American Journal of Epidemiology
Vol. 181, No. 12
DOI: 10.1093/aje/kwu469
Advance Access publication:
May 20, 2015

Practice of Epidemiology

Propensity Score Methods for Analyzing Observational Data Like Randomized Experiments: Challenges and Solutions for Rare Outcomes and Exposures

Michelle E. Ross*, Amanda R. Kreider, Yuan-Shung Huang, Meredith Matone, David M. Rubin, and A. Russell Localio

* Correspondence to Dr. Michelle E. Ross, Department of Biostatistics and Epidemiology, Perelman School of Medicine, University of Pennsylvania, 423 Guardian Drive, Room 512, Philadelphia, PA 19104 (e-mail: michross@upenn.edu).

Initially submitted July 11, 2014; accepted for publication December 19, 2014.

Randomized controlled trials are the "gold standard" for estimating the causal effects of treatments. However, it is often not feasible to conduct such a trial because of ethical concerns or budgetary constraints. We expand upon an approach to the analysis of observational data sets that mimics a sequence of randomized studies by implementing propensity score models within each trial to achieve covariate balance, using weighting and matching. The methods are illustrated using data from a safety study of the relationship between second-generation antipsychotics and type 2 diabetes (outcome) in Medicaid-insured children aged 10–18 years across the United States from 2003 to 2007. Challenges in this data set include a rare outcome, a rare exposure, substantial and important differences between exposure groups, and a very large sample size.

confounding; discrete-time failure analysis; inverse probability of treatment weighting; marginal effects; observational study; propensity score matching; randomized experiments

Abbreviation: RCT, randomized controlled trial.

Randomized controlled trials (RCTs) are the "gold standard" for estimating the causal effects of treatments. However, it is often not feasible to conduct such a trial because of ethical concerns or budgetary constraints (1). In addition, when randomized studies are longitudinal and when subjects drop out, do not adhere to assigned treatment, or receive post-randomization ancillary treatments or exposures, the strong benefits of randomization dissipate (2–4). Without randomization, however, differences in the distribution of baseline covariates between the treated and untreated subjects can confound the comparison of outcomes between the treatment groups.

Observational studies are routinely used to estimate the causal effects of treatment on outcomes. However, there are often systematic differences in the distribution of baseline characteristics between treated and untreated participants. Hence, outcomes cannot be compared directly between treatment groups (5). With many large databases being readily available, proper analyses of observational data are becoming increasingly important.

One recent approach to the analysis of longitudinal observational studies of initiators of new treatments formulates the problem as a sequence of what we call "pseudo-randomized" experiments over time (6–8). Inclusion and exclusion criteria are developed to determine which subjects are eligible for each trial, where individuals can contribute to more than 1 trial. All trials are then pooled into a single analysis by using conventional covariate adjustment and robust variance estimators to account for within-person correlation.

In the case of rare outcomes and rare exposures, traditional covariate adjustment leads to sparse data cells, resulting in biased estimates and inflated variances. Propensity score methods offer an alternative to conventional regression adjustment, although they rely on assumptions that investigators must recognize and meet. These methods first model exposure in a preliminary model and then use the resulting probability of exposure in a response model. The propensity score can take the form of a covariate, be categorized into subclasses for a stratified analysis, be transformed into weights for standardization, or be used for different forms of matching. Each approach attempts to balance confounders between exposure groups, thus reducing bias (9). The choice of propensity score methods will depend upon the question being asked, the size of the data set, the number of possible confounders, and the prevalences of the exposure and outcome. For example, Bobo et al. (10) used propensity score–based matching in a situation where the exposed were more frequent than the unexposed, but they did not implement the sequence of trials we outline that makes formal use of the principles of causal inference to mimic a RCT.

