Analyzing missing data in perinatal pharmacoepidemiology research: methodological considerations to limit the risk of bias

Lupattelli A,[1] Wood M,[2] Nordeng H[1,3]

[1]*PharmacoEpidemiology and Drug Safety Research Group, Department of Pharmacy, & PharmaTox Strategic Research Initiative, Faculty of Mathematics and Natural Sciences, University of Oslo, Norway*
[2]*Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, USA*
[3]*Department of Child Health and Development, Norwegian Institute of Public Health, Oslo, Norway*

Corresponding author: Angela Lupattelli, PhD. PharmacoEpidemiology and Drug Safety Research Group, Department of Pharmacy, & PharmaTox Strategic Research Initiative, Faculty of Mathematics and Natural Sciences, University of Oslo, Norway. PO Box 1068 Blindern, 0316 Oslo, Norway
e-mail: angela.lupattelli@farmasi.uio.no; phone: +47 41343628; fax: +4722854402

Contribution to authorship

HN, MW and AL have developed the research question, and drafted the commentary. All authors have all contributed to interpretation of the data presented in the commentary, and critically revised the commentary for intellectual content.

Conflict of Interest statement

The author HN declares no conflict of interest. The author MW declares no conflict of interest. The author AL declares no conflict of interest.

Introduction

Missing data is a global problem in human subjects research, and a serious threat to both validity and efficiency in effect estimation. The CONSORT 2010 Statement[1] advocates transparent reporting of the extent of missing data and how this issue was dealt with in the analysis, as this is crucial for readers to critically evaluate the study findings and potential biases. Recognition of the threat from these biases has resulted in calls for increased use of methods for dealing with missing data.[2] However, barriers exist that prevent applied pharmacoepidemiology researchers from assessing the potential gains to their own work, including understanding scenarios when simpler methods might be sufficient, or when complex approaches are needed. These barriers include a lack of resources that integrate missing data terminology and approaches with epidemiologic concepts, and a discussion of the strengths and weaknesses of the most common approaches.

We review the critical concepts for missing data problems, with the aim of integrating more traditional statistical language on missingness mechanisms with epidemiologic methods based on causal diagrams.[3] We have framed this commentary using examples from perinatal pharmacoepidemiology, including an applied example from the Norwegian Mother and Child Birth Cohort (MoBa): evaluating the effect of prenatal use of selective serotonin reuptake inhibitor (SSRI) antidepressants on preeclampsia in the presence of missing data on relevant confounders such as smoking status in gestation.

Missing data methods and mechanisms

Missing data are generally classified as missing completely at random (MCAR), missing at random (MAR), or missing not at random (MNAR)[2, 4, 5] as briefly described below.

Missing completely at random (MCAR)

Under this scenario, there are no systematic differences between the missing and the observed values.[2, 5] For example, if unexperienced health care personnel forget to ask about smoking during pregnancy, information about smoking will be missing at random in the pregnant woman's medical chart. The same occurs when study participants randomly forget to fill in or skip responses. There is no risk of bias with MCAR data, but there will be loss of precision.

Missing at random (MAR)

Missing at random is classified as any systematic difference between the missing values and the observed values, which can be explained by the observed data.[2, 5] For instance, depressed pregnant women may be less likely to report smoking than non-depressed.

Missing not at random (MNAR)

Missing not at random occurs in situations when systematic differences remain between the missing values and the observed values, even after the observed data are taken into account; missingness is thus related to unmeasured variables. For example, women who smoke during pregnancy may less likely report their smoking status. When missingness in a variable depends on the missing value itself, the unbiased estimate is not recoverable in observed data

Exploring extent and patterns of missingness

Although Little's test may help researchers to identify missingness that is MCAR vs. MAR, this test is not conclusive. In addition, no numerical diagnostics can differentiate MAR from MNAR. This means we are left with logical reasoning to inform us on the mechanism behind data missingness. Exploring the extent and pattern of missing data in one's own data sample (for

example by cross-tabulating variables with missing data against exposure and outcome), as well as using findings from previous studies and normative data (e.g. score distribution in a reference population) can give a hint of the underlying mechanism of missingness. This is important to appraise as it will guide decision making of missing data handling: the various approaches to missing data analysis require different assumptions about the underlying mechanisms.

