uitnodiging

voor het bijwonen
van de openbare verdediging
van het proefschrift

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

door
jungyeon choi

op donderdag
22 juni 2023
16:15 uur
academiegebouw
rapenburg 73, leiden

paranimfen
ruifang li & willem van wijk

graag aanmelden via
promotie.jungyeon.choi@gmail.com

Stellingen

behorende bij het proefschrift

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

1. There is no one optimal statistical method that can handle biases across every study setting. Each source of bias should be handled on the basis of context-specific knowledge. (this thesis)
2. Multiple imputation is not a panacea to handle missing values and should be used more consciously. (this thesis)
3. Incorporating experts' content knowledge is recommended to detect measurement errors in time-serial data rather than solely relying on automated approaches. (this thesis)
4. A research question such as 'what is the effect of X on Y?' requires further elaboration. One should consider whether and how medication use or other factors has affected the measurements of interest. (this thesis)
5. Problems of confounding, selection, and measurement bias can be addressed with a question; what is the missing information? This calls for unified perspectives for addressing these biases.
6. Conducting comparison studies evaluating existing methods should be incentivized. For many analysis problems, the issue is not a lack of available methods; rather, it is a lack of accessibility to available methods. (after STRATOS initiative)
7. Simulations allow empirical comparisons between available methods under various data structures and violation of assumptions. Utilizing simulation studies will benefit clinical researchers.
8. Even in the emergence of big data and machine learning, careful considerations of the research setting, clinical knowledge, and study designs remain highly important – possibly more than ever.
9. Prisoners in a cave we (epidemiologists) are, looking at shadows (data) cast upon the cave wall. The shadows reflect a fragment of the real world (medical reality). (after The Allegory of the Cave).
10. "Every new discovery is just a reminder - we are all small and stupid. [...] all of that exists inside of one universe out of who knows how many" (Everything Everywhere All At Once, 2022). Because nothing matters, everything we give meaning matters.

handling missing data, selection bias, and measurement error

in observational studies

jungyeon choi

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

Jungyeon Choi

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

Ph.D. Thesis. Department of Clinical Epidemiology, Leiden University Medical Center

Provided by thesis specialist Ridderprint, ridderprint.nl
Printing: Ridderprint
Layout and design: Wiebke Keck, persoonlijkproefschrift.nl
Cover design: illustration Dick Bruna © copyright Mercis bv, 1989

ISBN: 978-94-6483-142-9

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

Proefschrift

ter verkrijging van
de graad van doctor aan de Universiteit Leiden,
op gezag van rector magnificus prof.dr.ir. H. Bijl,
volgens besluit van het college voor promoties
te verdedigen op donderdag 22 juni 2023
klokke 16:15 uur

door

Jungyeon Choi
geboren te Seoul, Zuid-Korea
in 1990

Promotors

Prof. dr. S. le Cessie

Prof. dr. O.M. Dekkers

Ledencommissie

Prof. dr. R.H.H. Groenwold	
Prof. dr. M.J. de Rooij	Leiden University
Prof. dr. H.F. Lingsma	Erasmus University Medical Center
Dr. R. de Mutsert	

Table of Contents

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.

Chapter 2

A comparison of different methods to handle missing data in the context of propensity score analysis

Published in Eur J Epidemiol. 2019 Jan; 34(1): 23-26

Jungyeon Choi, Olaf M. Dekkers, Saskia le Cessie

Abstract

Propensity score analysis is a popular method to control for confounding in observational studies. A challenge in propensity methods is missing values in confounders. Several strategies for handling missing values exist, but guidance in choosing the best method is needed.

In this simulation study, we compared four strategies for handling missing covariate values in propensity matching and propensity weighting. These methods include complete case analysis, missing indicator method, multiple imputation, and combining multiple imputation and missing indicator method. Concurrently, we aimed to provide guidance in choosing the optimal strategy. Simulated scenarios varied regarding the missing mechanism, presence of effect modification, or unmeasured confounding. Additionally, we demonstrated how missingness graphs help clarify the missing structure.

When no effect modification existed, complete case analysis yielded valid causal treatment effects even when data were missing not at random. In some situations, complete case analysis was also able to partially correct for unmeasured confounding. Multiple imputation worked well if the data were missing (completely) at random, and if the imputation model was correctly specified. In the presence of effect modification, more complex imputation models than default options of commonly used statistical software were required. Multiple imputation may fail when data are missing not at random. Here, combining multiple imputation and the missing indicator method reduced the bias as the missing indicator variable can be a proxy for unobserved confounding.

The optimal way to handle missing values in covariates of propensity score models depends on the missing data structure and the presence of effect modification. When effect modification is present, default settings of imputation methods may yield biased results even if data are missing at random.

1. Introduction

Observational studies potentially suffer from confounding. First introduced by Rosenbaum and Rubin [1], Propensity score methods are increasingly used in medical research to handle confounding [2-5]. When the observed baseline characteristics are sufficient to correct for confounding bias and the propensity model is correctly specified, propensity score analysis creates conditional exchangeability between persons with the same propensity score. Numerous studies provide illustrations and discussions on the performance of different propensity score approaches [6, 3, 4, 7-11].

Besides confounding, observational studies often have missing values in covariates. Missing values can occur by different mechanisms: values are *missing completely at random* (MCAR) when the probability that a value is missing is independent of observed and unobserved information (e.g., a lab measurement is missing, because a technician dropped a tube), *missing at random* (MAR) where the probability of missing depends only on observed information (e.g., lab measurements are only performed when other measured variables were abnormal), or *missing not at random* (MNAR) where the probability of missing depends on unobserved information (e.g., lab measurements are only performed when a doctor judged that a patient was in a severe condition, while the severity is not well-registered.) [12]. It is difficult, however, to decide on the type of missing mechanism, especially when distinguishing whether the data are *missing at random* or *not at random* [13, 14]. Especially in routinely collected data, variables are often selectively measured based on a patient's characteristics which are often not well-specified [15]. If those ill-defined characteristics are associated with the variable with missing values, data is missing not at random. External knowledge or assumptions about the clinical setting are required to distinguish whether the missing is at random or not at random.

How to estimate propensity scores when there are missing values is a challenge when studying causal associations [16]. There are different strategies to handle missing data in a propensity score analysis. The simplest approach is to discard all observations with missing data, a so-called complete case analysis [12, 17]. Including a missing indicator in a statistical model is another simple method. However, various studies showed that the method generally introduces bias [18-21]. Multiple imputation is a standard method to deal with missing data. Many studies have shown the advantage of multiple imputation and its superiority over other methods [12, 19, 22]. In combination with propensity scores, however, several questions arise: Should we include the outcome in the imputation model? Can we use the imputation methods implemented in standard software? How should we combine the results of the different propensity scores estimated in each imputed dataset?

The aim of this simulation study is to investigate how different strategies of handling missing values of covariates in a propensity score model can yield valid causal treatment effect estimates. To limit the scope of the study, we deal only with missing values in the baseline characteristics, a rather common situation in routinely collected data. We create simulation scenarios varying in their missing data mechanisms, the presence of heterogeneous treatment effects, and unmeasured confounding. Subsequently, the results are used to provide guidance in choosing an optimal strategy to handle missing data in the context of propensity score analysis.

2. Simulation description

We generated simulated data with missing values in one of the confounders and compared effect estimates obtained by using several different strategies to deal with missing data. In Section 2.1, we considered a situation without unmeasured confounding and with an equal treatment effect for all subjects. In section 2.2, we introduced a heterogeneous treatment effect. In Section 2.3, the simulations were extended by adding unmeasured confounding.

2.1. Simulation setting 1: No unmeasured confounding and a homogeneous treatment effect

In this simulation series, for each subject, we generated two continuous covariates X_1 and X_2. X_1 follows a normal distribution of mean 0 and standard deviation of 1. X_2 depended on X_1, where for subject i,

$$X_{2i} = 0.5\, X_{1i} + \varepsilon_i \qquad \text{with } \varepsilon_i \sim N(0, 0.75)$$

In this way, the standard deviation of X_2 is also 1 and the correlation between X_1 and X_2 is equal to 0.5. The treatment T was generated from the binomial distribution, with the probability for subject i to receive the treatment being equal to:

$$logit\big(P(T_i = 1|X_{1i}, X_{2i})\big) = -0.8 + 0.5\, X_{1i} + 0.5\, X_{2i}$$

In this way about 33% of the generated subjects received treatment. A continuous outcome was generated with the mean linearly related to X_1 and X_2:

$$Y_i = X_{1i} + X_{2i} + \varepsilon_i, \qquad \text{with } \varepsilon_i \sim N(0, 1)$$

For ease of interpretation of the results, we assumed that treatment T had no effect on the outcome for any of the subjects. Missing data were generated for 50% of the X_2 values in three different ways:

- A missing completely at random (MCAR) scenario: 50% of values are randomly set to missing in X_2

- A missing at random (MAR) scenario: The higher the value of X_1, the more likely for the X_2 value to be missing. Denoting R as a missing indicator of X_2, the probability of a missing X_2 value was equal to:

$$logit\big(P(R_i = 1)\big) = X_{1i}$$

- A missing not at random (MNAR) scenario: The higher the value of X_2, the more likely that the value was missing. The probability of a missing X_2 value was:

$$logit\big(P(R_i = 1)\big) = X_{2i}$$

Missingness-graphs (*m*-graph, for short) of each missing scenario are depicted in Figure 1. The missingness graph is a graphical tool to represent missing data, proposed by Mohan et al. [23]. Guidance for practical users is given in Thoemmes, Mohan [24]. These graphs are extensions to causal directed acyclic graphs (DAGs) where nodes indicate covariates and arrows indicate causal relations. When a covariate contains missing values (X_2 in our simulations), it is expressed by a dashed rectangle around the node. The node R represents the missingness of X_2, and can be referred to as a missing indicator of X_2. The observed portion of X_2 is represented as X_2*. When R=0, X_2* is identical to X_2, and when R=1, X_2* is missing. In our simulations, we restricted ourselves to the situation where missing values occur only in one covariate. However, *m*-graphs can be extended to situations with multiple covariates having missing values and, accordingly, with multiple missing indicator variables.

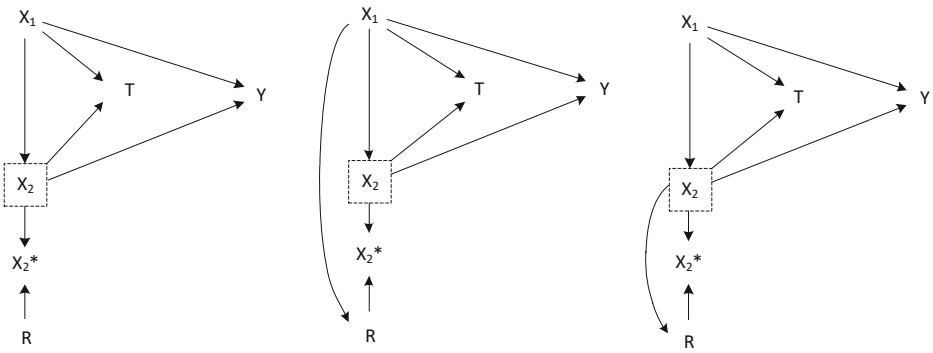


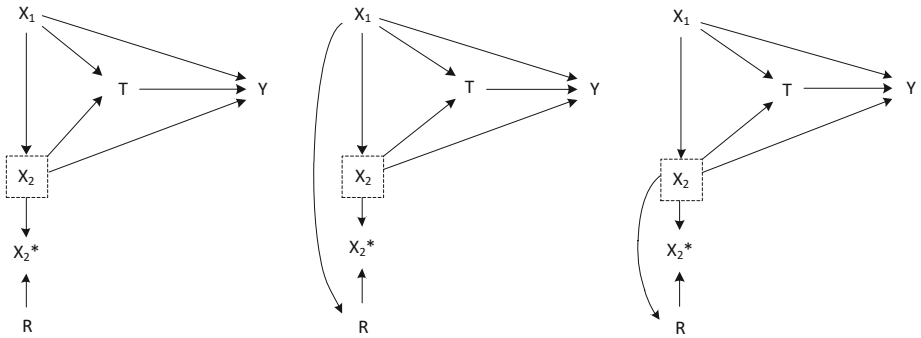
Figure 1. M-graphs for Simulation setting 1: MCAR scenario (a), MAR scenario (b), and MANR scenario (c)

2.2. Simulation setting 2: No unmeasured confounding and a heterogeneous treatment effect

The setup of this simulation series is the same as in Simulation setting 1, but here we assumed effect modification by X_2. That is,

$$Y_i = X_{1i} + X_{2i} + T_i X_{2i} + \varepsilon_i, \qquad \text{with } \varepsilon_i \sim N(0,1)$$

The average treatment effect in the population was equal to null as in Simulation setting 1. However, due to the effect modification by X_2, the average treatment effect was negative for subjects with $X_2 < 0$ and positive for subjects with $X_2 > 0$. Missing values were generated in the X_2 variable, following the same mechanisms as in Simulation setting 1. The *m*-graphs for each scenario are depicted in Figure 2. In these m-graphs, there is an arrow from the treatment assignment (T) to the outcome (Y) because, for some subjects, the treatment has an effect on their outcome.



(a) **(b)** **(c)**

Figure 2. M-graphs for Simulation setting 2: MCAR scenario (a), MAR scenario (b), and MANR scenario (c)

2.3. Simulation setting 3: Unmeasured confounding and a homogeneous treatment effect

In this series of simulations, we assumed an additional unobserved confounder U, normally distributed with a mean of 0 and standard deviation of 1 and independent from X_1. X_2 depended on X_1 and U, where for subject *i*,

$$X_{2i} = 0.5\, X_{1i} + 0.5\, U_i + \varepsilon_i \qquad \text{with } \varepsilon_i \sim N(0, 0.5)$$

The probability of receiving the treatment depended on X_1, X_2, and U as follows:

$$logit\big(P(T_i = 1 | X_{1i}, X_{2i}, U_i)\big) = -0.85 + 0.5\, X_{1i} + 0.5\, X_{2i} + 0.5\, U_i$$

This way, about 33% of the generated subjects received the treatment. The outcome now depended on X_1, X_2 and U:

$$Y_i = X_{1i} + X_{2i} + U_i + \varepsilon_i, \qquad \text{with } \varepsilon_i \sim N(0,1)$$

Here, we assumed a homogeneous treatment effect which was set to null. We considered two missing scenarios for X_2, one according to the MCAR mechanism and the other MNAR mechanism.

- A MCAR scenario: 50% of values are randomly set to be missing in X_2
- A MNAR scenario: Here, we considered a common situation in routinely collected health care data where the missing of X2 depended on the unobserved confounder U. We set the value of X2 to be missing if U > 0

A MAR scenario was not considered in this simulation setting. This is because we were interested in comparing a situation where an unmeasured confounder U affect the missingness of X_2 (MNAR) to a situation where it does not affect the missingness of X_2 (MCAR). The m-graphs for these scenarios are illustrated in Figure 3.

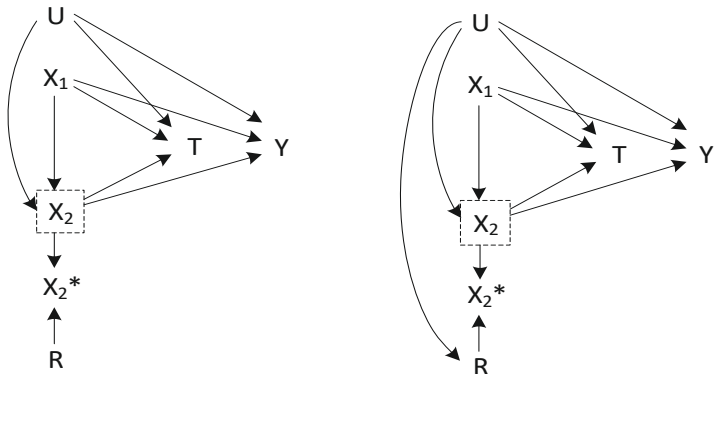


Figure 3. M-graphs for Simulation setting 3: MCAR scenario (a), MNAR scenario (b)

2.4. Analysis of the simulated datasets

In every simulated dataset, we estimated propensity scores by logistic regression. The treatment effect was estimated by i) propensity matching and ii) propensity weighting. For the matching procedure, we matched a treated subject to an untreated subject using one-to-one nearest neighbour matching without replacement and 0.1 caliper distance on the logit scale. In propensity weighting, the so-called inverse probability weighting, treated subjects are weighted by *1/propensity score*, and untreated subjects are weighted by *1/(1-propensity score)*. Note that causal effects estimated by propensity matching and

propensity weighting are different from each other. The matching estimates the average treatment effect in the *treated population*, while the weighting method estimates the average treatment effect in the *total population*. For handling missing values, we applied the following four different methods.

2.4.1. Complete case analysis

In this approach, only observations with complete information are used for analysis.

2.4.2. Missing indicator method

When a covariate contains missing values, they were replaced by one single value, for example, by the value 0. Additionally, a missing indicator variable was created, with 1 indicating that the corresponding value is missing and 0 indicating that it is observed. The missing indicator variable was then added as a covariate in a propensity score model. When there are multiple covariates with missing values, missing indicators will be created for each covariate which will be all added to a propensity score model.

2.4.3. Multiple imputation

The third method considered was multiple imputation. The chained equation (MICE) procedure was used, a commonly used imputation method that assumes data are missing at random [25]. We used the default options of MICE version 3.3.0 [26] in R version 3.5.1: predictive mean matching via a regression model with main effects of X_1, X_2, T, and with or without Y. In this way, the simulations reflect how most applied researchers using R would perform multiple imputation. Predictive mean matching is also readily available in SAS version 9.4, Stata version 15, and IBM SPSS version 25.0, and it is recommended when data contains both continuous and discrete values [27, 28]. As a sensitivity analysis, we repeated Simulation setting 2 using MICE with Bayesian linear regression, since many researchers will opt for this method when covariates and outcomes are continuous.

In Simulation setting 2, where a heterogeneous treatment effect exists, we additionally used a more extensive imputation model with three interaction terms included; the interaction between T and X_1, T and Y, and X_1 and Y. Adding interaction terms between the variables in a multiple imputation regression model is advocated by Tilling et al. [29]. For every multiple imputation, ten imputed datasets were generated. A treatment effect was estimated within each imputed dataset using the propensity score methods. Using Rubin's rule, the ten treatment effects were then combined into a single treatment effect. This method is referred to as the within method [30].

We explored whether the outcome should be included in the imputation model. The idea behind the propensity score methods is that the probability of receiving the treatment is modelled without knowing the outcome [16], which is why some researchers argue that the outcome should not be used in the imputation model [31]. The purpose of multiple

imputation, however, is a reconstruction of data to retain the original relationship between the covariates as much as possible, for which the outcome could provide valuable information [32-35]. This suggests that the outcome should be added to an imputation model.

2.4.4. Multiple imputation together with missing indicator

The fourth method was a combination of multiple imputation and the missing indicator method. Multiple imputation was used to impute the missing values. Afterward, both the imputed covariate and a missing indicator variable were added to the propensity score model [36]. Multiple imputation was performed following the same procedure as in Section 2.4.3, where the treatment effect is estimated by the within method.

2.5. Simulation summary

Each simulation run generated a thousand observations and was repeated a thousand times. We summarised the simulation results by calculating the mean treatment effects over the simulations and the standard deviation of the estimated treatment effects. As overall performance measures, we calculated the mean squared error, which is the squared distance between the estimated treatment effect and the true treatment effect averaged over the simulations.

In Simulation setting 1 and 3, the true treatment effect was null for all subjects, which means that mean estimated treatment effects deviating from 0 demonstrate bias has been introduced. In Simulation setting 2, the average treatment effect in the *population*; the causal effect estimated by propensity weighting, was also equal to null. However, due to the heterogeneous treatment effect, the average treatment effect in the *treated*; the causal effect estimated by propensity matching, differed from null. In this simulation setting, the treatment effect for individual *i* is equal to $X_{2i,}$ which implies the average treatment effect in the treated would be $E[X_2|T=1]$. In this simulated example, $E[X_2|T=1]$ was equal to 0.432.

3. RESULTS

3.1. Simulation setting 1: No unmeasured confounding and a homogeneous treatment effect

Figure 4 (left) displays the mean estimated effects of the propensity weighting analysis in Simulation setting 1 and their 5^{th} and 95^{th} percentile range. Table 1 shows the mean estimates with standard deviations and mean squared errors from the propensity matching and the propensity weighting. Complete case analysis yielded unbiased treatment effect estimates in all scenarios, even when data were missing not at random. The missing indicator method alone resulted in biased estimates in all scenarios. The

results suggested that the outcome should be included in an imputation model, because the imputation models not including the outcome resulted in bias. In the MCAR and MAR scenario, multiple imputation including the outcome yielded the smallest mean squared errors, and combining multiple imputation and the missing indicator method worked as efficiently. In the MNAR scenario, combining multiple imputation and the missing indicator method was slightly less biased than multiple imputation alone.

3.2. Simulation setting 2: No unmeasured confounding and a heterogeneous treatment effect

Figure 4 (middle) visualises the results of the propensity weighting analysis of Simulation setting 2, and Table 2 summarises the results of the propensity matching and the propensity weighting. Here, the complete case analysis yielded negatively biased results in the MAR or MNAR scenarios. This is because subjects with higher X_2 values, for whom the treatment was most beneficial, had a higher probability of being excluded from the analyses. The missing indicator method was still biased in all scenarios. The amount of bias, however, was relatively small in the MNAR scenario. We observed a remarkable result in the MAR scenario: the default multiple imputation method yielded biased effect estimates, even when the outcome was included in the imputation model and when a missing indicator was added to the propensity model. When more elaborate imputation regression models with specified interaction terms were used, the bias from the propensity weighting was much smaller, although a slight bias still remained (0.013).

The results of propensity matching, even in the situation without any missing values (0.327), deviated from the treatment effect in all treated subjects (0.432). This discrepancy is a general problem of propensity score matching [37-39]. A large caliper distance allows treated subjects with high propensity scores to be matched to untreated subjects with lower propensity scores, which will result in residual confounding. A smaller caliper distance reduces the confounding bias. However, many subjects, especially the subjects with a high propensity score, may not be matched. Therefore, the treatment effect in the treated *who are matched* may deviate from the treatment effect in *all* treated. The size of this discrepancy depends on the heterogeneity of the treatment effect. In this simulation setting, we used matching without replacement with a caliper distance of 0.1, which allows rather a tight matching. Thus, for some of the treated subjects with a high propensity score, whose treatment effect was more effective, no adequate untreated match could be found. As we were specifically interested in the additional bias under the different missing mechanisms, we used the estimate of propensity matching without any missing data (0.327) as a reference. Once more, we observed that multiple imputation with interaction terms performed best as it did in propensity weighting analysis.

The results of multiple imputation with Bayesian regression methods done in a sensitivity analysis did not largely differ from the results of predictive mean matching (see Appendix for the results in Simulation setting 2).

3.3. Simulation setting 3: Unmeasured confounding and a homogeneous treatment effect

Figure 4 (right) displays the results of the propensity weighting of Simulation setting 3, and Table 3 summarises the results of propensity matching and propensity weighting. Due to the unmeasured confounder U, bias remained in the propensity analyses even when there were no missing values. In the MNAR scenario where the missingness of X_2 depends on U, two methods were able to reduce the unmeasured confounding effect: the combined method and, somewhat surprisingly, the complete case analysis. The combined method partially adjusted for U by adding R to the propensity model; the complete case analysis used restriction to partially adjust for U, using only those with complete data. The results here were substantially less biased than the propensity analyses performed in complete data without missing values.

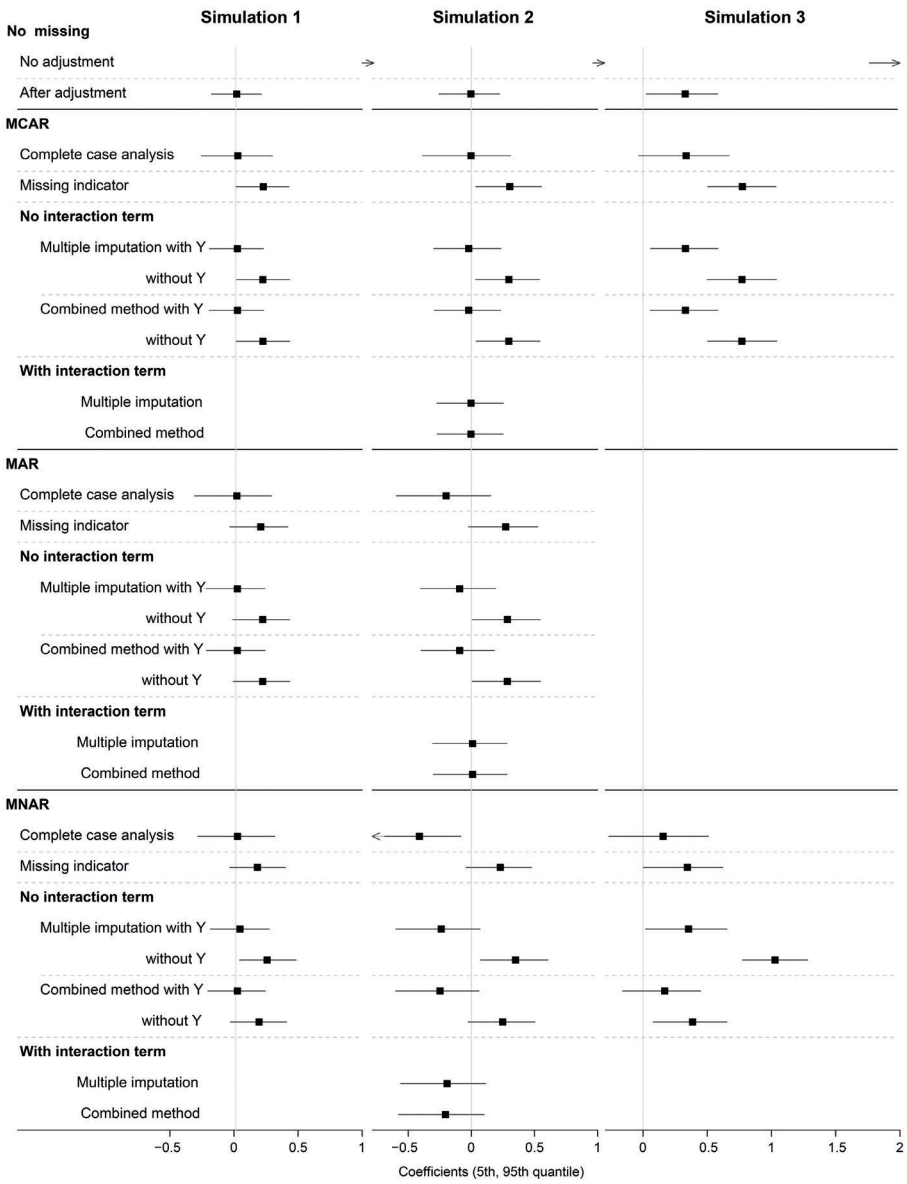


Figure 4. Mean treatment effects and their 5th and 95th percentile ranges estimated by propensity weighting in Simulation setting 1 (left), 2 (middle) and 3 (right). For each missing scenario, missing data are handled with complete case analysis, missing indicator method, multiple imputation, and the combination of multiple imputation and missing indicator method (Combined method). The vertical lines represent the true treatment effect.

Table 1. Results of treatment effect estimates from propensity matching and propensity weighting when assuming a homogeneous treatment effect and no unmeasured confounding. For each missing scenario, missing data are handled with complete case analysis, missing indicator method, multiple imputation, and the combination of multiple imputation and missing indicator (Combined method).

		Homogeneous treatment effect					
		Propensity matching			Propensity weighting		
		coefficient		MSE	coefficient		MSE
		mean	sd		mean	sd	
No missing	**No adjustment**	1.298	0.123	1.700	1.298	0.123	1.700
	After adjustment	0.044	0.085	0.009	0.006	0.109	0.012
MCAR	**Complete case analysis**	0.043	0.121	0.016	0.014	0.152	0.023
	Missing indicator	0.238	0.095	0.066	0.189	0.111	0.048
	Multiple imputation						
	with Y	0.047	0.086	0.010	0.011	0.113	0.013
	without Y	0.219	0.087	0.056	0.186	0.110	0.047
	Combined method						
	with Y	0.048	0.087	0.010	0.011	0.112	0.013
	without Y	0.218	0.087	0.055	0.187	0.110	0.047
MAR	**Complete case analysis**	0.024	0.128	0.017	0.007	0.165	0.027
	Missing indicator	0.259	0.099	0.077	0.172	0.123	0.044
	Multiple imputation						
	with Y	0.052	0.092	0.011	0.010	0.122	0.015
	without Y	0.244	0.090	0.068	0.185	0.120	0.049
	Combined method						
	with Y	0.050	0.092	0.011	0.010	0.122	0.015
	without Y	0.243	0.090	0.067	0.185	0.120	0.048
MNAR	**Complete case analysis**	0.025	0.129	0.017	0.012	0.166	0.028
	Missing indicator	0.231	0.098	0.063	0.149	0.122	0.037
	Multiple imputation						
	with Y	0.069	0.095	0.014	0.029	0.123	0.016
	without Y	0.248	0.091	0.070	0.215	0.118	0.060
	Combined method						
	with Y	0.052	0.093	0.011	0.011	0.122	0.015
	without Y	0.211	0.088	0.053	0.160	0.119	0.040

Table 2. Results of treatment effect estimates from propensity matching and propensity weighting when assuming X_2 is an effect modifier and no unmeasured confounder exists. Here, multiple imputation is done in two ways; the commonly used method (no interaction term) and the elaborated method (interaction terms included).

		Heterogeneous treatment effect						
		Propensity matching				Propensity weighting		
		coefficient		Bias	MSE	coefficient		MSE
		mean	sd			mean	sd	
No missing	**No adjustment**	1.736	0.156	1.409	2.011	1.736	0.156	3.040
	After adjustment	0.327	0.093	0.000	0.009	-0.003	0.152	0.023
MCAR	**Complete case analysis**	0.300	0.133	-0.027	0.018	-0.003	0.219	0.048
	Missing indicator	0.574	0.120	0.247	0.075	0.305	0.162	0.119
	No interaction term							
	Multiple imputation							
	with Y	0.315	0.103	-0.012	0.011	-0.021	0.168	0.029
	without Y	0.542	0.108	0.215	0.058	0.297	0.158	0.113
	Combined method							
	with Y	0.315	0.102	-0.012	0.011	-0.021	0.169	0.029
	without Y	0.541	0.110	0.214	0.058	0.297	0.158	0.113
	Interaction terms							
	Multiple imputation	0.316	0.103	-0.011	0.011	-0.002	0.166	0.028
	Combined method	0.316	0.104	-0.011	0.011	-0.003	0.166	0.028
MAR	**Complete case analysis**	0.129	0.147	-0.198	0.061	-0.200	0.241	0.098
	Missing indicator	0.620	0.122	0.293	0.101	0.272	0.179	0.106
	No interaction term							
	Multiple imputation							
	with Y	0.251	0.107	-0.076	0.017	-0.093	0.181	0.042
	without Y	0.579	0.112	0.252	0.076	0.286	0.173	0.111
	Combined method							
	with Y	0.250	0.108	-0.077	0.017	-0.092	0.182	0.042
	without Y	0.580	0.113	0.253	0.077	0.285	0.173	0.111
	Interaction terms							
	Multiple imputation	0.330	0.116	0.003	0.013	0.010	0.185	0.034
	Combined method	0.330	0.116	0.003	0.013	0.010	0.185	0.034

Table 2. Results of treatment effect estimates from propensity matching and propensity weighting when assuming X_2 is an effect modifier and no unmeasured confounder exists. Here, multiple imputation is done in two ways; the commonly used method (no interaction term) and the elaborated method (interaction terms included). *(continued)*

		Heterogeneous treatment effect						
		Propensity matching				Propensity weighting		
		coefficient		Bias	MSE	coefficient		MSE
		mean	sd			mean	sd	
MNAR	**Complete case analysis**	-0.111	0.141	-0.438	0.211	-0.411	0.224	0.219
	Missing indicator	0.588	0.121	0.261	0.082	0.230	0.171	0.082
	No interaction term							
	Multiple imputation							
	with Y	0.151	0.114	-0.176	0.044	-0.238	0.207	0.100
	without Y	0.586	0.112	0.259	0.080	0.350	0.165	0.150
	Combined method							
	with Y	0.140	0.111	-0.187	0.047	-0.248	0.206	0.104
	without Y	0.546	0.108	0.219	0.060	0.248	0.165	0.089
	Interaction terms							
	Multiple imputation	0.182	0.117	-0.145	0.035	-0.192	0.208	0.080
	Combined method	0.170	0.114	-0.157	0.038	-0.205	0.264	0.112

Table 3. Results of treatment effect estimates from propensity matching and inverse probability weighting when an unmeasured confounding exists.

		Homogeneous treatment effect / Unmeasured confounding					
		Propensity matching			Propensity weighting		
		coefficient		MSE	coefficient		MSE
		mean	sd		mean	sd	
No missing	**No adjustment**	2.011	0.154	4.068	2.011	0.154	4.068
	After adjustment	0.377	0.111	0.154	0.328	0.168	0.136
MCAR	**Complete case analysis**	0.362	0.152	0.154	0.336	0.233	0.167
	Missing indicator	0.870	0.138	0.776	0.774	0.171	0.628
	Multiple imputation						
	with Y	0.376	0.119	0.155	0.330	0.171	0.138
	without Y	0.807	0.119	0.665	0.771	0.165	0.621
	Combined method						
	with Y	0.375	0.119	0.155	0.330	0.171	0.138
	without Y	0.808	0.119	0.667	0.770	0.165	0.620
MNAR	**Complete case analysis**	0.145	0.163	0.048	0.157	0.255	0.089
	Missing indicator	0.514	0.117	0.277	0.345	0.197	0.158
	Multiple imputation						
	with Y	0.422	0.141	0.197	0.354	0.200	0.165
	without Y	1.003	0.129	1.023	1.028	0.154	1.079
	Combined method						
	with Y	0.240	0.114	0.071	0.169	0.191	0.065
	without Y	0.469	0.105	0.231	0.386	0.175	0.180

4. Guidance for the optimal strategy to handle missing values in baseline covariates in the context of propensity score analysis

The aim of a propensity score analysis is to obtain an average treatment effect in a certain population. To explain, we use the following notation in which every subject can have two potential outcomes:

- Y^1; the outcome if the person receives treatment 1
- Y^0; the outcome if the person receives treatment 0

Propensity weighting aims to estimate the average treatment effect in the *whole population* (ATE), which is equal to: ATE = $E[Y^1-Y^0]$. With propensity matching, where treated subjects are matched to untreated subjects, the aim is to estimate the average treatment effect in the *treated population* (ATT): ATT = $E[Y^1-Y^0[T=1]$ Several standard causal inference conditions, such as exchangeability, consistency, and positivity, should hold to estimate these causal effects without bias [40]. Whether the unbiased causal effects can still be estimated when missing values are present in the covariates of a propensity score depends on several elements: type of missingness, presence of effect modification, and the population of interest. In the following section, we discuss under which criteria the four methods dealing with missing values will yield a valid causal treatment effect in the context of propensity score analysis.

- **Complete case analysis, when does it work?**

When there is no unmeasured confounding, and the propensity score model is well specified, propensity weighting using complete cases will yield a valid estimate of a causal treatment effect, which will be the causal treatment effect in the *subjects without missing values:*

$$E[Y^1 - Y^0 | R = 0]$$

This means that propensity weighting using complete case analysis will yield valid estimates of the ATE in the population when the mean treatment effect in the fully observed subjects is equal to that of the subjects with missing values. That is:

$$E[Y^1 - Y^0 | R = 0] = E[Y^1 - Y^0 | R = 1] = E[Y^1 - Y^0] \qquad (1)$$

When data are missing completely at random, condition (1) will hold, because the probability of a missing value does not depend on any observed or unobserved variable. This means that the covariate with missing values is independent of its own missing indicator variable. The m-graphs may be helpful in identifying whether this independency holds. In the m-graphs in Figure 1a and 2a, these conditions hold because X_2 and R are unconditionally *d-separated,* meaning that there is no open path between X_2 and R.

When no effect modification and no unmeasured confounding is present, condition (1) will also hold since the treatment effect in the total population will be equal to the treatment effect in any subgroup regardless of the missing mechanism of data. This was the case in Simulation 1, where the effect of the treatment was constant across subjects. In this scenario, the complete case analysis yielded unbiased results even when the missing was not at random. Analogous arguments can be given for propensity matching using complete cases. The propensity matching will yield valid estimates if:

$$E[Y^1 - Y^0 | R = 0, T = 1] = E[Y^1 - Y^0 | R = 1, T = 1] = E[Y^1 - Y^0 | T = 1] \quad (2)$$

Even when there is unmeasured confounding, complete case analysis may be a useful way to handle missing values. Think of a situation where the severity of a disease determines whether certain laboratory tests will be performed. The severity of disease here may be an unmeasured confounder, which determines the values of observed covariates (in this case, the laboratory measurements) to be missing. This is a comparable situation to the MNAR scenario of Simulation setting 3. Here, the complete case analysis yielded less biased results. By restricting the analysis to subjects with R=0 (only the subjects with severe diseases who therefore have all lab measurements), the results were partially adjusted for the unmeasured confounder.

- **Missing Indicator, when does it work?**

In general, we do not recommend solely using the missing indicator method for handling missing values in confounders. The method is prone to bias because the information of the missing portion of the covariates is replaced by a dichotomous missing indicator R, consequently resulting in residual confounding. However, when data are missing not at random and the covariate with missing value is strongly associated with its missing indicator, the missing indicator variable in a propensity model may yield a smaller bias than the model without it. This was the case in the MNAR scenarios of Simulation setting 1 and 2. Similarly, when the missing of X_2 is strongly related to an unmeasured confounder U, the partial effect of U can be recovered by adding R to the propensity model. This was seen in the MNAR scenario of Simulation 3.

- **Multiple imputation, when does it work?**

The aim of multiple imputation is to recover the joint distribution of covariates, treatment, and outcome by reconstructing the missing values using the information from observed data. When there is no unmeasured confounding, multiple imputation in the context of propensity score analysis will be a valid approach under the following conditions:

i) Data are missing at random or completely at random, meaning the missing values are *recoverable* from the observed data. M-graphs can be used to visually determine whether the missing mechanism is at random. In m-graphs, the missing at random mechanism means that all paths between a covariate with missing values and its

missing indicator can be blocked by conditioning on measured variables. In DAG terms, it is said; two variables are *d-separated*. In our study, this was the case in Figure 1a, 1b, 2a, and 2b. Note that in Figure 1b and 2b, the path between X_2 and R can be blocked by conditioning on X_1.

ii) An imputation model should be correctly specified. This requires that:
   a. the outcome should be included in the imputation model.
   b. interaction terms between the covariates, treatment, and outcome should be included in the imputation model if a heterogeneous treatment effect is present.

In Simulation setting 1, multiple imputation yielded unbiased results even though it was used to impute a non-recoverable X_2. Note that the reason why multiple imputation worked well in this scenario was 1) the covariates, treatment, and the outcome in the model were linearly related, and 2) missing values in X_2 were generated probabilistically, which means the information of higher X_2 values could be gained in the data. This result is due to the simulation scenario we generated and should not be taken as showing multiple imputation can be used when data are missing not at random.

- **What to do in situations where complete case analysis or multiple imputation fails?**

We saw in the previous section it is important that a researcher is aware of the missing mechanism and whether strong heterogeneity is present. Depending on the missing mechanism and the heterogeneity in the treatment effect, both complete case analysis and multiple imputation may fail. Whether the treatment effect is heterogeneous can be explored by subgroup analysis and comparing the estimated effects across the groups. When there is a large difference across the subgroups, interaction terms should be specified in the multiple imputation. This was shown in Simulation setting 2.

The missing mechanism behind the data can be explored by drawing the expected causal structure and missing structure in an m-graph. When complete case analysis and multiple imputation are expected to fail, the combination of multiple imputation and the missing indicator method could be used to partially recover the effect of missing portions of covariates. For example, in the MNAR scenario of Simulation setting 3, the combined approach performed better than multiple imputation alone and even better than the analysis of the data without any missing values. When the relation between R and U is stronger, more of the effect of the unmeasured confounder will be recovered.

