Bias in regression analysis

Schuster, Noah Alexandra

2022

document version
Publisher's PDF, also known as Version of record

BIAS IN REGRESSION ANALYSIS
PROBLEMS AND SOLUTIONS

NOAH A. SCHUSTER

Bias in Regression Analysis

Problems and Solutions

Noah A. Schuster

This thesis was prepared within the Amsterdam Public Health Research Institute and the Department of Epidemiology and Data Science of the Amsterdam University Medical Center, location VU University Medical Center, Amsterdam, The Netherlands.

Printed by Ipskamp Printing

VRIJE UNIVERSITEIT

Bias in Regression Analysis

Problems and Solutions

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad Doctor of Philosophy aan
de Vrije Universiteit Amsterdam,
op gezag van de rector magnificus
prof.dr. J.J.G. Geurts,
in het openbaar te verdedigen
ten overstaan van de promotiecommissie
van de Faculteit der Geneeskunde
op woensdag 9 november 2022 om 13.45 uur
in een bijeenkomst van de universiteit,
De Boelelaan 1105

door
Noah Alexandra Schuster
geboren te Amsterdam

promotor	prof.dr. J.W.R. Twisk
copromotoren:	dr. M.W. Heymans dr. J.J.M. Rijnhart
promotiecommissie:	prof.dr. G.J.M.G. van der Heijden prof.dr. M. Huisman prof.dr. F.J. van Lenthe dr. M.R. de Boer dr. T. Hoekstra

Contents

CHAPTER 1

General introduction

Background

Epidemiologists are generally interested in the effect of an exposure on an outcome, also called the exposure effect. A well-known epidemiological example is the study by Doll and Hill linking smoking to lung cancer (1). Throughout the decades, the use of statistical methods to estimate exposure effects has increased substantially (2). Often, generalized linear models (GLMs) are used to estimate these effects. GLMs were introduced in 1972 by Nelder en Wedderburn and encompass multiple regression techniques, which all trace the outcome as a *linear* function of the exposure (3-5). The distribution of the outcome determines which regression technique is most appropriate to estimate the exposure effect. If the outcome is continuous (for example, blood pressure), then linear regression analysis can be used. For the analysis of a binary outcome (for example, hypercholesterolemia), logistic regression can be used. When analysing survival data (for example, the time till the development of depressive symptoms), then Cox regression analysis can be used. Logistic- and Cox regression require a transformation of the outcome variable to meet the linearity assumption. Because linear-, logistic- and Cox regression are the most common techniques applied in epidemiological research, these are described in more detail below.

Linear regression

A standard linear regression model is given by

$$Y_1 = i_1 + \beta_1 X + \varepsilon_1 \quad (1)$$

where Y_1 represents the continuous outcome of interest and X represents the exposure of any distribution. ε_1 is the error term (i.e., the variance of Y_1 not explained by exposure X), and i_1 and β_1 represent the intercept and exposure effect, respectively (3, 6). The intercept is the mean outcome value if the exposure is zero, and the exposure effect is the average difference in the outcome for every one unit difference in the exposure. Because a linear relation is assumed between the exposure and the outcome, the exposure effect is the same for all one unit differences in the exposure values. If exposure X is a binary variable coded as 0 for the non-exposed individuals and 1 for the exposed individuals, then the intercept is the mean outcome value for the non-exposed individuals, and the exposure effect is the average difference in the outcome between the exposed and the non-exposed individuals (3, 6). Using the coefficients from equation 1, it is possible to estimate the outcome for each individual given their exposure value.

Logistic regression

Whereas linear regression doesn't require any transformation to linearly relate the exposure to the outcome, in logistic regression the outcome is transformed into the natural logarithm of the odds (log odds or logit) of the outcome (3, 5-8). The odds of developing the outcome are given by $\frac{p}{1-p}$, where p represents the probability of developing the outcome of interest (i.e., $\Pr(Y = 1)$). A logistic model is given by

$$logit(\Pr(Y_2 = 1)) = i_2 + \beta_2 X \quad (2)$$

where $logit(\Pr(Y_2 = 1))$ represents the log odds of the outcome. Like in equation 1, X represents the exposure of any distribution and i_2 and β_2 represent the intercept and exposure effect, respectively. Using the coefficients from equation 2, it is possible to estimate the probability of developing the outcome for each individual (3, 8). This is done by

$$p = \frac{1}{1 + e^{-(i_2 + \beta_2 X)}} \quad (3)$$

The interpretation of the intercept and exposure effect in logistic regression are similar to that in linear regression. The intercept is the mean log odds of the outcome if the exposure is zero, and the exposure effect is the average difference in the log odds of the outcome corresponding to every one unit difference in the exposure. Because a linear relationship is assumed, the exposure effect is the same for every one unit difference in the exposure. If the exposure is binary, then the intercept is the mean log odds of the outcome for the non-exposed individuals, and the exposure effect is the average difference in the log odds of the outcome between the exposed and the non-exposed individuals (6, 8).

By taking the exponent of i_2 and β_2, the intercept is the odds of the outcome in the unexposed (i.e., $X = 0$), and the exposure effect is an odds ratio (OR): the ratio of odds of the outcome between exposed and non-exposed individuals. If the OR equals one, then there is no association between the exposure and the outcome. If the OR is larger than one, then the exposure is associated with a higher odds of the outcome, whereas the exposure is associated with a lower odds of the outcome if the OR is smaller than one (7, 9).

Cox regression

Like binary outcomes, survival outcomes have to be transformed to allow for the estimation of a linear relation between the exposure and the outcome. In Cox regression, the outcome is modelled as the natural logarithm of the hazard function, which is the instantaneous probability per unit of time for the event of interest to happen given that the individual has not experienced the event up to that moment (9-11). That is, the hazard is the probability that an individual that is under observation at a specific moment in time experiences an event at that exact time. A Cox regression model is given by

$$log(h(t)) = log(h_0(t)) * \beta_3 X \quad (4)$$

where $\log(h(t))$ represents the expected log hazard at time t, $\log(h_0)$ represents the baseline hazard (i.e., the hazard when the exposure X equals zero) which varies with time, and β_3 represents the exposure effect.

Here, the exposure effect is the average difference in the log hazard of the event of interest at any point in time corresponding to every one unit difference in the exposure. If the exposure is binary, then the exposure effect is the average difference in the log hazard of the event of interest at any point in time between the exposed and the non-exposed individuals. By taking the exponent, the exposure effect is a hazard ratio (HR): the ratio of hazards of the event of interest at any point in time between exposed and non-exposed individuals. If the HR is one, then there is no association between the exposure and the outcome. If the HR is larger than one, then the hazard of the event of interest increases, whereas the hazard decreases if the HR is smaller than one. In the latter case, the exposure has a protective effect on the outcome (10, 11). The HR is assumed to be constant over time (i.e., the proportional hazard assumption), meaning that the estimated effect is the same at every point in time (9).

Bias

In many epidemiological studies, the aim is to isolate the true effect of the exposure on the outcome. However, often the association between an exposure and an outcome is not entirely attributable to the exposure, i.e., the effect is *biased* (9, 12). Bias is defined as 'an error in the conception and design of a study – or in the collection, analysis, interpretation, reporting, publication, or review of data – leading to results or conclusions that are systematically (as opposed to randomly) different from truth' (13). Thus, bias can occur in practically all stages of a study and can be both negative or positive, resulting in

an under- or overestimation of the true effect. It can also reverse the apparent direction of the effect. In this thesis, I focus on bias that can occur in the analysis stage of a study as a result of the incorrect application of linear-, logistic- and Cox regression models. Below is a description of potential sources of bias that appear in the chapters in this thesis.

Non-linear effects

Each model has certain assumptions that must be met for the estimated exposure effect to be an *unbiased* estimate of the true effect. An assumption that GLMs share is that the exposure is linearly related to the outcome (4). If this assumption is not met, i.e., if a linear relation is modelled between the exposure and the outcome when in reality this relation is non-linear, then the effect estimate is not a good representation of the true underlying effect and bias is introduced.

In equations 1, 2 and 3, the effect estimates are obtained through univariable regression (Figure 1A), meaning that the outcome is modelled as a function of the exposure only. However, in observational data, associations are rarely that straightforward. There are multiple ways in which a third variable can distort the effect of the exposure on the outcome, one of which is if that variable is a confounder. A confounder is associated with both the exposure and the outcome, but does not lie in the causal pathway of the exposure on the outcome (Figure 1B). In this case, part of the effect of the exposure on the outcome is explained by the confounder (9). Ignoring the confounder leads to incorrect inference about the association between the exposure and the outcome. Thus, to obtain an unbiased exposure effect, it is necessary to adjust for this confounder. One of the ways in which this can be done is by regressing the outcome on the exposure and adding the confounder as a covariate, i.e., by using multivariable regression analysis (3). A multivariable linear regression model that does so is given by

$$Y_4 = i_4 + \beta^* X + \gamma C + \varepsilon_4 \quad (5)$$

where Y_4, X and C represent the outcome, exposure and confounder, respectively. i_4 represents the intercept, ε_4 is the error term and γ is the coefficient corresponding to confounder C. β^* represents the *confounder-adjusted* exposure effect.

The assumptions underlying a multivariable regression model are identical to the assumptions of a univariable regression model (3). Thus, in equation 5, the linearity

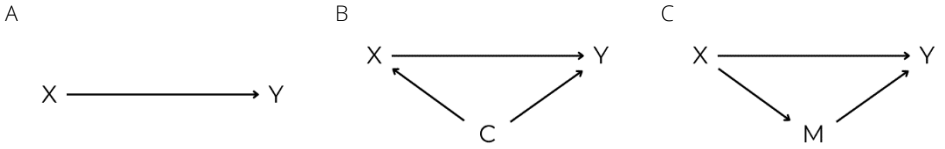

Figure 1 Figures illustrating a univariable association between an exposure and an outcome (panel A), a single-confounder model (panel B) and a single-mediator model (panel C)

assumption not only applies to the association between the exposure and the outcome, but also to the association between the confounder and the outcome. It is common practice to assess the linearity assumption of the association between the exposure and the outcome, and there is a substantial body of literature that covers this topic (6, 14). However, the linearity of the association between the confounder and the outcome is less commonly assessed in practice. This is problematic, because if it is incorrectly assumed that that the confounder is linearly related with the outcome, then, in an attempt to remove bias by adjusting for the confounder, bias may actually be introduced.

Other ways to adjust for confounding include propensity score methods, inverse probability weighting and double robust estimation (6, 15-18). Like with multivariable regression analysis, if the linearity assumptions underlying these methods are not met, bias may be introduced.

If the linearity assumption does not hold, then the non-linear associations present in the data have to be modelled explicitly. There are different ways to do so, of which some simple methods such as categorization of the exposure variable and the use of higher order terms are widely used in epidemiological research, largely due to historical precedent (19). A more advanced and flexible method to model non-linear relations is by the introduction of spline functions in the regression model. Spline functions are transformations of the continuous independent variable (i.e., the variable that the outcome is regressed on) and are available in different forms such as linear spline (LSP) functions and restricted cubic spline (RCS) functions (6). In both LSP and RCS regression, the independent variable is divided into multiple intervals, and for each interval the relationship between the exposure and outcome is estimated separately (6, 20). If for each interval a *linear* relationship is estimated, then the spline functions results in LSP regression. If for each interval a *third degree* relationship is estimated and the tails are restricted, then the spline functions result in RCS regression (6, 21). Both LSP and RCS regression are implemented in most software packages commonly used by

epidemiologists. However, its use lags behind in applied research, possibly because the theory behind these methods is often presented in a complex and mathematical way (22, 23).

