Propensity score methodology for confounding control in health care utilization databases

Elisabetta Patorno[1], Alessandra Grotta[2, 3], Rino Bellocco[2, 3], Sebastian Schneeweiss[1]

ABSTRACT

Propensity score (PS) methodology is a common approach to control for confounding in non-experimental studies of treatment effects using health care utilization databases. This methodology offers researchers many advantages compared with conventional multivariate models: it directly focuses on the determinants of treatment choice, facilitating the understanding of the clinical decision-making process by the researcher; it allows for graphical comparisons of the distribution of propensity scores and truncation of subjects without overlapping PS indicating a lack of equipoise; it allows transparent assessment of the confounder balance achieved by the PS at baseline; and it offers a straightforward approach to reduce the dimensionality of sometimes large arrays of potential confounders in utilization databases, directly addressing the "curse of dimensionality" in the context of rare events. This article provides an overview of the use of propensity score methodology for pharmacoepidemiologic research with large health care utilization databases, covering recent discussions on covariate selection, the role of automated techniques for addressing unmeasurable confounding via proxies, strategies to maximize clinical equipoise at baseline, and the potential of machine-learning algorithms for optimized propensity score estimation. The appendix discusses the available software packages for PS methodology. Propensity scores are a frequently used and versatile tool for transparent and comprehensive adjustment of confounding in pharmacoepidemiology with large health care databases.

Key words: Propensity score; Health care databases; Confounding; Epidemiologic methods; Non-experimental studies

(1) Division of Pharmacoepidemiology and Pharmacoeconomics, Department of Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, MA (U.S.A.)
(2) Unit of Biostatistics, Epidemiology and Publich Health, Department of Statistics and Quantitative Methods, University of Milano Bicocca, Milan, Italy
(3) Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden

CORRESPONDING AUTHOR: Elisabetta Patorno, Division of Pharmacoepidemiology and Pharmacoeconomics, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, 1620 Tremont Street (Suite 3030), Boston, MA 02120. Tel: 617 278-0930. Fax: 617 232-8602. e-mail: epatorno@partners.org

DOI: ***10.2427/8940***

INTRODUCTION

Large health care utilization databases are frequently used to study the effect of therapeutics on health outcomes [1]. Health care utilization data (1) reflect routine practice, which allows for the evaluation of real-world effectiveness and safety in large populations that include patients often under-represented in or completely excluded from clinical trials (e.g., the elderly, children, pregnant women); (2) are compiled in large enough bodies to make them useful in the study of rare events or newly marketed products; and (3) are readily available to researchers without the delays common to the collection of primary data [2, 3]. Despite their importance, studies of pharmacoepidemiologic claims data encounter specific issues that can compromise their validity. One of the principal threats to validity is confounding by indication [4], or drug channeling bias [5]. Physicians prescribe drug treatments in light of the diagnostic and prognostic information available at the time of prescribing. If predictors of patient outcomes are unevenly distributed among treatment groups, then failing to control for such factors will lead to confounding [6, 7]. Propensity score methodology is a suitable and commonly used analytical approach to control for confounding in database studies of treatment effects.

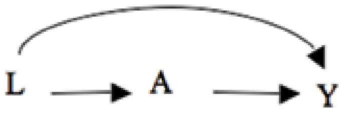
This article provides an overview of the use of propensity score methodology for pharmacoepidemiologic research based on large health care utilization databases. We begin with a description of confounding bias in pharmacoepidemiologic research and introduce the main limitations of conventional multivariable outcome models; we provide definitions and properties of propensity scores and discuss the advantages in pharmacoepidemiologic research using large health care databases; we describe criteria for covariate selection and discuss the role of proxy adjustment and automated adjustment techniques and illustrate strategies to balance treatment groups by their estimated baseline propensity scores; finally, we discuss the available software packages for PS methodology.

Confounding by indication and limitations of conventional outcome regression models

Physicians prescribe drug treatments in light of both clinical and non-clinical information available when patients present. For example, a physician's treatment choice will be driven by a specific diagnosis but may be also based on an evaluation of the patient's health status and prognosis, the physician's past experience with the medication, and an assessment of the patient's physical and cognitive ability and willingness to take a medication as prescribed [8]. While underlying consistency in the selection of treatment is fundamental for an appropriate prescribing process, it also means that patients receiving a specific therapeutic regimen versus another will differ in underlying clinical characteristics. If these characteristics are also predictors of patient outcomes (e.g., patients who preferentially receive a treatment are also at a higher risk for the study outcome), then an unconfounded assessment of a medication's effect on those outcomes will rely upon addressing the baseline differences between treated groups of patients that derive from the prescribing process [9]. This scenario is depicted in the directed acyclic graph (DAG) in Figure 1.

FIGURE 1

STRUCTURE OF CONFOUNDING [11]



The backdoor path between treatment A and outcome Y is open through their common cause L, e.g., any baseline characteristic unevenly distributed between treatment groups that is also an independent risk factor for the disease outcome.

The backdoor path between the treatment A and the outcome Y – any noncausal path between A and Y – is open through their common cause L, e.g., any baseline characteristic unevenly distributed between treatment groups that is also an independent risk factor for the disease outcome [10]. Causal inference involves the counterfactual comparison between the effect that treatment would have had on a patient and the effect that a different treatment would have had on the same patient at the same point in time. Randomization is a popular method to block all backdoor paths between treatment A and outcome Y, i.e., to control for confounding, because a random assignment of treatment is expected to produce groups that are comparable (exchangeable) with respect to (known and unknown) baseline characteristics [11]. This baseline comparability will allow the conclusion that any observed difference in the occurrence of outcomes during follow-up is due to the only characteristic that differs between the study groups by design, i.e., treatment.

In an observational setting, this baseline comparability is achievable only if all potential confounding variables L, leading to the prescription of a treatment vs. another, are identified, measured, and adjusted for, i.e., if the assumption of no unmeasured confounding is met. Under this assumption and the absence of other biases, causal statements can be drawn from observational data using multivariable modeling. Outcome regression models, however, do not explicitly examine the association between the factors entering into the prescribing decision and the treatment under study, and are constrained by the possibility that only a limited number of covariates can be accounted for per outcome [12]. This latter aspect is particularly in conflict with the common setting in pharmacoepidemiologic research of relatively few outcomes and the many potential covariates that may explain the prescribing process.

Propensity score definition and advantages in pharmacoepidemiologic research with large health care databases

An exposure propensity score is the estimated probability (propensity) of receiving treatment based on the measured covariates included in the propensity score model [13]. If A is an indicator for the exposure of interest, A=1 if a subject initiates a treatment, A=0 if a subject initiates another treatment, and L is a vector of potential determinants of treatment, then the propensity score is the conditional probability of receiving treatment given the covariates; that is, PS = Pr[A=1|L]. The propensity score is usually estimated by logistic regression of treatment, such as:

$$Logit\,(\mathrm{A} = 1) = \beta_0 + \beta_1\,(L_1) + \beta_2\,(L_2) + \beta_n\,(L_n)$$

However, other approaches are possible, including discriminant function analysis, classification and regression trees, or neural networks [14, 15].

Each patient is assigned an estimated probability of exposure ranging from 0 to 1, which can be viewed as a summary score that reflects the likelihood of being prescribed a given treatment, given all observable characteristics. Individuals with similar estimated propensity scores will have, on average, similar chances of receiving that treatment and overall a similar covariate distribution, although they may have different patterns of covariates at an individual level.

The propensity score can be used to reduce confounding via matching, stratification, regression adjustment, or any combination of these strategies [16].

