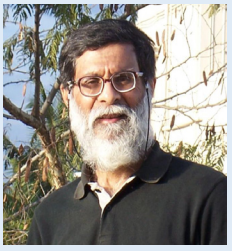
Clinical and Practical **Psychopharmacology**

Propensity Score Matching in Nonrandomized Studies: A Concept Simply Explained Using Antidepressant Treatment During Pregnancy as an Example

Chittaranjan Andrade, MD



Each month in his online column, Dr Andrade considers theoretical and practical ideas in clinical psychopharmacology with a view to update the knowledge and skills of medical practitioners who treat patients with psychiatric conditions.

Department of Psychopharmacology, National Institute of Mental Health and Neurosciences, Bangalore, India (candrade@psychiatrist.com).

ABSTRACT

In prospective and retrospective observational studies, such as those that examine the effects of antidepressant drugs for the treatment of depression during pregnancy, patients are not randomized to whatever treatments they do or do not receive. As a result, treatment groups may be unbalanced for a wide range of sociodemographic and clinical variables. In such studies, because the illness (and its correlates) for which the treatment is indicated may itself influence the study outcomes, the use of the treatment becomes indirectly linked to these outcomes (this is known as confounding by indication), and the effects of treatment and illness on the study outcomes are difficult if not impossible to separate. Case-control research designs, regression analyses that adjust for independent variables, and propensity score matching are ways in which baseline differences between groups and confounding by indication are statistically addressed. This article examines the concepts involved, explains what is done in propensity score matching procedures, and discusses the advantages and limitations of propensity score matching. Whereas propensity score matching can substantially reduce baseline differences between groups in observational studies, it can never correct for unmeasured confounds; therefore, cause-effect relationships can never be deduced from such studies. However, in situations in which gold standard randomized controlled trials are impractical, researchers must make do with statistical approaches such as propensity score matching.

J Clin Psychiatry 2017;78(2):e162–e165
https://doi.org/10.4088/JCP.17f11446

Introduction

Is it safe to use an antidepressant drug to treat depression during pregnancy? We do not have a definitive answer to this question because it has not been addressed in randomized controlled trials (RCTs) and may never be, due to ethical concerns. However, there are many nonrandomized studies on the subject. The data in some instances have been drawn from prospective, observational studies; however, for the most part, the data have been retrospectively ascertained from pregnancy registries, insurance claims databases, and other electronic record systems. There are concerns about conclusions drawn from prospective or retrospective nonrandomized studies, as the subsequent sections indicate. This article provides a simple explanation about certain statistical methods, especially propensity score matching, that may be used to address the concerns.

Biases Due to the Absence of Randomization

When subjects are not randomized to their respective treatments, there may be significant differences between groups. These differences may influence the study outcomes. For example, Boukhris et al[1] conducted a register-based study of women who were or were not prescribed an antidepressant drug during pregnancy. There were many significant sociodemographic differences between these 2 groups of women: those who received antidepressants were older, less well educated, more likely to be living alone, and more likely to be recipients of social assistance. They were also more likely to have diabetes and hypertension. There could also have been other, unmeasured, differences between the groups, including smoking and illicit substance use. Any of the measured or unmeasured differences, instead of the treatment that defined the groups, per se, could be responsible for adverse pregnancy and neurodevelopmental outcomes identified in the study.

Confounding by Indication

As a special case of bias due to the absence of randomization, conclusions drawn from nonrandomized studies are not definitive because of the possibility of confounding by indication.[2] That is (for example), if an adverse outcome is identified to be associated with antidepressant drug treatment, the adverse outcome may not be due to the drug; rather, it may arise from the indication for which the drug was prescribed.

