

*This is the peer reviewed version of the following article: Wood, M. E., Lapane, K. L., van Gelder, M. M. H. J., Rai, D., & Nordeng, H. M. E. (2017). Making fair comparisons in pregnancy medication safety studies: An overview of advanced methods for confounding control. *Pharmacoepidemiology and Drug Safety*, (February), 1–8, which has been published in final form at <https://doi.org/10.1002/pds.4336>. This article may be used for non-commercial purposes in accordance with Wiley Terms and Conditions for Use of Self-Archived Versions.*

Making fair comparisons in pregnancy medication safety studies: an overview of advanced methods for confounding control.

Mollie E. Wood¹, Kate L. Lapane², Marleen M.H.J. van Gelder^{3,4}, Dheeraj Rai⁵, Hedvig Nordeng^{1,6}

Affiliations:

1. PharmacoEpidemiology and Drug Safety Research Group, School of Pharmacy, University of Oslo, Norway
2. Department of Quantitative Health Sciences, University of Massachusetts Medical School, Worcester, MA, United States
3. Department for Health Evidence, Radboud Institute for Health Sciences, Radboud university medical center, Nijmegen, The Netherlands
4. Radboud REshape Innovation Center, Radboud university medical center, Nijmegen, The Netherlands
5. School of Social and Community Medicine, University of Bristol, United Kingdom
6. Department of Child Mental and Physical Health, Norwegian Institute of Public Health, Oslo, Norway

Current word count (main text only): 3098

References: 56

Figures: 2

Tables: 1

Supplemental Material: Box 1, Figure S1

*This is the peer reviewed version of the following article: Wood, M. E., Lapane, K. L., van Gelder, M. M. H. J., Rai, D., & Nordeng, H. M. E. (2017). Making fair comparisons in pregnancy medication safety studies: An overview of advanced methods for confounding control. *Pharmacoepidemiology and Drug Safety*, (February), 1–8, which has been published in final form at <https://doi.org/10.1002/pds.4336>. This article may be used for non-commercial purposes in accordance with Wiley Terms and Conditions for Use of Self-Archived Versions.*

Abstract.

Understanding the safety of medication use during pregnancy relies on observational studies: however, confounding in observational studies poses a threat to the validity of estimates obtained from observational data. Newer methods, such as marginal structural models and propensity calibration, have emerged to deal with complex confounding problems, but these methods have seen limited uptake in the pregnancy medication literature. In this article, we provide an overview of newer advanced methods for confounding control and show how these methods are relevant for pregnancy medication safety studies.

Key Points.

- Studies of the safety of medication use during pregnancy depend mainly on observational studies, which are subject to confounding bias.
- Novel methods for confounding control have seen limited uptake in the pregnancy medication safety literature.
- Application of novel methods is necessary to appropriately address the complex confounding scenarios found in pregnancy studies.

*This is the peer reviewed version of the following article: Wood, M. E., Lapane, K. L., van Gelder, M. M. H. J., Rai, D., & Nordeng, H. M. E. (2017). Making fair comparisons in pregnancy medication safety studies: An overview of advanced methods for confounding control. *Pharmacoepidemiology and Drug Safety*, (February), 1–8, which has been published in final form at <https://doi.org/10.1002/pds.4336>. This article may be used for non-commercial purposes in accordance with Wiley Terms and Conditions for Use of Self-Archived Versions.*

More than half of all pregnant women in western countries take medication during pregnancy,¹⁻³ making studies of medication safety a pressing public health concern. Studying medication safety in pregnancy presents particular challenges: effects of medications on fetal development can be unpredictable, vulnerability to exposure changes during pregnancy, and outcomes may occur early in fetal development but be detected later.⁴ In the general population, knowledge of medication efficacy and safety is primarily based on randomized controlled trials. However, randomized trials routinely exclude pregnant women due to uncertainties about the effects of medications on fetal development, meaning that studies of medication safety in pregnancy must rely on reproductive toxicity studies in animals and on observational data in humans. Several landmark cases, such as the thalidomide disaster, have taught us that animal models for teratogenicity do not necessarily translate to humans. Observational studies, using data from cohort studies, registries, and administrative databases,⁵ are opportunities for understanding the risks of medication use in pregnancy, and in 2005, the Food and Drug Administration (FDA) acknowledged that observational studies are the best method for assessing the maternal and fetal safety of using medication during pregnancy.⁶ However, confounding is a major source of bias in observational studies. Recent years have seen the rapid development of advanced methods for dealing with confounding; yet, uptake of these methods has been slow in the pregnancy medication literature. This is unfortunate, because in this field, it is arguably especially important that researchers use the best methods for confounding control, because the consequences for getting the wrong answer are so profound: failing to detect true effects of medication exposure can have enormous effects in the population, and falsely raising the alarm for a safe drug can result in women forgoing needed therapies and in some cases, terminating wanted pregnancies.⁶

*This is the peer reviewed version of the following article: Wood, M. E., Lapane, K. L., van Gelder, M. M. H. J., Rai, D., & Nordeng, H. M. E. (2017). Making fair comparisons in pregnancy medication safety studies: An overview of advanced methods for confounding control. *Pharmacoepidemiology and Drug Safety*, (February), 1–8, which has been published in final form at <https://doi.org/10.1002/pds.4336>. This article may be used for non-commercial purposes in accordance with Wiley Terms and Conditions for Use of Self-Archived Versions.*

In this paper, we advocate for a greater use of advanced methods for confounding control in the pregnancy medication safety research field, and provide an overview of these methods under the following framework:

- (1) How does this method help us to make fair comparisons between the exposed and unexposed groups?
- (2) How has this method been applied in the pregnancy medication literature?
- (3) How is the method used in practice?
- (4) What are the important assumptions for this method?
- (5) What are the major strengths and limitations of the method?

Table 1 provides an outline of pregnancy medication studies using advanced methods to deal with confounding. This paper gives a useful reference for both students and experienced researchers who wish to gain new skills in advanced methods for confounding control.

Confounding in pregnancy medication studies

Confounding control begins with a review of the literature and consultation with subject-area experts. Directed acyclic graphs (DAGs) provide a graphical means to represent the causal structure the investigator believes is present,⁷ and guide study design, data collection, and analysis. Figure 1 is an example DAG showing one possible causal model for prenatal antidepressant exposure and childhood neurodevelopment, with potential biasing paths, including confounders (other psychiatric illness, other psychiatric medication use, depression severity, and genetics) which should be controlled as far as possible, as well as a mediator (gestational age), and a collider (live birth). Several non-biasing paths, including a risk factor for the outcome that is unrelated to the exposure (child gender) and a predictor of exposure that is unrelated to the outcome (pre-pregnancy antidepressant use) are also shown. Obtaining unbiased effect estimates requires investigators to identify and control confounding, while avoiding bias from inappropriate control for colliders and mediators and loss of precision or confusing interpretation of estimates arising from control for factors only related to the exposure or

outcome but not both.⁸ The supplemental material contains a more comprehensive review of definitions of confounding, counterfactuals, and causal inference.

I. Methods for measured confounders

In Box 1 (supplemental material), we include a simplified illustration of confounding by measured factors and the methods to address confounding.

Confounder summary scores and marginal structural models work by reducing a large amount of information about an individual into a single summary score. Two individuals can have the same summary score but different individual confounder values (e.g., a woman with a propensity score for antidepressant use of 0.5 might be an unemployed smoker with anxiety, or a non-smoking lawyer with depression), but because their distribution of confounders is equivalent, any differences in outcome will be attributable only to exposure to the drug of interest. Fair comparisons between exposure groups can then be made because within each stratum of exposure, the distribution of common causes of exposure and outcome are the same.

I.A. Propensity Scores (and other confounder summary scores)

The propensity score (PS), which is the probability of exposure given observed confounders,⁹ reduces a large set of confounders to a single summary score. Propensity scores are commonly used in the medical literature; however, other summary score methods, including disease risk scores¹⁰ (preferred in the case of rare exposures) and polygenic risk scores¹¹ (useful for cases when genetic confounding) are available.