In this paper, we begin by treating an observational cohort as a sequence of nonrandomized trials and then extend this approach with a 2-stage modeling process to estimate the effect of the exposure among the exposed. The 2-stage modeling process involves implementing propensity score models, using both weighting and matching, within each trial to achieve balance. This work was motivated by a study of the relationship between second-generation antipsychotics and incident type 2 diabetes in Medicaid-insured children aged 10–18 years across the United States from 2003 to 2007, using Medicaid Analytic eXtract (MAX) data from the Centers for Medicare and Medicaid Services (11). A unique child identifier across years facilitated longitudinal analysis. The scientific question was the change in risk of diabetes if a child with no prior or recent use of second-generation antipsychotics was newly prescribed these medications (assuming they adhered to their assigned treatment), compared with the risk had the child not been prescribed second-generation antipsychotics. Additional challenges included a very large sample size and substantial and important differences between comparison groups.

METHODS

Defining pseudo-trials within the cohort

We adopt the approach described by Hernán et al. (6, 7) and Danaei et al. (8) and developed in parallel by Schaubel et al. (12) and Kennedy et al. (13) in which we mimic all possible initiation trials ("pseudo-trials") over time, where each month represented the start of a new pseudo-trial, and pool the data. Children who initiated a second-generation antipsychotic ("initiators") in a given pseudo-trial start month were then compared with children who otherwise qualified for that pseudo-trial but did not initiate. Eligibility criteria must have been satisfied for each pseudo-trial.

To assess pseudo-trial eligibility, we included only children whose data allowed for the same look-back period in which to ascertain exclusion and inclusion criteria (as would be the case in a hypothetical RCT of second-generation antipsychotic use). For this reason, each child must have had 12 months of observed Medicaid enrollment (from the start of data collection) before he/she could be found eligible for the study. We included children aged 10 years or more at the start of the look-back period who had a behavioral or psychiatric diagnosis. Those with a prior diagnosis of diabetes or with prescriptions for diabetes drugs or second-generation antipsychotics were excluded.

Children were included until diagnosis of type 2 diabetes or censoring due to loss to follow-up or administrative censoring and development of type 1 diabetes. This design considered the initiator to be continuously exposed until censoring, regardless of treatment change, just as in a randomized study with an "as randomized" (intention-to-treat) analysis (6). The process defined a series of sequential trials in which children were "enrolled" at each month on the basis of their eligibility for that month and followed until the end of observation. Once the trials within cohorts were defined, they were stacked to create a single overall cohort.

Controlling for confounding

As in a RCT, we were interested in estimating the marginal difference in outcome between 2 "exchangeable" populations, that is, 2 populations that were identical in all respects except in their treatment status (5). On a theoretical level, covariate regression adjustment in the response model does not make this marginal comparison (5). Moreover, this approach fails to mimic the spirit of a RCT, in which the outcome is known and analyzed only after the comparison groups are fully formed and described (1). On a practical level, a conventional analysis, which requires storing and manipulating the entire design matrix of covariates, would be problematic in terms of both the data storage and the execution time, owing to the large sample. Hence, to address both theoretical goals and practical constraints, we implemented a 2-step analysis approach: first, using propensity scores to mimic the assignment process of a randomized experiment, 1 pseudo-trial at a time; and second, using weighting and matching in a response model on the pooled set of pseudo-trials, mimicking the approach of Hernán et al. (and others) to the analysis of observational data as a sequence of trials. We defined the standardizing population at each month as those who initiated treatment in that month (14).

We considered several other balancing options. One option was stratifying by quintiles of the estimated propensity score, which requires a separate propensity score model for each pseudo-trial and quintiles defined within pseudo-trials. However, this approach defines propensity score quintiles by pseudo-trial as a 250-level class variable, which leads to a hugely expanded data set. A second option, conditional logistic regression, avoids the problem of estimating all the coefficients; however, such models can experience convergence problems when the strata are large, as in our application. In addition, with a rare outcome, many strata would be uninformative and lead to a loss of observations. A third option, defining a single set of quintiles for the stacked pseudo-trials, would not support balance within pseudo-trials. We found all approaches to be flawed in theory or problematic in application. Moreover, propensity score matching and inverse probability weighting using the propensity score have been found to induce better balance on baseline covariates compared with stratification on the propensity score, based on Monte Carlo simulations (5, 15, 16).

