Published in final edited form as:

Psychiatr Ann. 2009 July 1; 39(7): 41451. doi:10.3928/00485713-20090625-07.

Using Non-experimental Data to Estimate Treatment Effects

Elizabeth A. Stuart, PhD^{2,1}, Sue M. Marcus, PhD³, Marcela V. Horvitz-Lennon, MD⁴, Robert D. Gibbons, PhD⁵, and Sharon-Lise T. Normand, PhD⁶

² Johns Hopkins Bloomberg School of Public Health, Baltimore

³ Mount Sinai School of Medicine, New York

⁴ University of Pittsburgh School of Medicine, Department of Psychiatry, Pittsburgh

⁵ University of Illinois at Chicago, Chicago

⁶ Harvard Medical School and Harvard School of Public Health, Boston

Abstract

While much psychiatric research is based on randomized controlled trials (RCTs), where patients are randomly assigned to treatments, sometimes RCTs are not feasible. This paper describes propensity score approaches, which are increasingly used for estimating treatment effects in non-experimental settings. The primary goal of propensity score methods is to create sets of treated and comparison subjects who look as similar as possible, in essence replicating a randomized experiment, at least with respect to observed patient characteristics. A study to estimate the metabolic effects of antipsychotic medication in a sample of Florida Medicaid beneficiaries with schizophrenia illustrates methods.

Introduction

While much psychiatric research is based on randomized controlled trials (RCTs), where patients are randomly assigned to treatments, sometimes RCTs are not feasible. Ethical concerns might preclude randomization, such as randomizing subjects to smoke, 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 non-experimental study can be used to estimate treatment effects. While non-experimental 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 (1,2).

The primary challenge for estimation of treatment effects is the identification of subjects who are as similar as possible on all background characteristics other than the treatment of interest. By virtue of randomization, RCTs ensure, on average, the treatment and comparison groups are similar on background characteristics, measured and unmeasured. In non-experimental studies, there is no such guarantee. Treatment and comparison groups may systematically differ on factors that also affect the outcome, a problem referred to as “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 planned interventions such as baseline characteristics, prognostic factors, or concomitant interventions. For a factor to be a confounder, it must differ between the comparison groups and predict the outcome of interest” (3).

¹Corresponding author contact information: Johns Hopkins Bloomberg School of Public Health, 624 N Broadway, 8th Floor, Baltimore, MD, 21205; 410-502-6222 (phone); 410-955-9088 (fax); estuart@jhsph.edu.

Numerous design and analytical strategies are available to account for measured confounders but the major limitation is the potential for unmeasured confounders. Well-designed non-experimental studies make good use of measured confounders by creating treatment groups that look as similar as possible on the measured characteristics. Researchers then 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” (4–6).

We describe approaches that, through the careful design and analysis of non-experimental studies, 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. We describe the five key stages to this process (Table 1). A study that compares atypical and conventional antipsychotic medications with regard to their effect on adverse metabolic outcomes (dyslipidemia, Type II diabetes, and obesity) (16) illustrates the methods. The study uses data from Florida Medicaid beneficiaries (18 to 64 years), diagnosed with schizophrenia and continuously enrolled from 1997 to 2001. Although the bulk of the evidence on the causal associations of antipsychotics comes from studies using U.S. and U.K. administrative and medical databases, RCTs have been used to assess the metabolic effects of antipsychotic drugs (e.g., 17,18). Findings of these RCTs, however, are not regarded as representative of the adverse events of these drugs as used in routine practice. A possible exception is the CATIE trial (19), an effectiveness trial in that other than the randomization, every other aspect of the care was naturalistic. Conduct of this type of trial is costly and generally unfeasible.

I. 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 patients to receive an atypical medication. We then need to select an appropriate comparison condition. Because this study investigates 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 this time frame are treated with some type of antipsychotic agent and thus the key clinical question is not whether the patient should receive an antipsychotic medication, but rather, which type of antipsychotic medication should be used. We compare atypical antipsychotics (specifically, clozapine, olanzapine, quetiapine, and risperidone) to conventional antipsychotics (specifically, chlorpromazine, trifluoperazine, fluphenazine, perphenazine, thioridazine, haloperidol, and thiothixene). We use Medicaid claims data so that atypical (conventional) antipsychotic users are those subjects who *filled* at least one prescription for an atypical (conventional) antipsychotic. 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.

