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

Choi, J.

# Chapter 8

## Summary and general discussion

This thesis aimed to investigate methodological issues pervasive in epidemiological studies with observational data. We specifically focused on dealing with missing data in propensity score analysis, identifying measurement errors, and handling medication use, both statistically and conceptually. In this discussion chapter, we summarize the main findings of our research and discuss implications and future perspectives.

***Summary of the main findings***

In **Chapter 2**, we investigated how to optimally handle covariates with missing data in propensity score analysis. We generated several simulation scenarios by varying missing data mechanisms and the presence of an effect modification of the treatment. Our findings demonstrated that no single approach is universally optimal. Which methods to use depends on the data structure, such as the missing mechanism and presence of effect heterogeneity and/or unmeasured confounding. Importantly, complete case analysis or adding missing indicators in a model, methods that are considered 'naïve' and inappropriate to handle missing data, outperformed multiple imputation when missing is not at random. Multiple imputation performed best when data were missing at random, but only when the imputation model was correctly specified. This implies that the imputation model should include the outcome variable. When heterogeneity in the treatment effect is present, an interaction term should as well be added to the model.

**Chapter 3** examined methods to detect measurement errors possibly due to sample dilution in time-serial hormonal data where study participants' blood was drawn every 10 minutes for 24 hours. We compared four approaches for detecting measurements error: i) Eyeballing by physiological experts, which could be considered as a golden standard, ii) the stepwise approach, which incorporates physiological knowledge into standard deviation-based detection, iii) Tukey's fences method, which identifies errors based on interquartile ranges, and iv) the expectation-maximization (EM) algorithm which mathematically distinguishes the potential distributions of hormone levels measured with and without error. Based on the performance in the real-world setting and simulated data, we concluded that the stepwise approach, leveraging physiological background knowledge, outperformed fully automated data-driven methods, such as Tukey's fences and the EM algorithm. Tukey's fences performed especially unstably when the hormonal profile was mainly flat with few sudden pulses (e.g., growth hormone). The EM algorithm could not ensure whether the identified distributions truly distinguished outliers from non-outliers. On the other hand, the stepwise approach showed consistent performance under different types of hormonal trends.

**Chapter 4** studied how to handle variables affected by medication use when the research aim is in the variables if not treated. For instance, one may be interested in the effect of a genetic factor on blood pressure at a certain age or the effect of blood pressure on the risk of cardiovascular disease if no one with hypertension uses antihypertensive drugs.

We showed with simulations that which method to use is contingent upon whether the affected variable is the exposure, the outcome, or a confounder. When the exposure is affected, restricting the study population to the untreated individuals may yield a valid result. However, if effect heterogeneity is present, the result may not be extrapolated to the overall population. If external knowledge of the mean and standard deviation of the medication effect is known, regression calibration with adding the mean medication effect to the treated values could be used. When a confounder variable is affected by medication use, simple methods such as restricting the population to untreated individuals or adding an indicator variable for medication use in a regression model may work well. However, when the outcome is affected, simple methods will lead to bias. Instead, adding mean medication effect or using censored normal regression is appropriate. Based on the results, we encouraged researchers to critically consider the processes of medication prescription, the presence of effect heterogeneity, and what information on medication effects is available when handling medication use.

Several methods discussed in **Chapter 4** require external knowledge of the estimated effect of medication and, in some cases, its standard deviation of the medication effect. Randomized control trials on drugs may provide the information. However, populations in trials often do not represent a population of interest in observational research. Applying the medication effect acquired in trials to observational settings could introduce bias due to discrepancies in clinical settings between trials and the real world.

Hence, in **Chapter 5,** we aimed to describe changes in glucose and HbA1c levels after glucose-lowering medication from routinely collected data in the Netherlands Epidemiology of Obesity (NEO) study participants. Electronic Patient Records from general practitioners were used to identify incident diabetes cases and repeated measurements of glucose and HbA1c. We fitted linear mixed models with time as a categorical variable added as fixed and random effect. To avoid regression to the mean effect, we set 6 to 12 months *before* medication prescription as the reference, assuming that it would better represent the study participants' baseline glucose and HbA1c levels. The results showed that the effect of mediation was the largest at 6 to 12 months *after* medication use. The estimated effects were smaller than observed in RCTs, however, remained effective for more than two years after prescription. The effect of medication varied largely between individuals. We also observed that both glucose and HbA1c level increased shortly before medication use. This may reflect a random high measurement that led to a treatment decision in some individuals. Thus, using the last measurement before the start of medication as a reference could lead to a regression to the mean effect. The estimated mean changes can be used in further research in the NEO study when glucose or HbA1c level is the variable of interest. For instance, when they are the outcome of interest, one can add the estimated differences to the measurements of individuals using glucose-lowering medication. If they are the exposure of interest,

using regression calibration by adopting the mean difference and standard deviation could be an option. Routinely collected data allowed investigation of the long-term real-world effect of medication, which could not be easily obtained from RCTs. However, data collection and clinical decision-making processes in the routinely collected electronic health records were not clearly known, introducing challenges in our study.