Propensity scores are typically constructed using multivariable logistic regression, where exposure is the dependent variable and confounders are the independent variables. The PS model should include variables that are confounders or predictors of the outcome; inclusion of factors that are only predictors of exposure will increase variance without decreasing bias.¹² High dimensional propensity scores, which include thousands of variables identified through computational algorithms, may also be useful for adjusting for unmeasured confounders, if the measured variables are partial proxies for the

unmeasured confounders.¹³ The PS can be used to match, stratify, adjust, or weight the outcome model. Propensity scores, including high dimensional PS, have seen increased uptake in the pregnancy literature, i.e. safety studies on ondansetron,¹⁴ lithium,¹⁵ antidepressants,¹⁶ and statins¹⁷ in pregnancy, but their use is still minimal compared to multivariable regression (Table 1). Box 1, in the supplemental material, gives a simplified explanation of PS matching and weighting.

Assumptions: Use of PS requires several assumptions, including exchangeability (no unmeasured confounding) and positivity (nonzero probability of treatment). Neither assumption is formally testable. Positivity can be addressed by ensuring that the women in the sample all have the indication for the medication (i.e., if assessing safety of antidepressants, all women in the sample should be at risk for treatment) and that no individuals with clear contraindications are included. Exchangeability is never assured; however, sensitivity analyses can yield estimates for how vulnerable an effect estimate may be to unmeasured confounding.

Strengths and Limitations: PS are especially useful when working with a common treatment and rare outcome. They also separate the design of the study (modeling confounding) from modeling the outcome.¹⁸ However, for rare exposures, summary scores do not perform particularly well.¹⁹ In addition, use of PS methods may produce the appearance of effect modification and/or result in residual confounding in case control or case cohort studies²⁰ or in cohort studies where exposure is misclassified.²¹

I.B. Marginal Structural Models

Marginal structural models (MSM) address time-varying exposure and confounding.^{22,23} Rules for confounder adjustment state we must adjust for common causes of the exposure and outcome, but should not adjust for factors on the causal pathway. In the case of time-varying exposure and confounding, we encounter a double bind: factors that are confounders in one part of the causal structure are mediators in another part (Figure S1A). For example, when studying the safety of antidepressants, we may wish to control for depression severity. However, antidepressant use in earlier

*This is the peer reviewed version of the following article: Wood, M. E., Lapane, K. L., van Gelder, M. M. H. J., Rai, D., & Nordeng, H. M. E. (2017). Making fair comparisons in pregnancy medication safety studies: An overview of advanced methods for confounding control. *Pharmacoepidemiology and Drug Safety*, (February), 1–8, which has been published in final form at <https://doi.org/10.1002/pds.4336>. This article may be used for non-commercial purposes in accordance with Wiley Terms and Conditions for Use of Self-Archived Versions.*

pregnancy predicts depressive symptoms in later pregnancy, which will also predict subsequent antidepressant use. Standard adjustments for depression severity will always be biased in this scenario.

Central to the MSM is the inverse probability of treatment weight (IPTW). At each measurement time t , the investigator uses logistic regression to construct the numerator (probability of exposure) and denominator (probability of exposure, given baseline predictors and history of exposure at time $t-1$).²⁴ The total weight is the product of the weights at each time point, and analyses are conducted in the weighted population, or *pseudo-population*, in which individuals who are likely to be exposed are down-weighted, while those who are unlikely to be exposed are up-weighted, producing balance of measured confounders within strata of exposure.

Use of MSMs for pregnancy medication safety studies remains rare,^{25,26} despite examples where timing of exposure is of great importance, and exposure is conditional on time-varying confounders, such as other medication use, or changes in disease severity.

Assumptions: Under assumptions of positivity, exchangeability, and consistency, the MSM will give an unbiased estimate of the effect of the exposure on the outcome. These assumptions are not formally testable, although assessment of the positivity assumption may include evaluation of the IPTW for extreme weights, and progressive truncation of the weights to determine whether extreme weights are highly influential.²⁷ When important confounders are unmeasured or incompletely measured, marginal structural model methods will not provide unbiased effect estimates.

Strengths and Limitations: The key strength of the MSM is that it allows consideration of time-varying exposure and confounding, which is highly relevant in pregnancy research due to the changes in fetal vulnerability through the course of pregnancy, and the tendency of women to change their medication use during pregnancy.^{28,29} However, when the treatment-covariate association is very strong, MSMs can produce very wide confidence intervals which fail to include the true effect.²⁷

II. Methods for incomplete confounder data

*This is the peer reviewed version of the following article: Wood, M. E., Lapane, K. L., van Gelder, M. M. H. J., Rai, D., & Nordeng, H. M. E. (2017). Making fair comparisons in pregnancy medication safety studies: An overview of advanced methods for confounding control. *Pharmacoepidemiology and Drug Safety*, (February), 1–8, which has been published in final form at <https://doi.org/10.1002/pds.4336>. This article may be used for non-commercial purposes in accordance with Wiley Terms and Conditions for Use of Self-Archived Versions.*

Failure to adjust for unmeasured confounders results in biased effect estimates (Figure S1B). In some situations, the confounder of interest was not measured in the original dataset, but was measured in a similar sample. In this scenario, confounder adjustment is possible, even if the outcome has not been measured in this sample, using PS calibration.^{30–32} PS calibration is a method based on regression calibration³³ that offers an additional advantage over other methods of calibration,³⁴ by allowing for adjustment for multiple confounders. For example, in a study of triptan safety, we used a cross-sectional study to jointly adjust estimates for migraine severity and type.³⁵

In this method, two PS must be calculated: the error-prone PS (estimated in both the main and validation studies, including only the confounders available in the main study) and the gold-standard PS (estimated in the validation study, including all confounders). The outcome model is fitted using the difference between the error-prone and gold standard PS to calibrate effect estimates.

Assumptions: In addition to the assumptions of PS models, outlined previously, PS calibration also assumes that the validation sample is a reasonable stand-in for the main sample and that the measurement error model is correctly specified.^{30,31} PS calibration also assumes *surrogacy*, meaning that the error-prone PS is an adequate surrogate for the gold-standard PS.³⁶ If the outcome is not measured in the validation study, the surrogacy assumption is not testable. Violations of surrogacy occur when the direction of confounding differs between the main and validation studies,³⁰ and bias arising from violations of surrogacy can be predicted.³⁶

Other methods exist for unmeasured confounding, including weighting by the inverse probability of missingness, as well as standard imputation techniques, and a comparison of these methods with PS calibration showed little material differences in bias reduction.³⁷

Strengths and Limitations: The main strength of PS calibration allows for adjustment for multiple unmeasured confounders. However, calibration methods fail when unmeasured confounding is strong, and violations of the surrogacy assumption may result in increased bias.

III. Methods for unmeasured confounding

Information on confounders may be too difficult to measure (e.g., family environment or parenting style) or too costly (e.g. deep sequencing genetic data). The methods discussed below exploit aspects of observational data to control for measured and unmeasured confounders.

III.A. Sibling Comparison Designs

If the unmeasured confounders are shared between siblings (see Figure S1C for illustration), then studies examining with discordant exposure allows researchers to remove bias from shared confounders.^{38–40} If, for example, we believe that any differences in autism risk between children with and without prenatal exposure to antidepressants is due to inherited genetic risk, then comparing the autism diagnosis between pairs of siblings with different prenatal exposure should be less biased than comparing autism risk between unrelated exposed and unexposed groups.