Propensity score methodology offers some clear advantages in pharmacoepidemiologic research. (1) It focuses directly on the determinants of treatment, encouraging the researcher to explore the factors that predict treatment in more detail than would be the case for conventional multivariate models. An improved understanding of such factors and their inclusion in propensity score models may improve control of confounding. (2) It allows for the graphical comparison of the distribution of propensity scores for exposed and unexposed subjects and for the identification of areas of non-overlap. Patients with contraindications to use of a drug (or those with absolute indications) may have no comparable exposed subjects (or unexposed subjects) for valid exposure effect estimation. These subjects are not comparable and should not be considered for analysis. In contrast, parametric outcome models extrapolate into this parameter space, making unsupported assumptions. (3) Moreover, the propensity score offers a straightforward approach to reduce the dimensionality of the array of important confounders, directly addressing the “curse of dimensionality” issue [14]. This is particularly relevant in the evaluations of therapeutics, where we often deal with frequent exposure and rare outcomes [17].

By estimating the PS and analyzing the data within homogeneous levels of PS (e.g., stratification or matching on the propensity score), we can achieve a better balance of measured covariates between exposed and unexposed subjects than would be possible under randomized treatment assignment [18]. However, because propensity scores are conditional on measured covariates only, they might not control for unmeasured or imperfectly measured variables. Thus, unmeasured confounding bias cannot be excluded [2].

Criteria for covariate selection

A critical challenge for researchers using propensity score methodology is how to identify the relevant variables to be included in the propensity score model.

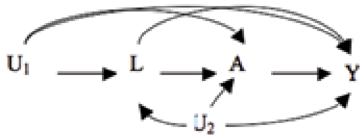
In general, any variable L that is thought to be a common cause of an exposure A and an outcome Y should be included in the propensity score model (Figure 1). In addition to this general rule, Rubin has recommended that all variables related to the outcome should be included in the propensity score model regardless of their association with exposure [19], a recommendation supported by subsequent simulations showing that the inclusion of these covariates decreases the variance of an estimated exposure effect [20]. These simulations and further work have also shown that variables associated with the exposure but not with the outcome (i.e., instrumental variables or IVs), should not be included in the propensity score model, as they increase the variance of exposure effect estimates and, when unmeasured confounders are present, may also increase bias compared with the crude estimate, i.e., Z-bias [20, 21]. In practice it is usually impossible to decide with certainty whether a variable is an IV or a confounder. In such cases it is recommended to adjust for the covariate instead of leaving it out [21].

Model specification is primarily guided by subject matter knowledge (e.g., a detailed understanding of how a particular drug treatment is assigned to patients with varying levels of outcome risk at baseline) and commonly relies on the identification of an *a priori* set of baseline risk factors to be included in the propensity score model. These covariates typically include demographic characteristics (age, gender, and race) calendar time, history of major medical conditions, measures of overall comorbidity [22-24], history of specific medication use, history of acute care hospitalizations, and measures of health care system use – such as number of hospitalization, physician's visits, and drugs dispensed over a given period of time before treatment initiation [25].

However, in practice a detailed understanding of how a particular drug treatment is assigned to patients is not always explicitly understood, and the available subject-matter knowledge is often inadequate to specify with any degree of certainty the complex causal connections between variables that determine exposure or outcome. This problem is exacerbated in the setting of large health care utilization databases, in which many variables are not directly measured (e.g., clinical disease severity, laboratory results, functional status, body mass index, smoking status, and over-the-counter medication use) and the meaning of diagnosis codes is not always clear [8]. Unknown, unmeasured, or residual confounding U variables (Figure 2) might limit the role of causal graph theory in understanding and describing potential bias.

FIGURE 2

STRUCTURE OF CONFOUNDING IN THE PRESENCE OF UNMEASURED CONFOUNDING



The backdoor path between treatment A and outcome Y is open through their common causes U_1 and U_2, i.e., any unobservable baseline characteristic unevenly distributed between treatment groups.

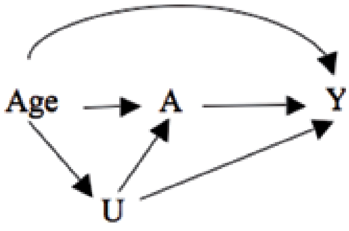
Proxy adjustment and automated techniques for variable selection

Longitudinal health claims data provide a rich source of information about patient health status and confounding beyond what is normally used in pharmacoepidemiologic research, and it has been suggested that they could be utilized as sources of proxies that indirectly describe the health status of patients. For example, Schneeweiss et al. [26] argue that the health status of a patient can be (a) assessed through the dispensing of a drug that was (b) prescribed by a physician who made a diagnosis in a (c) patient who came forward for medical care and (d) presented certain symptoms. Such a set of proxies can be influenced by many factors (e.g., access to care [27], severity of the condition, diagnostic ability of the physician, preference for one drug over another [28], the patient's ability to pay the medication co-payment [29], and the accurate recording of the dispensed medication), which are not directly observable in claims data.

In such a context, it may be argued that some recorded variables (e.g., patient characteristics, medical diagnoses, drug use, etc.) may function as sufficient proxies for unmeasured variables, and may be used to adjust for confounding. For example, old age serves as a proxy for comorbidity, frailty, cognitive decline, and many other factors to the extent it is correlated with these factors (Figure 3).

FIGURE 3

THE MECHANISM OF PROXIES TO ADJUST FOR UNMEASURED CONFOUNDING IN HEALTH CARE CLAIMS DATA



Measured confounders such as Age may serve as proxies for unmeasured confounders (U), e.g. frailty.

It has been suggested that adjusting for a perfect surrogate of an unmeasured factor is equivalent to adjusting for the factor itself [30], and that the degree to which a surrogate is related to an unobserved or imperfectly observed confounder is proportional to the degree to which adjustment can be achieved [31, 32]. Thus, a sufficiently large set of measured proxy covariates would likely be a good overall proxy for relevant unobserved confounding factors. The challenge is how to empirically identify these proxies out of the thousands of variables available in large health claims databases.

In such a context, an automated technique that assesses thousands of diagnoses, procedures, and drug treatment codes plus their clustering in time out of the high-dimensional variable space in longitudinal health care utilization databases can be helpful in identifying potential confounders or proxies for confounders [26]. These empirically identified covariates can then be used in addition to or in place of investigator-selected variables to estimate a PS.

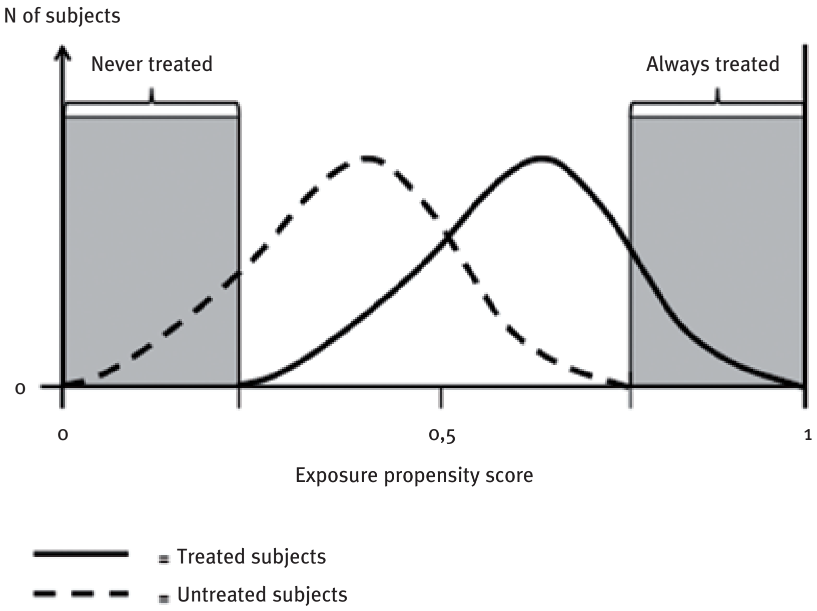
Overview of the high-dimensional propensity score algorithm for variable selection

