic side. Increasing restrictions to access to health care data sources makes often the quality of variables available in data sources less than ideal. Access to rich clinical data is not always avaialble (e.g., BMI, smoking history, laborator results, access to medical records for validation purposes). I'd try to balance the sentence here or in the discussion.*

RESPONSE TO ISPE MEMBER COMMENTS

This is a good point that our Statistics in Medicine reviewers also made. We have added a parenthetical to that sentence that reads *"(though these variables are often measured with error)."* to clarify that these data are far from perfect.

Comment. *It will be helpful to add the word 'counterfactual' somewhere in this paragraph as a universal term for the scenario being described here*

This term is definitely worth including; we have added a sentence to that paragraph reading *"Typically, a given person or group will experience only one of their potential outcomes, YT=A or YT=B, while the others are rendered unobservable (i.e. counterfactual)."*

Comment. *The usual descriptions tend to say 'influence' rather than 'cause' as it is seen as a somewhat more general term.*

We have changed "cause" to "influence" in that sentence.

Comment. *Should this be E(YA)-E(YB)?*

We have revised the wording there; rather than discuss "being more comfortable" we instead simply say *"if these conditions are met, then $\bar{Y}T=A - \bar{Y}T=B$ is equal to $E(\bar{Y}A) - E(\bar{Y}B)$."*

Comment. *This is not precise. Random allocation ensures independence from other factors irrespective of sample size. This then implies that imbalances between the groups tend in probability to zero with increasing sample size…*

Next comment in the chain: would not agree. On average on repeated sampling or in very large samples. Suggest instead to add "...renders treatment in the obtained sample..."

Based on these comments and those of the Statistics in Medicine reviewers, we have changed the language of this section and replaced "with a large enough sample" with "on average."

Comment. *This is too strong a statement. It is certainly true that, when we have an identified factor, it may be more efficient to use other strategies than randomisation but, of course, we are never sure we have all the relevant information. Hence exchangeability can not be assumed.*

This comment was also similar to several of those by our Statistics in Medicine reviewers; it is important to distinguish more clearly between exchangeability on measured variables and exchangeability on **all** baseline covariates. That sentence now reads: *"While randomization is the only way to balance unmeasured variables, there are other ways to achieve conditional exchangeability with respect to observed variables."*

Comment. *Again. Too strong. The point that needs to be emphasised from Rosenbaum and Rubin is that the entire argument is predicated on the phrase 'if treatment assignment is strongly ignorable given the observed covariates'. This is never known unless we make it so by randomisation. This should read 'create more balanced cohorts and estimate a variety of treatment effects with reduced bias'*

Next comment: Agree. You never have perfect data and model even for the measured confounding variables

RESPONSE TO ISPE MEMBER COMMENTS

Next comment: Without confounding bias? Just being explicit

Next comment: Isn't there a major risk that readers of this paper would cite it just to justify that a "PS-adjusted study is as good as an RCT" and never mention this caveat?

These comments all make a good and similar point; it should be clearer that the bias is only reduced for measured variables (well, and variables associated with them). We have revised that text to read: *"Each of these methods creates cohorts that are exchangeable on the measured variables, estimating a variety of treatment effects with reduced bias from these factors."*

Comment. *Perhaps "...not necessarily balance...", since there are arguments that some balance on unmeasured covariates can be achieved through covariance with the observed covariates.*

Statistics in Medicine reviewers had similar concerns. We have added *"…except to the extent that unmeasured variables are correlated with those that are part of the propensity score model.* To that sentence to include the potential for partial control of confounding.

Comment. *Any comment on which expected distribution of confounding may make PSs preferable to other methods for confounding control? Eg, if most confounding in a given study is expected to derive from age and smoking, we might prefer to adjust for those instead of using PSs.*

This is a good point, and one we believe is best addressed by our comment *"First, propensity score methods are preferable when it is easier or more plausible to identify the model for treatment than for the outcome, particularly in settings with few outcomes. In those cases, regression models for the outcome may be overfit with only a small number of confounding variables."*

Comment. *A large number of treated and untreated individuals; ie, with almost everyone treated, we may not choose to model exposure.*

Most PS methods can simply switch the core "treatment" in such cases; in fact, they can be preferred because outcome models may fare very poorly due to overfitting in the unexposed group.