There has been substantial uptake of sibling study designs in the pregnancy medication safety literature in recent years, particularly in studies examining the safety of antidepressants, where the main concern is separating the underlying genetic and familial components of depression from exposure to antidepressant medications.^{41,42}

Assumptions: Use of sibling designs is most appropriate when confounders that are shared between siblings are more important than unshared,³⁹ and there are no carryover effects between siblings.⁴³

Strengths and Limitations: Sibling designs control measured and unmeasured confounding that is shared between siblings. However, failing to control for unshared confounders increases bias; sibling studies are also more vulnerable to bias from measurement error than non-sibling studies.³⁹

III.B. Instrumental Variables

Instrumental variable (IV) methods^{44,45} require identifying a variable whose effect on the outcome occurs only through the exposure: an example of a perfect

*This is the peer reviewed version of the following article: Wood, M. E., Lapane, K. L., van Gelder, M. M. H. J., Rai, D., & Nordeng, H. M. E. (2017). Making fair comparisons in pregnancy medication safety studies: An overview of advanced methods for confounding control. *Pharmacoepidemiology and Drug Safety*, (February), 1–8, which has been published in final form at <https://doi.org/10.1002/pds.4336>. This article may be used for non-commercial purposes in accordance with Wiley Terms and Conditions for Use of Self-Archived Versions.*

instrument is a coin toss assigning an individual to exposure or non-exposure, while commonly used instruments include provider prescription preference and calendar time. One example is a study of antidepressant (AD) efficacy during pregnancy using provider preference, calendar time as a function of FDA recommendations, and geographic differences in AD use as instruments; however, these instruments were only weakly associated with the treatment, which may have contributed to the equivocal findings.⁴⁶ IV studies are often conducted using a two stage least squares methods, where in the first stage, the instruments are used as explanatory variables in a model predicting the exposure, and the predicted values from this first stage are used as predictors in the outcome model. Identifying a strong instrument that meets all assumptions is challenging, which has contributed to the slower adoption of this method. Mendelian randomization, which uses a genetic marker as an instrument, is a subtype of instrumental variable analysis;⁴⁷ while Mendelian randomization has not yet been used in pregnancy medication studies, studies estimating the effect of alcohol use during pregnancy on later neurocognitive outcomes have used the genetic variants encoding alcohol dehydrogenase, an enzyme that metabolizes alcohol, with some success.⁴⁸

Assumptions: Instrumental variable analyses allow for unbiased effect estimation under strict assumptions: (i) the instrument has a causal effect on the exposure of interest, (ii) the instrument affects the outcome only through the exposure, not through any other pathways, (iii) there are no common causes or confounders of the instrument-outcome pathway (Figure S1D).

Strengths and Limitations: Instrumental variable analyses control measured and unmeasured confounding, and so instruments that meet all the assumptions will mimic the results from a randomized trial. However, estimates are highly sensitive to violations of untestable assumptions, and violations may produce bias amplification.⁴⁴

Figure 2 guides readers through selecting a method or methods, based on characteristics of confounder data. The most important first step is to draw a DAG or

*This is the peer reviewed version of the following article: Wood, M. E., Lapane, K. L., van Gelder, M. M. H. J., Rai, D., & Nordeng, H. M. E. (2017). Making fair comparisons in pregnancy medication safety studies: An overview of advanced methods for confounding control. *Pharmacoepidemiology and Drug Safety*, (February), 1–8, which has been published in final form at <https://doi.org/10.1002/pds.4336>. This article may be used for non-commercial purposes in accordance with Wiley Terms and Conditions for Use of Self-Archived Versions.*

DAGs that represent the proposed causal mechanism, without regard to availability of data on confounders: if a confounder is important, it should be included in the DAG, even if the study did not collect data on it. Next, determine which confounders are available in your study, and whether the data support the analytic method. For example, if your DAG shows that medication use and confounders vary over time, but your data shows no such variation, a MSM approach should not be used; if the data cannot identify siblings, this method cannot be used. Most importantly, we urge researchers to consider potential sources of confounding regardless of whether they were measured in the data, and to choose the methods most suited to the data they have available: Figure 2 suggests a systematic way of approaching this process.

A reference to selected software for the methods discussed in this paper is included as part of the supplemental material.

Discussion

Studies of medication use during pregnancy use observational data to answer critical questions of safety and efficacy. More traditional methods for confounding control, such as stratification, restriction, matching, and adjustment have been described in great detail elsewhere, and because of this we have not discussed them here. These older methods have their place in observational research, but as our understanding of the complexities of bias has progressed, so has our understanding of the limitations of these methods. The methods described in this paper were developed to address specific confounding problems, and are necessary to reduce bias, and ultimately to produce the best information possible to health care providers and pregnant women. Using these methods can produce substantially different results from traditional methods, such as when we compare the cohort and sibling studies of antidepressant safety,^{41,42,49} the regression-adjusted to the MSM estimates for triptan safety,²⁵ PS calibrated estimates to standard PS methods,³⁵ or including standard regression, PS methods, sibling controls, and negative paternal controls in one study.⁵⁰

With few exceptions, these methods have seen slow uptake in the pregnancy medication literature. This may be due to a sense of caution about methods that can seem

*This is the peer reviewed version of the following article: Wood, M. E., Lapane, K. L., van Gelder, M. M. H. J., Rai, D., & Nordeng, H. M. E. (2017). Making fair comparisons in pregnancy medication safety studies: An overview of advanced methods for confounding control. *Pharmacoepidemiology and Drug Safety*, (February), 1–8, which has been published in final form at <https://doi.org/10.1002/pds.4336>. This article may be used for non-commercial purposes in accordance with Wiley Terms and Conditions for Use of Self-Archived Versions.*

opaque upon first encounter with the methods paper describing the technique. Caution is necessary when applying novel methods. However, it is also true that the standard regression methods require similar assumptions to the methods discussed in this paper. If readers find that their research question fits well with one of the scenarios described in this paper, we suggest approaching the problem by tackling the citations given for the technique. The techniques we describe in this paper have their roots in standard regression techniques and can be implemented with standard software.

While this paper focuses on bias due to confounding, other sources of bias such as exposure and/or outcome misclassification⁵¹ and selection bias⁵², as well as seasonal effects,⁵³ can also distort associations. This paper is not intended to be an exhaustive discussion of all possible methods for confounding control. New techniques are being developed all the time, and many of these, such as g-estimation^{54,55} and targeted maximum likelihood estimation,⁵⁶ have not yet been implemented in the pregnancy medication literature. Quantitative bias analysis can help researchers account for bias from systematic errors in their data.⁵⁷ Further, the methods discussed herein are not mutually exclusive, and can be used in combination with each other: combining propensity scores with instrumental variables⁴⁶ or marginal structural models with quantitative bias analysis²⁵ gives more information about the probable range of effect estimates than any single method.

Observational studies are vital to our understanding of medication safety in pregnancy, but great care must be taken in the analysis and interpretation of data to minimize confounding and bias. In all pharmacoepidemiological studies sources of bias should be acknowledged and discussed, and preferably quantified by performing sensitivity analysis of estimates under an array of assumptions about possible bias directions and magnitudes.

This is the peer reviewed version of the following article: Wood, M. E., Lapane, K. L., van Gelder, M. M. H. J., Rai, D., & Nordeng, H. M. E. (2017). Making fair comparisons in pregnancy medication safety studies: An overview of advanced methods for confounding control. *Pharmacoepidemiology and Drug Safety*, (February), 1–8, which has been published in final form at <https://doi.org/10.1002/pds.4336>. This article may be used for non-commercial purposes in accordance with Wiley Terms and Conditions for Use of Self-Archived Versions.

References

1. Mitchell AA, Gilboa SM, Werler MM, Kelley KE, Louik C, Hernández-Díaz S. Medication use during pregnancy, with particular focus on prescription drugs: 1976-2008. *Am J Obstet Gynecol*. 2011;205(1):51.e1-8. doi:10.1016/j.ajog.2011.02.029.
2. Headley J, Northstone K, Simmons H, Golding J. Medication use during pregnancy: data from the Avon Longitudinal Study of Parents and Children. *Eur J Clin Pharmacol*. 2004;60(5):355-361. doi:10.1007/s00228-004-0775-7.
3. Lupattelli A, Spigset O, Twigg MJ, et al. Medication use in pregnancy: a cross-sectional, multinational web-based study. *BMJ Open*. 2014;4(2):e004365. doi:10.1136/bmjopen-2013-004365.
4. Savitz DA, Hertz-Picciotto I, Poole C, Olshan AF. Epidemiologic measures of the course and outcome of pregnancy. *Epidemiol Rev*. 2002;24(2):91-101. doi:10.1093/epirev/mxf006.
5. Andrade SE, Bérard A, Nordeng HME, Wood ME, van Gelder MMHJ, Toh S. Administrative Claims Data Versus Augmented Pregnancy Data for the Study of Pharmaceutical Treatments in Pregnancy. *Curr Epidemiol Reports*. 2017:1-11. doi:10.1007/s40471-017-0104-1.
6. *Reviewer Guidance Evaluating the Risks of Drug Exposure in Human Pregnancies.*; 2005. <http://www.fda.gov/cber/guidelines.htm>.
7. Greenland S, Pearl J, Robins JM. Causal diagrams for epidemiologic research. *Epidemiology*. 1999;10(1):37-48. doi:10.1097/00001648-199901000-00008.
8. Bandoli G, Palmsten K, Flores KF, Chambers CD. Constructing Causal Diagrams for Common Perinatal Outcomes: Benefits, Limitations and Motivating Examples with Maternal Antidepressant Use in Pregnancy. *Paediatr Perinat Epidemiol*. 2016;30(5):521-528. doi:10.1111/ppe.12302.
9. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika*. 1983;70(1):41-55.
10. Arbogast PG, Ray WA. Use of disease risk scores in pharmacoepidemiologic studies. *Stat Methods Med Res*. 2009;18(1):67-80.