Modeling the propensity score. We fit a prespecified propensity score model for each pseudo-trial to avoid the assumption of constant association of covariates and treatment assignment across pseudo-trials, which facilitates better balance of covariates across exposure groups within each trial. Our design more faithfully mimics a sequence of clinical trials over time, each with somewhat different populations randomized separately. Each model adjusted for baseline covariates present at the start of each pseudo-trial, including age, sex, race, census division, mental health diagnosis, prior diagnosis of a lipid or metabolic disorder or hypertension, history of complex chronic conditions, Medicaid eligibility status, and other psychotropic medication classes used. We first cross-classified these covariates with each other and then with treatment assignment by creating a series of contingency tables to guard against aliasing of covariates and violations of the positivity assumption (17).

Propensity score weighting. In the inverse probability of treatment weighting method, the primary analysis, we used weights based on the propensity score on the probability scale for the ith child, e_i, where exposed children received a weight of 1, and unexposed children received a weight of $e_i/(1 - e_i)$ (14, 18, 19). The standardizing population was then the set of exposed children and allowed for the estimation of the average treatment effect among the treated (20). All weighting was performed by using the program “logit” in Stata (StataCorp LP, College Station, Texas) (21).

Matching. Because of the large pool of unexposed children (noninitiators) compared with exposed children (initiators), matching was an appropriate choice for analysis. For that reason, we investigated the option of propensity score–based matching and compared results with those from the weighting method described above. We implemented a 1:1 matching scheme within pseudo-trials with 2 types of matching: optimal and greedy.

Optimal matching (17) considers all possible matches before assigning an unexposed subject to an exposed subject in order to minimize the overall distance (in terms of propensity score) between exposed and unexposed pairs. It is reproducible with a given program seed.

Greedy matching, by contrast, selects 1 exposed child at a time and finds the closest available match (22). Subsequently selected children are matched to whichever unexposed children remain. Hence, depending on the order in which exposed children are selected, the matching program can select a different group of matches with each execution. This feature of greedy matching is especially problematic in our setting, because of the large pool of available unexposed children.

We implemented nearest-neighbor matching using a caliper of 0.25 standard deviations of the propensity score (on the log odds scale). To test greedy matching, we experimented with 4 different data set orders to examine the sensitivity of the response model to the matched data. All matching was performed using the program `MatchIt` in `R` (23, 24).

Assessing balance. In the weighted data set, checks for balance were performed within each pseudo-trial as well as across the stacked pseudo-trials. Conventional methods for balance checking, such as standardized differences or 2-by-k contingency tables, do not apply to weighted data. We dismissed the option of weighted contingency tables because of long execution times with our data set. Hence, we adopted a weighted regression approach in which each covariate was regressed on the treatment variable. We required small weighted differences between treatment groups for each covariate and developed threshold values that corresponded to clinically unimportant differences between treatment groups. For binary and categorical covariates, a threshold value of 0.2 was selected. For example, for a binary covariate with reference value 20%, a difference of 0.2 corresponds to a value of 23% in the comparison group.

For the matched data sets, we used standard balance checks. We elected to check for balance in the stacked pseudo-trials only. As all covariates included in the propensity score model were discrete, the proportions in each level of the covariate were compared in the 2 treatment groups using 2-by-k contingency tables.

Response model

As we used an intention-to-treat analysis, we did not account for treatment stoppage or switching. One drawback of this approach, in the context of measuring the risk of adverse events, is that it fails to consider changes in risk due to treatment switching. For example, once a child stops treatment, he/she may be at lower risk of the adverse event, and similarly, a child who initially is not on treatment but switches might have a change in risk. Hence, the intention-to-treat analysis can understate the association of exposure and outcome.