The next consideration is identification of confounders: factors that have previously been found to be associated with receipt of atypical antipsychotics and/or with metabolic outcomes. Key confounders in the Medicaid study include demographic and clinical variables, listed in Table 1, such as sex, age, race, and medical comorbidities. A good study will have a large set of measured confounders so that the assumption of no hidden bias is 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. Confounders are measured *before* treatment assignment to ensure that they are not affected by the treatment (20,21) and outcomes are measured *after* treatment assignment, to ensure temporal ordering. 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. Often it is not possible to have truly longitudinal data, and researchers instead use cross-sectional data where assumptions regarding the time ordering of the variables being measured are made. We analyze one measurement occasion for each subject, measured 12 months following antipsychotic initiation. See the paper by Marcus et al. in this series for methods for estimating causal effects with multiple outcome occasions (22).

II. Creating the Groups for Comparison

Table 2 (Columns 1–3) 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 (1,11,23). 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 African American (34% vs. 24%), as compared to the atypical users. Because of these differences between the groups, comparing the raw outcomes between the two treatment groups would result in bias (24). Statistical adjustments are required to deal with the differences in the observed confounders.

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—e.g., 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 (Table 3, Column 1).² A drawback of this approach is that if the atypical and conventional groups are very different on the observed covariates (e.g., with over 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 (24,25). Why does this pose a problem? First, the

²Although 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. Section IV provides more detail.

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, e.g., 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 balance between the two treatment groups. Other approaches 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 (6). 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 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.

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 score³ (26). 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 (27), and prioritizing certain variables (12). For example, if there are a large number of potential control subjects relative to the number of treated, it may be possible to get 2 or 3 good matches for each treated individual, which will increase the precision of estimates without sacrificing much balance (27,28). 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 (29).

Figure 1 Panel A 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. However, there are quite a few conventional users with low propensity scores who are left unmatched. This illustrates a

³Often the matches are based on the logits (the log-odds of the predicted probabilities) because the logits have better statistical properties.

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 (9,13,30). 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 in which 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), and 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 (31). 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 1 Panel B; Table 4). Because of the properties of propensity scores described above, within each subclass, the subjects look similar on the measured confounders.

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. While 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.

III. Assessing Potential 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 (1,11).

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% (11). Differences larger than 10% roughly imply that 8% or more of the area covered by atypical and conventional users combined is not overlapping.

⁴ For binary variables the absolute value of the difference in proportions is examined. These measures are generally calculated both in the full dataset (Table 2, Column 3), as well as in

the dataset after applying one of the propensity score methods described above (Table 2, Column 4); 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 (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.

IV. Estimating the 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, 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 based 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 (10,32–34), and we use that approach for the remaining methods.

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 subjects had obesity.

Alternatively, any analysis that would have been conducted on the full dataset can instead be conducted on the matched dataset (10). 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 (12). We find that atypical antipsychotics increased the risk of obesity, but not dyslipidemia or Type II diabetes (Table 3, Column 2), 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 (13). The results are consistent with those of the standard regression adjustment, indicating increased risk of dyslipidemia and obesity for those taking atypical antipsychotics (Table 3, Column 3).

Subclassification—With subclassification, treatment effects are first estimated separately within each subclass. Because of the potential for residual bias when the subclasses are

⁴The 10% threshold is a small effect size using Cohen's effect size criteria (21).

relatively large, it is particularly important to estimate these effects using regression adjustment within each subclass, controlling for the confounders (13). 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 (Table 4). 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 (Table 3, Column 4). An advantage of the subclassification approach is that it permits non-linear associations 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 (10).

V. Assessing Unmeasured Confounding

The final question in any non-experimental 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 (i.e., lose statistical significance; 35). 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 size 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.

VI. Discussion

This paper has provided an overview of the approaches for estimating treatment effects with non-experimental data, with a focus on propensity score methods that ensure comparison of similar individuals. While 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 (6). 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 on the patterns of missingness (36,37). A third approach is to use multiple imputation and undertake the propensity score matching and outcome analysis separately within each multiply imputed dataset (38).