Comment. *This may be a step too far also. It would be true to say that we gain insight into whether there is insufficient overlap to allow the possibility of comparative effectiveness research.*

This is a good point; gaining insight into when there is insufficient overlap is not the same as gaining insight into when there is sufficient overlap. We have revised that text to read *"Fourth, understanding the propensity score distribution in the treated and untreated can help researchers gain insight into whether there is insufficient overlap (empirical equipoise) between treatment groups to allow for meaningful comparative safety and effectiveness research."*

Comment. *Why only CER? This would seem to be relevant to safety or any other outcome too...*

We have added "safety" to the sentence.

RESPONSE TO ISPE MEMBER COMMENTS

Comment. *I would say effect rather than intervention*

We have changed the manuscript to read *"estimate a treatment effect."*

Comment. *And lifestyle choices and behaviors*

We have added *"lifestyle choices and behaviors"* to the list.

Comment. *I would mention that this is a general issue to consider in any study, independently of using propensity scores or not. Same applies to type of follow-up section.*

Next comment: The subsection Type of follow-up does not seem to focus on or directly relate with propensity scores. These are good points on general study design considerations.

We have tried to make it clear that these study considerations are important for any study via the beginning of the section, whose first paragraph concludes *"While these considerations are important in any study of medical treatments, the decisions below were often ignored in non-experimental work before propensity scores encouraged a focus on treatment assignment mechanisms;"*

Comment. *which design? new user? or prevalent user? or both? I would say this problem applies to both designs when comparing to non-users. This paragraph is also unclear since the first sentences seem to refer to treatment choice, i.e. comparing two treatments. Then suddenly this sentence referring to "this design" talks about non-user comparators. Some clarification of text is needed here.*

This is a good point. We have rephrased that section and it now reads *"Whether restricting to new users or prevalent users, it is often difficult to identify the start of follow-up for any non-user comparators."*

Comment. *selecting or defining...*

Next comment: setting?

We have changed the text to read "setting."

Comment. *should this rather be "...which, in real world settings, can diverge from the [true] biological effect of treatment more and more..."*

The term "biological" effect of treatment is probably a bit too confusing to include. Based on this comment and comments from the Statistics in Medicine reviewers, this text now reads: *"In real world settings, as time passes, that treatment effect can diverge more and more from the effects of initiation and continuous use of treatment."*

Comment. *Is "continued persistence and adherence to a given protocol" meant as an hypothetical quantity, i.e., the effect of a treatment if patients were fully adherent? I think it is what is estimated provided that censoring is independent*

We have specified that this is a "hypothetical" treatment effect.

RESPONSE TO ISPE MEMBER COMMENTS

Comment. *on the one hand…on the other hand…to clarify these are different things and not the same despite the use of "and"*

We have now split those clauses into two separate sentences, separated by the word *"Additionally,"* to make this clearer.

Comment. *In the discussion here or under implementation about target populations, it may be useful to somehow briefly note that an underlying assumption of RCTs (at least in the way their results are used) often is that the effect in the trial target population is representative of the effect in larger populations. I.e. that effect heterogeneity is low. At least that is what is generally the result, with an indication that extends to much broader populations than was actually studied. So this would correspond to the assertions that often ATT ≈ [P]ATE, which often seems to work fairly well in practice. A consequence of this (that could be relevant in PE settings) is that if you are interested in the ATE, but can estimate the ATT with less bias or better precision, then it may be better to focus on the ATT and rely on this assertion to undetrstand the ATE (which is essentially what the trialists do for RCTs). It is true that pragmatic and large-scale heterogeneous trials are increasingly being used, but this seems to often come at a price of more heterogeneity, less confidence in the effect and potentially more bias - even in trials.*