Consider a hypothetical study that used data drawn from an electronic database. The study compared women who did and did not receive selective serotonin reuptake inhibitor (SSRI) antidepressant drugs during pregnancy. This hypothetical study found that SSRI treatment was associated with an increased risk of the neonate being small for gestational age. It is possible that the SSRIs compromised fetal growth. It is also possible that depression was associated with poor appetite, decreased maternal weight gain during pregnancy, and/or compromised nutritional quality during pregnancy, and that these consequences of depression, rather than SSRI treatment, were responsible for the observed small

Clinical Points

- In prospective and retrospective observational studies that examine the effects of treatments, patients are not randomized to their respective treatments. This may often result in substantial baseline differences between treatment groups. These baseline differences can influence the outcomes of interest in the studies.
- Case-control research designs, regression analyses, and propensity score matching are ways in which baseline differences between groups are statistically addressed.
- In propensity score matching, a propensity score is first calculated using regression analysis, based on a large number of potential explanatory variables. This score is the probability that a subject will receive a particular treatment. Subjects in the different treatment groups are then matched, based on their propensity scores.
- Propensity score matching substantially reduces baseline differences between groups but can never adjust for unmeasured confounds. Propensity score matching, therefore, can never result in definitive conclusions about cause-effect relationships between treatments and outcomes. However, propensity matching is useful when randomized controlled trials, the gold standard, are impractical.

size for gestational age. The latter possibility exemplifies confounding by indication: a drug is blamed for an adverse outcome when it is the illness (for which the drug was prescribed) that was responsible for the adverse outcome.

As an aside, note that the opposite may also happen. For example, if tricyclic antidepressant drugs are deliberately not prescribed for patients with preexisting cardiac conditions, then their apparent cardiac safety, in persons who receive these drugs, may also be due to confounding by indication.

Comparing Treatment and No Treatment Specifically in Subjects With Illness

One way of getting around confounding by indication in nonrandomized studies of antidepressant treatment of depression during pregnancy is to compare only depressed women who did or did not receive an antidepressant drug. Because both groups of women are depressed, confounding by indication ought not to arise. This was done, for example, by Oberlander et al.[3] However, when subjects are not randomized to treatment or no treatment categories, biases may decide whether or not a subject receives treatment, and these biases may influence outcomes. For example, SSRIs may be prescribed only for women with more severe depression, and greater severity of depression may be associated with unmeasured behaviors (eg, poorer nutrition, poorer compliance with gynecologic/obstetric guidance) that worsen pregnancy outcome. So, the antidepressant drug is again linked to the poorer pregnancy outcomes. This, again, exemplifies confounding by indication; the problem is not eliminated by restricting analysis to depressed women.

In the study referred to above,[3] women who received SSRIs had a larger number of psychiatric visits during the year before becoming pregnant, had a larger number of depression diagnoses in the previous year, and had a larger number of non-depression diagnoses in the previous year. These indicated that the women who received SSRIs had had a worse illness course and may have had more severe depression during pregnancy, explaining why the SSRIs were prescribed.

Selecting Matched Controls for Every Case

Investigators may try to get around the problem of nonrandomization by selecting matched controls from the same population from which the cases were drawn. For example, if a woman who received an antidepressant is considered a case, then 1, 5, or even 10 antidepressant-untreated controls per case can be selected from the database with control-to-case matching performed for age, diagnosis, and other variables. The problems here are many. Matching on a few variables will certainly not suffice to adjust for important group differences, and matching for many variables will make it hard to find appropriate controls for every case. Next, some group differences on measured variables are likely to persist despite the best efforts at matching. Finally, group differences could also be present on unmeasured variables, including genetic variables, and variables whose contextual importance has not yet been discovered. Thus, case-control matching is not a solution for confounding.[4]

A Very Simple Introduction to Regression

As boys become older, they grow taller. If data are obtained from a sample of growing boys, a simple equation can very easily be derived to describe the mathematical relationship between age and height. This is linear regression; 1 independent variable (age) is modeled with 1 dependent variable (height).

The same concept can be extended to a set of independent variables. Thus, the effects of age, sex, race, family income, and other relevant variables can be mathematically modeled to predict height in growing children. This is multiple regression: several independent variables are modeled with 1 dependent variable.

When the dependent variable is dichotomous, as in alive/dead, euthymic/relapsed, or received/did not receive SSRIs during pregnancy, the analysis performed is known as logistic regression.