Noncollapsibility

To identify confounders to adjust for in the analysis, researchers often use statistical methods to quantify the confounding bias. This is mostly done by comparing exposure effect estimates between a univariable- and a multivariable regression model (e.g., by comparing β_1 from equation 1 with β^* from equation 5). This is also called the change-in-estimate criterion (24-26). Typically, a 10% difference in the exposure effect estimates is used in practice as an arbitrary threshold indicating confounding that needs to be adjusted for (25, 27).

However, in logistic regression, adjusting for a third variable may lead to a change in the exposure effect estimate regardless of whether that variable is actually a confounder. That is because in logistic regression there are two mechanisms in which a third variable may affect the effect estimates: through confounding if that variable is associated with both the exposure and the outcome, and through noncollapsibility if that variable is associated with the outcome (28). This noncollapsibility effect stems from a change in scales that occurs in logistic regression when variables are added to the model (28). As a result of the change in scales, negative exposure effects become more negative, and positive exposure effects become more positive. Thus, in logistic regression, the difference between univariable- and multivariable exposure effect estimates may not only represent confounding bias but also a noncollapsibility effect (29, 30). Relying on the change-in-estimate criterion may then lead to wrong conclusions about the presence and magnitude of confounding bias.

Causal mediation analysis

Another way in which a third variable can affect the association between an exposure and an outcome is if that variable acts as a mediator (Figure 1C). A mediator (partially) explains the effect of the exposure on the outcome, as the exposure causes the mediator, and the mediator in turn causes the outcome (31). In contrast to confounders, mediators do not bias the exposure effect and adjustment for a mediator leads to overadjustment bias. Instead of adjusting for a mediator, mediation analysis can be used to decompose the total effect of the exposure on the outcome into an indirect effect through the mediator and a direct effect after removing the influence of the mediator

(32, 33). Two general methods for mediation analysis have been described in the literature: traditional mediation analysis and causal mediation analysis. In contrast to traditional mediation analysis, causal mediation analysis separates the causal effect definitions from the causal effect estimates. This allows for different estimation methods to be used to estimate the causal indirect-, direct- and total effect (34). These methods include regression, simulation, imputation and weighting.

To estimate the causal effect of an exposure on an outcome, potential outcomes are compared between two exposure values (35, 36). These two exposure values are also called the causal contrast. Only the regression- and the simulation-based approach require the selection of a causal contrast, which is also reflected in the interpretation of the results: the indirect effect from the regression- and simulation-based approaches only apply to the two values selected for the causal contrast, whereas the imputation- and weighting-based approaches return the average difference in the outcome for *every one unit difference* in the exposure through the mediator.

If the mediator and the outcome are both continuous, then all four estimation approaches provide the same causal effect estimates (37). That is, if all pathways in the mediation model are linear, then the effect estimates are the same for every causal contrast. However, if the exposure is continuous and the mediator is binary, then the different estimation approaches no longer provide the same effect estimates. In addition, because of the now non-linear relationship between the exposure and the mediator, for the regression- and simulation-based approaches the estimated mediation effects depend on the selected causal contrast. If researchers are unaware of the different approaches and the role of the causal contrast, then the mediation effect estimates may be interpreted incorrectly.

Competing risks

A potential source of bias in survival analysis are competing events. A competing event is an event that prevents the event of interest from happening (38). For example, if the event of interest is time till the development of depressive symptoms, then death is a competing event. In epidemiological research competing events are often ignored, and individuals that experience a competing event are censored. Censoring occurs when the exact survival time of an individual (i.e., the time till the event of interest) is unknown. This happens, for example, if an individual withdraws from the study or does not experience the event of interest before the end of the study (38, 39). One of the assumptions of Cox

regression is that of independent or noninformative censoring, meaning that individuals who are censored have the same future risk of the event of interest as the individuals that remain under observation (9). However, if someone experiences a competing event, then that individual does no longer have the same future risk for the event of interest. Thus, censoring these individuals goes against the assumption of noninformative censoring, and failing to account for competing risks generally leads to an overestimation of the effect of the exposure on the outcome (39, 40). Therefore, specific competing risk analysis should be applied to analyse survival data in the presence of competing risks.

Aim

Even though regression modelling is widely used in epidemiological research, researchers are often unaware of the many ways in which the incorrect application of these methods can introduce bias. Existing literature on these topics often contain a high level of technical and mathematical details (21-23, 29, 30, 41-43), which hampers the understanding, application and interpretation of correct methods by applied researchers. Therefore, the aim of this thesis is to describe situations in which bias can occur in regression analysis in a non-technical and non-mathematical way, and to propose solutions where possible.

Data

In each chapter, the theory is illustrated using an empirical data example. The data comes from the Longitudinal Aging Study Amsterdam (LASA) and the Amsterdam Growth and Health Longitudinal Study (AGHLS). LASA is an ongoing prospective cohort study among older adults in the Netherlands. The study started in 1992 and focusses on the determinants, trajectories and consequences of physical, cognitive, emotional and social functioning. A new round of measurements is conducted approximately every three years (44, 45). The AGHLS is an ongoing prospective cohort study that started in 1976 with the aim to examine growth and health among Amsterdam teenagers. Later measurement rounds focus on the association between health and lifestyle measures, on the determinants of chronic diseases and on parameters for the investigation of deterioration in health with age (46). In addition to empirical data examples, some chapters also contain a simulation study. In a simulation study, data is created by pseudo-random sampling in which 'true' effects are known. This allows for the evaluation of model performance and for the comparison of methods, for example in terms of bias (47, 48). Bias can then be expressed, among other things, as *absolute* bias (i.e., the absolute

difference between the exposure effect estimate and the true exposure effect) or *relative* bias (the absolute bias relative to the true exposure effect) (47).

Outline

Chapter two of this thesis demonstrates the importance of correctly specifying the confounder-exposure and confounder-outcome associations to estimate unbiased exposure effects if the confounder is a continuous variable. Four different confounder-adjustment methods are reviewed and researchers are provided with an overview of tools to examine and correctly specify the functional form of the associations. One of the tools to correctly specify the functional form is the introduction of spline functions in the regression model. Chapter three compares spline regression to more traditional methods to deal with non-linear exposure effects and explains in detail what spline functions are, how they can be applied and how the results should be interpreted. Chapter four describes how the traditional change-in-estimate criterion to determine the presence of confounding bias can lead to wrong conclusions when applied to logistic regression coefficients. The role of noncollapsibility in logistic regression is clarified and guidance is provided in determining the presence of confounding bias. Chapter five illustrates the difference between the causal estimation methods for mediation models with a continuous exposure and a binary outcome. Four estimation approaches are compared in terms of their performance and interpretation: the regression-, simulation-, imputation- and weighting-based approach. Chapter six introduces competing risk analysis to deal with competing events in survival data and explains how to analyse and interpret survival data in the presence of these competing events. Finally, chapter seven contains a discussion of the results presented in this thesis and provides recommendations for practice.

References

1. Doll R, Hill AB. Smoking and Carcinoma of the Lung. British Medical Journal. 1950;2(4682):739.
2. Levy PS, Stolte K. Statistical methods in public health and epidemiology: a look at the recent past and projections for the next decade. Statistical Methods in Medical Research. 2000;9(1):41-55.
3. Fox J. Applied Regression Analysis and Generalized Linear Models. 3rd ed: SAGE Publications, Inc; 2016.
4. Nelder JA, Wedderburn RWM. Generalized Linear Models. Journal of the Royal Statistical Society. 1972;135(3):370-84.
5. Dobson AJ, Barnett AG. An Introduction to Generalized Linear Models. 4th ed: Chapman & Hall/CRC; 2018.
6. Harrell FE. Regression Modeling Strategies. 2nd ed: Springer; 2015.
7. Bland JM, Altman DG. The odds ratio. BMJ. 2000;320(7247):1468.
8. Kleinbaum DG, Kupper LL, Chambless LE. Logistic regression analysis of epidemiologic data: theory and practice. Communications in Statistics - Theory and Methods. 1982;11(5):485-547.
9. Newman SC. Biostatistical Methods in Epidemiology: John Wiley & Sons, Inc.; 2001.
10. Kleinbaum DG, Klein M. Survival Analysis: A Self-Learning Text. 3rd ed. New York: Springer-Verlag; 2012.
11. Cox DR. Regression Models and Life-Tables. Journal of the Royal Statistical Society: Series B (Methodological). 1972;34(2):187-202.
12. Hernán MA, Hernández-Díaz S, Robins JM. A Structural Approach to Selection Bias. Epidemiology. 2004;15(5):615-25.
13. Porta M. A Dictionary of Epidemiology. 6th ed: Oxford University Press; 2014.
14. Rothman KJ, Greenland S, Lash TL. Modern Epidemiology. 3rd Edition ed: Lippincott Williams & Wilkins; 2008.
15. 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.
16. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. Biometrika. 1983;70(1):41-55.
17. Funk MJ, Westreich D, Wiesen C, Stürmer T, Brookhart MA, Davidian M. Doubly Robust Estimation of Causal Effects. American Journal of Epidemiology. 2011;173(7):761-7.

18. Kang JD, Schafer JL. Demystifying double robustness: a comparison of alternative strategies for estimating a population mean from incomplete data (with discussion). Stat Sci. 2008;22(4):523-80.
19. Bennette C, Vickers A. Against quantiles: categorization of continuous variables in epidemiologic research, and its discontents. BMC Medical Research Methodology. 2012;12(1):21.
20. Greenland S. Dose-Response and Trend Analysis in Epidemiology: Alternatives to Categorical Analysis. Epidemiology. 1995;6(4):356-65.
21. Durrleman S, Simon R. Flexible regression models with cubic splines. (0277-6715 (Print)).
22. de Boor CR. A Practical Guide to Splines: Springer-Verlag New York; 1978.
23. Smith PL. Splines As a Useful and Convenient Statistical Tool. The American Statistician. 1979;33(2):57-62.
24. Greenland S, Pearce N. Statistical foundations for model-based adjustments. (1545-2093 (Electronic)).
25. Kleinbaum DG, Sullivan KM, Barker ND. A Pocket Guide to Epidemiology: Springer Science + Business Media, LLC; 2007.
26. Miettinen OS, Cook EF. Confounding: essence and detection. American Journal of Epidemiology. 1981;114(4):593-603.
27. Lee PH. Is a Cutoff of 10% Appropriate for the Change-in-Estimate Criterion of Confounder Identification? Journal of Epidemiology. 2014;24(2):161-7.
28. Mood C. Logistic Regression: Why We Cannot Do What We Think We Can Do, and What We Can Do About It. European Sociological Review. 2009;26(1):67-82.
29. Janes H, Dominici F, Zeger S. On quantifying the magnitude of confounding. Biostatistics. 2010;11(3):572-82.
30. Pang M, Kaufman JS, Platt RW. Studying noncollapsibility of the odds ratio with marginal structural and logistic regression models. Stat Methods Med Res. 2016;25(5):1925-37.
31. MacKinnon DP, Fairchild AJ, Fritz MS. Mediation Analysis. Annual Review of Psychology. 2006;58(1):593-614.
32. Hernan MA, Robins JM. Causal Inference: What If. Boca Raton: Chapman & Hall/CRC; 2020.
33. MacKinnon DP. Introduction to Statistical Mediation Analysis. New York: Lawrence Erlbaum Associates; 2008.
34. Pearl J. The Causal Mediation Formula—A Guide to the Assessment of Pathways and Mechanisms. Prevention Science. 2012;13(4):426-36.