This is an interesting point. We have added a passage stating *"…as heterogeneity in treatment effect can result in a difference in study findings depending on population."* that emphasizes the importance of heterogeneity in study populations. That said, we think that giving this topic the lengthy treatment it deserves is outside the scope of this general review, as it can vary markedly depending on the final use for the data; if the data is for estimating the effect of some policy change or intervention on a population, the ATE is a useful quantity; if you want to know exactly how much people were harmed, the ATT is a more useful quantity; etc. And, of course, we may want to use one study to estimate effects in any number of other target populations via generalizability and transportability methods, in which case heterogeneity is fine so long as the variables associated with it are measured.

Comment. *It would be good to point out that reports of such analyses should clearly state that the results have not been shown to apply to those patients similar to those who have been 'trimmed'.*

Next comment: I would also mention that excessive trimming may indicate you included an instrument in the propensity score. One other benefit about trimming may be to characterize the trimmed population to identify what makes them different, and may be if enough people, see if the effects on the trimmed looks different (heterogeneity)

Next comment: It may be useful to emphasize that in trimming tails and limiting very large weights, we are really only recognizing that we do not have sufficient data for that section of the population, and so we decide to either refrain from even trying or limit the influence of this poor data section on our results.

Our Statistics in Medicine reviewers concurred with these comments; we should emphasize the potential lack of information. That section now reads *"Researchers should also be sure to*

describe those who were trimmed and make it clear that they have limited evidence about effects in them, and (if one characteristic is strongly predictive of being trimmed) consider explicitly reframing the study question to exclude those individuals."

Comment. *Please consider adding a comment about the number of variables that can be included in a PS model (ie, is overfitting a problem?)*

Based on this comment and comments from our Statistics in Medicine reviewers, we have added more detail to that sentence. It now reads *"The choice of covariates is critical, as including variables that only predict treatment reduces study efficiency, and that cost has to be weighted against gains in validity."*

Comment. *It may be worth pointing out that the acyclic graphs depict the assumed true state, and not data patterns.*

Our Statistics in Medicine reviewers concurred. We have added "assumed" to the text.

Comment. *Near-instrument has not been defined*

We have added a definition.

Comment. *caused may not be the best word here. How about "...led to spurious associations with treatment in the data". I assume this is what you mean...*

This section now reads: *"…random sampling error has led to spurious associations with treatment in the study sample."*

Comment. *Generally, what I see as the first approach, is "scientific/clinical a priori knowledge". Is that what you are referring to, optionally supplemented with acyclic graphs to provide structure?*

That knowledge is used to construct the graphs, yes.

Comment. *Two immediate questions follow from this: - when will it NOT be associated with outcome via treatment?; - how then can instrumental variables be identified and eliminated? Or can they not? May be useful to be explicit on these points.*

This is a good point, and our Statistics in Medicine reviewers echoed the first concern. That sentence now reads *"Notably, establishing this ranking by simply estimating marginal associations between variables and the outcome does not always eliminate instrumental variables if treatment affects the outcome, since the instruments will, in expectation, be associated with the outcome through treatment."* making it much clearer how this association will be created (if treatment effects the outcome).

Comment. *I would like to have a note on whether correlated variables should or should not be included in a propensity score model, for instance, a typical example is including prior diabetes diagnosis and treatment for diabetes.*

Generally, collinearity/correlation are not problems for propensity score models-except insofar as one may double-close confounding paths and increase variance for no gain in bias-as the

coefficients should generally not be interpreted. While this is an interesting point, we do not think there is a good location to discuss it in the text.

Comment. *treatment of interest*

That line now reads *"probability of receiving the treatment of interest."*

Comment. *or categorization of continuous covariables*

We have added a line reading *"…including whether to categorize continuous covariates."*

Comment. *I think that checking balance could be a separate paragraph, and may be between score estimation and balance checking, trimming should be discussed.*

We have split those two paragraphs. We prefer to focus on trimming in the Study Design section, as we believe it is generally better to have trimming cutpoints specified prior to estimating the propensity score.

