Handling missing data, selection bias, and measurement error in observational studies

Choi, J.

Chapter 1

Introduction

In this thesis, we address potential threats to the validity of observational epidemiological studies. Examples of these potential sources of bias are confounding, missing data, selection bias, and measurement error. Although various methods have been developed to mitigate these biases, it is often unclear which methods can be used in which empirical settings. It is also common that issues discussed in methodological studies are overlooked in clinical research. Thus, we aim to investigate problems of missing data, selection bias, and measurement error occurring in several specific observational settings and discuss how to optimally handle them.

Observational studies

Observational studies are widely used in epidemiological research. The strength of observational research, in contrast to a randomized control design, is that it can be used in settings where the manipulation of the exposure of interest by investigators is not feasible (1). When properly designed, observational research has the potential to provide evidence with greater external validity than a randomized control study. Especially nowadays, so-called big data collected via routine care, such as electronic health records or disease registry, have become increasingly available, which broaden the possibilities of conducting observational studies (2, 3).

Strengths and weaknesses are two sides of the same coin. Unlike in randomized control trials, the exposure of interest is not randomly assigned in observational studies. Whether it is a treatment, a lifestyle factor, or a biomarker, there are many known or unknown factors that affect why a certain individual has a particular exposure value. Often, these factors are also related to the prognosis of the person (4). This introduces a major well-known threat to the validity of observational research: bias due to confounding (5, 6). Numerous publications have discussed the mechanism of confounding (4, 6) and how to identify confounding factors clinically and statistically (7-10). Widely known methods to adjust for confounding include but are not limited to stratified analyses (11, 12), regression modelling (13), probability weighting (14, 15), propensity score analysis (16, 17), and g-methods (18).

Besides confounding, methodological and statistical challenges remain as epidemiological studies often face other issues that may jeopardize the validity. Typically, these issues are missing data, selection bias, and measurement error.

Missing data

Missing data is inevitable in medical research, and observational studies are especially susceptible to it (19). Missing data can occur by three different mechanisms: data are *missing completely at random* (MCAR) when the probability that a value is missing is independent of observed and unobserved information, *missing at random* (MAR) where the probability of missing depends only on observed information, or *missing not at random* (MNAR) where the probability of missing depends on unobserved information

(20, 21). Ignoring missing data compromises precision and statistical power. More detrimentally, it could lead to an invalid estimation of parameters due to selection bias (22).

Depending on the assumed missing mechanism of the data, appropriate methods to mitigate the missing data problem differ. Multiple imputation is a technique to impute missing values based on the observed data. By generating multiple datasets with plausible values, its strength lies in the reflection of the uncertainty of an imputation model (23). Many studies have shown the superiority of multiple imputation over other methods, such as complete case analysis or adding missing indicator variables in a model (20, 24-28) when data is MCAR or MAR. Although less known, maximum likelihood estimation (29, 30) or inverse probability weighting can also be used for handling data MAR (31-33). However, it is often difficult to discern the missing mechanism of the data, especially whether the data are MAR or MNAR (34, 35). Extra caution is needed when discerning missing data mechanisms in routinely collected data. For instance, some biomarkers may be selectively measured only when considered necessary by clinicians (e.g., albumin is measured only in patients with signs of liver or kidney diseases) (36). Results can be substantially biased when the methods for handling MAR are wrongly used for MNAR without tailored adjustment (19, 37).

One particular problem is missing data in the context of propensity score analysis. Propensity score analysis, first introduced by Rosenbaum and Rubin (38), rapidly gained popularity in the past decade as a method for adjusting confounding in observational settings (39). The method aims to mimic a randomized control study; when variables associated with exposure distribution are available and the propensity model is correctly specified, the method creates conditional exchangeability between persons with the same propensity score (16, 17). Missing values in covariates introduce a challenge in propensity score analysis as propensity scores require that all covariates are fully observed (40). Several studies have shown that when covariates are missing (completely) at random, multiple imputation performs better than complete case analysis or adding a missing indicator in the context of propensity score analysis (20, 24, 25). Yet, questions remain on how best to implement multiple imputation in conjunction with propensity score analysis or which methods to use when missing (unmeasured) confounding exist. **Chapter 2** of this thesis discusses how to optimally handle missing data when performing propensity score analysis under different missing data mechanisms.