The stacked pseudo-trials were the basis for our analyses, where time-to-event methods (such as Cox proportional hazards models), accelerated failure-time analysis, or discrete-time failure analysis could be used. Time-to-event methods and accelerated failure-time models proved to be beyond our computing capacity with our data set. Hence, we pursued discrete-time failure models, which were fit to both the weighted and matched data sets by using pooled logistic regression (25). Models included main effects for treatment and prespecified time periods: 0–6 months, 7–12 months, 13–18 months, 19–24 months, and 25 or more months. Sensitivity analyses were performed using models that additionally included interactions between time and treatment. Because children could contribute to multiple pseudo-trials, all analyses used robust variance estimates to account for repeated observations of the same child to multiple pseudo-trials, as recommended in Hernán et al. (6).

Analyses were performed using SAS, version 9.3 (SAS Institute, Inc., Cary, North Carolina); Stata/MP, version 13.1; and R, version 3.0 (R Foundation for Statistical Computing, Vienna, Austria), with 64-bit applications on 2 servers, the first with 2 Intel Xeon (Intel Corporation, Santa Clara, California) 2.2 MHz processors with 256 GB of random access memory (RAM) and the second with 2 Xeon 2.6 MHz processors with 8 GB of RAM.

RESULTS

The analysis data set

The data consisted of 1,958,808 subjects, of which 1,328,985 were eligible for at least 1 pseudo-trial out of a total of 50 pseudo-trials. Of these eligible children, 107,551 ever initiated second-generation antipsychotics, while 1,221,434 never initiated second-generation antipsychotics, during the observation period. Children could be eligible for multiple pseudo-trials, provided the inclusion and exclusion criteria were satisfied at different observation months. For example, a noninitiating child eligible for the first pseudo-trial could be

Table 1. Demographic and Clinical Characteristics of Eligible Children at First Eligible Pseudo-Trial by Drug Initiation Status[a], US Medicaid, 2003–2007

Characteristic	Ever SGA Initiators (*n* = 107,551), %	Never SGA Initiators (*n* = 1,221,434), %
Sex		
Male	60.9	56.1
Female	39.1	43.9
Age group		
Younger (10–14 years)	59.8	65.2
Older (15–18 years)	40.2	34.8
Race		
White	56.0	53.8
Black	25.8	26.6
Hispanic	9.7	11.7
Other	2.7	3.1
Unknown	5.9	4.8
Eligibility group		
Foster care	23.4	14.0
TANF/other	49.6	68.7
SSI	27.0	17.3
Psychiatric diagnoses		
ADHD	53.1	32.1
Autism	7.6	2.6
Anxiety disorder	11.3	7.4
Bipolar disorder	18.9	2.6
Conduct disorder	45.0	19.3
Depression	39.8	17.0
Developmental delay	14.7	16.1
Intellectual disability	10.1	6.9
Other diagnoses		
Metabolic disorder, lipid disorder, or hypertension	4.0	3.0
Complex chronic condition	20.3	13.0
Other psychotropic medication classes		
Stimulant	44.6	23.8
Antidepressant	49.2	13.3
Mood stabilizer	22.8	4.0
α-Agonist	12.0	3.0
Anxiolytic	3.5	1.5
Sedative/hypnotic	1.7	0.4
Census division		
Division 1	2.8	4.3
Division 2	10.4	7.2
Division 3	20.6	19.7
Division 4	7.4	5.8
Division 5	23.1	27.1
Division 6	7.4	7.3
Division 7	18.3	17.1
Division 8	2.8	3.0
Division 9	7.2	8.4
Diabetes[b]	0.4	0.2

Abbreviations: ADHD, attention deficit hyperactivity disorder; SGA, second-generation antipsychotic; SSI, Supplemental Security Income; TANF, Temporary Assistance for Needy Families.