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 vs. high doses. Second, a new area of methodological research has developed generalized propensity scores for use with non-binary treatments (5,16,39).

A final concern with any non-experimental 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 effects of 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.

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

- Consider whether the results are plausible (40),
- Examine whether the groups being compared are similar on the relevant variables,
- Consider whether there are potentially important confounders that were not measured.

When estimating treatment effects using non-experimental methods, researchers should:

- Be clear about the treatment and comparison conditions,
- Identify data that has a large set of potential confounders measured,
- 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 non-experimental 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.

Acknowledgments

Dr. Stuart's effort was supported by the 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). Dr. Normand's effort was supported by Grant MH61434 from NIMH. Dr. Gibbons' effort was supported by NIMH Grant R56-MH078580, and Dr. Horvitz-Lennon's by NIMH Grant P50-MH073469. 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.

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, Mamdani M, Streiner DL, Garfinkel S, Normand SLT, Anderson GM. Reader's guide to critical appraisal of cohort studies: 1. Role and Design. *British Medical Journal* 2005;330:895–897. [PubMed: 15831878]
3. Altman DG, Schulz KF, Moher D, Egger M, Davidoff F, Elbourne D, et al. The revised CONSORT statement for reporting randomized trials: explanation and elaboration. *Annals of Internal Medicine* 2001;134:663–694. [PubMed: 11304107]
4. Imbens GW. Nonparametric estimation of average treatment effects under exogeneity: a review. *Review of Economics and Statistics* 2004;86(1):4–29.
5. Rosenbaum, PR. *Observational Studies*. 2. New York: Springer Verlag; 2002.
6. Rosenbaum PR, Rubin DB. The Central Role of the Propensity Score in Observational Studies for Causal Effects. *Biometrika* 1983;70:41–55.
7. Holland PW. Statistics and causal inference. *Journal of the American Statistical Association* 1986;81:945–960.
8. Rubin DB. The design *versus* the analysis of observational studies for causal effects: Parallels with the design of randomized trials. *Statistics in Medicine* 2007;26:20–36. [PubMed: 17072897]
9. Hernán MA, Robins JM. Estimating causal effects from epidemiological data. *Journal of Epidemiology and Community Health* 2006;60:578–586. [PubMed: 16790829]
10. 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.
11. Mamdani M, Sykora K, Li P, Normand SLT, Streiner DL, Austin PC, Rochon PA, Anderson GM. Reader's guide to critical appraisal of cohort studies: 2. Assessing potential for confounding. *British Medical Journal* 2005;330:960–962. [PubMed: 15845982]
12. Stuart, EA.; Rubin, DB. Best Practices in Quasi-Experimental Designs: Matching methods for causal inference. In: Osborne, J., editor. *Best Practices in Quantitative Social Science*. Thousand Oaks, CA: Sage Publications; 2007. p. 155-176.
13. Lunceford JK, Davidian M. Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. *Statistics in Medicine* 2004;23:2937–2960. [PubMed: 15351954]
14. Copas JB, Li HG. Inference for non-random samples. *Journal of the Royal Statistical Society, Series B* 1997;59(1):55–95.
15. 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.
16. Tchernis R, Horvitz-Lennon M, Normand SLT. On the use of discrete choice models for causal inference. *Statistics in Medicine* 2005;24:2197–2212. [PubMed: 15887310]
17. Conley RR, Mahmoud R. A randomized double-blind study of risperidone and olanzapine in the treatment of schizophrenia or schizoaffective disorder. *American Journal of Psychiatry* 2001;158:765–774. [PubMed: 11329400]
18. Lindenmayer JP, Czobor P, Volavka J, et al. Changes in glucose and cholesterol levels in patients with schizophrenia treated with typical or atypical antipsychotics. *American Journal of Psychiatry* 2003;160:290–296. [PubMed: 12562575]
19. 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 I. *Schizophrenia Research* 2008;101:273–286. [PubMed: 18258416]
20. Frangakis CE, Rubin DB. Principal stratification in causal inference. *Biometrics* 2002;58:21–29. [PubMed: 11890317]
21. 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.
22. Marcus SM, Siddique J, Gibbons RD, Normand SLT. Balancing treatment comparisons in longitudinal studies. *Psychiatric Annals*. 2008