5. Discussion

Our simulations showed that there is no single method to handle missing values in covariates of a propensity score model which would perform optimally in all situations. The optimal strategy depends on the missing data structure and whether there is effect modification or unmeasured confounding. We focussed on missing values in covariates, because in routinely collected data, baseline patient characteristics are often incomplete while prescribed treatments and important outcomes of patients will be more generally recorded.

Our results cannot be generalized to situations when there are missing values in the treatment assignment or the outcome. An example of this is that under homogenous treatment effect and no unmeasured confounding, complete case analysis will yield biased results if the outcome is missing not at random.

Propensity score analysis mimics randomized control studies by creating conditional exchangeability between the subjects with the same propensity score. Both propensity weighting and matching aim to obtain valid estimates of marginal treatment effects. This is different from outcome regression analysis which estimates conditional treatment effects. Unlike outcome regression models, no assumptions about treatment-outcome relation and the effect of the confounders on the outcome have to be made in propensity score analysis; only the propensity score model has to be correctly specified. This is an advantage, especially when the outcome is rare, in which case fitting an extensive outcome model is not possible.

When using multiple imputation, the advantage of not having to formulate a treatment-outcome relation model disappears. In our simulations, we showed that all variables associated with the covariates with missing values, including the outcome, should be included in the imputation model. Furthermore, when effect modification is present, the interaction terms between the variables should be correctly specified in the imputation model as well. The results correspond to the idea that imputation models should reflect the complexity of the data analysis procedure [41, 42]. When complex modelling is needed for multiple imputation, an alternative to propensity score analysis could be to use an outcome regression model with specified interaction terms. By fitting this outcome regression model, one can predict potential outcomes under treatment and no treatment for *every individual*. Then, the *average* potential outcomes can be estimated by integrating over the covariate distribution and used to obtain the average treatment effect in the population [40].

Multiple imputation is not a panacea to handle missing values and should be used more consciously. In our simulations, we demonstrated that a default option for multiple imputation in commonly used software such as SAS, Stata, SPSS, or R yielded

biased results (based on Simulation 2) even when data were missing at random and no unmeasured confounding was present.

Complete case analysis may often be a good method to deal with missing values in covariates. Although statistical efficiency is lost, estimated effects still have a causal interpretation if there is no unmeasured confounding. In these cases, it is up to the researcher to determine how generalizable these results are to the general population of interest. In the case of substantial heterogeneity of treatment effects, generalizability cannot be taken for granted.

When unmeasured confounding is present, all standard missing data methods fail to provide valid estimates. Complete case analysis, however, may reduce the bias by controlling the unmeasured confounding by restriction. The use of an indicator variable (with or without multiple imputation) may also reduce the bias, because the indicator variable functions as a proxy for the unmeasured confounding.

A recent systematic review on how missing data are addressed with propensity score methods in observational comparative effectiveness studies showed that among 167 studies conducted from 2010 to 2017, only 86 (51%) discussed missing data issues and only 12 (7%) provided reasons for missingness [43]. Our simulation study showed that it is important to make assumptions about the expected relationship between the unobserved and observed covariates. This allows one to understand the expected missing structure of the data and to handle missing values more cautiously. We recommend researchers to use m-graphs to draw their assumption between the covariates and their missing indicator explicitly. In summary, in the context of propensity score analysis, we urge researchers to consciously choose missing data strategies while considering the missing data mechanisms, possible unmeasured confounding, and heterogeneity of treatment effects.

References

1. Rosenbaum PR, Rubin DB. The Central Role of the Propensity Score in Observational Studies for Causal Effects. Biometrika. 1983;70(1):41-55. doi:10.2307/2335942.
2. Austin PC. A critical appraisal of propensity-score matching in the medical literature between 1996 and 2003. Stat Med. 2008;27(12):2037-49. doi:10.1002/sim.3150.
3. Williamson E, Morley R, Lucas A, Carpenter J. Propensity scores: from naive enthusiasm to intuitive understanding. Stat Methods Med Res. 2012;21(3):273-93. doi:10.1177/0962280210394483.
4. Austin PC. The relative ability of different propensity score methods to balance measured covariates between treated and untreated subjects in observational studies. Med Decis Making. 2009;29(6):661-77. doi:10.1177/0272989X09341755.
5. Austin PC. Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples. Stat Med. 2009;28(25):3083-107. doi:10.1002/sim.3697.
6. 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.
7. Austin PC. The performance of different propensity score methods for estimating marginal odds ratios. Stat Med. 2007;26(16):3078-94. doi:10.1002/sim.2781.
8. Austin PC. The performance of different propensity-score methods for estimating relative risks. J Clin Epidemiol. 2008;61(6):537-45. doi:10.1016/j.jclinepi.2007.07.011.
9. Austin PC, Grootendorst P, Anderson GM. A comparison of the ability of different propensity score models to balance measured variables between treated and untreated subjects: a Monte Carlo study. Stat Med. 2007;26(4):734-53. doi:10.1002/sim.2580.
10. Austin PC, Schuster T. The performance of different propensity score methods for estimating absolute effects of treatments on survival outcomes: A simulation study. Stat Methods Med Res. 2016;25(5):2214-37. doi:10.1177/0962280213519716.
11. d'Agostino RB. Tutorial in biostatistics: propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. Stat Med. 1998;17(19):2265-81.
12. Donders AR, van der Heijden GJ, Stijnen T, Moons KG. Review: a gentle introduction to imputation of missing values. J Clin Epidemiol. 2006;59(10):1087-91. doi:10.1016/j.jclinepi.2006.01.014.
13. White IR, Royston P, Wood AM. Multiple imputation using chained equations: Issues and guidance for practice. Statistics in Medicine. 2011;30(4):377-99. doi:10.1002/sim.4067.
14. Horton NJ, Lipsitz SR. Multiple Imputation in Practice. The American Statistician. 2001;55(3):244-54. doi:10.1198/000313001317098266.
15. Benchimol EI, Smeeth L, Guttmann A, Harron K, Moher D, Petersen I et al. The REporting of studies Conducted using Observational Routinely-collected health Data (RECORD) Statement. PLOS Medicine. 2015;12(10):e1001885. doi:10.1371/journal.pmed.1001885.
16. 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. doi:10.1080/01621459.2000.10474263.
17. White IR, Thompson SG. Adjusting for partially missing baseline measurements in randomized trials. Stat Med. 2005;24(7):993-1007. doi:10.1002/sim.1981.

18. 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. doi:10.1093/oxfordjournals.aje.a117592.

19. unvan der Heijden GJ, Donders AR, Stijnen T, Moons KG. 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. doi:10.1016/j.jclinepi.2006.01.015.

20. Knol MJ, Janssen KJM, Donders ART, Egberts ACG, Heerdink ER, Grobbee DE 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. doi:http://dx.doi.org/10.1016/j.jclinepi.2009.08.028.

21. Groenwold RHH, White IR, Donders ART, Carpenter JR, Altman DG, Moons KGM. Missing covariate data in clinical research: when and when not to use the missing-indicator method for analysis. CMAJ. 2012;184(11):1265-9. doi:10.1503/cmaj.

22. 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. doi:10.1002/sim.3944.

23. Mohan K, Pearl J, Tian J. Graphical models for inference with missing data. Advances in Neural Information Processing System. 2013;26(1277-1285).

24. Thoemmes F, Mohan K. Graphical Representation of Missing Data Problems. Structural Equation Modeling: A Multidisciplinary Journal. 2015;22(4):631-42. doi:10.1080/10705511.2014.937378.

25. Azur MJ, Stuart EA, Frangakis C, Leaf PJ. Multiple imputation by chained equations: what is it and how does it work? International Journal of Methods in Psychiatric Research. 2011;20(1):40-9. doi:10.1002/mpr.329.

26. Buuren Sv, Groothuis-Oudshoorn K. mice: Multivariate imputation by chained equations in R. Journal of statistical software. 2010:1-68.

27. Kleinke K. Multiple Imputation Under Violated Distributional Assumptions: A Systematic Evaluation of the Assumed Robustness of Predictive Mean Matching. Journal of Educational and Behavioral Statistics. 2017;42(4):371-404. doi:10.3102/1076998616687084.

28. van Buuren S. Multiple imputation of discrete and continuous data by fully conditional specification. Statistical Methods in Medical Research. 2007;16(3):219-42. doi:10.1177/0962280206074463.

29. Tilling K, Williamson EJ, Spratt M, Sterne JAC, Carpenter JR. Appropriate inclusion of interactions was needed to avoid bias in multiple imputation. Journal of Clinical Epidemiology. 2016;80:107-15. doi:10.1016/j.jclinepi.2016.07.004.

30. Penning de Vries B, Groenwold R. A comparison of approaches to implementing propensity score methods following muliple imputation. Epidemiology Biostatistics and Public Health. 2017;14(4). doi:10.2427/12630.

31. 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. doi:10.1177/0962280212445945.

32. Leyrat C, Seaman SR, White IR, Douglas I, Smeeth L, Kim J et al. Propensity score analysis with partially observed covariates: How should multiple imputation be used? Stat Methods Med Res. 2017:962280217713032. doi:10.1177/0962280217713032.

33. Penning de Vries B, Groenwold R. Comments on propensity score matching following multiple imputation. Stat Methods Med Res. 2016;25(6):3066-8. doi:10.1177/0962280216674296.

34. Mattei A. Estimating and using propensity score in presence of missing background data: an application to assess the impact of childbearing on wellbeing. Statistical Methods and Applications. 2009;18(2):257-73. doi:10.1007/s10260-007-0086-0.

35. Moons KGM, Donders RART, Stijnen T, Harrell FE. Using the outcome for imputation of missing predictor values was preferred. Journal of Clinical Epidemiology. 2006;59(10):1092-101. doi:http://dx.doi.org/10.1016/j.jclinepi.2006.01.009.

36. 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. doi:10.1080/03610926.2012.700371.

37. Lunt M. Selecting an Appropriate Caliper Can Be Essential for Achieving Good Balance With Propensity Score Matching. American Journal of Epidemiology. 2014;179(2):226-35. doi:10.1093/aje/kwt212.

38. King G, Nielsen R. Why propensity scores should not be used for matching. Copy at http://j mp/1sexgVw Download Citation BibTex Tagged XML Download Paper. 2016;378.

39. Austin PC. Optimal caliper widths for propensity-score matching when estimating differences in means and differences in proportions in observational studies. Pharm Stat. 2011;10(2):150-61. doi:10.1002/pst.433. (Accessed date: 15th May 2018)

40. Hernan MA, Robins JM. Causal inference. CRC Boca Raton, FL:; 2010.

41. Meng X-L. Multiple-Imputation Inferences with Uncongenial Sources of Input. Statistical Science. 1994;9(4):538-58.

42. Quartagno M, Carpenter JR. Multiple imputation for IPD meta-analysis: allowing for heterogeneity and studies with missing covariates. Statistics in Medicine. 2016;35(17):2938-54. doi:10.1002/sim.6837.

43. Malla L, Perera-Salazar R, McFadden E, Ogero M, Stepniewska K, English M. Handling missing data in propensity score estimation in comparative effectiveness evaluations: a systematic review. Journal of Comparative Effectiveness Research. 2018;7(3):271-9. doi:10.2217/cer-2017-0071.

Appendix

Results of Simulation setting 2 where the multiple imputation by chained equations (MICE) with Bayesian linear regression is used for a sensitivity analysis.

	Heterogeneous treatment effect						
	Propensity matching				Propensity weighting		
	coefficient		Bias	MSE	coefficient		MSE
	mean	sd			mean	sd	
No adjustment	1.730	0.158	1.410	2.012	1.730	0.158	3.019
After adjustment	0.321	0.096	0.000	0.009	-0.017	0.156	0.025
No interaction term							
Multiple imputation							
with Y	0.304	0.095	-0.017	0.009	-0.041	0.170	0.031
without Y	0.536	0.101	0.215	0.056	0.292	0.142	0.105
Combined method							
with Y	0.303	0.095	-0.018	0.009	-0.042	0.172	0.031
without Y	0.537	0.104	0.216	0.058	0.294	0.143	0.107
Interaction terms							
Multiple imputation	0.315	0.094	-0.006	0.009	-0.014	0.169	0.029
Combined method	0.315	0.096	-0.006	0.009	-0.015	0.171	0.029
No interaction term							
Multiple imputation							
with Y	0.220	0.103	-0.101	0.021	-0.116	0.192	0.050
without Y	0.568	0.110	0.247	0.073	0.264	0.158	0.095
Combined method							
with Y	0.220	0.101		0.010	-0.116	0.190	0.049
without Y	0.568	0.111	0.248	0.074	0.264	0.157	0.094
Interaction terms							
Multiple imputation	0.330	0.101	0.009	0.010	0.002	0.199	0.040
Combined method	0.331	0.103	0.010	0.011	0.001	0.198	0.039
No interaction term							
Multiple imputation							
with Y	0.102	0.110	-0.219	0.060	-0.269	0.213	0.118
without Y	0.570	0.110	0.249	0.074	0.325	0.153	0.129
Combined method							
with Y	0.095	0.103	-0.225	0.061	-0.275	0.211	0.120
without Y	0.537	0.105	0.216	0.058	0.233	0.149	0.076
Interaction terms							
Multiple imputation	0.173	0.101	-0.147	0.032	-0.197	0.220	0.087
Combined method	0.169	0.103	-0.151	0.034	-0.206	0.215	0.089

Chapter 3

Comparing methods for measurement error detection in serial 24-hour hormonal data

Published in J Biol Rhythms. 2019 Aug;34(4): 347-363

Evie van der Spoel*, Jungyeon Choi*, Ferdinand Roelfsema, Saskia le Cessie , Diana van Heemst, Olaf M. Dekkers

*Contributed equally to this work

Abstract

Measurement errors commonly occur in 24-hour hormonal data and may affect the outcomes of such studies. Measurement errors often appear as outliers in such datasets; however, no well-established method is yet available for their automatic detection.

In this study, we aimed to compare the performances of different methods for outlier detection in hormonal serial data. Hormones (glucose, insulin, thyroid stimulating hormone (TSH), cortisol, and growth hormone (GH)) were measured in blood sampled every 10 minutes for 24 hours in 38 participants of the Leiden Longevity Study. Four methods for detecting outliers were compared: i) eyeballing, ii) Tukey's fences, iii) Stepwise approach, and iv) the Expectation-Maximization (EM) algorithm. Eyeballing detects outliers based on experts' knowledge, and Stepwise approach incorporates physiological knowledge with a statistical algorithm. Tukey's fences and the EM algorithm are data-driven methods, using interquartile range and a mathematical algorithm to identify underlying distribution, respectively. The performance of the methods was evaluated based on the number of outliers detected and the change in statistical outcomes after removing detected outliers. Eyeballing resulted in the lowest number of outliers detected (1.0% of all data points), followed by Tukey's fences (2.3%), Stepwise approach (2.7%), and the EM algorithm (11.0%). In all methods, the mean hormone levels did not materially change after removing outliers. However, their minima were affected by outlier removal. Although removing outliers affected the correlation between glucose and insulin on the individual level, when averaged over all participants, none of the four methods influenced the correlation.

Based on our results, the EM algorithm is not recommended given the high number of outliers detected, even where data points are physiologically plausible. Since Tukey's fences is not suitable for all types of data, and eyeballing is time-consuming, we recommend Stepwise approach for outlier detection which combines physiological knowledge and an automated process.

1. Introduction

Many physiological parameters such as hormones or metabolites exhibit rhythmicity. These rhythms are regulated by different systems. The most prominent rhythm is the circadian rhythm, which is induced by the biological clock located in the suprachiasmatic nucleus of the brain. The biological clock does not only synchronize molecular clocks in peripheral cells, but it also orchestrates many physiological functions, including blood pressure, core body temperature, and hormone secretion. An example of a hormone that exhibits strong circadian rhythmicity is cortisol. The sleep-wake cycle is another form of rhythm, and although similar to the circadian rhythm, it has other effects on hormone secretion than the biological clock. The secretion of growth hormone, for example, is more strongly influenced by sleep than by clock time. External cues, including food intake and physical activity, also can influence hormone secretion, such as the secretion of insulin (Oike *et al*., 2014).

Hormones and metabolites are measured for different purposes; e.g., in clinical settings to make a diagnosis or to evaluate the effect of treatment and in research settings to investigate how these parameters change upon interventions or differ between groups. Different cues can elicit changes in hormone secretion, amongst which circadian time, nutrient availability and food intake, physical activity, and sleep. Circulating concentrations of many hormones change over time, because these hormones are secreted in a pulsatile fashion and have a relatively short half-life (Spiga *et al*., 2015). Therefore, to obtain reliable hormonal time series data, hormones need to be measured in blood that is sampled frequently. For some hormones, such as insulin, the preferred sampling frequency is 2 minutes because of its short half-life (Porksen *et al*., 1997). Other hormones, including thyroid stimulating hormone (TSH), can be measured every 20 minutes to obtain reliable profiles (Odell et al., 1967; Grossmann et al., 1997). To take into account practical possibilities, half-lives, costs, and ethics, most studies investigating hormone secretion are performed with a sampling frequency of every 10 minutes during 24 hours, as reviewed by Veldhuis *et al.* and Roelfsema *et al.* (Veldhuis *et al*., 2016; Roelfsema *et al*., 2017).

When measuring hormones frequently over time, measurement errors are likely to occur. Measurement errors can be caused by pre-analytical experimental variation of various sources, including sample dilution (possibly because of flushing the intravenous line with heparinized saline), or the presence of a blood clot in the sample. Measurement error can influence the outcomes of studies with serial hormonal data. Therefore, it is important to identify measurement errors. Measurement errors are likely to be outliers (Grubbs, 1969), which deviate largely from the overall trend of the data. The challenge is that there is no clear-cut distinction between measurement errors and true biological variation. The starting point to detect measurement errors, however, is by identifying outliers.

No well-established method is yet available to automatically detect measurement errors. Therefore, we aimed to compare four methods to detect outliers likely due to measurement errors in 24-hour hormonal data: eyeballing (relying on experts' opinions), Tukey's fences (identifying outliers based on inter-quartile ranges), Stepwise approach (identifying outliers based on standard deviations), and the Expectation Maximization (EM) algorithm (using a mathematical algorithm based on disentangling the two different distributions of outliers and non-outliers). Furthermore, we studied the influence of removing the detected outliers on the assessment of statistical features of 24-hour hormonal data such as mean, minimum, maximum, and cross-correlation.

For this study, we used data on the pituitary hormones growth hormone (GH), adrenocorticotropic hormone (ACTH) and TSH, the adrenal hormone cortisol, as well as data on the metabolic signals insulin, and glucose, which were all measured during 24 hours every 10 minutes in serum from 38 participants of the Switchbox Leiden Study (Jansen *et al.*, 2015).

2. Methods

2.1. Data collection

Study population

The Leiden Longevity Study comprises 421 families with at least two long-lived Caucasian siblings fulfilling the age criteria (men ≥89 years and women ≥91 years) without selection on health or demographics (Westendorp *et al.*, 2009). In the current study, the Switchbox Leiden Study, we included 20 offspring of long-lived families from the Leiden Longevity Study together with 18 partners of the offspring as environmental and age-matched controls. The primary aim of the Switchbox Leiden Study was to compare the levels and dynamics of hormones and metabolites and their interplay between offspring of long-lived families and controls. In- and exclusion criteria were described previously in detail (Jansen *et al.*, 2015). Participants were middle-aged (52–76 years) and had a stable body mass index (BMI) between 18 and 34 kg/m^2. The Switchbox Leiden Study was approved by the Medical Ethical Committee of the Leiden University Medical Centre and was performed according to the Helsinki declaration. All participants gave written informed consent for participation.

24-hour blood sampling

The 24-hour blood sampling procedure started with placing a catheter in a vein of the forearm of the non-dominant hand, and blood withdrawal started around 9:00h (Akintola *et al.*, 2015). Samples of 2 ml serum and 1.2 ml EDTA plasma were withdrawn every 10 min. To prevent blood clotting, heparinized saline (0.9% NaCl)

was continuously infused via an infusion pump at a rate of 20 ml per hour. Before each blood withdrawal, 5 ml of saline/heparin mixed with blood was collected (without disconnecting the syringe from the blood withdrawal system) to prevent contamination of heparin/saline in the blood samples. After blood withdrawal, this 5 ml was flushed back into the subject to reduce the total amount of blood that would be withdrawn. Participants received standardized feeding consisting of 600 kcal Nutridrink (Nutricia Advanced Medical Nutrition Zoetermeer, The Netherlands) at three fixed times during the day. Participants were not allowed to sleep during the day, and except for lavatory use, no physical activity was allowed during the study period. Lights were switched off for approximately 9 hours (circa between 23:00h to 08:00h) to allow the participants to sleep.

Assays

All laboratory assays were performed with fully automated equipment and diagnostics from Roche Diagnostics (Almere, The Netherlands) at the Department of Clinical Chemistry and Laboratory Medicine of the Leiden University Medical Centre in The Netherlands.

Thyroid-stimulating hormone (TSH), cortisol, insulin, and glucose were measured in the same serum tube. Growth hormone (GH) was also measured in the same serum tube but after one additional freeze/thaw cycle. TSH and cortisol were measured by ElectroChemoLuminescence ImmunoAssay (ECLIA) using a Modular E170 Immunoanalyzer from Roche (Roche Diagnostics, Almere, The Netherlands). For TSH, the overall interassay coefficients of variation (CV) ranged in our study between 1.41–4.16%, and the overall CV of cortisol ranged between 2.4–5.1%. Human GH with a molecular mass of 22 kDa and insulin were measured using an IMMULITE® 2000 Xpi Immunoassay system (Siemens Healthcare diagnostics). The interassay CV of GH ranged between 5.4% at 5.43 mU L^{-1} and 7.2% at 25.0 mU $L^{-1,}$ and the overall CV of insulin ranged between 3.19–7.69%. Glucose was measured using Hitachi Modular P800 from Roche Diagnostics (Almere, the Netherlands), and the overall interassay CV of glucose ranged between 0.90–7.44%. If a measurement was below the detection limit, half of the lower detection limit was taken as a result.

Although ACTH was also measured, we did not take along these data in our mathematical models because this hormone was measured in EDTA plasma, so in another tube than the other hormones. However, we used ACTH data for the eyeballing, because they were instrumental for inspecting physiologically abnormal points in the cortisol data.

2.2. Physiological considerations

Since hormones are secreted in a pulsatile manner, a sudden increase is more likely to occur than a sudden decrease. Also, glucose < 2.8 mmol/L does not occur in healthy persons without an accompanying strong stress response (cortisol and GH pulses).

ACTH stimulates the secretion of cortisol. Therefore, cortisol should show a pulse following an (extreme) increase in ACTH. If an outlier is caused by sample dilution, then all hormones measured in that sample should be lower than expected. These physiological considerations could be taken into account in measurement error detection.

2.3. Methods of detecting outliers

In the following section, we will discuss four methods for outlier detection: i) eyeballing, ii) Tukey's fences, iii) Stepwise approach, and iv) the EM algorithm. The procedures of these methods are visualized in Figure 1.

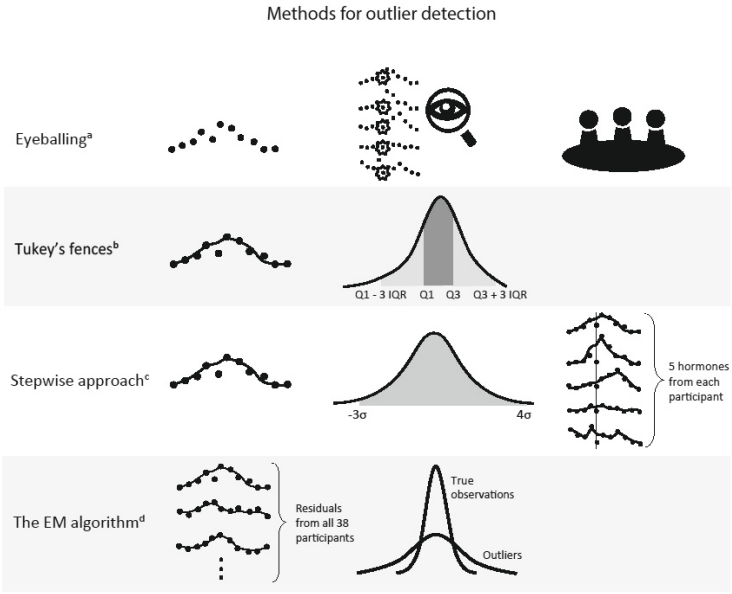


Figure 1. (a) Eyeballing detects outliers without fitting smooth curves. By visual inspection, individual experts detect outliers by taking into account that some hormones were measured in the same sample. Afterward, a consensus meeting is held, and the experts discuss all data points with conflicting detection results. (b) Tukey's fences starts with fitting a moving average curve to per-person per-hormone data and taking residuals of all data points. Then the interquartile range (IQR = Q3–Q1) of the residuals is calculated. The data points lying outside the range between Q1 − 31QR and Q3 + 3IQR are detected as outliers. (c) The stepwise approach fits the moving average curve to per-person per-hormone data, and standardized residuals of all data points are calculated (step 1). The data points lying outside the range between −3 and 4 standard deviations are detected as outliers (step 2). Then, the residuals of 5 hormones measured at the same time points are summed. When the sum of the residuals is smaller than −8, the data points are detected as outliers (step 3). Afterward, steps 1 and 3 are repeated (step 4). (d) The expectation-maximization (EM) algorithm first fits a smoothing curve to per-person per-hormone data, and the residuals are calculated. Then, all the residuals of a hormone from all 38 participants are put in the EM algorithm. The algorithm then identifies two distinguishable distributions and yields the probability of each data point being an outlier.

Eyeballing

Eyeballing was based on a visual inspection of a graphical display of individual hormone profiles from all 38 patients. This was performed by four reviewers with expert knowledge in endocrinology (EvdS, FR, OMD, and DvH). Hard copies of the 24-hour trajectories of all hormones measured per participant were provided. Three reviewers examined all 38 participants' hormone profiles, and one reviewer checked half of the participants. Information about which hormones were measured in the same tube was given verbally. Reviewers were also explicitly told that dilution of the sample may have led to measurement errors in all hormones from the same tube. After reviewing the data separately, a consensus meeting was held to reach an agreement on data points which only one (out of three or four) or two out of four reviewers had marked as an outlier.

Tukey's fences

For this algorithmic approach of outlier detection, we made the following assumptions: i) A hormone trajectory of a person follows a smooth general trend over 24 hours while measurement errors may deviate clearly from the trend, and ii) Hormone levels cannot abruptly decrease within 10 minutes. If a measurement is vastly distant from the adjacent measurements before and after, that measurement is likely to be a measurement error. Thus, by fitting a smooth curve to the data points and measuring the distance between the curve and each measurement, the algorithm can detect outliers expected to be measurement errors.

Tukey's fences is a non-parametric method developed to detect observations out of the normal range by using interquartile ranges (Tukey, 1977), and it is often used for detecting outliers in various fields (Muraleedharan *et al.*, 2016; Pham and Eggleston, 2016; Luo *et al.*, 2018; O'Brien *et al.*, 2018). Before performing Tukey's fences, the normality of the data was checked before fitting the curve. The distributions of insulin and GH data were highly skewed. Therefore, these data were log-transformed prior to applying the algorithm. Afterward, Tukey's fences was implemented using the following two steps:

I. Hormone data were smoothened over time by fitting moving average curves for every hormone per-person separately. Moving average is a method commonly applied for smoothing time series data (Montgomery *et al.*, 2015). The moving average with window size n (with n an odd number) at a certain time point is the average of the current, the -½(n-1) previous, and ½(n-1) subsequent measurements in time. In our analyses, moving averages were calculated using a window of five points. Residuals were calculated for all data points. We defined a residual as the vertical distance between an original data point and a fitted moving average curve.

II. between the first quartile and the third quartile ($Q_1 - Q_3$), and the median (Q_2) were identified. The ranges between Q_2 $k(Q_3 Q_1)$ and $Q_2 + k(Q_3 - Q_1)$ are referred to as fences. The data points that are below the lower fence or above the higher fence are identified as outliers. The value k determines the width of the fences. The larger the value of k, the lower the number of outliers that will be detected. In our analyses, we set k=3, which according to the literature, implies that the data point is "far out" (Tukey, 1977). To use the method as it was originally suggested and commonly applied, we did not adjust the value of k=3 (Horn et al., 1988; Hung and Yang, 2006; Kimenai et al., 2016).

Stepwise approach

Stepwise approach is an automatic detection process based on an algorithm that incorporates physiological knowledge and statistical methods comprising three steps as described below. We aim to detect potential outliers within a 24-hour hormone trajectory in several steps. As in Tukey's fences, the insulin and GH data were log-transformed.

I. Step 1: Fitting smoothed curves
Likewise to Tukey's fences, a moving average curve is fitted to each participant's 24-hour hormone data using a window of 5 points. By computing the distance between each data point and the fitted curve, residuals are acquired. The residuals are standardized to have a mean of 0 and a standard deviation of 1.

II. Step 2: Detecting outliers within a 24-hour hormone trajectory
Data points with standardized residuals smaller than -3 or larger than 4 are detected as outliers. The cut-off of 3 standard deviations is a commonly applied empirical rule for detecting outliers in normally distributed data. However, asymmetrical cut-offs are chosen to be more liberal for the upper boundary, as hormones are secreted in a pulsatile fashion which makes rapid increases in hormone levels biologically more plausible than rapid decreases since clearance of the hormone will occur slower. Note that this cut-off boundary is wider than the width of Tukey's fences with k=3. Furthermore, data points where glucose < 2.8 mmol/L were detected as outliers as discussed under *Physiological considerations*.

III. Step 3: The standardized residuals of all hormones measured in the same serum tube are added up for each participant. If the sum of the standardized residuals is lower than -8, all data points measured in that tube are detected as outliers. This means that the residuals of the five hormones are, on average, below the 5th percentile of standard normal distribution (1.64 standard deviation). This step allows detecting measurement errors due to the dilution of the samples. The underlying assumption is that when samples were diluted, levels of the hormones measured in the same sample are likely to all be lower at the same time point. In this step, we aim to detect these types of measurement errors which occur across the hormones.

IV. Step 4: Repeat step 1 and step 3
After all outliers detected so far are removed, a new moving average curve is fitted and step 1 and 3 are repeated once. If already detected outliers are removed, the newly fitted curves will be flatter than the fitted curve from the original data, which will allow detecting potential outliers that were missed in the previous steps.

The EM algorithm

Another approach is to estimate the probability for a data point to reflect measurement error, rather than using a dichotomous division. This starts with assuming two distinguishable data distributions: true measurement variation and background noise due to measurement errors. Based on this assumption, we expect the residuals of the true measurements to be normally distributed with standard deviations close to 0, while those of the erroneous measurements would be normally distributed with a larger standard deviation. The expectation maximization (EM) algorithm is a method that can be used to identify these two distinguishable distributions. The algorithm estimates model parameters when data is incomplete or when the model depends on a latent variable; a variable that is not directly observed but can be inferred by other observed variables (Dempster *et al*., 1977), and the method was suggested for detecting outliers (Aitkin and Wilson, 1980). The EM algorithm was applied in R version 3.5.1, using the normalmixEM function of the package mixtools (Benaglia *et al.,* 2009). In our situation, the latent variable of interest would be whether a data point is a true measurement or a measurement error. Further technical details about the EM algorithm can be found in Supplementary Material, Appendix 1.

The EM algorithm has the advantage that detected outliers do not have to be removed. Instead, the probabilities can later be used as weights for estimating outcomes, such as mean hormone levels or cross-correlations.

The outlier detection method using the EM algorithm followed the steps below. Again, insulin and GH data were log-transformed.

I. As in Tukey's fences and Stepwise approach, a moving average curve per 24-hour hormone profile for each participant was fitted. Afterward, residuals were calculated and standardized for each data point.
II. The EM algorithm was applied for each hormone with residuals of all participants together taken into account in one model.

2.4. Comparing methods on statistical outcomes

Since we do not know with certainty which data points reflect measurement errors, it is not possible to ascertain which of the four methods performed best. Therefore, we compared the number of outliers detected which were counted *per time point* and in *total data points*. In addition, the overlap in detected outliers between the four methods

was visually presented with Venn diagrams (Larsson, 2018). We chose these parameters since these descriptive statistics give a transparent description of the data and will give an insight into how removing outliers have an impact on general measures.

Furthermore, we analyzed statistical outcomes of 24-hour hormonal data before and after removing the outliers as detected by the four different methods. In this way, we could investigate whether removing outliers influenced the statistical outcome and how different methods may do so differently. Therefore, the 24-hour means, median, minima, and maxima of the five hormones were assessed, which provides a transparent description of the data and insights into how removing outliers impacts general measures. Another relevant analysis is the cross-correlation between two hormones. Cross-correlation estimates the temporal relationship between two hormonal concentrations. It is a common analysis performed with data from two simultaneously measured hormonal time series (Vis *et al.,* 2014). Therefore, it could be of interest for researchers to know to which extent measurement error would affect the estimates, especially since this method might be sensitive to the presence of outliers that co-occur in different time series data, for example, due to the dilution of a sample. Two relevant outcome measures are the strongest correlation coefficient (the maximal correlation) and the correlation coefficient at lag time 0. For the purpose of this paper, we performed cross-correlation on concentrations of glucose and insulin, which are expected to display strong cross-correlation (Feneberg *et al.*, 1999). When estimating the mean and cross-correlations after outlier removal by the EM algorithm, the weighted mean and weighed correlation are calculated, with the weight equal to the probability of each data point being an outlier. All statistical analyses were performed using the software program R, version 3.5.1.

3. Results

For each of the 38 participants, blood samples were collected at 144 time points over 24 hours, with five hormones being measured in the same serum tube. After discarding missing data, the total number of data points was 21,467. We counted detected outliers *per time point* and in *total data points*. If counted *per time point*, at least one outlier was detected in a time point among all hormones assayed in serum (i.e., glucose, insulin, TSH, cortisol, and growth hormone). In the case of a complete series, a single participant has 144 time points for each hormone. If counted in *total data points*, every data point is counted individually. In the case of a complete dataset, one participant has in total 720 data points, that is, 144 time points times five hormones.

3.1. Number of detected outliers

Table 1 summarizes the mean percentage of outliers detected per time point and in total data points. The results are averaged across 38 participants. Since the EM algorithm yields continuous probability as its outcome, we defined a data point in which its probability of being an outlier is higher than 0.9 as an outlier. For the percentage of detected outliers, we observed some differences between the four methods. Eyeballing resulted in the smallest percentage of detected outliers both per time point (mean=1.7%) as well as for total data points (1.0%), followed by Stepwise approach (per time points: 5.1%, total data points: 2.7%). Tukey's fences yielded more outliers per time point (9.3%) but a similar percentage in total data points (2.3%). The EM algorithm method yielded the largest percentage of outliers (per time points: 40.3%, total data points: 11.0%).

In Figure 2, the numbers of detected outliers for each hormone averaged over all participants are presented. The EM algorithm and Tukey's fences both detected more outliers in cortisol and GH compared to other hormones. Eyeballing and Stepwise approaches detected a similar number of outliers across the different hormones.

Table 1. The percentage of time points with at least one detected outlier among the hormones measured, and the percentage of total data points detected as outliers among the same set of hormones. The mean and standard deviation of the 38 participants are given.

	mean (sd); n=38	
	Time points detected to contain an outlier (%)	**Total data points detected to be outliers (%)**
Eyeballing	1.7 (2.1)	1.0 (1.4)
Tukey's fences	9.3 (5.6)	2.3 (1.4)
Stepwise approach	5.1 (1.5)	2.7 (1.5)
EM algorithm*	40.3 (7.7)	11.0 (2.8)

*For the EM algorithm results, the measurement points where the probability of being an outlier > 0.9 was counted.

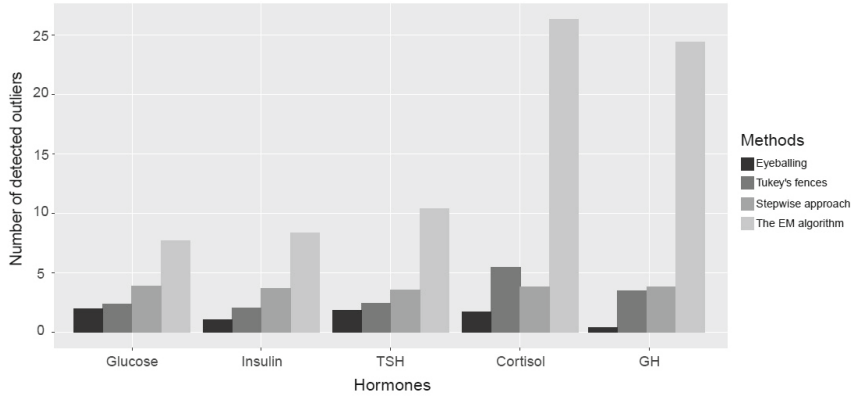


Figure 2. The mean number of data points detected per hormone per method across all participants.

3.2. Overlap in detected outliers

Figure 3 displays Venn diagrams presenting the number of outliers detected by eyeballing, Stepwise approach, and Tukey's fences and their overlap. We did not include the results of the EM algorithm in the Venn diagrams for two reasons i) the EM algorithm detected an implausibly large number of outliers (per time point=1,590 and in total data points =2,728), and (ii) three sets of data is the maximum to draw a proportional Venn diagram in two-dimensional space. Figure 3a presents the number of outliers per time point, and Figure 3b presents that of the total data points. In Figure 3a, most of the outliers detected by eyeballing were also detected by the other two methods, while the overlap is larger with Stepwise approach. In Figure 3b, the overlap between eyeballing and Stepwise approach is again larger than the overlap between eyeballing and Tukey's fences. Here, Stepwise approach and Tukey's fences detected a similar number of outliers. However, the overlap is relatively small, which indicates that they are detecting different data points. Eyeballing detected 47 total data points, which were not detected by Stepwise approach or Tukey's fences. Among outliers per time point detected by eyeballing, Stepwise approach, and Tukey's fences, 95.8% overlapped with the outliers detected by the EM algorithm (data not shown). Additionally, 70.1% of the total data points detected by the three methods overlapped with the outliers detected by the EM algorithm (data not shown).

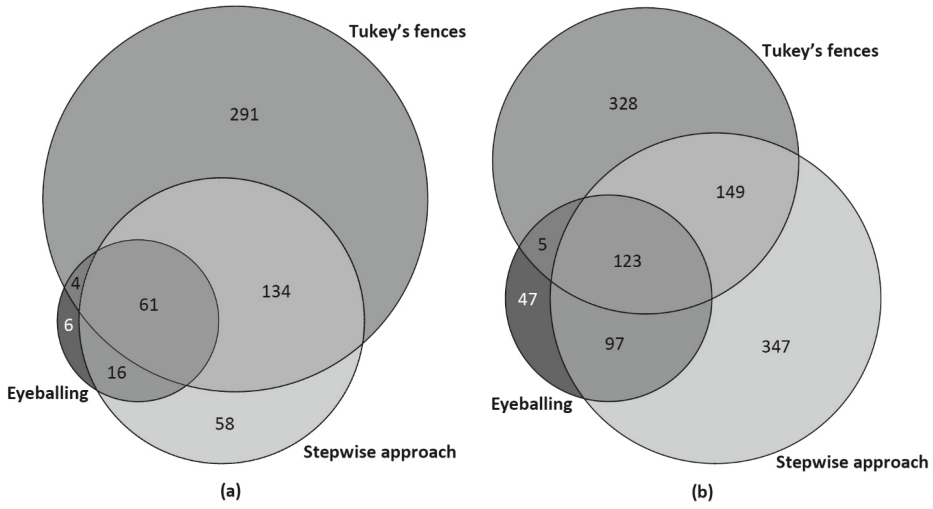


Figure 3. Venn diagrams visualizing the number of measurement errors detected by each method (eyeballing, Stepwise approach, and Tukey's fences) and their overlap counted in total time points (a) and in all data points (b). The overlap with the EM algorithm is not presented here for the reasons mentioned in the results section.