Propensity Score Matching

Propensity score matching is one way of getting around the problem of differences between groups that result from nonrandomization. As a first step, the investigators use a large number of potential explanatory variables in a logistic regression procedure to mathematically model the relationship between these variables and the grouping variable.[5,6] For example, Oberlander et al[3] used over a dozen explanatory variables to mathematically model women who did or did not receive an SSRI to treat depression during pregnancy. These explanatory variables included age, income, number of prenatal visits, number of physician

visits, number of visits to a psychiatrist, number of times diagnosed as depressed, number of times diagnosed with a mental health disorder other than depression, and so on.[3]

Using the results of the logistic regression equation, and based on the explanatory variables entered in the equation, the investigators next calculate a propensity score for each subject. This propensity score is the probability that the subject would belong to the group of interest, based on the explanatory variables studied. Propensity scores, like probability values, can range from 0 to 1.

In the example cited above,[3] the explanatory variables listed were used to generate a propensity score that indicated the probability that a woman would receive an SSRI for the treatment of her depression. So, a high propensity score could identify a woman who did receive an SSRI, or it could identify a woman who, based on the explanatory variables, was a likely candidate to receive an SSRI, but did not (for whatever reason).

After the propensity scores are generated, propensity score matching can be performed. For every woman with a high propensity score who did receive an SSRI, a woman with a high propensity score who did not receive an SSRI is selected. Similarly, for every woman with a low propensity score who did receive an SSRI, a woman with a low propensity score who did not receive an SSRI is selected. Thus, case to control matching is based on propensity scores.

As a result of propensity score matching, treatment groups tend to be closely similar on all the important measured independent variables. As an example, Vlenterie et al[7] described an observational study of the effects of gestational exposure to paracetamol on child neurodevelopment at 18 months. They found a very large number of baseline differences in the independent variables between pregnancies that were (n = 1,787) and were not (n = 30,451) exposed to paracetamol. In the propensity-matched subsample (n = 1,630 in each group), none of the previously observed between-groups differences in the independent variables remained significant. Thus, the situation approximated that of an RCT in which the groups are similar at baseline. Note that propensity matching resulted in an attenuation of the samples because not every woman could be propensity matched.

Propensity Score Matching Methods

There are many different ways in which propensity scores are matched.[6,8] Here are 2 simple examples:

1. On the basis of identical or very similar propensity scores; all the matched pairs are then compared in subsequent analyses.
2. On the basis of membership in the same stratum (eg, propensity score quintile); matched pairs are then compared within strata in separate analyses to understand how the treatment influences the study outcome within each stratum of probability of receiving the treatment. Alternately, the analyses can be pooled across strata.

Propensity scores can be used in other ways, too.[6] For example, the effect of treatment on outcome can be mathematically modeled with the propensity score as a covariate; this reduces confounding by adjusting for the probability that the subject would receive the treatment. Expressed otherwise, the analysis would tell us that, regardless of the biases that influence the likelihood that the treatment would be prescribed, the treatment has (or does not have) a certain effect on the outcome of interest in the study.

Propensity scores can also be used to generate statistical weights for each subject to create a sample in which the distribution of the potential confounding variables is independent of the treatment. As a result, the effect of treatment on outcome can be modeled with reduced bias.[6] There are other methods of matching, too. Each method has its strengths and limitations. Because different methods result in different subsets of subjects, different methods can yield different findings.[8]

Advantages of Propensity Score Matching

Propensity score matching, as already stated, results in groups that are similar in most if not all measured baseline characteristics; the expectation is that, once the propensity score is accounted for, confounding by indication is attenuated or eliminated.[8] A further advantage is that propensity matching can substantially reduce the number of independent variables that need to be used in the final analysis; that is, rather than having many covariates, they are combined into a single propensity score. Lastly, whereas propensity score matching cannot establish cause-effect relationships, as can RCTs, it can provide useful guidance for situations in which RCTs cannot be conducted.[6]

Limitations of Propensity Score Matching

The propensity score is derived from a large number of independent variables that have the potential to influence the grouping variable (eg, received vs did not receive an SSRI during pregnancy). However, it cannot account for variables that have not been measured or that do not exist in the database. This means that, for example, if severity of the depression was not measured, then the severity of the depression cannot be accounted for in the propensity score. That is, confounding for indication, where severity of depression is the confound, remains uneliminated. Oberlander et al[3] used several proxies for severity of depression as described earlier, but proxies are no guarantee that propensity score matching will result in groupings that approximate what is observed in RCTs.