Selection bias

Selection bias broadly refers to bias introduced due to a systematic discrepancy between the target population and the observed population. Consequently, estimated associations in the selected sample will differ from the association in those initially targeted (41). Various terms refer to selection bias occurring for different reasons; for

example, healthy workers bias, Berkson's bias, non-response bias, or loss to follow-up bias. Although seemingly in different forms, a principal shared is that the bias is introduced due to conditioning on common effects (41, 42). For example, healthy workers bias refers to a situation where workers exposed to specific environmental hazards are wrongly estimated to be in better health status than the general population. The bias occurs when selecting the study population from the workers still working in the field. The problem here is that workers who were exposed to the hazards and could not work anymore would not be included in the study. At the same time, people in the general population who were unfit would have not been hired to work. Thus, selecting only the workers who are still at the field leads to a conditioning on comment effects of the exposure (environmental hazard) and the outcome (health status) of the study (43) and results in selection bias.

Selection bias can be seen as a particular type of missing data problem; information is missing for some individuals of the target population. A fundamental issue in selection bias and missing data problems is that information needed to describe a population of interest is missing from the observed data. Similar to the missing data problem, ignoring a selection of a particular demographic would lead to bias unless the selection of a study sample is a random selection of the target population. Statistical methods suggested to correct selection bias are the inverse probability of sampling weighting(6), g-formula (44), or Heckman's sample selection model (45). The idea behind both the inverse probability of sampling weight and g-formula is to generate a pseudo population by weighting the observed individuals, where the weights are estimated from the representative distribution in the target population. On the other hand, Heckman's sample selection model does not require data from the target population. Instead, it relies on a correct model specification of the outcome regression model and a selection model (42).

Measurement error

Measurement error, also termed misclassification bias if a categorical variable is measured with error, is another common source of bias in epidemiological settings (46-48). Measurement error can happen in any variable, whether in the exposure, other covariates, or the outcome. Depending on the mechanism, measurement error can be classified as *non-differential* when the error is independent of the outcome conditional to covariates; otherwise, *differential* (46). When measurement error occurs, the observed values fail to reflect the true underlying values correctly. Consequently, using variables measured with error in statistical analyses without adequately handling the error would likely result in bias, even when the error is non-differential (49). Statistical methods for handling measurement errors have been discussed extensively (46, 50-52). For example, simple approaches that can be used when the exposure or other covariates in a regression model are measured with errors are regression calibration (53) and simulation extrapolation (SIMEX) (54). The idea of regression calibration is to substitute

the error-prone values for expected values without error, which is derived from a validation dataset (55). SIMEX evaluates the impact of adding more error to a variable on the target parameter and uses this information to extrapolate the scenario without the error (46). More advanced methods include likelihood-based methods (56) or Bayesian correction methods (57). When approaching from a missing data perspective, a multiple imputation approach can as well be used (58); variables measured without error are missing and can be estimated by observed data

Not only the correction of measurement error, but identifying which measurements were measured with error is also a challenge. Measurement errors sometimes occur under specific study settings. Therefore, identifying the errors requires approaches tailored to the setting. Yet, methods are not always readily available, and what is the most suitable method is unknown. One particular example is measurement error in serial hormonal data. The hormonal levels of a person change throughout the day. Although natural variation may occur, the levels would follow an underlying smooth trend, which can be captured by measuring hormones regularly throughout the 24-hour cycle. When measuring hormones, however, errors can occur from various sources, including sample dilution or blood clots in the sample. Such types of measurement errors lead to an underreporting of the hormonal level than it would have been without the error. Reasonably, we may assume hormonal levels deviating largely from a smooth trend are results of measuring error.

Ignoring the measurement error would lead to bias. For instance, one of the statistical measures often used in hormonal research is cross-correlation, which assesses the relative strength of hormonal secretion between two simultaneously measured hormonal series (59). Ignoring hormone levels measured with error will distort a time-serial trend in hormonal secretion and consequently underestimate the cross-correlation. Therefore, in **Chapter 3**, we investigate methods for random measurement error detection in this setting.