3.3. Representative 24-hour hormone figures presented with detected outliers

Figures 4a-d display the detected outliers in glucose, insulin, TSH, cortisol, and GH for eyeballing, Tukey's fences, Stepwise approach, and the EM algorithm, respectively in one representative participant. By eyeballing (Figure 4a), four data points are detected as outliers in glucose, TSH, and cortisol, and these four outliers are all in the same time points. Of these four time points, outliers in insulin were detected in three time points and GH in one time point. Tukey's fences (Figure 4b) detected the same outliers for glucose, insulin, TSH, and cortisol but detected several more than eyeballing. In both TSH and cortisol between time points 110 to 130, several points that are biologically unlikely to be measurement errors were detected. No outliers were detected in GH. Stepwise approach (Figure 4c) identified the same outliers as eyeballing. However, several extra points were detected as well. Here in several time points (42nd, 76th, and 114th), outliers were detected in all hormones, which is a result of Step 3 of the Stepwise approach. The EM algorithm (Figure 4d, note that the points are only marked if the probability of being an outlier is higher than 0.9) resulting in many detected outliers in the pulses that are unlikely to be outliers. Remarkably in GH, data points close to detection limits were detected as outliers.

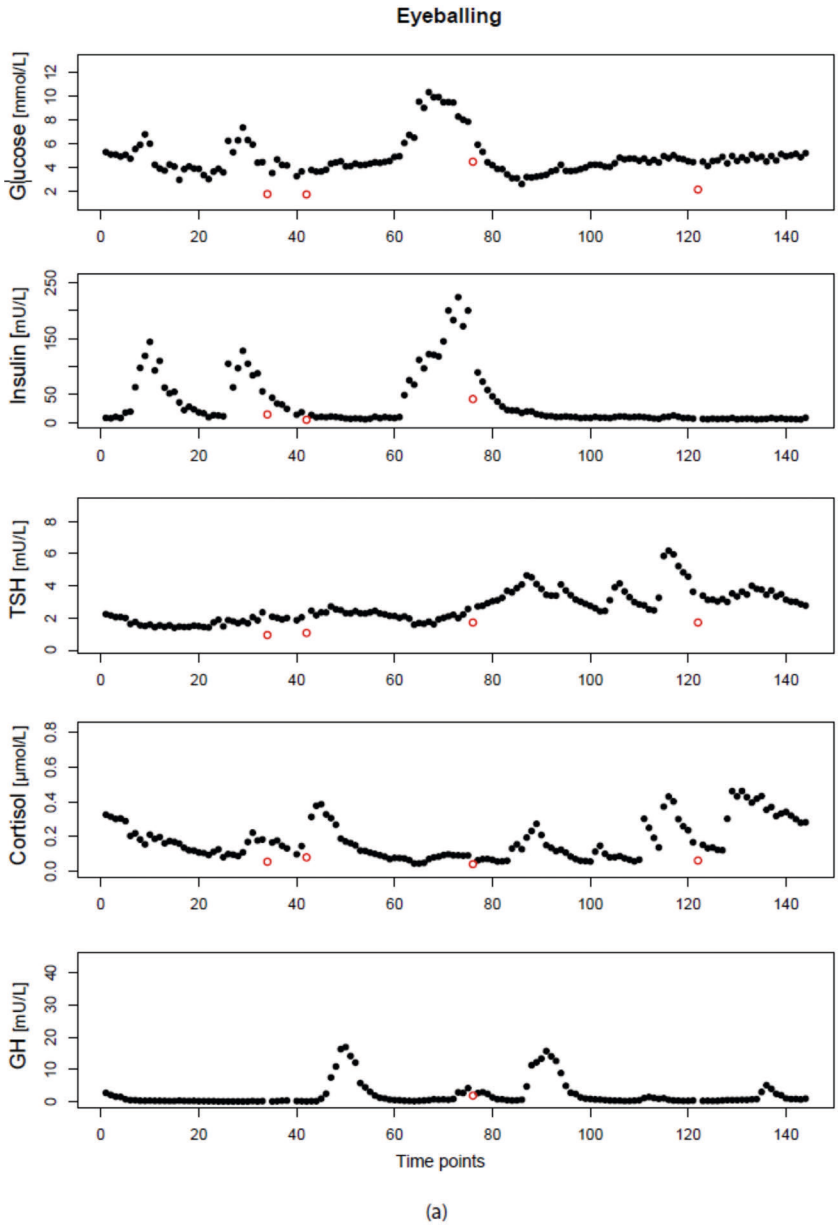


Figure 4a. The results of outlier detection by eyeballing in glucose, insulin, TSH, cortisol, and growth hormone of participant 19. Red hollow data points indicate detected outliers.

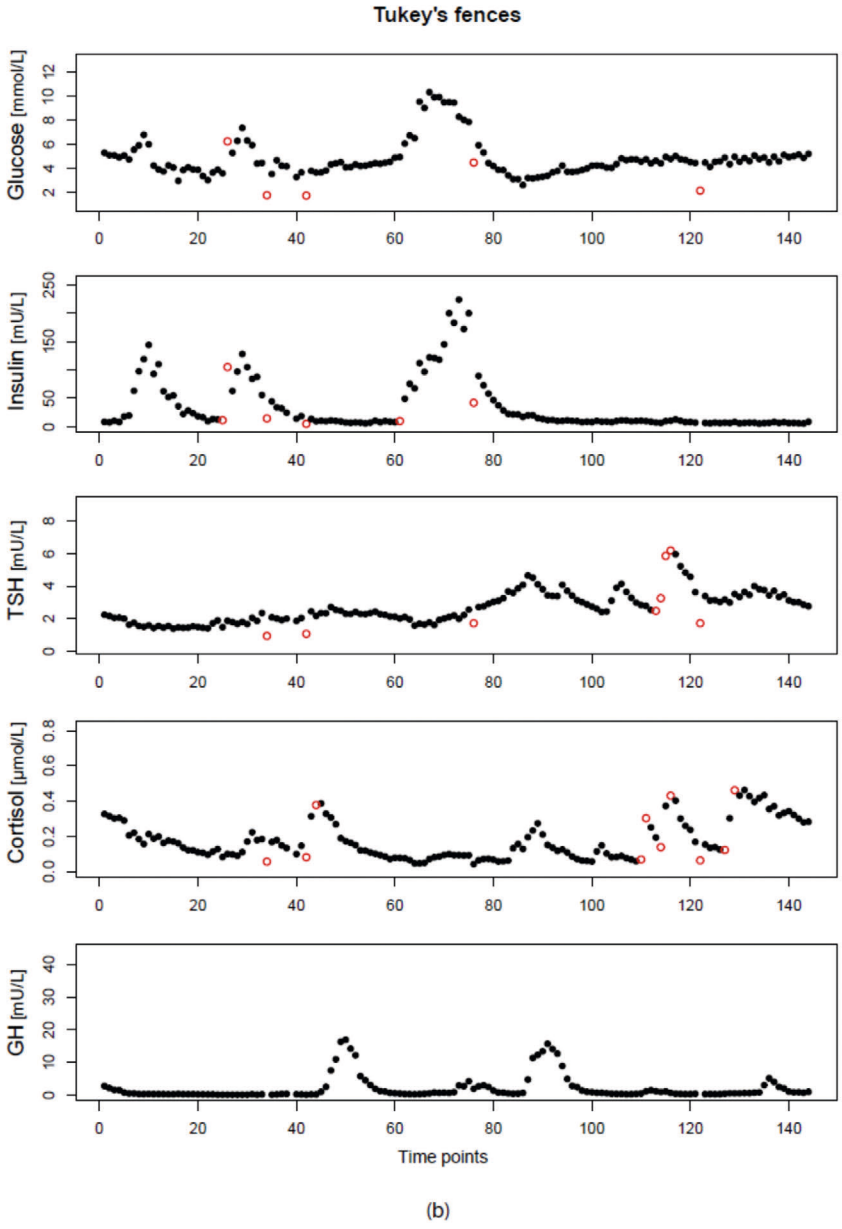


Figure 4b. The results of outlier detection by Tukey's fences in glucose, insulin, TSH, cortisol, and growth hormone of participant 19. Red hollow data points indicate detected outliers.

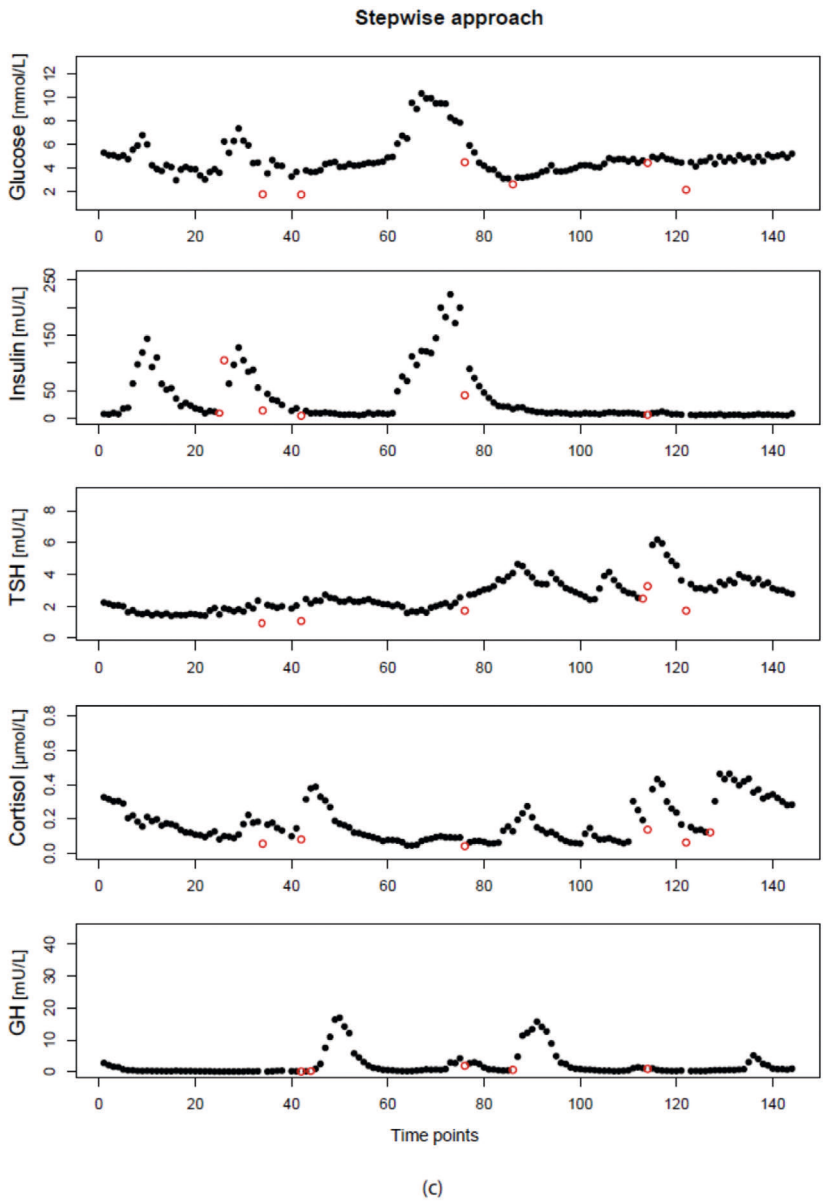


Figure 4c. The results of outlier detection by Stepwise approach in glucose, insulin, TSH, cortisol, and growth hormone of participant 19. Red hollow data points indicate detected outliers.

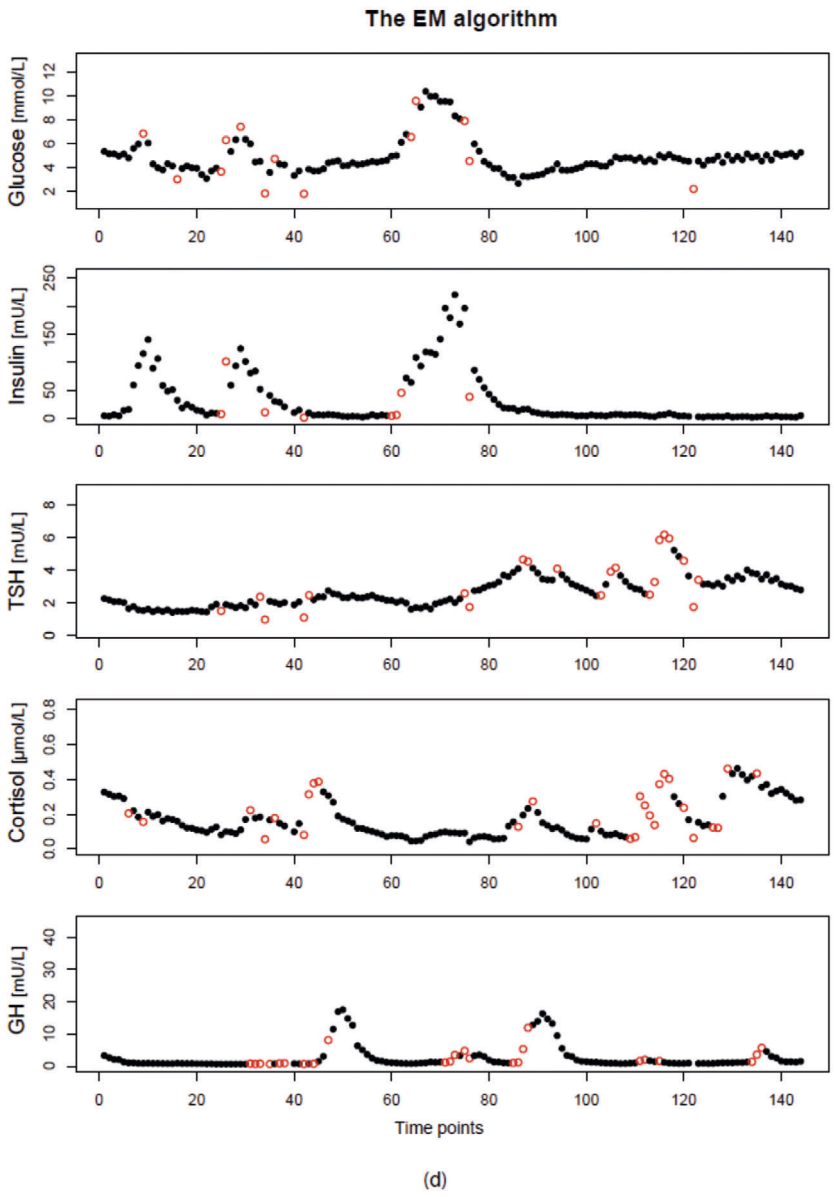


Figure 4d. The results of outlier detection by the EM algorithm in glucose, insulin, TSH, cortisol, and growth hormone of participant 19. Red hollow data points indicate the probability of the data point being an outlier is higher than 0.9.

3.4. Effects of removing outliers on statistical outcomes

Descriptive statistics: 24-hour mean, median, minimum, and maximum

The mean, median, minimum, and maximum values for every hormone were calculated over time before and after removing outliers detected by the four methods. This is shown in Table 2. Mean and median values did not change substantially after outlier removal. Minimum values changed for glucose and TSH after removing outliers by all four methods, while in insulin, the value did not change much after eyeballing. The EM algorithm had the largest influence on maximum values in all hormones.

Cross-correlation of glucose and insulin

In Table 3 cross-correlations between glucose and insulin are presented before and after removing outliers. Overall, removing outliers did not have a major influence on the cross-correlation of glucose and insulin, and on the lag time at the maximum cross-correlation. Figure 5 shows the individual changes in correlation at lag time 0. In Figure 5, we observe large differences between participants. Especially the first participant shows a big change in correlation after removing outliers by all methods. Overall, the changes after eyeballing, Tukey's fences, and Stepwise approach were mostly small, and the changes were not in one direction dominantly. However, after removing outliers detected by the EM algorithm, cross-correlation decreased in most cases.

Table 2. Mean, median, minimum, and maximum values for glucose, insulin, TSH, cortisol, and growth hormone in 24 hours, before (raw data) and after outlier removal (Eyeballing, Tukey's fences, Stepwise approach, and the EM algorithm). The mean and standard deviation of the 38 participants are given.

	mean (sd); n=38											
	Glucose [mmol/L]				Insulin [mU/L]				TSH [mU/L]			
	Mean	**Median**	**Min**	**Max**	**Mean**	**Median**	**Min**	**Max**	**Mean**	**Median**	**Min**	**Max**
Raw data	5.09 (.36)	4.80 (.39)	2.76 (.70)	9.51 (1.52)	19.90 (10.11)	9.66 (5.51)	2.76 (2.41)	91.61 (54.41)	2.02 (1.05)	1.92 (1.01)	1.01 (.61)	3.57 (1.89)
Eyeballing	5.11 (.36)	4.81 (.39)	3.16 (.53)	9.48 (1.47)	19.96 (10.14)	9.66 (5.52)	2.80 (2.46)	91.61 (54.41)	2.03 (1.05)	1.93 (1.02)	1.21 (.67)	3.57 (1.89)
Tukey's fences	5.07 (.37)	4.80 (.39)	3.04 (.62)	9.21 (1.42)	19.96 (10.16)	9.69 (5.52)	3.39 (2.39)	91.34 (54.69)	2.02 (1.04)	1.93 (1.02)	1.19 (.63)	3.49 (1.82)
Stepwise approach	5.12 (.36)	4.80 (.39)	3.29 (.41)	9.40 (1.48)	20.47 (10.43)	10.21 (5.86)	3.54 (2.31)	91.03 (54.26)	2.02 (1.05)	1.92 (1.01)	1.21 (.64)	3.50 (1.81)
EM algorithm*	5.00 (.37)	4.77 (.40)	3.14 (.58)	9.08 (1.47)	20.05 (10.35)	10.16 (5.97)	3.74 (2.44)	87.65 (49.73)	1.98 (1.01)	1.90 (1.00)	1.24 (.73)	3.33 (1.64)
	Cortisol [µmol/L]				GH [mU/L]							
	Mean	**Median**	**Min**	**Max**	**Mean**	**Median**	**Min**	**Max**				
Raw data	.21 (.05)	.18 (.05)	.05 (.03)	.57 (.09)	2.49 (1.51)	.95 (.94)	.16 (.22)	20.63 (10.31)				
Eyeballing	.21 (.05)	.18 (.05)	.05 (.03)	.57 (.09)	2.48 (1.58)	.95 (.94)	.16 (.22)	20.63 (10.31)				
Tukey's fences	.20 (.05)	.18 (.05)	.05 (.03)	.55 (.09)	2.47 (1.55)	.95 (.94)	.17 (.22)	20.27 (10.67)				
Stepwise approach	.21 (.05)	.18 (.05)	.05 (.03)	.56 (.09)	2.51 (1.54)	.96 (.95)	.17 (.22)	20.27 (10.59)				
EM algorithm*	.18 (.04)	.16 (.05)	.05 (.03)	.50 (.08)	2.24 (1.48)	.94 (1.02)	.18 (.22)	18.90 (11.13)				

* For the EM algorithm results, weighted mean and standard deviation is used.

Table 3. Cross correlations between glucose and insulin. Mean and standard deviation across 38 participants.

	mean (sd); n=38		
	Correlation at lag time 0	**Maximum cross corr.**	**Lag time at maximum cross corr. (min)**
Raw data	0.74 (0.12)	0.74 (0.12)	-4.7 (7.3)
Eyeballing	0.74 (0.11)	0.75 (0.12)	-5.3 (7.6)
Tukey's fences	0.73 (0.14)	0.74 (0.14)	-6.3 (8.2)
Stepwise approach	0.74 (0.12)	0.75 (0.12)	-5.0 (8.0)
EM algorithm*	0.71 (0.12)	0.73 (0.17)	-9.5 (9.8)

*For the EM algorithm results, weighted correlation is used.

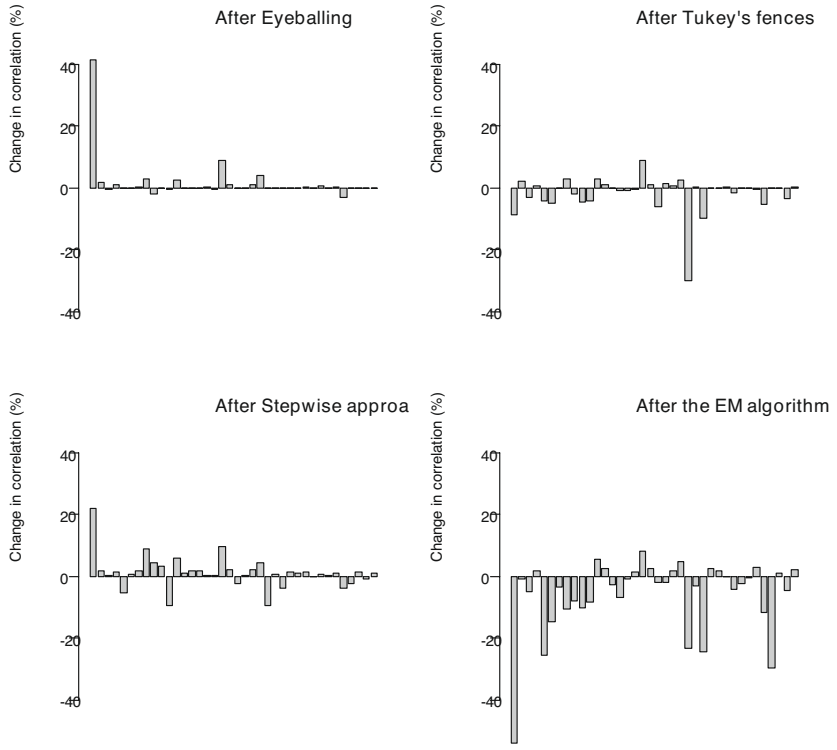


Figure 5. Change in correlation at lag time 0 (%) after removal of measurement errors detected by the four methods; eyeballing, Tukey's fences, Stepwise approach, and the EM algorithm. Each bar represents an individual participant.

4. Discussion

In this study, we aimed to evaluate and compare different methods to detect outliers in 24-hour hormonal data since no specific methods were routinely available for this purpose. We assumed that measurement errors would deviate largely from the physiological curves of hormones. By identifying outliers in the data, therefore, we expected to detect likely measurement errors. The main outcomes of this study were that human-judgement (eyeballing) defined fewer data points as an outlier than the other three automatic approaches. Among the automatic approaches, the data-driven methods (Tukey's fences and the EM algorithm) were prone to detect more outliers likely to be true measurements than the method involving subject-specific knowledge (Stepwise approach). The mean, minima, and maxima of the hormones did not change much after removing outliers. However, the minima of glucose and TSH did change, and the EM algorithm had a large influence on maximum values in all hormones. The effect of removing outliers on the correlation between glucose and insulin can be large within an individual but had no major impact on a group level.

A relatively low number of outliers were detected by eyeballing. This may be an advantage of this method, as only truly deviating points will be discarded in the analysis. Another advantage of eyeballing is that the data points detected as outliers are based on physiological arguments and are not data-driven. This allows eyeballing to detect (i) a sequence of data points that was physiologically implausible to display the same pattern in several hormones, and (ii) outliers at the beginning or end of a time series. These types of outliers cannot be detected by fitting smoothing curves, which explains the 47 data points that were exclusively detected by eyeballing, and not by Stepwise approach or Tukey's fences. However, a disadvantage of eyeballing is that it is time-consuming and depends on individual reviewers' background knowledge and subjective decision. If the number of reviewers is large enough and a consensus meeting is held, the precision may increase. However, the amount of time to reach a unanimous decision would take longer. Also, eyeballing is a one-off process that cannot be generalized to other settings.

Although Tukey's fences are advocated as a non-parametric approach, the method did not perform well in our case when applied with moving median curves instead of moving average curves. Especially when the hormone profile is mostly flat with sudden pulses, such as GH, Tukey's fences with moving median curves detected a biologically implausible number of outliers (54.6% of the total data points). Therefore, when using Tukey's fences to detect outliers, we suggest researchers to be aware of the type of their data and smoothing methods.

We introduced Stepwise approach as a new method to detect measurement errors in 24-hour hormonal data. The advantage of Stepwise approach is that by using the

standardized residuals, it facilitates the detecting of measurement errors caused by dilution, which may not have been identified by only looking into individual hormones. Additionally, it is expected to be a more objective method than eyeballing, as it explicitly incorporates the information from multiple hormones and applies the same cut-off values of standard deviations to every hormone. Furthermore, it is less time-consuming than eyeballing and can relatively easily be applied to different hormonal datasets. Compared to Tukey's fences, Stepwise approach has more flexibility to incorporate physiological knowledge, such as adopting asymmetrical cut-off or removing glucose measurements lower than 2.8 mmol/L. However, the performance of the method may depend on parameters such as a time window for moving average, or cut-off points of standard deviations. These parameters still require decisions and need to be chosen with care; the decisions should also be clearly reported. Another disadvantage of Stepwise approach, which also applies to eyeballing and Tukey's fences, is that we discard data according to a dichotomous division. Whether a data point is an outlier or not is often dependent on the degree of belief instead of a clear dichotomous distinction. Furthermore, this dichotomous distinction reduces the statistical power in further analyses.

The strength of the EM algorithm is that, instead of the dichotomous distinctions, it gives probabilities of each point being an outlier. Therefore, we acquire extra information which can be incorporated into further analysis, such as for probability weighting. Additionally, the EM algorithm requires less prior knowledge compared to the previously discussed methods. However, a critical disadvantage of the EM algorithm is that we cannot ensure whether the two identified distributions are actually distinguishing outliers and non-outliers. In our dataset, the detected points were often not plausible to be detected as outliers from a physiological perspective.

It is worth to mention the performances of Tukey's fences, Stepwise approach, and the EM algorithm depend on which smoothing technique is applied. Moving average, which was used in the study, does not require extensive modeling and can capture local fluctuations of hormone concentration. However, it may smooth out the transient increase of hormone concentration and lead to detect true measurements as outliers. Stepwise approach takes this shortcoming of moving the average into account by setting different cut-off values for positive and negative residuals. There are more advanced model-based smoothing techniques, such as deconvolution analysis, which takes the underlying dynamics of hormone secretions into account (Brown et al., 2001; Faghih et al., 2014). These methods were not considered in this study as our aim was to compare outlier detection methods that could be easily adopted by applied researchers in a pre-analysis phase.

To test the efficacy of the outlier detection methods, we simulated 24-hour hormonal data and measurement errors as comparable as possible to real data. The advantage of

the simulation study is that we know which data points are true measurement errors. We compared the performance of Stepwise approach, Tukey's fences, and the EM algorithm. The simulation description and the results are attached as an appendix (see Supplementary Material, Appendix 2). The EM algorithm resulted in a high percentage of true measurements wrongly detected as errors, especially when a simulated hormone has a higher variation during the day than during the night. Most methods yielded relatively low percentages of true error detected. This could be due to the fact that some simulated errors are close to fitted curves, while the methods we are comparing are based on detecting errors deviating from the curves. For detecting dilution errors, Stepwise approach performed better than other methods. This is because Stepwise approach could detect dilution errors that were not deviating much from the curves by taking the sum of the residuals from all hormones.

In this study, the effect of removing outliers on the cross-correlation between glucose and insulin had no major impact on a group level. Note that these results may not be generalized to other statistical outcomes, such as deconvolution analysis and approximate entropy analysis, which are also common analyses for 24-hour hormonal data. Furthermore, glucose and insulin are strongly cross-correlated; however, when two hormones are less strongly correlated, the impact of removing outliers may be higher.

5. Conclusions

Based on our results, we generally recommend methods that incorporate physiological knowledge over data-driven methods. The EM algorithm is not recommended for outlier detection in 24-hour hormonal data since the method seems to falsely distinguish true biological variations due to circadian factors, such as meal response or day-night differences, as outliers. Tukey's fences, the other data-driven method, is not recommended in 24-hour hormonal data. Since no statistical assumptions have to be made and fewer data points will be removed, eyeballing could be a good method for detecting outliers. However, since it is time-consuming (depending on the number of participants studied), it might not always be practical. The strengths and limitations of each method are presented in Table 4.

Table 4. Methods for detecting measurement errors

	Eyeballing	Tukey's fences	Stepwise approach	The EM algorithm
Underlying assumptions	• Researchers' expert knowledge is reliable.	• From how much standard deviations (or interquartile range) away from a smoothing curve is considered to be an outlier should be decided by researchers.		• Normal distributions
Efficiency and generalizability of the method	• Relatively time-consuming process. • Different experts' knowledge is required for different types of data.	• Although it needs several adjustments for different types of time series (e.g., parameters for smoothing curves), the processes can be easily applied to different settings.		
Limitations	• Explicit knowledge and clear physiological reasoning behind the detection process. • Disagreement between experts may happen.	• The method is highly affected by smoothing techniques and type of data, especially when the hormone levels are mostly constant over time.	• Measurement error within a hormone and within a sampling method (serum) can both be detected	• Yields a probability • Need a large sample to be able to distinguish between the distributions

In conclusion, we recommend Stepwise approach for detecting outliers in serial 24-hour hormonal data since this method combines both physiological knowledge and an automated process. However, decisions such as which cut-offs of standard deviation should be applied or which hormones can be used together in the method should be supported by solid physiological knowledge. Stepwise approach is especially suitable for data of several hormone measurements from the same tube and when dilution is a possible cause of measurement errors. In this case, the outlier detection process can improve by taking along a reference measurement together with the hormonal measurements, whose concentration is stable over the day, such as creatinine or urea.

Although the methods showed different performances in outlier detection, this had little impact on the statistical outcomes. Overall, 24-hour means and cross-correlations did not materially change, but on an individual basis, correlations might change. The influence of outliers may depend on the study's sample size and outcome of interest. We recommend researchers be aware of the potential influence of measurement errors in their study and consciously decide which method to choose for outlier detection and whether it is necessary to remove outliers at all.

References

1. Aitkin M, and Wilson GT (1980) Mixture Models, Outliers, and the EM Algorithm. Technometrics 22:325-331.
2. Akintola AA, Jansen SW, Wilde RB, Hultzer G, Rodenburg R, and van Heemst D (2015) A simple and versatile method for frequent 24 h blood sample collection in healthy older adults. MethodsX 2:33-38.
3. Benaglia T, Chauveau D, Hunter DR, and Young D (2009). mixtools: An R Package for Analyzing Finite Mixture Models. Journal of Statistical Software, 32(6), 1-29. URL http://www.jstatsoft.org/v32/i06/.
4. Brown EN, Meehan PM, and Dempster AP (2001) A stochastic differential equation model of diurnal cortisol patterns. American Journal of Physiology-Endocrinology and Metabolism 280:E450-E461.
5. Dempster AP, Laird NM, and Rubin DB (1977) Maximum likelihood from incomplete data via the EM algorithm. Journal of the royal statistical society Series B (methodological):1-38.
6. Faghih RT, Dahleh MA, Adler GK, Klerman EB, and Brown EN (2014) Deconvolution of Serum Cortisol Levels by Using Compressed Sensing. PLOS ONE 9:e85204.
7. Feneberg R, Sparber M, Veldhuis JD, Mehls O, Ritz E, and Schaefer F (1999) Synchronous fluctuations of blood insulin and lactate concentrations in humans. J Clin Endocrinol Metab 84:220-227.
8. Grossmann M, Wong R, Szkudlinski MW, and Weintraub BD (1997) Human thyroid-stimulating hormone (hTSH) subunit gene fusion produces hTSH with increased stability and serum half-life and compensates for mutagenesis-induced defects in subunit association. J Biol Chem 272:21312-21316.
9. Grubbs FE (1969) Procedures for Detecting Outlying Observations in Samples. Technometrics 11:1-21.
10. Horn PS, Britton PW, and Lewis DF (1988) On The Prediction of a Single Future Observation from a Possibly Noisy Sample. Journal of the Royal Statistical Society Series D (The Statistician) 37:165-172.
11. Hung W-L, and Yang M-S (2006) An omission approach for detecting outliers in fuzzy regression models. Fuzzy Sets and Systems 157:3109-3122.
12. Jansen SW, Akintola AA, Roelfsema F, van der Spoel E, Cobbaert CM, Ballieux BE, Egri P, Kvarta-Papp Z, Gereben B, Fekete C, Slagboom PE, van der Grond J, Demeneix BA, Pijl H, Westendorp RG, and van Heemst D (2015) Human longevity is characterised by high thyroid stimulating hormone secretion without altered energy metabolism. Sci Rep 5:11525.
13. Kimenai DM, Henry RM, van der Kallen CJ, Dagnelie PC, Schram MT, Stehouwer CD, van Suijlen JD, Niens M, Bekers O, Sep SJ, Schaper NC, van Dieijen-Visser MP, and Meex SJ (2016) Direct comparison of clinical decision limits for cardiac troponin T and I. Heart 102:610-616.
14. Larsson J (2018) eulerr: Area-Proportional Euler and Venn Diagrams with Ellipses. R package version 4.1.0. In.
15. Luo J, Frisken S, Machado I, Zhang M, Pieper S, Golland P, Toews M, Unadkat P, Sedghi A, Zhou H, Mehrtash A, Preiswerk F, Cheng C-C, Golby A, Sugiyama M, and Wells WM (2018) Using the variogram for vector outlier screening: application to feature-based image registration. International Journal of Computer Assisted Radiology and Surgery.

16. Montgomery DC, Jennings CL, and Kulahci M (2015) Introduction to time series analysis and forecasting. John Wiley & Sons.

17. Muraleedharan G, Lucas C, and Guedes Soares C (2016) Regression quantile models for estimating trends in extreme significant wave heights. Ocean Engineering 118:204-215.

18. O'Brien JD, Kahn RM, Zenko Z, Fernandez JR, and Ariely D (2018) Naïve models of dietary splurges: Beliefs about caloric compensation and weight change following non-habitual overconsumption. Appetite 128:321-332.

19. Odell WD, Utiger RD, Wilber JF, and Condliffe PG (1967) Estimation of the secretion rate of thyrotropin in man. J Clin Invest 46:953-959.

20. Oike H, Oishi K, and Kobori M (2014) Nutrients, Clock Genes, and Chrononutrition. Curr Nutr Rep 3:204-212.

21. Pham NM, and Eggleston K (2016) Prevalence and determinants of diabetes and prediabetes among Vietnamese adults. Diabetes Research and Clinical Practice 113:116-124.

22. Porksen N, Nyholm B, Veldhuis JD, Butler PC, and Schmitz O (1997) In humans at least 75% of insulin secretion arises from punctuated insulin secretory bursts. Am J Physiol 273:E908-914.

23. Roelfsema F, Boelen A, Kalsbeek A, and Fliers E (2017) Regulatory aspects of the human hypothalamus-pituitary-thyroid axis. Best Pract Res Clin Endocrinol Metab 31:487-503.

24. Spiga F, Walker JJ, Gupta R, Terry JR, and Lightman SL (2015) 60 YEARS OF NEUROENDOCRINOLOGY: Glucocorticoid dynamics: insights from mathematical, experimental and clinical studies. J Endocrinol 226:T55-66.

25. Tukey JW (1977) Exploratory data analysis. Reading, Mass.

26. Veldhuis J, Yang R, Roelfsema F, and Takahashi P (2016) Proinflammatory Cytokine Infusion Attenuates LH's Feedforward on Testosterone Secretion: Modulation by Age. J Clin Endocrinol Metab 101:539-549.

27. Vis DJ, Westerhuis JA, Hoefsloot HC, Roelfsema F, van der Greef J, Hendriks MM, and Smilde AK (2014) Network identification of hormonal regulation. PLoS One 9:e96284.

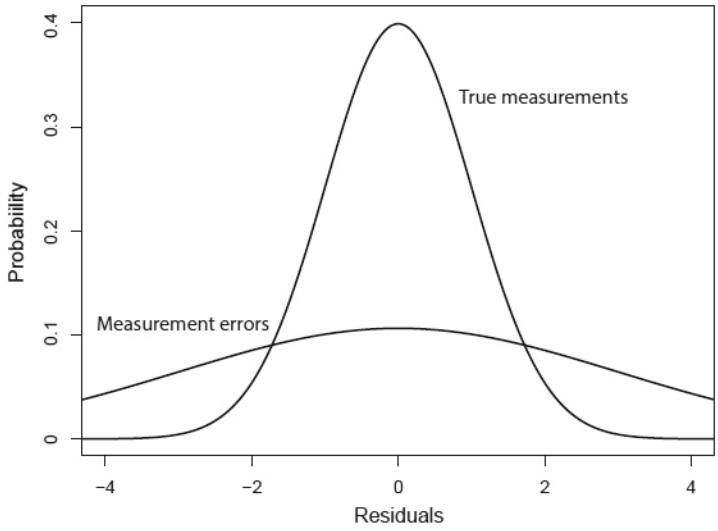
28. Westendorp RG, van Heemst D, Rozing MP, Frolich M, Mooijaart SP, Blauw GJ, Beekman M, Heijmans BT, de Craen AJ, Slagboom PE, and Leiden Longevity Study G (2009) Nonagenarian siblings and their offspring display lower risk of mortality and morbidity than sporadic onagenarians: The Leiden Longevity Study. J Am Geriatr Soc 57:1634-1637.

Appendix 1

Details on the EM algorithm to detect outliers

For each of the hormones separately, the EM algorithm was applied to the residuals of all subjects simultaneously, where the residual of the ith measurement of subject j was calculated as $R_{ij}= Y_{ij}- \hat{Y}_{ij}$, with $Y_{ij,}$ the observed measurement and $\hat{Y}_{ij}$, the moving average smoothed estimate.

We assumed that there were two types of measurements: true measurements and erroneous measurements. We expected that the residuals of the true measurements had standard deviations close to 0, while erroneous measurements had a much larger standard deviation.



The (unobserved) indicator variable Z denotes whether a measurement is an error, with Z_{ij}=1 if the ith measurement of subject j is an error and Z_{ij} =0 if it is a true measurement. The proportion of erroneous measurements $Pr(Z_{ij}=1)$ is denoted by π_e . We assumed that residuals R of true measurements were normally distributed with mean 0 and standard deviation σ_1 while the residuals of the erroneous measurements were normally distributed with mean 0 and standard deviation σ_2 , with $\sigma_2 >> \sigma_1$. The proportion of erroneous parameters π_e and the standard deviations σ_1 and σ_2, can be estimated using maximum likelihood. The complete likelihood of the data is

$$L(\sigma_1, \sigma_2; R, Z) = \prod_{ij} f(R_{ij};\ \sigma_1)^{(1-Z_{ij})}\ f(R_{ij};\ \sigma_2)^{Z_{ij}},$$

with $f(R_{ij}; \sigma_i)$, the normal density with mean 0 and standard deviation σ_i . Because the Z_{ij} are unobserved, the EM algorithm is applied with following EM steps:

E step: given current estimates p_e s_1 and s_2 for π_e σ_1 and σ_2 , the expected probability of being an error is estimated using Bayes formula:

$$\Pr(Z_{ij}=1|\ R_{ij}) = \frac{p_e f(R_{ij};\, s_2)}{(1-p_e) f(R_{ij};\, s_1)+p_e f(R_{ij};\, s_2)} \quad (1)$$

M step: the likelihood function where the Z_{ij} are replaced by the expected probabilities that Z_{ij} is 1, is maximized.

The EM steps are repeated until convergence. The final estimates p_e, s_1, and s_2 are filled in in equation (1). This yields for each measurement an estimated probability of being an error measurement.

The EM algorithm was applied in R version 3.5.1, using the normalmixEM function of the package mixtools.

Reference

Benaglia T, Chauveau D, Hunter DR, Young D (2009). mixtools: An R Package for Analyzing Finite Mixture Models. Journal of Statistical Software, 32(6), 1-29. URL http://www.jstatsoft.org/v32/i06/.

Appendix 2

Detecting outliers in 24-hour hormonal data: a simulation study

1. Data generation

We simulate measurements for five hormones; glucose, insulin, thyroid stimulating hormone (TSH), cortisol, and growth hormone (GH), according to their physiological characteristics and the laboratory setting where our sample was drawn. This setting was reproduced in simulation as described below:

- 24 hours with measurements every 10 minutes, in total 144 measurements per hormone and person.
- Three meals at time 0, 18, 54.
- Night from time 84 to time 138.