Methods to handle missing data

Multiple methods for handling missing data are used in perinatal pharmacoepidemiology research. These methods fall into two broad categories: analyze the observed data (complete case analysis), or use some principled method for filling in the missing data (imputation). In complete case analysis (CCA), observations with missing data on relevant variables are dropped from the analysis. This approach will always produce unbiased results under the MCAR assumption, and may produce unbiased results under MAR or MNAR. CCA is commonly used in perinatal pharmacoepidemiology due to its simplicity (Table 1). In database linkage studies where study size is large, the loss of data has less impact on precision than in smaller size or different design studies.[6-10]

Single imputation comprises a set of techniques where missing value are replaced by a value from the observed data, for instance the mean or mode. The imputed values are assumed to be equal to the values that would have been observed if data had been complete. This method, however, underestimates uncertainty about the missing values and will therefore result in standard errors that are too small.[2, 5] In the study by Panchaud et al,[11] gestational age was conditionally imputed for 6% of the pregnancies based on the sample mean. In the study by Pasternak et al,[12] missing information on several baseline maternal characteristics was replace

using the mode. In longitudinal studies with repeated variable measurement, for example using questionnaires at several time points in pregnancy, the "last value carried forward" technique can be used to replace missing values with the last measured value of the individual, as done by Norby et al.[13] This method assumes that the observation of the individual remains the same since the last measured observation. Due to well-established shortcomings,[2, 5] single imputation techniques are less used in perinatal pharmacoepidemiology.

More advanced model-based methods for handling missing data have become more accessible to researchers in recent years through packages in standard statistical software. The two most common model-based methods are maximum likelihood using the expectation maximization (EM) algorithm and multiple imputation.[4, 14, 15] These are considered model-based methods since the researcher must make assumptions about the joint distribution of all variables in the model (including both outcomes and predictors).

Maximum likelihood methods using the EM algorithm uses each observation's available data to compute maximum likelihood estimates, rather than filling in the missing values. It runs until the algorithm converges to the "best fit" model for a set of data. The multiple imputation (MI) method fills in missing values by averaging from the distribution of the missing data given the observed data in a way that accounts for the uncertainty associated with the missing values. In MI by chained equations (MICE) a series of regression models are run whereby each variable with missing data is modeled conditional upon the other variables in the data.[14] At the end of one cycle, all missing values have been replaced with predicted values (imputations). The process is repeated for a number of cycles, with the imputations being updated at each cycle, finally resulting in one imputed dataset. The number of imputed datasets is generally between 5 and 20. Standard errors are calculated using Rubin's rules.[15, 16] The MI approach produces valid

estimates under the MAR assumption. This is a weaker assumption than MCAR and more likely to hold in observational studies. MI is a computationally intensive method which is increasingly used in perinatal pharmacoepidemiology research (Table 1).[7, 17-21] Yet, this needs to be applied after careful reflection about the missing data to avoid misleading conclusions. For a comprehensive review of multiple imputation for missing data in epidemiological studies we recommend the papers by Sterne et al,[5] and Perkins et al.[2] Recent research has also shown that the proportion of missing data should not be the major driver for the decision on how to handle missing data.[22] In fact, even when the extent of missing data is large, results can still be unbiased provided that the MAR assumption is met and methods to handle missing data have been adequately applied.

Missing data approaches in recent medication in pregnancy literature

Table 1 summarizes the reporting and handling of missing data in recent perinatal pharmacoepidemiology studies, by type of data source utilized. Of note, this study overview serves as common ground for appraising current methodological gaps, and it is not a comprehensive, systematic extract of the literature. Transparent reporting of the extent and handling of missing data, and the uptake of multiple imputation methods, remains limited. For instance, in multiple cases we computed the extent of missing data in a study using baseline characteristic data of the study sample based on numbers reported in each manuscript; in some studies, it was unclear what missing data approach was used. The majority of studies reported missing data on confounding variables, in different extent (from <1% to 65%) depending on the data source utilized.

On the basis of the missing data definition used by study authors, and the information reported, missing data do not seem to be a major problem in health registry, administrative claims, or pregnancy registry. This contrasts with studies set in birth cohorts, teratology information services, or general practice databases, which often have to contend with much higher levels of missingness, and with patterns that are likely to be informative. The substantial problem of missing data in these study types has promoted important methodological research on the topic,[23, 24] as well as a greater uptake of multiple imputation methods by researchers using this type of data (Table 1). Simpler approaches to handle missing data such as indicator variable, were not often reported in the papers we evaluated; this is encouraging given the well-established shortcomings of the method. Study authors rarely stated any assumptions they made about the underlying mechanism of missingness in the literature we reviewed.