In **Chapter 6**, we performed a systematic review of how variables affected by medication use were handled in clinical research. We showed that a majority of the studies ambiguously reported whether their research aim is in the values as observed regardless of medication use or if not affected by medication. Even when the aim was clear, many studies used invalid methods for handling medication use. Especially when the outcome variable was affected, methods that are invalid regardless of the research aim, such as restricting a study population to untreated individuals or adding an indicator for medication use, were frequently used. More advanced methods described in methodological literature were rarely adopted. These results indicated that the importance of establishing a clear research question regarding medication use is often overlooked, and appropriate methods to handle medication use are not well-known to clinical researchers.

**Chapter 7** set out to discuss how medication use can be differently incorporated into a research question when the exposure or outcome of interest is affected by medication use in some people. Under each possibility, we discussed the assumptions on relationships between variables and the potential clinical relevance behind them. Some questions could be formulated within a causal framework, where emulating a target trial could help crystalize the question. Other questions are not suitable for a causal estimation but may still provide etiological insight. Concurrently, medication use should be handled differently in the analysis of each question, and different methodological considerations are required.

***Implications and future perspectives***

- **There is no one optimal method for all situations: all decisions made in a study depend on contextual knowledge**

Numerous decisions have to be made when conducting an epidemiological study, from setting a research question and designing a study to analyzing collected data and interpreting the results. Every decision should be made consciously according to the aims and the population of interest. For the analysis, it is essential to understand the structure of the collected data. Whether a certain method is considered default or commonly used should not be a reason to routinely choose the method. This is also the case when handling confounding, missing data, selection bias, and measurement error in observational research.

For example, we showed in **Chapter 2** that using the default settings of multiple imputation software could lead to biased results even when the data are missing at random if the default regression models are not sufficient to capture the complete data structure (3-5). Also, when handling measurement errors, it is essential to know how the data were collected. In our particular example of **Chapter 3,** we utilized the information that multiple hormones were processed simultaneously, which was known from context-dependent background knowledge (6). Intercurrent events should also be dealt with according to one's research question and corresponding target estimand (7-9). We discussed this in a specific context of medication use (see **Chapter 3, Chapter 4, and Chapter 7**). However, in **Chapter 6**, we observed that many clinical studies applied prevalently used but invalid methods.

The increasing availability of electronic health records and disease registries facilitates conducting a broad range of observational studies with so-called big data. As sample sizes are getting bigger and data structures are becoming more challenging to grasp, machine-learning approaches are thought of as attractive alternatives to traditional statistical modeling (6). With machine learning approaches, computer algorithms can learn and improve themselves to grasp the complex structure and patterns of the data. This development may lead to the thought that context-specific knowledge of the research setting is redundant as long as 'big data' to run a machine learning algorithm is available.

Big data, however, also is affected by issues regarding measurement error, selection bias, confounding, and missing data, if not more so (10). For instance, in **Chapter 5**, we encountered challenges when using electronic health records, such as selective medication prescription within the individuals diagnosed with the same diseases or irregular measurements of the outcomes between individuals. These issues will remain and will not magically disappear simply by increasing the sample size. The machine does not learn itself and will likely derive invalid causal estimates without appropriate input in the algorithm about the data collection process and various sources of potential error (11, 12). Unless these are adequately addressed, one should be skeptical about the interpretability, reproducibility, and reliability of the results from machine learning (13, 14). Even in the emergence of big data and machine learning, careful considerations of the research setting, clinical knowledge, and study designs remain highly important.



- **Simulation studies should be used more often in clinical research**

In several chapters, we conducted simulation studies to compare the performances of different statistical methods and to find an optimal approach in different scenarios (see **Chapter 2, Chapter, 3 and Chapter 4**). A simulation study is a widely used tool in statistical research due to its advantage of providing empirical results on how specific methods would perform under various settings as opposed to theoretical evidence from mathematical derivations (15).