Measurements affected by medication use

Variables affected by medication use are often encountered in epidemiological studies with observational data, where the data consists of medication users and non-users. Medication use can be considered an intercurrent event that occurred during the follow-up of a study. Handling intercurrent events in causal inference has recently received much attention (60, 61). It is emphasized that intercurrent events should be incorporated into the well-defined research question. If not, the estimated effect cannot be precisely defined (61). Accordingly, it is essential to choose a statistical method for handling medication use based on the target question and not to make an arbitrary decision.

When choosing which statistical method to use for handling mediation use, the problem can be approached from various angles. It can be viewed as a measurement error

problem when the research interest is in the values not affected by medication use. Measurements of those under medication are 'systematically measured lower' than the values if they had been observed under no medication use. Or, it can be seen as a missing data problem because the true underlying value of interest is not observed. Selection bias may also play a role, as many researchers will only select medication non-users. It can also be seen as a censored data problem when assuming that measurements, if not affected by medication, are always higher than the values observed after medication use. Several methodological studies have illustrated statistical methods for handling medication use and demonstrated that inappropriate methods might lead to substantial bias (62-72).

Despite the suggestions from the existing literature, however, the importance of incorporating medication use in one's research question seems to be overlooked in a majority of clinical research. Consequently, medication use is likely inadequately handled in the analysis (65). Such practice would lead to an arbitrary interpretation of the results and undermine the scientific validity of the study. In this light, in **Chapter 4 to Chapter 7**, we investigate the potential problem of variables affected by medication use and discuss appropriate methods from a practical analytical stage to a conceptual step of setting a research question.

Outline of this thesis

Chapter 2 investigates the handling of missing covariates in propensity score analysis. We conduct a simulation study where we vary missing data mechanisms in a covariate and the presence of effect heterogeneity. Based on the simulation results and missingness graphs, we aim to provide guidance. **Chapter 3** explores how to detect measurement errors that appear in the form of outliers in the time-serial hormonal data of the Leiden Longevity Study. We compare several approaches, from fully relying on experts' knowledge to automated methods, and identify the most well-performing method in empirical and simulated data. From **Chapter 4 to Chapter 7**, we aim to investigate the problem of variables affected by medication use. We start in **Chapter 4** by discussing how to optimally handle a measurement affected by medication use in an analysis by using a simulation study. We vary simulation scenarios based on which variable of interest is affected by medication use and compare various methods, from so-called naïve methods to more advanced methods. Several methods discussed in Chapter 4 require external knowledge of medication use. Therefore, in **Chapter 5**, we attempt to describe the patterns of fasting glucose and HbA1c measurements over time and estimate the effect of glucose-lowering drugs on these measurements in the Netherlands Epidemiology of Obesity study participants. In **Chapter 6**, we conducted a literature review on how medication use is being handled in clinical research. By reviewing clinical studies published in cardiology, diabetes, and epidemiological fields, we aim to describe which methods are being used in practice and evaluate the validity of the methods based on the recommendations from previous methodological studies.

Chapter 7 brings the discussion of handling medication use to a conceptual level. We address several research questions that could be of interest when data contains a mixture of medication users and non-users. For each question, we discuss how medication use is incorporated in the estimand and where the potential methodological challenges lie.

