Using Non-experimental Data to Estimate Treatment Effects

Elizabeth A. Stuart, PhD; Sue M. Marcus, PhD; Marcela V. Horvitz-Lennon, MD;
Robert D. Gibbons, PhD; Sharon-Lise T. Normand, PhD; and C. Hendricks Brown, PhD

Psychiatric Annals, Volume 39, Issue 7, July 2009

CME EDUCATIONAL OBJECTIVES

1. Define the strengths and weaknesses of observational studies.
2. List methods for the identification and control of bias in observational studies.
3. Recognize propensity score matching methods.

ABOUT THE AUTHOR

Elizabeth A. Stuart, PhD, is with Johns Hopkins Bloomberg School of Public Health, Baltimore. Sue M. Marcus, PhD, is with Mount Sinai School of Medicine, New York. Marcela V. Horvitz-Lennon MD, is with the Department of Psychiatry, University of Pittsburgh School of Medicine. Robert D. Gibbons, PhD, is with the Center for Health Statistics, University of Illinois at Chicago. Sharon-Lise T. Normand, PhD, is with the Department of Health Care Policy, Harvard Medical School, and Department of Biostatistics, Harvard School of Public Health, Boston. C. Hendricks Brown, PhD, is with the Department of Epidemiology and Public Health, University of Miami.

Address correspondence to: Elizabeth A. Stuart, PhD, Johns Hopkins Bloomberg School of Public Health, 624 N Broadway, Eighth Floor, Baltimore, MD, 21205; fax 410-955-9088; or e-mail estuart@jhsph.edu.

Dr. Stuart has disclosed the following relevant financial relationships: Center for Prevention and Early Intervention, jointly funded by the National Institute of Mental Health (NIMH) and the National Institute on Drug Abuse (Grant MH066247; PI: N. Ialongo), and NIMH grant K25-MH083846: research grant recipient. Dr. Marcus has disclosed no relevant financial relationships. Dr. Horvitz-Lennon has disclosed the following relevant financial relationships: research grant recipient (NIMH Grant P50-MH073469). Dr. Gibbons has disclosed the following relevant financial relationship: research grant recipient (NIMH Grant R56-MH078580 and R01-MH8012201). Dr. Normand has disclosed the following relevant financial relationship: research grant recipient (NIMH Grant MH61434). Dr. Brown has disclosed the following relevant financial relationship: research grant recipient (NIMH Grant R01-MH040859).

The authors are indebted to Larry Zaborski, MS, Harvard Medical School, for earlier programming help, and to Richard Frank, PhD, Harvard Medical School, for generously providing the Medicaid data. Dr. Zaborski and Dr. Frank have disclosed no relevant financial relationships.

doi: 10.3928/00485713-20090625-07

PARTICIPANT ATTESTATION

___ I certify that I have read the article(s) on which this activity is based, and claim credit commensurate with the extent of my participation.

COMMERCIAL BIAS EVALUATION

Please rate the degree to which the content presented in this activity was free from commercial bias. No bias Significant bias
5 4 3 2 1

Comments regarding commercial bias: ________________________________
____

INSTRUCTIONS

1. Review the stated learning objectives of the CME articles and determine if these objectives match your individual learning needs.

2. Read the articles carefully. Do not neglect the tables and other illustrative materials, as they have been selected to enhance your knowledge and understanding.

3. The following quiz questions have been designed to provide a useful link between the CME articles in the issue and your everyday practice. Read each question, choose the correct answer, and record your answer on the CME REGISTRATION FORM at the end of the quiz. Retain a copy of your answers so that they can be compared with the correct answers should you choose to request them.

4. Type your full name and address and your date of birth in the space provided on the CME REGISTRATION FORM.

5. Complete the evaluation portion of the CME REGISTRATION FORM. Forms and quizzes cannot be processed if the evaluation portion is incomplete. The evaluation portion of the CME REGISTRATION FORM will be separated from the quiz upon receipt at PSYCHIATRIC ANNALS. Your evaluation of this activity will in no way affect the scoring of your quiz.

6. Your answers will be graded, and you will be advised whether you have passed or failed. Unanswered questions will be considered incorrect. A score of at least 80% is required to pass. Your certificate will be mailed to you at the mailing address provided. Upon receiving your grade, you may request quiz answers. Contact our customer service department at (856) 994-9400.

7. Be sure to complete the CME REGISTRATION FORM on or before July 31, 2010. After that date, the quiz will close. Any CME REGISTRATION FORM received after the date listed will not be processed.

8. This activity is to be completed and submitted online only.

Indicate the total time spent on the activity (reading article and completing quiz). Forms and quizzes cannot be processed if this section is incomplete. All participants are required by the accreditation agency to attest to the time spent completing the activity.

CME ACCREDITATION

This CME activity is primarily targeted to patient-caring physicians specializing in psychiatry. There are no specific background requirements for participants taking this activity. Learning objectives are found at the beginning of each CME article.

This activity has been planned and implemented in accordance with the Essential Areas and policies of the Accreditation Council for Continuing Medical Education through the joint sponsorship of Vindico Medical Education and PSYCHIATRIC ANNALS. Vindico Medical Education is accredited by the ACCME to provide continuing medical education for physicians.

Vindico Medical Education designates this educational activity for a maximum of 3 *AMA PRA Category 1 Credits™.* Physicians should only claim credit commensurate with the extent of their participation in the activity.

FULL DISCLOSURE POLICY

In accordance with the Accreditation Council for Continuing Medical Education's Standards for Commercial Support, all CME providers are required to disclose to the activity audience the relevant financial relationships of the planners, teachers, and authors involved in the development of CME content. An individual has a **relevant financial relationship** if he or she has a financial relationship in any amount occurring in the last 12 months with a commercial interest whose products or services are discussed in the CME activity content over which the individual has control. Relationship information appears at the beginning of each CME-accredited article in this issue.

UNLABELED AND INVESTIGATIONAL USAGE

The audience is advised that this continuing medical education activity may contain references to unlabeled uses of FDA-approved products or to products not approved by the FDA for use in the United States. The faculty members have been made aware of their obligation to disclose such usage.

HOW TO OBTAIN CME CREDITS BY READING THIS ISSUE

This CME activity is primarily targeted to patient-caring physicians specializing in psychiatry. Physicians can receive *AMA PRA Category 1 Credits™* by reading the CME articles in Psychiatric Annals and successfully completing the quiz at the end of the articles. Complete instructions are given subsequently. Educational objectives are found at the beginning of each CME article.

CME ACCREDITATION

This activity has been planned and implemented in accordance with the Essential Areas and policies of the Accreditation Council for Continuing Medical Education through the joint sponsorship of Vindico Medical Education and Psychiatric Annals. Vindico Medical Education is accredited by the ACCME to provide continuing medical education for physicians.

Vindico Medical Education designates this educational activity for a maximum of 3 *AMA PRA Category 1 Credits™*. Physicians should only claim credit commensurate with the extent of their participation in the activity.