35. Holland PW. Statistics and Causal Inference. Journal of the American Statistical Association. 1986;81(396):945-60.
36. Rubin DB. Estimating causal effects of treatments in randomized and nonrandomized studies. Journal of Educational Psychology. 1974;66(5):688-701.
37. Valente MJ, Rijnhart JJM, Smyth HL, Muniz FB, MacKinnon DP. Causal Mediation Programs in R, Mplus, SAS, SPSS, and Stata. Struct Equ Modeling. 2020;27(6):975-84.
38. Austin PC, Lee DS, Fine JP. Introduction to the analysis of survival data in the presence of competing risks. Circulation. 2016;133(6):601-9.
39. Putter H, Fiocco M, Geskus RB. Tutorial in biostatistics: competing risks and multi-state models. Stat Med. 2007;26(11):2389-430.
40. Koller MT, Raatz H, Steyerberg EW, Wolbers M. Competing risks and the clinical community: irrelevance or ignorance? Stat Med. 2012;31(11-12):1089-97.
41. Lunn M, McNeil D. Applying Cox Regression to Competing Risks. Biometrics. 1995;51(2):524-32.
42. Imai K, Keele L, Tingley D. A general approach to causal mediation analysis. Psychol Methods. 2010;15(4):309-34.
43. Muthén BO, Muthén LK, Asparouhov T. Regression and Mediation Analysis using Mplus. Los Angeles, CA: Muthén & Muthén; 2017.
44. Hoogendijk EO, Deeg DJ, Poppelaars J, van der Horst M, Broese van Groenou MI, Comijs HC, et al. The Longitudinal Aging Study Amsterdam: cohort update 2016 and major findings. Eur J Epidemiol. 2016;31(9):927-45.
45. Hoogendijk EO, Deeg DJH, de Breij S, Klokgieters SS, Kok AAL, Stringa N, et al. The Longitudinal Aging Study Amsterdam: cohort update 2019 and additional data collections. Eur J Epidemiol. 2019.
46. Wijnstok NJ, Hoekstra T, van Mechelen W, Kemper HC, Twisk JW. Cohort profile: the Amsterdam Growth and Health Longitudinal Study. Int J Epidemiol. 2013;42(2):422-9.
47. Burton A, Altman DG, Royston P, Holder RL. The design of simulation studies in medical statistics. Statistics in Medicine. 2006;25(24):4279-92.
48. Morris TP, White IR, Crowther MJ. Using simulation studies to evaluate statistical methods. Statistics in Medicine. 2019;38(11):2074-102.

CHAPTER 2

Misspecification of confounder-exposure and confounder-outcome associations leads to bias in effect estimates

Noah A. Schuster
Judith J.M. Rijnhart
Lisa C. Bosman
Jos W.R. Twisk
Thomas Klausch
Martijn W. Heymans

Submitted for publication

Abstract

Background

Confounding is a common issue in epidemiological research. Commonly used confounder-adjustment methods include multivariable regression analysis and propensity score methods. Although it is common practice to assess the linearity assumption for the exposure-outcome effect, most researchers do not assess linearity of the relationship between the confounder and the exposure and between the confounder and the outcome before adjusting for the confounder in the analysis. Failing to take the true non-linear functional form of the confounder-exposure and confounder-outcome associations into account may result in an under- or overestimation of the true exposure effect. Therefore, this paper aims to demonstrate the importance of correctly specifying the confounder-exposure and confounder-outcome associations to estimate unbiased exposure effects.

Methods

A Monte Carlo simulation study was used to assess and compare the performance of confounder-adjustment methods when the functional form of the confounder-exposure and confounder-outcome associations were misspecified (i.e., linearity was wrongly assumed) and correctly specified (i.e., linearity was rightly assumed) under multiple sample sizes. An empirical data example was used to illustrate that the misspecification of confounder-exposure and confounder-outcome associations leads to bias.

Results

The simulation study illustrated that the exposure effect estimate will be biased when for propensity score (PS) methods the confounder-exposure association is misspecified. For methods in which the outcome is regressed on the confounder or the PS, the exposure effect estimate will be biased if the confounder-outcome association is misspecified. In the empirical data example, correct specification of the confounder-exposure and confounder-outcome associations resulted in smaller exposure effect estimates.

Conclusion

When attempting to remove bias by adjusting for confounding, misspecification of the confounder-exposure and confounder-outcome associations might actually introduce bias. It is therefore important that researchers not only assess the linearity of the exposure-outcome effect, but also of the confounder-exposure or confounder-outcome associations depending on the confounder-adjustment method used.

Introduction

Unlike in randomized controlled trials, the observed exposure values in observational studies are often influenced by the characteristics of the study subjects. As a result, there might be an unintended difference in baseline characteristics between exposed and unexposed individuals. If these characteristics are also associated with the outcome, then these covariates are confounders of the exposure-outcome effect. In other words, a confounder is a common cause of the exposure and the outcome (1). A simple comparison of the outcome between exposure groups then results in a biased effect estimate (2, 3). Therefore, in observational studies, to obtain an unbiased estimate of the exposure effect it is necessary to remove the spurious part of the exposure-outcome effect caused by the confounders.

There are different methods to obtain confounder-adjusted exposure effect estimates, such as multivariable regression analysis and various propensity score (PS) methods. In multivariable regression analysis the confounders are added to the model in which the outcome is regressed on the exposure (4). This way, the confounder-outcome association is removed. In propensity score methods a balancing score is created which can subsequently be used to adjust, stratify, or weight the exposure-outcome effect (2, 5). By creating this balancing score, the confounder-exposure association is removed and an unbiased exposure effect estimate can be obtained (6).

When multivariable regression analysis is used to adjust for a continuous confounder in order to obtain an unbiased exposure effect estimate, both the exposure-outcome effect and the confounder-outcome association are assumed to be linear. It is common practice to assess the linearity assumption for the exposure-outcome effect and there is a substantial body of literature that covers this topic (4, 7). However, it seems less common practice to also assess the linearity of the confounder-exposure and confounder-outcome associations (8, 9). When it is incorrectly assumed that the confounders are linearly related with the exposure and outcome (i.e., if the associations are misspecified), the exposure effect estimate might be over- or underestimated. Thus, in an attempt to remove bias, bias may actually be introduced. The bias that remains (or is introduced) after adjusting for confounding is also called residual confounding (7, 8, 10).

In this study, we demonstrate the importance of correctly specifying the confounder-exposure and confounder-outcome associations to estimate unbiased exposure effects if the confounder is a continuous variable. First, we describe how to examine the linearity

assumption for any association. Next, we review four well-known confounder-adjustment methods and lay out their respective functional form assumptions. Then, we illustrate the importance of the correct specification using a Monte Carlo simulation and an empirical data example. Finally, we discuss methods that can be used to correctly specify the confounder-exposure and confounder-outcome associations.

Examination of the linearity assumption

There are several ways to assess linearity of the effects. Assume there are two continuous variables A and B and that one is interested in examining the linearity of the A-B effect. The easiest way to assess linearity is by visual inspection: a scatterplot with variable A on the X-axis and variable B on the Y-axis provides an indication of the nature of the relationship between A and B (11). Figure 1 provides a hypothetical example of a linear relationship between variables A and B (panel A), and a non-linear relationship between those variables (panel B). In both panels, the dotted line represents the linear regression line, i.e., the line that describes a linear relationship between variables A and B. In panel A, the regression line fits the data very well. In panel B, however, the linear regression line is not a good representation of the non-linear relationship between A and B. Then, failing to take the non-linear nature of the relationship into account leads to a biased estimate of the A-B effect.

Non-visual ways to assess linearity include adding a non-linear term to the model and categorization of the continuous independent variable. When adding a non-linear term (e.g. a quadratic function) to the model, variable B is modelled as a function of variable A and the non-linear term of that same variable A using linear regression. If the A-B effect

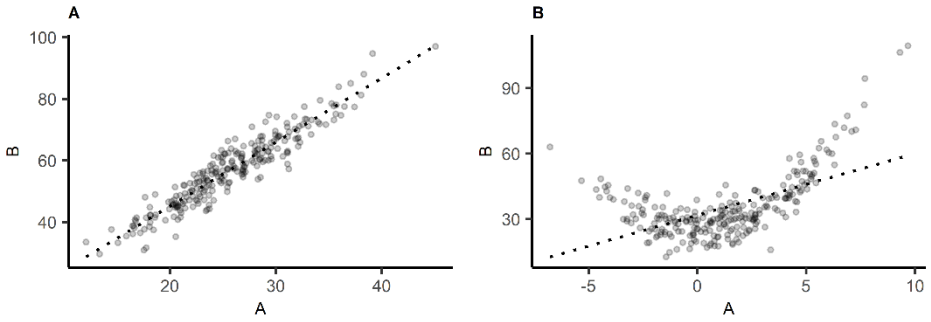

Figure 1 Hypothetical example of the relationship between continuous variables A and B, where each point represents an observation. Panel A: linear relationship. Panel B: non-linear relationship. The dotted line represents the linear regression line for the relationship between variables A and B

is truly linear, then the coefficient corresponding to the non-linear term will be zero (4). Often, if the non-linear term is not significant, the effect is considered linear. When using categorization to assess linearity, variable B is modelled as a function of a categorized variable A. If the regression coefficients corresponding to the categories of variable A do not increase linearly, then this is also an indication of a non-linear A-B effect (7). Both non-visual ways to assess linearity can also be applied when the outcome of interest is not continuous, but is, for example, dichotomous. In that case, logistic regression will be used to model continuous variable A and its non-linear term, or to model dichotomous variable B as a function of the categorized variable A.

Confounder-adjustment methods

Studies are often interested in estimating the average effect of an exposure on an outcome. In terms of potential outcomes, the average effect of the exposure on the outcome is defined as the difference between two expected outcome values under two exposure values, i.e., $E[Y(1) - Y(0)]$. To obtain an unbiased estimate of this exposure effect it is necessary to adjust for any confounding. In this study we discuss four confounder-adjustment methods: multivariable regression analysis, covariate adjustment using the propensity score (PS), inverse probability weighting (IPW) and double robust (DR) estimation. As assessing the linearity assumption for the exposure-outcome effect is common practice, throughout this paper we assume that the exposure-outcome effect is always correctly specified as linear. However, we believe that the information in this paper also applies to models in which the exposure-outcome effect is (correctly specified as) non-linear. Table 1 shows which association (i.e., the confounder-exposure or the confounder-outcome association, or both) has to be correctly specified for each method in order to obtain unbiased exposure effect estimates.

Multivariable regression analysis

With multivariable regression analysis, the outcome is modelled as a function of the exposure and the confounders (4) (equation 1):

$$E(Y|X, C) = i_1 + \beta_1 X + \beta_2 C_1 + \cdots + \beta_{n+1} C_n \quad (1)$$

where Y and X represent the continuous outcome and a dichotomous exposure, respectively, and i_1 represents the intercept term. β_1 is the multivariable confounder adjusted exposure effect estimate and β_2 to β_{n+1} are the coefficients that correspond to the continuous confounding variables C_1 to C_n.