References

1. Black N. Why we need observational studies to evaluate the effectiveness of health care. *BMJ* 1996;312(7040):1215-8.
2. Nicholls SG, Langan SM, Benchimol EI. Routinely collected data: the importance of high-quality diagnostic coding to research. *CMAJ : Canadian Medical Association journal = journal de l'Association medicale canadienne* 2017;189(33):E1054-E5.
3. Hemkens LG, Contopoulos-Ioannidis DG, Ioannidis JPA. Routinely collected data and comparative effectiveness evidence: promises and limitations. *Canadian Medical Association Journal* 2016;188(8):E158-E64.
4. Kyriacou DN, Lewis RJ. Confounding by Indication in Clinical Research. *JAMA* 2016;316(17):1818-9.
5. GREENLAND S, NEUTRA R. Control of Confounding in the Assessment of Medical Technology. *International Journal of Epidemiology* 1980;9(4):361-7.
6. Hernán M, Robins J. *Causal inference: What if*. Boca Raton: Chapman & Hall/CRC; 2020.
7. MICKEY RM, GREENLAND S. THE IMPACT OF CONFOUNDER SELECTION CRITERIA ON EFFECT ESTIMATION. *American Journal of Epidemiology* 1989;129(1):125-37.
8. Groenwold RH, Klungel OH, Grobbee DE, et al. Selection of confounding variables should not be based on observed associations with exposure. *European journal of epidemiology* 2011;26(8):589.
9. Bursac Z, Gauss CH, Williams DK, et al. Purposeful selection of variables in logistic regression. *Source Code for Biology and Medicine* 2008;3(1):17.
10. Greenland S. Modeling and variable selection in epidemiologic analysis. *American Journal of Public Health* 1989;79(3):340-9.
11. Tripepi G, Jager KJ, Dekker FW, et al. Stratification for Confounding – Part 1: The Mantel-Haenszel Formula. *Nephron Clinical Practice* 2010;116(4):c317-c21.
12. Tripepi G, Jager KJ, Dekker FW, et al. Stratification for Confounding – Part 2: Direct and Indirect Standardization. *Nephron Clinical Practice* 2010;116(4):c322-c5.
13. McNamee R. Regression modelling and other methods to control confounding. *Occupational and Environmental Medicine* 2005;62(7):500-6.
14. Curtis LH, Hammill BG, Eisenstein EL, et al. Using Inverse Probability-Weighted Estimators in Comparative Effectiveness Analyses with Observational Databases. *Medical Care* 2007;45(10):S103-S7.
15. Mansournia MA, Altman DG. Inverse probability weighting. *BMJ* 2016;352:i189.
16. Williamson E, Morley R, Lucas A, et al. Propensity scores: from naive enthusiasm to intuitive understanding. *Stat Methods Med Res* 2012;21(3):273-93.
17. 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.
18. Naimi AI, Cole SR, Kennedy EH. An introduction to g methods. *International Journal of Epidemiology* 2016;46(2):756-62.
19. Lee KJ, Tilling KM, Cornish RP, et al. Framework for the treatment and reporting of missing data in observational studies: The Treatment And Reporting of Missing data in Observational Studies framework. *Journal of Clinical Epidemiology* 2021;134:79-88.