FULL DISCLOSURE POLICY

In accordance with the Accreditation Council for Continuing Medical Education's Standards for Commercial Support, all CME providers are required to disclose to the activity audience the **relevant financial relationships** of the planners, teachers, and authors involved in the development of CME content. An individual has a relevant financial relationship if he or she has a financial relationship in any amount occurring in the last 12 months with a commercial interest whose products or services are discussed in the CME activity content over which the individual has control. Relationship information appears at the beginning of each CME-accredited article in this issue.

UNLABELED AND INVESTIGATIONAL USAGE

The audience is advised that this continuing medical education activity may contain references to unlabeled uses of FDA-approved products or to products not approved by the FDA for use in the United States. The faculty members have been made aware of their obligation to disclose such usage.

Financial Disclosures: Stanley Caroff, MD, has disclosed no relevant financial relationships. John M. Davis, MD, has disclosed no relevant financial relationships. Jan Fawcett, MD, has disclosed the following relevant financial relationships: Merck: Member of Editorial Advisory Board, Merck Manual. Paula Hensley, MD, has disclosed no relevant financial relationships. Andrew A. Nierenberg, MD, has disclosed the following relevant financial relationships: Abbott, Brain Cells Inc., Bristol Myers Squibb (BMS), Eli Lilly, GlaxoSmithKline (GSK), Innapharma, PGx, Ortho-McNiel Janssen, Novartis, Pfizer, Sepracor, Shire, and Somerset: Member of Advisory Board/Consultant: Massachusetts General Hospital (MGH) Structured Interview Guide to the Montgomery-Åsberg Rating Scale (MADRS), and the Clinical Positive Affect Scale, licensed exclusively to the MGH Clinical Trials Network and Institute. Copyright Holder: BMS, Cederroth, Cyberonics, Eli Lilly, Forest, GSK, Ortho-McNeil Janssen; Lichtwer, NARSAD, National Institute of Mental Health; Pfizer, Stanley Foundation, Wyeth: Research Grant Recipient; and BMS, Cyberonics, Eli Lilly, Forest, GSK, Wyeth: Member of Speakers' Bureau. The staff of *Psychiatric Annals* have disclosed no relevant financial relationships.

INSTRUCTIONS

1. Review the stated learning objectives of the CME articles and determine if these objectives match your individual learning needs.

2. Read the articles carefully. Do not neglect the tables and other illustrative materials, as they have been selected to enhance your knowledge and understanding.

3. The following quiz questions have been designed to provide a useful link between the CME articles in the issue and your everyday practice. Read each question, choose the correct answer, and record your answer on the CME REGISTRATION FORM at the end of the quiz. Retain a copy of your answers so that they can be compared with the correct answers should you choose to request them.

4. Type or print your full name and address and your date of birth in the space provided on the CME REGISTRATION FORM.

5. Complete the evaluation portion of the CME REGISTRATION FORM. Forms and quizzes cannot be processed if the evaluation portion is incomplete. The evaluation portion of the CME REGISTRATION FORM will be separated from the quiz upon receipt at PSYCHIATRIC ANNALS. Your evaluation of this activity will in no way affect the scoring of your quiz.

6. Send the completed form, with your $25 payment (check, money order, or credit card information) to: VINDICO MEDICAL EDUCATION, PO Box 36, Thorofare NJ 08086. Payment should be made in US dollars drawn on a US bank.

7. Your answers will be graded, and you will be advised whether you have passed or failed. Unanswered questions will be considered incorrect. A score of at least 80% is required to pass. Upon receiving your grade, you may request quiz answers. Contact our customer service department at (856) 994-9400.

8. Be sure to mail the CME REGISTRATION FORM on or before the deadline listed. After that date, the quiz will close. Any CME REGISTRATION FORM received after the date listed will not be processed.

Indicate the total time spent on the activity (reading article and completing quiz). Forms and quizzes cannot be processed if this section is incomplete. All participants are required by the accreditation agency to attest to the time spent completing the activity.

EDUCATIONAL OBJECTIVES OVERVIEW

If any readers of *Psychiatric Annals* read any psychiatric articles in any journal so that they can keep up to date and use the best-available evidence in their clinical practice, then the statistical papers in this issue are essential not only to read, but also to re-read, study, and master. What do these excellent statistical papers have to do with clinical psychiatric practice? They contain the tools that will allow you to critically appraise and interpret the psychiatric literature. They demystify the statistics used to generate the evidence and help you to become statistically literate. They give you the tools to turn data into information and knowledge.

TABLE OF CONTENTS

RESPONSIBILITY FOR STATEMENTS

All opinions expressed by authors and quoted sources are their own and do not necessarily reflect the opinions of the editors, publishers, or editorial boards of *Psychiatric Annals* or its employees, Vindico Medical Education or its employees, or the University of New Mexico. The acceptance of advertising in no way implies endorsement by the editors, publishers, or editorial boards of *Psychiatric Annals* .

The material presented at or in any *Psychiatric Annals* or Vindico Medical Education continuing education activity does not necessarily reflect the views and opinions of Vindico Medical Education or *Psychiatric Annals*. Neither *Psychiatric Annals*, Vindico Medical Education, nor the faculty endorse or recommend any techniques, commercial products, or manufacturers. The faculty/authors may discuss the use of materials and/or products that have not yet been approved by the U.S. Food and Drug Administration. Articles are intended for informational purposes only and should not be used as the basis of patient treatment. All readers and continuing education participants should verify all information before treating patients or utilizing any product.

CME

© iStockphoto.com

Using Non-experimental Data to Estimate Treatment Effects

CME EDUCATIONAL OBJECTIVES

1. Define the strengths and weaknesses of observational studies.
2. List methods for the identification and control of bias in observational studies.
3. Recognize propensity score matching methods.

Elizabeth A. Stuart, PhD, is with Johns Hopkins Bloomberg School of Public Health, Baltimore. Sue M. Marcus, PhD, is with Mount Sinai School of Medicine, New York. Marcela V. Horvitz-Lennon MD, is with the Department of Psychiatry, University of Pittsburgh School of Medicine. Robert D. Gibbons, PhD, is with the Center for Health Statistics, University of Illinois at Chicago. Sharon-Lise T. Normand, PhD, is with the Department of Health Care Policy, Harvard Medical School, and Department of Biostatistics, Harvard School of Public Health, Boston. C. Hendricks Brown, PhD, is with the Department of Epidemiology and Public Health, University of Miami.

Address correspondence to: Elizabeth A. Stuart, PhD, Johns Hopkins Bloomberg School of Public Health, 624 N Broadway, Eighth Floor, Baltimore, MD, 21205; fax 410-955-9088; or e-mail estuart@jhsph.edu.