This is the peer reviewed version of the following article: Wood, M. E., Lapane, K. L., van Gelder, M. M. H. J., Rai, D., & Nordeng, H. M. E. (2017). Making fair comparisons in pregnancy medication safety studies: An overview of advanced methods for confounding control. *Pharmacoepidemiology and Drug Safety*, (February), 1–8, which has been published in final form at <https://doi.org/10.1002/pds.4336>. This article may be used for non-commercial purposes in accordance with Wiley Terms and Conditions for Use of Self-Archived Versions.

doi:10.1177/0962280208092347.

11. Chatterjee N, Shi J, García-Closas M. Developing and evaluating polygenic risk prediction models for stratified disease prevention. *Nat Rev Genet*. 2016;17(7):392-406. doi:10.1038/nrg.2016.27.
12. Brookhart MA, Schneeweiss S, Rothman KJ, Glynn RJ, Avorn J, Stürmer T. Variable selection for propensity score models. *Am J Epidemiol*. 2006;163(12):1149-1156.
13. Schneeweiss S, Rassen J, Glynn RJ, Avorn J, Mogun H, Brookhart MA. High-dimensional propensity score adjustment in studies of treatment effects using health care claims data. *Epidemiology*. 2009;20(4):512-522. doi:10.1097/EDE.0b013e3181a663cc.High-dimensional.
14. Pasternak B, Svanström H, Hviid A. Ondansetron in pregnancy and risk of adverse fetal outcomes. *N Engl J Med*. 2013;368(9):814-823. doi:10.1056/NEJMoa1211035.
15. Paterno E, Huybrechts KF, Bateman BT, et al. Lithium Use in Pregnancy and the Risk of Cardiac Malformations. *N Engl J Med*. 2017;376(23):2245-2254. doi:10.1056/NEJMoa1612222.
16. Huybrechts KF, Palmsten K, Avorn J, et al. Antidepressant use in pregnancy and the risk of cardiac defects. *N Engl J Med*. 2014;370(25):2397-2407. doi:10.1056/NEJMoa1312828.
17. Bateman BT, Hernandez-diaz S, Fischer MA, et al. Statins and congenital malformations: cohort study. *Br Med J*. 2015;350. doi:10.1136/bmj.h1035.
18. Austin PC. An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies. *Multivariate Behav Res*. 2011;46(3):399-424. doi:10.1080/00273171.2011.568786.
19. Glynn RJ, Schneeweiss S, Stürmer T. Indications for propensity scores and review of their use in pharmacoepidemiology. *Basic Clin Pharmacol Toxicol*. 2006;98(3):253-259. doi:10.1111/j.1742-7843.2006.pto_293.x.
20. Maansson R, Joffe MM, Sun W, Hennessy S. On the Estimation and Use of

This is the peer reviewed version of the following article: Wood, M. E., Lapane, K. L., van Gelder, M. M. H. J., Rai, D., & Nordeng, H. M. E. (2017). Making fair comparisons in pregnancy medication safety studies: An overview of advanced methods for confounding control. *Pharmacoepidemiology and Drug Safety*, (February), 1–8, which has been published in final form at <https://doi.org/10.1002/pds.4336>. This article may be used for non-commercial purposes in accordance with Wiley Terms and Conditions for Use of Self-Archived Versions.

- Propensity Scores in Case-Control and Case- Cohort Studies. *Am J Epidemiol*. 2007;166(3):332-339. doi:10.1093/aje/kwm069.
21. Wood ME, Chrysanthopoulou SA, Nordeng HME, Lapane KL. The impact of nondifferential exposure misclassification on the performance of propensity scores for continuous and binary outcomes: a simulation study [ACCEPTED, IN PRESS]. *Med Care*. 2017.
 22. Robins JM, Hernan M, Brumback B. Marginal structural models and causal inference in epidemiology. *Epidemiology*. 2000;11(5):550-560.
 23. Robins JM. Marginal Structural Models versus Structural Nested Models as Tools for Causal Inference. 1999;(1986):1-30.
 24. Hernán MA, Brumback BA, Robins JM. Estimating the causal effect of zidovudine on CD4 count with a marginal structural model for repeated measures. *Stat Med*. 2002;21(12):1689-1709. doi:10.1002/sim.1144.
 25. Wood ME, Lapane K, Frazier JA, Ystrom E, Mick EO, Nordeng H. Prenatal Triptan Exposure and Internalising and Externalising Behaviour Problems in 3-Year-Old Children: Results from the Norwegian Mother and Child Cohort Study. *Paediatr Perinat Epidemiol*. November 2015. doi:10.1111/ppe.12253.
 26. Bodnar L, Davidian M, Siega-Riz AM, Tsiatis A. Marginal Structural Models for Analyzing Causal Effects of Time-dependent Treatments: An Application in Perinatal Epidemiology. *Am J Epidemiol*. 2004;159(10):926-934. doi:10.1093/aje/kwh131.
 27. Cole SR, Hernán MA. Constructing inverse probability weights for marginal structural models. *Am J Epidemiol*. 2008;168(6):656-664. doi:10.1093/aje/kwn164.
 28. Lupattelli A, Spigset O, Björnsdóttir I, et al. Patterns and Factors Associated With Low Adherence To Psychotropic Medications During Pregnancy-a Cross-Sectional, Multinational Web-Based Study. *Depress Anxiety*. 2015;11:n/a-n/a. doi:10.1002/da.22352.
 29. Harris G-ME, Wood M, Eberhard-Gran M, Lundqvist C, Nordeng H. Patterns and

This is the peer reviewed version of the following article: Wood, M. E., Lapane, K. L., van Gelder, M. M. H. J., Rai, D., & Nordeng, H. M. E. (2017). Making fair comparisons in pregnancy medication safety studies: An overview of advanced methods for confounding control. *Pharmacoepidemiology and Drug Safety*, (February), 1–8, which has been published in final form at <https://doi.org/10.1002/pds.4336>. This article may be used for non-commercial purposes in accordance with Wiley Terms and Conditions for Use of Self-Archived Versions.

- predictors of analgesic use in pregnancy: A longitudinal drug utilization study with special focus on women with migraine. *BMC Pregnancy Childbirth*. 2017;17(1). doi:10.1186/s12884-017-1399-0.
30. Stürmer T, Schneeweiss S, Rothman KJ, Avorn J, Glynn RJ. Performance of propensity score calibration- a simulation study. *Am J Epidemiol*. 2007;165(10):1110-1118.
 31. Stürmer T, Schneeweiss S, Avorn J, Glynn RJ. Adjusting effect estimates for unmeasured confounding with validation data using propensity score calibration. *Am J Epidemiol*. 2005;162(3):279-289. doi:10.1093/aje/kwi192.
 32. Sturmer T, Glynn RJ, Rothman KJ, Avorn J, Schneeweiss S. Adjustments for unmeasured confounders in pharmacoepidemiologic database studies using external information. *Med Care*. 2007;45(10(S)):1-15.
 33. Spiegelman D, Mcdermott A, Rosner B. Regression calibration methods for correcting measurement error bias in nutritional epidemiology. *Am J Clin Nutr*. 1997;65:1179S-86S.
 34. Schneeweiss S. Sensitivity analysis and external adjustment for unmeasured confounders in epidemiologic database studies of therapeutics. *Pharmacoepidemiol Drug Saf*. 2006;15(5):291-303. doi:10.1002/pds.1200.
 35. Wood ME, Frazier JA, Nordeng HME, Lapane KL. Prenatal triptan exposure and parent-reported early childhood neurodevelopmental outcomes: an application of propensity score calibration to adjust for unmeasured confounding by migraine severity. *Pharmacoepidemiol Drug Saf*. November 2015. doi:10.1002/pds.3902.
 36. Lunt M, Glynn RJ, Rothman KJ, Avorn J, Stürmer T. Propensity score calibration in the absence of surrogacy. *Am J Epidemiol*. 2012;175(12):1294-1302. doi:10.1093/aje/kwr463.
 37. Toh S, Garcia Rodriguez LA, Hernan M. Analyzing partially missing confounding information in comparative effectiveness and safety research of therapeutics. *Pharmacoepidemiol Drug Saf*. 2012;21(0 2):13-20. doi:10.1002/pds.3248.Analyzing.