For each hormone, we generated measurements. The mean hormone value at time t, Y(t) consisted of a constant baseline level and one or more peaks using an absorption/elimination model. A peak starting at t_s has the form:

$$Y(t) = C_0 + C_1 (t>t_s)(e^{-\lambda_e t} - e^{-\lambda_a t}),$$

where C_0 determines the minimum hormone values over time, C_1 the peak value and λ_a and λ_e the rate of absorption and elimination of the hormone in the blood. The latter is directly related to the half-life of the hormone by λ_e=ln(2)/half-life. Random between and within-person variation was added to the generated mean values. The specific minimum, location and duration of peaks, and the random intra/inter-person variation were based on the observed patterns in our data. Specific features of each hormone are:

- Glucose: Three clear peaks after meals, where the third one is slightly higher than others. At night, the hormone level is stable and low, and the variation is smaller. Physiologically, glucose levels cannot be below 2.8 mmol/L.
- Insulin: Three clear peaks after each meal, and the hormone is highly correlated with glucose (corr.=0.75). At night, the hormone level is stable and low, and the variation is smaller.
- TSH: One prominent peak, where the hormone builds up in the evening from 6 pm (t=54) with the highest levels at 11 pm (t=84), with large variation.
- Cortisol: Peaks at the end of the night.
- GH : Sharp peaks and the number of peaks varies from 0 to 20 across the individuals.

Inter-person variation is generated by varying the highest concentration reached during peaks, following a normal distribution (specific parameters are provided in the table

below). For TSH, cortisol, and GH, the location of the peaks also varies across people. In this way, we generated 24-hour hormonal data for 38 simulated subjects. Table A1 shows the specific parameters used for simulating the 24-hour hormonal data of 38 individuals.

In each individual, for each hormone, we generated measurement errors at 14 time points. To generate random measurement errors in each hormone at seven randomly selected time points (5% out of 144 points), we replaced the true measurement with an error measurement drawn from a uniform distribution with a wide range (-10 x intra-person SD to 15 x intra-person SD). Furthermore, we generated related dilution errors at seven time points which were the same across all hormones for one individual. The dilution errors were generated by dividing the original measurement by 2.

Table A1. Parameters for generating 24-hour glucose, insulin, TSH, cortisol and GH data

	Glucose [mmol/L]	Insulin [mU/L]	TSH [mU/L]	Cortisol [μmol/L]	GH [mU/L]
Starting value (C_0)	3.8	6.6	1	0.05	1
Number and location of peaks	3 peaks, increase starts at mealtimes	3 peaks, increase starts at mealtimes	One wide peak, increase starts between t=45 and 65	3 peaks, Increase starts between (i) t=75 and 100, (ii) between t=100 and 124, and (iii) between 124 and 140	0 to 20 peaks, increase starts from t=0 and 143
Half-life	35 min	35 min	120 min	50 min	10 min
Intra-person variation (SD)	Day 0.50, Night 0.25	Day 6.5 Night 3.2	0.17	0.03	0.27
Mean and SD of peaks: first peak (i), second peak (ii), third peak (iii), with inter person Sd	(i) 4 (0.5), (ii) 4 (0.5), (iii) 7 (0.7)	90 (5)	2.5 (0.5)	(i) 0.3 (0.1), (ii) 0.4 (0.1), (iii) 0.5 (0.1)	15 (1)
Remarks	Values <2.8 are set to 2.8	Values <2.8 are set to 2.8	Values <1 are set to 1	Values <0.05 are set to 0.05	Values <0.2 are set to 0.2
Absorption/ elimination rate	$\lambda_a = 1.1\,\lambda_e$	$\lambda_a = 1.1\,\lambda_e$	$\lambda_a = 1.1\,\lambda_e$	$\lambda_a = 2\,\lambda_e$	$\lambda_a = 1.1\,\lambda_e$
Comments		Log transformation			Log transformation

2. Simulation results

Figure A1 shows simulated 24-hour hormonal data for glucose, insulin, TSH, cortisol, and GH of the first two generated individuals are shown. The hormone-specific measurement errors are indicated by a red dot. The dilution errors are indicated by a green dot.

Figure A2 displays how many points are indicated as measurement errors by each method averaged across the 38 simulated subjects. The EM algorithm indicated the highest number of measurement errors, followed by the stepwise approach. Especially for the hormones where the intra-person variation was larger during the day than during the night (glucose and insulin), the EM algorithm indicated high numbers of measurement errors.

Table A2 shows what percentage of true errors (random errors and dilution errors) were detected by each method and how many non-errors were identified as errors by each method. When it comes to detecting a true error, the EM algorithm performed best. However, the EM algorithm also indicated the most non-errors as measurement errors. Especially for insulin, the number of true measurements falsely indicated as errors was extremely high. This is explained by the fact that the intra-person variation in insulin differed between day and night, and the insulin residuals were not normally distributed without log transformation. The percentage of non-error detected as measurement error was much lower in Stepwise approach and Tukey's fences than in the EM algorithm. Stepwise approach is to be preferred when detecting dilution errors.

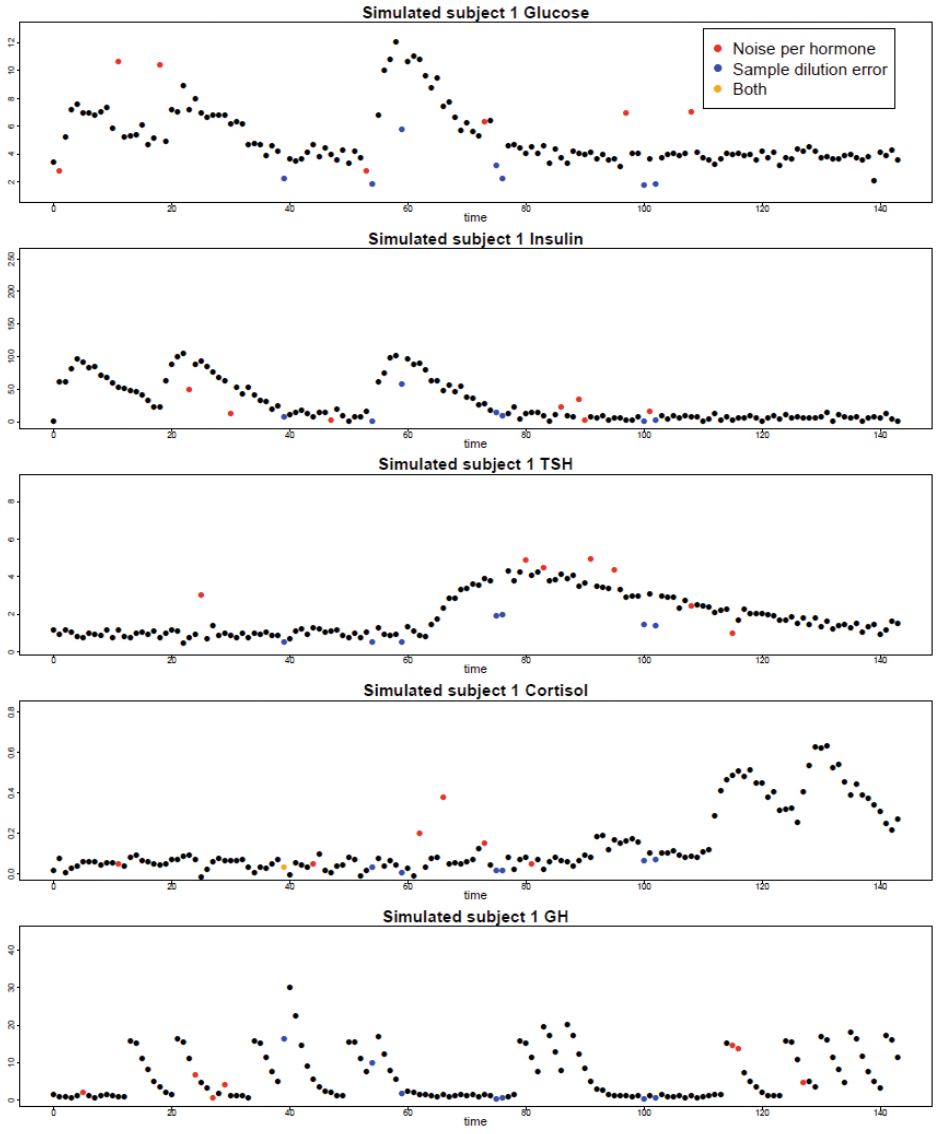


Figure A1. Simulated 24-hour glucose, insulin, TSH, cortisol, and GH data of the first two generated individuals

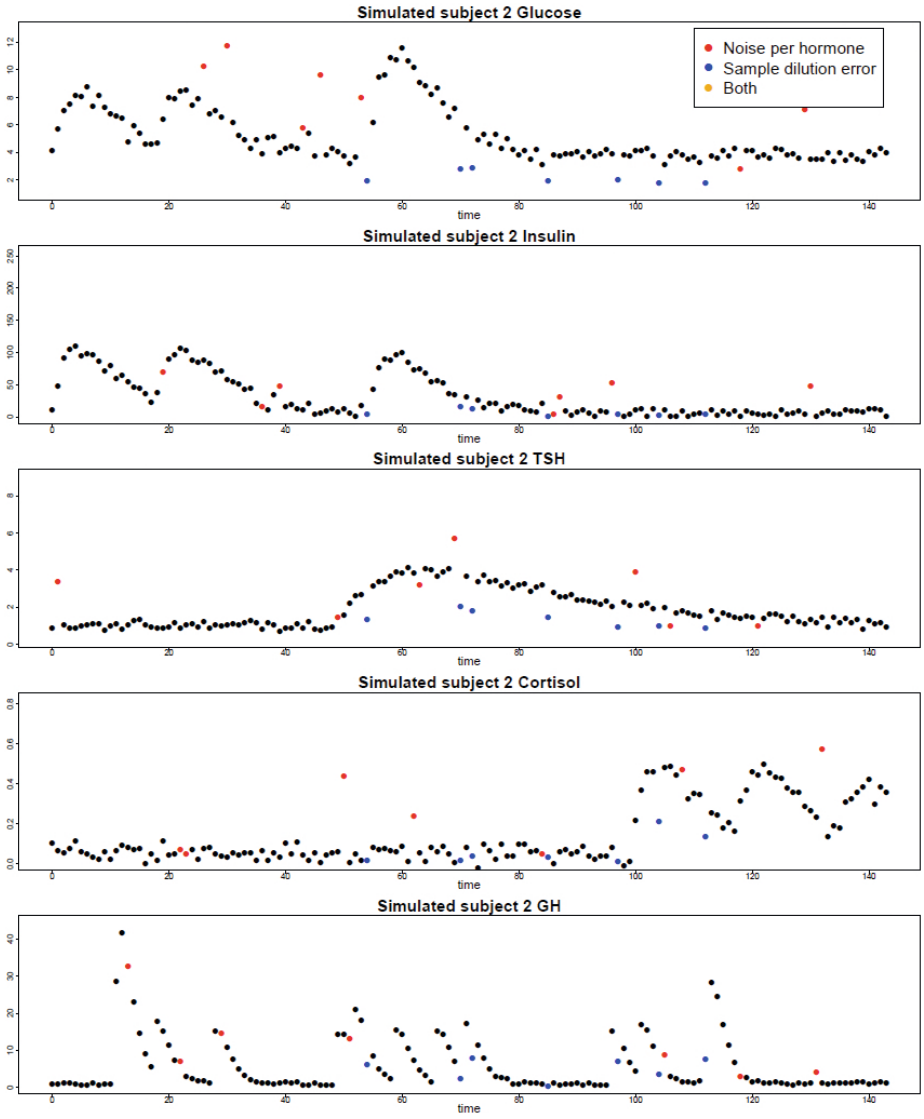


Figure A1 (cont'd)

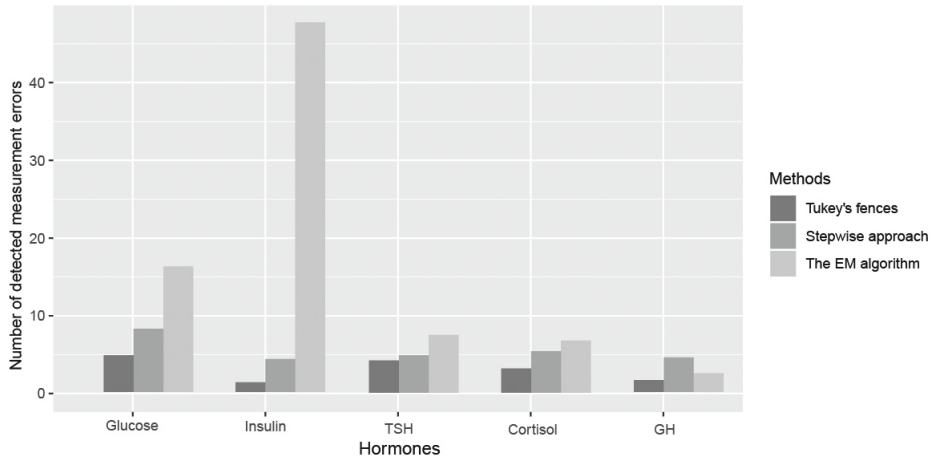


Figure A2. Simulated 24-hour glucose, insulin, TSH, cortisol, and GH data of the first two generated individuals

Table A2. Percentage of true errors detected and true measurement wrongly indicated as an error by each method stratified by random error and dilution error.

		Random error		Dilution error	
		True errors detected (%)	True measurements wrongly indicated as error (%)	True errors detected (%)	True measurements wrongly indicated as error (%)
Stepwise approach	Glucose	22.18	4.80	92.86	1.19
	Insulin	4.51	2.86	49.25	0.58
	TSH	18.42	2.59	53.76	0.79
	Cortisol	24.06	2.67	49.62	1.36
	GH	8.27	2.82	49.25	0.73
	mean	15.49	3.15	58.95	0.93
Tukey's fences	Glucose	34.96	1.67	29.32	1.96
	Insulin	7.89	0.50	4.14	0.69
	TSH	31.58	1.42	27.82	1.61
	Cortisol	31.95	0.65	7.52	1.90
	GH	7.89	0.71	2.63	0.98
	mean	22.86	0.99	14.29	1.43

Table A2. Percentage of true errors detected and true measurement wrongly indicated as an error by each method stratified by random error and dilution error. *(continued)*

		Random error		Dilution error	
		True errors detected (%)	True measurements wrongly indicated as error (%)	True errors detected (%)	True measurements wrongly indicated as error (%)
The EM algorithm	Glucose	60.53	8.70	90.98	7.15
	Insulin	74.81	30.89	77.82	30.73
	TSH	49.25	2.92	46.24	3.07
	Cortisol	43.98	2.65	18.80	3.94
	GH	8.65	1.31	4.14	1.54
	mean	47.44	9.29	47.59	9.29

Chapter 4

A comparison of different methods for handling measurements affected by medication use

Ready to be submitted

Jungyeon Choi, Olaf M. Dekkers, Saskia le Cessie

Abstract

In epidemiological research, it is common to encounter measurements affected by medication use, such as blood pressure lowered by antihypertensive drugs. When one is interested in the relation between the variables not affected by medication, ignoring medication use can cause bias. Several methods have been proposed, but the problem is often ignored or handled with generic methods, such as excluding individuals on medication or adjusting for medication use in the analysis.

This study aimed to investigate methods for handling measurements affected by medication use when one is interested in the relation between the unaffected variables and to provide guidance for how to handle the problem optimally. We focused on linear regression and distinguished between the situation where the affected measurement is an exposure, confounder, or outcome. In the Netherlands Epidemiology of Obesity study and several simulated settings, we compared generic and more advanced methods; such as substituting or adding a fixed value to the treated values, regression calibration, censored normal regression, Heckman's treatment model, and multiple imputation methods.

For an exposure affected by medication, restricting the analysis to untreated individuals could yield unbiased estimates. Regression calibration is an alternative, but the mean and standard deviation of the medication effect should be known. For an outcome affected by medication, adding the mean medication effect, censored normal regression, and imputation using censored regression worked well. For a confounder affected, selecting untreated individuals worked well, as well as adjusting for medication use, adding mean medication effect, and censored normal regression imputation. In conclusion, methods for handling medication effects should be carefully chosen based on which variable is affected by medication and available information of the clinical setting.

1. Introduction

Measurements affected by medication use are commonly encountered in epidemiological research. Examples are glucose levels lowered by glucose-lowering medications or blood pressure relieved by antihypertensive drugs. Depending on the research questions, these measurements can be an outcome of interest or covariates.

Although researchers often are interested in the effect of certain drugs, the relation between the values not affected by medication can also be the primary scientific interest. However, the value of a variable had an individual not been treated is often not available. Using the values affected by medication instead may lead to biased results. In clinical research, however, medication use is often ignored or handled with naïve methods such as excluding medication users or adjusting for medication use. For outcomes affected by medication use, these naïve methods may introduce bias (1-4).

Several methods have been proposed to handle measurements affected by medication use. Relatively simple methods are adding an expected medication effect to treated values or substituting the treated values for other values (1, 4, 5). More sophisticated methods include censored normal regression, Heckman's treatment model, quantile regression, measurement error methods, or advanced imputation techniques (1, 2, 6, 7). However, these methods are seldom used in applied research. Additionally, many of the suggested methods are limited to outcomes affected by medication, and little has been known about how to handle exposures or confounders affected by medication.

This study aims to investigate methods for handling measurements affected by medication use when the unaffected values are of interest. We focused on etiological studies where effects are estimated by linear regression. We discuss different methods and compare these methods in a large cross-sectional study of the Netherlands Epidemiology of Obesity (NEO) study and several simulation scenarios generated based on the NEO data. The scenarios vary on whether the exposure, confounder, or outcome is affected by medication use. Based on the results of the simulation study, we provide guidance on how to handle the medication effect optimally.

2. Methods to handle measurements affected by medication use

We will consider the situation where for some individuals a variable is affected by medication use (e.g., blood pressure affected by antihypertensive drugs), while the relation between variables when no one is affected by medication is of interest. For convenience, we assume that medication is taken when values are high, aiming to lower

the values. Depending on the research question, the variable affected by medication use can be the exposure, a confounder, or the outcome in an analysis.

The problem of measurements affected by medication can be viewed from different perspectives; it can be viewed as a missing data problem, because for people on medication, their untreated values are unobserved. It may be viewed as a measurement error problem, as the observed values differ systematically from the values had the treated individuals not been treated. It could also be viewed as a censoring problem if we assume that the unobserved untreated values are at least as high as the observed values under treatment. Depending on how one approaches the problem, methods for missing data, measurement error, or censored observations can be used.

Table 1 summarizes methods for handling measurements affected by medication use. The methods can be categorized as generic methods [M1-M5], a method for the exposure affected by medication [M6], methods for the outcome affected by medication [M7-M10], and multiple imputation approaches [M11-M13]. Detailed descriptions of each method and underlying assumptions are available in Appendix 1. All methods are applied to empirical and simulated data in the following sections.

Table 1. Overview of methods for Handling Measurements Affected by Medication use

	Methods	**Description**
Generic methods	**[M1]** Ignoring medication use	Medication use is ignored.
	[M2] Restricting to untreated individuals	The analysis is performed in the subgroup of individuals who are not receiving medication.
	[M3] Binary adjustment for medication use	An indicator for medication use is added as a covariate in the regression model.
	[M4] Substituting measurement of treated individuals with a fixed value	Measurements affected by medication are substituted with a prespecified value.
	[M5] Adding a constant value to observations of treated individuals	A prespecified treatment effect is added to the observed measurements of treated individuals.
For exposures affected by medication	**[M6]** Regression calibration	Measurement error methods are used. Based on the expected mean treatment effect and its standard deviation, the observed measurements affected by medication are corrected.

Table 1. Overview of methods for Handling Measurements Affected by Medication use *(continued)*

	Methods	**Description**
For outcomes affected by medication	**[M7]** Inverse probability weighting	Treated individuals are removed from the analysis, and a reweighted analysis is performed where more weight is given to individuals who are untreated but have a similar profile as treated individuals
	[M8] Quantile regression	The method assumes that the untreated values of individuals on medication would have been above the median, conditional on covariates. The median outcome is modelled as a function of covariates.
	[M9] Censored normal regression	Measurements of treated individuals are considered to be censored observations, where the untreated values are assumed to be at least as high as the observed values affected by treatment, or in more complex censoring mechanisms, at least as high as the observed values and a clinical guideline at which treatment is prescribed.
	[M10] Heckman's treatment model	Treatment assignment is assumed to be dependent on the untreated values, and the treatment results in a "structural shift" of the mean outcome.
Multiple imputation approaches	**[M11]** Predictive mean matching	A default multiple imputation option in commonly used statistical software. It assumes that the observations of treated individuals are missing at random.
	[M12] Censored normal imputation	Censored normal regression is used in the imputation algorithm to predict the untreated values of those on treatment.
	[M13] Heckman's model imputation	Heckman's model is used in the imputation algorithm to predict the untreated values of those on treatment

3. Example: the Netherlands Epidemiology of Obesity Study

The Netherlands Epidemiology of Obesity (NEO) study is a population-based study designed to investigate pathways that lead to obesity-related diseases. From 2008 to 2012, 6,671 individuals aged 45–65 years were included in the study. Participants brought all medication they were using to the NEO study site, which was coded using the Anatomical Therapeutic Chemical Classification (8). Details can be found elsewhere (9). The NEO study data includes several measurements affected by medication; for example, 31% of the participants used antihypertensive medication, and 15% used lipid-lowering medication.

To illustrate the effect of different methods for handling medication use, we use data collected at baseline and consider three research questions:

i) The effect of systolic blood pressure (SBP) on the intima-media thickness (IMT), where the exposure is affected by medication.
ii) The effect of BMI on SBP, where the outcome is affected by medication.
iii) The effect of BMI on IMT, adjusted for SBP, where the confounder is affected by medication.

All methods described in Table 1 were applied to estimate the regression models corresponding to the three research questions stated above. The analyses were adjusted for potential confounders: BMI, sex, age, education level, and smoking status.

In the Netherlands, physicians prescribe blood pressure medication generally aiming at values below 140 mmHg (10). Therefore, we replaced treated SBP values with 150 mmHg in the substitution method [M4] and repeated it using 170 mmHg. For adding medication effect [M5], we followed previous literature using the values 10 mmHg and 15 mmHg (4, 11). For regression calibration [M6], the assumed mean treatment effect was 15 mmHg; SD=10 mmHg.

For inverse probability weighting [M7], logistic regression was used to estimate the probability of medication use based on 21 covariates (see Appendix 2 for details). The same covariates were used in the probit part of Heckman's treatment model [M10] and in the multiple imputation approaches [M11-M13]. For quantile regression [M8], the values of treated individuals were replaced by 150, 170, and 190 mmHg. For censored regression [M9] and imputation [M12], we used 140 mmHg and 160 mmHg as a clinical threshold for treatment prescription. For research questions i) and iii) the outcome variable IMT was added to the imputation models (12). Ten imputed datasets were created in each imputation.

All analyses were performed using R version 3.6.1, with packages Survival v3.1-8 for (13) [M9], SampleSelection v1.2-6 (function treatreg) (14) for [M10], Quantreg v5.54 (15) for [M11], MICE v3.7.0 (16) with default options for [M11] and miceMNAR (17) for [M13]. R code for [M12] is provided in Appendix 2.

Figure 1 presents effect estimates from the different methods for the three research questions. The results show that different methods can lead to quite different effect estimates in all three considered situations. This signals that choosing an appropriate method for handling measurements affected by medication use is essential for the validity of study results.

4. Simulation studies

To understand the results of the NEO study and provide recommendations, we performed several so-called real-life simulation studies. To mimic the NEO study as closely as possible, we used the baseline variables of the NEO participants (BMI, sex, age education, and LDL cholesterol). We simulated SBP, antihypertensive medication prescription, and IMT values based on the other baseline variables directly from the NEO study. We generated different scenarios where blood pressure could be the exposure (scenario 1), the outcome (scenario 2), or the confounder (scenario 3). In each scenario, we considered the research questions i), ii), and iii) of Section 3, respectively.

4.1 Simulation setting 1: Medication effect on the exposure

In this simulation setting, we are interested in the effect of SBP on IMT, with SBP affected by antihypertensive drugs in some individuals. The *untreated SBP* depended linearly on *BMI, sex* (*man*=0, *women*=1), *age,* and *education* (*low*=0, *high*=1), with parameter values closely corresponding to observed values in the NEO study:

$$Untreated\ SBP\ = 90 + 0.8\,BMI - 8.0\,Sex + 0.6\,Age$$

with the residual error ε_{SBP} normally distributed with mean 0 and SD 15.9 mmHg. The probability of receiving medication depended on *BMI*, *sex*, *education,* and the *untreated SBP* values:

$$logit\big(pr(Medication = 1)\big) =$$
$$-16 + 0.01\,BMI - 0.5\,Sex - 0.3\,Education + 0.1\,Untreated\ SBP$$

In this way, approximately 28% of the participants were treated for high SBP. For a SBP of 150 mmHg, the probability of receiving medication was approximately 11%, while for 180 mmHg, the probability was 88%. The *Observed SBP* was lowered when medication was used:

$$Observed\ SBP = Untreated\ SBP - medication\ effect, \quad if\ Medication = 1$$
$$Observed\ SBP = Untreated\ SBP, \quad if\ Medication = 0,$$

where the *medication effect* was generated from a normal distribution (30 mmHg, SD=10 mmHg). The outcome, IMT was generated as:

$$IMT = 31 + 0.2\ Untreated\ SBP + 0.3\ BMI + 2.8\ Sex + 0.4\ Age + 0.8\ Fasting$$

with ε_{IMT} following a normal distribution (0, SD= 9.2 mm). The relation between medication use and IMT is confounded by sex and BMI.

4.2 Simulation setting 2: Medication effect on the outcome

In Simulation setting 2, we consider the effect of BMI on untreated SBP. BMI was taken directly from the NEO data. Untreated SBP, medication prescription, and the observed SBP were generated in the same way as in Simulation setting 1.

4.3 Simulation setting 3: Medication effect on a confounder

Here, we consider the effect of BMI on IMT measurement when adjusted for SBP. Untreated and observed SBP, medication prescription, and IMT were generated as in Simulation setting 1.

4.4 Alternative simulation scenarios

Simulation setting 1, 2, and 3 were repeated while changing three parameters: i) The size of the mean treatment effect decreased from 30 mmHg to 10 mmHg. In this simulation, 16% of the treated individuals' SBP increased after medication. ii) The standard deviation of the treatment effect changed from 10 mmHg to 1 mmHg. iii) The percentage of individuals on medication increased from approximately 28% to 50% by changing the intercept of the logistic model for medication use.

4.5 Analysis

All methods [M1-M13] were applied to the simulated data sets in the same way as described in Section 3, except we used 20 mmHg and 30 mmHg to add to the treated SBP in [M5]. Analyses were adjusted for BMI, sex, age, education level, and smoking status. Each simulation was repeated 1000 times. The estimates obtained from using untreated SBP values were considered a reference. Mean bias and mean squared error were calculated as an overall measure of performance.

ε_{IMT},

5. Results

5.1 Simulation setting 1: Medication effect on the exposure

Figure 2 (left) and Table 2 display the results of simulation setting 1. The results show that medication use cannot be ignored [M1]. Restricting the analysis to untreated individuals [M2] yielded estimates very close to the true values. In this setting, medication use was affected by the exposure and several covariates, in which case one should adjust for all variables both affecting medication use and the outcome to prevent selection bias (18). Furthermore, there was no effect modification, meaning that the effect of SBP on the outcome in the subgroup of untreated individuals is the same as in the total population.

Binary adjustment for medication use [M3] did not work well. In our simulation, the medication effect was generated with large variability. This random variability in medication effect attenuated the association between SBP and IMT in the *treated* individuals and led to a bias toward the null in the overall effect. The method worked better when the variance of the medication effect was smaller (Appendix 3). Substituting treated values [M4] did not perform well in any scenarios. The method cannot reconstruct the original distribution of the exposure and, therefore in general, will yield biased results.

Adding 30 mmHg [M5], which was the true mean medication effect in our simulations, did not perform well either. The reason is that the medication effect was generated with SD=10 mmHg. Therefore, by adding 30 mmHg to all treated SBP values, we reconstruct untreated SBP with random measurement error. Random measurement error in exposures will bias the estimates in a regression model (14). The method performed better when the random variation of the medication effect was smaller (Appendix 3). Regression calibration [M6] yielded unbiased results in all our simulations scenarios, assuming that true medication effect and standard deviation are known.

None of the multiple imputation methods [M11-M13] yielded valid results. A possible explanation is that the imputation models included the outcome, which does not correspond to how medication use was generated in our simulations.

5.2 Simulation setting 2: Medication effect on the outcome

Figure 2 (middle) and Table 2 show the results of Simulation setting 2. Ignoring medication use [M1], restricting to untreated subgroup [M2], and binary adjustment for medication use [M3] yielded biased results. As the outcome determines medication use directly, adjusting or selecting based on medication use [M2 & M3] implies selection based on outcome values, which will generally lead to selection bias (19, 20).

Substituting method [M4] using 150 mmHg led to a large underestimation. It performed better when 170 mmHg was used, which was slightly higher than the mean untreated SBP in the treated individuals (164 mmHg). Regardless of the substituting values, however, the method cannot reconstruct the original distribution of the outcome.

Adding 30 mmHg [M5] yielded unbiased results in all simulation settings (Appendix 4). Unlike in Simulation setting 1, adding the true mean medication effect yields valid results irrespective of the amount of variance in the medication effect.

Inverse probability weighting [M7] resulted in a large bias. Quantile regression [M8] performed poorly for all replacement values. In our simulation setting, more than 50% were using antihypertensive drugs among individuals with very high BMI. Therefore, the median SBP conditional on high BMI was affected by the substituting values.

Censored normal regression [M9] performed reasonably well when the simple censoring method was used or when clinical guideline set to 140 mmHg was applied. However, in alternative scenarios with a smaller medication effect, the results were off (Appendix 4). One reason is that the treated SBP was sometimes higher in these scenarios than the untreated SBP. This violates the assumption that untreated values are at least as high as untreated values (1). Heckman's treatment model [M10] performed less well in our main scenario, which contrasts with the results reported by Spieker et al. (2, 7). Heckman's treatment model assumes that the residual variances of two linear regression models, one for untreated individuals and the other for treated individuals, are equal. This assumption was violated in our main simulation scenario, as we simulated a medication effect with large random variability. This reflects the reported instability of Heckman's treatment (21, 22). In the scenarios with a smaller variance in the medication effect, Heckman's treatment model outperformed the censored regression (Appendix 4).

Multiple imputation with predictive mean matching [M11] resulted in bias. Results of multiple imputation with censored regression [M12] were only slightly biased, but for smaller medication effects, the method performed less well. Multiple imputation with Heckman's model [M13] sometimes yielded a large underestimation of the effect.

5.3 Simulation setting 3: Medication effect in a confounder

Figure 2 (right) and Table 2 show the results of Simulation setting 3. Ignoring the medication effect [M1] resulted in bias. Restricting to untreated individuals [M2] performed well, which is the same as adjusting for confounding by restriction. The method will yield valid estimation under the conditions as in Simulation setting 1, that is, with proper adjustment for variables affecting both medication use and the outcome. Binary adjustment for medication use [M3] yielded results close to the truth. Substitution methods [M4] were biased, because the distribution of untreated SBP could not be correctly reconstructed.

Adding 30 mmHg [M5] yielded a very small upward bias. This is due to the random measurement error introduced by the method. It has been known random measurement error in exposures attenuates the effect, while random measurement in confounders can lead to overestimation (23, 24).

Multiple imputation with censored regression [M12] yielded results close to the truth, especially when clinical guideline information was incorporated, and performed better than multiple imputation with Heckman's model [M13]. All results were consistent in the alternative simulation scenarios (Appendix 5).

Table 2. Mean Coefficient, Standard Deviation, Bias, and Mean Squared Error (MSE) for Three Main Simulation Settings

	Variable affected by medication use:											
	Exposure affected[1]				Outcome affected[2]				Confounder affected[3]			
Methods	**Mean**	**SD**	**Bias**	**MSEx1000**	**Mean**	**SD**	**Bias**	**MSEx10**	**Mean**	**SD**	**Bias**	**MSEx100**
All untreated values known (true coefficient)	0.200	0.006	0.000	0.004	0.800	0.053	0.000	0.028	0.292	0.025	0.000	0.063
Generic methods												
[M1] Ignoring medication use	0.165	0.007	-0.035	0.127	0.483	0.048	-0.317	1.028	0.372	0.025	0.080	0.703
[M2] Restricting to untreated individuals	0.200	0.008	0.000	0.006	0.564	0.054	-0.236	0.586	0.295	0.030	0.003	0.091
[M3] Binary adjustment for medication use	0.182	0.007	-0.018	0.037	0.533	0.048	-0.267	0.736	0.302	0.025	0.010	0.073
[M4] Substituting treated values												
to 150 mmHg	0.223	0.007	0.023	0.058	0.532	0.040	-0.268	0.734	0.333	0.026	0.041	0.236
to 170 mmHg	0.180	0.006	-0.020	0.044	0.743	0.052	-0.057	0.060	0.318	0.025	0.026	0.130
[M5] Adding a constant value												
20 mmHg	0.200	0.006	0.000	0.004	0.694	0.051	-0.106	0.138	0.313	0.025	0.021	0.107
30 mmHg (true value)	0.187	0.006	-0.013	0.021	0.800	0.056	0.000	0.031	0.302	0.025	0.010	0.073
Methods for exposure affected												
[M6] Regression calibration	0.199	0.006	-0.001	0.004	-	-	-	-	-	-	-	-
Methods for outcome affected												
[M7] Inverse probability weighting	-	-	-	-	0.521	0.056	-0.279	0.810	-	-	-	-

Table 2. Mean Coefficient, Standard Deviation, Bias, and Mean Squared Error (MSE) for Three Main Simulation Settings *(continued)*

	Variable affected by medication use:											
	Exposure affected[1]				**Outcome affected[2]**				**Confounder affected[3]**			
[M8] Quantile regression												
k = 150 mmHg	-	-	-	-	0.429	0.037	-0.371	1.390	-	-	-	-
k = 170 mmHg	-	-	-	-	0.948	0.070	0.148	0.268	-	-	-	-
k = 190 mmHg	-	-	-	-	1.052	0.100	0.252	0.735	-	-	-	-
[M9] Censored normal regression												
standard censoring	-	-	-	-	0.749	0.054	-0.051	0.055	-	-	-	-
with guideline at 140 mmHg	-	-	-	-	0.756	0.054	-0.044	0.049	-	-	-	-
with guideline at 160 mmHg	-	-	-	-	0.879	0.063	0.079	0.102	-	-	-	-
[M10] Heckman's treatment model	-	-	-	-	0.660	0.083	-0.140	0.265	-	-	-	-
Multiple imputation methods												
[M11] Predictive mean matching	0.208	0.008	0.008	0.013	0.555	0.056	-0.245	0.632	0.332	0.026	0.040	0.228
[M12] Censored normal regression												
standard censoring	0.221	0.007	0.021	0.049	0.750	0.055	-0.050	0.055	0.286	0.025	-0.006	0.066
with guideline at 140 mmHg	0.212	0.007	0.012	0.019	0.756	0.055	-0.044	0.050	0.291	0.025	-0.001	0.063
[M13] Heckman's model	0.182	0.008	-0.018	0.039	0.737	0.127	-0.063	0.201	0.311	0.030	0.019	0.126

[1]Scenario 1: Effect of systolic blood pressure on IMT measurement. [2] Scenario 2: Effect of BMI on systolic blood pressure. [3] Scenario 3: Effect of BMI on IMT measurement, where systolic blood pressure is one of the confounders. In all scenarios, systolic blood pressure was the variable affected by medication.

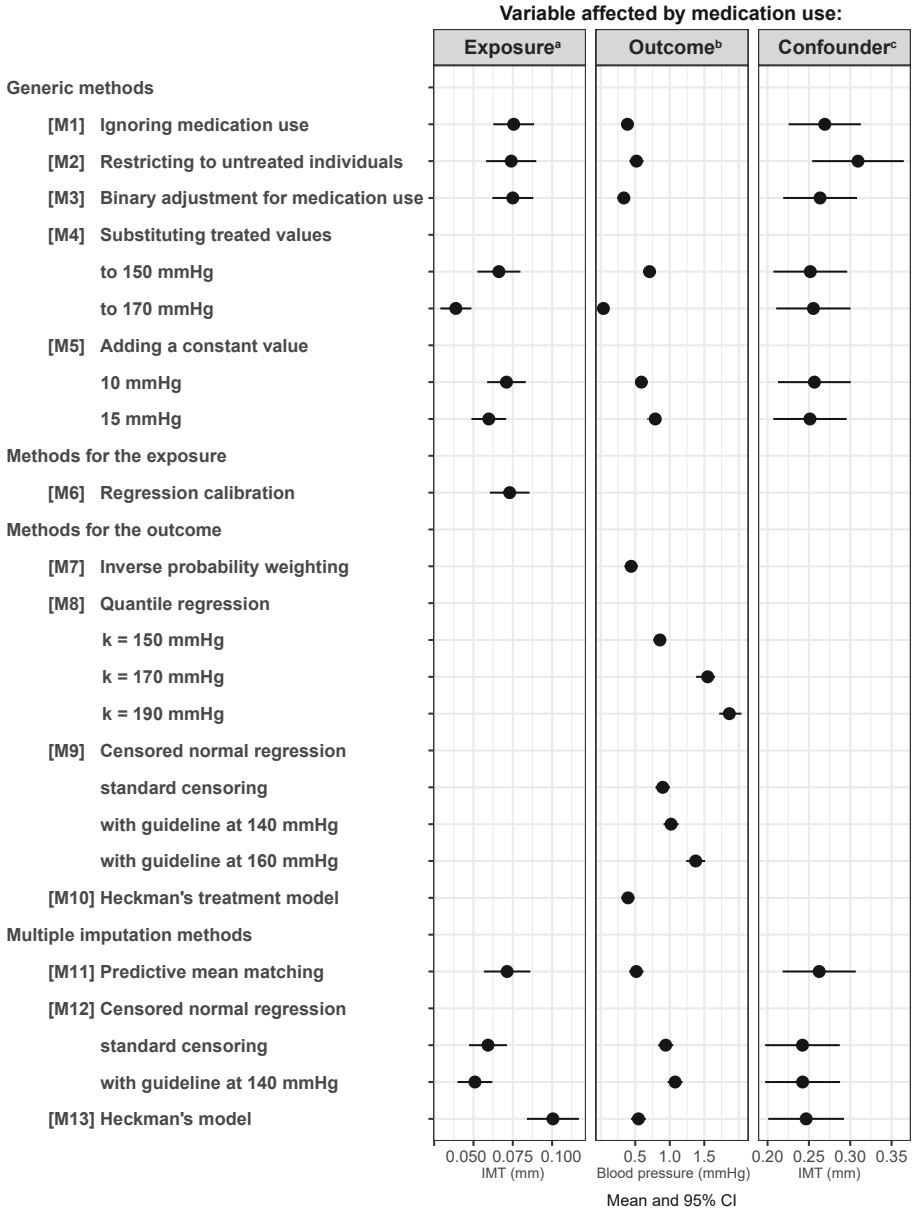


Figure 1. Regression coefficients and their 95% confidence interval estimated from the NEO data using the different methods to handle medication effect. In all analyses, SBP was the variable affected by medication. [a]Question 1: effect of SBP (mmHg) on IMT (mm). [b]Question 2: effect of BMI (kg/m2) on SBP. [C]Question 3: effect of BMI on IMT where SBP is a confounder.