Table 1 Confounder-adjustment methods and the association(s) that need to be correctly specified to obtain an unbiased estimate of the exposure effect

Confounder-adjustment methods	Confounder-exposure association	Confounder-outcome association
Multivariable regression analysis	n/a	√
Covariate adjustment using the PS[§]	√	√*
IPW[§]	√	n/a
DR estimation[§]	Both associations need to be specified but estimators are consistent if either is *correctly* specified	

Abbreviations: PS: propensity score; IPW: inverse probability weighting; DR: double robust; *: PS-outcome effect; n/a: not applicable, §: requires a correctly specified propensity score (i.e., the log odds of the exposure is linear in the confounders)

Multivariate regression analysis adjusts for confounding of the exposure-outcome effect by adding confounders C_1 to C_n to the equation. As a result, β_1 represents the difference in the outcome between the exposed and unexposed groups, holding the confounders at the same value (4, 11). Strictly speaking, β_1 is an estimate of the exposure-outcome effect conditional on confounders, i.e., $E[Y(1) - Y(0)|C = c]$, which is different from the average exposure-outcome effect defined earlier, i.e., $E[Y(1) - Y(0)]$. However, β_1 is an estimate of $E[Y(1) - Y(0)]$ when equation 1 is estimated with linear regression (12).

In equation 1, a linear association is assumed between the exposure and the outcome, and between each confounding variable and the outcome (11). The confounder-exposure association is not modelled, therefore no assumptions are made about the functional form of that association.

Propensity score adjustment

The PS is the predicted probability of endorsing exposure (equation 2):

$$PS = P(X = 1|C_1, \dots, C_n) = \frac{1}{1 + e^{-(i_2 + \lambda_1 C_1 + \cdots + \lambda_n C_n)}} \quad (2)$$

where X represents the dichotomous exposure, i_2 is the model intercept and λ_1 to λ_n are regression coefficients corresponding to confounders C_1 to C_n.

The propensity score is estimated in two steps. First, the exposure is modelled as a function of the confounders C_1 to C_n using a logistic regression model. Second, each individual's predicted probability of endorsing the exposure is estimated, which is the

propensity score (2, 6, 13). The PS can be used in different ways to adjust for confounding. In this paper we discuss three of these methods: covariate adjustment with the PS, inverse probability weighting and double robust estimation. All three methods assume that the propensity score is correctly specified, i.e., that the log odds of the exposure is linear in the confounders. Details on the computation of the PS in general and other PS methods such as matching and stratification can be found elsewhere (2, 6, 13-20).

Covariate adjustment using the propensity score

Because the PS contains information on the confounders, it is possible to adjust for confounding by modelling the outcome as a function of the exposure and the PS (2, 13). Thus, instead of conditioning on confounding variables C_1 to C_n as in equation 1, we now condition on the PS (equation 3):

$$E(Y|X, PS) = i_3 + \beta_1^* X + \beta_2^* PS \quad (3)$$

where Y and X represent the continuous outcome and the dichotomous exposure, respectively, and i_3 represents the intercept term. β_1^* is the PS confounder-adjusted exposure effect estimate and β_2^* is the coefficient that corresponds to the propensity score PS.

Because in equation 3 the outcome is regressed on the exposure and the propensity score, linearity assumptions apply both to the exposure-outcome effect and the PS-outcome association. Whereas all PS methods require the PS to be adequately specified, this is the only PS method that additionally makes assumptions about the linearity of the PS-outcome association (2, 4).

Inverse probability weighting

Inverse probability weighting uses weights based on the PS to create a pseudo-population in which each confounder combination is balanced between the exposed and unexposed groups. When there is perfect confounder balance between the groups there is no longer an association between confounders C_1 to C_n and the exposure (4). With weighting, individuals who are underrepresented get larger weights assigned, whereas individuals who are overrepresented get smaller weights assigned.

For exposed individuals the weight is calculated as $\frac{1}{PS}$, whereas for unexposed individuals the weight is calculated as $\frac{1}{1-PS}$ (2, 21). A potential issue with IPW is that the weights can

be unstable. This is because individuals with a PS close to 0 receive very large weights, whereas individuals with a PS close to 1 receive very small weights. Subjects with these large weights will then dominate the weighted analysis, resulting in a large variance of the IPW estimator (22). As an alternative, stabilized weights have been proposed (2). This reduces the weights of the treated individuals with a small PS and the untreated individuals with a large PS. For exposed individuals, these stabilized weights are calculated as $\frac{p}{PS}$ and for unexposed individuals stabilized weights are calculated as $\frac{1-p}{1-PS}$, with p being the probability of exposure without considering the confounders (2). After calculating the weights for all individuals the IPW confounder-adjusted exposure effect is estimated by performing a weighted regression analysis with the exposure as the only independent variable.

IPW does not make any linearity assumptions about the confounder-outcome or PS-outcome association (20). Thus, IPW only assumes a correctly specified propensity model. If the propensity model is misspecified this results in inappropriate weights and possibly a biased IPW confounder-adjusted exposure effect estimate (23).

Double robust estimation

Double robust estimation combines multivariable regression analysis and IPW and is done in two steps: first, a propensity model is specified and stabilized weights are calculated. Second, a weighted analysis in which the outcome is regressed on the exposure and the confounders is performed.

Because the model is weighted by the stabilized weights, an adequately specified propensity model is needed. In addition, because the confounders are included in the regression analysis, linearity assumptions about the confounder-outcome association are made. However, only one of these two associations (i.e., either the confounder-exposure associations in the propensity model or the confounder-outcome associations in the multivariable regression model) has to be correctly specified to obtain an unbiased exposure effect estimate (20, 23, 24). However, if both effects are misspecified, the DR exposure effect estimate may be even more biased than the estimate of a less robust single confounder-adjustment method such as multivariable regression or IPW (25, 26).

Simulation study

Simulation methods

A simulation study was designed to assess and compare the performance of the four confounder-adjustment methods. Four different scenarios were considered based on

the (mis)specification of the confounder-exposure and confounder-outcome association (see Table 2). The R programming language version 4.0.3 was used to generate and analyse the data (27).

To model both misspecified and correctly specified confounder-exposure and confounder-outcome associations, first two continuous confounders were generated. Confounder Z was generated from a standard normal distribution, and confounder C was its corresponding squared terms. The dichotomous exposure was generated from a binomial distribution conditional on confounder Z and its squared term C (equation 4), and the continuous outcome was a function of the exposure and confounders Z and C (equation 5).

$$P(X = 1|Z, C) = \frac{1}{1 + e^{-(i_4 + \beta_1 Z + \beta_1 C)}} \quad (4)$$

$$E(Y|X, Z, C) = i_5 + \beta_1 X + \beta_2 Z + \beta_2 C \quad (5)$$

This way, the exposure and the outcome had a quadratic relation with each of the confounders. Next, we estimated the confounder-adjusted exposure-outcome effect using the four confounder-adjustment methods. In the scenarios in which the non-linearity of the confounder-exposure and confounder-outcome association were correctly specified, the analysis was adjusted for confounders Z and C. This way, the underlying quadratic relation was modelled. In the scenarios in which the effects were misspecified, only confounder Z was included in the analysis. This way, only the incorrect linear relation was modelled. Sample sizes were 200, 500 and 1000. The parameter value for the exposure- outcome effect was set to 0.59 to mimic a large effect size. The parameter values for the confounder-exposure and confounder-outcome association were set to -0.14, -0.39, -0.59 and 0.14, 0.39 and 0.59 to mimic negative and positive

Table 2 Overview of simulated scenarios

Scenario	Confounder-exposure association	Confounder-outcome association
Scenario 1	Correctly specified	Correctly specified
Scenario 2	Correctly specified	Misspecified
Scenario 3	Misspecified	Correctly specified
Scenario 4	Misspecified	Misspecified

When effects are correctly specified, confounders Z and C are adjusted for in the analysis. When effects are misspecified, only confounder Z is adjusted for.

small, medium and large effect sizes, respectively (28). In total, 72 conditions were simulated (4 scenarios; 3 sample sizes; 6 confounder-exposure and confounder-outcome effect sizes) with 1,000 repetitions per condition, resulting in 72,000 observations.

The performance of the confounder-adjustment methods was compared based on the absolute bias (AB) and the relative bias (RB) (29). AB is the absolute difference between the estimated exposure effect and the true exposure-outcome effect of 0.59. RB is the ratio of AB to the true exposure-outcome effect (29, 30). For both performance measures a lower score corresponds to a better performance. The simulation code is available in additional file A.

In additional file E we show an extra condition in which the direction of the exposure effect changes if the non-linearity of the confounder-exposure and confounder-outcome associations is not modelled correctly.

Simulation results

Table 3 shows the mean estimated exposure effect, AB and RB for all models across the four simulated scenarios based on a sample size of 500 and positive confounder-exposure and confounder-outcome associations. Results for sample sizes 200 and 1000 can be found in additional files B and C, respectively.

In scenario 1, where both the confounder-exposure and confounder-outcome associations were correctly specified, multivariable regression analysis, PS adjustment and DR estimation all performed well. When the confounder-outcome association was misspecified (scenario 2), multivariable regression analysis and DR estimation resulted in biased exposure effect estimates. PS adjustment still performed well, but had the PS-outcome association been misspecified as well, then residual bias would also have been observed for that method. In both scenarios 1 and 2, bias was observed for IPW as IPW is a large sample technique (3). Increasing the sample size resulted in exposure effect estimates closer to the true effect. In scenario 3, where the confounder-exposure association was misspecified but the confounder-outcome association was correctly specified, multivariable regression analysis and DR estimation performed well, whereas PS adjustment and IPW resulted in biased exposure effect estimates. When both associations were misspecified (scenario 4), all methods resulted in biased exposure effect estimates. In all scenarios, the amount of bias depended on the strength of the

Table 3 Model performance across all simulated scenarios, n = 500

	Parameter values for the confounder-exposure and confounder-outcome associations								
	0.14			0.39			0.59		
	$\hat{\beta}$	AB	RB	$\hat{\beta}$	AB	RB	$\hat{\beta}$	AB	RB
Scenario 1: correct specification of cx-association & correct specification of cy-association									
Multivariable regression analysis	0.5900	0.0000	0.0000	0.5900	0.0000	0.0000	0.5900	0.0000	0.0000
Covariate adjustment using the PS	0.5901	0.0001	0.0001	0.5907	0.0007	0.0012	0.5909	0.0009	0.0015
Stabilized IPW	0.5903	0.0003	0.0005	0.6030	0.0130	0.0220	0.6417	0.0517	0.0876
DR estimation	0.5900	0.0000	0.0000	0.5900	0.0000	0.0000	0.5900	0.0000	0.0000
Scenario 2: correct specification of cx-association & misspecification of cy-association									
Multivariable regression analysis	0.6263	0.0363	0.0615	0.8051	0.2151	0.3646	0.9859	0.3959	0.6711
Covariate adjustment using the PS	0.5901	0.0001	0.0001	0.5907	0.0007	0.0012	0.5909	0.0009	0.0015
Stabilized IPW	0.5903	0.0003	0.0005	0.6030	0.0130	0.0220	0.6417	0.0517	0.0876
DR estimation	0.5905	0.0005	0.0008	0.6064	0.0164	0.0278	0.6465	0.0565	0.0957
Scenario 3: misspecification of cx-association & correct specification of cy-association									
Multivariable regression analysis	0.5900	0.0000	0.0000	0.5900	0.0000	0.0000	0.5900	0.0000	0.0000
Covariate adjustment using the PS	0.6267	0.0367	0.0622	0.8147	0.2247	0.3808	1.0155	0.4255	0.7212
Stabilized IPW	0.6276	0.0376	0.0638	0.8339	0.2439	0.4134	1.0676	0.4776	0.8096
DR estimation	0.5900	0.0000	0.0000	0.5900	0.0000	0.0000	0.5900	0.0000	0.0000
Scenario 4: misspecification of cx-association & misspecification of cy-association									
Multivariable regression analysis	0.6263	0.0363	0.0615	0.8051	0.2151	0.3646	0.9859	0.3959	0.6711
Covariate adjustment using the PS	0.6267	0.0367	0.062	0.8147	0.2247	0.3808	1.0155	0.4255	0.7212
Stabilized IPW	0.6276	0.0376	0.0638	0.8339	0.2439	0.4134	1.0676	0.4776	0.8096
DR estimation	0.6273	0.0373	0.0533	0.8261	0.2361	0.4002	1.0456	0.4556	0.7723

Abbreviations: n: sample size; cx-association: confounder-exposure association; cy-association: confounder-outcome association; $\hat{\beta}$: mean estimated exposure effect; AB: absolute bias; RB: relative bias; PS: propensity score; IPW: inverse probability weighting; DR: double robust

confounder-exposure and confounder-outcome associations: the weaker the associations were, the less biased was observed. The same patterns can be observed for negative confounder-exposure and confounder-outcome associations. For detailed results see additional file D

Empirical data example

To demonstrate the consequences of misspecification of the confounder-exposure and confounder-outcome association we used an illustrative example from the Amsterdam Growth and Health Longitudinal Study (AGHLS). The AGHLS is an ongoing cohort study that started in 1976 to examine growth and health among teenagers. In later measurement rounds, health and lifestyle measures, determinants of chronic diseases and parameters for the investigation of deterioration in health with age were measured (31). For this demonstration we use data collected in 2000, when the participants were in their late 30s.