This is the peer reviewed version of the following article: Wood, M. E., Lapane, K. L., van Gelder, M. M. H. J., Rai, D., & Nordeng, H. M. E. (2017). Making fair comparisons in pregnancy medication safety studies: An overview of advanced methods for confounding control. *Pharmacoepidemiology and Drug Safety*, (February), 1–8, which has been published in final form at <https://doi.org/10.1002/pds.4336>. This article may be used for non-commercial purposes in accordance with Wiley Terms and Conditions for Use of Self-Archived Versions.

38. Carlin JB, Gurrin LC, Sterne JAC, Morley R, Dwyer T. Regression models for twin studies: A critical review. *Int J Epidemiol*. 2005;34(5):1089-1099. doi:10.1093/ije/dyi153.
39. Frisell T, Öberg S, Kuja-Halkola R, Sjölander A. Sibling comparison designs: bias from non-shared confounders and measurement error. *Epidemiology*. 2012;23(5):713-720. doi:10.1097/EDE.0b013e31825fa230.
40. Keyes KM, Smith GD, Susser E. On sibling designs. *Epidemiology*. 2013;24:473-474. doi:10.1097/EDE.0b013e31828c7381.
41. Nezvalová-Henriksen K, Spigset O, Brandlistuen RE, Ystrom E, Koren G, Nordeng H. Effect of prenatal selective serotonin reuptake inhibitor (SSRI) exposure on birthweight and gestational age: a sibling-controlled cohort study. *Int J Epidemiol*. 2016;(Ci):dyw049. doi:10.1093/ije/dyw049.
42. Viktorin A, Lichtenstein P, Lundholm C, et al. Selective serotonin re-uptake inhibitor use during pregnancy: association with offspring birth size and gestational age. *Int J Epidemiol*. 2016;(January):dyv351-. doi:10.1093/ije/dyv351.
43. Sjölander A, Frisell T, Kuja-Halkola R, Öberg S, Zetterqvist J. Carry-over effects in sibling comparison designs. *Epidemiology*. 2016;27(6). doi:10.1097/EDE.0000000000000541.
44. Hernan MA, Robins JM. Instruments for Causal Inference: an epidemiologist's dream? *Epidemiology*. 2006;17(4):360-372. doi:10.1097/01.ede.0000222409.00878.37.
45. Angrist JD, Imbens GW, Rubin DB. Identification of Causal Effects Using Instrumental Variables. *J Am Stat Assoc*. 1996;91(434):444-455.
46. Swanson SA, Hernandez-Diaz S, Palmsten K, Mogun H, Olfson M, Huybrechts KF. Methodological considerations in assessing the effectiveness of antidepressant medication continuation during pregnancy using administrative data. *Pharmacoepidemiol Drug Saf*. 2015;24(9):934-942. doi:10.1002/pds.
47. Smith GD, Ebrahim S. "Mendelian randomization": Can genetic epidemiology contribute to understanding environmental determinants of disease? *Int J*

This is the peer reviewed version of the following article: Wood, M. E., Lapane, K. L., van Gelder, M. M. H. J., Rai, D., & Nordeng, H. M. E. (2017). Making fair comparisons in pregnancy medication safety studies: An overview of advanced methods for confounding control. *Pharmacoepidemiology and Drug Safety*, (February), 1–8, which has been published in final form at <https://doi.org/10.1002/pds.4336>. This article may be used for non-commercial purposes in accordance with Wiley Terms and Conditions for Use of Self-Archived Versions.

- Epidemiol.* 2003;32(1):1-22. doi:10.1093/ije/dyg070.
48. Zuccolo L, Lewis SJ, Smith GD, et al. Prenatal alcohol exposure and offspring cognition and school performance. A mendelian randomization natural experiment. *Int J Epidemiol.* 2013;42(5):1358-1370. doi:10.1093/ije/dyt172.
 49. Brandlistuen RE, Ystrom E, Eberhard-Gran M, Nulman I, Koren G, Nordeng H. Behavioural effects of fetal antidepressant exposure in a Norwegian cohort of discordant siblings. *Int J Epidemiol.* 2015;44(4):1397-1407. doi:10.1093/ije/dyv030.
 50. Rai D, Lee BK, Dalman C, Newschaffer C, Lewis G, Magnusson C. Antidepressants during pregnancy and offspring autism: population-based cohort study. *Br Med J.* 2017;(article i).
 51. Grzeskowiak LE, Gilbert AL, Morrison JL. Exposed or not exposed? Exploring exposure classification in studies using administrative data to investigate outcomes following medication use during pregnancy. *Eur J Clin Pharmacol.* 2012;68(5):459-467. doi:10.1007/s00228-011-1154-9.
 52. Liew Z, Olsen J, Cui X, Ritz B, Arah OA. Bias from conditioning on live birth in pregnancy cohorts: An illustration based on neurodevelopment in children after prenatal exposure to organic pollutants. *Int J Epidemiol.* 2015;44(1):345-354. doi:10.1093/ije/dyu249.
 53. Hutcheon JA, Fell DB, Jackson ML, et al. Detectable Risks in Studies of the Fetal Benefits of Maternal Influenza Vaccination. *Am J Epidemiol.* 2016;184(3):227-232. doi:10.1093/aje/kww048.
 54. Hernán MA, Lanoy E, Costagliola D, Robins JM. Comparison of dynamic treatment regimes via inverse probability weighting. *Basic Clin Pharmacol Toxicol.* 2006;98(3):237-242. doi:10.1111/j.1742-7843.2006.pto_329.x.
 55. Robins J. A new approach to causal inference in mortality studies with a sustained exposure period-application to control of the healthy worker survivor effect. *Math Model.* 1986;7(9-12):1393-1512. doi:10.1016/0270-0255(86)90088-6.
 56. van der Laan MJ. Targeted maximum likelihood based causal inference: Part I. *Int*

*This is the peer reviewed version of the following article: Wood, M. E., Lapane, K. L., van Gelder, M. M. H. J., Rai, D., & Nordeng, H. M. E. (2017). Making fair comparisons in pregnancy medication safety studies: An overview of advanced methods for confounding control. *Pharmacoepidemiology and Drug Safety*, (February), 1–8, which has been published in final form at <https://doi.org/10.1002/pds.4336>. This article may be used for non-commercial purposes in accordance with Wiley Terms and Conditions for Use of Self-Archived Versions.*

J Biostat. 2010;6(2):Article 3. doi:10.2202/1557-4679.1241.

57. Lash TL, Fox MP, MacLehose RF, Maldonado G, McCandless LC, Greenland S. Good practices for quantitative bias analysis. *Int J Epidemiol.* 2014;(July):1-17. doi:10.1093/ije/dyu149.

This is the peer reviewed version of the following article: Wood, M. E., Lapane, K. L., van Gelder, M. M. H. J., Rai, D., & Nordeng, H. M. E. (2017). Making fair comparisons in pregnancy medication safety studies: An overview of advanced methods for confounding control. *Pharmacoepidemiology and Drug Safety*, (February), 1–8, which has been published in final form at <https://doi.org/10.1002/pds.4336>. This article may be used for non-commercial purposes in accordance with Wiley Terms and Conditions for Use of Self-Archived Versions.