[a] The first eligible pseudo-trial is defined as the first month of eligibility with the specified status (SGA initiator or noninitiator). For ever initiators, this will be the baseline month for the first pseudo-trial in which they initiate. For never initiators, this will be their first month of eligibility. For all variables, we take the status in the given month, with the exception of diabetes.

[b] Proportion of children who go on to develop type 2 diabetes.

included in the second pseudo-trial as an initiator if he/she started second-generation antipsychotics in the second month of observation, assuming all other eligibility criteria were satisfied. When the pseudo-trials were formed, the resulting data set contained 446,360,161 child-month observations, each weighted for use in the response model of the association of second-generation antipsychotic use and diabetes.

Table 1 displays demographic and clinical characteristics for all eligible children at the start of their first eligible pseudo-trial. For "ever" initiators, this was defined as the baseline month of the first pseudo-trial in which they both were eligible and initiated second-generation antipsychotics. For "never" initiators, this was their first month of eligibility. The ever initiators had slightly higher proportions of males and older children. They had substantially more foster care and Supplemental Security Income children and a higher proportion with complex chronic conditions. In general, the ever initiators had much higher proportions of psychiatric diagnoses and of other psychotropic medication class usage. The proportion of ever initiators who went on to develop type 2 diabetes was double that of never initiators, although the proportion was relatively small (0.4%).

The number of initiators in each pseudo-trial ranged from 1,485 to 3,116 and totaled 110,985 across all trials. For each pseudo-trial, we performed a series of unweighted regressions, treating the baseline covariates of interest as the outcome and using treatment as the exposure to assess the degree of imbalance prior to implementing propensity score methods. In general, coefficients varied substantially across pseudo-trials and rarely fell below the chosen threshold to indicate balance. For many covariates, the coefficient estimates switched signs, and the range of values for the estimates tended to be quite large (as much as 1.01) across pseudo-trials. Coefficients for psychiatric diagnoses and other psychotropic medication class usage were particularly large—often around 1.0—with some as large as 2.2 on the log odds scale, indicating a large degree of imbalance.

The weighted analysis

With over 400 million observations and without access to huge computing resources, we were somewhat limited in our software choice for the weighted analysis. We proceeded in SAS, as neither R nor Stata/MP could handle the data set because these programs store data in central memory. We ran the analysis using the SURVEYLOGISTIC procedure.

For each pseudo-trial, observations were reweighted so that the sum of the weights in the unexposed group equaled the sample size in the exposed group by multiplying the weights in the unexposed group by a constant, a minimal adjustment. In the assessment of balance within pseudo-trials, all coefficients in the weighted regression models were substantially smaller than 0.20 in absolute value. The same was true in the stacked pseudo-trials.

The matched analysis

We attempted optimal matching within each pseudo-trial without success. The algorithm stores a matrix of distances of all possible matches within pseudo-trials in order to minimize the overall distance among matched pairs. Our data set was too large to store such a matrix using our available computing resources.

Greedy matching, by contrast, was possible even with stringent caliper constraints. In 3 of the 4 matched data sets, we were unable to match 52 initiators out of a total of 110,985 initiators, while we were unable to match 53 initiators in the fourth matched data set.

In all of the matched data sets, the distribution of proportions of the potential confounders was balanced between initiators and noninitiators. In general, the proportions in the 2 treatment groups differed by less than 2%. There was a slightly higher proportion of children with mood and bipolar disorders in the initiator group than in the noninitiator group in all 4 matched data sets. The difference between the proportions for these 2 disorders was approximately 4.2% and 4.3%, respectively, in each matched data set.

Type 2 diabetes

The results of the matched and weighted analyses agree well; however, there was some variability in the matched results (Table 2). The estimated log odds ratio of diabetes in a given month comparing initiators with noninitiators was 0.41 (95% confidence interval: 0.30, 0.52) from the weighted analysis. The estimated log odds ratio from the matched analyses ranged from 0.29 (95% confidence interval: 0.14, 0.44) to 0.42 (95% confidence interval: 0.26, 0.58) by comparison. There was no evidence of an interaction between time and treatment in either the weighted or matched analyses.