Using data from the AGHLS, we estimated the effect of overweight (BMI ≥ 25) on systolic blood pressure. We adjusted this effect for confounding by alcohol consumption (measured in number of glasses per week) and cardiorespiratory fitness (VO2max). Only subjects with complete data on all variables were included in the analyses (n = 359). Note that this data example is included for illustrative purposes only and therefore represents a simplified scenario. In reality, it is likely that there will be additional confounders and time-varying confounders. As a result, substantive interpretations should be approached with caution.

First, we examined the linearity of the confounder-exposure and the confounder-outcome associations. We did this by categorizing alcohol consumption and cardiorespiratory fitness, and separately regressing overweight and systolic blood pressure on the categorized confounders. In both cases, the regression coefficients corresponding to the categories of alcohol consumption and respiratory fitness did not increase linearly. Thus, both confounder-exposure and confounder-outcome associations were non-linear. Second, to demonstrate the consequences of misspecification, we modelled systolic blood pressure as a function of overweight, adjusting for alcohol consumption and cardiorespiratory fitness. We did this first by (falsely) assuming a linear relation between the confounders and overweight and between the confounders and systolic blood pressure. Next, we took these non-linear associations into account by adjusting for alcohol consumption and cardiorespiratory

Table 4 The effect of overweight on systolic blood pressure, adjusted for alcohol consumption. 2nd column: linear confounder-exposure and confounder-outcome associations are assumed. 3rd column: non-linear confounder-exposure and confounder-outcome associations are modelled

	Linearity assumed β (95% CI)	Linearity not assumed β (95% CI)
Multivariable regression analysis	3.589 (0.686; 6.493)	3.022 (0.136; 5.908)
Covariate adjustment using the PS	3.739 (0.822; 6.656)	3.062 (0.164; 5.960)
Stabilized IPW	4.121 (1.110; 7.132)	3.813 (0.807; 6.819)
DR estimation	3.983 (1.262; 6.704)	3.585 (0.879; 6.291)

Abbreviations: PS: propensity score; IPW: inverse probability weighting; DR: double robust; β: regression coefficient; CI: confidence interval

fitness using 3-knot restricted cubic spline (RCS) regression, which has the ability to fit non-linear shapes. A detailed explanation of RCS regression can be found elsewhere (4).

Although implementing RCS regression might still not equal perfect specification of both effects, it provides a better representation of the true non-linear relations than simply assuming linear confounder-exposure and confounder-outcome associations. The results of these analyses can be found in Table 4.

With all four methods, the estimated exposure effects were greater when linearity was assumed than when non-linear confounder-exposure and confounder-outcome associations were modelled. The difference in estimated exposure effects between the two scenarios was largest for covariate adjustment using the PS and smallest for IPW.

Discussion

This paper aimed to emphasize the importance of checking and modelling the (non-)linearity of the confounder-exposure and confounder-outcome association when adjusting for a continuous confounder. Many epidemiologists are unaware that the functional form assumptions (e.g., the linearity assumption in regression analysis) also apply to the confounder-exposure and confounder-outcome associations. If these associations are incorrectly specified as linear, then bias might be introduced in an attempt to remove bias. Our simulation study showed that bias is introduced if the confounder-exposure and/or the confounder-outcome association are misspecified. The amount of bias also depended on the confounder-adjustment method and the strength of the confounder-exposure and confounder-outcome association. This was also illustrated in our empirical data example, in which we modelled the effect of overweight

on systolic blood pressure, adjusted for alcohol consumption and cardiorespiratory fitness. Taking the non-linearity of the confounder-exposure and confounder-outcome associations into account by using restricted cubic spline regression to model the effects resulted in smaller exposure effect estimates for all methods. The simulation study and the empirical data example both showed that merely adjusting for confounding is not enough, but that correct specification of *all* effects in the model is crucial to obtain unbiased exposure effect estimates.

Correct specification of effects

There are several methods that can be used to model the non-linear shape of an effect, such as categorization, the use of higher order terms and the use of spline functions. An overview of the methods, their application and advantages and disadvantages can be found in Table 5.

A great disadvantage of modelling a non-linear effect by categorization is that it assumes homogeneity of effects within groups (32-35). More concretely, this means that we assume that all individuals in a category have the same confounder-exposure or confounder-outcome association. Thus, a potential non-linear association *within* a category is not captured in the analysis. An example of the use of higher order terms can be found in the data generation process of our simulation study, where we generated the outcome as a function of the exposure, linear confounder Z and its quadratic term C. Adding higher order terms increases the flexibility of the model, but also reduces the interpretability of the results (36). However, using higher order terms to approximate the non-linearity of the confounder-exposure or confounder-outcome association does not affect the interpretability of the exposure effect that's our main interest. With spline regression, the confounding variable is also categorized, but a higher power function is fitted for each category separately making spline regression more flexible (4, 11, 37). The boundaries of the categories are called *knots*. For the 3-knot restricted cubic spline function in the empirical data example, the confounding variable alcohol consumption was first categorized into 4 categories, then cubic functions were fitted in each category and restricted in the tails. Next, a single spline variable was added to our model so that the outcome was regressed on the exposure and this spline function. Like with higher order terms, the interpretation of the coefficients can be complicated when spline functions are used (11). However, because we are not necessarily interpreting the coefficients of the confounder-exposure or confounder-outcome associations, spline

Table 5 Methods to approximate true non-linear effects

Method	Explanation	Advantages	Disadvantages
Categorization	The confounder is grouped (e.g. on pre-specified percentile values such as quartiles) and subsequently the outcome is regressed on the exposure and the now categorical confounding variable	Easy to apply	Homogeneity of the effects is assumed within groups, resulting in severe loss of information and possibly residual confounding
Higher order terms	The outcome is regressed on the confounder and the non-linear term of that same confounder, e.g., a quadratic term	Easy to apply Adding higher order terms increases the flexibility of the model	Coefficients are difficult to interpret*
Linear spline regression	First, the confounding variable is categorized and subsequently a first power function is fitted for each category separately. After fitting the spline functions these are added to the regression model	Good approximation of the true effect Coefficients are easy to interpret	
Restricted cubic spline regression	Same as linear spline regression, but instead a more flexible third power function is fitted for each category separately. To avoid instability in the tails where there's not much data, *restricted* cubic splines are often used where at the tails a line is fitted rather than a curve.	Good approximation of the true effect Adding higher order terms increases the flexibility of the model	Coefficients are difficult to interpret*

* This is not a hindrance when these methods are used to model non-linear confounder-exposure or confounder-outcome associations as the corresponding coefficients will not be interpreted

functions are a good and efficient way to approximate the non-linear shapes of those effects.

Reporting of confounding

The results in this paper demonstrate that misspecification of the confounder-exposure and confounder-outcome associations may lead to additional bias. However, in practice residual confounding may often go unnoticed, as inappropriate reporting makes it difficult to assess the reliability and validity of study results. In 2007 the STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) initiative published a checklist of items that should be addressed in reports of observational studies, including two items that address confounding (9 'Bias' and 12 'Statistical methods') (38). The explanatory and elaboration document of STROBE acknowledges that adjusting for confounding may involve additional assumptions about the functional form of the studied associations (39). Despite the publication of the STROBE checklist, the overall quality of reporting of confounding remains suboptimal (40, 41). To increase transparency on the risk of residual confounding, we advise researchers to report how the functional form of the confounder-exposure and confounder-outcome association was assessed and taken into account.

Limitations

The simulation study in this paper is a simplified representation of real world scenarios. We adjusted for one confounder, whereas in reality there might be multiple confounders. If there are multiple confounders, then the confounder-exposure and confounder-outcome association of each of the confounders needs to be assessed and non-linear effects need to be modelled for confounders that are not linearly related to either the exposure or the outcome. In the PS methods, the PS-outcome association was linear, so no additional bias was observed in scenarios in which the confounder-outcome association was misspecified. However, if the PS-outcome association is also misspecified, residual bias would be observed. Therefore, the linearity of the relation between the PS and the outcome should always be checked. IPW is known to perform less well in small samples, which was also confirmed in our simulation (3). Last, in this paper we assume associations are either misspecified or correctly specified, whereas in reality, naturally, everything exists in shades of grey. In addition, there are other important contributors to residual confounding that researchers should be aware of, such as unobserved and mismeasured confounders. These contributors are described in detail elsewhere (42, 43).

Conclusion

To summarize, in this study we showed the importance of correctly specifying the confounder-exposure and confounder-outcome associations to obtain unbiased exposure effect estimates. When these effects are misspecified, bias might actually be introduced in an attempt to remove bias. Thus, to estimate unbiased effects it is important to examine the linearity of the confounder-exposure or confounder-outcome association depending on the confounder-adjustment method used and to adjust the model accordingly.

References

1. Pearl J. Causality. 2nd ed: Cambridge University Press; 2009.
2. 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.
3. Hernan MA, Robins JM. Causal Inference: What If. Boca Raton: Chapman & Hall/CRC; 2020.
4. Harrell FE. Regression modeling strategies: with applications to linear models, logistic and ordinal regression, and survival analysis. 2 ed: Springer International Publishing AG Switzerland; 2003.
5. Guo S, Fraser MW. Propensity Score Analysis: Statistical Methods and Applications. United States of America: SAGE Publications, Inc.; 2014.
6. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. Biometrika. 1983;70(1):41-55.
7. Lash TL, VanderWeele TJ, Haneuse S, Rothman KJ. Modern Epidemiology. 4 ed: Wolters Kluwer; 2020.
8. Groenwold RHH, Klungel OH, Altman DG, van der Graaf Y, Hoes AW, Moons KGM, et al. Adjustment for continuous confounders: an example of how to prevent residual confounding. CMAJ. 2013;185(5):401-6.
9. Groenwold RHH, Van Deursen AMM, Hoes AW, Hak E. Poor Quality of Reporting Confounding Bias in Observational Intervention Studies: A Systematic Review. Annals of Epidemiology. 2008;18(10):746-51.
10. Becher H. The concept of residual confounding in regression models and some applications. Statistics in Medicine. 1992;11(13):1747-58.
11. Cohen J, Cohen P, West SG, Aiken LS. Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences. 3 ed: Routledge; 2002.
12. Greenland S, Pearl J, Robins JM. Confounding and collapsibility in causal inference. Statistical science. 1999;14(1):29-46.
13. D'Agostino Jr RB. Propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. Statistics in Medicine. 1998;17(19):2265-81.
14. Stuart EA, Lee BK, Leacy FP. Prognostic score–based balance measures can be a useful diagnostic for propensity score methods in comparative effectiveness research. Journal of Clinical Epidemiology. 2013;66(8, Supplement):S84-S90.e1.
15. Normand S-LT, Landrum MB, Guadagnoli E, Ayanian JZ, Ryan TJ, Cleary PD, et al. Validating recommendations for coronary angiography following acute

myocardial infarction in the elderly: A matched analysis using propensity scores. Journal of Clinical Epidemiology. 2001;54(4):387-98.