Dr. Stuart has disclosed the following relevant financial relationships: Center for Prevention and Early Intervention, jointly funded by the National Institute of Mental Health (NIMH) and the National Institute on Drug Abuse (Grant MH066247; PI: N. Ialongo), and NIMH grant K25-MH083846: research grant recipient. Dr. Marcus has disclosed no relevant financial relationships. Dr. Horvitz-Lennon has disclosed the following relevant financial relationships: research grant recipient (NIMH Grant P50-MH073469). Dr. Gibbons has disclosed the following relevant financial relationship: research grant recipient (NIMH Grant R56-MH078580 and R01-MH8012201). Dr. Normand has disclosed the following relevant financial relationship: research grant recipient (NIMH Grant MH61434). Dr. Brown has disclosed the following relevant financial relationship: research grant recipient (NIMH Grant R01-MH040859).

The authors are indebted to Larry Zaborski, MS, Harvard Medical School, for earlier programming help, and to Richard Frank, PhD, Harvard Medical School, for generously providing the Medicaid data. Dr. Zaborski and Dr. Frank have disclosed no relevant financial relationships.

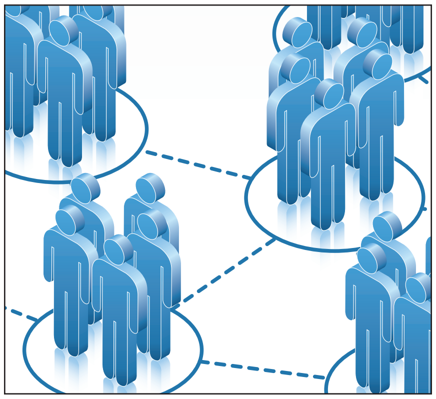
doi: 10.3928/00485713-20090625-07

Elizabeth A. Stuart, PhD; Sue M. Marcus, PhD; Marcela V. Horvitz-Lennon, MD; Robert D. Gibbons, PhD; Sharon-Lise T. Normand, PhD; and C. Hendricks Brown, PhD

Although much psychiatric research is based on randomized controlled trials (RCTs), where patients are randomly assigned to treatments, sometimes RCTs are not ethical nor feasible. Ethical concerns might preclude randomization, such as when evaluating whether "light cigarettes" produce less health risk by potentially randomizing subjects to smoke different brands, or it may be impractical, such as when the treatment of interest is widely available and commonly used. When RCTs are unethical or infeasible, a carefully constructed nonexperimental study can often be used to estimate treatment effects. Although nonexperimental studies are disadvantaged by lack of randomization, the study costs may be lower, the study sample may be broader, and follow-up may be longer, as compared to an RCT.[1-3]

The causal effect of a treatment can be unambiguously identified in studies where we are confident that the only difference between those subjects who take one treatment versus another is exposure to the intended treatments. By virtue of randomization, RCTs ensure, on average, the treatment and comparison groups are similar on baseline characteristics, both those that are measured as well as unmeasured ones. In nonexperimental studies, there is no such guarantee, but as we will see, there are some analytic approaches that can reduce differences. If treatment and comparison groups systematically differ on baseline factors that are correlated with the outcome, we say there is "selection bias." Selection bias leads to confounding, "a situation in which the estimated intervention effect is biased because of some difference between the comparison groups (apart from the intended interventions), such as baseline characteristics or prognostic factors. For a factor to be a confounder, it must differ between the comparison groups and predict the outcome of interest."[4]

Numerous design and analytic strategies are available to account for measured confounders. Well-designed nonexperimental studies make good use of measured confounders by creating treatment groups that look as similar as possible on the measured characteristics. Researchers then generally need to assume that, given comparability (or balance) between the groups on measured confounders, there are no measured or unmeasured differences, other than treatment received. This assumption has many names: "unconfounded treatment assignment," "no hidden bias," "ignorable treatment assignment," or "selection on observables."[5-7]



Numerous design and analytic strategies are available to account for measured confounders.

In this article, we describe nonexperimental approaches that create balance between treatment groups. The key idea is to use relatively recently developed techniques, known as propensity score methods, to ensure that the treatment and comparison subjects are as similar as possible. The goal is to replicate a randomized experiment, at least with respect to the measured confounders, by making the treatment and comparison groups look as if they could have been randomly assigned to the groups, in the sense of having similar distributions of the confounders. The sections of this article describe the five key stages to this process (see Table 1, page 721). The first section (Step 1) discusses defining and identifying the groups, then (Step 2) we outline the methods available for adjusting for covariate differences. We will also provide ways of assessing those differences (Step 3). We will also show how to estimate the treatment effects (Step 4), and then we discuss potential unobserved confounding (Step 5). We end with future considerations and conclusions.

We illustrate these propensity score methods using a study that compares atypical and conventional antipsychotic medications with regard to their effect on adverse metabolic outcomes (dyslipidemia, type II diabetes, and obesity).[17] The study uses data from Florida Medicaid beneficiaries (18 to 64 years), who were diagnosed with schizophrenia and continuously enrolled from 1997 to 2001. There have been a few RCTs that have been used to assess the metabolic effects of antipsychotic drugs.[18,19] Findings of these RCTs, which exclude a large portion of schizophrenic patients and test single drugs, are generally regarded as unrepresentative of the adverse events of these drugs as used in routine practice. A notable exception is Clinical Antipsychotic Trials of Intervention Effectiveness (CATIE),[20] a randomized effectiveness trial that enrolled a much broader group of patients. However, by far the bulk of the evidence on the causal associations of antipsychotics comes from studies using U.S. and U.K. administrative and medical databases.[21]

STEP 1: DEFINING THE TREATMENT AND COMPARISON GROUPS

The first step involves clearly specifying the treatment of interest and identifying individuals who experienced that treatment. One way to address this is to consider what treatment would be randomized if randomization were possible. For example, we could randomly assign some patients to receive an atypical medication. We then need to select an appropriate comparison condition (ie, what will the remaining patients receive?). In the motivating example of this article, which is investigating the metabolic effects of atypical antipsychotics, the relevant question is whether the comparison of interest is another type of medication, no medication, or either? Virtually all subjects with schizophrenia during the time frame during which the data are available (1997-2001) are treated with some type of antipsychotic agent. The key clinical question is thus not whether the patient should receive an antipsychotic medication, but rather, which type of antipsychotic medication should be used. In particular, we compare atypical antipsychotics (specifically, clozapine, olanzapine, quetiapine, and risperidone) to conventional antipsychotics (specifically, chlorpromazine, trifluoperazine, fluphenazine, perphenazine, thioridazine, haloperidol, and thiothixene). In this observational study, we use Medicaid claims data to examine outcomes for patients with filled prescriptions of these antipsychotic medications. We classify atypical (conventional) antipsychotic medication users as those subjects who filled at least one prescription for an atypical (conventional) antipsychotic during the time period examined. Prescribing information is unavailable and so only subjects who were written an antipsychotic prescription and filled it are included. Like an intent-to-treat analysis, we only know that the prescription was filled and not whether the medication was actually taken. See Tchernis et al[17] for more details on the sample and measures.

TABLE 1.