20. Donders AR, van der Heijden GJ, Stijnen T, et al. Review: a gentle introduction to imputation of missing values. *J Clin Epidemiol* 2006;59(10):1087-91.
21. Thoemmes F, Mohan K. Graphical Representation of Missing Data Problems. *Structural Equation Modeling: A Multidisciplinary Journal* 2015;22(4):631-42.
22. Westreich D. Berkson's Bias, Selection Bias, and Missing Data. *Epidemiology* 2012;23(1).
23. Rubin DB. Multiple Imputation after 18+ Years. *Journal of the American Statistical Association* 1996;91(434):473-89.
24. unvan der Heijden GJ, Donders AR, Stijnen T, et al. Imputation of missing values is superior to complete case analysis and the missing-indicator method in multivariable diagnostic research: a clinical example. *J Clin Epidemiol* 2006;59(10):1102-9.
25. White IR, Carlin JB. Bias and efficiency of multiple imputation compared with complete-case analysis for missing covariate values. *Stat Med* 2010;29(28):2920-31.
26. Greenland S, Finkle WD. A Critical Look at Methods for Handling Missing Covariates in Epidemiologic Regression Analyses. *American Journal of Epidemiology* 1995;142(12):1255-64.
27. Knol MJ, Janssen KJM, Donders ART, et al. Unpredictable bias when using the missing indicator method or complete case analysis for missing confounder values: an empirical example. *Journal of Clinical Epidemiology* 2010;63(7):728-36.
28. Groenwold RHH, White IR, Donders ART, et al. Missing covariate data in clinical research: when and when not to use the missing-indicator method for analysis. *CMAJ* 2012;184(11):1265-9.
29. Allison PD. Handling missing data by maximum likelihood. Presented at SAS global forum2012.
30. Baraldi AN, Enders CK. An introduction to modern missing data analyses. *Journal of School Psychology* 2010;48(1):5-37.
31. Seaman S, White I. Inverse Probability Weighting with Missing Predictors of Treatment Assignment or Missingness. *Communications in Statistics - Theory and Methods* 2014;43(16):3499-515.
32. Li L, Shen C, Li X, et al. On weighting approaches for missing data. *Statistical Methods in Medical Research* 2011;22(1):14-30.
33. Vansteelandt S, Carpenter J, Kenward M. Analysis of Incomplete Data Using Inverse Probability Weighting and Doubly Robust Estimators. *Methodology: European Journal of Research Methods for The Behavioral and Social Sciences* 2010;6:37-48.
34. White IR, Royston P, Wood AM. Multiple imputation using chained equations: Issues and guidance for practice. *Statistics in Medicine* 2011;30(4):377-99.
35. Horton NJ, Lipsitz SR. Multiple Imputation in Practice. *The American Statistician* 2001;55(3):244-54.
36. Benchimol EI, Smeeth L, Guttmann A, et al. The REporting of studies Conducted using Observational Routinely-collected health Data (RECORD) Statement. *PLOS Medicine* 2015;12(10):e1001885.
37. Sterne JAC, White IR, Carlin JB, et al. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *BMJ* 2009;338:b2393.
38. Rosenbaum PR, Rubin DB. The Central Role of the Propensity Score in Observational Studies for Causal Effects. *Biometrika* 1983;70(1):41-55.

39. Stürmer T, Joshi M, Glynn RJ, et al. A review of the application of propensity score methods yielded increasing use, advantages in specific settings, but not substantially different estimates compared with conventional multivariable methods. *Journal of clinical epidemiology* 2006;59(5):437. e1-. e24.
40. D'Agostino RB, Rubin DB. Estimating and Using Propensity Scores with Partially Missing Data. *Journal of the American Statistical Association* 2000;95(451):749-59.
41. Hernán M, S H-D, Robins J. A Structural Approach to Selection Bias. *Epidemiology* 2004;15(5):615-25.
42. Infante-Rivard C, Cusson A. Reflection on modern methods: selection bias—a review of recent developments. *International Journal of Epidemiology* 2018;47(5):1714-22.
43. Eisen EA, Robins JM. Healthy Worker Effect. *Encyclopedia of Environmetrics*, 2001.
44. Lesko CR, Buchanan AL, Westreich D, et al. Generalizing study results: a potential outcomes perspective. *Epidemiology (Cambridge, Mass)* 2017;28(4):553.
45. Heckman JJ. Sample selection bias as a specification error. *Econometrica: Journal of the econometric society* 1979:153-61.
46. Keogh RH, Shaw PA, Gustafson P, et al. STRATOS guidance document on measurement error and misclassification of variables in observational epidemiology: part 1—basic theory and simple methods of adjustment. *Statistics in medicine* 2020;39(16):2197-231.
47. Brakenhoff TB, Mitroiu M, Keogh RH, et al. Measurement error is often neglected in medical literature: a systematic review. *Journal of Clinical Epidemiology* 2018;98:89-97.
48. Keogh RH, White IR. A toolkit for measurement error correction, with a focus on nutritional epidemiology. *Statistics in medicine* 2014;33(12):2137-55.
49. van Smeden M, Lash TL, Groenwold RHH. Reflection on modern methods: five myths about measurement error in epidemiological research. *International Journal of Epidemiology* 2019;49(1):338-47.
50. Carroll RJ. Measurement Error in Epidemiologic Studies. *Wiley StatsRef: Statistics Reference Online*, 2014.
51. Buonaccorsi JP. *Measurement error: models, methods, and applications*. Chapman and Hall/CRC; 2010.
52. Shaw PA, Gustafson P, Carroll RJ, et al. STRATOS guidance document on measurement error and misclassification of variables in observational epidemiology: Part 2—More complex methods of adjustment and advanced topics. *Statistics in Medicine* 2020;39(16):2232-63.
53. Carroll RJ, Stefanski LA. Approximate Quasi-likelihood Estimation in Models with Surrogate Predictors. *Journal of the American Statistical Association* 1990;85(411):652-63.
54. Cook JR, Stefanski LA. Simulation-Extrapolation Estimation in Parametric Measurement Error Models. *Journal of the American Statistical Association* 1994;89(428):1314-28.
55. Fraser GE, Stram DO. Regression Calibration in Studies with Correlated Variables Measured with Error. *American Journal of Epidemiology* 2001;154(9):836-44.
56. Bartlett JW, De Stavola BL, Frost C. Linear mixed models for replication data to efficiently allow for covariate measurement error. *Statistics in Medicine* 2009;28(25):3158-78.
57. Bartlett JW, Keogh RH. Bayesian correction for covariate measurement error: A frequentist evaluation and comparison with regression calibration. *Statistical Methods in Medical Research* 2016;27(6):1695-708.
58. Cole SR, Chu H, Greenland S. Multiple-imputation for measurement-error correction. *International Journal of Epidemiology* 2006;35(4):1074-81.