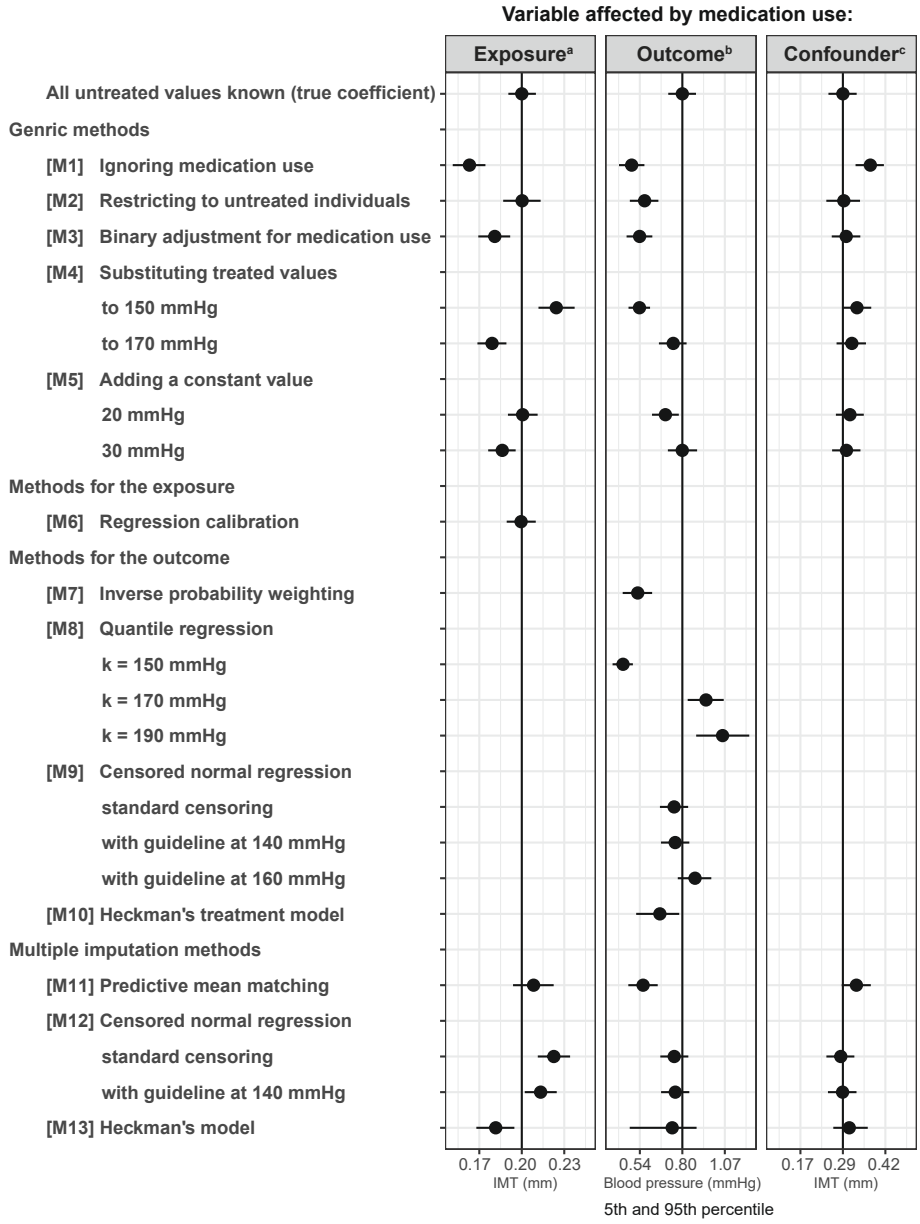


Figure 2. Regression coefficients and their 5th and 95th percentile estimated from simulation setting 1, 2 and 3. Results are standardized based on the mean and standard deviation of the true coefficients in each simulation setting. One grid unit represents 2.5 standard deviation. In all scenarios, SBP was the variable affected by medication use. [a]Question 1: effect of SBP (mmHg) on IMT (mm). [b]Question 2: effect of BMI (kg/m2) on SBP. [c]Question 3: effect of BMI on IMT where SBP is a confounder.

6. Guidance on how to optimally handle measurements affected by medication use

When interest is in the relation between the unaffected variables, ignoring medication use will in general yield biased results regardless of whether the exposure, outcome, or confounder is affected by medication. To obtain valid estimates, adequate methods for handling medication use are needed.

What to do when exposure is affected by medication?

- Performing analysis on the untreated individuals [M2] is a valid approach and will not lead to a large loss in power if the number of treated individuals is relatively low. However, there are two things to consider when applying this method: i) One should adjust for variables that both affect medication use and the outcome. ii) The result cannot be generalized to the total population if the effect of the exposure on the outcome is heterogeneous.
- Regression calibration [M6] may be used but requires an external estimate of the medication effect with its standard deviation.

What to do when the outcome is affected by medication?

- When an estimate of the mean medication effect is available, it could be added to the measurements of treated individuals. This method was also advocated by Tobin et al. (1). Like them, we also highly recommend performing sensitivity analysis with several different values to determine the stability of effect estimates.
- Quantile (median) regression [M8] can be used when less than 50% of the individuals are treated at any value of the exposure. The method does not require knowledge of the medication effect and can yield robust estimates but with lower power (6) than other methods.
- The advantage of censored normal regression [M9] or multiple imputation with censored normal regression [M12] is that no treatment effect needs to be specified. However, the method assumes that the observed values are lower than the untreated values, which could be violated when the treatment is ineffective. Furthermore, the method assumes non-informative censoring, which is likely to be violated in most clinical settings. In our simulation, we relaxed this assumption by incorporating knowledge from a clinical guideline into a censoring mechanism. Both in the study of Tobin et al. and in our main simulation study, the method was rather robust against the violation of the non-informative censoring assumption.
- Heckman's treatment model works well only if the treatment effect has a small variance.

What to do when the confounder is affected by medication?

- Restricting the analysis to untreated individuals [M2] is a valid approach with the same considerations as for the exposure affected by treatment.
- Using a binary indicator [M3] is a reasonable solution.
- Adding the true mean medication effect to the treated individuals [M5] performs relatively well.

7. Discussion

Our simulation study showed that the problem of variables affected by medication use should not be ignored, and proper methods are needed to avoid potential bias. Different methods are needed depending on whether the exposure, the outcome, or a confounder is affected by medication. Additional information, such as medication prescription patterns in clinical settings and the presence of effect heterogeneity, should also be considered carefully. Accordingly, all methods need to be used with caution.

One important consideration is the trade-off between the robustness of a method and the availability of external information. Methods that use external information on the medication effect, such as adding the mean medication effect or regression calibration, performed well when the external information was correct. However, such information is not always available. Other methods, such as censored regression, Heckman's treatment model, or multiple imputation methods, do not require assumptions on the medication effect. However, these methods rely on other assumptions and can perform suboptimally if the assumptions are violated.

We aimed our simulation scenarios to resemble realistic clinical situations instead of creating an ideal scenario for a particular method. Likely, assumptions required for statistical methods will not all be met in clinical data. Therefore, knowing which methods are robust against violation of assumptions is relevant. We encourage researchers to perform real-life simulations more often, as we did when generating simulations based on the NEO study data.

One limitation of our study is that we did not consider situations where more than one variable is affected by medication. Additionally, our study focused on the methods applicable to cross-sectional analyses. Other approaches may be available; for example, when there is an interaction by medication (25), when effect modifiers are associated with medication use (7), in longitudinal settings (26), or in the presence of interaction or mediation by time-varying treatment (27). Furthermore, we focussed on linear regression models, but our recommendations for exposures and confounders will also hold for regression models with a binary or survival outcome.

In summary, the optimal strategy for handling measurements affected by medication depends on whether the medication effect is on the exposure, the outcome, or a confounder. When deciding which strategy to use, we urge researchers to critically consider the processes of medication prescription and what information on medication effects is available.

References

1. 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.
2. 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.
3. Tanamas SK, Hanson RL, Nelson RG, et al. Effect of different methods of accounting for antihypertensive treatment when assessing the relationship between diabetes or obesity and systolic blood pressure. Journal of Diabetes and its Complications 2017;31(4):693-9.
4. Cui JS, Hopper JL, Harrap SBJH. Antihypertensive treatments obscure familial contributions to blood pressure variation. 2003;41(2):207-10.
5. Hunt Steven C, Ellison RC, Atwood Larry D, et al. Genome Scans for Blood Pressure and Hypertension. Hypertension 2002;40(1):1-6.
6. 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.
7. 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.
8. NEO codebook baseline measurements. (https://www.lumc.nl/org/neo-studie/researchers/variables/codebook-neo1/medicalhistory/). (Accessed).
9. de Mutsert R, den Heijer M, Rabelink TJ, et al. The Netherlands Epidemiology of Obesity (NEO) study: study design and data collection. European journal of epidemiology 2013;28(6):513-23.
10. NHG-Standaaren. Cardiovasculair risicomanagement (CVRM). 2019. (https://www.nhg.org/standaarden/samenvatting/cardiovasculair-risicomanagement). (Accessed).
11. Neaton JD, Grimm RH, Jr, Prineas RJ, et al. Treatment of Mild Hypertension Study: Final Results. JAMA 1993;270(6):713-24.
12. Moons KGM, Donders RART, Stijnen T, et al. Using the outcome for imputation of missing predictor values was preferred. Journal of Clinical Epidemiology 2006;59(10):1092-101.
13. Therneau TM, Lumley T. Package 'survival'. 2015;128:112.
14. Henningsen A, Toomet O, Henningsen MA, et al. Package 'sampleSelection'. 2019.
15. Koenker R, Portnoy S, Ng PT, et al. Package 'quantreg'. 2019.
16. Buuren Sv, Groothuis-Oudshoorn K. mice: Multivariate imputation by chained equations in R. Journal of statistical software 2010:1-68.
17. Galimard J-E, Chevret S, Protopopescu C, et al. A multiple imputation approach for MNAR mechanisms compatible with Heckman's model. Statistics in Medicine 2016;35(17):2907-20.
18. Hernán M, Robins J. Causal inference: What if. Boca Raton: Chapman & Hall/CRC; 2020.
19. Hernán M, S H-D, Robins J. A Structural Approach to Selection Bias. Epidemiology 2004;15(5):615-25.
20. Cole SR, Platt RW, Schisterman EF, et al. Illustrating bias due to conditioning on a collider. International Journal of Epidemiology 2009;39(2):417-20.

21. Stolzenberg RM, Relles DA. Theory Testing in a World of Constrained Research Design:The Significance of Heckman's Censored Sampling Bias Correction for Nonexperimental Research. 1990;18(4):395-415.

22. Lillard L, Smith JP, Welch F. What Do We Really Know about Wages? The Importance of Nonreporting and Census Imputation. 1986;94(3, Part 1):489-506.

23. White IR. Commentary: Dealing with measurement error: multiple imputation or regression calibration? International Journal of Epidemiology 2006;35(4):1081-2.

24. Brakenhoff TB, van Smeden M, Visseren FLJ, et al. Random measurement error: Why worry? An example of cardiovascular risk factors. PLOS ONE 2018;13(2):e0192298.

25. 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.

26. 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.

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

Appendix 1

Detailed description of methods for handling medication effect

We consider the situation where a linear relationship between variables when no one is affected by medication is of interest and a variable is affected by medication use (e.g., blood pressure affected by antihypertensive drugs) for some individuals. For convenience, we assume people take medication when values are high, aiming to lower the values. Depending on the research question, the variable(s) affected by medication use can be the exposure, a confounder, or the outcome in an analysis. The different methods to handle medication use are :

NAÏVE METHODS

M1. ***Ignoring medication use***
Measurements affected by medication are used in the analysis as they are observed.

M2. ***Selecting untreated individuals***
Only the individuals who are not receiving medication are included in the analysis.

M3. ***Adjusting for medication use by adding a binary indicator variable to the regression model***
An indicator for medication use is added as a covariate in the regression model.

M4. ***Substituting measurements of treated individuals with a fixed value***
As Hunt et al. (1) suggested, measurements affected by medication are substituted with a pre-specified value. For example, when guidelines indicate that antihypertensive durgs should be prescribed for blood pressures over 140 mmHg, a value higher than 140 mmHg can be used as a substitution.

M5. ***Adding a constant value to observations of treated individuals***
When the effect of medication on the variable of interest is approximately known, the mean treatment effect can be added to the observed measurements of treated individuals (2, 3). For blood pressure, for example, some authors added 10 mmHg to the systolic blood pressure and 5 mmHg to the diastolic blood pressure when individuals are using antihypertensive medication (4, 5). These values were based on known average treatment effects from a clinical trial (6). However, this is not a set rule and could be adapted

METHODS FOR A MEDICATION EFFECT ON EXPOSURE

M6. ***Regression calibration***
A vast amount of literature addresses measurement error in the covariates of a regression model (7, 8). A simple method is regression calibration, where the untreated

values of the treated individuals (thus, unobserved) replace the measurements affected by medication. The untreated values are estimated by the observed values and other covariates. The method needs an educated guess of the mean *and* standard deviation of the medication effect. These may be obtained from previous clinical trials or observational studies where the effect of treatment is studied.

For individuals on medication, their observed measurement X is replaced by $\lambda(X-\overline{X})+\overline{X}$ + mean medication effect ; with $\overline{X}$, the mean value of X for those using medication and λ, so-called reliability ratio (9). The reliability ratio is equal to $\lambda = 1 - SD(med)^2/SD(X|Z)^2$; with *SD(med)*, the standard deviation of the medication effect and *SD(X|Z)*, the standard deviation of *X* for the medication users adjusted for *Z*, a set of other covariates in the regression model.

METHODS FOR A MEDICATION EFFECT ON THE OUTCOME

M7. **Inverse probability weighting (Sampling weights)**

In this approach, treated individuals are removed from the analysis, and more weight is given to untreated individuals with a similar profile (10, 11). First, the probability of receiving medication for each individual is estimated by logistic regression. Then the untreated individuals are weighted by 1/(1-probability to receive medication). This creates a pseudo-population with the same characteristics as the original population but where no one is treated.

M8. ***Quantile regression***

White et al. (12) proposed to use quantile regression for outcomes affected by medication use. In this approach, the median outcome is modeled as a function of covariates. The method assumes the untreated values would have been above the median conditional on covariates for individuals on medication. The treated individuals' outcome values are replaced by *k,* that is, any value higher than the conditional median, after which a median regression model can be fitted.

M9. ***Censored normal regression***

An alternative approach is to use methods for censored outcomes (2, 3), such as censored normal regression, which assumes a normal underlying distribution of the untreated outcome. This method is also known as tobit regression. Measurements of treated individuals are considered to be censored observations, where the untreated values are assumed to be at least as high as the observed values affected by treatment. An advantage of this method is that no assumptions on the treatment effect size are needed. However, non-informative censoring is assumed. The non-informative censoring implies that conditional on covariates, the probability of receiving treatment does not depend on the untreated values. This assumption is likely to be invalid, as individuals with higher values are more likely to be treated. Previous simulations showed good

performance in realistic scenarios (2). However, recent literature showed that the method performed poorly under certain scenarios (13).

More complex censoring mechanisms can also be used to resemble realistic clinical settings. For example, when a clinical guideline suggests starting treatment for values above a certain threshold δ this information can be incorporated. In this case, the untreated values are assumed to be higher than the observed measurements *and* higher than the threshold. That is, for the treated observations, we assume that:

$$\begin{cases} Untreated\ value\ \geq \delta, & if\ observed\ value < \delta \\ Untreated\ value\ \geq observed\ value, & if\ observed\ value \geq \delta \end{cases}$$

The threshold value of δ is obtained using external knowledge of the clinical setting.

M10. ***Heckman's treatment effects model***

Heckman's treatment effects model originates from economics and can account for non-random sample selection (13-15). Spieker et al. (13, 15) used this model for handling outcomes affected by medication use. This model assumes that treatment assignment depends on the untreated values where higher values are more likely to be treated and treatment results in a "structural shift" of the mean outcome. In the standard treatment effect model, this treatment effect does not depend on covariates (13), but it is possible to extend this model to incorporate effect modification (15).

Technically, the method assumes that there is an unobserved latent variable that determines treatment. If its value is above 0, treatment is prescribed. The latent variable is correlated with the original untreated values, so people with higher untreated values are more likely to be treated. Parameters are estimated by joint modeling of i) a linear regression model for the effect of exposure on the untreated blood pressure, ii) the same linear regression model for the effect of exposure on the treated blood pressure, with a lower constant term which reflects the effect of treatment and iii) a probit model for the probability of medication prescription (13, 15). Both the linear regression model and the probit model may depend on other covariates.

MULTIPLE IMPUTATION APPROACHES

Untreated values of individuals on treatment can be considered missing, and multiple imputation methods can be used to handle these missing values. The method can be applied in many different ways under different assumptions. We considered three multiple imputation approaches that are based on various assumptions.

M11. ***Multiple imputation with predictive mean matching via a linear regression model***

For a numerical variable with missing values, the default multiple imputation option is chained equations with predictive mean matching via a linear regression model with

the main effects of the covariates. This imputation method is readily available in many standard statistical software packages. Note that the method assumes that the data is missing at random.

M12. ***Multiple imputation with censored normal regression***

Instead of using linear regression as imputation model, censored normal regression may be used to predict missing values (16). This may be done under the different censoring mechanisms we discussed in [M9]. A regular censored normal regression can only be used when medication effect is on the outcome. However, multiple imputation with censored normal regression does not have this restriction.

M13. ***Multiple imputation with Heckman's model***

Galimard et al. developed an imputation approach for missing not at random data using Heckman's model (17). Again, the multiple imputation approach can be used regardless of whether the outcome, exposure, or a confounder is affected.

References

1. Hunt Steven C, Ellison RC, Atwood Larry D, et al. Genome Scans for Blood Pressure and Hypertension. Hypertension 2002;40(1):1-6.
2. 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.
3. 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.
4. Cui J, Hopper JL, Harrap SBJH. Genes and family environment explain correlations between blood pressure and body mass index. 2002;40(1):7-12.
5. Cui JS, Hopper JL, Harrap SBJH. Antihypertensive treatments obscure familial contributions to blood pressure variation. 2003;41(2):207-10.
6. Neaton JD, Grimm RH, Jr, Prineas RJ, et al. Treatment of Mild Hypertension Study: Final Results. JAMA 1993;270(6):713-24.
7. Carroll RJ. Measurement Error in Epidemiologic Studies. Wiley StatsRef: Statistics Reference Online, 2014.
8. Buonaccorsi JP. Measurement error: models, methods, and applications. Chapman and Hall/CRC; 2010.
9. Carroll RJ, Ruppert D, Stefanski LA, et al. Measurement error in nonlinear models: a modern perspective. Chapman and Hall/CRC; 2006.
10. 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.
11. Hernán M, Robins J. Causal Inference. Harvard TH Chan School of Public Health (July 27, 2019).
12. 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.

13. 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.
14. Certo ST, Busenbark JR, Woo H-s, et al. Sample selection bias and Heckman models in strategic management research. Strategic Management Journal 2016;37(13):2639-57.
15. 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.
16. Gartner H, Rässler S. Analyzing the changing gender wage gap based on multiply imputed right censored wages. 2005.
17. Galimard J-E, Chevret S, Protopopescu C, et al. A multiple imputation approach for MNAR mechanisms compatible with Heckman's model. Statistics in Medicine 2016;35(17):2907-20.

Appendix 2

Detailed application of the methods

1) Covariates used for inverse probability weighting [M7] and the imputation methods [M11-M13]

Sex, age, BMI, total body fat, waist circumference, hip circumference, education level, income, smoking status, ethnicity, alcohol intake, the total amount of leisure, glucose, insulin, glycated hemoglobulin A1C, triglycerides, HDL cholesterol, LDL cholesterol and medication use for glucose, lipid, and depression.

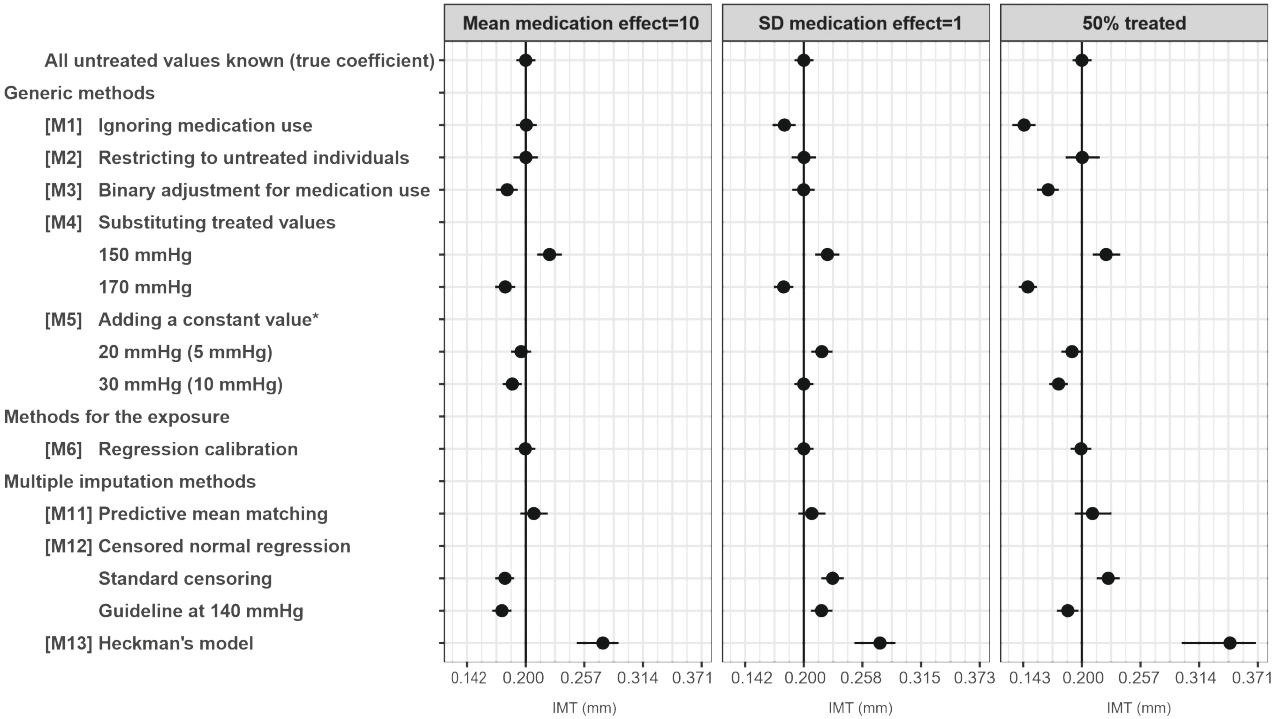
2) R syntax for censored normal imputation [M12]

```
# Function to draw from a truncated normal distribution, range lwb-upb
rnorm.trunc <- function(n,mean,sd, low=-Inf, upp=Inf)
{U <- runif(n,0,1)
qnorm(pnorm(low, mean = mean, sd = sd)+
        (pnorm(upp, mean = mean, sd = sd)-pnorm(low, mean = mean, sd = sd))*U, mean = mean, sd
= sd)
}

# impute censored normal
mice.impute.censnorm <-
  function (y, ry, x, wy = NULL,ycens, ...)
  {
    #1 prepare data
    wy <- !ry # wy= TRUE indicates that value should be imputed
    x <-  as.matrix(x)
    m <- ncol(x)+1

    # 2. estimate coefficients censored model
    fit <- survreg(Surv(ycens, ry) ~ x, dist='gaussian')
    beta <- coefficients(fit)
    sigma <- fit$scale
    #    print(fit)

    #3. generate new beta and sigma for bayesian drawings
    df <- max(length(y[ry]) - ncol(x), 1)
    rv <- t(chol((vcov(fit)[1:m,1:m])))
    beta.star <- beta + rv %*% rnorm(ncol(rv))
    sigma.star <- sqrt(df*sigma^2/rchisq(1, df))

    #4. Draw new observations
    mean.star <- cbind(1,x[wy, , drop = FALSE]) %*% beta.star
    vec<- rnorm.trunc(nrow(mean.star),mean.star,sd=sigma.star, low=ycens[wy])
    return(vec)
  }
```

Appendix 3

Regression coefficients and 5th and 95th percentile estimated from modified scenarios of Simulation setting 1, where we estimated the effect of SBP on IMT measurement. One grid unit represents 2.5 standard deviation of the true coefficient, estimated from 1000 simulation runs. *We added 5 mmHg and 10 mmHg for the scenario where mean medication effect was 10 mmHg,

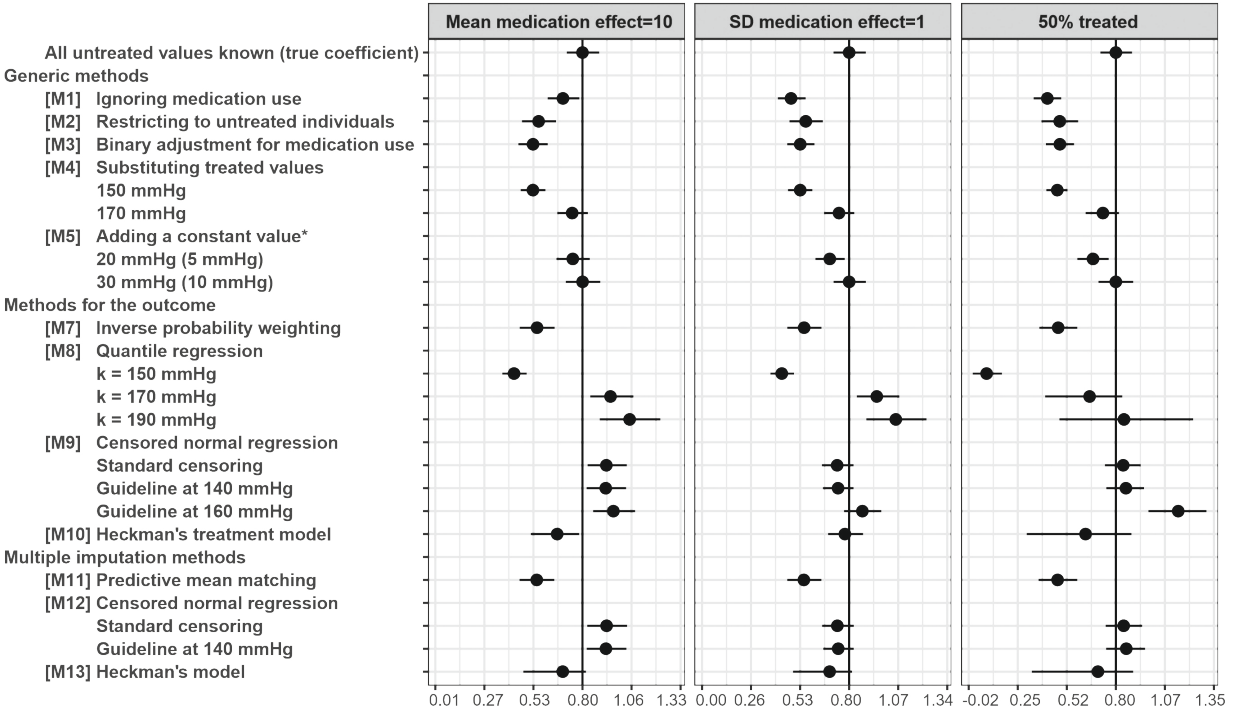
Exposure affected by medication (modified Simulation setting 1)



Appendix 4

Regression coefficients and 5th and 95th percentile estimated from modified scenarios of Simulation setting 2, where we estimated the effect of BMI on SBP. One grid unit represents 2.5 standard deviation of the true coefficient, estimated from 1000 simulation runs. *We added 5 mmHg and 10 mmHg for the scenario where mean medication effect was 10 mmHg,

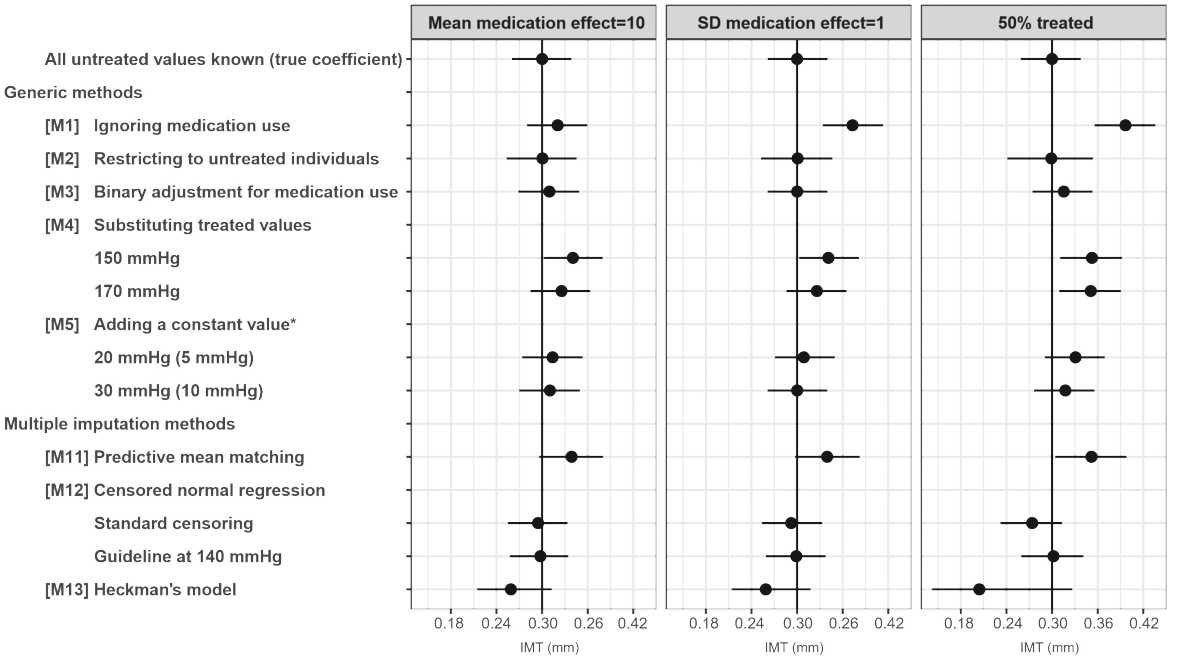
Outcome affected by medication (modifed Simulation setting 2)



Appendix 5

Regression coefficients and 5th and 95th percentile estimated from modified scenarios of simulation setting 3, where we estimated the effect of BMI on IMT while SBP is one of the confounders. One grid unit represents 2.5 standard deviation of the true coefficient, estimated from 1000 simulation runs. *We added 5 mmHg and 10 mmHg for the scenario where mean medication effect was 10 mmHg,

Confounder affected by medication (modifed Simulation setting 3)



Chapter 5

Estimating medication effects using routinely collected electronic health records: changes in blood glucose and HbA1c levels after glucose-lowering medication prescription in the Netherlands Epidemiology of Obesity study participants

Ready to be submitted

Jungyeon Choi, Renée de Mutsert, Jeroen van der Velde, Ester van Eekelen, Olaf M. Dekkers, Saskia le Cessie

Abstract

Background

Measurements affected by medication use, such as glucose levels alleviated after glucose-lowering medication, are commonly encountered in epidemiological studies. Potential methods for validly handling these measurements affected by medication use are incorporating the information of the mean medication effect and, sometimes, its standard deviation. In this study, we aim to describe changes in blood glucose and HbA1c levels after glucose-lowering medication prescription from routinely collected data.

Method

Participants from the Netherlands Epidemiology of Obesity (NEO) study who developed type 2 diabetes during the follow-up period were included. The patients were identified using general practitioners' Electronic Patient Records (EPR). The EPRs were also used to obtain repeated measurements of blood glucose and HbA1c. We fitted linear mixed models with glucose and HbA1c as the outcomes. Time as a categorical variable was added as a fixed effect and random effect.

Results

In total, 127 incident diabetes cases were included in the analyses. In general, we observed a sharp increase in glucose and HbA1c levels shortly before the medication prescription. After the prescription, levels of both decreased. The lowest values were observed at 6-12 months after prescription, which were 1.76 mmol/L lower in glucose [CI: -2.54, -0.99] and 0.80% lower in HbA1c [CI: -1.61, -0.45] than 6-12 months before prescription. After one year, glucose and HbA1c levels increased, but even after two years, levels were significantly lower than before starting medication. Variation in medication effect between individuals was large.

Conclusion

The sharp increase in glucose and HbA1c shortly before medication prescription likely reflects random high values. Considering a longer period before the medication prescription is needed to obtain a better estimate of the medication effect. The estimated medication effects were smaller than observed in RCTs, yet on average, treatment remained effective for more than two years after prescription. Routinely collected data can provide insights into medication effects in the real-world which may not be easily obtained from RCTs.

1. Introduction

Population-based observational studies are often used to provide insight into the real-world relationships between clinical measurements and the effects of various treatments. Population-based studies, by their nature, include a wide range of individuals with various clinical features. Thus, in a population-based cohort, it is common to encounter that some measurements are affected by medication use in a subgroup of the study population. Examples are cholesterol levels controlled by cholesterol-lowering medication or blood pressure levels lowered by antihypertensive medication.

Glucose-lowering medication is a commonly used treatment for (pre)diabetes to regulate blood glucose levels. It was recently reported that 10.2% of the US population had diabetes (1). From 2007 to 2010, 88% of people aged ≥ 20 years with diagnosed diabetes were reported to be treated with insulin and/or oral medications (2). In the database of the UK Biobank, a widely known prospective cohort recruited from the general UK population aged 40–69 years (3), approximately 4% reported using glucose-lowering medication for type 2 diabetes (T2D) (4).

Medication use is not of concern when one is interested in measurements as observed regardless of whether medication is used. Sometimes, however, researchers may be interested in the measurements if untreated so that the estimated result would reflect the natural relationship between the variables of interest. However, the untreated values cannot be observed for those who are on medication. Consequently, appropriate methods to correctly adjust for medication effects would be needed.

Several studies have shown that effect estimates can be substantially biased if the measurements affected by medication use are handled with invalid methods (5-8). When the affected measurement is an outcome variable, adding an estimated mean medication effect to the treated values is an appropriate method (5, 7). When the exposure is affected, a valid method could be a regression calibration approach (7). To apply these methods, information on the mean (and standard deviation) of the medication effect, acquired from external information, is needed.

The mean medication effect and standard deviation may be acquired from randomized controlled trials (RCTs). Meta-analyses of RCTs on glucose-lowering medication showed that using a single type of medication reduced HbA1c levels by, on average, 0.66% to 1.11% (values aligned to the assay used in the Diabetes Control and Complication Trial; DCCT), depending on the drug classes (9) or approximately 1% over the course of the studies (10, 11). Trials on glucose-lowering medication found an effect of 2-4 mmol/L lowering blood glucose on average (12-14).

Although the effects of glucose-lowering medication are known from RCTs, these may not reflect how blood glucose and HbA1c levels change before and after medication prescription in real-world settings. Populations eligible for trials may not represent the population of interest in an observational cohort study. Eligibility criteria and the recruitment process of population-based cohort studies could be vastly different from RCTs, where study participants are usually recruited in a restrictive manner. Furthermore, randomization of treatment by no means reflects how medications are prescribed and administered in real-world settings. Additionally, follow-up in RCTs generally starts shortly before or at the start of the prescription, and the follow-up period is often less than one year (10), providing limited information on long-term medication effects. A possible approach to circumvent these issues is to estimate the medication effect directly from the population of interest in a real-world setting.

In this study, we explore how observational routinely collected data can be used to describe and estimate changes in blood glucose and HbA1c levels change over time before and after glucose-lowering medication prescription. Therefore, we use data from the Netherlands Epidemiology of Obesity (NEO) study and its follow-up data routinely collected by general practitioners. Using these data, we estimate the effect of medication use on blood glucose and HbA1c levels and discuss the results and the advantages and pitfalls of using observational routinely collected data to estimate the effect of glucose-lowering treatments.

2. Method

Study population

The Netherlands Epidemiology of Obesity (NEO) study is a population-based prospective cohort study designed to investigate pathways leading to obesity-related conditions and diseases. The study recruited men and women aged 45 to 65 years living in the greater area of Leiden, the Netherlands, with an oversampling of individuals with a BMI of 27 kg/m^2 or higher. Details of the design and inclusion criteria of the NEO study can be found elsewhere (15). The first wave of data collection started in September 2008 and was completed in September 2012.

Follow-up of the NEO study participants

During the follow-up of the NEO study participants, clinical endpoints were collected thorough electronic patient records (EPR) of general practitioners (GP). The EPR contains basic data of care provided and recorded by general practitioners, such as disease diagnosis, treatment prescription, test results, and referrals. The records are encoded with International Classification of Primary Care (ICPC) codes. Medication prescriptions are coded according to the Anatomical Therapeutic Chemical (ATC)

classification. We used the EPRs extracted in October 2017 - May 2018 to obtain repeated measurements of blood glucose and HbA1 and diagnosis of T2D of the NEO study participants.

From the EPR, those who were not diagnosed with T2D at the first NEO visit but were diagnosed during the follow-up (i.e., incident diabetes cases) were identified. Ascertainment of T2D was performed based on three components: 1) the presence of ICPC code T90 or T90.02, and/or 2) a prescription of glucose-lowering medication, defined by ATC codes starting with A10, and/or 3) the presence of keywords for glucose-lowering medication, such as insulin, metformin, or any generic names in free text (complete list of keywords is provided in a Appendix 1). The general practitioner was contacted if it remained unclear whether a participant was correctly diagnosed with T2D. We then excluded participants i) whose medication prescription date was unknown, ii) who did not have blood test results for glucose or HbA1c, or iii) whose blood test results were only available more than 12 months before the first medication prescription date.

Statistical analysis

HbA1c levels were standardized to HbA1c DCCT (%) values. Biologically unrealistic low values (HbA1c < 4% or blood glucose=0 mmol/L) were set to be missing. Time was centralized to the first prescription date of the antidiabetic treatment (time 0: date of the first prescription).

Descriptive statistics of the participants' characteristics at the NEO visit were presented as the mean and standard deviation for continuous variables and frequencies for categorical variables. To explore the change in glucose and HbA1c over time, we used spaghetti plots to display individual-level data and box plots to visualize all data.

We fitted linear mixed models to estimate changes in glucose and HbA1c levels over time after starting medication. Dependent variables were repeated blood glucose and HbA1c measurements. Time was added as a fixed effect with the following categorization: (up and until) 6 to (less than) 12 months *before* the first prescription/ 3 to 6 months *before*/ 0 (including the date of a first prescription) to 3 months *before*/ (more than) 0 to (up an until) 3 months *after*/ 3 to 6 months *after*/ 6 to 12 months *after*/ 12 to 24 months *after*/ more than 24 months *after*. Categorization was done such that the mean value at each time category contrasts with the mean value at 6 to 12 months *before* the prescription. As random effects, we added a random person effect plus a random effect for different periods after medication prescription categorized as follows: (more than) 0 to 3 (up and until) months *after the first prescription*/ 3 to 6 months *after*/ 6 to 12 months *after*/ 12 to 24 months *after*/ more than 24 months *after*. Figure 1 visualizes the timeline of the glucose and HBA1c measurements and the NEO visit.

We further explored whether the mean changes in glucose and HbA1c after medication prescription were dependent on age, BMI, or sex. For this, we fitted three models, where we respectively added BMI at the NEO visit (continuous), sex, or age at the first prescription date (continuous) as fixed effects, with an interaction term with medication prescription.

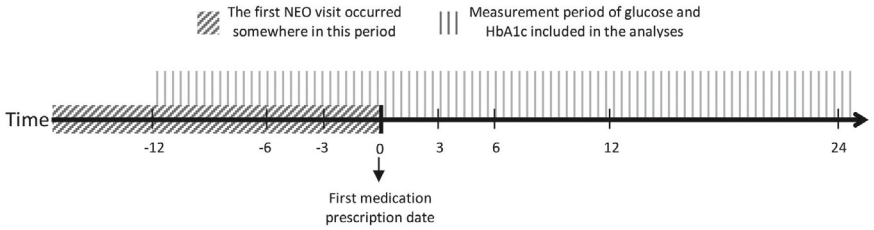


Figure 1. Timeline of the glucose and HBA1c measurements and the NEO visit. Time 0 is the date of the first prescription. The date of the first NEO visit, which was before time 0, varies between individuals (for some individuals, it was less than 12 months before the first prescription). In the analyses, glucose and HbA1c measures from 12 months before the first prescription were used.

3. Results

In total, 6671 individuals were included in the NEO study. Among the participants who did not use any antidiabetic medication at the NEO visit, 297 participants were identified as incident type 2 diabetes cases from the EPR. Participants who did not have information on the medication prescription date (n=126), did not have laboratory measurements for glucose or HbA1c (n=41), or had only laboratory measurements more than 12 months before the medication prescription (n=3) were excluded. In total, 127 individuals remained. The mean number of repeated measurements for blood glucose was 12 (IQR: [7, 20]; maximum: 105), and for HbA1c was eight (IQR: [5, 14]; maximum: 58). Figure 2 illustrates a flow chart of the selection of the study sample.