Recommended Steps in Analyzing Nonexperimental Studies

Step	Rationale
1	Define the treatment and comparison group(s)[8,9]
2	Create the treatment groups[2,9,10]
3	Assess the potential for confounding using standardized differences and plots[11,12]
4	Estimate the treatment effect on the treatment groups created in Step 2[13,14]
5	Determine robustness of conclusions to unmeasured confounders[6,15,16]

The next consideration is identification of confounders: factors that have previously been found to be associated with receipt of atypical versus typical antipsychotics and with metabolic outcomes. In addition, confounders should be measured before treatment assignment to ensure that they are not affected by the treatment.[22,23] Key confounders in the Medicaid study include demographic and clinical variables, listed in Table 2 (see page 722), such as sex, age, race, and medical comorbidities. A good study will use a dataset that has a large set of measured confounders so that the assumption of no hidden bias is more likely to be satisfied.

Once the treatment group, comparison group, and potential confounders are identified, researchers need to identify data on those groups and the confounders. The particular data elements necessary are: subjects, some of whom received the treatment (atypical antipsychotics) and others the comparison condition (conventional antipsychotics), an indicator for which subject is in which group, potential confounders, and outcomes. Ideally, the confounders are measured before the treatment and the outcomes after the treatment, to ensure temporal ordering. Unfortunately, often it is not possible to have truly longitudinal data, and researchers instead use cross-sectional data and make assumptions regarding the time ordering of the variables being measured. In the Medicaid study, we determined periods during which an individual had some minimal exposure to an antipsychotic drug, at least 6 months of Medicaid enrollment preceding treatment initiation (from which we obtained the covariate information), and a 12-month follow-up period to examine incidence of metabolic outcomes. We analyze one measurement occasion for each subject, measured 12 months following antipsychotic initiation. (See Marcus et al for methods for estimating causal effects with multiple outcome occasions.)[24]

STEP 2: METHODS FOR ADJUSTING FOR COVARIATE DIFFERENCES

Table 2 (Columns 1-3; see page 722) compares the means of the potential confounders between atypical and conventional antipsychotic users. The differences in percentages (for binary variables) or standardized differences (for continuous variables) are also reported. The standardized difference is the difference in means divided by the standard deviation of the confounder among the full set of conventional users.[1,12,25] We then multiply by 100 to express the difference as a percentage. The conventional users are older on average (by 26% of a standard deviation) and more likely to be black (34% vs. 24%), as compared with the atypical users. Because of these differences between the groups, comparing the raw outcomes between the two treatment groups would likely result in bias.[26] Statistical adjustments are required to deal with the differences in the observed confounders.

TABLE 2.

Characteristics of Individuals Taking Atypical vs. Conventional Antipsychotics

Characteristic	Type of Antipsychotic		Difference (Atypical — Conventional)	
	Atypical	Conventional	Full Cohort[a]	Matched Pairs
Male	48%	50%	-2.0%	0%
Mean age, yrs	38.4	41.2	-26%	5%
Race				
White	43%	36%	7.0%	-1.0%
Black	24%	34%	-10.0%	1.0%
Other race	33%	30%	4.0%	0%
SSI benefits[‡]	93%	96%	-2.0%	0%
Bipolar disorder	11%	8%	3.0%	3.0%
Substance abuse disorder	11%	9%	2.0%	3.0%
Hypertension	11%	12%	-1.0%	2.0%
Other chronic medical comorbidities[¶]	20%	16%	3.0%	3.0%
On other medications[£]	27%	27%	0%	1.0%
Mean # of inpatient days	2.1	1.5	12%	10%
Number of subjects	3,384	3,367	6,751	3,384[§]

a: For binary variables, defined as the difference in percentages. For continuous variables, defined as the difference in means divided by the standard deviation among those taking conventional antipsychotics and multiplied by 100 to express as a percentage.

‡: Social Security income benefit due to disability.

¶: Defined as 1 or more claims with a diagnosis of cardiovascular, respiratory, endocrine, liver, or pancreatic disease; HIV/AIDS; cancer; and/or seizure disorder.

£: Defined as 2 or more claims for beta-blockers; corticosteroids; valproate; loop and thiazide diuretics; and/or protease inhibitors.

§: Number of matched pairs.

Ideally, we want to compare atypical and conventional users who have "exactly" the same values for all the confounders. Assuming no unmeasured confounders, any difference in the outcomes could then be attributed to the treatment. However, exact matching on all of the covariates is often infeasible given the large number of covariates and relatively small number of subjects available. In the Medicaid study, if we were to make each of our 11 confounders binary, we would have 2048 (= 2^{11}) distinct strata and need to have both atypical and conventional antipsychotic users in each. Because this is not feasible, a reasonable strategy is to make the "distributions" of the confounders similar between the atypical and conventional antipsychotic users (eg, similar age, similar race, similar chronic medical comorbidity status). There are several general strategies to create comparable groups.

Regression Adjustment

A common approach to adjusting for confounders is regression adjustment, whereby the treatment effect is estimated by regressing the outcome of interest on an indicator for the treatment received and the set of confounders. The coefficient on the treatment indicator provides an estimate of the treatment effect (see Table 3, Column 1, page 723). A drawback to this approach is that if the atypical and conventional groups are very different on the observed covariates (eg, with more than a 25% standard deviation difference on average age, as seen in Table 2), the regression adjustment relies heavily on the particular model form and extrapolates between the two groups.[26,27] Why does this pose a problem? First, the regression approach will provide a prediction of what would have happened to atypical users had they instead used conventional antipsychotics using information from a set of conventional users who are very different from (eg, older than) those atypical users. Second, in most cases, the regression approach assumes a linear relationship between the measured covariates and the outcome of interest — an assumption that may not be true and is often difficult to test. Third, the output of standard regression analysis provides no information regarding covariate bal-

TABLE 3.

Estimated Absolute Risk (%) of Adverse Metabolic Outcomes of Atypical Compared with Conventional Antipsychotic Medication Use

Outcome	Regression Adjustment (# Subjects = 6,751)	Propensity Score-Based Analyses		
		Matching (# Pairs = 3,384)	Weighting (# Subjects = 6,751)	Subclassification§ (# Subjects = 6,751)
Dyslipidemia	1.67 (0.03)	1.04 (0.26)	1.66 (0.03)	1.92 (0.01)
Type 2 diabetes	0.27 (0.53)	0.06 (0.90)	0.31 (0.49)	0.23 (0.61)
Obesity	1.27 (0.00)	1.39 (0.00)	1.22 (0.00)	1.27 (0.00)

P *value in parentheses. Numbers greater than 0 indicate higher risk for individuals taking atypical antipsychotics.*

§ Average effect calculated by taking a precision-weighted average of the subclass-specific effects shown in Table 4 (see page 725).

**Although our outcomes are binary, we present results from a linear regression model. This was for comparability with the analyses described for the propensity score approaches with weights. If a logistic regression model is used, the difference in absolute risk can be obtained by comparing predictions of the outcomes for the full sample under each of the treatment conditions. In this study, the results are virtually identical. Step 4 (see page XXX) provides more detail.*

ance between the two treatment groups. Other approaches, which we now turn to, avoid these problems by ensuring that the comparisons are made between groups that are similar.