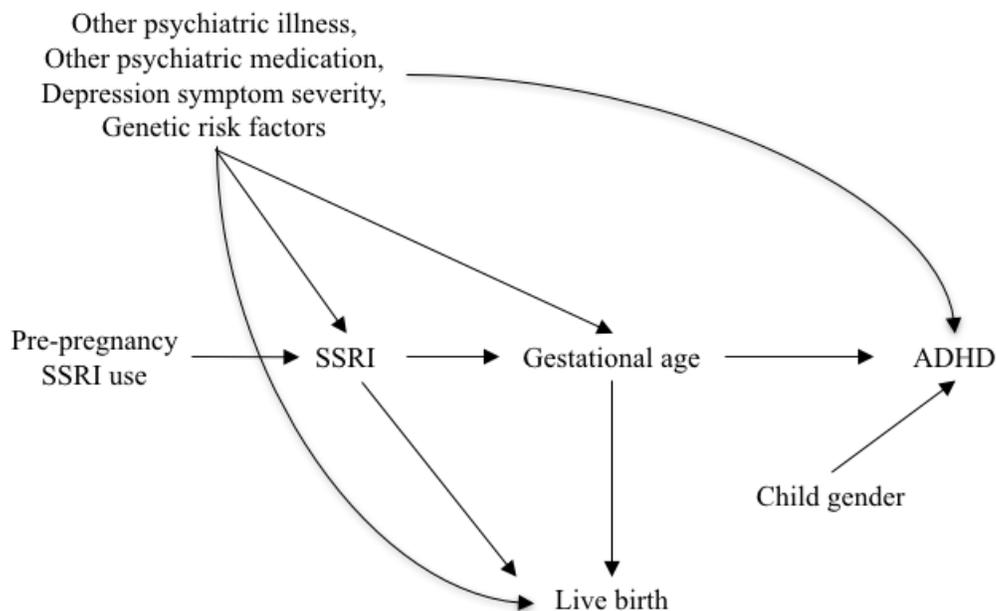


Figure 1. Conceptual model for the effect of prenatal SSRI exposure on attention deficit/hyperactivity disorder (ADHD), including a set of important confounders (depression severity, concomitant medication use, genetics), a potential mediator (gestational age), a collider (live birth), and factors related only to the exposure (pre-pregnancy SSRI use) or the outcome (child gender).

This is the peer reviewed version of the following article: Wood, M. E., Lapane, K. L., van Gelder, M. M. H. J., Rai, D., & Nordeng, H. M. E. (2017). Making fair comparisons in pregnancy medication safety studies: An overview of advanced methods for confounding control. *Pharmacoepidemiology and Drug Safety*, (February), 1–8, which has been published in final form at <https://doi.org/10.1002/pds.4336>. This article may be used for non-commercial purposes in accordance with Wiley Terms and Conditions for Use of Self-Archived Versions.

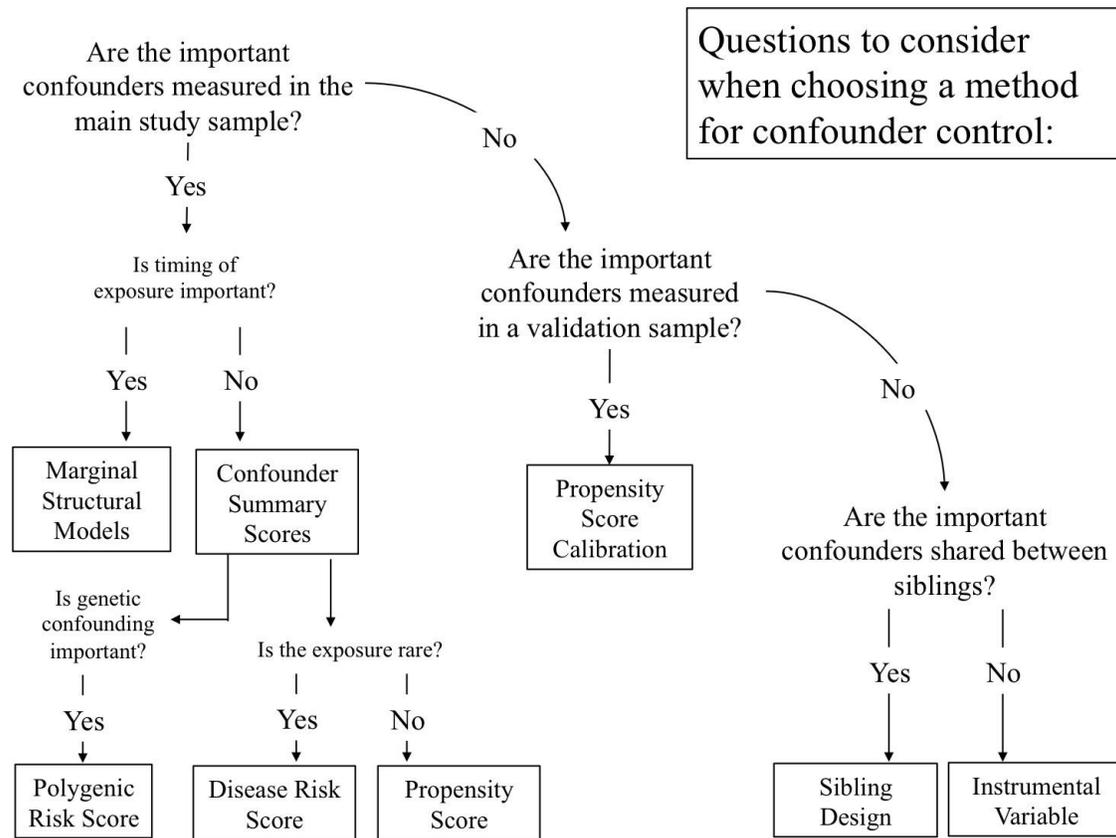


Figure 2. Choosing methods for confounding control

This is the peer reviewed version of the following article: Wood, M. E., Lapane, K. L., van Gelder, M. M. H. J., Rai, D., & Nordeng, H. M. E. (2017). Making fair comparisons in pregnancy medication safety studies: An overview of advanced methods for confounding control. *Pharmacoepidemiology and Drug Safety*, (February), 1–8, which has been published in final form at <https://doi.org/10.1002/pds.4336>. This article may be used for non-commercial purposes in accordance with Wiley Terms and Conditions for Use of Self-Archived Versions.

Table 1. Examples of application of advanced confounding control methods in the pregnancy medication safety literature.

Medication & Study Reference	Outcome	Confounder(s)	Confounding Problem(s)			Method(s) Employed					
			Time-varying	Complex/ High-dimensional	Unmeasured confounders	Propensity scores/ summary scores	structural models	Marginal	Propensity calibration	Sibling/family studies	Instrumental variables
Ondansetron (Pasternak, Svanström, & Hviid, 2013)	Malformations	Nausea/vomiting; maternal characteristics, comorbidities, other medications, pregnancy history.		x		x					
Lithium (Paterno et al., 2017)	Cardiac Malformations	Maternal comorbidities, other medications, maternal characteristics.		x		x					
Statins (Bateman et al., 2015)	Malformations	Maternal characteristics, obstetric and medical conditions, other medications.		x		x					
Triptans (Wood, Lapane, et al., 2015)	Neurodevelopment	Other medications (time-varying), maternal characteristics; migraine severity.	x				x				
Iron supplementation (Bodnar, Davidian, Siega-Riz, & Tsiatis, 2004)	Anemia	Maternal baseline characteristics; gastric symptoms; serum ferritin and hemoglobin concentration.	x				x				
Triptans (Wood, Frazier, Nordeng, & Lapane, 2015)	Neurodevelopment	Other medications, maternal characteristics; migraine severity, attitudes about medication use.		x	x	x		x			
SSRI (Nezvalová-Henriksen et al., 2016; Viktorin et al., 2016)	Gestational age, birth weight	Family factors, maternal depression; illnesses, other medications.		x		x				x	
Anti-epileptic drugs (Bech et al., 2014)	Spontaneous abortion	Severity of maternal epilepsy; maternal characteristics, environmental exposures, comorbidities.			x					x	
SSRI (Swanson et al., 2015)	Maternal depression relapse	Maternal depression severity; comorbidities, other medications, maternal characteristics, proxies for severity.		x	x	x					x

*This is the peer reviewed version of the following article: Wood, M. E., Lapane, K. L., van Gelder, M. M. H. J., Rai, D., & Nordeng, H. M. E. (2017). Making fair comparisons in pregnancy medication safety studies: An overview of advanced methods for confounding control. *Pharmacoepidemiology and Drug Safety*, (February), 1–8, which has been published in final form at <https://doi.org/10.1002/pds.4336>. This article may be used for non-commercial purposes in accordance with Wiley Terms and Conditions for Use of Self-Archived Versions.*

Supplemental Material

Part 1. Discussion of confounding.