Table 2. Estimated Log Odds Ratios and 95% Confidence Intervals for Type 2 Diabetes From Weighted and Matched Analyses Comparing Initiators With Noninitiators, US Medicaid, 2003–2007

Method[a]	Log OR	95% CI
Weighting	0.41	0.30, 0.52
Matching 1	0.29	0.14, 0.44
Matching 2	0.38	0.22, 0.53
Matching 3	0.29	0.14, 0.44
Matching 4	0.42	0.26, 0.58

Abbreviations: CI, confidence interval; OR, odds ratio.

[a] We implemented nearest neighbor matching on 4 different orderings of the data using 1:1 matching with a caliper of 0.25 standard deviations of the propensity score (on the log odds scale). Each different ordering of the data set resulted in a different set of matched pairs (“Matching 1,” “Matching 2,” “Matching 3,” “Matching 4”). We also implemented inverse probability of treatment weighting with weights based on the propensity score and using the set of exposed children as the standardizing population.

DISCUSSION

We have described an approach for the analysis of large observational data sets to mimic randomized studies, which combines several methods. Traditional regression adjustment would not only neglect this important randomization paradigm but, in the case of this large data set, would have been problematic because of changing numbers of initiators and their varying characteristics over the observation period. Our experiences and approaches will likely be replicable with large observational data sets.

Large databases, such as The Health Improvement Network (THIN), Medicare, and Medicaid, and the advent of research networks, such as The National Patient-Centered Clinical Research Network (PCORnet), bring opportunities for the analysis of rare outcomes. With those opportunities, however, come challenges. Big data might be “the next big thing in epidemiology” (26, p. 351), but the sheer size of the data set does not overcome the potential for bias from confounding.

Matching has become a routine method for confounding control in spite of its well-known limitations in the analysis of observational data (27). When a sequence of pseudo-trials is used, matching ideally should occur within, rather than across, trials. The reason is simple: Each pseudo-trial should mimic its own RCT, but the sheer size of the data set limited our matching options. Optimal matching would have guaranteed the same matched data set with repeated use, but the data set size exceeded the capabilities of a leading matching program. On the other hand, greedy matching was unable to find suitable matches for all children and led to a small loss of treated children in the sample. This loss of treated children preserved close matches and thus produced estimates with minimal bias. However, large losses relative to the sample size could impair the generalizability of findings.

We implemented inverse probability of treatment weighting in addition to matching, because each method has its strengths, and the use of both adds robustness to the findings. Inverse probability weighting and matching were also convenient choices due to the size of the data set. Although results from the matched studies varied somewhat, the results from the weighted and matched analyses agreed well. Exposed children with a very low propensity score or unexposed children with a high propensity score lead to large weights (5). In this case, stabilized or trimmed weights can be used to address potential instability due to very large weights (28, 29).

The size of our data set, and of many data sets to come, contraindicates conventional regression adjustment. Assembling covariates for an otherwise long data set challenges software. Our data set had over 400 million rows because months were repeated within pseudo-trials, and each child could qualify for multiple pseudo-trials. Propensity-score development, testing within the pseudo-trial, and matching break the computing problem down into tractable pieces.

However, propensity score methods are not always appropriate, and the choice of method for controlling for confounding must take into account the scientific question of interest. We found the 2-stage approach of propensity scores superior to traditional regression adjustment because of our interest in a specific estimate: the effect of medication in the population of children taking second-generation antipsychotics (treatment effect among the treated). By separating the 2 modeling exercises, we find that this approach also mimics RCTs, where patients are first stratified and randomized to make the treatment groups exchangeable and then followed for their response.