We suggest clinical researchers utilize simulation studies in collaboration with analytical experts when it is unclear which statistical method to use in their research setting. Simulation studies can provide information on the magnitude and the direction of the bias and/or the robustness of methods under the violation of assumptions (16). From such information, one can evaluate the validity of adopting a particular method in a given setting. The validity of the methods can also be easily compared in several different data structures by modifying simulation parameters. For instance, previous methodological studies suggested that Heckman's treatment model could be a suitable method for handling measurements affected by medication. However, through simulation studies in **Chapter 4**, we observed that the method may not be suitable when the medication effect varies largely between individuals, which is likely the case in the NEO study setting (shown for the glucose-lowering medication effect in **Chapter 5**).

One of the pitfalls of simulation studies is that the simulation cannot fully reflect real-world settings. The complexity of a real-world setting may be mitigated by incorporating real-world data into the simulation study. For example, in **Chapter 4**, we used several variables directly from the NEO study data in our simulation so that the simulated data would reflect the relationship between the variables in the real world. Performing simulation studies would enable researchers to make analytical decisions more consciously and enhance transparent reporting on the rationale behind using a specific statistical method over another. Several studies discussed how to set up and conduct a sound simulation study (15, 17-19).

- **More focus on bridging the gap between statistical advances and clinical research is needed**

The statistical methods compared in our simulation studies were not newly developed methods but have already been discussed in methodological literature. Our focus in this thesis was to compare available statistical methods and provide guidance on when and how to properly apply them in specific observational research contexts.

Unfortunately, advances in statistical methods mainly remain within the methodological research domain. It often takes a long time, if ever, before new methodological advances are adopted in applied research. For example, none of the more advanced methods to handle medication use, which we studied in the simulation of **Chapter 4**, were applied in the clinical studies that we reviewed in **Chapter 6**. Also, pitfalls of commonly used methods known in methodological research are easily neglected in applied research, leading to potentially flawed results (20, 21). Using multiple imputation without a correct model specification is one of the examples shown in this thesis (see **Chapter 2**).

Possible reasons for the gap between the methodological and clinical research could be a lack of understanding of the technical backgrounds of the problems, a difficulty in programming in statistical software, or an absence of guidance on when to use which

methods in practical settings. To overcome this, there should be a constant focus not only on developing new methods but on bridging the gap between existing methods and applied epidemiological research (16), to which we hope to have contributed with this thesis. Other systematic efforts are being made. For instance, some epidemiological journals provide a corner, such as the education corner in the International Journal of Epidemiology, to introduce methodological development in an accessible (22).

- **Confounding, missing data, selection bias, and measurement error are interrelated**

Our research investigated methodological challenges due to missing data, selection bias, and measurement errors in several specific observational study settings. Although these biases are mostly addressed as separate issues, they are closely related (23, 24). Selection bias is closely related to a missing data problem; a part of the data needed to make a valid conclusion about the target population is not observed. Ignoring missing data when data are missing at random or not at random would lead to selection bias. Missing data is an extreme form of measurement error, and differential measurement error may lead to selection bias. Thus, a methodological issue that seemingly originates from one type of bias can be approached from multiple angles. Subsequently, a method developed to handle one type of problem may be used for handling another one.

For instance, it was demonstrated that a method we used for handling data missing not at random in the context of propensity score analysis (see **Chapter 2**) is also applicable when handling bias due to sample selection (25). Although it was unsuccessful, we showed that the EM algorithm, which is usually considered in statistical modeling when missing values exist, can also be applied in detecting measurement errors (see **Chapter 3**). Also, for handling medication use, we could adopt methods rooted from different angles (see **Chapter 4**). By approaching the problem from a selection bias perspective, we used inverse probability weighting or Heckman's treatment model. From a measurement error perspective, we used regression calibration or adding a constant value. From a missing data perspective, we used multiple imputation methods. From a censored data perspective, we used quantile regression and censored normal regression.

Efforts have been made to provide a unified understanding of the biases by adopting a potential outcome framework (23, 24, 26, 27). From a more practical angle, several authors also provided statistical methods to simultaneously address different sources of bias. These include but are not limited to, multiple imputation methods, Bayesian models, g-formula, and inverse probability weighting (27-31). We believe seeking solutions from broader and more flexible perspectives than approaching each bias in an isolated manner will lead to a better possibility of finding an appropriate solution in one's research setting.

## Conclusions

There is no one best method that can be universally applied to mitigate the problems of confounding, missing data, selection bias, and measurement error in various settings of observational research. No analytical decision should be taken for granted, and each source of bias should be handled on the basis of context-specific knowledge. A constant pursuit of connecting the methodological and clinical worlds and broadening the perspectives on handling biases will contribute to the validity of observational research.