Simply expressed, propensity score matching can only adjust for measured independent variables. It cannot adjust for (known or unknown) unmeasured independent variables. What is unmeasured can result in what is known as residual confounding. As an example, if smoking and illicit drug use were not measured during pregnancy, then, despite propensity matching, we cannot exclude smoking and illicit drug use as possible explanations for adverse outcomes in depressed women who received an SSRI during pregnancy.

That failure of propensity score matching is not a hypothetical possibility was well demonstrated by Freemantle et al[5]: a propensity-matched analysis of observational data on the effect of spironolactone on all-cause mortality yielded strikingly different results from those obtained from an RCT; confounding by indication was a likely explanation.

Finally, another limitation is that if the propensity score distributions are very different in the two groups, only subjects whose scores overlap can be included. This will result in a reduction in sample size and in potential biases in terms of who is included in the study.[8]

General Notes

Statistical adjustments performed on observational data in regression analyses, propensity-matched analyses, and other analyses, can fail for several reasons[2]:

1. All important independent variables may not have been measured. The variables missed are therefore sources of residual confounding.
2. The independent variables may not have been measured using an appropriate instrument, or may be too broadly defined; the observations are therefore approximations of the actual data, leaving scope for residual confounding.

Some authors have recommended using different approaches to propensity score matching analysis of the same set of data; the additional analyses would be sensitivity analyses.[8]

Concluding Notes

Using large sample RCTs with simple or stratified randomization into the groups of interest is the best approach to balance groups for measured as well as unmeasured confounds. Propensity score matching is an approximation method that can be applied to observational data in situations, such as pregnancy, in which RCTs are unlikely to be performed. Readers need to keep in mind the possibility that confounding by indication and residual confounding may generate misleading results when observational data are analyzed using propensity score matching. Propensity score matching cannot establish cause-effect relationships, as can RCTs. In propensity score matching studies, if (for example) an antidepressant drug is linked to a particular outcome, rather than assume a cause-effect relationship, it would be more meaningful to consider that the drug is a marker for that outcome.

Acknowledgment: Dr Andrade thanks Prof David Streiner, PhD, CPsych, Department of Psychiatry and Behavioral Neurosciences, McMaster University, and Department of Psychiatry, University of Toronto, for his careful reading of a previous version of this manuscript and suggestions for its improvement.

REFERENCES

1. Boukhris T, Sheehy O, Mottron L, et al. Antidepressant use during pregnancy and the risk of autism spectrum disorder in children. *JAMA Pediatr*. 2016;170(2):117–124.
2. Kyriacou DN, Lewis RJ. Confounding by indication in clinical research. *JAMA*. 2016;316(17):1818–1819.
3. Oberlander TF, Warburton W, Misri S, et al. Neonatal outcomes after prenatal exposure to selective serotonin reuptake inhibitor antidepressants and maternal depression using population-based linked health data. *Arch Gen Psychiatry*. 2006;63(8):898–906.
4. Pearce N. Analysis of matched case-control studies. *BMJ*. 2016;352:i969.
5. Freemantle N, Marston L, Walters K, et al. Making inferences on treatment effects from real world data: propensity scores, confounding by indication, and other perils for the unwary in observational research. *BMJ*. 2013;347:f6409.
6. Haukoos JS, Lewis RJ. The propensity score. *JAMA*. 2015;314(15):1637–1638.
7. Vlenterie R, Wood ME, Brandlistuen RE, et al. Neurodevelopmental problems at 18 months among children exposed to paracetamol in utero: a propensity score matched cohort study [published online ahead of print August 31, 2016]. *Int J Epidemiol*.
8. Streiner DL, Norman GR. The pros and cons of propensity scores. *Chest*. 2012;142(6):1380–1382.