Propensity Score Methods

A useful tool to achieve comparable confounder distributions is the "propensity score," defined as the probability of receiving the treatment given the measured covariates.[7] A property of the propensity score makes it possible to select subjects based on their similarity with respect to the propensity score (a single number summary of the covariates, similar to a comorbidity score) in order to achieve comparability on all of the measured confounders, rather than having to consider each confounder separately. If a group of subjects have similar propensity scores, then they have similar probabilities of receiving the treatment, given the measured confounders. Within a small range of propensity score values, the atypical and conventional users should only differ randomly on the measured confounders, in essence replicating a randomized experiment.

Because the true propensity score for each subject is unknown, it is estimated with a model, such as a logistic regression, predicting treatment received given the measured confounders. Each subject's propensity score is their predicted probability of receiving the treatment, generated from the model. The diagnostics for propensity score estimation are not the standard logistic regression diagnostics, as concern is not with the parameter estimates or predictive ability of the model. Rather, the success of a propensity score model (and subsequent matching or stratification procedure) is determined by the covariate balance achieved. Below we describe several ways that propensity scores can be used.

Nearest Neighbor Matching

One of the simplest ways of ensuring the comparability of groups is to select for each treated individual the comparison individual with the closest propensity score.[28] (Often the matches are based on the logits, the log-odds of the predicted probabilities, because the logits have better statistical properties.) We illustrate a 1:1 matching algorithm where one conventional antipsychotic user is selected for each atypical antipsychotic user. Variations on this algorithm include selecting multiple matches for each atypical user, matching atypical users to a variable number of conventional users,[29] and prioritizing certain variables.[13] For example, if there are a large number of potential control subjects relative to the number of treated, it may be possible to get two or three good matches for each treated individual, which will increase the precision of estimates without sacrificing much balance.[29,30] In our study, because the numbers of conventional and atypical users are nearly equal, we used matching with replacement, meaning that each conventional user could be used as a match multiple times.[31]

The Figure, Panel A (see page 724), illustrates the resulting matches in the Medicaid study, with 1,809 conventional users matched to the 3,384 atypical users. The x-axis reflects the propensity scores; the y-axis is used to group the subjects into atypical (treated) vs. conventional (control), and matched vs. unmatched; the vertical spread of the symbols within each grouping is done to show the symbols more clearly. The figure shows the relative weight different subjects receive in the analyses of the outcomes, with the relative size of the symbols reflecting the number of times a subject was matched. Thus, conventional users selected as a match multiple times have larger symbols. The goal is to see good "overlap" between the propensity scores of the atypical and conventional users, which we have.

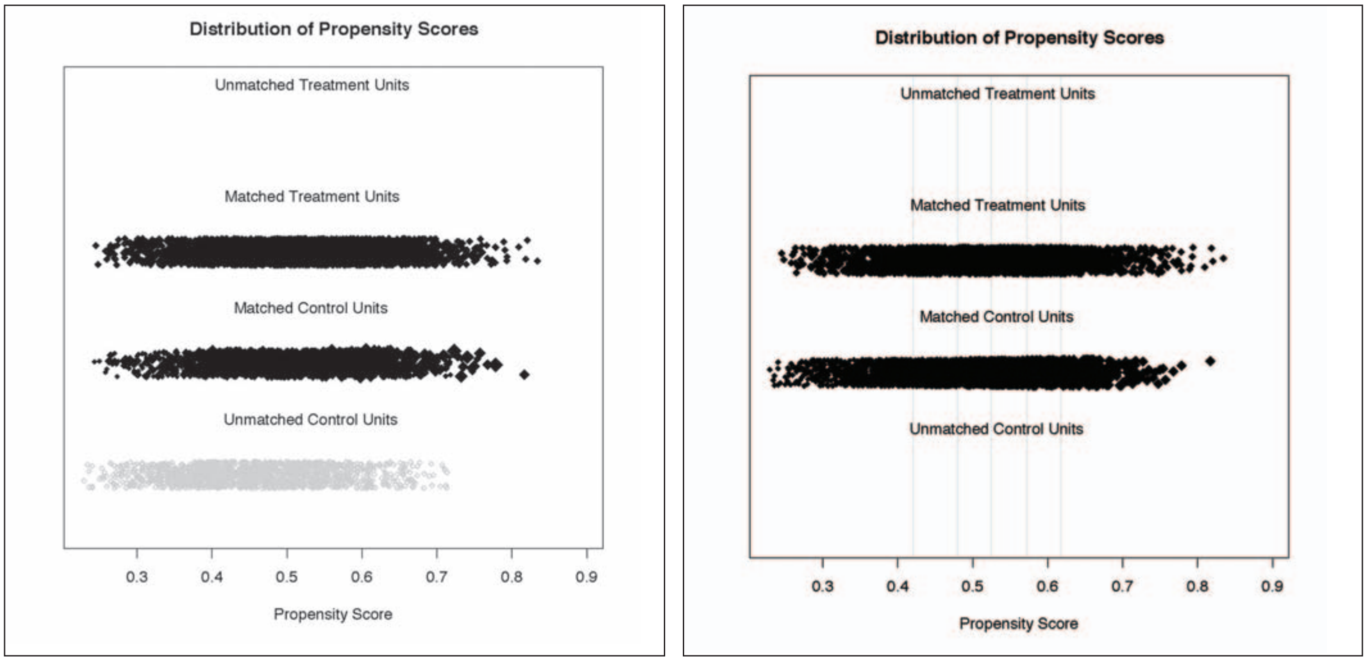


Figure. Results of 1:1 nearest neighbor matching with replacement and subclassification. Left: 1:1 nearest neighbor matching. Right): Subclassification. Propensity scores on x-axis; y-axis used to group subjects into atypical (treated) vs. conventional (control) and matched vs. unmatched. Matched subjects in black; unmatched in grey. The relative sizes of the diamonds reflect the relative weights subjects receive. Propensity score predicts atypical use given covariates; higher values indicate a higher likelihood of using atypical antipsychotics as compared with conventional antipsychotics. At right, vertical lines indicate subclass dividers.

However, there are quite a few conventional users with low propensity scores who are left unmatched. This illustrates a common drawback of nearest neighbor matching, in that sometimes subjects are unmatched, including some with propensity scores similar to those in the other group.

Weighting

A second approach, inverse probability of treatment weighting (IPTW), avoids this problem by using data from all subjects.[10,14,32] The idea of IPTW is similar to that of survey sampling weights, where individuals in a survey sample are weighted by their inverse probabilities of selection so that they then represent the full population from which the sample was selected. In our setting, we treat each of the treatment groups (the atypical users and the conventional users) as a separate sample and weight each up to the "population," which in this case is all study subjects. Each subject receives a weight that is the inverse probability of being in the group that they are in. However, instead of having known survey sampling probabilities, we use the estimated propensity scores. In particular, atypical users are weighted by one over their probability of receiving an atypical antipsychotic (the propensity score) (ie, the weight for a user of atypical antipsychotics with a propensity score of 0.1 would be 10). Conventional users are weighted by one over their probability of receiving a conventional antipsychotic (one minus the propensity score). In the Medicaid study, the conventional users with low probabilities of receiving a conventional antipsychotic will receive relatively large weights, because they actually look more similar to the atypical users, thus providing good information about what would happen to the atypical users if they had instead taken conventional antipsychotics.