## References


1. Mitra R, Reiter JP. A comparison of two methods of estimating propensity scores after multiple imputation. *Stat Methods Med Res* 2016;25(1):188-204.
2. outPenning de Vries B, Groenwold R. Comments on propensity score matching following multiple imputation. *Stat Methods Med Res* 2016;25(6):3066-8.
3. 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.
4. Hughes RA, Heron J, Sterne JAC, et al. Accounting for missing data in statistical analyses: multiple imputation is not always the answer. *International Journal of Epidemiology* 2019;48(4):1294-304.
5. Tilling K, Williamson EJ, Spratt M, et al. Appropriate inclusion of interactions was needed to avoid bias in multiple imputation. *Journal of Clinical Epidemiology* 2016;80:107-15.
6. 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.
7. 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.
8. 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.
9. 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.
10. Cahan EM, Hernandez-Boussard T, Thadaney-Israni S, et al. Putting the data before the algorithm in big data addressing personalized healthcare. *npj Digital Medicine* 2019;2(1):78.
11. Bi Q, Goodman KE, Kaminsky J, et al. What is Machine Learning? A Primer for the Epidemiologist. *American Journal of Epidemiology* 2019;188(12):2222-39.
12. Wiemken TL, Kelley RR. Machine Learning in Epidemiology and Health Outcomes Research. *Annual Review of Public Health* 2020;41(1):21-36.
13. Dunson DB. Statistics in the big data era: Failures of the machine. *Statistics & Probability Letters* 2018;136:4-9.
14. Keil AP, Edwards JK. You are smarter than you think: (super) machine learning in context. *European Journal of Epidemiology* 2018;33(5):437-40.
15. Morris TP, White IR, Crowther MJ. Using simulation studies to evaluate statistical methods. *Statistics in Medicine* 2019;38(11):2074-102.
16. Boulesteix AL, Binder H, Abrahamowicz M, et al. On the necessity and design of studies comparing statistical methods. *Biometrical journal Biometrische Zeitschrift* 2017;60(1):216-8.
17. Boulesteix A-L, Groenwold RH, Abrahamowicz M, et al. Introduction to statistical simulations in health research. *BMJ Open* 2020;10(12):e039921.
18. Burton A, Altman DG, Royston P, et al. The design of simulation studies in medical statistics. *Statistics in Medicine* 2006;25(24):4279-92.
19. Sigal MJ, Chalmers RP. Play it again: Teaching statistics with Monte Carlo simulation. *Journal of Statistics Education* 2016;24(3):136-56.

20. Sauerbrei W, Abrahamowicz M, Altman DG, et al. STRengthening Analytical Thinking for Observational Studies: the STRATOS initiative. *Statistics in Medicine* 2014;33(30):5413-32.
21. Economist T. Unreliable research: trouble at the lab. *Economist* 2013.
22. Michels KB, Saracci R, Lynch J, et al. The Education Corner: updates on new and established core concepts and methods in epidemiology. *International Journal of Epidemiology* 2012;41(2):333-4.
23. Edwards JK, Cole SR, Westreich D. All your data are always missing: incorporating bias due to measurement error into the potential outcomes framework. *International Journal of Epidemiology* 2015;44(4):1452-9.
24. Howe CJ, Cain LE, Hogan JW. Are All Biases Missing Data Problems? *Current Epidemiology Reports* 2015;2(3):162-71.
25. Bonander C, Strömberg U. Methods to handle missing values and missing individuals. *European Journal of Epidemiology* 2019;34(1):5-7.
26. Westreich D. Berkson's Bias, Selection Bias, and Missing Data. *Epidemiology* 2012;23(1):159-64.
27. Lesko CR, Buchanan AL, Westreich D, et al. Generalizing study results: a potential outcomes perspective. *Epidemiology (Cambridge, Mass)* 2017;28(4):553.
28. Blackwell M, Honaker J, King G. A Unified Approach to Measurement Error and Missing Data: Overview and Applications. *Sociological Methods & Research* 2015;46(3):303-41.
29. Cole SR, Chu H, Greenland S. Multiple-imputation for measurement-error correction. *International Journal of Epidemiology* 2006;35(4):1074-81.
30. van Smeden M, Penning de Vries BBL, Nab L, et al. Approaches to addressing missing values, measurement error, and confounding in epidemiologic studies. *Journal of Clinical Epidemiology* 2021;131:89-100.
31. Keogh RH, White IR. A toolkit for measurement error correction, with a focus on nutritional epidemiology. *Statistics in medicine* 2014;33(12):2137-55.