Confounders and Causal Inference

First, a definition of confounding: a confounder is a factor in a study that (a) is associated with (and precedes) the medication exposure, (b) is a risk factor for the outcome, and (c) does not lie on the causal pathway between medication use and outcome. In randomized studies, we can be reasonably confident that confounders (both measured and unmeasured) are balanced between exposed and unexposed groups, whereas in observational studies, this is almost certainly not the case. If the distribution of confounders is different for the exposed and unexposed groups, estimates of effect may be biased. Taking confounding into account is essential in observational studies. Further, it is important for researchers to carefully consider the timing of the confounder relative to the exposure and outcome.

Second, we focus on the definition of causal effects arising from the *counterfactual*: that is, we would like to understand what the experience of women who took a medication during pregnancy would have been, had she not taken the medication. In this definition, confounding exists when women who did not take the medication of interest are not adequate stand-ins for the women who did. The counterfactual approach to confounding clarifies a concept that researchers understand intuitively: that we must strive to make fair comparisons between exposed and unexposed groups. For example, for researchers studying the effects of antidepressant exposure on congenital malformations, comparing a group of women using antidepressants who also have severe depression to a group of women with no antidepressant use and no history of depression would be an unfair comparison: the unexposed group is not an adequate stand-in for the experience of women using antidepressants.

*This is the peer reviewed version of the following article: Wood, M. E., Lapane, K. L., van Gelder, M. M. H. J., Rai, D., & Nordeng, H. M. E. (2017). Making fair comparisons in pregnancy medication safety studies: An overview of advanced methods for confounding control. *Pharmacoepidemiology and Drug Safety*, (February), 1–8, which has been published in final form at <https://doi.org/10.1002/pds.4336>. This article may be used for non-commercial purposes in accordance with Wiley Terms and Conditions for Use of Self-Archived Versions.*

Confounding in pregnancy medication studies

Specific types of confounders arise often enough that they are grouped together. *Confounding by indication* challenges the researcher to distinguish whether the outcome of interest is caused by the drug under study or the disorder being treated. Including a disease comparison group (women with the same disease and ideally the same disease severity, but not treated with the drug), or comparing drug use across different indications, offers advantages over studies comparing exposed cases to healthy controls only. *Confounding by concomitant medication use* occurs when women use multiple medications, and it is the concomitant medication, not the index drug, that is responsible for the observed effects. Other potential confounders include maternal factors such as lifestyle, paternal characteristics, genetic risk, and complex familial factors like home environment. These confounders may be measured with varying degrees of completeness or quality, depending on the data source, study design, and difficulty of measurement.

This is the peer reviewed version of the following article: Wood, M. E., Lapane, K. L., van Gelder, M. M. H. J., Rai, D., & Nordeng, H. M. E. (2017). Making fair comparisons in pregnancy medication safety studies: An overview of advanced methods for confounding control. *Pharmacoepidemiology and Drug Safety*, (February), 1–8, which has been published in final form at <https://doi.org/10.1002/pds.4336>. This article may be used for non-commercial purposes in accordance with Wiley Terms and Conditions for Use of Self-Archived Versions.

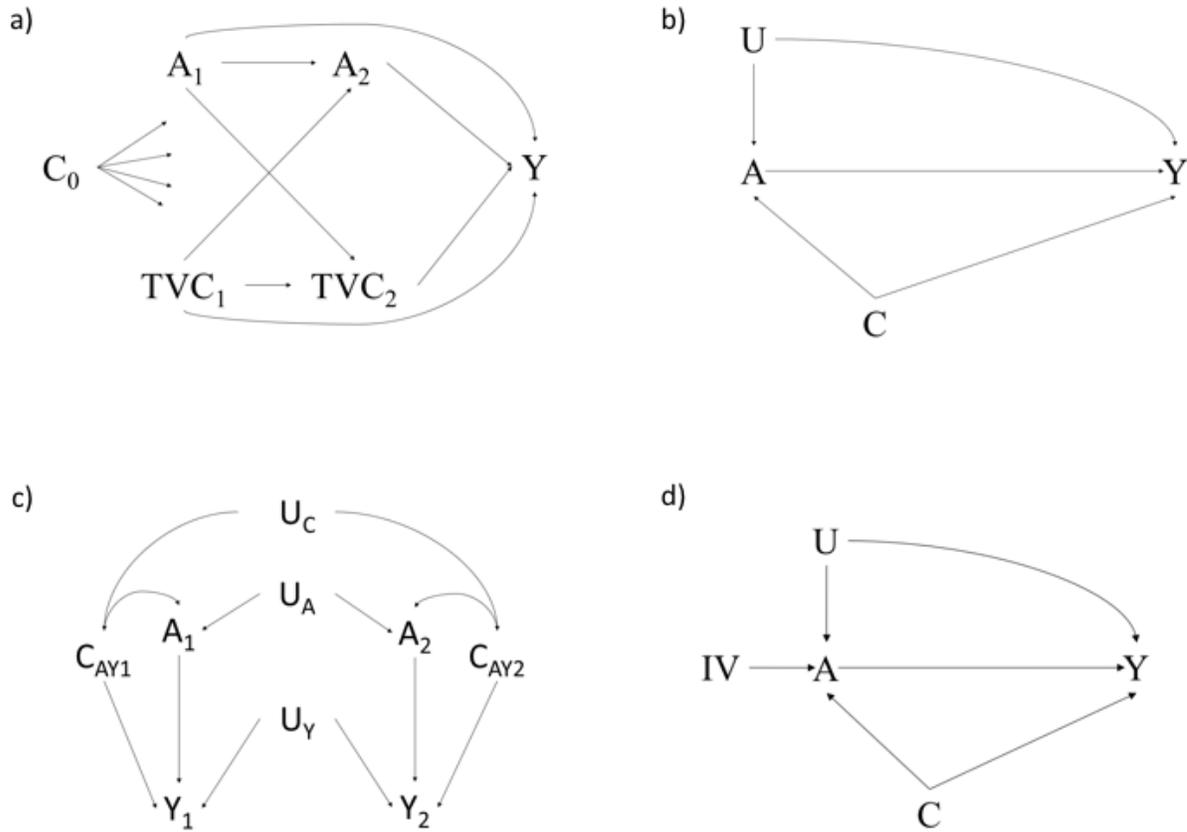


Figure S1. Directed acyclic graphs (DAGs) for **(a)** Time varying confounding: time-varying exposure A , outcome Y , baseline confounders C and time-varying confounders TVC at times 0, 1, and 2; **(b)** Unmeasured confounding: exposure A , outcome Y , and measured C and unmeasured U confounders; **(c)** Sibling study design, for siblings (1 and 2), with exposure A , outcome Y , and confounders C of AY , and shared unmeasured factors which cause C , A , and Y ; **(d)** Instrumental variable (IV) which affects the outcome Y only through the exposure A and therefor controls both measured confounders C and unmeasured confounders U .

This is the peer reviewed version of the following article: Wood, M. E., Lapane, K. L., van Gelder, M. M. H. J., Rai, D., & Nordeng, H. M. E. (2017). Making fair comparisons in pregnancy medication safety studies: An overview of advanced methods for confounding control. *Pharmacoepidemiology and Drug Safety*, (February), 1–8, which has been published in final form at <https://doi.org/10.1002/pds.4336>. This article may be used for non-commercial purposes in accordance with Wiley Terms and Conditions for Use of Self-Archived Versions.