Subclassification

Subclassification, also called stratification, is a method that also uses all subjects, by forming groups (subclasses) of individuals with similar propensity scores.[33] This is done by sorting the propensity scores and forming some number of subclasses (eg, 10), based on the percentiles (eg, deciles) in which each individual's propensity score falls. In the Medicaid study, the subclasses were created to have approximately the same number of subjects taking atypical antipsychotics (about 565); the number of conventional users in each subclass ranges from 287 to 933 (Figure, Panel B; also see Table 4, page 725). Because of the properties of propensity scores described above, within each subclass, the subjects look similar on the measured confounders. When a relatively small number of subclasses are formed, there are sometimes still differences within subclasses;[14] in those cases it may make sense to create more subclasses or use more sophisticated methods such as full matching, which forms many subclasses in a way that minimizes the differences in the propensity scores.[29]

TABLE 4.

Estimated Absolute Risk (%) of Adverse Metabolic Outcomes of Atypical Compared with Conventional Antipsychotic Medication Use Stratified by Propensity Score Subclass

Outcome	Subclass Group					
	1§	2	3	4	5	6§
Dyslipidemia	1.89 (0.26)	0.93 (0.61)	1.17 (0.56)	2.45 (0.20)	2.87 (0.18)	2.38 (0.24)
Type 2 diabetes	1.14 (0.31)	0.20 (0.84)	-0.73 (0.47)	-1.05 (0.31)	1.81 (0.12)	0.27 (0.82)
Obesity	0.72 (0.28)	2.56 (0.00)	1.87 (0.06)	0.66 (0.54)	1.97 (0.14)	-0.76 (0.59)
# of atypical users	560	567	565	563	563	566
# of conventional users	933	749	572	476	350	287

§ Subclass 1 includes subjects with the lowest propensity of receiving atypical antipsychotics while those in Subclass 6 have the highest propensity of receiving atypical antipsychotics.
P *value in parentheses. Numbers greater than 0 indicate higher risk for atypical users.*

Remarks

Is it better to match or to stratify/weight? The answer depends on whether the investigator is more concerned about bias or about having enough power to detect an effect. Matching approaches are often used when it is important to reduce as much as possible differences between treatment groups and consequently, not all subjects are used, reducing the total sample size available to find differences. Although subclassification and weighting retain all subjects (generally yielding some efficiency gain), there is a risk of making comparisons between individuals who are not as alike as desired.[14]

STEP 3: ASSESSING MEASURED CONFOUNDING

How do we know if the atypical and conventional groups are "similar," at least on the measured covariates? After using one of the approaches described above, the crucial next step is to check the resulting "balance": the similarity of the confounders between the treatment and comparison groups. Common (and sometimes misguided) measures used for balance checks are standard hypothesis tests, such as t-tests. The danger in using test statistics is that they conflate changes in balance with changes in the sample size; comparing *P* values before and after matching can be misleading, implying that balance has improved, when in fact it has not.[1,12]

A good balance measure, and the one we suggest, is the standardized difference in means. This is most appropriate for continuous variables. A general rule of thumb is that an acceptable standardized difference is less than 10%.[12] Differences larger than 10% roughly imply that 8% or more of the area covered by atypical and conventional users combined is not overlapping. For binary variables the absolute value of the difference in proportions is examined. These measures are generally calculated both in the full dataset (see Table 2, Column 3, page 723,), as well as in the dataset after applying one of the propensity score methods described above (see Table 2, Column 4, page 722); if the propensity score method was successful the standardized differences and differences in proportions should be smaller than they were in the original data set. After 1:1 matching (see Table 2, Column 4) the largest standardized difference is 3%, which is a good situation. Similar balance was achieved with weighting and subclassification. In contrast, the largest standardized difference prior to matching was 26%, which is clearly an unacceptable situation. In some cases adequate balance may not be achieved with the available data. This is an indication that estimating the treatment effect with that data may be unreliable. It may be necessary to add interactions of the measured covariates in the propensity score model, seek additional data sources, or reconsider the question of interest.

STEP 4: ESTIMATING THE AVERAGE TREATMENT EFFECT

Once adequate balance is achieved, the next step is to estimate the treatment effect. Note that this is the first time that the outcome is used; the propensity score method itself is not selected or implemented using the metabolic outcome measures, beyond the idea of selecting confounders that may be correlated with the outcome(s).

Regression Adjustment

One method of estimating the treatment effect is to regress the outcomes for subjects in the original (unmatched) dataset on the measured confounders. In the antipsychotic study, we estimated a linear regression, where the coefficient of the atypical antipsychotic variable represents the increase (or decrease) in risk for atypical users. The results of this approach are shown in Table 3, Column 1 (see page 723), where atypical antipsychotic use increases the risk of dyslipidemia and of obesity. This regression is easy to conduct, but has the drawbacks discussed above, particularly when the treatment groups are far apart on the covariates. However, despite these limitations of regression adjustment in general, in fact, combining it with the propensity score methods described above has been found to be a very effective approach,[11,34-36] and we use that approach for the remaining methods.

TABLE 5.

Sensitivity of Atypical Antipsychotic Effect on Obesity to an Unmeasured Confounder

Sensitivity Parameter	Lower *P* value	Upper *P* value
1 (No hidden bias)	.00	.00
1.25	.00	.01
1.5	.00	.12
1.75	.00	.43
2.0	.00	.75
3.0	.00	.99

Sensitivity parameter represents the odds by which individuals with the same measured confounders differ in receiving atypical antipsychotics due to hidden bias. P values shown are 1-sided; the sum of P values > .05 indicates the odds of atypical use that would change the conclusions of the study in terms of making the effect insignificant. For the risk of obesity, this occurs at a value of 1.5.

Nearest Neighbor Matching

Outcome analysis after 1:1 nearest neighbor matching is very straightforward. With paired data and binary outcomes, a natural method is McNemar's test. McNemar's test indicates a statistically significant adverse effect of atypical antipsychotics on obesity (chi^2 = 14.61 on 1 degree of freedom P = 0.0001): 5% of the 3,384 pairs had discordant outcomes and in 65% of the discordant pairs, the atypical subject had obesity.

Alternatively, any analysis that would have been conducted on the full dataset can instead be conducted on the matched dataset.[11] We estimated a regression model with each metabolic outcome predicted by whether someone took an atypical antipsychotic and the measured confounders, using the matched sample. Because the matching was done with replacement, the regression analysis was run using weights to account for that design.[13] We find that atypical antipsychotics increased the risk of obesity, but not dyslipidemia or type 2 diabetes (see Table 3, Column 2, page 723), consistent with the results found using McNemar's test.