The size of the weighted data set became prohibitive at times. Execution times were substantial, even for simple descriptive statistics and analyses. Performing these tasks across 50 pseudo-trials meant exceedingly long computation times. However, the weighted analysis is completely reproducible, and all individuals in the data set are included in the analysis. On the other hand, greedy matching resulted in more tractable computation times, even for very large samples, which may be preferable for some. In this case, investigators should use several matched sets to explore the sensitivity of the response model to the matches, especially when the pool of potential matches is large and the outcome is rare. Optimal matching is advantageous because it produces the same set of matched pairs for a given program seed; however, by its very nature, it requires considerable computing resources.

Our study has several limitations. First, limited observational data cannot replace randomization to control for bias. In particular, RCTs balance on both measured and unmeasured confounders. Here, we have assumed there are no unmeasured confounders. Additional patient-level covariates can reduce bias, provided they offer additional information about treatment assignment choices and outcome. However, when data sets are large, additional covariates will compound some of the data storage and program execution problems we report. In addition, although we developed separate propensity scores and then matched patients on those scores within the pseudo-trial, additional strata for model development and matching might be indicated. For example, with large data sets from multiinstitutional settings, matching within institution and within pseudo-trial might be warranted if the goal is to mimic a sequence of stratified RCTs. Although we were able to fit propensity score models and find matches in our setting of a low prevalence of second-generation antipsychotic use, these methods are not a panacea for problems of rare exposures. Propensity score models must estimate expected probabilities of exposure that lie within the range (0,1) and not on the boundaries (the positivity assumption), and the expected probabilities among the exposed and unexposed must overlap to some degree to support matching. Finally, we were unable to study the performance of greedy matching in contrast to optimal matching. The performance of greedy matching might depend heavily on the distributions of propensity scores among potential controls, the number of available matches, the matching frequency, and the prevalence of the outcome of interest among potential controls. Moreover, we cannot speculate, at this point, as to the number of different sorting patterns that the investigator may need to try.

This paper focused on a design that mimics a RCT and estimates the effect of treatment as randomized. If treatment changes, either due to treatment switching or stoppage, the analysis methods we outlined might not answer the research question of interest. Nevertheless, our reported approach using weighting and/or matching followed by a discrete-time failure response model will lay the foundation for estimating the effect of exposure "per protocol" (the effect of exposure assuming a child were to continue indefinitely on his or her initiated treatment) or the effect "as treated" (allowing for time-varying treatment) (6). Both sets of estimates can follow from our approach by means of second-stage weighting after propensity score weighted approaches or by means of weighted analyses of matched data sets. For that reason, the methods we described build upon existing designs for mimicking adherence-adjusted RCTs.

ACKNOWLEDGMENTS

Author affiliations: Department of Biostatistics and Epidemiology, Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania (Michelle E. Ross, A. Russell Localio); PolicyLab, The Children's Hospital of Philadelphia, Philadelphia, Pennsylvania (Amanda R. Kreider, Meredith Matone, David M. Rubin); Healthcare Analytics Unit, The Children's Hospital of Philadelphia, Philadelphia, Pennsylvania (Yuan-Shung Huang); and Department of Pediatrics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania (David M. Rubin).

This work was supported by grant 5R01HS018550 from the Agency for Healthcare Research and Quality.

Conflict of interest: none declared.

REFERENCES

1. Rubin DB. The design versus the analysis of observational studies for causal effects: parallels with the design of randomized trials. *Stat Med*. 2007;26(1):20–36.
2. White IR, Carpenter J, Horton NJ. Including all individuals is not enough: lessons for intention-to-treat analysis. *Clin Trials*. 2012;9(4):396–407.
3. Hernán MA, Hernández-Díaz S, Robins JM. Randomized trials analyzed as observational studies. *Ann Intern Med*. 2013; 159(8):560–562.
4. Shrier I, Steele RJ, Verhagen E, et al. Beyond intention to treat: What is the right question? *Clin Trials*. 2014;11(1):28–37.
5. Austin PC. The use of propensity score methods with survival or time-to-event outcomes: reporting measures of effect similar to those used in randomized experiments. *Stat Med*. 2014; 33(7):1242–1258.