The high-dimensional propensity score (hd-PS) algorithm grounds on automated technique that examines thousands of covariates among different claims data dimensions in the study population. Each dimension describes an aspect of care and has different information content, such as recorded diagnoses in hospitals versus outpatient, performed procedures, and dispensed drugs. For each dimension, the top *n* most prevalent codes are identified and classified into three levels of within-patient frequency of occurrence during the baseline period: code occurred ≥ 1 times (once), ≥ median number of times the code was observed for the cohort (sporadic), and ≥ 75th percentile number of times the code was observed for the cohort (frequent). A code classified as frequent would have a "true" value for all three levels of frequency of occurrence. These codes are subsequently transformed into binary covariates and then individually assessed for selection into a propensity score. For example, with five dimensions (hospital diagnoses, hospital procedures, doctor's office diagnoses, doctor's office procedures, and pharmacy prescription fills), the 200 most prevalent codes from each dimension (n=200), and three levels of within-patient frequency of occurrence of each code (once, sporadic, or frequent), there are up to 3 000 possible indicator variables that could be added to a propensity score. By default, the hd-PS algorithm then prioritizes each of these variables by its potential to bias the exposure-outcome relationship under study ("bias" ranking criterion) on the basis of the formula by Bross [33] and includes the top k=500 of these covariates in a propensity score. Other methods for selecting variables have also been developed [34]. In situations with few exposed outcomes, the top k variables may be selected on the basis of the empirical confounder-exposure association ("exposure-only" ranking criterion); in situations of frequent outcomes, selection may consider the confounder-outcome association ("outcome-only" ranking criterion) resembling a disease risk score [35]. The issue of few exposed cases can be addressed by zero-cell correction for the covariate-exposure and covariate-outcome associations, although small numbers of exposed subjects remain challenging [36].

This algorithm will not only empirically identify most of the investigator pre-defined covariate codes but will also identify additional potential confounders. These covariates are hoped to be proxies for constructs that are difficult to measure in claims data or are factors that the investigators did not consider and together have been shown to improve adjustment for confounding in several example studies [37]. The hd-PS algorithm and its associated Statistical

FIGURE 4

DISTRIBUTION OF ESTIMATED PROPENSITY SCORES FOR TREATED AND UNTREATED SUBJECTS



Adapted from Stürmer T, et al. [38]

Analysis System (SAS) code are available at www.hdpharmacoepi.org [34].

Achieving clinical equipoise at baseline with estimated propensity scores

As with RCTs, clinical equipoise, i.e., equivalence of patients at baseline, is desirable in observational studies. Plotting and comparing the distribution of estimated propensity scores for exposed and unexposed subjects can be informative in assessing clinical equipoise at baseline and should be standard practice in database analyses using propensity score methodology (Figure 4) [38].

The amount of non-overlap of these two curves on the extreme ends of the distribution identifies (1) patients who have a very low probability of treatment and are never treated, e.g. patients with important contraindications, and (2) patients who have a very high probability of treatment and are always treated, e.g., patients with absolute indications. These patients are not equally plausible candidates for the treatment under study, i.e., because there is no clinical equipoise, it is questionable whether these patients should be included in an analysis [2].

Conversely, the area of overlap will identify patients who have comparable propensity scores among treated and untreated subjects, and are therefore better candidates for inclusion in a comparative analysis.

In this regard, a very high C-statistic generated by the propensity score model indicates small overlap in the distribution of propensity scores between treatment groups due to limited clinical equipoise in the population, making comparisons between treatment groups that do overlap (and are comparable) less precise (Figure 5) [39].

The estimated propensity score can be used to reduce confounding via matching, stratification, weighting or regression adjustment. However, matching or stratification with trimming, which exclude subjects with more extreme propensity score values, will guarantee equipoise between treatment groups more often than inclusion of a propensity score in a multivariate regression model. Matching and stratification on the propensity score will be further discussed in the next sections.

Matching on propensity score

Matching on the propensity score entails forming matched sets of treated and untreated

FIGURE 5

DISTRIBUTION OF ESTIMATED PROPENSITY SCORES FOR TREATED AND UNTREATED SUBJECTS WITH MINOR OVERLAP

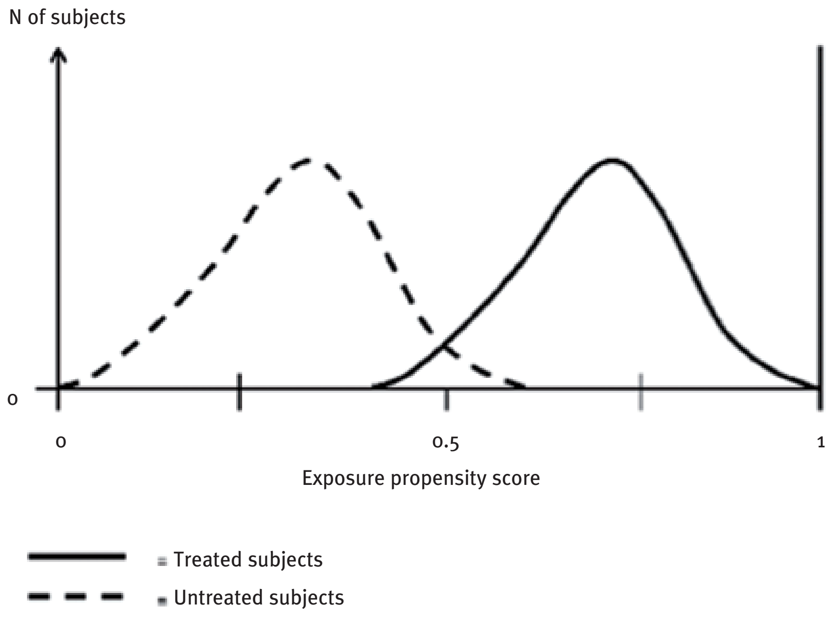
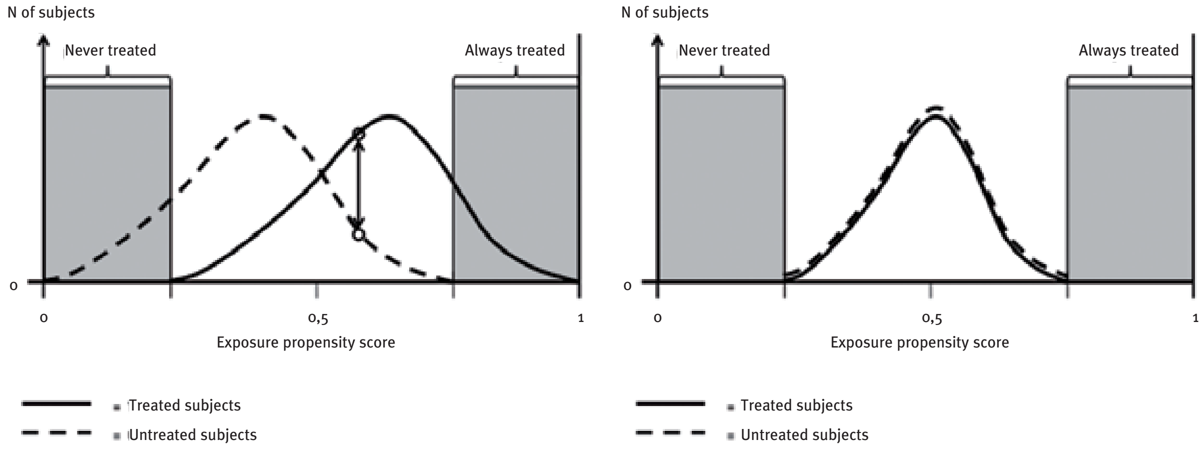


FIGURE 6

PROPENSITY SCORE MATCHING PROCESS



Before matching

After matching

Adapted from Schneeweiss S. [39]

subjects who have a similar probability of receiving treatment [40] (Figure 6).