Weighting

After constructing IPTW weights, the effect estimate is obtained by estimating a weighted regression model using the IPTW weights.[14] The results are consistent with those of the standard regression adjustment, indicating increased risk of dyslipidemia and obesity for those taking atypical antipsychotics (see Table 3, Column 3, page 723).

Subclassification

With subclassification, treatment effects are first estimated separately within each subclass. Because of the potential for residual bias when the subclasses are relatively large, it is particularly important to estimate these effects using regression adjustment within each subclass, controlling for the confounders.[14] If the treatment effects are similar across subclasses, it may make sense to combine the subclass-specific estimates to obtain an overall estimate. The results for the antipsychotic study do not indicate substantial treatment differences across subclasses (see Table 4, page 725). After combining the subclass results by taking a precision-weighted average of the effects within each subclass, we find that the overall effects are similar to those from the simple regression adjustment and from weighting (see Table 3, Column 4, page 723). An advantage of the subclassification approach is that it permits a non-linear pattern in the effects across the subclasses.

Remarks

Selection of matching versus subclassification or weighting involves a bias/variance trade-off. One-to-one matching generally yields more closely matched samples and thus lower bias, but higher variance because of the smaller sample size used. The better balance generally obtained by matching also sometimes yields smaller point estimates of effects. In our example, the lack of a statistically significant finding on dyslipidemia when using 1:1 matching but a significant finding when using other approaches appears to be a result of a combination of these factors. In comparison with the effect on obesity, the effect of dyslipidemia is much weaker: for dyslipidemia, 53% of the discordant pairs had an atypical user with dyslipidemia (chi^2 = 2.613 on 1 degree of freedom; P = 0.11), for obesity, 65% of the discrepant pairs had an atypical user with obesity. The discrepancy in results also indicates the value in assessing sensitivity by trying a few different approaches; those that yield the best covariate balance should be used.[11]

STEP 5: ASSESSING UNMEASURED CONFOUNDING

The final question in any nonexperimental study is how sensitive are the results to a potential unmeasured confounder. We illustrate an approach that determines how strongly related to the decision to fill an atypical antipsychotic medication an unmeasured confounder would have to be to make the observed effect go away (ie, lose statistical significance).[37] We illustrate the approach using the matched pairs from 1:1 matching using the obesity outcome. Table 5 indicates that for two subjects who appear similar on the measured covariates, if their odds of filling an atypical antipsychotic medication differ by a factor of 1.5 or larger, then the treatment effect becomes statistically insignificant. The size of these odds needs to be interpreted in the context of the particular problem. In our analyses, the largest observed odds ratio was 1.75 (95% CI: 1.55, 1.98) reflecting an increased odds of receiving an atypical antipsychotic for white subjects relative to black subjects. Given this sized odds ratio observed, the small number of confounders available in the data, and knowing that the results are sensitive at an odds of 1.5, makes us cautious in concluding that atypical antipsychotic use increases the risk of obesity compared to conventional antipsychotic use. These results need to be replicated in other studies.

FUTURE CONSIDERATIONS AND CONCLUSIONS

This article has provided an overview of the approaches for estimating treatment effects with nonexperimental data, with a focus on propensity score methods that ensure comparison of similar individuals. Although in this study the propensity score approaches gave results similar to those of traditional regression adjustment, we can have more confidence because of the balance obtained by the matching, weighting, and subclassification methods. The methods generally imply increased risk of dyslipidemia and obesity for individuals on atypical antipsychotics and no increased risk of Type II diabetes. However, we should interpret these results with caution, as the effect on dyslipidemia was sensitive to the particular method used and even the (stronger) effect on obesity is potentially sensitive to an unmeasured confounder.

There are a number of complications that researchers may encounter when designing an observational study. The first is missing data: rarely do researchers measure all of the variables of interest for all study subjects. If there are not many patterns of missing data, a first solution is to estimate separate propensity scores for each missing data pattern.[7] A second approach is to include missing data indicators in the propensity score model; this will essentially match individuals on both the observed values (when possible) and the patterns of missingness.[38,39] A third approach is to use multiple imputation and undertake the propensity score matching and outcome analysis separately within each multiply imputed dataset.[40]

A second complication involves questions where the treatment of interest is not a simple binary comparison. Interest might be in the effect of different types or dosages of antipsychotic medications. Two solutions exist in this type of setting. First, if scientifically interesting, focus can be shifted to a binary comparison, for example comparing low compared with high doses. Second, a new area of methodological research has developed generalized propensity scores for use with non-binary treatments.[5,16,41]

Another concern in observational studies is that an outcome of interest, such as obesity, may be subject to ascertainment bias: a differential probability of ascertaining a condition based solely on the medication that is used. Ascertainment bias can also potentially occur in randomized trials when there is increased physician contact because of elevated side effects on active medication versus placebo.[42] When ascertainment bias occurs, propensity score methods are not sufficient to provide appropriate adjustments.

A final concern with any nonexperimental study is that of unmeasured confounding: there may be some unmeasured variable related to both which treatment an individual receives and their outcome. Using propensity score approaches to deal with measured confounders is an important step, but there is always concern about potential unmeasured confounders. One approach to assess whether this could be a problem is to examine an outcome that should not be affected by the treatment of interest; if an effect is actually found, that may indicate the presence of unmeasured confounding. We have also illustrated here a statistical sensitivity analysis, which can be used to assess how important such an unmeasured confounder may be with respect to the study conclusions. Instrumental variables methods, also known as "natural experiments," are another type of nonexperimental study that can be used when unmeasured confounding is of particular concern. Instrumental variables analyses do not rely on the assumption of no unmeasured confounding and instead rely on finding some "instrument" that affects the receipt of the treatment of interest but does not directly affect the outcomes.[43]

What are the primary lessons? When reading a study that uses nonexperimental data, readers should:

- Consider whether the results are plausible,[44]
- Examine whether the groups being compared are similar on the relevant variables, and
- Consider whether there are potentially important confounders that were not measured.

When estimating treatment effects using nonexperimental methods, researchers should:

- Be clear about the treatment and comparison conditions,
- Identify data that has a large set of potential confounders measured, and
- Ensure comparisons are made using similar individuals by using one of the propensity score methods described above.

In conclusion, propensity score approaches such as matching, weighting, and subclassification are an important step forward in the estimation of treatment effects using observational data. Whenever treatment effects are estimated using nonexperimental studies, particular care should be taken to ensure that the comparison is being done using treated and comparison subjects who are as similar as possible; propensity scores are one way of doing so. Propensity score methods can thus help researchers, as well as users of that research, to have more confidence in the resulting study findings.