DAG framework with missing data

Directed acyclic graphs (DAGs) can provide helpful insights into potential biases from assuming various missingness mechanisms. Figure 1 introduces a simplified causal model for the effect of prenatal SSRI exposure on preeclampsia. In this model, we assume a causal effect of depression severity on SSRI use and on smoking, and that smoking has an effect on preeclampsia risk. If these assumptions hold, we could estimate the effect of SSRI use on preeclampsia by conditioning on smoking and depression severity. If some fraction of the study sample lacks data on smoking, assumptions about the mechanism that explains the missingness will point to different strategies for analyzing our data. In Figure 1A, smoking is missing completely at random (MCAR), and we can fit a model for the effect of SSRI use on preeclampsia risk, adjusting for depression severity and smoking, in the complete case sample only, without risk of bias. Figures 1B shows that if missingness in smoking status is explained by depression severity,

we can also estimate unbiased effects in the complete case sample, as the covariates required for confounding control also block bias paths from missingness to the outcome. For missing data mechanisms where the missingness is predicted by the missing values, as in Figure 1C, or when the probability of being a complete case depends on the outcome, as in Figures 1D and 1E, complete case analysis will result in a biased estimate. Finally, the presence of an auxiliary variable (that is, a variable that predicts missingness but is unrelated to the causal mechanism being considered) allows for unbiased and efficient effect estimation via multiple imputation.

Applied example: prenatal antidepressant use and risk of preeclampsia

As a motivating example, we present recent work on the association between use of selective serotonin re-uptake inhibitor (SSRI) antidepressants during gestation and risk of late-onset preeclampsia, using data from the MoBa cohort study.[25] MoBa is a nation-wide, population-based pregnancy cohort study conducted by the Norwegian Institute of Public Health, with recruitment occurring between 1999-2008.[23] Pregnant women were recruited from all over Norway at the time of their routine ultrasound at 17-18 weeks of gestation. Data were gathered prospectively by self-administered questionnaires. The cohort now includes 114500 children, 95200 mothers and 77300 fathers, all of whom are followed as long as they continue to participate in the study[23] MoBa has a license from the Norwegian Data Inspectorate and approval from The Regional Committee for Medical Research Ethics. All participants gave their written informed consent prior to participation.

In our study, we first explored patterns of missing data on important confounders by exposure and outcome strata. Missing values on these confounders ranged from 1-3% for maternal smoking and body mass index, to 7-8% for education and weight gain. Missing information on

maternal depressive and anxiety symptom severity in pregnancy, measured via the Hopkins Symptom Checklist-25 (SCL-25) at gestational week 17 (5 items, SCL-5) and 30 (8 items, SCL-8),[26, 27] was as follows: 5% and 10% on at least one of the SCL-5 or SCL-8 items, respectively; 15% total missing information simultaneously on either scale. However, only few women (< 3%) completed none or less than a half of the items composing the individual SCL scales. The missing data mechanism in our study seemed to be linked to maternal age and to the extent of completion of the SCL items, but importantly, it did not seem to be associated with the outcome, late-onset preeclampsia. Based on this and under the MAR assumption,[25] we conducted three sets of analyses: i) complete case analysis; ii) multiple imputation of missing data on the two SCL scales only (approach I); and iii) multiple imputation of missing data on the two SCL scales and on other maternal confounders (approach II). As shown in Table 2, the adjusted and weighted association measures were higher and less precise in the complete case analysis than in the other two sets. However, the results of the complete case analysis expanded to pregnancies with only SCL imputed values (approach I) were similar to those obtained in the fully imputed models (approach II). Increasing sample size and higher statistical power following multiple imputation can indeed explain these discrepancies. The extent of missing data on confounders other than the SCL between the complete-case and approach I (31.9% vs 24.1%) analysis was however not substantial. Hence, because in this example missing data seemed to relate to the extent of completion of the SCL items, we could not exclude the possibility that a complete case analysis approach would yield biased estimates.[5, 28, 29] In the context of this applied example, results from approach II were thereby considered as those least biased.