After matching, the treated and untreated groups tend to have, on average, the same distribution of observed baseline covariates that were considered in the propensity score model. The most common implementation of propensity score matching is pair matching or 1:1 matching, in which matched pairs of treated and untreated subjects are formed.

Once matched pairs have been identified, matched treated and untreated subjects can easily be compared by calculating crude relative risks or relative differences, which will provide estimates of the average treatment effect in the treated subjects.

When feasible, matching on the propensity score offers investigators several advantages. Close matching of subjects across exposure groups on a propensity score ensures almost perfect overlap in the propensity score distributions after matching. This is because matching excludes all subjects falling in the non-overlapping ranges of the score, i.e., those with no comparable controls. In this regard, the matching process serves a function similar to propensity score trimming and improves the validity of the estimates [41, 42]. In addition, matching on PS offers investigators the ability to balance treatment groups across all potential confounders and to inspect the achieved balance across measured covariates, by comparing these variables before and after matching in a similar manner to the comparison of randomized treatment groups from a randomized clinical trial. This can be done graphically or by comparing differences between groups [40, 43].

One limitation of matching is that not all subjects will be matched. We have already mentioned that in order to achieve high validity, it may be necessary to exclude exposed subjects for whom no comparable subjects are found in the referent group. However, in 1:1 matching of few exposed and many unexposed subjects, many unexposed subjects may remain unmatched and excluded from the analysis, yielding less precise estimates [14]. Variable ratio matching can improve precision, although it has been observed that strong asymmetry in the size of the treatment groups may be required for meaningful improvements in precision [44].

Matching algorithms

Pair matching or 1:1 matching is commonly performed without replacement, in which an untreated subject who has been matched with a treated subject is no longer available for consideration as a potential match for other treated subjects. As a result, each untreated subject is included in at most one matched set. In contrast, matching with replacement allows a given untreated subject to be included in more than one matched set. When matching with replacement is used, variance estimation must account for the fact that the same untreated subject may be in multiple matched sets [45, 46]. Commonly used matching algorithms for 1:1 matching are greedy matching and nearest neighbor matching [47]. With greedy matching [48], a random treated subject and the nearest untreated subject are selected for matching. The untreated subject is selected even if it would better serve as a match for a subsequent treated subject. With nearest-neighbor matching [49], pairs of treated and untreated subjects are formed to minimize the total within-pair differences in the propensity score. The use of nearest-neighbor matching has been partly limited due to high computational intensity. However, Rassen et al. have shown that a variation of the nearest-neighbor matching algorithm, the pairwise nearest-neighbor matching algorithm, is computationally fast and provides better balance among treatment groups [42]. Additional information can be found in the Appendix.

One-to-many (1:n) matching

One-to-many matching is an extension of 1:1 matching in which a treated subject is matched to *n* untreated subjects, with either a fixed or variable ratio.

Matching at ratios of 1:n aims to increase the number of untreated subjects included in the analysis, limiting the loss of information characteristic of 1:1 matching and augmenting the precision of the estimated exposure effect.

Fixed ratio matching is not optimal, since it does not account for the fact that some treated individuals may have many close matches while others have very few. Ming and Rosenbaum [50] proposed a form of ratio matching that allows for a variable number of untreated subjects to be matched to each treated subject. They also found that matching with a variable number of controls reduced bias as compared with fixed ratio matching. More recently it has been shown that augmenting the number of untreated subjects matched to each treated subject increased the bias in the estimated treatment effect, arguably because matches subsequent to the first will lead to lower quality matches [51]. However, further simulations found that increasing the match ratio beyond 1:1 increased precision in cohort studies at a small cost in bias [42].

Full matching or many-to-many matching, i.e., forming matched sets consisting of either one treated subject and at least one untreated subject or one untreated subject and at least one treated subject, has also been proposed [52-54].

A difficulty in using one-to-many or many-to-many matching is that because of the differing numbers of patients in each matched set, a simple "Table 1" of a variable ratio matched cohort will not show how well baseline covariates are balanced, and diagnostics for assessing such a balance are less well-developed than in the setting of 1-1 matching. Weighting each patient's characteristics by matched set size has been proposed [55]; this method will describe balance but will lack transparency compared to 1:1 matching. Presenting a Table 1 with each matched set's single best match (1:1 "best matches") or a hybrid approach, with both the 1:1 "best matches" displayed alongside a weighted population, have also been suggested [42].

Three-way matching

Pharmacoepidemiological studies generally assess the relative benefits and risks of one medication versus another. However, for many medical conditions, three or more appropriate treatment choices may be available. As traditionally defined, propensity scores predict patients' probabilities of receiving one treatment versus a single alternative. Imbens suggests computing the conditional probability of receiving a particular level of the treatment given baseline covariates, i.e., the generalized propensity score, to account for multiple levels of treatment [56]. However, the same author also notes that matching approaches are probably less suited to multiple levels of treatment than other propensity score adjustment methods. Rassen et al. have recently developed an algorithm for simultaneously matching groups of three patients on propensity score, in a three-way matching approach [44]. This algorithm creates cohorts of exchangeable patients among whom the comparative effects of three exposure groups could be studied. Revised software for handling more than three study groups is currently in preparation.

Stratification on the propensity score after trimming

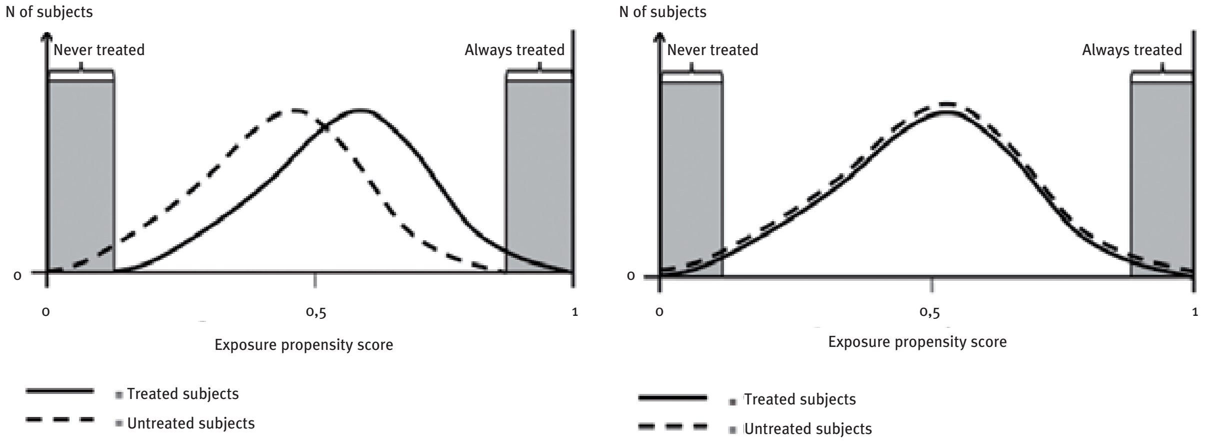
Stratification is frequently used to control for confounding in epidemiologic research and involves grouping subjects into mutually exclusive strata that are determined on the basis of the propensity score. A common approach is to define the strata using specified percentiles of the propensity score distribution. Within each stratum, treated and untreated subjects will have roughly similar propensity scores. Therefore, if the propensity score has been correctly specified, the distribution of measured baseline covariates will be similar between treated and untreated subjects within the same stratum [45]. The researcher will have to decide how many strata should be used in the analysis. Based on Cochran's [57] observation that a stratified analysis with five strata is sufficient to remove at least 90% of the bias for most continuous distributions, Rosenbaum and Rubin [58] proposed stratifying the propensity score into quintiles. It has been suggested that increasing the number of strata improves bias reduction until matching is reached, although the marginal reduction in bias decreases as the number of strata increases [57, 59].