REFERENCES

1. Imai K, King G, Stuart EA. Misunderstandings among experimentalists and observationalists about causal inference. *Journal of the Royal Statistical Society, Series A.* 2008;171:481-502.
2. Rochon PA, Gurwitz JH, Sykora K, et al. Reader's guide to critical appraisal of cohort studies: 1. Role and Design. *BMJ.* 2005;330:895-897.
3. West SG, Duan N, Pequegnat W, et al. Alternatives to the randomized controlled trial. *Am J Public Health.* 2008;98(8):1359-1366.
4. Altman DG, Schulz KF, Moher D, et al. The revised CONSORT statement for reporting randomized trials: explanation and elaboration. *Ann Intern Med.* 2001;134(8):663-694.
5. Imbens GW. Nonparametric estimation of average treatment effects under exogeneity: a review. *Review of Economics and Statistics.* 2004;86(1):4-29.
6. Rosenbaum PR. *Observational Studies.* 2nd ed. New York, NY: Springer Verlag; 2002.
7. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika.* 1983;70:41-55.
8. Holland PW. Statistics and causal inference. *Journal of the American Statistical Association.* 1986;81:945-960.
9. 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.
10. Hernán MA, Robins JM. Estimating causal effects from epidemiological data. *J Epidemiol Community Health.* 2006;60(7):578-586.
11. Ho DE, Imai K, King G, Stuart EA. Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Political Analysis.* 2007;15:199-236.
12. Mamdani M, Sykora K, Li P, et al. Reader's guide to critical appraisal of cohort studies: 2. Assessing potential for confounding. *BMJ.* 2005;330(7497):960-962.
13. Stuart EA, Rubin DB. Best practices in quasi-experimental designs: matching methods for causal inference. In: J. Osborne. *Best Practices in Quantitative Social Science.* Thousand Oaks, CA: Sage Publications; 2007:155-176.
14. 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.
15. Copas JB, Li HG. Inference for non-random samples. Journal of the Royal Statistical Society, Series B. 1997;59(1):55-95.
16. Rosenbaum PR, Rubin DB. Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome. *Journal of the Royal Statistical Society, Series B.* 1983;45(2):212-218.
17. Tchernis R, Horvitz-Lennon M, Normand SL. On the use of discrete choice models for causal inference. *Stat Med.* 2005;24(14):2197-2212.
18. Conley RR, Mahmoud R. A randomized double-blind study of risperidone and olanzapine in the treatment of schizophrenia or schizoaffective disorder. *Am J Psychiatry.* 2001;158(5):765-774.
19. Lindenmayer JP, Czobor P, Volavka J, et al. Changes in glucose and cholesterol levels in patients with schizophrenia treated with typical or atypical antipsychotics. *Am J Psychiatry.* 2003;160(2):290-296.
20. Meyer JM, Davis VG, Goff DC, et al. Change in metabolic syndrome parameters with antipsychotic treatment in the CATIE Schizophrenia Trial: prospective data from phase 1. *Schizophr Res.* 2008;101(1-3):273-286.
21. Jin H, Meyer JM, Jeste DV. Atypical antipsychotics and glucose dysregulation: a systematic review. *Schizophr Res.* 2004;71(2-3):195-212.
22. Frangakis CE, Rubin DB. Principal stratification in causal inference. *Biometrics.* 2002;58(1):21-29.
23. Rosenbaum PR. The consequences of adjustment for a concomitant variable that has been affected by the treatment. *Journal of the Royal Statistical Society, Series A.* 1984;147:656-666.
24. Marcus SM, Siddique J, Ten Have TR, Gibbons RD, Stuart EA, Normand SL. Balancing treatment comparisons in longitudinal studies. *Psychiatric Annals.* 2008;38(12):805-811.
25. Cohen J. Statistical power analysis for the behavioral sciences. 2nd ed. Philadelphia, PA: Lawrence Erlbaum; 1988.
26. Rubin DB. Using propensity scores to help design observational studies: application to the tobacco litigation. *Health Services & Outcomes Research Methodology.* 2001;2:169-188.
27. Robins J, Greenland S. The role of model selection in causal inference from nonexperimental data. *Am J Epidemiol.* 1986;123(3):392-402.
28. Rubin DB. Matching to remove bias in observational studies. *Biometrics.* 1973;29:159-184.
29. Stuart EA, Green KM. Using full matching to estimate causal effects in non-experimental studies: examining the relationship between adolescent marijuana use and adult outcomes. *Developmental Psychology.* 2008;44(2):395-406.
30. Smith H. Matching with multiple controls to estimate treatment effects in observational studies. *Sociological Methodology.* 1997;27:325-353.
31. Dehejia RH, Wahba S. Propensity score matching methods for non-experimental causal studies. *Review of Economics and Statistics.* 2002;84:151-161.
32. McCaffrey DF, Ridgeway G, Morral AR. Propensity score estimation with boosted regression for evaluating causal effects in observational studies. *Psychol Methods.* 2004;9(4):403-425.
33. Rosenbaum PR, Rubin DB. Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association.* 1984;79:516-524.
34. Cochran WG, Rubin DB. Controlling bias in observational studies: a review. Sankhya: *The Indian Journal of Statistics, Series A.* 1973;35:417-446.
35. Heckman JJ, Ichimura H, Todd PE. Matching as an econometric evaluation estimator: Evidence from evaluating a job training program. *Review of Economic Studies.* 1997;64:605-654.
36. Robins JM, Rotnitzky A. Comment on the Peter J. Bickel and Jaimyoung Kwon, 'Inference for semiparametric models: Some questions and an answer.' *Statistica Sinica.* 2001;11:920-936.
37. Rosenbaum PR. Sensitivity analysis for certain permutation inferences in matched observational studies. *Biometrika.* 1987;74:13-26.
38. D'Agostino RB, Lang W, Walkup M, Morgan T. Examining the impact of missing data on propensity score estimation in determining the effectiveness of self-monitoring of blood glucose (SMBG). *Health Services & Outcomes Research Methodology.* 2001;2:291-315.
39. Haviland A, Nagin DS, Rosenbaum PR. Combining propensity score matching and group-based trajectory analysis in an observational study. *Psychol Methods.* 2007;12(3):247-267.
40. Song J, Belin TR, Lee MB, Gao X, Rotheram-Borus MJ. Handling baseline differences and missing items in a longitudinal study of HIV risk among runaway youths. *Health Services & Outcomes Research Methodology.* 2001;2:317-329.
41. Imai K, van Dyk DA. Causal inference with general treatment regimes: Generalizing the propensity score. *Journal of the American Statistical Association.* 2004;99(467):854-866.
42. Posner K, Oquendo MA, Gould M, Stanley B, Davies M. Columbia Classification Algorithm of Suicide Assessment (C-CASA): Classification of suicidal events in the FDA's pediatric suicidal risk analysis of antidepressants. *Am J Psychiatry.* 2007;164(7):1035-1043.
43. Hernan MA, Robins JM. Instruments for causal inference: an epidemiologist's dream? *Epidemiology.* 2006;17(4):360-372.
44. Normand SL, Sykora K, Li P, Mamdani, M, Rochon PA, Anderson GM. Reader's guide to critical appraisal of cohort studies: 3. Analytical strategies to reduce confounding. *BMJ.* 2005;330:1021-1023.