Implications for applied researchers

Methods for identifying, analyzing, and mitigating bias from missing data have advanced significantly in recent years, and are seeing greater uptake in applied perinatal pharmacoepidemiology research. Based on our survey of the literature, we have several recommendations for applied researchers who need to analyze data with missing values. These recommendations are made bearing in mind that there is no missing data handling solution that fits all research contexts.

First: where possible, limit missingness during collection of data. Recognize that no statistical method can make up for careful study design and data curation. Sometimes the assumptions a specific case of missing data require are simply so unrealistic that the effect estimate is unlikely to be informative. *Second*: carefully diagnose missingness, and use subject-area knowledge as well as exploratory and descriptive data analysis to understand plausible mechanisms of missingness. We suggest that a minimum standard for missing data analysis should be a complete reporting of missingness within strata of exposure and outcome. Researchers should consider the use of causal graphs to make their assumptions about missingness mechanisms explicit. *Third*: Be aware that the proportion of missing data should not be the major driver for the decision on how to handle missing data, but rather the assumed mechanism as to why data are missing. *Fourth*: include a statistical analyst with expertise in missing data methods. Inappropriate analyses using these complex methods can result in seriously biased results. *Finally*: apply strategies for missing data mitigation under different assumptions, and include evaluations of robustness results under these assumptions. For example, including both the complete case analysis and the multiply imputed results can allow readers to decide which estimate they prefer, depending on assumptions about the missingness mechanism.

Careful attention to missing data, and to the assumptions required for analysis of missing data, is necessary in all areas of research, including perinatal pharmacoepidemiology. With transparent reporting of the extent and assumed mechanisms of missing data, and by applying strategies for missing data mitigation under different assumptions, future research can avoid the problems that result from failure to consider this important source of bias.

Figure Legends

Figure 1. Causal diagrams showing relationships between prenatal SSRI use, maternal depression severity, preeclampsia, and smoking, as well as a binary indicator, $Miss_{SMK}$, denoting missing information in the smoking variable.

References

1. Moher D, Hopewell S, Schulz KF, Montori V, Gøtzsche PC, Devereaux PJ, et al. Consort 2010 explanation and elaboration: Updated guidelines for reporting parallel group randomised trials. *BMJ.* 2010;340.
2. Perkins NJ, Cole SR, Harel O, Tchetgen Tchetgen EJ, Sun B, Mitchell EM, et al. Principled approaches to missing data in epidemiologic studies. *Am J Epidemiol.* 2018;187:568-75.
3. Daniel RM, Kenward MG, Cousens SN, De Stavola BL. Using causal diagrams to guide analysis in missing data problems. *Stat Methods Med Res.* 2012;21:243-56.
4. Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the em algorithm. *J R Stat Soc Series B Stat Methodol.* 1977;39:1-38.
5. Sterne JA, White IR, Carlin JB, Spratt M, Royston P, Kenward MG, et al. Multiple imputation for missing data in epidemiological and clinical research: Potential and pitfalls. *BMJ.* 2009;338:b2393.
6. Bateman BT, Heide-Jorgensen U, Einarsdottir K, Engeland A, Furu K, Gissler M, et al. Beta-blocker use in pregnancy and the risk for congenital malformations: An international cohort study. *Ann Intern Med.* 2018;169:665-73.
7. Caniglia EC, Patel K, Huo Y, Williams PL, Kapetanovic S, Rich KC, et al. Atazanavir exposure in utero and neurodevelopment in infants: A comparative safety study. *AIDS.* 2016;30:1267-78.
8. Dhalwani NN, Szatkowski L, Coleman T, Fiaschi L, Tata LJ. Stillbirth among women prescribed nicotine replacement therapy in pregnancy: Analysis of a large uk pregnancy cohort. *Nicotine Tob Res.* 2019;21:409-15.
9. Elkjaer LS, Bech BH, Sun Y, Laursen TM, Christensen J. Association between prenatal valproate exposure and performance on standardized language and mathematics tests in school-aged children. *JAMA Neurol.* 2018;75:663-71.
10. Ernst A, Brix N, Lauridsen LLB, Olsen J, Parner ET, Liew Z, et al. Acetaminophen (paracetamol) exposure during pregnancy and pubertal development in boys and girls from a nationwide puberty cohort. *Am J Epidemiol.* 2019;188:34-46.
11. Panchaud A, Rousson V, Vial T, Bernard N, Baud D, Amar E, et al. Pregnancy outcomes in women on metformin for diabetes or other indications among those seeking teratology information services. *Br J Clin Pharmacol.* 2018;84:568-78.
12. Pasternak B, Svanstrom H, Hviid A. Ondansetron in pregnancy and risk of adverse fetal outcomes. *N Engl J Med.* 2013;368:814-23.
13. Norby U, Kallen K, Shemeikka T, Korkmaz S, Winbladh B. Pregnant women's view on the swedish internet resource drugs and birth defects intended for health care professionals. *Acta Obstet Gynecol Scand.* 2015;94:960-8.
14. Azur MJ, Stuart EA, Frangakis C, Leaf PJ. Multiple imputation by chained equations: What is it and how does it work? *Int J Methods Psychiatr Res.* 2011;20:40-9.
15. Rubin DB. *Multiple imputation for nonresponse in surveys*. Wiley, 1987.
16. Rubin DB. Inference and missing data. *Biometrika.* 1976;63:581-92.
17. Beau AB, Montastruc JL, Lacroix I, Montastruc F, Hurault-Delarue C, Damase-Michel C. Atropinic burden of drugs during pregnancy and psychological development of children: A cohort study in the efemeris database. *Br J Clin Pharmacol.* 2016;82:478-86.