16. Austin PC. The Relative Ability of Different Propensity Score Methods to Balance Measured Covariates Between Treated and Untreated Subjects in Observational Studies. Medical Decision Making. 2009;29(6):661-77.
17. Ho D, Imai K, King G, Stuart E. Matching as Nonparametric Preprocessing for Reducing Model Dependence in Parametric Causal Inference. Political Analysis. 2007;15:199–236.
18. Imai K, King G, Stuart E. Misunderstandings Among Experimentalists and Observationalists about Causal Inference. Journal of the Royal Statistical Society, Series A. 2008;171, part 2:481–502.
19. Morgan SL, Todd JJ. A Diagnostic Routine for the Detection of Consequential Heterogeneity of Causal Effects. Sociological Methodology. 2008;38(1):231-82.
20. Schafer JL, Kang J. Average causal effects from nonrandomized studies: A practical guide and simulated example. Psychological Methods. 2008;13(4):279-313.
21. Rosenbaum PR. Model-Based Direct Adjustment. Journal of the American Statistical Association. 1987;82(398):387-94.
22. Robins JM, Hernan MA, Brumback B. Marginal Structural Models and Causal Inference in Epidemiology. Epidemiology. 2000;11(5):550-61.
23. Funk MJ, Westreich D, Wiesen C, Stürmer T, Brookhart MA, Davidian M. Doubly Robust Estimation of Causal Effects. American Journal of Epidemiology. 2011;173(7):761-7.
24. Bang H, Robins JM. Doubly robust estimation in missing data and causal inference models. Biometrics. 2005;61(4):962-73.
25. Kang JD, Schafer JL. Demystifying double robustness: a comparison of alternative strategies for estimating a population mean from incomplete data (with discussion). Stat Sci. 2008;22(4):523-80.
26. Robins JM, Rotnitzky A, Zhao LP. Estimation of Regression Coefficients When Some Regressors are not Always Observed. Journal of the American Statistical Association. 1994;89(427):846-66.
27. R Core Team. R: A language and environment for statistical computing. Vienna: R Foundation for Statistical Computing; 2020.
28. Cohen J. Statistical power analysis for the behavioral sciences. Hillsdale, NJ: Lawrence Erlbaum Associates; 1988.

29. Burton A, Altman DG, Royston P, Holder RL. The design of simulation studies in medical statistics. Statistics in Medicine. 2006;25(24):4279-92.
30. Morris TP, White IR, Crowther MJ. Using simulation studies to evaluate statistical methods. Statistics in Medicine. 2019;38(11):2074-102.
31. Wijnstok NJ, Hoekstra T, van Mechelen W, Kemper HCG, Twisk JWR. Cohort Profile: The Amsterdam Growth and Health Longitudinal Study. International Journal of Epidemiology. 2013;42(2):422-9.
32. Greenland S. Avoiding Power Loss Associated with Categorization and Ordinal Scores in Dose-Response and Trend Analysis. Epidemiology. 1995;6(4):450-4.
33. Greenland S. Dose-response and trend analysis in epidemiology: alternatives to categorical analysis. Epidemiology. 1995;6(4):356-65.
34. Bennette C, Vickers A. Against quantiles: categorization of continuous variables in epidemiologic research, and its discontents. BMC Medical Research Methodology. 2012;12(21).
35. Royston P, Altman DG, Sauerbrei W. Dichotomizing continuous predictors in multiple regression: a bad idea. Stat Med. 2006;25(1):127-41.
36. Ruppert D, Wand MP, Carroll RJ. Semiparametric Regression. United States of America: Cambridge University Press; 2003.
37. Durrleman S, Simon R. Flexible regression models with cubic splines. Statistics in Medicine. 1989;8(5):551-61.
38. von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP, et al. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. Bull World Health Organ. 2007;85(11):867-72.
39. Vandenbroucke JP, von Elm E, Altman DG, Gøtzsche PC, Mulrow CD, Pocock SJ, et al. Strengthening the Reporting of Observational Studies in Epidemiology (STROBE): Explanation and Elaboration. Epidemiology. 2007;18(6).
40. Pouwels KB, Widyakusuma NN, Groenwold RHH, Hak E. Quality of reporting of confounding remained suboptimal after the STROBE guideline. Journal of Clinical Epidemiology. 2016;69:217-24.
41. Hemkens LG, Ewald H, Naudet F, Ladanie A, Shaw JG, Sajeev G, et al. Interpretation of epidemiologic studies very often lacked adequate consideration of confounding. Journal of Clinical Epidemiology. 2018;93:94-102.
42. Fewell Z, Davey Smith G, Sterne JAC. The Impact of Residual and Unmeasured Confounding in Epidemiologic Studies: A Simulation Study. American Journal of Epidemiology. 2007;166(6):646-55.

43. Vanderweele TJ, Arah OA. Bias formulas for sensitivity analysis of unmeasured confounding for general outcomes, treatments, and confounders. Epidemiology. 2011;22(1):42-52.

SUPPLEMENTARY MATERIALS FOR CHAPTER 2

Additional file A Simulation code

Step 1 – Generate data

```
generate_data <- function(seed, reps, n, ix, iy, cx, xy, cy){

  # define total number of rows required to store data
  rows <- reps * n

  # create data frame to store data in
  df <- as.data.frame(matrix(NA, nrow = rows, ncol = 13))
  colnames(df) <- c("ID",           # ID through entire data set
                    "repnr",        # for each repetition
                    "ID_repnr",     # ID through each repetition
                    "n",            # number of observations
                    "ix",           # intercept exposure
                    "iy",           # intercept outcome
                    "cx",           # confounder-exposure effect
                    "xy",           # exposure-outcome effect
                    "cy",           # confounder-outcome effect
                    "C",            # continuous confounder (correctly
                                      specified)
                    "Z",            # continuous confounder
                                      (misspecified)
                    "X",            # dichotomous exposure
                    "Y")            # continuous outcome

  # define simulation parameters
  df[, "ID"] <- seq(1:rows)
  df[, "repnr"] <- rep(1:reps, each = n)
  df[, "ID_repnr"] <- rep(seq(1, n), reps)
  df[, "n"] <- n

  # define intercepts ix and iy
  df[, "ix"] <- ix
  df[, "iy"] <- iy

  # define coefficients cx, xy and cy
  df[, "cx"] <- cx
  df[, "xy"] <- xy
  df[, "cy"] <- cy

  # generate confounder Z from a standard normal distribution
  # with mean = 0 and sd = 1 (misspecified)
  df[, "Z"] <- rnorm(n = rows)

  # generate confounder C (correctly specified)
  df[, "C"] <- df[, "Z"]^2
```

```
# generate dichotomous exposure X
  lpx <- ix + cx * df[, "Z"] + cx * df[, "C"]
  prx <- 1/(1 + exp(-lpx))
  df[, "X"] <- rbinom(n = rows, size = 1, prob = prx)

  # generate continuous outcome Y
  df[, "Y"] <- iy + xy * df[, "X"] + cy * df[, "Z"] + cy * df[, "C"] +
  rnorm(n = 1)

  # return data frame
  return(df)

  rm(lpx, prx)

}

# define simulation parameters
seed <- 20220718
reps <- 1000
n <- c(200, 500, 1000)
ix <- 0
iy <- 0
confounder_effects <- c(-0.59, -0.39, -0.14, 0.14, 0.39, 0.59)
xy <- 0.59

# generate data sets for all parameters defined above and save each set
# in folder '220119 Step 1 - Generated datasets'

for(j in n){
  for(k in confounder_effects){

    df <- generate_data(seed = seed,
                        reps = reps,
                        n = j,
                        ix = ix,
                        iy = iy,
                        cx = k,
                        xy = xy,
                        cy = k)

    # save each file in folder 'Step 1 - Generated datasets'
    save(df, file = paste0("Step 1 - Generated datasets\\",
                           "n = ", j, ", cx = ", k, ", cy = ", k,
                           ".RData"))
  }
}
```