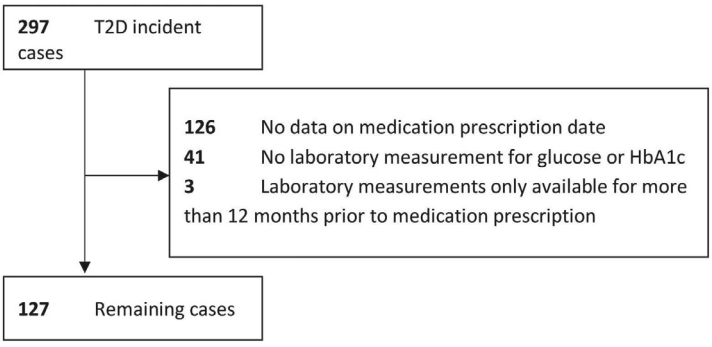


Figure 2. Sample selection process and the number of individuals included in the analyses

Table 1 summarizes the general characteristics of the 127 individuals measured at the first visit of the NEO study. Mean fasting glucose (7.0 mmol/L, SD: 1.8), HbA1c level (6.0 %, SD: 0.9), and HOMA-IR (6.0, SD: 3.8) indicated that a large number of the included participants were already prediabetic, defined as fasting glucose level between 5.6–6.9 mmol/L or HbA1c level between 5.7–6.4% (18), at the first NEO visit. The selected individuals also had high mean BMI (33.6 kg/m2, SD: 5.4), and many were hypertensive (50%). Time from first measurement to prescription varied largely between individuals (121 days, IQR: [7, 260]). Types of first-prescribed glucose-lowering medication are summarized in Table 2. Metformin was most often prescribed as the first glucose-lowering medication.

Table 1. Participants' characteristics at the first NEO visit. Mean and standard deviation was used for continuous variables [mean (sd)]. Frequency and percentage were used for categorical variables [N (%)]

Measurements at the NEO visit (N=127)	
Sex	
Male	65 (51.2%)
Age in years	56.0 (5.8)
Age in years at the date of prescription	60.6 (6.2)
Education	
High	50 (39.4%)
Hypertension	
Yes	63 (49.6%)
BMI (kg/m2)	33.6 (5.4)
Glucose (mmol/L)	7.0 (1.8)
Insulin (mU/L)	19.5 (11.8)
HOMA1-IR	6.0 (3.8)
HbA1c (%)	6.0 (0.9)
Total cholesterol (mmol/L)	5.7 (1.1)
Triglycerides (mmol/L)	1.9 (1.3)
HDL (mmol/L)	1.3 (0.3)
LDL (mmol/L)	3.6 (1.0)
Lipid-lowering drugs use	
Yes	19 (15.0%)
Hypertension drugs use	
Yes	54 (42.5%)
Time from NEO visit to prescription (days)	121 (IQR: 7, 260)

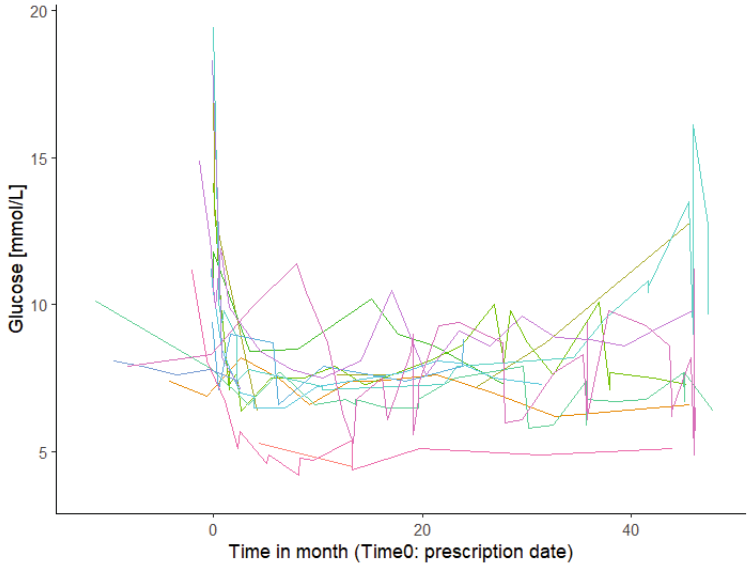
Table 2. Types of medication prescribed at the first prescription date for 127 individuals

Prescribed medication types*	**Frequency**
Gliclazide	2
Glimepiride	1
Insulin injection pen	4
Metformin	117
Sitagliptin	1
Tolbutamide	3
Others/ unknown	5
Total	**133**

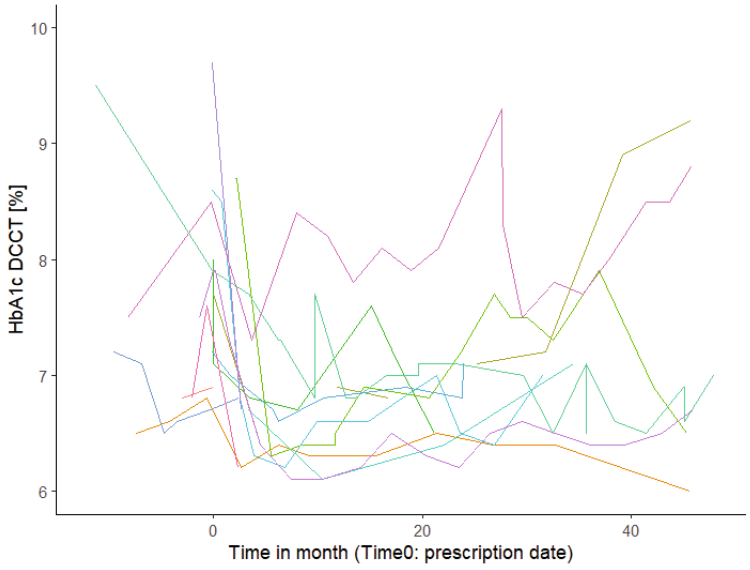
* Six individuals have been prescribed two types of medication on their first prescription date.

Appendix 2 compares the characteristics of the individuals included (n=127) and excluded (n=170) from the analyses. Glucose and HbA1c levels were on average lower (0.4 mmol/L and 0.2%, respectively) in the excluded individuals compared to those included.

Figure 3 displays changes in blood glucose and HbA1c levels of 15 randomly chosen individuals, showing considerable variation in patterns over time. The observed overall means over time are visualized in Figure 4. The figures indicate that glucose and HbA1c levels peaked near the first medication prescription date.

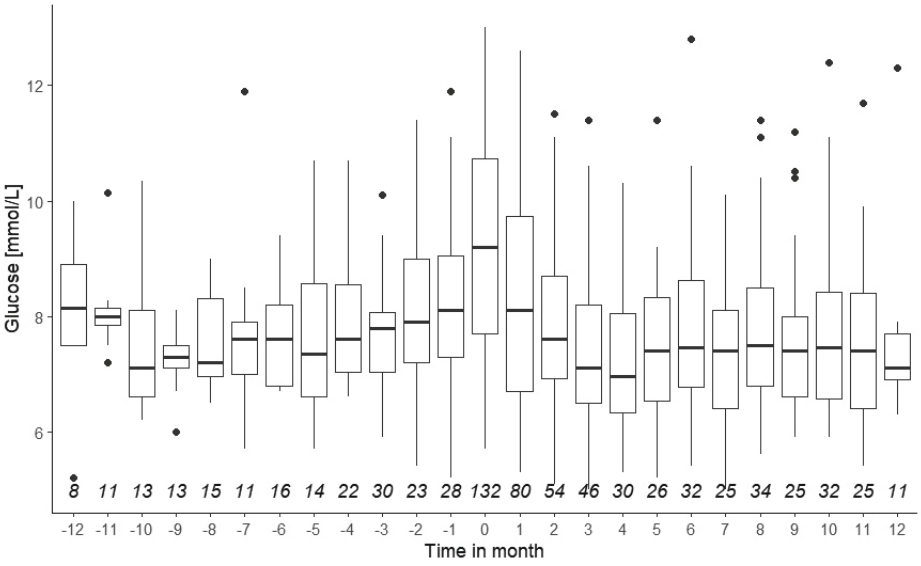


3a.

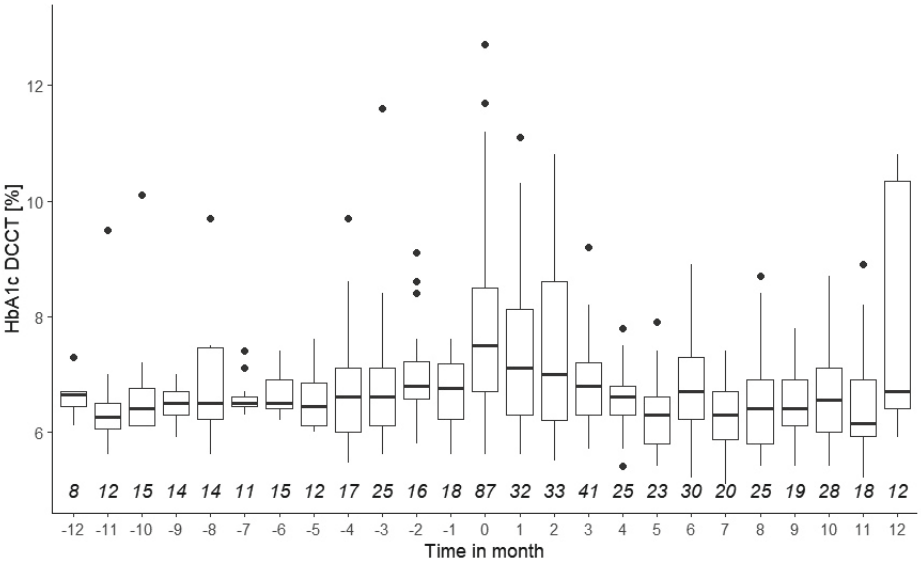


3b.

Figure 3. Spaghetti plots for glucose (a) and HbA1c (b) levels of 15 randomly chosen individuals



4a.



4b.

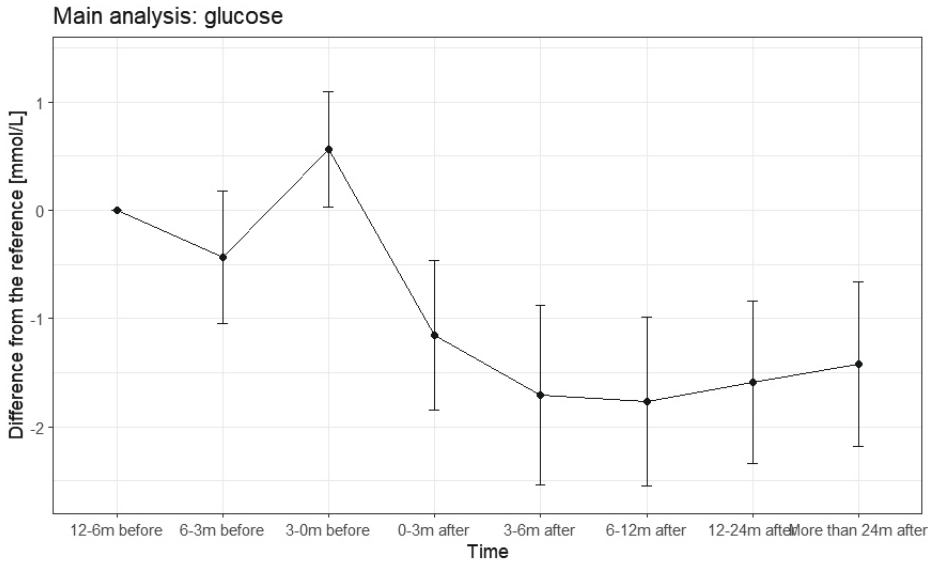
Figure 4. The median and interquartile range of glucose (a) and HbA1c (b) measurements at each time on a monthly scale. Numbers in italic fonts represent the number of available observations at each point.

Main analyses

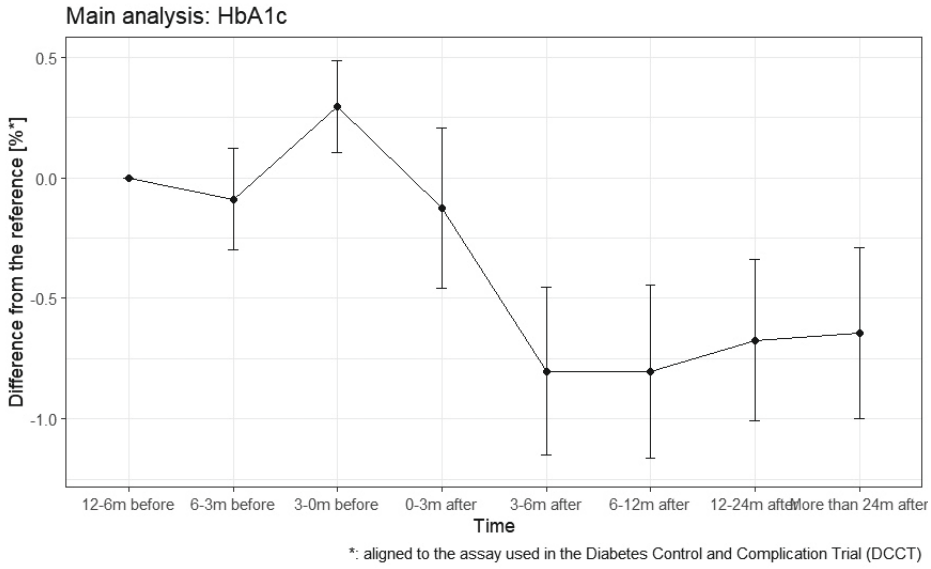
Table 3 summarizes the results of the fitted linear mixed models. Figure 5 visualizes the estimated mean differences in glucose and HbA1c at each time category compared to the level at 6-12 months *before* the prescription and their confidence intervals.

For glucose, the mean level increased shortly before the first medication prescription; that is, the level at 0 to 3 months before prescription was 0.56 mmol/L *higher* [CI: 0.03, 1.10] compared to 6-12 months before prescription. The mean level decreased in 0-3 months *after* prescription, which was 1.15 mmol/L *lower* [CI: -1.85, -0.46] than 6-12 months before prescription. The decrease in glucose levels was the largest 6-12 months *after* prescription; the levels were on average 1.76 mmol/L *lower* [CI: -2.54, -0.99] compared to 6-12 months before prescription. Glucose levels slightly increased after 12 months and more than 24 months after prescription, the level was after 24 months 1.42 mmol/L *lower* [CI: -2.18, -0,66] than 6-12 months before prescription. The effect of medication on glucose varied largely between the individuals, with the standard deviation equal to 2.30 mmol/L at 0-3 months *after* prescription and 3.09 mmol/L at more than 24 months after prescription. Between-individual before medication and within-individual variability were also relatively large (SD: 3.23 mmol/L and 1.63 mmol/L, respectively).

The trend was similar for the HbA1c measurements. The mean level at 0-3 months *before* the prescription was 0.30% *higher* [CI: 0.10, 0.49] than 6-12 months before prescription. The HbA1c level decreased after the prescription. The largest decrease was shown at 3-6 months *after* prescription, which was 0.80% lower than 6-12 months *before* prescription, [CI: -1.15, -0.45] and 6-12 months *after prescription* [CI: -1.16, -0.45]. The mean HbA1c level slightly increased at later time points. At more than 24 months *after* prescription, the mean level was 0.65% *lower* [CI: -1.00, -0.29] than 6-12 months before treatment. Variations in the prescription effect were large and tended to be larger when the follow-up time increased.



5a.



5b.

Figure 5. Change in glucose (a) and HbA1c (b) levels at each time point compared to the levels at 6 to 12 months before medication use

Table 3. Results of fitting linear mixed models where the outcome was glucose or the HbA1c measurement. Time as a categorical variable was added as a fixed effect, and medication use was added as a random effect

	Glucose (mmol/L)	HbA1c (%)
Fixed effects (estimate [CI])		
Intercept (mean at time 0)	9.45 [8.71, 10.18]	7.36 [7.05, 7.67]
6 - 12m before prescription	-	-
3 - 6m before prescription	-0.44 [-1.05, 0.18]	-0.09 [-0.3, 0.12]
0 - 3m before prescription	0.56 [0.03, 1.10]	0.30 [0.10, 0.49]
0 - 3m after prescription	-1.15 [-1.85, -0.46]	-0.13 [-0.46, 0.21]
3 - 6m after prescription	-1.71 [-2.54, -0.88]	-0.80 [-1.15, -0.45]
6 - 12m after prescription	-1.76 [-2.54, -0.99]	-0.80 [-1.16, -0.45]
12 - 24m after prescription	-1.59 [-2.34, -0.84]	-0.67 [-1.01, -0.34]
More than 24m after prescription	-1.42 [-2.18, -0.66]	-0.65 [-1.00, -0.29]
Random effects (SD)		
Between-person variation	3.23	1.45
Variation in the mean difference		
at 0 - 3m after prescription	2.30	1.24
at 3 - 6m after prescription	3.30	1.52
at 6 -12m after prescription	3.06	1.61
at 12 - 24m after prescription	3.07	1.51
More than 24m after prescription	3.09	1.55
Within-person variation	1.63	0.53

Interaction effects

When adding *BMI at the first visit* and its interaction with medication use in the model, we observed that people with higher BMI at the NEO visit had a larger decrease in both glucose and HbA1c levels after medication prescription; 0.13 mmol/L [CI: 0.03; 0.23] *lower* for glucose and 0.05% *lower* [CI: 0.00; 0.09] for HbA1c per 1kg/m^2 increase in BMI. We did not observe an interaction effect between *sex* and medication use. When adding age at the first prescription date, we observed different directions of the effect for HbA1c and glucose; people who were older, on average, had a higher decrease in glucose but a smaller decrease in HbA1c after medication prescription.

4. Discussion

This study explored changes in blood glucose and HbA1c levels before and after glucose-lowering medication prescription from observation study data. We used the data from the NEO study and the routinely collected electronic health record data of its participants. We observed that glucose and HbA1c levels sharply increased shortly before prescription. The decrease in the outcome levels was the largest at 6-12 months after prescription; on average, 1.76 mmol/L lower in glucose and 0.80% lower in HbA1c compared to 6-12 months before starting medication. After one year, glucose and HbA1c levels increased slightly. The levels of both, however, remained significantly lower than before medication use. Similar to previous studies, we observed considerable within-person variations (17). The effect of medication on glucose and HbA1c varied largely between the individuals. The effects of the medication were larger in people with higher BMI.

Our results showed that the estimated medication effect depends on the period before medication use is chosen as a reference. The effect of medication would seem larger when considering 0-3 months before medication use as the reference, because an increase in glucose and HbA1c levels was observed shortly before medication prescription. It is known that the variability of glucose is large (17), and it might have occurred that a randomly high measurement of blood glucose level led to a decision to prescribe medication for some patients. Only comparing the last measurement before the start of medication to the measurements after the start of medication could lead to a regression to the mean effect, i.e., the phenomenon that extreme observations are followed by observations closer to the mean (19, 20). Thus, to not overestimate the medication effect, it is important to consider trends in measurements over a longer period. In our study, HbA1c and glucose measurements in 6-12 months before the treatment seemed to better reflect the clinical condition of an individual compared to the measurements shortly before the medication prescription and thus more appropriate to be set as the reference.

Advantages and pitfalls of estimating the effect of medication use from observational data

The average reduction in glucose and HbA1c after the first prescription estimated in our study was somewhat lower than the medication effects obtained from RCTs, which varied between 2-4mmol/L lower values for glucose and 0.66-1.5% for HbA1c depending on medication types (10, 12-14). These discrepancies may reflect the differences between observational settings and RCTs.

Compared to RCTs, routinely collected data better reflects how the population of interest behaves in practice, which has a consequence on the effectiveness of medication in the real world. It is known that the adherence rate of the routinely administered oral

treatment for chronic diseases, such as diabetes, is low (25, 26), likely leading to a lower average reduction in glucose and HbA1c in this study than shown in RCTs. Such tendency was also, in part, reflected when looking at long-term effects. Although the mean levels also after the first year remained significantly below the levels before medication use, the levels increased after the first year of medication use. Between-individual behavioral variations might have contributed to the large standard deviations in medication effect. As the variability in the real world is well-represented, routinely collected data may provide more realistic information about the medication effect in one's population of interest.

However, several considerations should be made when utilizing routinely collected data. The main concern is that clinical decisions made in real-world settings are often not clearly known and the recording of data may have been done selectively or inaccurately. For instance, some individuals in our study had much more frequent measurements of glucose and HbA1c than others. This suggests that a selected group of T2D patients were much more closely monitored than others.

We also observed that many individuals did not have information on medication prescriptions even though they were identified as type 2 diabetes patients. This could indeed reflect the diagnosis and prescription process of the real world. In the Netherlands, for example, the initial action of a GP when a person is diagnosed with T2D is lifestyle intervention, where individuals are advised to change their behaviors by exercising or controlling their diet. The fact that the average glucose and HbA1c levels were lower in the individuals excluded from the analyses (see Appendix 2) may indicate that medication use was not needed for some of these individuals. On the other hand, it may be that medication prescriptions were not recorded and that the date of the first prescription was wrong or missing in some individuals. Discrepancies in medical recordings, such as omitting prescribed medication or wrongly recording administration timing, commonly occur (27, 28). It is challenging to know what appears in the data is whether a true reflection of the real world or an error in the recording process.

Insufficient contextual knowledge introduces challenges in knowing to what extent the effect estimates in our study can be generalized. Hence, more detailed knowledge of how GPs prescribed medication in routine care would greatly help in modelling medication effects and understanding the generalizability of the estimated effects.

Recommendations

The estimated changes in blood glucose and HbA1c levels after medication prescription can be used to account for medication use in population cohort studies when untreated glucose or HbA1c values are of interest. For instance, when glucose or HbA1c is used as the outcome in the analysis, excluding individuals on glucose-lowering medication

or adding an indicator variable for medication use would lead to selection bias (21-23). Instead, one can add the values estimated from our study to the outcome values of the individuals using glucose-lowering mediation, an approach recommended in the literature (5, 7, 8). We suggest adding the estimated mean medication effects of this paper to the measurements of the individuals using medication based on their period of medication use. For example, for glucose, one can add 1.15 mmol/L to glucose measurements of the people who were on glucose-lowering medication for 0-3 months and 1.71 to those on medication for 3-6 months.

If glucose or HbA1c is the exposure or a confounding variable, one may either exclude individuals using glucose-lowering medication to estimate results among the non-medication users. As an alternative, researchers may use regression calibration together with adding the mean effect of medication use (7, 24). For this, one can use the estimated mean changes and standard deviations in this paper.

The sample size was limited in our study; therefore, we did not further investigate medication effects in different subgroups. Also, the type of medication used was homogenous, where 90% of the first prescribed medication was metformin. Studies with different populations may show different trends in types of prescribed medication.

5. Conclusion

This study explored changes in blood glucose and HbA1c after glucose-lowering medication prescription using routinely collected electronic health records. We observed that mean glucose and HbA1c levels increased shortly before the first prescription, which may reflect random high values. The medication effects were largest at 6-12 months after the prescription and smaller than what was known from RCTs. Routinely collected observational data allow investigation of real-world effects of medication over a longer period which could not be easily obtained from RCTs. However, challenges remain as clinical decisions and data recording processes in real-world settings are not always clearly known.

Reference

1. Control CfD, Prevention. National diabetes statistics report, 2020. Atlanta, GA: Centers for Disease Control and Prevention, US Department of Health and Human Services 2020:12-5.
2. Saydah SH. Medication use and self-care practices in persons with diabetes. Diabetes in America 3rd edition 2018.
3. Sudlow C, Gallacher J, Allen N, et al. UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. PLoS medicine 2015;12(3):e1001779.
4. Eastwood SV, Mathur R, Atkinson M, et al. Algorithms for the Capture and Adjudication of Prevalent and Incident Diabetes in UK Biobank. PLOS ONE 2016;11(9):e0162388.
5. 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.
6. 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.
7. Choi J, Dekkers OM, le Cessie S. A comparison of different methods for handling measurements affected by medication use. (ready to submit).
8. 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.
9. Palmer SC, Mavridis D, Nicolucci A, et al. Comparison of Clinical Outcomes and Adverse Events Associated With Glucose-Lowering Drugs in Patients With Type 2 Diabetes: A Meta-analysis. JAMA 2016;316(3):313-24.
10. Bennett WL, Maruthur NM, Singh S, et al. Comparative Effectiveness and Safety of Medications for Type 2 Diabetes: An Update Including New Drugs and 2-Drug Combinations. Annals of Internal Medicine 2011;154(9):602-13.
11. Bennett WL, Balfe LM, Faysal JM. AHRQ's comparative effectiveness research on oral medications for type 2 diabetes: a summary of the key findings. J Manag Care Pharm 2012;18(1 Suppl A):1-22.
12. Bailey C. The Current Drug Treatment Landscape for Diabetes and Perspectives for the Future. Clinical Pharmacology & Therapeutics 2015;98(2):170-84.
13. Maloney A, Rosenstock J, Fonseca V. A Model-Based Meta-Analysis of 24 Antihyperglycemic Drugs for Type 2 Diabetes: Comparison of Treatment Effects at Therapeutic Doses. Clinical Pharmacology & Therapeutics 2019;105(5):1213-23.
14. Feingold KR. Oral and Injectable (Non-Insulin) Pharmacological Agents for Type 2 Diabetes. Endotext [Internet] 2020.
15. de Mutsert R, den Heijer M, Rabelink TJ, et al. The Netherlands Epidemiology of Obesity (NEO) study: study design and data collection. European Journal of Epidemiology 2013;28(6):513-23.
16. WHO Collaboration Centre for Drug Statistics Methodology. Anatomical Therapeutic Chemical classification system.; 2012. (http://www.whocc.no/atc/). (Accessed November 1 2012).

17. Selvin E, Crainiceanu CM, Brancati FL, et al. Short-term Variability in Measures of Glycemia and Implications for the Classification of Diabetes. Archives of Internal Medicine 2007;167(14):1545-51.
18. American Diabetes A. Standards of medical care in diabetes--2010. Diabetes care 2010;33 Suppl 1(Suppl 1):S11-S61.
19. Barnett AG, van der Pols JC, Dobson AJ. Regression to the mean: what it is and how to deal with it. International Journal of Epidemiology 2004;34(1):215-20.
20. Morton V, Torgerson DJ. Regression to the mean: treatment effect without the intervention. Journal of Evaluation in Clinical Practice 2005;11(1):59-65.
21. Hernán M, S H-D, Robins J. A Structural Approach to Selection Bias. Epidemiology 2004;15(5):615-25.
22. Hernán M. Invited Commentary: Selection Bias Without Colliders. American Journal of Epidemiology 2017;185(11):1048-50.
23. Elwert F, Winship C. Endogenous Selection Bias: The Problem of Conditioning on a Collider Variable. Annu Rev Sociol 2014;40:31-53.
24. Carroll RJ, Ruppert D, Stefanski LA, et al. Measurement error in nonlinear models: a modern perspective. Chapman and Hall/CRC; 2006.
25. Briesacher BA, Andrade SE, Fouayzi H, et al. Comparison of drug adherence rates among patients with seven different medical conditions. Pharmacotherapy 2008;28(4):437-43.
26. Cramer JA. A Systematic Review of Adherence With Medications for Diabetes. Diabetes Care 2004;27(5):1218-24.
27. England G, Davin D, Lavin P, et al. MO923: Electronic Medication Record Accuracy and Clinical Pharmacist Intervention in Haemodialysis Outpatient Settings. Nephrology Dialysis Transplantation 2022;37(Supplement_3).
28. Pearce R, Whyte I. Electronic medication management: is it a silver bullet? Aust Prescr 2018;41(2):32-3.

Appendix 1

Keywords used for identifying prescription of glucose-lowering medication

- Keywords used for insulin:

abasaglar; actraphane; actrapid; apidra; fiasp; humalog; humuline; insulatard; insulin; insuman; lantus; levemir; liprolog; mixtard; novomix; novorapid; protaphane; ryzodeg; semglee; suliqua; toujeo; tresiba; xultophy

- Keywords used for other type of glucose lowering medication:

acarbose; actos; aloglip; amaryl; amglidia; avandamet; avandia; bydureon; byetta; canaglif; competact; dapaglif; diamicron; diastabol; dulaglut; ebymect; edistride; efficib; empaglif; enyglid; eucreas; exenat; fertin; forxiga; galvus; glibencl; gliclaz; glidipion; glimepiride; glubrava; glucient; glucobay; glucovance; glustin; glyxambi; icandra; incresync; invokana; jalra; janumet; januvia; jardiance; jentadueto; komboglyze; linaglip; liraglut; metfocell; metform; metnova; miglitol; nateglin; novonorm; onglyza; ozempic; pioglit; prandin; qtern; repaglin; ristaben; ristfor; rosiglit; saxaglip; saxenda; semaglut; sitaglip; starlix; steglujan; synjardy; tandemact; tesavel; tolbutam; trajenta; trulicity; velmetia; victoza; vildaglip; vipdomet; vipidia; vokanamet; xelevia; xigduo; xiliarx; yalformet; zomarist

Appendix 2

A comparison of characteristics at the first NEO visit between the individuals who were included and excluded from the analyses. Mean and standard deviation was used for continuous variables [mean (sd)]. Frequency and the percentage was used for categorical variables [N (%)].

	Included participants (n=127)	**Excluded participants (n=170)**
Sex		
Male	65 (51.2%)	72 (42.4%)
Age in years	56.0 (5.8)	56.7 (5.5)
Education		
High	50 (39.4%)	50 (29.4%)
Hypertension		
Yes	63 (49.6%)	93 (54.7%)
BMI (kg/m2)	33.6 (5.4)	32.6 (4.7)
Glucose (mmol/L)	7.0 (1.8)	6.6 (1.2)
Insulin (mU/L)	19.5 (11.8)	20.8 (23.9)
HOMA1-IR	6.0 (3.8)	6.2 (8.4)
HbA1c (%)	6.0 (0.9)	5.8 (0.7)
Total cholesterol (mmol/L)	5.7 (1.1)	5.6 (1.2)
Triglycerides (mmol/L)	1.9 (1.3)	2.0 (1.2)
HDL (mmol/L)	1.3 (0.3)	1.2 (0.3)
LDL (mmol/L)	3.6 (1.0)	3.5 (1.1)
Lipid-lowering drugs use		
Yes	19 (15.0%)	32 (18.8%)
Hypertension drugs use		
Yes	54 (42.5%)	79 (46.5%)

Chapter 6

How measurements affected by medication use are reported and handled in observational research: a literature review

Published in Pharmacoepidemiol Drug Saf. 2022; 31(7): 739- 748

Jungyeon Choi, Olaf M. Dekkers, Saskia le Cessie

Abstract

Purpose

In epidemiological research, measurements affected by medication, e.g., blood pressure lowered by antihypertensives, are common. Different ways of handling medication are required depending on the research questions and whether the affected measurement is the exposure, the outcome, or a confounder. This study aimed to review the handling of medication use in observational research.

Methods

PubMed was searched for etiological studies published between 2015 to 2019 in fifteen high-ranked journals from cardiology, diabetes, and epidemiology. We selected studies that analyzed blood pressure, glucose, or lipid measurements (whether exposure, outcome, or confounder) by linear or logistic regression. Two reviewers independently recorded how medication use was handled and assessed whether the methods used were in accordance with the research aim. We reported the methods used per variable category (exposure, outcome, confounder).

Results

127 articles were included. Most studies did not perform any method to account for medication use (exposure 58%, outcome 53%, confounder 45%). Restriction (exposure 22%, outcome 23%, confounders 10%), or adjusting for medication use using a binary indicator were also used frequently (exposure: 18%, outcome: 19%, confounder: 45%). No advanced methods were applied. In 60% of studies, the methods' validity could not be judged due to ambiguous reporting of the research aim. Invalid approaches were used in 28% of the studies, mostly when the affected variable was the outcome (36%).

Conclusion

Many studies ambiguously stated the research aim and used invalid methods to handle medication use. Researchers should consider a valid methodological approach based on their research question.

Key points

- Methodological studies stressed the importance of adequately handling variables affected by medication use and showed that using invalid methods may lead to substantial bias. However, we found that many clinical studies did not consider this issue.
- A large proportion of the studies did not provide information on whether their interest was in the observed or the untreated underlying values. Without clear reporting on research aims, the interpretation of the results will be ambiguous.
- Methods that have been shown invalid, such as restricting a study population to non-medication users when the outcome variable was affected by medication use, are still often used.
- Justification on methods used for handling medication use was seldom given.

1. Introduction

Measurements affected by medication use are a commonly encountered feature in epidemiological research. For example, blood pressure is lowered by antihypertensive drugs or glucose levels by glucose-lowering drugs. Several methods for handling medication use have been proposed and compared (1-9). Studies have shown that different methods may lead to substantially different effect estimates (2-5, 8-10), and the optimal method depends on i) the research aim and ii) whether the medication effect is on the exposure, outcome, or a confounder (10). If the method used for handling medication effect does not match the research question, substantial bias can be introduced, and the interpretation of results will be unclear (11).

Thus, it is essential to carefully think about the research question when some individuals in a study population use medication that affects the variables in the dataset. In some situations, the research interest could be in the observed measurements, regardless of whether some individuals' measurements are lowered due to antihypertensive medication use; for instance, when the effect of current blood pressure on the course of the disease for patients infected with Covid-19 is considered. In other cases, blood pressure values that would have been observed if the medication was not administered (sometimes referred to as *underlying values* (2, 12)) could be the primary interest, for example, if the effect of genetic factors on blood pressure is examined. In this instance, a method to correct for the medication effect should be used.

Handling medication use in epidemiological research has received attention, although this was mainly in methodological papers (1-9, 13, 14). There are studies that adopted some of the methods suggested (15, 16). However, a majority seems to overlook the

potential bias due to inadequate handling of medication use (4). To our knowledge, there has been no systematic review of how medication use is being handled in research practice. Therefore, In this literature review, we aim to investigate which methods are used in observational studies to handle measurements affected by medication, assess how often methods used correspond to the research aims stated in these studies, and evaluate the validity of the methods used.

2. Methods

Search strategy

Our search aimed to identify observational studies that included measurements that have been affected by medication use. The search covered three different journal fields; cardiology, diabetes, and epidemiology, thereby focusing on blood pressure, glucose, or lipid measurements. Five journals with the highest impact factors were selected for each journal field. Table 1 lists the selected journals.

To select the publications, we searched PubMed for studies published in the 15 selected journals between January 1st, 2015 to December 31st, 2019 that used logistic or linear regression. The full search strategies for this step can be found in Online supplementary material 1.

Table 1. List of selected journals and the number of articles returned from the PubMed search

Cardiology journals (n=258)	Diabetes journals (n=331)	Epidemiology journals (n=688)
Cardiovascular Research (4)	Diabetes (25)	American Journal of Epi. (212)
Circulation Research (7)	Diabetes Care (169)	Epidemiology (108)
Circulation (89)	Diabetes, Obesity & Metabolism (35)	European Journal of Epi. (62)
European Heart Journal (39)	Diabetologia (84)	International Journal of Epi. (228)
Hypertension (119)	The Lancet Diabetes & Endocrinology (18)	Journal of Clinical Epi. (78)

The full-text of the identified papers was screened, and papers that met the following inclusion criteria were selected for review: 1) observational studies in adults, 2) sample size larger than 100, 3) aimed to answer etiological questions, 4) performed linear or logistic regression (including linear mixed modelling), and 5) inclusion of any of the following variables: blood pressure-related measurements (e.g., systolic or diastolic blood pressure, pulse wave velocity), glucose-related measurements (e.g., glucose level, insulin level, HbA1c, HOMA index) and lipid levels (e.g., cholesterol measures, triglycerides). For studies on type 1 diabetes patients, glucose measurements were not considered because there is no variation in glucose medication use in these

patients as insulin treatment is mandatory and unavoidable. If blood pressure related measurements or lipid measurements were used, these studies could be included.

Among the studies that met the inclusion criteria, we selected a maximum of 50 articles to be reviewed from each field. If a specific journal (five per field) contained less than ten articles meeting the inclusion criteria, all articles from that journal were selected to be reviewed. The rest of the studies were randomly selected until the sample size per journal field met 50 or no more articles were left to be selected. If two or more studies used the same study population within a field, the latest publication was considered.

Data extraction

Data extraction for all 127 papers was independently performed by two reviewers, JC (a Ph.D. candidate in clinical epidemiology) and SlC (a senior statistician and epidemiologist). Disagreements between the two reviewers were resolved during a consensus meeting involving the third reviewer, OMD (a senior epidemiologist and endocrinologist). For each paper, the following general information was extracted:

i) Authors, journal, name of the study/cohort/database
ii) Study population and sample size
iii) Research question with exposure(s) and outcome of interest
iv) Whether linear, logistic regression, or both were performed

For information related to medication use, we extracted the following:

v) Measurements that may have been affected by medication use (blood pressure, glucose, and/or lipid). 'Medication use' was defined as the use of drugs that aim to lower blood pressure, glucose, or lipid level.
vi) Whether the measurement potentially affected by medication was an exposure, an outcome, or a confounder. We used the following rules:
    a. When the measurement was mentioned as an 'independent variable' and the effect of the variable on the outcome was specifically discussed in the paper, it was coded as an exposure.
    b. In Mendelian randomization studies, the exposures in the research questions are the outcomes in the corresponding regression analyses. In this case, we coded the variable as an outcome.
vii) Percentage of individuals using medication
viii) Whether details on medication information were given (e.g., type and dose of medication, duration of use)
ix) Methods used for handling medication use for each affected variable
    a. If different variables had the same role and were handled by the same method, the method was recorded once (e.g., if a study had blood pressure and glucose

level as confounding variables and medication use for both variables was handled by a restriction method, the method was recorded once).

b. When multiple models were used to evaluate the same relationship, the most complex model was considered (e.g., when both unadjusted and adjusted analyses were performed to estimate the relationship between the same variables, the adjusted analysis was considered).

x) Justification for the chosen method

xi) Sensitivity analyses for handling medication use

Assessment of research aims and the validity of the methods used

We evaluated the validity of methods used for handling medication use based on the research aims of the study and which variable was affected by medication use. Figure 1 displays our evaluation process. In detail, the following steps were taken.

Step I: For each variable affected by medication use, we first evaluated the research aim as stated by the authors, which was categorized as follows:

1) The interest is in the observed values as they are.
2) The interest is in the values that would be observed if no medication was administered (we refer to this as 'values if untreated' or 'untreated values' in the further text).
3) The interest is ambiguously reported.

Step II: The validity of the method used for each variable was evaluated in relationship to the research aim and whether the affected variable was an exposure, an outcome, or a confounder. The assessment of whether the methods used are in general valid or invalid was based on recommendations from previous methodological studies (2-6, 10, 11, 17-26). For example, restricting the study population to non-medication users was considered valid when the exposure or a confounder is affected by medication use regardless of whether the research aim is in the values as observed or if untreated. This is because the restriction on a proxy variable of the exposure or a confounder (medication use, in this case) in general would lead to a selection of a subgroup without introducing selection bias (21). Contrarily, the restriction method was considered invalid regardless of the research aim when the outcome is affected by medication use. Selection on medication use, an event that occurred after the follow-up started and related to the outcome, would introduce selection bias (2, 21, 22, 24-26). A complete discussion of all possible options can be found in the appendix.