Within each stratum, the effect of treatment on outcomes can be estimated by directly comparing treated and untreated subjects. The stratum-specific estimates of treatment effect can then be pooled across strata to estimate the overall average treatment effect [58], which is a weighted average with weights equal to the proportion of individuals within that stratum.

Stratification on the propensity score offers several advantages. Stratified analyses based on the propensity score are transparent, as within-stratum balance can be readily assessed [44] and treated, and untreated subjects in each stratum can be directly compared using simple methods. Stratification allows for an explicit evaluation of potential for effect measure modification (i.e., effect modification by propensity score) by comparing the individual stratum-specific effect estimates before estimating the overall treatment effect. Finally, in contrast to matching, all study participants can be used to estimate the treatment effect, thereby limiting the loss of information characteristic of propensity score matching. However, the inclusion of all subjects might introduce bias due to inclusion of individuals with a propensity score outside the overlapping range of scores among treated and untreated, i.e., patients who are never or always treated (Figure 4). These extreme observations may be overly influential and problematic in estimating the effect of treatment

FIGURE 7

REPRESENTATION OF PROPENSITY SCORE DISTRIBUTION TRIMMING



Trimming non-overlapping regions of the propensity score distribution

Trimming with symmetric range restriction

because of minimal covariate overlap between exposed and unexposed subjects [14].

This bias may be reduced by truncating the data in the tails of the distribution, i.e., excluding unexposed patients with a propensity score lower than the lowest propensity score observed in exposed patients and vice versa, or symmetrically trimming according to the extreme X% of the propensity score distribution in the overall population (e.g., exclusion of the extreme 2.5% in both the tails of the propensity score distribution) (Figure 7).

Trimming with asymmetric range restriction

Trimming with asymmetric range restriction is an extension of trimming the non-overlapping regions of the propensity score distribution. It entails trimming asymmetrically according to the percentile of the propensity score in treated patients at the lower end and in untreated patients at the upper end (e.g., excluding the 5th and 95th percentiles of the PS distribution in the treated and untreated patients, respectively). Thus, asymmetric trimming will also exclude patients who were treated or untreated most contrary to prediction (Figure 8).

Sturmer et al. [41] hypothesize that unmeasured confounding such as frailty may be at the root of prescribing patterns contrary to prediction ("treatment withheld" or "last resort") and may explain the treatment effect heterogeneity observed in epidemiologic studies addressing elderly populations [60, 61]. The authors show that under the assumption of unmeasured confounding driving prescribing patterns contrary to prediction, increasing asymmetric PS trimming may increase the validity of the treatment effect estimates.

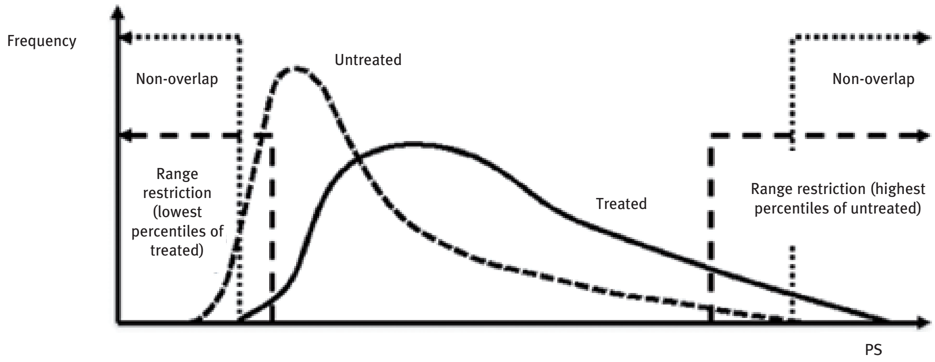
Algorithmic approaches

With the increasing availability of high-dimensional data to researchers, machine learning procedures are becoming object of increasing interest as tools to efficiently identify optimal methods for confounding adjustment [62]. Examples of algorithms for improved variable selection for predicting exposure and outcomes risk beyond standard logistic regression include random forest, [63] neural networks, [64, 65], lasso regression [66], and many more. Few studies have examined the performance of these methods for the estimation of propensity scores [15, 67].

Despite the wide range of estimating algorithms available, it has been argued that in practice the estimation of outcome regressions or propensity scores frequently uses parametric estimators like logistic regression, which are often misspecified and therefore may lead to biased treatment effect estimates [68]. Van der Laan et al. have proposed a data adaptive algorithm, the super-learner algorithm, for optimizing the prediction of treatment choice and endpoints [69, 70]. The super-

FIGURE 8

SCHEMATIC OF ASYMMETRIC RANGE RESTRICTION


Frequency
Non-overlap
Untreated
Non-overlap
Range restriction (lowest percentiles of treated)
Treated
Range restriction (highest percentiles of untreated)
PS

PS, propensity score

From Stürmer T, et al. [41]

learner algorithm combines several estimating algorithms and chooses the best combination using cross validation. It is claimed to perform at least as well as the best single candidate algorithm [69]. This machine-learning approach allows the researcher to assess many candidate estimating algorithms – ranging from machine-learning algorithms to parametric estimators proposed by subject matter experts – prior to looking at the data, without having to choose only one prediction that may not perform well in a particular application. The super-learning methodology may be coupled with targeted maximum likelihood estimation (TMLE) or collaborative TMLE (CTMLE), a doubly robust (DR) estimator that combine estimators of the outcome regression and the propensity score and maximize the precision of a selected target parameter, e.g., the relative risk of the drug outcome association [69].

CONCLUSIONS

The propensity score is an increasingly common methodology to control for confounding in observational studies using large health care utilization databases. Specific features make this methodology particularly suitable for epidemiological research of treatment effects. Propensity scores address directly the determinants of treatment, “forcing” researchers to think through the clinical decision-making process and the potential sources of confounding of the exposure-outcome association; allow for the straightforward examination of PS distributions and the easy assessment of the confounder balance at baseline; and are particularly useful in situations of frequent exposure and rare outcomes. Proxy adjustment and automated techniques for variable selection assist the researcher in addressing and limiting unmeasurable confounding in the context of high-dimensional data. *Ad hoc* matching algorithms and truncation of subjects without overlapping PS distributions maximize clinical equipoise at baseline and guarantee less biased contrasts among more comparable treatment groups. Machine-learning algorithms are currently in development to optimize propensity score estimation.