18. Coomarasamy A, Devall AJ, Cheed V, Harb H, Middleton LJ, Gallos ID, et al. A randomized trial of progesterone in women with bleeding in early pregnancy. *N Engl J Med.* 2019;380:1815-24.
19. Magnus MC, Karlstad O, Haberg SE, Nafstad P, Davey Smith G, Nystad W. Prenatal and infant paracetamol exposure and development of asthma: The norwegian mother and child cohort study. *Int J Epidemiol.* 2016;45:512-22.
20. Radojcic MR, El Marroun H, Miljkovic B, Stricker BHC, Jaddoe VWV, Verhulst FC, et al. Prenatal exposure to anxiolytic and hypnotic medication in relation to behavioral problems in childhood: A population-based cohort study. *Neurotoxicol Teratol.* 2017;61:58-65.
21. Scherneck S, Schlinke N, Beck E, Grupe K, Weber-Schoendorfer C, Schaefer C. Pregnancy outcome after first-trimester exposure to metformin: A prospective cohort study. *Reprod Toxicol.* 2018;81:79-83.
22. Madley-Dowd P, Hughes R, Tilling K, Heron J. The proportion of missing data should not be used to guide decisions on multiple imputation. *J Clin Epidemiol.* 2019;110:63-73.
23. Marston L, Carpenter JR, Walters KR, Morris RW, Nazareth I, Petersen I. Issues in multiple imputation of missing data for large general practice clinical databases. *Pharmacoepidemiol Drug Saf.* 2010;19:618-26.
24. Petersen I, Welch CA, Nazareth I, Walters K, Marston L, Morris RW, et al. Health indicator recording in uk primary care electronic health records: Key implications for handling missing data. *Clin Epidemiol.* 2019;11:157-67.
25. Lupattelli A, Wood M, Lapane K, Spigset O, Nordeng H. Risk of preeclampsia after gestational exposure to selective serotonin reuptake inhibitors and other antidepressants: A study from the norwegian mother and child cohort study. *Pharmacoepidemiol Drug Saf.* 2017;26:1266-76.
26. Strand BH, Dalgard OS, Tambs K, Rognerud M. Measuring the mental health status of the norwegian population: A comparison of the instruments scl-25, scl-10, scl-5 and mhi-5 (sf-36). *Nord J Psychiatry.* 2003;57:113-8.
27. Fink P, Orbol E, Hansen MS, Sondergaard L, De Jonge P. Detecting mental disorders in general hospitals by the scl-8 scale. *J Psychosom Res.* 2004;56:371-5.
28. Moodie EE, Delaney JA, Lefebvre G, Platt RW. Missing confounding data in marginal structural models: A comparison of inverse probability weighting and multiple imputation. *Int J Biostat.* 2008;4:Article 13.
29. Mojaverian N, Moodie EE, Bliu A, Klein MB. The impact of sparse follow-up on marginal structural models for time-to-event data. *Am J Epidemiol.* 2015;182:1047-55.

Acknowledgements

The authors acknowledge Prof Kate Lapane and Dr Shao-Hsien Liu for the fruitful discussions on missing data methods in observational studies.

Conflict of Interest statement

The authors declare no conflict of interest.