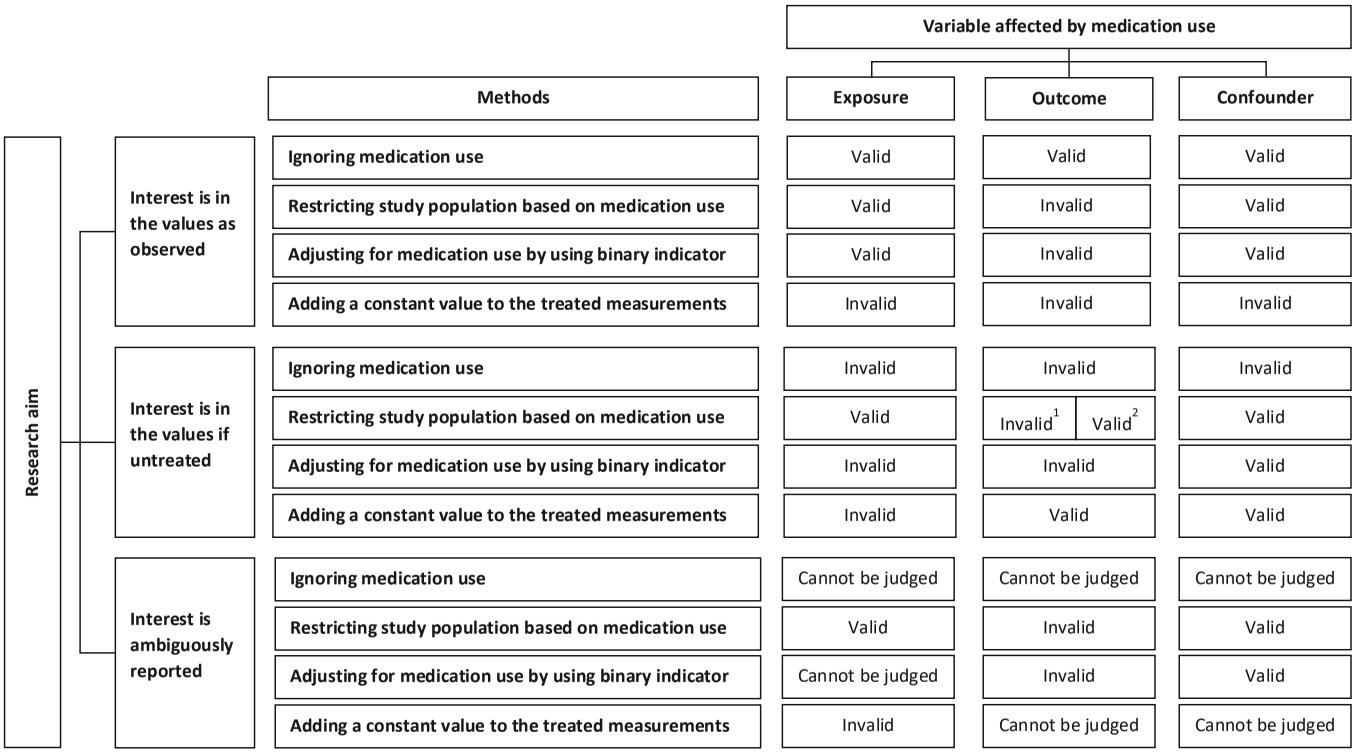
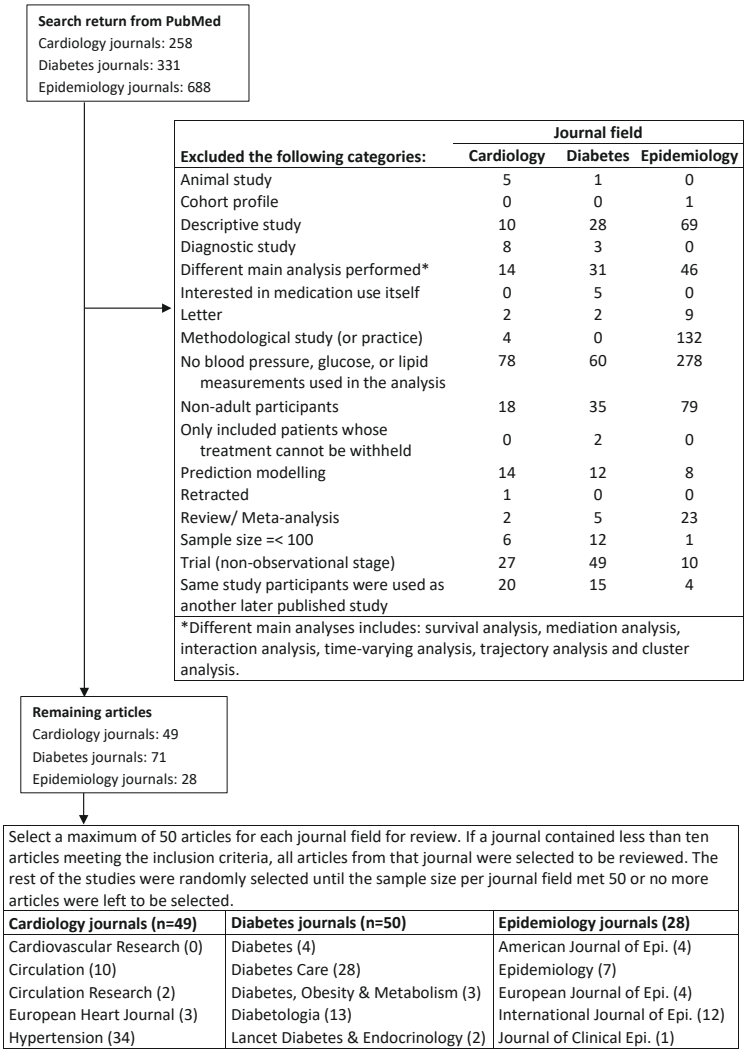


Figure 1. A flow chart for an assessment of valid and invalid approaches for handling medication use. Details on the assessment of the validity of the methods used can be found in Appendix 2. 1Invalid in cross-sectional settings. 2Valid in a follow-up setting where the restricting is based on the data before the start of follow-up.

3. Results

Our search strategy in PubMed retrieved 258 articles in cardiology journals, 331 articles in diabetes journals, and 688 articles in epidemiology journals (see Table 1 for the number of papers and Figure 2 for the flowchart). After the screening process, 49 articles in the cardiology field, 73 articles in the diabetes field, and 28 articles in the epidemiology field remained. For the diabetes field, a subset of 50 articles was selected, as described in the methods section. We included 49 articles from cardiology journals, 50 articles from diabetes journals, and 28 articles from epidemiology journals for a total of 127 studies.



Excluded the following categories:	Journal field: Cardiology	Diabetes	Epidemiology
Animal study	5	1	0
Cohort profile	0	0	1
Descriptive study	10	28	69
Diagnostic study	8	3	0
Different main analysis performed*	14	31	46
Interested in medication use itself	0	5	0
Letter	2	2	9
Methodological study (or practice)	4	0	132
No blood pressure, glucose, or lipid measurements used in the analysis	78	60	278
Non-adult participants	18	35	79
Only included patients whose treatment cannot be withheld	0	2	0
Prediction modelling	14	12	8
Retracted	1	0	0
Review/ Meta-analysis	2	5	23
Sample size =< 100	6	12	1
Trial (non-observational stage)	27	49	10
Same study participants were used as another later published study	20	15	4

*Different main analyses includes: survival analysis, mediation analysis, interaction analysis, time-varying analysis, trajectory analysis and cluster analysis.

Select a maximum of 50 articles for each journal field for review. If a journal contained less than ten articles meeting the inclusion criteria, all articles from that journal were selected to be reviewed. The rest of the studies were randomly selected until the sample size per journal field met 50 or no more articles were left to be selected.

Cardiology journals (n=49)	Diabetes journals (n=50)	Epidemiology journals (28)
Cardiovascular Research (0)	Diabetes (4)	American Journal of Epi. (4)
Circulation (10)	Diabetes Care (28)	Epidemiology (7)
Circulation Research (2)	Diabetes, Obesity & Metabolism (3)	European Journal of Epi. (4)
European Heart Journal (3)	Diabetologia (13)	International Journal of Epi. (12)
Hypertension (34)	Lancet Diabetes & Endocrinology (2)	Journal of Clinical Epi. (1)

Figure 2. Flow chart of the literature search and screening process

Summaries of reviewed articles

Online supplementary material 2 displays the complete list of the reviewed articles and extracted information from each article. Table 2 provides a summary of the included studies. Overall, the measurement affected by medication use was most often a confounding variable (In 56% of the studies), followed by an outcome (42%) and/or an exposure (35%). In the epidemiology journals, affected outcomes were more often present (64%). Sample sizes varied largely between the reviewed articles and were generally larger in the epidemiology journals. Included studies performed linear regression analysis (59%), logistic regression analysis (40%), and/or linear mixed modelling (9%).

Overall, a majority of the studies did not report the percentage of medication users (47%) or only reported medication use for part of the variables affected (14%). Among the studies which fully or partially provided information on the percentage of medication users, the median percentage of medication users was 32%. The percentage of medication users ranged from 0 to 100, because some studies restricted their study population to medication users or non-users. Details of medication use, such as dose or prescription frequency, were seldom given (7%).

Table 2. Summaries of reviewed articles

	Journal field			
	All journals (n=127)	**Cardiology (n=49)**	**Diabetes (n=50)**	**Epidemiology (n=28)**
Affected variables in the analysis* [n(%)]				
Exposure	45 (35.4)	21 (42.9)	20 (40.0)	4 (14.3)
Outcome	53 (41.7)	17 (34.7)	18 (36.0)	18 (64.3)
Confounder	71 (55.9)	29 (59.2)	33 (66.0)	9 (32.1)
Sample size [Median [min, max]]	1540	1746	1147	2514
	[122, 615035]	[122, 615035]	[122, 222773]	[277, 486936]
Type of analysis* [n(%)]				
Linear regression	75 (59.1)	29 (59.2)	26 (52.0)	20 (71.4)
Logistic regression	51 (40.2)	19 (38.8)	25 (50.0)	7 (25.0)
(Generalized) Linear mixed model	12 (9.4)	6 (12.2)	4 (8.0)	2 (7.1)
Percentage of medication use				
Reported or traceable for all variables	49 (38.6)	24 (49.0)	19 (38.0)	6 (21.4)
Reported for some variables	18 (14.2)	6 (12.2)	9 (18.0)	3 (10.7)
Not reported	60 (47.2)	19 (38.8)	22 (44.0)	19 (67.9)
Medication user percentage among the reported [median [min, max]]	32.0 [0, 100]	22.0 [0, 91]	54.6 [0, 100]	11.7 [1.3, 59]
Details of medication information reported	9 (7.1)	6 (12.2)	2 (4.0)	1 (3.6)

*Exceed 100% when added up because some studies performed more than one analysis.

Methods used for handling medication use

Table 3 summarizes the methods used for handling measurements affected by medication use. Lists of the studies using each method can be found in Online supplementary material 3. A large number of studies did not use any method specifically for handling medication use (58% when medication use was in the exposure, 53% when in the outcome, and 45% when in a confounder). Restricting the analysis to a certain subpopulation was frequently used (for exposure: 22%, outcome: 23%, and confounder: 10%). Some studies restricted their study population to medication users or non-medication users. Others restricted the analyses to subgroups that were partly

defined based on medication use, such as individuals without hypertension, defined as people not using antihypertensive drugs *and* having normal blood pressure levels.

A binary covariate in a regression model was the next most used method for exposures (18%) and outcomes (19%). For confounders, it was one of the most used methods (45%). The binary variable used for the adjustment was often 'using medication (yes/no)'. However, one study adjusted for 'using medication *or* having high value (yes/no)' (e.g., hypertension vs. no hypertension, while defining hypertension as taking antihypertensive drugs *or* having blood pressure above a certain level).

Adding an estimate of the mean medication effect to treated values was adopted only in four studies. One study used this method for handling medication use in the exposure. No study used any of the more advanced methods suggested in the literature, such as quantile regression (3), censored normal regression (2), or Heckman's treatment model (4, 5).

In total, only ten studies (8%) explicitly provided justification for the chosen methods for handling medication use. Given justifications, however, may not reflect the validity of the methods used. Sensitivity analyses were performed in 21 studies (16%) in total. A list of methods used in the sensitivity analyses can be found in Online supplementary material.

Table 3. Frequency of methods used for handling medication use in main analyses [n(%)]

Affected measurement		Journal field			
	Methods	**All journals**	**Cardiology**	**Diabetes**	**Epidemiology**
Exposure	**Ignoring medication use**	26 (57.8)	13 (61.9)	9 (45.0)	4 (100)
	Restricting study population	10 (22.2)	4 (19)	6 (30.0)	-
	to medication users	1	-	1	-
	To medication non-users	3	2	1	-
	to medication non-users AND having normal values	6	2	4	-
	Adjusting as a binary covariate	8 (17.8)	3 (14.3)	5 (25.0)	-
	using medication (yes/no)	7	2	5	-
	using medication OR having high values (yes/no)	1	1	-	-
	Adding a constant value to the treated measurements	1 (2.2)	1 (4.8)	-	-
	Subtotal	**45 (100)**	**21 (100)**	**20 (100)**	**4 (100)**
Outcome	**Ignoring medication use**	28 (52.8)	9 (52.9)	8 (44.4)	11 (61.1)
	Restricting study population	12 (22.6)	1 (5.9)	7 (38.9)	4 (22.2)
	to medication non-users	7	1	3	3
	to medication non-users AND having normal clinical values	5	-	4	1
	Adjusting as a binary covariate	10 (18.9)	5 (29.4)	3 (16.7)	2 (11.1)
	using medication (yes/no)	10	5	3	2

Table 3. Frequency of methods used for handling medication use in main analyses [n(%)] *(continued)*

Affected measurement	Journal field			
	Methods	**All journals**	**Cardiology**	**Diabetes**	**Epidemiology**
	Adding a constant value to the treated measurements	3 (5.7)	2 (11.8)	-	1 (5.6)
	Subtotal	**53 (100)**	**17 (100)**	**18 (100)**	**18 (100)**
Confounder	**Ignoring medication use**	32 (45.1)	9 (31.0)	17 (51.5)	6 (66.7)
	Restricting study population	7 (9.9)	2 (6.9)	5 (15.2)	-
	to medication users	3	-	3	-
	to medication non-users	1	-	1	-
	to medication non-users AND having normal clinical values	3	2 (6.9)	1	-
	Adjusting as a binary covariate	32 (45.1)	18 (62.1)	11 (33.3)	3 (33.3)
	using medication (yes/no)	26	13	10	3
	using medication OR having high values (yes/no)	6	5	1	
	Subtotal	**71 (100)**	**31 (100)**	**33 (100)**	**9 (100)**
Justification for the chosen method given		12 (9.4)	6 (12.2)	3 (6.0)	3 (10.7)
Sensitivity analysis performed		22 (17.3)	9 (18.4)	8 (16.0)	5 (17.9)

The sum of the subtotals exceeds the total number of articles included for each journal field (49 for cardiology journals, 50 for diabetes journals, and 28 for epidemiology journals) because more than one variable was affected by medication use in some studies.

Assessment of research aim-analysis match and validity of the methods used
The results of the assessment of the methods used for handling medication use are summarized in Table 4. In a majority of the studies, it was unclear whether the research interest was in the values as observed or in untreated values. Thus, the validity of the used methods often could not be judged properly (exposure: 56%, outcome: 36%, confounder: 45%). Overall, no noticeable difference in performance was observed across the journal fields.

In all studies where the interest explicitly was in observed exposure values, medication use was also ignored in the analyses. When interest was in untreated exposure values (11 analyses), most often, the analysis was restricted to untreated individuals, which is considered in general a valid approach. However, in 5/11 analyses, invalid approaches were used; such as ignoring the treatment, adjusting for medication use as binary covariates, or adding a constant value. In 3/28 analyses where the research aim for the exposure variable was ambiguous, the study population was restricted to untreated individuals, which we considered a valid approach for all research aims.

When the outcome was an affected variable, we found only three out of 53 analyses that were undoubtedly interested in the values as observed. Among these, two analyses ignored medication use accordingly. However, one used a valid method which is adjusting for medication use as a binary covariate. More often, the studies were found to be interested in the outcome values if untreated. However, in most cases (19/21 analyses), invalid approaches, such as restricting the study population or adjusting using a binary covariate, were used. When the research aim regarding the outcome variable was ambiguous, the affected outcome was often handled with methods that are prone to yield biased causal effects regardless of the research aim; for example, restricting the study population in a cross-sectional setting or adjusting using a binary covariate.

For confounders affected, only in eight out of 71 cases, it was clear whether interest was in observed values (n=4) or untreated values (n=4). Valid methods were used in these cases. When the aim was unclear, often (31/63) medication use was added as an additional covariate to the regression model. This approach is considered valid both when interest is in observed values (where medication use could be an extra confounder) and also when interest is in unaffected values (in which case adding both medication use and the observed value will account for most of the confounding of the underlying unaffected values).

Table 4. Assessment of the research question-analysis match and the validity of used methods [n (%)]. The percentage adds up to 100 per affected variable.

			Journal field			
Affected variable	**Research aim**	**Validity**	**All journals (n=127)**	**Cardiology (n=49)**	**Diabetes (n=50)**	**Epidemiology (n=28)**
Exposure	As observed	Valid	6 (13.3)	1 (4.8)	5 (25.0)	-
		Invalid	-	-	-	-
	If untreated	Valid	6 (13.3)	3 (14.3)	3 (15.0)	-
		Invalid	5 (11.1)	3 (14.3)	1 (5.0)	1 (25.0)
	Ambiguously reported	Valid	3 (6.7)	1 (4.8)	2 (10.0)	-
		Invalid	-	-	-	-
		Cannot be judged	25 (55.6)	13 (61.9)	9 (45.0)	3 (75.0)
Outcome	As observed	Valid	2 (3.8)	-	2 (11.1)	-
		Invalid	1 (1.9)	-	1 (5.6)	-
	If untreated	Valid	2 (3.8)	1 (5.9)	-	1 (5.6)
		Invalid	19 (35.8)	4 (23.5)	10 (55.6)	5 (27.8)
	Ambiguously reported	Valid	-	-	-	-
		Invalid	11 (20.8)	4 (23.5)	2 (11.1)	5 (27.8)
		Cannot be judged	18 (34.0)	8 (47.1)	3 (16.7)	7 (38.9)
Confounder	As observed	Valid	4 (5.6)	1 (3.4)	3 (9.1)	-
		Invalid	-	-	-	-
	If untreated	Valid	4 (5.6)	3 (10.3)	1 (3.0)	-
		Invalid	-	-	-	-
	Ambiguously reported	Valid	31 (43.7)	16 (55.2)	12 (36.4)	3 (33.3)
		Invalid	-	-	-	-
		Cannot be judged	32 (45.1)	9 (31.0)	17 (51.5)	6 (66.7)

4. Discussion

In this review, we empirically assessed how variables affected by medication use are handled in observational etiological studies. Our review showed that a large proportion of the studies did not provide clear research aims stating whether their interest was in the observed or the untreated underlying values and methods in general considered invalid, such as restricting the study population to non-medication users when the outcome is affected by medication use, were often used. Notably, a justification for the chosen method was rarely given, and the number of medication users was not reported or insufficiently reported in more than half of the studies. These findings suggest that there is low awareness of potential bias by medication use.

The median percentage of medication users in our review was 31%, in which case the estimated effect may differ considerably depending on whether the interest is in the observed values or the underlying unaffected values. Even when the number of medication users is low, differences can still be substantial if the effect of the medication is large. More information on the direction and magnitude of bias when interest is in the underlying unaffected values can be found in several methodological studies (2, 3, 10). Factors that may play a role include but are not limited to, different types of medication and doses, heterogeneity of medication effect across the individuals, medication effect being cancelled/ enhanced by other interventions, or time-varying aspects of medication use. Such information heavily relies on content knowledge. Thus, we urge clinical researchers to provide and discuss relevant information on the medication used in their study population.

We found that invalid methods were especially prevalent when the affected variable was the outcome. Often the analysis was performed conditional on medication use. Although the bias due to selecting events related to the outcome has been discussed extensively in the literature (2, 21, 22, 24-26), it seemed that such consideration was often not taken into account. We also observed that the research aim was most often ambiguously reported for confounding variables affected by medication use. This is not surprising since confounders are mostly not the variables of main interests. However, inadequately handling medication use in confounding variables can lead to bias (10).

We noticed that recommendations in methodological papers were seldom applied. For example, Tobin et al. (2) recommended adding a constant value to measurements of treated values of an outcome variable when interest is in the underlying unaffected values and stressed the necessity of sensitivity analysis to determine the robustness of the particular choice of constant. In our review, none of the four studies which applied this method tested the robustness of their choice of constant. Additionally, no study was found to use any of the more advanced statistical methods previously suggested (2, 3, 5, 10). This may call for methodological papers in clinical journals that provide

practical guidelines and tutorials on when and how to apply corrections for medication use in applied clinical research.

We only included studies that used linear regression, logistic regression, and mixed linear models. However, potential bias due to measurements affected by medication use is present in any study where a mixed study population of medication users and non-users exists (14). In complex settings, such as when medication use is an effect modifier or a mediator or when there is time-varying medication use, extra caution would be needed (1, 13). Handling medication use also plays a role when continuous variables are being categorized. For example, when categorizing glucose values as high versus normal, the distinction could be made based on untreated values, where patients on medication are classified as high glucose even if their glucose levels are regulated. These approaches would be considered valid once medication users are classified correctly; however, the power may be lower (3, 27, 28).

5. Conclusion

Our review has shown that potential bias due to medication use is often overlooked and that decisions on handling medication use are frequently made without valid justification. We urge researchers to provide clear information on medication use, consciously decide on a method for handling medication use based on their research question, and communicate the rationale behind their decision.

References

1. 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.
2. Tobin MD, Sheehan NA, Scurrah KJ, Burton PR. Adjusting for treatment effects in studies of quantitative traits: antihypertensive therapy and systolic blood pressure. Statistics in Medicine. 2005;24(19):2911-35.
3. 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.
4. 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.
5. 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.
6. Cui JS, Hopper JL, Harrap SBJH. Antihypertensive treatments obscure familial contributions to blood pressure variation. 2003;41(2):207-10.
7. Levy D, DeStefano AL, Larson MG, O'Donnell CJ, Lifton RP, Gavras H, et al. Evidence for a gene influencing blood pressure on chromosome 17. Hypertension. 2000;36:477-83.
8. Balakrishnan P, Beaty T, Young JH, Colantuoni E, Matsushita K. Methods to estimate underlying blood pressure: The Atherosclerosis Risk in Communities (ARIC) Study. PLOS ONE. 2017;12(7):e0179234.
9. McClelland RL, Kronmal RA, Haessler J, Blumenthal RS, Goff DC. Estimation of risk factor associations when the response is influenced by medication use: An imputation approach. Statistics in Medicine. 2008;27(24):5039-53.
10. Choi J, Dekkers OM, le Cessie S. A comparison of different methods for handling measurements affected by medication use. (ready to submit).
11. Choi J, le Cessie S, Dekkers OM. Tying research question and analytical strategy when variables are affected by medication use. Submitted to European Journal of Epidemiology.
12. Hulman A, Witte DR, Vistisen D, Balkau B, Dekker JM, Herder C, et al. Pathophysiological Characteristics Underlying Different Glucose Response Curves: A Latent Class Trajectory Analysis From the Prospective EGIR-RISC Study. Diabetes care. 2018;41(8):1740-8.
13. Schmidt AF, Heerspink HJL, Denig P, Finan C, Groenwold RHH. When drug treatments bias genetic studies: Mediation and interaction. PLOS ONE. 2019;14(8):e0221209.
14. van Geloven N, Swanson SA, Ramspek CL, Luijken K, van Diepen M, Morris TP, et al. Prediction meets causal inference: the role of treatment in clinical prediction models. European Journal of Epidemiology. 2020;35(7):619-30.
15. Laaksonen J, Mishra PP, Seppälä I, Lyytikäinen L-P, Raitoharju E, Mononen N, et al. Examining the effect of mitochondrial DNA variants on blood pressure in two Finnish cohorts. Scientific Reports. 2021;11(1):611.
16. Norris T, Cole TJ, Bann D, Hamer M, Hardy R, Li L, et al. Duration of obesity exposure between ages 10 and 40 years and its relationship with cardiometabolic disease risk factors: A cohort study. PLOS Medicine. 2020;17(12):e1003387.

17. Tanamas SK, Hanson RL, Nelson RG, Knowler WC. Effect of different methods of accounting for antihypertensive treatment when assessing the relationship between diabetes or obesity and systolic blood pressure. Journal of Diabetes and its Complications. 2017;31(4):693-9.
18. Laird EJ, McNicholas T, O'Halloran AM, Healy M, Molloy AM, Carey D, et al. Vitamin D Status Is Not Associated With Orthostatic Hypotension in Older Adults. Hypertension (Dallas, Tex : 1979). 2019;74(3):639-44.
19. VanderWeele TJ. Principles of confounder selection. European Journal of Epidemiology. 2019;34(3):211-9.
20. Carroll RJ, Ruppert D, Stefanski LA, Crainiceanu CM. Measurement error in nonlinear models: a modern perspective: Chapman and Hall/CRC; 2006.
21. Hernán M, Robins J. Causal inference: What if. Boca Raton: Chapman & Hall/CRC; 2020.
22. Hernán M, S H-D, Robins J. A Structural Approach to Selection Bias. Epidemiology. 2004;15(5):615-25.
23. Rubin DB. For objective causal inference, design trumps analysis. The Annals of Applied Statistics. 2008;2(3):808-40, 33.
24. Hernán M. Invited Commentary: Selection Bias Without Colliders. American Journal of Epidemiology. 2017;185(11):1048-50.
25. Elwert F, Winship C. Endogenous Selection Bias: The Problem of Conditioning on a Collider Variable. Annu Rev Sociol. 2014;40:31-53.
26. Cole SR, Platt RW, Schisterman EF, Chu H, Westreich D, Richardson D, et al. Illustrating bias due to conditioning on a collider. International Journal of Epidemiology. 2009;39(2):417-20.
27. Bennette C, Vickers A. Against quantiles: categorization of continuous variables in epidemiologic research, and its discontents. BMC Medical Research Methodology. 2012;12(1):21.
28. Altman DG, Royston P. The cost of dichotomising continuous variables. BMJ. 2006;332(7549):1080.

Appendix

Details on the assessment of the validity of used methods.

1. When the affected variable is an exposure
   - 1.1. When the interest is in the values as observed
     - 1.1.1. Ignoring medication use is valid.
     - 1.1.2. Restricting the study population based on medication use is considered valid. The method can yield unbiased estimates in the selected subpopulation, given that variables affecting both medication use and the outcome are correctly adjusted (1, 2). Results cannot be extrapolated to the excluded population when effect heterogeneity is present.
     - 1.1.3. Adjusting for medication use as a binary covariate is considered valid, as the medication use occurs before the exposure variable is measured. Adjusting for medication use may be needed if it also affects the outcome, in which case medication use is a confounder (3).
     - 1.1.4. Adding a constant value (the estimated mean medication effect) to the treated measurements) is considered invalid. This is because the method does not account for the variability in medication effect between medication users (2, 4).
   - 1.2. When the interest is in the values if untreated
     - 1.2.1. Ignoring medication use is invalid.
     - 1.2.2. Restricting the study population based on medication use is considered valid. The method will yield a valid estimate of the effect of the exposure on the outcome in the subpopulation under the same considerations as 1.1.2.
     - 1.2.3. Adjusting for medication use as a binary covariate is considered invalid. The effect of the medication is ,in general, not the same in all individuals. Therefore, applying this method will likely lead to an underestimation of the association between the exposure and the outcome due to the fact that the medication effect on the exposure level cannot be completely accounted for (2, 4).
     - 1.2.4. Adding a constant value to the treatment is considered invalid, because the method does not account for the variability in medication effect between medication users. This phenomenon is described in literature on measurement error (see reference) (2, 4).
   - 1.3. When the interest is ambiguous
     - 1.3.1. The validity of ignoring medication use cannot be judged.
     - 1.3.2. Restricting the study population based on medication use is considered valid, because it is a valid approach in either case where the research aim is in the values as observed or the values if untreated.
     - 1.3.3. The validity of adjustment for a binary indicator cannot be judged.
     - 1.3.4. Adding a constant value to the treated measurements is invalid.

2. When the affected variable is an outcome
   2.1. When the interest is in the values as observed
      2.1.1. Ignoring medication use is valid.
      2.1.2. Restricting the study population based on medication use is invalid, because selection on intercurrent events may lead to selection bias (1, 5, 6).
      2.1.3. Adjusting for medication use with a binary indicator is considered invalid. The method may lead to selection bias (collider bias) due to indirect conditioning on intercurrent events and the outcome variable (1, 5, 6).
      2.1.4. Adding a constant value is invalid.
   2.2. When the interest is in the values, if untreated
      2.2.1. Ignoring medication use is invalid.
      2.2.2. Restricting the study population based on medication use is invalid as the method introduces selection bias. (1, 7-9)
      2.2.3. Adjusting for medication use is invalid as it introduces collider bias (5, 8).
      2.2.4. Adding a constant value to the treated measurement is a valid approach (2, 8, 10, 11).
   2.3. When the interest is ambiguous
      2.3.1. The validity of ignoring medication use cannot be judged.
      2.3.2. Restricting the study population based on medication use is invalid, because it is invalid for both when interest was in the values as observed or the values if untreated values.
      2.3.3. Adjusting for medication use is invalid when interest was in observed and in untreated values.
      2.3.4. The validity of adding a constant value to the treated measurement cannot be judged.

3. When the affected variable is a confounder
   3.1. When the interest is in the values as observed
      3.1.1. Ignoring medication use is valid.
      3.1.2. Restricting the study population based on medication use is considered valid under the same considerations as 1.1.2. The method serves as accounting for confounding by restriction (5).
      3.1.3. Adjusting for medication use is valid. The method is comparable to adjusting for a proxy confounder (5, 12).
      3.1.4. Adding a constant value to treated measurements is considered invalid.
   3.2. When the interest is in the values, if untreated
      3.2.1. Ignoring medication use is invalid.

3.2.2. Restricting the study population based on medication use is valid for the same reason as 3.1.2.
3.2.3. Adjusting for medication use is valid for the same reason as 3.1.3.
3.2.4. Adding a constant value to the treatment measurement is considered valid because simulation showed that this approach would handle most of the confounding (2).

3.3. When the interest is ambiguous
3.3.1. The validity of ignoring medication use cannot be judged.
3.3.2. Restricting the study population based on medication use is valid for the same reason as 1.1.2.
3.3.3. Adjusting for medication use is valid for the same reason as 3.1.3.
3.3.4. The validity of adding a constant value to the treated measurement cannot be judged.

More advanced methods are available (not shown in Figure 1) when the study interest is in the underlying value that is *not* affected by medication use. For instance, censored normal regression (8), quantile regression (10), and Heckman's treatment model (13) could be used under certain assumptions when the outcome is affected. Methods for correcting differential measurement error (e.g., regression calibration with adding mean treatment effect) could be used (4, 14) for an exposure affected by medication use. However, a judgment about the validity of these methods cannot be made if the study aim is ambiguous.

References

1. Hernán M, S H-D, Robins J. A Structural Approach to Selection Bias. Epidemiology 2004;15(5):615-25.
2. Choi J, Dekkers OM, le Cessie S. A comparison of different methods for handling measurements affected by medication use. (ready to submit).
3. Rubin DB. For objective causal inference, design trumps analysis. The Annals of Applied Statistics 2008;2(3):808-40, 33.
4. Carroll RJ, Ruppert D, Stefanski LA, et al. Measurement error in nonlinear models: a modern perspective. Chapman and Hall/CRC; 2006.
5. Hernán M, Robins J. Causal inference: What if. Boca Raton: Chapman & Hall/CRC; 2020.
6. Cole SR, Platt RW, Schisterman EF, et al. Illustrating bias due to conditioning on a collider. International Journal of Epidemiology 2009;39(2):417-20.
7. Hernán M. Invited Commentary: Selection Bias Without Colliders. American Journal of Epidemiology 2017;185(11):1048-50.
8. 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.
9. Elwert F, Winship C. Endogenous Selection Bias: The Problem of Conditioning on a Collider Variable. Annu Rev Sociol 2014;40:31-53.

10. 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.
11. Tanamas SK, Hanson RL, Nelson RG, et al. Effect of different methods of accounting for antihypertensive treatment when assessing the relationship between diabetes or obesity and systolic blood pressure. Journal of Diabetes and its Complications 2017;31(4):693-9.
12. VanderWeele TJ. Principles of confounder selection. European Journal of Epidemiology 2019;34(3):211-9.
13. 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.
14. Hutcheon JA, Chiolero A, Hanley JA. Random measurement error and regression dilution bias. BMJ 2010;340:c2289.

Online supplementary materials

Online supplementary materials 1 to 4 are available online at: https://github.com/Yeon-Choi-git/Chapter-6-online-materials

Chapter 7

Tying research question and analytical strategy when variables are affected by medication use

Published in Pharmacoepidemiol Drug Saf. 2023; 1- 10. doi:10.1002/pds.5599

Jungyeon Choi, Olaf M. Dekkers, Saskia le Cessie

Abstract

Ill-defined research questions could be particularly problematic in an epidemiological setting where measurements fluctuate over time due to intercurrent events, such as medication use. When a research question fails to specify how medication use should be handled methodologically, arbitrary decisions may be made during the analysis phase, which likely leads to a mismatch between the intended question and the performed analysis. The mismatch can result in vastly different or meaningless interpretations of estimated effects. Thus, a research question such as 'what is the effect of X on Y?' requires further elaboration, and it should consider whether and how medication use has affected the measurements of interest.

In our study, we will discuss how well-defined questions can be formulated when medication use is involved in observational studies. We will distinguish between a situation where an exposure is affected by medication use and where the outcome of interest is affected by medication use. For each setting, we will give examples of different research questions that could be asked depending on how medication use is considered in the estimand and discuss methodological considerations under each question.

Keywords
Research question; Medication effect; Well-defined question; Estimand; Causal inference;

Key points

- An overview is given of well-defined research questions that can be formulated in an epidemiological study where the exposure or the outcome values may be affected by medication use.
- Different ways of handling medication use in the analysis can lead to vastly different estimated effects with different interpretations.
- Some commonly used approaches, such as deleting patients using medication when the outcome is affected by medication, yield estimates which do not have a meaningful interpretation
- Researchers are advised to consciously set research questions and corresponding analytic strategies for handling medication use based on the clinical aims of the study.

Introduction

A well-defined research question is the cornerstone of research. Depending on the research question, different theoretical considerations and statistical analyses are required, and most importantly, estimated effects should be interpreted differently [1, 2]. Unfortunately, researchers may start performing statistical analyses before their research question is settled with sufficient detail. Analyses are done first, and the meaning of the estimated effect remains vague [3].

Ill-defined research questions are particularly problematic in an epidemiological setting where measurements fluctuate or change over time. Medication use is one important cause for this change, as it is prescribed to target specific measures. A research question that fails to specify how medication use should be handled methodologically may lead to arbitrary decisions during the analysis phase, and a subsequent mismatch between the intended research question and the performed analysis.

Suppose that different researchers are interested in the effect of blood pressure (BP) on myocardial infarction (MI) risk. Some researchers may exclude individuals using antihypertensive drugs. The result would be interpreted as the effect of BP on MI in the subset of medication non-users, and it may not be transportable to medication users. Others may be interested in untreated BP values and take a modelling approach to reconstruct BP values without medication; for example, by using methods to account for measurement error [4]. Again, others may ignore the medication information and consider the effect of observed BP, which might have been lowered by medications in the total population. Similar problems arise when blood pressure is studied as an outcome. Thus, a research question such as ‘what is the effect of X on Y?’ requires

further elaboration, and it should consider whether and how medication use has affected the measurements of interest.

Numerous authors in causal inference have stressed that exposures should be well-defined [5-8]. Moreover, the handling of intercurrent events in causal inference has recently achieved considerable attention. Young et al. have recently proposed a causal framework where they discuss different causal estimands under competing events. In the field of randomized trials, the European Medicines Agency (EMA) released a guideline proposing several different estimands for intercurrent events such as post-randomization medication use [9].

As practical guidance, several authors [4, 10-12] discussed statistical methods that could be used when measurements are affected by medication use. However, our recent review of the handling of medication use in medical papers [13] demonstrated that a majority of studies featured vaguely formulated research questions and unclear research aims. Invalid methods were often used, and a justification for the chosen method was rarely given. Despite the efforts to raise awareness, medication use as intercurrent events was overlooked in majority of reviewed papers.

Therefore, in this paper, we emphasize the importance of further elaborating on ostensibly straightforward research questions when the exposure or the outcome variable is affected by medication use. We describe several types of research questions of interest to applied researchers; some are formulated within the framework of causal inference, and others are more explorative in nature. When considering a cause, we take a practical pluralistic perspective; not only manipulable interventions but also 'states', such as having a certain level of blood pressure, can be studied as causes [14, 15]. We discuss how medication use is incorporated into each research question and which potential design considerations or methodological challenges may occur. Additionally, we warn against some common approaches to handling medication use that generally fail to yield interpretable results.

We start this paper by discussing a situation where an exposure, possibly time-varying, is affected by medication use by considering five different research aims. Following, we consider five research aims when the outcome of interest may be affected by medication. We conclude with a general discussion.

Situation 1: The exposure is affected by medication use

Imagine a researcher interested in the effect of blood pressure (BP) on the severity of COVID-19 in patients who just tested positive for the coronavirus. The time of the positive test is indicated by *t*. The outcome, severity of COVID-19, is measured at a certain moment after *t*. Individuals' BP levels have changed over time before time t, and some people have started using antihypertensive drugs at a certain moment before time *t*. Depending on research settings, BP may have been measured repeatedly before time *t* or only at *t*.

The initial research question, 'the effect of BP on the severity of COVID-19', is not well defined; it ignores the fact that BP varies over time and does not specify which BP values are of interest. For simplicity of the further discussion, let us assume three categories of study participants (Figure 1a). In category A, individuals had a high BP for a prolonged period and never used antihypertensive drugs. Individuals in category B also had a history of high BP but started using antihypertensive drugs before *t*. Thus, at time *t*, their blood pressure is lower than before taking the medication. In category C, individuals had normal blood pressure over time without medication. We use this example to discuss different possible research questions of interest. Throughout the paper, we assume that all confounding factors are measured and dealt with appropriately. Table 1 summarizes the different research questions.

Table 1. Summary of Section 1 (the exposure is affected by medication use) and Section 2 (the outcome is affected by medication use)

Section 1		
The interest is in	**Research question example**	**When or why**
the currently observed exposure value	What is the effect of the currently observed BP value on the severity of COVID-19?	BP values observed at a certain time point reflect a patient's health status.
the exposure trajectory before time *t*	What is the effect of the history of BP on the severity of COVID-19?	Regardless of antihypertensive medication use, history of BP values manifests an accumulated effect on the outcome.
the untreated exposure value	What is the effect of untreated BP at time t on the severity of COVID-19?	Untreated BP values at time *t* better reflect the medical condition than the observed BP after medication.
the effect of an intervention on the exposure	What would have happened if no one had been treated with antihypertensive drugs?	A causal effect of intervening on BP on the relationship between BP and the outcome is of interest.

Table 1. Summary of Section 1 (the exposure is affected by medication use) and Section 2 (the outcome is affected by medication use) *(continued)*

Section 1		
The interest is in	**Research question example**	**When or why**
the untreated population only	What is the effect of BP on the severity of COVID-19 among people who did not use antihypertensive drugs?	The subpopulation of medication non-users is of interest.
Section 2		
The interest is in	**Research question example**	**When or why**
the observed value of the outcome	What is the difference in observed BP at age 40 between individuals born with and without genetic factor?	The total effect of gene A on BP that may be partly mediated by using antihypertensive drugs is of interest.
the outcome value unaffected by medication use	What is the effect of the genetic factor A on BP at age 40 if no one had used antihypertensive drugs?	The biological effect of gene A on BP is of interest, and antihypertensive drug use is considered to have altered the effect of interest.
medication use as part of the outcome	What is the effect of the genetic factor A on the risk of hypertension at age 40?	The fact that a person started using antihypertensive medication is a part of the outcome.
in the outcome values while being untreated	What is the difference in BP between individuals born with and without genetic factor A while being untreated?	Only the measurements before treatment may be of interest. More meaningful in situations where measurement after intercurrent events is undefined; i.e., quality of life between the treatment group compared over time only in those still alive.
the untreated population	What is the difference in BP between individuals born with and without genetic factor A in those untreated at age 40?	It resembles a per-protocol analysis of an RCT Questionable whether this approach corresponds to any sensible and clinically relevant estimand.

Figure 1. Visual representations of research questions 1.1 to 1.5 (: medication use) We consider three persons; person A developed high blood pressure (BP) before time t but did not start medication, person B developed high BP and started medication, and person C had low BP.

1.1 The interest is the currently observed exposure value

It may occur that BP values observed at a certain time point reflect a patient's health status. In this case, one may ask: w*hat is the effect of the currently observed BP value on the severity of COVID-19*? This question hypothesizes that the current BP value determines COVID-19 severity; for example, people with higher BP values are at a higher risk (e.g., because of inflammation or vessel wall stress), and people with lower values (whether controlled naturally or by antihypertensive medication) are at a lower risk. This is illustrated in Figure 1a, where the BP measurements as they are observed at time *t* are used as the exposure in the analysis. The analysis here is relatively straightforward. In principle, medication use does not need to be added as an extra variable in the model unless the medication affects the outcome independently of blood pressure (i.e., medication use is a confounder).