Propensity scores are a versatile tool for transparent and extensive confounding adjustment in pharmacoepidemiological research using large health care databases. Such distinctive characteristics will continue to encourage further research and novel applications in the context of this methodology.

APPENDIX

AVAILABLE STATISTICAL PACKAGES FOR PROPENSITY SCORE METHODOLOGY

Due to the spread of PS methodologies in medical and epidemiological research, many software applications have been developed for their implementation. The majority of the available tools are written for Stata, SAS, and R [71-73]. Several authors illustrated their programs using data from LaLonde [74], and Dehejia and Wahba [75]. Main programs reviews can be found in Stuart [76] and Yang et al. [77].

PS estimation and balance checking. When implementing PS analyses, the researcher must first identify the set of variables to estimate the PS. Variable selection is usually based on a priori knowledge on the relationships between covariates and both the outcome and the treatment variables [20]. Rassen et al. developed a toolbox for SAS and R that implements automatic variable selection to estimate high-dimensional PS [34]. This tool is based on algorithms proposed by Schneeweiss et al. [26] and is particularly useful when dealing with large healthcare utilization databases [36]. After variable selection the researcher estimates the PS and verifies the balance of baseline characteristics across treatment groups. When the treatment is dichotomous, propensity scores are usually estimated through standard logistic regression. The code for PS estimation using SAS can be found in Leslie and Thiebaud [78] and in Kleinman and Horton [79].

Advanced applications that automatically estimate the PS and check the balance at baseline are also available. Becker and Ichino developed the program *pscore.ado* in Stata, which estimates the PS via a logit (or probit) procedure and checks the balancing property through an iterative algorithm [80]. First, PS quantiles are used to stratify subjects in blocks with equal average PS among treated subjects and controls. Second, t-tests are used to assess the equality of means of each covariate in treated and controls within each block. Balance checking is not carried out for higher moments of the covariate distribution. If the test rejects the balancing hypothesis, the user has to specify a less parsimonious PS model by adding squared terms and two-way interactions of covariates that were not balanced. A common support option is available when estimating the PS.

More advanced PS diagnostic tools are available in the *PSAgraphics* package for R, developed by Helmreich and Pruzek [81]. This package offers a wide variety of numerical and graphical tools to assess the balance for both categorical and continuous covariates in PS-based strata. For categorical covariates, the programs provide Fisher's exact test and display bars showing the proportion of subjects in different covariate-levels. For continuous covariates, it is possible to display pairs of box-plots for treated and control groups, and to check the equality of covariate distributions through the Kolmogorov-Smirnov statistic.

PS can be also estimated using models other than standard logistic regression. Non-parametric estimation techniques can be implemented using the *Twang* package for R which enables the user to estimate the PS via generalized boosted regression [82]. The iterative estimating process ends when the stopping rule (based on the Kolmogorov-Smirnov statistic or on the absolute standardized bias) is satisfied. For more details on underlying theoretical issues, see McCaffrey et al. [83]. In settings where multiple treatments are considered, multiple propensity scores can be estimated using the Twang package through the R function mnps (multinomial propensity score).

Implementation of PS-based techniques. After propensity scores are estimated, the researcher must choose the effect measure of interest. The most common effect measures are the average treatment effect (ATE) and the average treatment effect on the treated (ATT). ATE estimates the average difference in the outcomes comparing the setting in which everyone was treated to the setting in which no one was treated. ATT quantifies instead the average difference in the outcomes if everyone in the sample who was treated had been untreated. An analogous definition can be given for the average effect on the untreated (ATU) [84]. Many computing tools are available to integrate PS with matching, weighting, and stratification. There are no automatic procedures that implement regression using the PS as covariate, but examples of code for SAS and R can be found in Kleinman and Horton [79].

APPENDIX (CONTINUED)

AVAILABLE STATISTICAL PACKAGES FOR PROPENSITY SCORE METHODOLOGY

Matching. As described in the manuscript, matching is the most popular PS-based method. Thus, the majority of programs developed for PS analysis implement matching techniques. Though different matching algorithms have been proposed in the literature, greedy nearest-neighbor matching is the simplest and most common. Many variants of this procedure are available: with caliper, with replacement of control units, with many controls for each treated subject, and with variable ratio matching. Radius, kernel, local linear regression, and spline matching are other common algorithms. As an alternative, optimal matching was proposed by Rosembaum to obtain the matching configuration which minimizes the within-pair PS distance [47]. Some reviews of the main matching algorithms can be found in Stuart [76] and in Caliendo and Kopeinig [85].

Many macros that implement nearest-neighbor matching are available in SAS: Parsons and Coca-Perraillon developed, respectively, the *%OneToManyMTCH* and the *%PSMatching* macros, which allow greedy 1:n matching [86,87], Feng et al. developed the macro *%match*, which combines PS and Mahalanobis metric distance matching [88]. Kosanke and Bergstralh provided routines for the implementation of optimal matching with variable ratio [89]. The SAS toolkit developed by Rassen et al. includes software to implement different types of matching algorithms, including optimal 1:n matching with variable ratio [42]. Moreover, when three treatments need to be compared, the toolkit enables the researcher to perform 1:1:1 matching. Theoretical details can be found in Rassen et al. [44].

In Stata, Becker and Ichino developed the programs *attnd.ado* and attnw.ado for nearest-neighbor matching, *attr.ado* for radius matching and *attk.ado* for kernel matching [80]. These programs estimate ATT. Standard errors for these estimators can be obtained analytically (only for nearest neighbor and radius matching) or by bootstrapping. Note that if the user specifies a set of variables to estimate the PS, then bootstrap standard errors will encompass the variability due to the PS estimation process. In Stata also, Leuven and Sianesi developed the program *psmatch2.ado* to implement 1:n nearest-neighbor, radius, kernel, spline, and local linear regression matching [90]. This macro allows to estimate ATE, ATT, and ATU. In addition, it allows estimating the PS through a logistic model, but the user can specify a pre-computed score. Common support and trimming options are available. This macro includes *psgraph.ado*, which is used to display the PS distribution in treated and control groups and *pstest.ado*, which provides an assessment of the PS balancing property in the matched cohorts by computing t-tests, standardized bias, and other balance indexes before and after the matching procedure.

In R, Ho et al. developed the package *MatchIt*, which implements different types of matching algorithms (nearest-neighbor, optimal, full, and genetic) and offers several advanced numerical diagnostic indexes as alternatives to the traditional two-sample t-tests. Graphical tools for the assessment of achieved balance in the matched groups are also available [91]. A sophisticated procedure based on genetic algorithms was developed by Sekhon to achieve optimal balance in the matched cohorts. This algorithm is implemented in the R package *Matching*, which also provides many different indexes and statistics to assess balance degree [92].

Weighting. PS weighting can be implemented after PS estimation. Examples of code for SAS can be found in Leslie and Thiebaud [78]. The R package Twang performs weighting to estimate both ATT and ATE. The program displays tables and different types of plots to assess whether the weighting procedure was successful in balancing pre-treatment covariates in treatment and control groups.

Stratification. PS stratification can be implemented in Stata through the program *atts.ado* [80]. This procedure estimates ATT as a weighted average of stratum-specific effects. Standard errors can be computed either analytically or by bootstrapping. In R, PS stratification can be implemented through the *MatchIt* package [91]. This command performs stratification based on PS and allows an overall estimation of the ATE across strata. Balance checking within stratum is assessed for mean, squares, and two-way interactions of the covariates used to estimate PS. Balance can also be checked through graphical tools. Moreover, stratification can be used in conjunction with nearest-neighbor matching: in this case, nearest-neighbor matches are grouped in PS-based strata. Kleinman and Horton gave an example of the application of PS stratification in SAS and in R [79].