Comment. *It may be worth adding that if stratification of PS is used and imbalance in only found in the tails, further trimming can be tried and see if balance is achieved.*

While this is an interesting thought, we prefer to keep this score estimation section relatively method-agnostic and not delve into specifics of implementation approaches prior to describing them in detail.

Comment. *Please specify option for what, as there is ambiguity as to using these models for estimating the PS vs measuring balance.*

That section now reads *"option for estimating the propensity score."*

Comment. *I Two reflections: 1. For a uniform treatment effect (which may be the true state, and is often an analysis assumption), will they then answer the same question? 2. Even for a non-uniform effect, I would think the summary effect estimated MAY be the same, if the integrated total of the non-uniform effects happens to be identical in different populations. And in practice - they may be so similar it is impossible to determine if differences are true or due to random error. So they may be answering different questions, but get identical or similar summary results.*

Our focus here is generally on the target population as it relates to the causal question, but the point about distinguishing the actual estimate from the causal question is an interesting one. Generally, the option of exact cancelling/near cancelling here is similar to the idea of exact cancelling/near cancelling of confounding relationships. While we appreciate this thought, we prefer to avoid discussion of this "pseudo-faithlessness" in this comparatively introductory paper.

Comment. *For one to many matching, I think the caveats about potential bias due to matching "to one side" should be mentioned, which would also tend to increase with increasing number of matches. There has been work describing advantage from picking matches alternatively from one and the other side of the closest matches.*

Next comment: Is there a good reference for this method to add here? Naively, with replacement, you would assume with this brief description that the same match would be picked as the best one each time for a particular subject, which does not add additional information.

These comments are definitely an interesting consideration. We have added a clause discussing such one to many with replacement matching as well as "one-sided" matching with a reference for further reading: *"for such matching a balanced matching strategy can help potential bias from "one-sided" matching."*

Comment. *What is meant with "validity" here. Of what? Also, "may" in the first half of the sentence, does that relate to the case of varying matching ratio? If the standard error changes with adjusting, both unadjusted and adjusted can hardly be valid.*

This passage has more or less been completely retooled based on feedback from the reviewers at Statistics in Medicine and now refers only to the estimate, not the standard error. It now reads: *"After matching, outcomes in the two groups can generally be compared directly since matching leads to exchangeability on measured variables and therefore removes (measured) confounding."*

Comment. *Do you mean because the dataset becomes smaller/small? If so, might be good to say so explicitly. At which sample size is this realistic potential problem though?*

We have added *"…in the progressively smaller data set…"* to that sentence; we hope that future statistical work can give a sense of what sample size this will be problematic; current simulation work is somewhat limited.

Comment. *I this term is not clear. Either there is overlap or there is not, so how can it be limited? Do you mean lack of comparators in the high-propensity strata?*

This section has been revised based on this comment and thoughts from reviewers at Statistics in Medicine to add additional clarity. It now reads: *"This can lead to considerable residual confounding because of poorer within-strata balance in the low propensity score strata."*

Comment. *I think it may be worth mentioning that balance needs to be achieved, so checked, within each stratum*

We have added a comment to that effect *"(and balance within strata should be checked)."*

Comment. *In my experience, I think it's better to use a categorization of the PS than using the continuous variable. Maybe mentioning as an option, as I don't have reference for this...*

While this is an interesting point, we think it somewhat conflicts with the concept of the complex multivariable relationship of the score with the outcome thanks to its multivariable nature and may confuse readers.

Comment. *Suggest "modeling of PS" for clarity. (as outcome modeling can also be combined with these)*

Replaced with "*propensity score modeling."*

RESPONSE TO ISPE MEMBER COMMENTS

Comment. *Even if this approach is not commonly used, one advantage not mentioned is the ability to adjust for a specific variable, which you want to control separately from the PS including it in the model together with the PS.*

This is an interesting point worth making explicitly; that sentence now reads *"However, propensity score modeling can be combined with propensity score stratification, matching, or weighting, and researchers can attempt to reduce residual confounding from measured variables in the score by including them in the multivariable regression."*

Comment. *or odds-weights*

We added odds-weights and ATT weights as potential alternative names.