59. Veldhuis JD, Keenan DM, Pincus SM. Motivations and Methods for Analyzing Pulsatile Hormone Secretion. *Endocrine Reviews* 2008;29(7):823-64.

60. Young JG, Stensrud MJ, Tchetgen Tchetgen EJ, et al. A causal framework for classical statistical estimands in failure-time settings with competing events. *Statistics in Medicine* 2020;39(8):1199-236.

61. ICH E9 (R1): addendum on estimands and sensitivity analysis in clinical trials to the guideline on statistical principles for clinical trials. European Medicines Agency, 2020.

62. Masca N, Sheehan NA, Tobin MD. Pharmacogenetic interactions and their potential effects on genetic analyses of blood pressure. *Statistics in Medicine* 2011;30(7):769-83.

63. Tobin MD, Sheehan NA, Scurrah KJ, et al. Adjusting for treatment effects in studies of quantitative traits: antihypertensive therapy and systolic blood pressure. *Statistics in Medicine* 2005;24(19):2911-35.

64. White IR, Koupilova I, Carpenter J. The use of regression models for medians when observed outcomes may be modified by interventions. 2003;22(7):1083-96.

65. Spieker AJ, Delaney JA, McClelland RL. A method to account for covariate-specific treatment effects when estimating biomarker associations in the presence of endogenous medication use. *Statistical Methods in Medical Research* 2018;27(8):2279-93.

66. Spieker AJ, Delaney JAC, McClelland RL. Evaluating the treatment effects model for estimation of cross-sectional associations between risk factors and cardiovascular biomarkers influenced by medication use. *Pharmacoepidemiology and drug safety* 2015;24(12):1286-96.

67. Cui JS, Hopper JL, Harrap SBJH. Antihypertensive treatments obscure familial contributions to blood pressure variation. 2003;41(2):207-10.

68. Levy D, DeStefano AL, Larson MG, et al. Evidence for a gene influencing blood pressure on chromosome 17. *Hypertension* 2000;36:477-83.

69. Balakrishnan P, Beaty T, Young JH, et al. Methods to estimate underlying blood pressure: The Atherosclerosis Risk in Communities (ARIC) Study. *PLOS ONE* 2017;12(7):e0179234.

70. Schmidt AF, Heerspink HJL, Denig P, et al. When drug treatments bias genetic studies: Mediation and interaction. *PLOS ONE* 2019;14(8):e0221209.

71. van Geloven N, Swanson SA, Ramspek CL, et al. Prediction meets causal inference: the role of treatment in clinical prediction models. *European Journal of Epidemiology* 2020;35(7):619-30.

72. McClelland RL, Kronmal RA, Haessler J, et al. Estimation of risk factor associations when the response is influenced by medication use: An imputation approach. *Statistics in Medicine* 2008;27(24):5039-53.