References

[1] Arana A, Rivero E, Egberts TCG. What do we show and who does so? An analysis of the abstracts presented at the 19th ICPE. Pharmacoepidemiol Drug Saf. 2004; 13: S330-1

[2] Schneeweiss S, Avorn J. Using health care utilization databases for epidemiologic research on therapeutics. J Clin Epidemiol. 2005; 58: 323-37

[3] Strom BL, Carson JL. Use of automated databases for Pharmacoepidemiology research. Epidemiol Rev. 1990; 12: 87-107

[4] Walker AM. Confounding by indication. Epidemiology. 1996; 7(4): 335-6

[5] Petri H, Urquhart J. Channeling bias in the interpretation of drug effects. Stat Med. 1991; 10(4): 577-81

[6] Walker AM. Observation and inference. An Introduction to the Methods of Epidemiology. Newton Lower Falls, MA: Epidemiology Resources Inc.; 1991: 119-28

[7] Schneeweiss S, Avorn J. A review of uses of health care utilization databases for epidemiologic research on therapeutics. J Clin Epidemiol. 2005; 58(4): 323-37

[8] Brookhart MA, Sturmer T, Glynn RJ, et al. Confounding control in healthcare database research: challenges and potential approaches. Med Care 2010; 48(6 Suppl): S114-20

[9] Seeger JD, Kurth T, Walker AM. Use of propensity score technique to account for exposure related covariates: an example and lesson. Med Care. 2007; 45(Supl 2): S143–8

[10] Greenland S, Pearl J, Robins JM. Causal diagrams for epidemiologic research. Epidemiology. 1999 Jan; 10(1): 37-48

[11] Hernan MA, Robins JM. Causal Inference (Chapman & Hall/CRC, 2013). http://www.hsph.harvard.edu/miguel-hernan/causal-inference-book/

[12] Peduzzi P, Concato J, Kemper E, et al. A simulation study of the number of events per variable in logistic regression analysis. J Clin Epidemiol. 1996; 49(12): 1373-9

[13] Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. Biometrika. 1983; 70(1): 41-55

[14] Glynn RJ, Schneeweiss S, Stürmer T. Indications for Propensity Scores and Review of Their Use in Pharmacoepidemiology. Basic Clin Pharmacol Toxicol. 2006; 98(3): 253-9

[15] Setoguchi S, Schneeweiss S, Brookhart MA, et al. Evaluating uses of data mining techniques in propensity score estimation: a simulation study. Pharmacoepidemiol Drug Saf 2008; 17: 546.5

[16] D'Agostino RB Jr. Propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. Stat Med 1998; 17(19): 2265-81

[17] Braitman LE, Rosenbaum PR. Rare outcomes, common treatments: analytic strategies using propensity scores. Ann Intern Med 2002; 137: 693-5

[18] Joffe MM, Rosenbaum PR. Propensity Scores. American Journal of Epidemiology 1999; 150:327-33

[19] Rubin DB. Estimating causal effects from large data sets using the propensity score. Ann Intern Med 1997; 127: 757-63

[20] Brookhart MA, Schneeweiss S, Rothman KJ, et al. Variable selection for propensity score models. Am J Epidemiol. 2006; 163(12): 1149-56

[21] Myers JA, Rassen JA, Gagne JJ, et al. Effects of Adjusting for Instrumental Variables on Bias and Precision of Effect Estimates. Am J Epidemiol. 2011; 174(11): 1213-22

[22] Charlson ME, Pompei P, Ales KL, et al. A new method of classifying prognostic comorbidity in longitudinal studies: development and validation. J Chronic Dis. 1987; 40: 373-83

[23] Romano PS, Roos LL, Jollis JG. Adapting a clinical comorbidity index for use with ICD-9-CM administrative data: differing perspectives. J Clin Epidemiol. 1993; 46: 1075-79, discussion 81–90

[24] Gagne JJ, Glynn RJ, Avorn J, et al. A combined comorbidity score predicted mortality in elderly patients better than existing scores. J Clin Epidemiol. 2011; 64(7): 749-59

[25] Schneeweiss S, Seeger JD, Maclure M, et al. Performance of comorbidity scores to control for confounding in epidemiologic studies using claims data. Am J Epidemiol. 2001; 154(9): 854-64

[26] Schneeweiss S, Rassen JA, Glynn RJ, et al. High-dimensional propensity score adjustment in studies of treatment effects using health care claims data. Epidemiology. 2009; 20(4): 512-22

[27] Anderson RM. Revisiting the behavioral model and access to medical care: Does it matter? J Health Social Behav. 1995; 36: 1-10

[28] Schneeweiss S, Glynn RJ, Avorn J, Solomon DH. A Medicare database review found that physician preferences increasingly outweighed patient characteristics as determinants of first-time prescriptions for COX-2 inhibitors. J Clin Epidemiol. 2005; 58: 98-102

[29] Roblin DW, Platt R, Goodman MJ, et al. Effect of increased cost-sharing on oral hypoglycemic use in five managed care organizations: how much is too much? Med Care. 2005; 43: 951-9

[30] Wooldridge, JM. Econometric Analysis of Cross Section and Panel Data. MIT Press; Cambridge, MA: 2001

[31] Greenland S. The effect of misclassification in the presence of covariates. Am J Epidemiol. 1980; 112: 564-9

[32] Greenland S, Robins JM. Confounding and misclassification. Am J Epidemiol. 1985; 122: 495-506

[33] Bross ID. Spurious effects from an extraneous variable. J Chronic Dis. 1966; 19(6): 637-47

[34] Rassen JA, Doherty M, Huang W, Schneeweiss S. Pharmacoepidemiology Toolbox. Boston, MA. http://www.hdpharmacoepi.org

[35] Arbogast PG, Kaltenbach L, Ding H, Ray WA. Adjustment for multiple cardiovascular risk factors using a summary risk score. Epidemiology. 2008; 19(1): 30-7

[36] Rassen JA, Glynn RJ, Brookhart MA, Schneeweiss S. Covariate selection in high-dimensional propensity score analyses of treatment effects in small samples. Am J Epidemiol. 2011; 173(12): 1404-13

[37] Rassen JA, Schneeweiss S. Using high-dimensional propensity scores to automate confounding control in a distributed medical product safety surveillance system. Pharmacol Drug Saf. 2012; 21(S1): 41-9

[38] Stürmer T, Joshi M, Glynn RJ, et al. 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. J Clin Epidemiol. 2006; 59(5): 437-47

[39] Schneeweiss S. A basic study design for expedited safety signal evaluation based on electronic healthcare data. Pharmacoepidemiol Drug Saf. 2010; 19: 858-68

[40] Rosenbaum PR, Rubin DB. Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. The American Statistician 1985; 39: 33-8

[41] Sturmer T, Rothman KJ, Avorn J, Glynn RJ. Treatment Effects in the Presence of Unmeasured Confounding: Dealing With Observations in the Tails of the Propensity Score Distribution - A Simulation Study. Am J Epidemiol 2010; 172: 843-54

[42] Rassen JA, Shelat AA, Myers JA, et al. One-to-many propensity score matching in cohort studies. Pharmacoepidemiol Drug Saf. 2012; 21(S2): 69-80

[43] Austin PC. Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples. Statist. Med. 2009; 28: 3083-3107