1.2 The interest is the exposure trajectory before time *t*

Researchers may hypothesize that the history of BP values may affect a certain health outcome. They may be interested in whether COVID-19 patients with a history of high BP in the last 12 months are at greater risk than comparable patients with a history of lower BP. This translates into the following research question: *what is the effect of the history of BP on the severity of COVID-19?* This implies that the history of BP values, regardless of antihypertensive medication use, manifests an accumulated effect on the outcome. To address this research question, repeated measurement of BP is required to estimate the trajectories of BP for each individual (see Figure 1b).

Still, the "effect of the history of BP" is vaguely defined and needs to be specified. For example, one could be interested in the cumulative BP values during a certain period before *t* (estimated by the area under the curve), the mean value of BP in a specific period, or the increase in BP over a certain period. In any case, the length of the period of interest before time *t* should be well defined. Notably, medication use is not added as a variable in the model, but the effect of medication use is incorporated in the analysis through its effect on subsequent BP levels. Furthermore, in this scenario, confounders should be measured at the time when the follow-up starts.

1.3 The interest is the untreated exposure value

In a third scenario, it may be hypothesized that the untreated exposure values at time *t* better reflect the medical condition of interest than the observed exposure value after medication. For example, a history of high BP may alter vessel wall conditions. While antihypertensive medication may quickly alleviate one's BP level, it takes a longer period for the damaged vessel wall to recover. If vessel wall condition affects COVID-19 severity, BP values measured shortly after treatment initiation are less informative than pre-treatment values. In this case, for those who started treatment in a certain time frame before time t, BP measurements that would have been observed under no treatment can be a proxy for the unmeasured vessel wall difference. The corresponding

research question here would be: *what is the effect of untreated BP at time t on the severity of COVID-19?* The effect of an intervention on BP, directly applicable in medical decision-making, is not under inquiry here. However, the intended research question could provide a valuable etiologic perspective [15].

Answering question 1.3 is not straightforward because the BP level without treatment at time *t* is unobserved for treated individuals. When repeated BP measurements are available, measurements before medication use could be used. For example, as depicted in Figure 1c, we may use the last BP measurement of person B before starting medication as a proxy for the untreated value at time t or extrapolate the untreated BP trajectory of B until time t (under the assumption that individual A and B are exchangeable with respect to BP trajectory). When no previous BP measurements are available, external information on the effect of medication and/or the prescription process is needed to reconstruct the untreated BP at time *t*. For instance, the mean and standard deviation of medication effect can be acquired from randomized control trials. These parameters can be used in a regression calibration method to reconstruct the untreated BP with the uncertainty around it.

If treatment started not long before *t*, such research questions seem especially sensible. Yet when there is a mixture of long-term and short-term medication users, it becomes more complicated; for example, the antihypertensive drug may have improved the vessel wall condition in long-term medication users. When this is the case, time since medication use should be incorporated into the analysis.

One simple solution to answer question 1.3 could be to remove individuals on medication from the analysis. However, when there are many medication users, the estimated effect may be less precise. Furthermore, if there is an effect modification by BP medication use or other characteristics associated with medication use, the average effect in the untreated subpopulation may differ from the average effect in the total population. Finally, one should be aware that selection bias may occur if medication users differ from non-users with high BP in terms of other characteristics and this should be properly accounted for [16].

1.4 Interest in the effect of an intervention on the exposure

The previous sections 1.2 and 1.3 are not anchored to a clear time zero, as the time of starting medication use may differ between patients. The questions are, therefore, not formulated sharply enough to fit within a causal inference framework. In this section, we consider how causal research questions can be formulated as interventions on BP before time t. For example, we may wonder *what would have happened if no one had been treated with antihypertensive drugs*. Alternatively: *what would have been the effect of BP on COVID-19 severity if we had intervened on everyone with high BP with antihypertensive drugs?* While Section 1.3 is interested in the (unobserved) untreated BP values at one particular

time point, Section 1.4 considers the effect of intervening on BP on the relationship between BP and the outcome.

These types of research questions consider hypothetical intervention scenarios as the untreated BP level at time *t,* and the corresponding untreated outcome is unobserved for treated people. Similarly, the BP level and the outcome under treatment are unobserved for people untreated for their high BP. These research questions can be formulated in a counterfactual framework using the concept of a target trial [5, 17, 18]. In a target trial, a study population would be defined at time *t*0 when the follow-up starts, and confounders would also be measured at time *t*0. In our example, *t*0 could be one year before the start of the COVID-19 epidemic. The interventions of interest may be, for example, "prescribe medication if BP is above a certain level" versus "prescribe no medication at all, even if BP is high". People are followed until they are infected by COVID-19 and experience severe or less severe symptoms of COVID-19. There are several approaches for estimating the effect of a possible time-varying intervention (see Hernan and Robins[5], chapter 21 for an overview), such as the use of inverse probability weighting [19, 20]. Ideally, all individuals should be followed from the beginning of the trajectory to the final measures; otherwise, loss to follow-up should be accounted for, for example, by using censoring weights [20, 21].

1.5 The interest is the untreated population only

Another aim may be to estimate the effect of BP on the severity of COVID-19 *among people who did not use antihypertensive drugs*. To answer this research question, one would restrict the analysis to individuals without medication use, as illustrated in Figure 1e. While previous questions are interested in the *total* population, the interest here is the subpopulation of medication non-users. Individuals in this subpopulation might be under antihypertensive treatment and may be more likely to have higher BMI. The subpopulation could therefore have different characteristics than the total population. If BMI were an effect modifier for the association between BP and the severity of COVID-19, the estimated effect would only be valid for the population untreated at time *t*.

Selection bias may occur if one does not adjust for confounding between medication use and the outcome. Individuals using antihypertensive medication could, for instance, be more health-conscious than individuals with untreated high BP. This implies that health-conscious people with high blood pressure will be underrepresented in the selected subpopulation of medication non-users. Therefore, health consciousness should be adjusted for in the analysis [16].

The appendix displays simple numerical examples of each research aim depicted in Section 1.1 to 1.5.

Situation 2: The outcome is affected by medication use

Now, let us consider a scenario where we have an exposure determined at a certain time (t_0) and a continuous outcome that could change throughout life. During follow-up, the outcome levels of some individuals may have been influenced by medication use. For this example, we pick the exposure to be genetic variant A rather than a treatment or another intervention to avoid confusion with the intercurrent medication use. The outcome is BP. In our example, the follow-up starts at adulthood (t_0), and some individuals with high BP have started using antihypertensive drugs between time t_0 and *t, t* being the end of the follow-up.

As an illustration, we consider four hypothetical individuals in Table 2 and Figure 2a. Person a1 and b1 were both born with gene A, which causes high BP. Individual b1 starts using medication. Person a0 and b0 are identical to a1 and b1, respectively, except that they both were born without the gene and did not develop high BP. Person a0 and b0 share identical characteristics, and the difference in Figure 2 only reflects random inter-variability. A summary of the research interests is given in Table 1.

Table 2. Four different hypothetical individuals under a scenario where the interest is estimating the effect of the gene A on blood pressure (BP) at time t, while some individuals started antihypertensive medication use before time t.

Individual	Genetic variant	BP before time t	Medication use
a1	Gene A	High	No
b1	Gene A	High	Yes
a0	No gene A	Low	No
b0	No gene A	Low	No

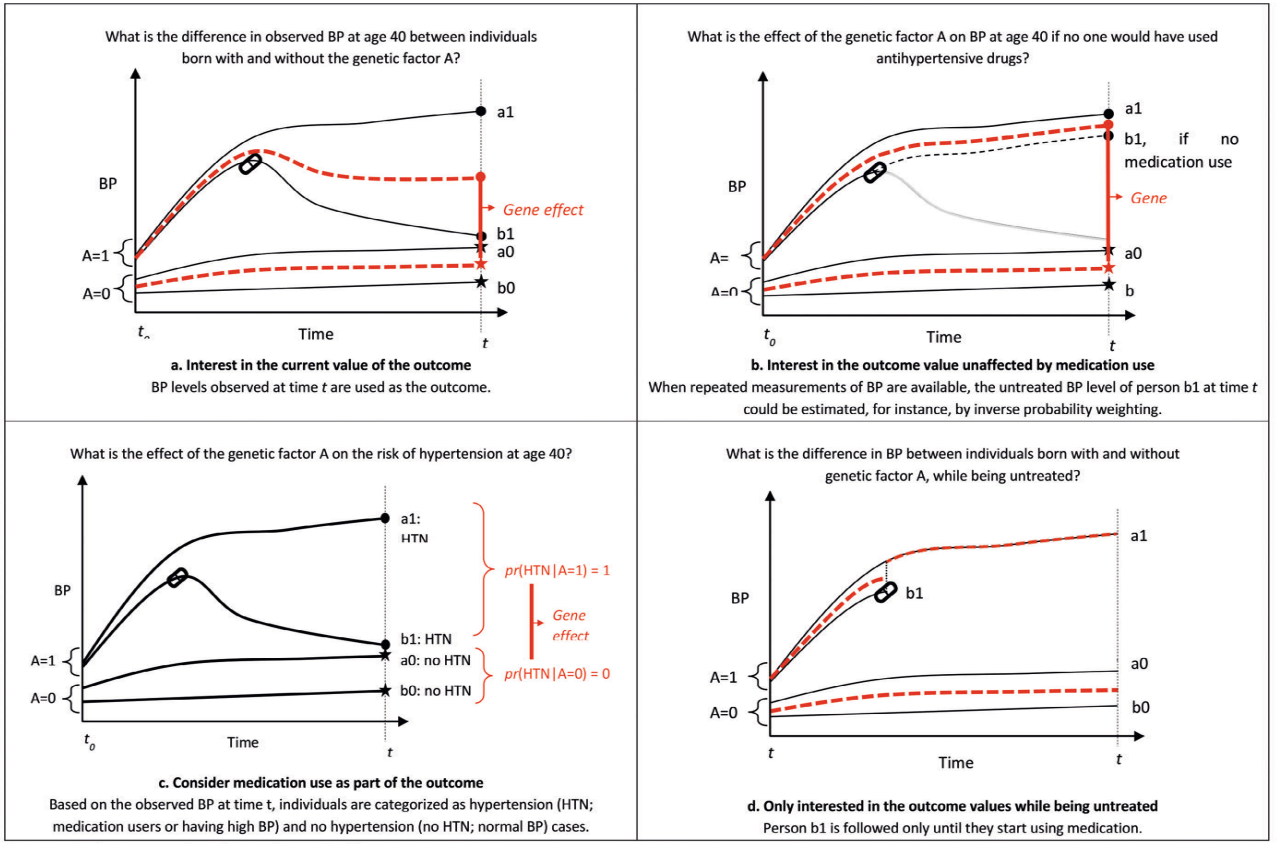


Figure 2. Visual representations of research question from 2.1 to 2.4

Four individuals are assigned to having a certain genetic factor A (A=1) or not (A=0) at time t0. The follow-up starts at time t0 and continues until time t, where t0 represents the start of adulthood and t represents the age of 40. Black lines represent the BP trajectories of the individuals in Table 2. Red dotted lines represent the average BP levels in each exposure group (the upper line for group A=1 and the lower line for group A=0). Person a0 and b0 share identical characteristics; the difference only reflects random inter-variability.

2.1 The interest is the observed value of the outcome

Firstly, the BP levels as observed can be the outcome of interest (Figure 2a). For example, we may want to compare observed BP levels at age 40 of individuals with gene A to similar individuals born without the gene. In this type of research question, one is interested in the total effect of the exposure on the outcome; that is, an effect that may be partly mediated by using antihypertensive drugs. In counterfactual notation, we are interested in the average total effect of A on the outcome: $E[Y^{A=1}] - E[Y^{A=0}]$, where $Y^{A=1}$ is the potential outcome when setting A to 1 and $Y^{A=0}$ is the potential outcome when setting A to 0. Young et al. referred to this contrast as the "effect without elimination of competing events". In the clinical trial context [9], this is referred to as "*treatment policy strategy-estimand"* [9]. The principle of such analysis corresponds to an intention-to-treat analysis in an RCT, as the data is analyzed using the observed outcomes ignoring any intercurrent event or protocol deviation. Therefore, under question 2.1, medication use would be ignored in the analysis.

2.2. The interest is the outcome value unaffected by medication use

Alternatively, the interest could be the biological effect of gene A on BP, where antihypertensive drug use may alter this effect. Here we would ask research questions such as, *what is the effect of the genetic variant A on BP at age 40 if no one would have used antihypertensive drugs?* In counterfactual notation, we are interested in the effect: $E[Y^{A=1,med=0}] - E[Y^{A=0,med=0}]$, with $Y^{A=1,med=0}$ the potential outcome of Y when A is set to 1 and no medication would have been used. This is called "the effect under elimination of competing events" [22]. In a clinical trial context, it is referred to as "*hypothetical strategy-estimand"* [9]. Figure 2b depicts this scenario.

Suppose repeated measurements of BP are available and all factors influencing medication use are measured. In that case, the estimand can be estimated using repeated measurement methods, such as linear mixed models or generalized estimation equation methods with inverse probability weighting [5, 21]. The BP levels after medication use will not be used in these analyses. If no repeated measurements of BP are available, other methods for handling an outcome variable affected by medication use, such as adding the mean medication effect to the treated measurements or fitting a censored regression model [4, 10-12, 23, 24] may be used.

2.3. Considering medication use as part of the outcome

Medication use can be incorporated into the definition of the outcome when the use of antihypertensive medication provides information about a person's condition. For example, we may use hypertension (yes/no) as a dichotomous outcome. The research question then is: *what is the effect of the genetic factor A on the risk of hypertension at age 40*? In this case, the outcome is dichotomized into *hypertension* (high BP and/or using antihypertensive medication) and *no hypertension* (normal BP and no medication use). This is illustrated in Figure 2c. In other scenarios, using an ordinal scale could be an

alternative (e.g., categorizing fasting glucose level into *normal glucose, impaired glucose,* and *diabetes*, where diabetes is defined as glucose level above a certain level or use of diabetes medication). In clinical trials, this type of scenario is called "*composite variable strategy-estimand"*.

2.4. Only interested in the outcome values while being untreated

In Section 2.2, the interest was in the effect of the gene on untreated BP measurements in the total population. In this section and Section 2.5, we consider two strategies that restrict the population based on medication use. Sometimes only the measurements before treatment may be of interest. In that case, one could compare outcomes between the exposure groups at each time point using only the individuals still untreated at that time. In other words, comparing different exposure groups conditionally on being untreated (Figure 2d). This approach may be called the "while untreated strategy", analogous to the EMA guideline where the "while on treatment-estimand" and "while alive-estimand" are discussed.

In general, this comparison will not answer a causal research question because of selection bias; the comparison only involves individuals who are still untreated at the time of comparison. Suppose people born with the genetic variant A (exposed group) are more likely to use antihypertensive drugs. As time passes, more people in the exposed group will be excluded from the comparison, and the remaining individuals in the exposed and unexposed groups are no longer comparable. This issue will arise even if the groups are exchangeable at baseline.

However, combined with comparing the percentages of individuals starting medication, this comparison may still yield valuable clinical information. It provides an answer to a combination of two questions: i) what is the effect of the genetic factor A on the probability of starting antihypertensive medication, and ii) what is the difference in blood pressure levels effect in those still untreated at the time of comparison? These types of combined questions occur, for example, in quality-of-life studies in cancer research, where the quality-of-life measurements are compared over time only in those still alive at that time because the quality of life after death is undefined [25, 26].

When persons can go on and off treatment (treatment episodes), defining a "while untreated strategy" becomes even more complicated, as also measurements in an untreated period after a period of taking the drug may be considered in some instances as "while untreated". The definition of "while untreated" should in this case, be carefully considered with the clinical context in mind.

2.5 Could the interest be only in the untreated population?

Some studies exclude all measurements of individuals who started medication during follow-up from their analysis, including the measurements before starting medication

use. A difference with Section 2.4 is that here the measurements before medication use are removed as well.

This approach resembles a per-protocol analysis of an RCT where only the participants who completed the follow-up without protocol deviation are included in the analysis [27]. Defining whether an individual belongs to a population of interest (i.e., people who are untreated at any time point) based on an event happening after the follow-up started (i.e., medication use) is risky. If the follow-up time increases, more people will start using medication and consequently be excluded from the comparisons, even for the time before using medication. Consequently, this approach can lead to substantial selection bias [28, 29]. It is questionable whether this approach corresponds to any sensible and clinically relevant estimand.

Discussion

Clinical measurements affected by medication use are commonly encountered in epidemiological research. In this paper, we discussed different research questions that could be of interest when the exposure or the outcome variable is affected by medication use. We argued that each question is driven by different assumptions and clinical aims. Concurrently, each requires a tailored strategy for handling medication use in the analysis. Even with causal inference experts emphasizing the importance of well-defined research questions, the role of medication use is often overlooked, resulting in arbitrary decisions regarding its handling in statistical analysis and vague interpretations of its estimated effects.

Some causal inference experts may argue that BP is not an *intervention* due to its nature of having multiple ways to be manipulated and, therefore, cannot be studied causally [30, 31]. In practice, however, *states* such as having a certain level of BP or glucose are frequently studied as causal risk factors, and they can provide valuable etiological knowledge. In this paper, therefore, we took a practical pluralistic perspective based in research practice and also discussed research questions that are not directly causal interventional.

Still, emulating a target trial can greatly help in crystallizing a research question and choosing a valid analytical strategy [18, 32]. A vaguely defined exposure or outcome variable would not be acceptable in RCTs. For RCTs, protocols are written in advance and demand a clear research question and a detailed statistical analysis plan. A definition of the treatment or the outcome would (and should) not change based on arbitrary decisions made during an analysis phase. Deciding how to handle medication use at the stage of formulating a research question applies equally to observational

studies. For this reason, the importance and benefits of writing a protocol and defining a target estimand prior to conducting observational studies have been stressed [33, 34]. Connecting a research question to a target trial could also contribute to identifying potential sources of bias. For instance, question 2.5 would be analogous to being interested in the effect only in the participants adhering to the protocol until the end of a randomized trial. Compared to an RCT setting, it becomes clear that this type of research question would suffer from selection bias and would rarely yield clinically meaningful results.

One of the estimands mentioned by the EMA is the "principal stratum-estimand", which is the effect in subpopulations where a particular intercurrent event would or would not occur. In our example, a principal stratum could be individuals who would not use hypertension medication when their blood pressure would be elevated (e.g., because they have an aversion to medication or are not aware that their BP is too high). We decided not to discuss this in detail as research questions using potential medication use to define a subpopulation are rarely considered. The corresponding analysis is challenging because whether a person is a medication non-user can be observed only if their BP becomes high during the follow-up.

Situations with medication use can be much more complex as multiple medications can be used simultaneously and/or switching between medications may occur. It is also possible that both the exposure and the outcome measurements are affected by medication use. In addition to medication use, behavioral changes (e.g., starting exercising regularly to regulate high BP) after the baseline could also affect measurements of interest. Needless to say, examples are not limited to blood pressure and blood pressure medication but could be other measurements, such as glucose or lipid levels, and other types of drugs. Numerous sources of potential bias outside those discussed in this paper should be critically considered as well (e.g., how to properly adjust for confounding or ill-defined time zero of follow-up: immortal time bias) [28, 35].

The complexity of the situation, however, should not discourage tackling the problem of measurements affected by medication use. Rather, it requires additional caution when defining research questions and more rigorous planning on how medication should be handled in the analysis. In any given case, we advise researchers to consciously set a research question and corresponding analytic strategy for handling medication use based on the clinical aim and underlying assumptions.

Acknowledgment
We thank Dr. Sonja Swanson (Erasmus University Medical Center) for a valuable discussion and for providing comments on the manuscript.

References

1. Thabane L, Thomas T, Ye C, Paul J. Posing the research question: not so simple. Canadian Journal of Anesthesia/Journal canadien d'anesthésie. 2008;56(1):71. doi:10.1007/s12630-008-9007-4
2. Bragge P. Asking good clinical research questions and choosing the right study design. Injury. 2010;41:S3-S6. doi:https://doi.org/10.1016/j.injury.2010.04.016
3. Goetghebeur E, le Cessie S, De Stavola B, Moodie EE, Waernbaum I, initiative obottgCIotS. Formulating causal questions and principled statistical answers. Statistics in Medicine. 2020;39(30):4922-48. doi:https://doi.org/10.1002/sim.8741
4. Choi J, Dekkers OM, le Cessie S. A comparison of different methods for handling measurements affected by medication use. medRxiv. 2022:2022.04.23.22273899. doi:10.1101/2022.04.23.22273899
5. Hernán M, Robins J. Causal inference: What if. Boca Raton: Chapman & Hall/CRC; 2020.
6. Maldonado G, Greenland S. Estimating causal effects. International Journal of Epidemiology. 2002;31(2):422-9. doi:10.1093/ije/31.2.422
7. VanderWeele TJ. On Well-defined Hypothetical Interventions in the Potential Outcomes Framework. Epidemiology. 2018;29(4):e24-e5. doi:10.1097/ede.0000000000000823
8. Cole SR, Frangakis CE. The Consistency Statement in Causal Inference: A Definition or an Assumption? Epidemiology. 2009;20(1):3-5. doi:10.1097/EDE.0b013e31818ef366
9. 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.
10. Tobin MD, Sheehan NA, Scurrah KJ, Burton PR. Adjusting for treatment effects in studies of quantitative traits: antihypertensive therapy and systolic blood pressure. Statistics in Medicine. 2005;24(19):2911-35. doi:doi:10.1002/sim.2165
11. 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. doi:10.1002/pds.3876
12. 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. doi:10.1002/sim.1408
13. Choi J, Dekkers OM, le Cessie S. How measurements affected by medication use are reported and handled in observational research: A literature review. Pharmacoepidemiology and Drug Safety. 2022. doi:https://doi.org/10.1002/pds.5437
14. Vandenbroucke JP, Broadbent A, Pearce N. Causality and causal inference in epidemiology: the need for a pluralistic approach. International Journal of Epidemiology. 2016;45(6):1776-86. doi:10.1093/ije/dyv341
15. Glymour C, Glymour MR. Commentary: Race and Sex Are Causes. Epidemiology. 2014;25(4):488-90. doi:10.1097/ede.0000000000000122
16. Hernán M, S H-D, Robins J. A Structural Approach to Selection Bias. Epidemiology. 2004;15(5):615-25.
17. 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. doi:10.1007/s10654-020-00636-1
18. Labrecque JA, Swanson SA. Target trial emulation: teaching epidemiology and beyond. European Journal of Epidemiology. 2017;32(6):473-5. doi:10.1007/s10654-017-0293-4

19. Mansournia MA, Etminan M, Danaei G, Kaufman JS, Collins G. Handling time varying confounding in observational research. BMJ. 2017;359:j4587. doi:10.1136/bmj.j4587

20. Cain LE, Cole SR. Inverse probability-of-censoring weights for the correction of time-varying noncompliance in the effect of randomized highly active antiretroviral therapy on incident AIDS or death. Statistics in Medicine. 2009;28(12):1725-38. doi:https://doi.org/10.1002/sim.3585

21. Robins J, Hernán M, Brumback B. Marginal Structural Models and Causal Inference in Epidemiology. Epidemiology. 2000;11(5).

22. Young JG, Stensrud MJ, Tchetgen Tchetgen EJ, Hernán MA. A causal framework for classical statistical estimands in failure-time settings with competing events. Statistics in Medicine. 2020;39(8):1199-236. doi:https://doi.org/10.1002/sim.8471

23. 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. doi:10.1177/0962280216680240

24. Tanamas SK, Hanson RL, Nelson RG, Knowler WC. Effect of different methods of accounting for antihypertensive treatment when assessing the relationship between diabetes or obesity and systolic blood pressure. Journal of Diabetes and its Complications. 2017;31(4):693-9. doi:https://doi.org/10.1016/j.jdiacomp.2016.12.013

25. Kurland BF, Johnson LL, Egleston BL, Diehr PH. Longitudinal Data with Follow-up Truncated by Death: Match the Analysis Method to Research Aims. Stat Sci. 2009;24(2):211. doi:10.1214/09-sts293

26. Kurland BF, Heagerty PJ. Directly parameterized regression conditioning on being alive: analysis of longitudinal data truncated by deaths. Biostatistics. 2005;6(2):241-58. doi:10.1093/biostatistics/kxi006

27. ICH E9 Statistical Principles for Clinical Trials. European Medicines Agency; 1998.

28. Suissa S. Immortal Time Bias in Pharmacoepidemiology. American Journal of Epidemiology. 2007;167(4):492-9. doi:10.1093/aje/kwm324

29. Hernán MA, Sauer BC, Hernández-Díaz S, Platt R, Shrier I. Specifying a target trial prevents immortal time bias and other self-inflicted injuries in observational analyses. Journal of Clinical Epidemiology. 2016;79:70-5. doi:https://doi.org/10.1016/j.jclinepi.2016.04.014

30. Hernán M. Invited Commentary: Selection Bias Without Colliders. American Journal of Epidemiology. 2017;185(11):1048-50. doi:10.1093/aje/kwx077

31. Hernán MA, Taubman SL. Does obesity shorten life? The importance of well-defined interventions to answer causal questions. International Journal of Obesity. 2008;32(3):S8-S14. doi:10.1038/ijo.2008.82

32. García-Albéniz X, Hsu J, Hernán MA. The value of explicitly emulating a target trial when using real world evidence: an application to colorectal cancer screening. European Journal of Epidemiology. 2017;32(6):495-500. doi:10.1007/s10654-017-0287-2

33. Vandenbroucke JP, von Elm E, Altman DG, et al. Strengthening the Reporting of Observational Studies in Epidemiology (STROBE): explanation and elaboration. Int J Surg. 2014;12(12):1500-24. doi:10.1016/j.ijsu.2014.07.014

34. Yang W, Zilov A, Soewondo P, Bech OM, Sekkal F, Home PD. Observational studies: going beyond the boundaries of randomized controlled trials. Diabetes Research and Clinical Practice. 2010;88:S3-S9. doi:https://doi.org/10.1016/S0168-8227(10)70002-4

35. Lévesque LE, Hanley JA, Kezouh A, Suissa S. Problem of immortal time bias in cohort studies: example using statins for preventing progression of diabetes. BMJ. 2010;340:b5087. doi:10.1136/bmj.b5087

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.

Appendix

Dutch Summary

In dit proefschrift onderzoeken wij de potentiële uitdagingen die de validiteit van observationeel epidemiologisch onderzoek kunnen aantasten. Deze potentiële bronnen van vertekening, zoals confounding, ontbrekende gegevens, selectiebias en meetfouten, vormen aanzienlijke obstakels voor een accurate interpretatie van onderzoeksresultaten. Hoewel talrijke methoden zijn ontwikkeld om het effect van vertekeningen te verminderen, blijft het bepalen van de meest geschikte benadering voor specifieke empirische contexten een complexe taak. Bovendien krijgen problemen die worden besproken in de methodologische literatuur, in klinisch onderzoek vaak beperkte aandacht. Daarom analyseren we de problemen omtrent ontbrekende gegevens, selectiebias en meetfouten die optreden in diverse specifieke observationele situaties, en bespreken we de meest geschikte methoden om deze kwesties te behandelen.

In **Hoofdstuk 2** hebben wij systematisch onderzocht hoe covariabelen met ontbrekende gegevens op de meest optimale manier in een propensity score-analyse kunnen worden behandeld door verschillende simulatiescenario's te genereren. Hierbij hebben we rekening gehouden met diverse mechanismen voor ontbrekende gegevens en de aanwezigheid van modificatie van het behandelingeffect. Onze bevindingen tonen aan dat er geen universeel optimale aanpak is, aangezien de geschikte methode afhangt van de datastructuur, zoals het ontbrekende mechanisme, effectheterogeniteit en ongemeten confounding. Opmerkelijk is dat een complete case-analyse of het toevoegen van ontbrekende indicatoren aan een model, die doorgaans als 'naïef' en ongeschikt voor het omgaan met ontbrekende gegevens beschouwd worden, beter presteerden dan meervoudige imputatie bij niet-random ontbrekende gegevens. Meervoudige imputatie was het meest effectief wanneer gegevens willekeurig ontbraken, maar alleen wanneer het imputatiemodel correct gespecificeerd was. Dit impliceert dat de uitkomstvariabele in het imputatiemodel moet worden opgenomen en, indien van toepassing, een interactieterm om rekening te houden met heterogeniteit in het behandeleffect.

In **Hoofdstuk 3** hebben we methoden onderzocht voor het opsporen van meetfouten die mogelijk te wijten zijn aan het verdunnen van monsters in tijdreeks hormoongegevens. We vergeleken vier methoden: i) beoordeling door fysiologische experts, dat als gouden standaard gezien kan worden, ii) de stapsgewijze aanpak, die fysiologische kennis integreert met op standaarddeviatie gebaseerde detectie, iii) de methode van Tukey's fences, die fouten identificeert op basis van interkwartielafstanden, en iv) het verwachting-maximalisatie (EM) algoritme, dat de verdelingen van hormoonspiegels met en zonder fout wiskundig onderscheidt. Op basis van de prestaties in de praktijk en gesimuleerde gegevens, concludeerden we dat de stapsgewijze methode, die gebruikmaakt van fysiologische kennis, beter presteerde dan volledig geautomatiseerde datagedreven methoden zoals Tukey's fences en het EM-algoritme.

In **Hoofdstuk 4** zijn strategieën voor het omgaan met variabelen die worden beïnvloed door medicijngebruik onderzocht, voor de situatie waarbij interesse is effecten zonder behandeling. Een voorbeeld is het onderzoeken van de invloed van een genetische factor op de bloeddruk op een bepaalde leeftijd of de invloed van bloeddruk op het risico op hart- en vaatziekten zonder gebruik van antihypertensiva bij personen met hypertensie. Uit simulaties bleek dat de keuze van de methode afhangt van of de beïnvloede variabele de blootstelling, de uitkomst of een confounder is. Als de blootstelling wordt beïnvloed, kan het reduceren van de onderzoekspopulatie tot onbehandelde individuen tot een valide resultaat leiden. Echter, als er effectheterogeniteit aanwezig is, kunnen de resultaten mogelijk niet naar de algemene populatie worden geëxtrapoleerd. Als externe kennis over het gemiddelde en de standaardafwijking van het medicijneffect bekend is, kan regressiecalibratie met het toevoegen van het gemiddelde medicijneffect aan de behandelde waarden worden gebruikt. Als de uitkomst wordt beïnvloed, leiden simpele gangbare methoden tot vertekening. In plaats daarvan is het optellen van het gemiddelde medicijneffect of het gebruik van gecensureerde normale regressie geschikt. Op basis van de resultaten adviseren we onderzoekers om kritisch na te denken over het proces van medicijnvoorschrijving, de aanwezigheid van effectheterogeniteit en welke informatie over medicijneffecten beschikbaar is.

Verschillende methoden die in **Hoofdstuk 4** worden besproken, vereisen externe kennis over het geschatte effect van medicatie en in sommige gevallen de standaardafwijking van het medicijneffect. Deze gegevens kunnen uit gerandomiseerde gecontroleerde onderzoeken (RCT's) verkregen worden. Populaties in dergelijke onderzoeken zijn vaak niet representatief voor de populaties in observationeel onderzoek. Door verschillen tussen klinische settings in RCT's en praktijksettings, kan het toepassen van het medicijneffect verkregen in RCT's in observationele settings vertekening geven.

In **Hoofdstuk 5** hebben we veranderingen in glucose- en HbA1c-niveaus na het gebruik van glucoseverlagende medicatie beschreven met behulp van routinematig verzamelde gegevens van deelnemers aan de Nederlandse Epidemiologie van Obesitas (NEO) studie. Elektronische Patiëntendossiers werden gebruikt om nieuwe gevallen van diabetes en herhaalde metingen van glucose en HbA1c te identificeren. Lineaire gemengde modellen werden toegepast, waarbij tijd als een categorische variabele werd gebruikt voor zowel vaste als willekeurige effecten. Om regressie naar het gemiddelde te beperken, werd het referentiepunt ingesteld op 6 tot 12 maanden vóór het voorschrijven van medicatie. De resultaten toonden aan dat de meest aanzienlijke medicatie-effecten optraden 6 tot 12 maanden na de start van de behandeling, en dat de medicatie ook na twee jaar effectief bleef. Deze effecten waren echter kleiner dan die waargenomen in RCT's. We zagen verhoogde glucose- en HbA1c-waarden kort voor het gebruik van medicatie, wat mogelijk wijst op willekeurig hoge metingen die leiden tot behandelingsbeslissingen en daaropvolgende regressie naar het gemiddelde.

In **Hoofdstuk 6** is systematisch in de wetenschappelijke literatuur onderzocht hoe variabelen die door medicatie worden beïnvloed, worden behandeld in klinisch onderzoek. Hierbij kwamen we ambiguïteit in onderzoeksdoelen en het frequente gebruik van ongeldige methoden tegen. Geavanceerde methodologieën werden zelden toegepast, wat wijst op een behoefte aan meer kennis onder klinische onderzoekers en de noodzaak van een duidelijke onderzoeksvraag met betrekking tot medicijngebruik.

In **Hoofdstuk 7** zijn verschillende manieren onderzocht om medicijngebruik op te nemen in onderzoeksvragen wanneer blootstelling- of uitkomstvariabelen worden beïnvloed door medicatie bij sommige individuen. We bespreken aannames over de relaties tussen variabelen van belang en de mogelijke klinische relevantie daarvan. Sommige vragen passen in een causaal kader, waar het nabootsen van gerandomiseerd onderzoek de onderzoeksvraag kan verhelderen. Andere vragen zijn mogelijk niet causaal, maar kunnen wel etiologische inzichten bieden. Medicijngebruik moet voor elk type vraag op een andere manier worden behandeld en de overwegingen voor het omgaan met medicatie en methodologische kwesties variëren per vraag.

In **Hoofdstuk 8** concludeerden wij dat er geen universeel toepasbare methode bestaat om de problemen van confounding, ontbrekende gegevens, selectiebias en meetfouten in observationele onderzoekssettings te verminderen. Analytische beslissingen zouden niet als vanzelfsprekend moeten worden beschouwd en elke bron van vertekening moet worden behandeld op basis van contextspecifieke kennis. Een constant streven naar het verbinden van de methodologische en klinische werelden en het verbreden van de perspectieven op het omgaan met biases zal bijdragen aan de validiteit van observationeel onderzoek.

A

Publications

Liu, L. Choi, J., Musoro, J.Z., Sauerbrei, W., Amdal, C.D., Alanya, A., Barbachano, Y., Cappelleri, J.C., Falk, R.S., Fiero, M.H., Regnault, Reijneveld, J.C., Sandin, R., Thomassen, Roychoudhury, S., Goetghebeur, E., le Cessie, S.; on behalf of work package 3 of the Setting International Standards in Analyzing Patient-Reported Outcomes and Quality of Life Endpoints Data Consortium (SISAQOL). Single-arm studies involving patient-reported outcome data in oncology: a literature review on current practice. *Lancet Oncology.* (2023). 24(5). doi:10.1016/S1470-2045(23)00110-9

Choi, J., Dekkers, O.M., le Cessie, S. Tying research question and analytical strategy when variables are affected by medication use. *Pharmacoepidemiol Drug Saf.* (2023). 32(6): 661-670. doi:10.1002/pds.5599

Choi, J., Dekkers, O.M., le Cessie, S. How measurements affected by medication use are reported and handled in observational research: A literature review. *Pharmacoepidemiol Drug Saf.* (2022). 31(7): 739- 748. doi:10.1002/pds.5437

Choi, J., Dekkers, O.M. & le Cessie, S. Authors' Reply: A comparison of different methods to handle missing data in the context of propensity score analysis. *Eur J Epidemiol.* (2020). 35, 89–91. doi: 0.1007/s10654-019-00553-y

van der Spoel, E., Choi, J., Roelfsema, F., le Cessie, S., van Heemst, D., Dekkers, O.M. Comparing Methods for Measurement Error Detection in Serial 24-h Hormonal Data. *Journal of Biological Rhythms*. (2019). 34(4): 347-363. doi:10.1177/0748730419850917

Choi, J., Dekkers, O.M. & le Cessie, S. A comparison of different methods to handle missing data in the context of propensity score analysis. *Eur J Epidemiol.* (2019). 34, 23–36. doi:10.1007/s10654-018-0447-z

To be submitted

Choi, J., de Mutsert, R., van der Velde, J.H.P.M., van Eekelen, E., Dekkers, O.M., le Cessie, S. Estimating medication effects using routinely collected electronic health records: changes in blood glucose and HbA1c levels after glucose-lowering medication prescription in the Netherlands Epidemiology of Obesity study participants

Choi, J., Dekkers, O.M., le Cessie, S. A comparison of different methods for handling measurements affected by medication use

Curriculum Vitae

Jungyeon Choi was born in Seoul, South Korea, on the 14th of September 1990. She obtained a bachelor in Economics and Psychology from Sungkyunkwan University in 2014. During her bachelor's program, she did a semester at Leiden University as an exchange student. She pursued a master's degree in Methodology and Statistics in Psychology at Leiden University. She worked on her master's thesis at the Nederlands Forensisch Instituut. After obtaining her master's degree, she started a Ph.D. under the supervision of Prof. dr. S. le Cessie and Prof. dr. O.M. Dekkers, at the department of Clinical Epidemiology, LUMC. Her research addressed commonly encountered biases in observational studies and compared available statistical methods. While finalizing her Ph.D., she worked as a researcher for a consortium, 'Setting International Standards for Analyzing Quality of Life (SISAQOL)'. Afterward, she worked as a data scientist at Julius Clinical, Zeist, contributing to a VAC4EU project. From May 2023, Jungyeon Choi started as a biostatistician in the department of Biostatistics and Data Science of the Julius Center, UMC Utrecht.

Acknowledgment

Finalizing this thesis would not have been possible without the countless help and support from colleagues, friends, and family. With this page, I would like to take the opportunity to express my gratitude.

Prof. le Cessie and Prof. Dekkers, Saskia and Olaf, thank you for giving me an opportunity to pursue a PhD under your guidance. You were patient and understanding, and I could learn both statistical and clinical perspectives in every meeting. It was incredibly fortunate to have you both as my supervisors.

Colleagues from the department of Clinical Epidemiology at the LUMC, thank you for all the chats, lunches, borrels, uitjes, and coffee & cookies hours. I felt comfortable and included in the department.

Mona Shahab and Amir Zamanipoor, my time at the department could not have been as fun without you two. I cherish all our past and future hangouts, dramas, and hugs. You are my biggest findings.

Ruifang Li, I admire your passion for research, but what amazes me even more is your humor and genuineness. Thank you for always being kind and caring.

Ajda Bedene and Vid Prijatelj, thank you for feeding and pampering me beyond count. I am excited to see us growing together in our friendship and professional life.

Linda Nab and Kim Luijken, thank you for all our fun times at the conferences. I am looking forward to our reunion in the near future.

Sebastiaan Bonne, thank you for making hard times lighter with gossip, jokes, sarcasm, and animal videos.

Tahani Alsheri, you always welcomed me whenever I knocked on your office door. Thank you for all our relaxing conversations and an endless supply of chocolate.

Denise Ginkel-Zielinski, thank you for your cupcakes and laughter - and especially for your love and care for those whom the Netherlands is not yet home.

Bunga Pratiwi, how lucky I was to be your classmate eight years ago! Coming back from talking to you always makes me feel warm and understood. Thank you very much.

Limin Liu, it was so natural and easy to be friends with you. You made working from home during the pandemic bearable for me. Thank you so much.

Colleagues at Julius Clinical Data Science team, working with such a smart, fun, kind, and supportive team was a privilege. You raised the bar for future colleagues very high.

Yujin Cho, Sunyoung Hwang, Jinah Kim, and Doyoun Lee, despite being in the Netherlands, you never made me feel distant from you. Thank you for allowing me to be myself around you guys for the last 14 years.

All my small friends, thank you for giving me chances to see life from different perspectives and never judging me on who I am.

Family van Wijk, thank you for making the Netherlands home for me. Knowing that I have family here is extremely special.

My beloved Choi & Woo family, knowing that I forever have a place to go back encourages me to step out to the unknown. Mom and Dad, I never once doubted your trust in me, which has been and will be the biggest gift in my life. Thank you for your unwavering support, curiosity, and enthusiasm in every step I made.

Willem van Wijk, my partner in crime, best friend, and biggest fan, thank you for all the singing and dancing in every moment we shared. I cannot think of the counterfactual of my life without you.