Step 2 – Generate Models

```
# in scenario 1, both the confounder-exposure and the confounder-
# outcome effect are correctly specified
scenario1 <- function(df){

  estimates <- as.data.frame(matrix(NA, nrow = 1, ncol = 4))

  # 1. multivariable regression analysis
  model_multivar <- glm(Y ~ X + Z + C, data = df)
  estimates[1, 1] <- model_multivar$coefficients[2]

  # 2. covariate adjustment using the ps
  ps <- predict(glm(X ~ Z + C, family = "binomial", data = df), type =
        "response")
  model_covadj <- glm(Y ~ X + ps, data = df)
  estimates[1, 2] <- model_covadj$coefficients[2]

  # stabilized IPW
  ipw <- ifelse(df$X == 1, 1/ps, 1/(1-ps))
  sipw <- ipw/sum(ipw)
  model_sipw <- glm(Y ~ X, weights = sipw, data = df)
  estimates[1, 3] <- model_sipw$coefficients[2]

  # DR estimation
  model_dr <- glm(Y ~ X + Z + C, weights = sipw, data = df)
  estimates[1, 4] <- model_dr$coefficients[2]

  return(estimates)

}

# in scenario 2, the confounder-exposure effect is correctly specified
# and the confounder-outcome effect is misspecified
scenario2 <- function(df){

  estimates <- as.data.frame(matrix(NA, nrow = 1, ncol = 4))

  # 1. multivariable regression analysis
  model_multivar <- glm(Y ~ X + Z, data = df)
  estimates[1, 1] <- model_multivar$coefficients[2]

  # 2. covariate adjustment using the ps
  ps <- predict(glm(X ~ Z + C, family = "binomial", data = df), type =
        "response")
  model_covadj <- glm(Y ~ X + ps, data = df)
  estimates[1, 2] <- model_covadj$coefficients[2]

  # stabilized IPW
  ipw <- ifelse(df$X == 1, 1/ps, 1/(1-ps))
  sipw <- ipw/sum(ipw)
```

```
  model_sipw <- glm(Y ~ X, weights = sipw, data = df)
  estimates[1, 3] <- model_sipw$coefficients[2]

  # DR estimation
  model_dr <- glm(Y ~ X + Z, weights = sipw, data = df)
  estimates[1, 4] <- model_dr$coefficients[2]

  return(estimates)

}

# in scenario 3, the confounder-exposure effect is misspecified and the
# confounder-outcome effect is correctly specified
scenario3 <- function(df){

  estimates <- as.data.frame(matrix(NA, nrow = 1, ncol = 4))

  # 1. multivariable regression analysis
  model_multivar <- glm(Y ~ X + Z + C, data = df)
  estimates[1, 1] <- model_multivar$coefficients[2]

  # 2. covariate adjustment using the ps
  ps <- predict(glm(X ~ Z, family = "binomial", data = df), type =
        "response")
  model_covadj <- glm(Y ~ X + ps, data = df)
  estimates[1, 2] <- model_covadj$coefficients[2]

  # stabilized IPW
  ipw <- ifelse(df$X == 1, 1/ps, 1/(1-ps))
  sipw <- ipw/sum(ipw)
  model_sipw <- glm(Y ~ X, weights = sipw, data = df)
  estimates[1, 3] <- model_sipw$coefficients[2]

  # DR estimation
  model_dr <- glm(Y ~ X + Z + C, weights = sipw, data = df)
  estimates[1, 4] <- model_dr$coefficients[2]

  return(estimates)

}

# in scenario 4, both the confounder-exposure and the confounder-
# outcome effect are misspecified
scenario4 <- function(df){

  estimates <- as.data.frame(matrix(NA, nrow = 1, ncol = 4))

  # 1. multivariable regression analysis
  model_multivar <- glm(Y ~ X + Z, data = df)
  estimates[1, 1] <- model_multivar$coefficients[2]
```

```
  # 2. covariate adjustment using the ps
  ps <- predict(glm(X ~ Z, family = "binomial", data = df), type =
        "response")
  model_covadj <- glm(Y ~ X + ps, data = df)
  estimates[1, 2] <- model_covadj$coefficients[2]

  # stabilized IPW
  ipw <- ifelse(df$X == 1, 1/ps, 1/(1-ps))
  sipw <- ipw/sum(ipw)
  model_sipw <- glm(Y ~ X, weights = sipw, data = df)
  estimates[1, 3] <- model_sipw$coefficients[2]

  # DR estimation
  model_dr <- glm(Y ~ X + Z, weights = sipw, data = df)
  estimates[1, 4] <- model_dr$coefficients[2]

  return(estimates)

}

# function generate_models returns for each repetition the simulation
# details and the estimated treatment effects
generate_models <- function(df){

  # create data frame to store effect estimates in
  effects <- data.frame(matrix(NA, nrow = max(df$repnr) * 4, ncol =
             10))
  colnames(effects) <- c("scenario",
                         "repnr",
                         "n",
                         "cx",
                         "xy",
                         "cy",
                         "coef_multivar",
                         "coef_covadj",
                         "coef_sipw",
                         "coef_dr")

  # store simulation characteristics
  effects$scenario <- rep(seq(c(1:4)), max(df$repnr))
  effects$repnr <- rep(unique(df$repnr), each = 4)
  effects$n <- unique(df$n)
  effects$cx <- unique(df$cx)
  effects$xy <- unique(df$xy)
  effects$cy <- unique(df$cy)

  # for loop to iterate through each repetition
  repnr <- unique(effects$repnr)
  for(i in repnr){

    temp <- df[df$repnr == i, ]
```

```
    # estimate exposure effects under each scenario
    effects[effects$repnr == i & effects$scenario == 1, c(7:10)] <-
    scenario1(temp)
    effects[effects$repnr == i & effects$scenario == 2, c(7:10)] <-
    scenario2(temp)
    effects[effects$repnr == i & effects$scenario == 3, c(7:10)] <-
    scenario3(temp)
    effects[effects$repnr == i & effects$scenario == 4, c(7:10)] <-
    scenario4(temp)

  }

  # return data frame with all simulation details and exposure effect
  # estimates
  return(effects)

}

# save path
path <- "Step 1 - Generated datasets\\"

# save all file names in files
files <- list.files(path = path,
                    pattern = "*.RData")

# START FOR LOOP - loop through each file in the folder
for(i in files){

  # load the data into the environment
  load(paste0(path, i))

  # run function
  effects <- generate_models(df)

  # save each file in folder 'Step 2 - Generated models'
  save(effects, file = Step 2 - Generated models\\", i))

}
```

Step 3 – Model performance

```
performance_measures <- function(data){

  # for each scenario, all estimates will be stored in a matrix
  performance <- matrix(NA, nrow = 4, ncol = 3)
  colnames(performance) <- c("mean(b)",
                             "AB",
                             "RB")
  rownames(performance) <- c("Multivariable regression analysis",
                             "Covariate adjustment using the PS",
```

```
                              "Standardized IPW",
                              "DR estimation")

  # functions to calculate the performance measures
  # 1. absolute bias
  AB <- function(data, variable){
    return(mean(variable - data$xy))
  }

  # 2. relative bias
  RB <- function(data, variable){
    return(mean((variable - data$xy)/data$xy))
  }

  # run for loop
  for(i in unique(data$scenario)){

    df <- data[data$scenario == i, ]

    # mean exposure effect
    performance[1, "mean(b)"] <- mean(df$coef_multivar)
    performance[2, "mean(b)"] <- mean(df$coef_covadj)
    performance[3, "mean(b)"] <- mean(df$coef_sipw)
    performance[4, "mean(b)"] <- mean(df$coef_dr)

    # absolute bias
    performance[1, "AB"] <- AB(df, df$coef_multivar)
    performance[2, "AB"] <- AB(df, df$coef_covadj)
    performance[3, "AB"] <- AB(df, df$coef_sipw)
    performance[4, "AB"] <- AB(df, df$coef_dr)

    # relative bias
    performance[1, "RB"] <- RB(df, df$coef_multivar)
    performance[2, "RB"] <- RB(df, df$coef_covadj)
    performance[3, "RB"] <- RB(df, df$coef_sipw)
    performance[4, "RB"] <- RB(df, df$coef_dr)

    # round to 4 digits
    performance <- round(performance, 4)

    # return scenario number and performance measures matrix
    print(paste0("scenario number ", i))
    print(performance)

  }
}

# save path
path <- "Step 2 - Generated models\\"

# save all file names in files
```

```
files <- list.files(path = path,
                    pattern = "*.RData")

# START FOR LOOP - loop through each file in the folder
for(i in files){

  # load the data into the environment
  load(paste0(path, i))

  # print scenario and all performance measures
  print(i)
  performance_measures(effects)

}
```

Additional file B Model performance for sample size 200

Table B1 Model performance across all simulated scenarios. n = 200

	Parameter values for the confounder-exposure and confounder-outcome associations								
		0.14			**0.39**			**0.59**	
	$\hat{\beta}$	AB	RB	$\hat{\beta}$	AB	RB	$\hat{\beta}$	AB	RB
Scenario 1: correct specification of cx-association & correct specification of cy-association									
Multivariable regression analysis	0.5900	0.0000	0.0000	0.5900	0.0000	0.0000	0.5900	0.0000	0.0000
Covariate adjustment using the PS	0.5902	0.0002	0.0004	0.5917	0.0017	0.0029	0.5917	0.0017	0.0028
Stabilized IPW	0.5921	0.0021	0.0036	0.6182	0.0282	0.0478	0.6715	0.0815	0.1382
DR estimation	0.5900	0.0000	0.0000	0.5900	0.0000	0.0000	0.5900	0.0000	0.0000
Scenario 2: correct specification of cx-association & misspecification of cy-association									
Multivariable regression analysis	0.6256	0.0356	0.0603	0.8031	0.2132	0.3611	0.9813	0.3913	0.6632
Covariate adjustment using the PS	0.5902	0.0002	0.0004	0.5917	0.0017	0.0029	0.5917	0.0017	0.0028
Stabilized IPW	0.5921	0.0021	0.0036	0.6182	0.0282	0.0478	0.6715	0.0815	0.1382
DR estimation	0.5920	0.0020	0.0034	0.6201	0.0301	0.0510	0.6710	0.0810	0.1373
Scenario 3: misspecification of cx-association & correct specification of cy-association									
Multivariable regression analysis	0.5900	0.0000	0.0000	0.5900	0.0000	0.0000	0.5900	0.0000	0.0000
Covariate adjustment using the PS	0.6260	0.0360	0.0611	0.8127	0.2227	0.3774	1.0107	0.4207	0.7131
Stabilized IPW	0.6274	0.0374	0.0633	0.8323	0.2423	0.4107	1.0639	0.4739	0.8032
DR estimation	0.5900	0.0000	0.0000	0.5900	0.0000	0.0000	0.5900	0.0000	0.0000
Scenario 4: misspecification of cx-association & misspecification of cy-association									
Multivariable regression analysis	0.6256	0.0356	0.0603	0.8031	0.2131	0.3611	0.9813	0.3913	0.6632
Covariate adjustment using the PS	0.6260	0.0360	0.0611	0.8127	0.2227	0.3774	1.0107	0.4207	0.7131
Stabilized IPW	0.6274	0.0374	0.0633	0.8323	0.2423	0.4107	1.0639	0.4739	0.8032
DR estimation	0.6271	0.0371	0.0629	0.8250	0.2350	0.3983	1.0424	0.4524	0.7668

Abbreviations: n: sample size; cx-association: confounder-exposure association; cy-association: confounder-outcome association; $\hat{\beta}$: mean estimated exposure effect; AB: absolute bias; RB: relative bias; PS: propensity score; IPW: inverse probability weighting; DR: double robust

Table B2 Model performance across all simulated scenarios, n = 200

	Parameter values for the confounder-exposure and confounder-outcome associations								
		-0.14			-0.39			-0.59	
	$\hat{\beta}$	AB	RB	$\hat{\beta}$	AB	RB	$\hat{\beta}$	AB	RB
Scenario 1: correct specification of cx-association & correct specification of cy-association									
Multivariable regression analysis	0.5900	0.0000	0.0000	0.5900	0.0000	0.0000	0.5900	0.0000	0.0000
Covariate adjustment using the PS	0.5902	0.0002	0.0003	0.5919	0.0019	0.0032	0.5930	0.0030	0.0051
Stabilized IPW	0.5913	0.0013	0.0022	0.6221	0.0321	0.0545	0.6671	0.0771	0.1307
DR estimation	0.5900	0.0000	0.0000	0.5900	0.0000	0.0000	0.5900	0.0000	0.0000
Scenario 2: correct specification of cx-association & misspecification of cy-association									
Multivariable regression analysis	0.6265	0.0365	0.0619	0.7999	0.2099	0.3558	0.9754	0.3854	0.6533
Covariate adjustment using the PS	0.5902	0.0002	0.0003	0.5919	0.0019	0.0032	0.5930	0.0030	0.0051
Stabilized IPW	0.5913	0.0013	0.0022	0.6221	0.0321	0.0545	0.6671	0.0771	0.1307
DR estimation	0.5917	0.0017	0.0029	0.6216	0.0316	0.0536	0.6717	0.0817	0.1384
Scenario 3: misspecification of cx-association & correct specification of cy-association									
Multivariable regression analysis	0.5900	0.0000	0.0000	0.5900	0.0000	0.0000	0.5900	0.0000	0.0000
Covariate adjustment using the PS	0.6270	0.0370	0.0627	0.8105	0.2205	0.3737	1.0042	0.4142	0.7020
Stabilized IPW	0.6283	0.0383	0.0648	0.8320	0.2420	0.4102	1.0561	0.4661	0.7900
DR estimation	0.5900	0.0000	0.0000	0.5900	0.0000	0.0000	0.5900	0.0000	0.0000
Scenario 4: misspecification of cx-association & misspecification of cy-association									
Multivariable regression analysis	0.6265	0.0365	0.0619	0.7999	0.2099	0.3558	0.9754	0.3854	0.6533
Covariate adjustment using the PS	0.6270	0.0370	0.0627	0.8105	0.2205	0.3737	1.0042	0.4142	0.7020
Stabilized IPW	0.6283	0.0383	0.0648	0.8320	0.2420	0.4102	1.0561	0.4661	0.7900
DR estimation	0.6280	0.0380	0.0644	0.8239	0.2339	0.3965	1.0350	0.4450	0.7543