[44] Rassen JA, Shelat AA, Myers JA, Glynn RJ, Solomon DH, Schneeweiss S. Matching by Propensity Score in Cohort Studies with Three Treatment Groups. Epidemiology 2013; 24(3): 401-9

[45] Austin PC. An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies. Multivariate Behavioral Research 2011; 46: 399-424

[46] Hill J, Reiter JP. Interval estimation for treatment effects using propensity score matching. Statistics in Medicine 2006; 25: 2230-56

[47] Austin PC. Some Methods of Propensity-Score Matching had Superior Performance to Others: Results of an Empirical Investigation and Monte Carlo simulations. Biometrical Journal 2009; 51: 171-84

[48] Parsons L. Reducing bias in a propensity score matched-pair sample using greedy matching techniques. In: Proceedings of the 26th Annual SAS User Group International Conference. SAS Institute Inc, 2001. http://www2.sas.com/proceedings/sugi26/p214-26.pdf. Accessed May 8, 2010

[49] Rosenbaum PR. Observational Studies. Springer-Verlag, New York:1995

[50] Ming K, Rosenbaum PR. Substantial Gains in Bias Reduction from Matching with a Variable Number of Controls. Biometric 2000; 56S: 118-24

[51] Austin PC. Statistical Criteria for Selecting the Optimal Number of Untreated Subjects Matched to Each Treated Subject When Using Many-to-One Matching on the Propensity Score. Am J Epidemiol 2010; 172: 1092-7

[52] Gu XS, Rosenbaum PR. Comparison of multivariate matching methods: Structures, distances, and algorithms. Journal of Computational and Graphical Statistics 1993; 2: 405-20

[53] Hansen BB. Full matching in an observational study of coaching for the SAT. Journal of the American Statistical Association 2004; 99: 609-18

[54] Rosenbaum PR. A characterization of optimal designs for observational studies. Journal of the Royal Statistical Society 1991; Series B 53: 597-610

[55] Austin PC. Assessing balance in measured baseline covariates when using many-to-one matching on the propensity-score. Pharmacoepidemiol Drug Saf. 2008; 17: 1218-25

[56] Imbens G. The role of the propensity score in estimating dose-response functions. Biometrika. 2000; 87: 706-10

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

[58] Rosenbaum PR, Rubin DB. Reducing bias in observational studies using subclassification on the propensity score. Journal of the American Statistical Association 1984; 79: 516-24

[59] Huppler Hullsiek K, Louis TA. Propensity score modeling strategies for the causal analysis of observational data. Biostatistics 2002; 3: 179-93

[60] 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. Am J Epidemiol. 2006; 163(3): 262-70

[61] Lunt M, Solomon D, Rothman K, et al. Different methods of balancing covariates leading to different effect estimates in the presence of effect modification. Am J

Epidemiol. 2009; 169(7): 909-17
[62] van der Laan MJ, Rose S. Statistics ready for a revolution: next generation of statisticians must build tools for massive data sets. 2010 Amstat News
[63] Breiman L. Random forests. Machine Learning 2001; 45: 5-32
[64] Cross SS, Harrison RF, Kennedy RL. Introduction to neural networks. Lancet 1995; 346(8982): 1075-9
[65] Baxt WG. Application of artificial neural networks to clinical medicine. Lancet 1995; 346(8983): 1135-8
[66] Tibshirani R. Regression Shrinkage and Selection via the Lasso. Journal of the Royal Statistical Society 1996; Series B 58(1): 267-88
[67] Lee BK, Lessler J, Stuart EA. Improving propensity score weighting using machine learning. Statistics in Medicine 2010; 29: 337-46
[68] Lendle SD, Fireman B, van der Laan MJ. Targeted Maximum Likelihood Estimation in Safety Analysis. 2013; in press.
[69] van der Laan MJ, Rose S. Targeted Learning: Causal Inference for Observational and Experimental Data. New York: Springer; 2011
[70] van der Laan MJ, Polley EC, Hubbard AE. Super learner. Statistical applications in genetics and molecular biology. 2007; 6(1): 1-21
[71] StataCorp. 2011. Stata Statistical Software: Release 12. College Station, TX: StataCorp LP
[72] SAS Institute Inc. 2011. SAS 9.3. Cary, NC: SAS Institute, Inc
[73] R Development Core Team (2011). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org
[74] LaLonde R. Evaluating the Econometric Evaluations of Training Programs with Experimental Data. Am Econ Rev 1986; 76: 604-20
[75] Dehejia RH, Wahba S. Causal effects in non-experimental studies: Reevaluating the evaluation of training programs. J Am Statist Assoc 1999; 94: 1053-62
[76] Stuart EA. Matching methods for causal inference: a review and a look forward. Statist Sci 2010; 25(1): 1-21
[77] Yang G, Stemkowsky S, Saunders W. A review of propensity score application in healthcare outcome and epidemiology. Premier Inc. Working Paper PR02; 2007
[78] Leslie S, Thiebaud P. Using propensity scores to adjust for treatment selection bias. SAS Global Forum, Statistics and Data Analysis. Working Paper 184-2007
[79] Kleinman K, Horton NJ. Accessed May 28, 2013, at http://www.math.smith.edu/sasr
[80] Becker S, Ichino A. Estimation of average treatment effects based on propensity scores. The Stata Journal 2002; 2: 358-77
[81] Helmreich JE, Pruzek RM. PSAgraphics: An R package to support propensity score analysis. J Stat Softw 2009; 29(6): 1-23
[82] Ridgeway G, McCaffrey DF, Morral AR. Twang: Toolkit for weighting and analysis of nonequivalent groups. Software for using matching methods in R. 2006. Available at http://cran.rproject.org/src/contrib/Descriptions/twang.html
[83] McCaffrey DF, Ridgeway G, Morral AR. Propensity score estimation with boosted regression for evaluating causal effects in observational studies. Psychol Methods 2004; 9(4): 403-25
[84] Fang G, Brooks JM, Chrischilles EA. Apples and oranges? Interpretations of risk adjustment and instrumental variable estimates of intended treatment effects using observational data. Am J Epidemiol 2012; 175(1): 60-5
[85] Caliendo M, Kopeinig S. Some practical guidance for the implementation of propensity score matching. J Econ Surv 2005; 22: 31-72
[86] Parsons LS. Performing a 1:N Case-Control Match on Propensity Score. Proceedings of the Twenty-Sixth Annual SAS Users Group International Conference, Montreal, Canada. 2004
[87] Coca-Perraillon, M. Local and global optimal propensity score matching. SAS Global Forum, Statistics and Data Analysis. Working Paper 185-2007
[88] Feng W, Jun Y, Xu R. A Method/Macro Based on Propensity Score and Mahalanobis Distance to Reduce Bias in Treatment Comparison in Observational Study. SAS Technical Report 2006: 1-11. paper PR05
[89] Kosanke J, Bergstralh E. SAS Macro. Accessed May 28, 2013, at http://mayoresearch.mayo.edu/mayo/research/biostat/upload/vmatch.sas
[90] Leuven E, Sianesi B. PSMATCH2: Stata module to perform full Mahalanobis and propensity score matching, common support graphing, and covariate imbalance testing, Statistical Software Components S432001, Boston College Department of Economics. 2003
[91] Ho DE, Imai K, King G et al. MatchIt: Nonparametric preprocessing for parametric causal inference. J Stat Softw 2011; 42(8): 1-28
[92] Sekhon JS. Multivariate and propensity score matching software with automated balance optimization: the Matching package for R. J Stat Softw 2011; 42(7): 1-52

*