23. Cohen, J. Statistical power analysis for the behavioral sciences. 2. New Jersey: Lawrence Erlbaum; 1988.
24. Rubin DB. Using propensity scores to help design observational studies: application to the tobacco litigation. *Health Services & Outcomes Research Methodology* 2001;2:169–188.
25. Robins J, Greenland S. The role of model selection in causal inference from nonexperimental data. *American Journal of Epidemiology* 1986;123(3):392–402. [PubMed: 3946386]
26. Rubin DB. Matching to remove bias in observational studies. *Biometrics* 1973;29:159–184.
27. 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. [PubMed: 18331131]
28. Smith H. Matching with multiple controls to estimate treatment effects in observational studies. *Sociological Methodology* 1997;27:325–353.
29. Dehejia RH, Wahba S. Propensity Score Matching Methods for Non-Experimental Causal Studies. *Review of Economics and Statistics* 2002;84:151–161.
30. McCaffrey DF, Ridgeway G, Morral AR. Propensity score estimation with boosted regression for evaluating causal effects in observational studies. *Psychological Methods* 2004;4:403–425. [PubMed: 15598095]
31. 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.
32. Cochran WG, Rubin DB. Controlling bias in observational studies: A review. *Sankhya: The Indian Journal of Statistics, Series A* 1973;35:417–446.
33. Heckman JJ, Ichimura H, Todd PE. Matching as an econometric evaluation estimator: Evidence from evaluating a job training program. *Review of Economic Studies* 1997;64:605–54.
34. 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.
35. Rosenbaum PR. Sensitivity analysis for certain permutation inferences in matched observational studies. *Biometrika* 1987;74:13–26.
36. D’Agostino RB Jr, 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.
37. Haviland A, Nagin DS, Rosenbaum PR. Combining propensity score matching and group-based trajectory analysis in an observational study. *Psychological Methods* 2007;12(3):247–267. [PubMed: 17784793]
38. 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.
39. 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.
40. Normand SLT, 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. *British Medical Journal* 2005;330:1021–1023. [PubMed: 15860831]

A: 1:1 nearest neighbor matching

B: Subclassification

Figure 1.

Results of 1:1 nearest neighbor matching with replacement and 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 to conventional antipsychotics. In Panel B, vertical lines indicate subclass dividers.

Table 1

Recommended Steps in Analyzing Non-Experimental Studies

Step	Rationale
1	Define the treatment and comparison group(s) (7,8).
2	Create the treatment groups (2,8,9).
3	Assess the potential for confounding using standardized differences and plots (10,11).
4	Estimate the treatment effect on the treatment groups created in Step 2 (12,13).
5	Determine robustness of conclusions to unmeasured confounders (5,14,15).

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%
African American	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 [§]

^aFor 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 pancreas disease; HIV/AIDS; cancer; and/or seizure disorder.

[£]Defined as 2 or more claims for β blockers; corticosteroids; valproate; loop and thiazide diuretics; and/or protease inhibitors.

[§]Number of matched pairs.

Table 3

Estimated Absolute Risk (%) of Adverse Metabolic Outcomes of Atypical Compared to Conventional Antipsychotic Medication Use. P-value in parentheses. Numbers greater than 0 indicate higher risk for individuals taking atypical antipsychotics.

Outcome	Regression adjustment (# Subjects = 6751)	Propensity Score-Based Analyses		
		Matching (# Pairs = 3384)	Weighting (# Subjects = 6751)	Subclassification [§] (# Subjects = 6751)
Dyslipidemia	1.67(0.03)	1.04 (0.26)	1.66 (0.03)	1.92 (0.01)
Type II 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)

[§] Average effect calculated by taking a precision-weighted average of the subclass-specific effects shown in Table 4.

Estimated Absolute Risk (%) of Adverse Metabolic Outcomes of Atypical Compared to Conventional Antipsychotic Medication Use Stratified by Propensity Score Subclass. P-value in parentheses. Numbers greater than 0 indicate higher risk for atypical users.

Table 4

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 II 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.

Table 5

Sensitivity of atypical antipsychotic effect on obesity to an unmeasured confounder. 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.

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