Box 1. Illustration of confounding control using propensity matching and weighting.

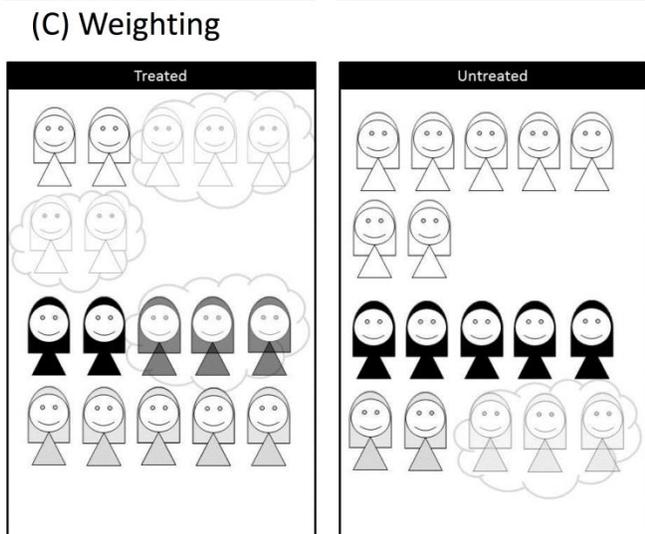
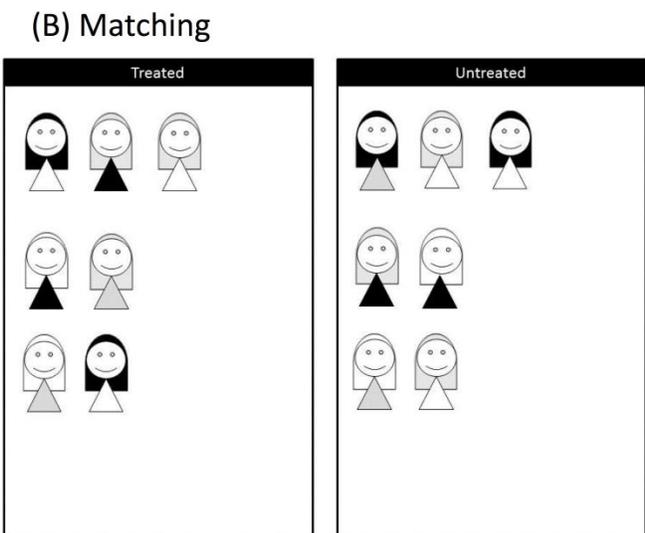
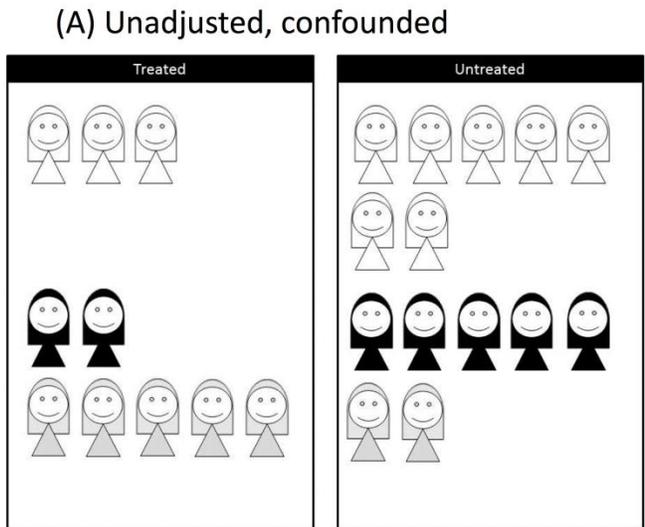
Women who use a medication during pregnancy (“treated”) often have substantially different characteristics than women who do not use a medication (“untreated”); in this illustration, different characteristics are shown through shading and patterns in the figures. Part A shows that the treated and untreated groups are confounded by several participant characteristics.

For example:

Participant 1 has severe depression and drug abuse problems.



Participant 2 has mild anxiety and a history of smoking.



When we use propensity score matching to control for measured confounders, we may be matching two participants who are not similar in terms of specific confounders, but whose propensity for treatment is the same.

When we compare the treated and untreated groups in Part B, we see that the individual participants in the treated and untreated groups look quite different. However, the distribution of characteristics in each group is the same.

It is also important to note that in the matched sample, a large number of the untreated group has been discarded.

Another use of the propensity score is to use the propensity score to weight the study sample. Weights in the sample vary between zero and infinity, and care must be taken to avoid bias from extreme weights.

Weighted analyses are conducted in a *pseudo-population*. In this pseudo-population, participants with rare combinations of confounders are up-weighted (e.g., they count for more than one participant in the analysis) and participants with common combinations of confounders are down-weighted (e.g., they count for less than one participant in the analysis). This results in balanced confounders within each stratum of treatment.

In this illustration (Part C), the participant characteristics are balanced within each stratum of treatment, and we have used lighter drawings to illustrate the presence of a pseudo-population.

This is the peer reviewed version of the following article: Wood, M. E., Lapane, K. L., van Gelder, M. M. H. J., Rai, D., & Nordeng, H. M. E. (2017). Making fair comparisons in pregnancy medication safety studies: An overview of advanced methods for confounding control. *Pharmacoepidemiology and Drug Safety*, (February), 1–8, which has been published in final form at <https://doi.org/10.1002/pds.4336>. This article may be used for non-commercial purposes in accordance with Wiley Terms and Conditions for Use of Self-Archived Versions.

Software resources

Method	Software	References
Propensity Scores Notes: all major software packages can output predicted probabilities after logistic regression to give propensity score, which can be used to calculate weights and strata. Additional software can assist with matching,	SAS	Macros “psmatch_multi”, “cem” available for public use
	Stata	teffects command available in v13 and later; user-written commands “pscore”, “match”, “cem”, “psmatch2” for earlier versions
	SPSS	No formal support; “PS Matching” add-on calls “matchit” package from R.SPSS only supports frequency weights except through the complex survey sampling add-on! Software may round propensity weight to the nearest whole number and use it as a frequency weight without returning an error message.
	R	Packages “twang” “cem” “optmatch” “matchit” and “matching” available for download
Marginal Structural Models Notes: weights estimated from pooled logistic regression, similarly to PS methods above. Notes above regarding weights in SPSS also apply here.	SAS	After creating joint IPTW in data step, use “weight” option in standard outcome analyses.
	Stata	After creating joint IPTW in data step, use “pweight” option in standard outcome analyses.
	SPSS	See notes above on propensity score weighting
	R	After creating joint IPTW in data step, specify appropriate probability weight option (varies by outcome model type)
Propensity Calibration Notes: technique developed using SAS macros. Other software supports regression calibration (noted at right) but has not been testing in the unmeasured confounding application.	SAS	Carry out through regression calibration; “Blinplus” macro available for download
	Stata	User-written command “rcal” performs regression calibration
	SPSS	Not available.
	R	“iWeigReg” package after propensity score estimation.
Sibling Comparison Notes: implemented using standard mixed model	SAS	PROC MIXED and PROC NL MIXED
	Stata	Xtreg command
	SPSS	MIXED command
	R	Multiple options for downloadable

This is the peer reviewed version of the following article: Wood, M. E., Lapane, K. L., van Gelder, M. M. H. J., Rai, D., & Nordeng, H. M. E. (2017). Making fair comparisons in pregnancy medication safety studies: An overview of advanced methods for confounding control. *Pharmacoepidemiology and Drug Safety*, (February), 1–8, which has been published in final form at <https://doi.org/10.1002/pds.4336>. This article may be used for non-commercial purposes in accordance with Wiley Terms and Conditions for Use of Self-Archived Versions.

software		packages, including “lme” and “nlme”
Instrumental Variable(s)	SAS	PROC SYSLIN, PROC MODEL, PROC CALIS
Notes: multiple approaches to IV analyses. Two stage least squares approaches are most appropriate for linear outcomes; non-linear outcomes require special consideration, particularly with respect to standard errors.	Stata	“IVRegress” included from v13 on; user written package “ivreg2” fits 2sls models; other packages available for extended applications.
	SPSS	2SLS command
	R	“tsls” package