Comment. *i.e the treated population restricted to those that could be matched?*

We have added 'treated" to that sentence describing the target population of match weights.

Comment. *The discussion on confidence intervals could be considerably expanded: problems with correct CI's do not only result from inability of software to handle weighted data, but also from the fact that absolute weights depend on the relative amount of data in the two groups being compared (i.e. the "average propensity score") - which is quite arbitrary. In IPTW there is stabilized weights - which restore the original sample size and yield approximately correct CI's.. unclear (for me) how this works with SMRW.. probably bootstrapping would be an applicable solution there as well.*

At the suggestion of the Statistics in Medicine reviewers we have added an entirely new sub-section of implementation dealing with estimating variance.

Comment. *Is this discussion particularly pertinent to [P]ATE? Worth pointing out?*

We have added *"particularly when estimating the population average treatment effect"* to that sentence.

Comment. *This terminology has always confused me. Not sure why it is undue. It is large but correct from a weighting perspective BUT reflects the lack of information (sample size) in that section of the PS. So it is problematic (because we don't have data) but not really undue in my view. And if they are special and lack important data on confounders, not only do we have lack of sample size, but differential lack of or deficient quality of data in that section. I think these concept could be expressed with more care.*

This is a good point that was echoed by our Statistics in Medicine reviewers. We have replaced "undue" with "large" for exactly the reason describing by this reviewer.

Comment. *I miss some reference to what if some variable for the PS have missing values*

This is an interesting question. Because we have added considerably to the length of the manuscript in this revision, we are wary of introducing the question of treating missing data appropriately and adding yet another concept-unlike the other study design considerations we cover, it has relatively little "unique" aspects brought to like by propensity score methods.

RESPONSE TO ISPE MEMBER COMMENTS

Comment. *Any citation regarding missing values? Is multiple imputation the gold-standard?*

See the comment above. Because we have added considerably to the length of the manuscript in this revision, we are wary of introducing the question of treating missing data appropriately and adding yet another concept-unlike the other study design considerations we cover, it has relatively little "unique" aspects brought to like by propensity score methods.

Comment. *and to what extent and how inference to other target populations may be possible.*

Next comment: Again - the question of inference to other populations should also be discussed, otherwise research will become meaningless if we never can make inferences to other populations

This is a good point. We have added to the text and it now reads *"Researchers should also be clear about the treatment effect they are estimating with their analysis to ensure readers can understand to whom it applies and to what other populations the inference could extend."*

Comment. *I think that the amount of trimming, and a brief description of the population excluded by trimming should be reported. As well as, apart from the "typical" table 1 with the description of the initial population, is important to report the characteristics of the population after trimming, which is the population where the effect estimates applies.*

This is a good point. We have added an additional sentence to that paragraph which reads *"It is especially helpful to know the number and covariate distributions of individuals in each treatment group that were excluded from the analysis due to their propensity score or a failure to find a match, as they may be important in defining future study populations."*

Comment. *I think that this section may go before the balance check and together with trimming discussion.*

That is an interesting thought. Based on the other changes we have made, we think the paper flows slightly better as is.

Comment. *Would the preference score be even more useful in this context?*

We have added a discussion of the preference score to the text in that portion.

Comment. *Not sure what this "however" is making a contrast to. Omit? My (implicit) reading of these sentences is that if nonoverlap increases, then revisit your PS model to try to remove instruments. Is that what you are trying to say? So why is it stated here under Reporting?*

We have restructured that sentence for clarity.

Comment. *One final comment - would it be useful to add a few sentences on the topic of balancing more than two groups?*

We have added this and other topics to a brand new section titled "Evolving usage of propensity scores."