6. Hernán MA, Robins JM, García Rodríguez LA. Discussion on: "Statistical issues arising in the Women's Health Initiative." *Biometrics*. 2005;61(4):922–930.
7. 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*. 2008;19(6):766–779.
8. Danaei G, Rodríguez LA, Cantero OF, et al. Observational data for comparative effectiveness research: an emulation of randomised trials of statins and primary prevention of coronary heart disease. *Stat Methods Med Res*. 2013;22(1): 70–96.
9. 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):2265–2281.
10. Bobo WV, Cooper WO, Stein CM, et al. Antipsychotics and the risk of type 2 diabetes mellitus in children and youth. *JAMA Psychiatry*. 2013;70(10):1067–1075.
11. Rubin DM, Kreider AR, Matone M, et al. Risk for incident diabetes mellitus following initiation of second-generation antipsychotics among Medicaid-enrolled youths. *JAMA Pediatr*. 2015;169(4):e150285.
12. Schaubel DE, Wolfe RA, Port FK. A sequential stratification method for estimating the effect of a time-dependent experimental treatment in observational studies. *Biometrics*. 2006;62(3):910–917.
13. Kennedy EH, Taylor JM, Schaubel DE, et al. The effect of salvage therapy on survival in a longitudinal study with treatment by indication. *Stat Med*. 2010;29(25): 2569–2580.
14. Sato T, Matsuyama Y. Marginal structural models as a tool for standardization. *Epidemiology*. 2003;14(6):680–686.
15. Austin PC. The relative ability of different propensity score methods to balance measured covariates between treated and untreated subjects in observational studies. *Med Decis Making*. 2009;29(6):661–677.
16. Lunceford JK, Davidian M. Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. *Stat Med*. 2004;23(19):2937–2960.
17. Stuart EA. Matching methods for causal inference: a review and a look forward. *Stat Sci*. 2010;25(1):1–21.
18. Rosenbaum PR. Model-based direct adjustment. *J Am Stat Assoc*. 1987;82(398):387–394.
19. Vittinghoff E. *Regression Methods in Biostatistics Linear, Logistic, Survival, and Repeated Measures Models*. New York, NY: Springer; 2012.
20. Williamson E, Morley R, Lucas A, et al. Propensity scores: from naive enthusiasm to intuitive understanding. *Stat Methods Med Res*. 2012;21(3):273–293.
21. StataCorp LP. *Stata Statistical Software, Release 13*. College Station, TX: StataCorp LP; 2013.
22. Hansen BB. Full matching in an observational study of coaching for the SAT. *J Am Stat Assoc*. 2004;99(467):609–618.
23. Ho DE, Imai K, King G, et al. MatchIt: nonparametric preprocessing for parametric causal inference. *J Stat Softw*. 2011;42(8).
24. R Development Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing; 2008.
25. D'Agostino RB, Lee ML, Belanger AJ, et al. Relation of pooled logistic regression to time dependent Cox regression analysis: the Framingham Heart Study. *Stat Med*. 1990;9(12):1501–1515.
26. Toh S, Platt R. Is size the next big thing in epidemiology? *Epidemiology*. 2013;24(3):349–351.
27. Smith JA, Todd PE. Does matching overcome LaLonde's critique of nonexperimental estimators? *J Econom*. 2005; 125(1-2):305–353.
28. Cole SR, Hernán MA. Adjusted survival curves with inverse probability weights. *Comput Methods Programs Biomed*. 2004; 75(1):45–49.
29. Lee BK, Lessler J, Stuart EA. Weight trimming and propensity score weighting. *PLoS One*. 2011;6(3):e18174.