Abbreviations: n: sample size; cx-association: confounder-exposure association; cy-association: confounder-outcome association; $\hat{\beta}$: mean estimated exposure effect; AB: absolute bias; RB: relative bias; PS: propensity score; IPW: inverse probability weighting; DR: double robust

Table C1 Model performance across all simulated scenarios, n = 1,000

	Parameter values for the confounder-exposure and confounder-outcome associations								
	0.14			**0.39**			**0.59**		
	$\hat{\beta}$	AB	RB	$\hat{\beta}$	AB	RB	$\hat{\beta}$	AB	RB
Scenario 1: correct specification of cx-association & correct specification of cy-association									
Multivariable regression analysis	0.5900	0.0000	0.0000	0.5900	0.0000	0.0000	0.5900	0.0000	0.0000
Covariate adjustment using the PS	0.5900	0.0000	0.0000	0.5903	0.0003	0.0006	0.5908	0.0008	0.0013
Stabilized IPW	0.5902	0.0002	0.0003	0.5971	0.0071	0.0120	0.6212	0.0312	0.0528
DR estimation	0.5900	0.0000	0.0000	0.5900	0.0000	0.0000	0.5900	0.0000	0.0000
Scenario 2: correct specification of cx-association & misspecification of cy-association									
Multivariable regression analysis	0.6263	0.0363	0.0615	0.8027	0.2127	0.3606	0.9863	0.3963	0.6718
Covariate adjustment using the PS	0.5900	0.0000	0.0000	0.5903	0.0003	0.0006	0.5908	0.0008	0.0013
Stabilized IPW	0.5902	0.0002	0.0003	0.5971	0.0071	0.0120	0.6212	0.0312	0.0528
DR estimation	0.5903	0.0003	0.0004	0.6002	0.0102	0.0173	0.6277	0.0377	0.0639
Scenario 3: misspecification of cx-association & correct specification of cy-association									
Multivariable regression analysis	0.5900	0.0000	0.0000	0.5900	0.0000	0.0000	0.5900	0.0000	0.0000
Covariate adjustment using the PS	0.6267	0.0367	0.0621	0.8124	0.2224	0.3770	1.0145	0.4245	0.7195
Stabilized IPW	0.6275	0.0375	0.0636	0.8316	0.2416	0.4095	1.0640	0.4740	0.8033
DR estimation	0.5900	0.0000	0.0000	0.5900	0.0000	0.0000	0.5900	0.0000	0.0000
Scenario 4: misspecification of cx-association & misspecification of cy-association									
Multivariable regression analysis	0.6263	0.0363	0.0615	0.8027	0.2127	0.3606	0.9863	0.3963	0.6718
Covariate adjustment using the PS	0.6267	0.0367	0.0621	0.8124	0.2224	0.3770	1.0145	0.4245	0.7195
Stabilized IPW	0.6275	0.0375	0.0636	0.8316	0.2416	0.4095	1.0640	0.4740	0.8033
DR estimation	0.6272	0.0372	0.0631	0.8237	0.2337	0.3960	1.0423	0.4523	0.7667

Abbreviations: n: sample size; cx-association: confounder-exposure association; cy-association: confounder-outcome association; $\hat{\beta}$: mean estimated exposure effect; AB: absolute bias; RB: relative bias; PS: propensity score; IPW: inverse probability weighting; DR: double robust

Table C2 Model performance across all simulated scenarios, n = 1000

	Parameter values for the confounder-exposure and confounder-outcome associations								
		-0.14			-0.39			-0.59	
	$\hat{\beta}$	AB	RB	$\hat{\beta}$	AB	RB	$\hat{\beta}$	AB	RB
Scenario 1: correct specification of cx-association & correct specification of cy-association									
Multivariable regression analysis	0.5900	0.0000	0.0000	0.5900	0.0000	0.0000	0.5900	0.0000	0.0000
Covariate adjustment using the PS	0.5900	0.0000	0.0001	0.5904	0.0004	0.0006	0.5905	0.0005	0.0009
Stabilized IPW	0.5903	0.0003	0.0005	0.5980	0.0080	0.0136	0.6188	0.0288	0.0489
DR estimation	0.5900	0.0000	0.0000	0.5900	0.0000	0.0000	0.5900	0.0000	0.0000
Scenario 2: correct specification of cx-association & misspecification of cy-association									
Multivariable regression analysis	0.6260	0.0360	0.0611	0.8030	0.2130	0.3610	0.9859	0.3959	0.6711
Covariate adjustment using the PS	0.5900	0.0000	0.0001	0.5904	0.0004	0.0006	0.5905	0.0005	0.0009
Stabilized IPW	0.5903	0.0003	0.0005	0.5980	0.0080	0.0136	0.6188	0.0288	0.0489
DR estimation	0.5904	0.0004	0.0006	0.6013	0.0113	0.0192	0.6286	0.0386	0.0654
Scenario 3: misspecification of cx-association & correct specification of cy-association									
Multivariable regression analysis	0.5900	0.0000	0.0000	0.5900	0.0000	0.0000	0.5900	0.0000	0.0000
Covariate adjustment using the PS	0.6264	0.0364	0.0617	0.8126	0.2226	0.3773	1.0141	0.4241	0.7189
Stabilized IPW	0.6273	0.0373	0.0632	0.8315	0.2415	0.4094	1.0635	0.4735	0.8025
DR estimation	0.5900	0.0000	0.0000	0.5900	0.0000	0.0000	0.5900	0.0000	0.0000
Scenario 4: misspecification of cx-association & misspecification of cy-association									
Multivariable regression analysis	0.6260	0.0360	0.0611	0.8030	0.2130	0.3610	0.9859	0.3959	0.6711
Covariate adjustment using the PS	0.6264	0.0364	0.0617	0.8126	0.2226	0.3773	1.0141	0.4241	0.7189
Stabilized IPW	0.6273	0.0373	0.0632	0.8315	0.2415	0.4094	1.0635	0.4735	0.8025
DR estimation	0.6270	0.0370	0.0627	0.8237	0.2337	0.3962	1.0420	0.4520	0.7661

Abbreviations: n: sample size; cx-association: confounder-exposure association; cy-association: confounder-outcome association; $\hat{\beta}$: mean estimated exposure effect; AB: absolute bias; RB: relative bias; PS: propensity score; IPW: inverse probability weighting; DR: double robust

Table D1 Model performance across all simulated scenarios, n = 500

	Parameter values for the confounder-exposure and confounder-outcome associations								
		-0.14			-0.39			-0.59	
	$\hat{\beta}$	AB	RB	$\hat{\beta}$	AB	RB	$\hat{\beta}$	AB	RB
Scenario 1: correct specification of cx-association & correct specification of cy-association									
Multivariable regression analysis	0.5900	0.0000	0.0000	0.5900	0.0000	0.0000	0.5900	0.0000	0.0000
Covariate adjustment using the PS	0.5901	0.0001	0.0001	0.5909	0.0009	0.0015	0.5914	0.0014	0.0023
Stabilized IPW	0.5900	0.0000	0.0000	0.6092	0.0192	0.0326	0.6444	0.0544	0.0922
DR estimation	0.5900	0.0000	0.0000	0.5900	0.0000	0.0000	0.5900	0.0000	0.0000
Scenario 2: correct specification of cx-association & misspecification of cy-association									
Multivariable regression analysis	0.6259	0.0359	0.0608	0.8031	0.2131	0.3612	0.9882	0.3982	0.6750
Covariate adjustment using the PS	0.5901	0.0001	0.0001	0.5909	0.0009	0.0015	0.5914	0.0014	0.0023
Stabilized IPW	0.5900	0.0000	0.0000	0.6092	0.0192	0.0326	0.6444	0.0544	0.0922
DR estimation	0.5904	0.0004	0.0007	0.6094	0.0194	0.0330	0.6460	0.0560	0.0949
Scenario 3: misspecification of cx-association & correct specification of cy-association									
Multivariable regression analysis	0.5900	0.0000	0.0000	0.5900	0.0000	0.0000	0.5900	0.0000	0.0000
Covariate adjustment using the PS	0.6263	0.0363	0.0615	0.8130	0.2230	0.3780	1.0168	0.4268	0.7234
Stabilized IPW	0.6272	0.0372	0.0631	0.8329	0.2429	0.4117	1.0671	0.4771	0.8086
DR estimation	0.5900	0.0000	0.0000	0.5900	0.0000	0.0000	0.5900	0.0000	0.0000
Scenario 4: misspecification of cx-association & misspecification of cy-association									
Multivariable regression analysis	0.6259	0.0359	0.0608	0.8031	0.2131	0.3612	0.9882	0.3982	0.6750
Covariate adjustment using the PS	0.6263	0.0363	0.0615	0.8130	0.2230	0.3780	1.0168	0.4268	0.7234
Stabilized IPW	0.6272	0.0372	0.0631	0.8329	0.2429	0.4117	1.0671	0.4771	0.8086
DR estimation	0.6269	0.0369	0.0626	0.8250	0.2350	0.3983	1.0451	0.4551	0.7714

Abbreviations: n: sample size; cx-association: confounder-exposure association; cy-association: confounder-outcome association; $\hat{\beta}$: mean estimated exposure effect; AB: absolute bias; RB: relative bias; PS: propensity score; IPW: inverse probability weighting; DR: double robust

Additional file E Sign change as a results of misspecification of the confounder-exposure and confounder-outcome associations

In some scenarios, incorrect modelling of non-linear confounder-exposure and confounder-outcome associations may lead to a change of the direction of the exposure effect. In this illustration, we consider an exposure effect of -0.14 (a small negative effect), and confounder-exposure and confounder-outcome associations of 0.59 (a large positive effect). The sample size and the number of repetitions are both 1,000.

Table E1 shows that, if the confounder-exposure association is misspecified (scenario 2), multivariable regression analysis results in a positive exposure effect estimate. If the confounder-outcome association is misspecified, a sign change occurs for covariate adjustment using the PS and stabilized IPW. If both associations are misspecified, all methods estimate a positive exposure effect.

Table E1 Model performance across simulated scenarios for sample size 1000 and exposure effect -0.14

	Parameter value for the confounder-exposure and confounder-outcome associations: 0.59		
	$\hat{\beta}$	AB	RB
Scenario 1: correct specification of cx-association & correct specification of cy-association			
Multivariable regression analysis	-0.1400	0.0000	0.0000
Covariate adjustment using the PS	-0.1402	-0.0002	0.0016
Stabilized IPW	-0.1271	0.0129	-0.0925
DR estimation	-0.1400	0.0000	0.0000
Scenario 2: correct specification of cx-association & misspecification of cy-association			
Multivariable regression analysis	0.2556	0.3956	-2.8261
Covariate adjustment using the PS	-0.1402	-0.0002	0.0016
Stabilized IPW	-0.1471	0.0129	-0.0925
DR estimation	-0.1087	0.0313	-0.2236
Scenario 3: misspecification of cx-association & correct specification of cy-association			
Multivariable regression analysis	-0.1400	0.0000	0.0000
Covariate adjustment using the PS	0.2844	0.4244	-3.0313
Stabilized IPW	0.3339	0.4739	-3.3849
DR estimation	-0.1400	0.0000	0.0000
Scenario 4: misspecification of cx-association & misspecification of cy-association			
Multivariable regression analysis	0.2556	0.3956	-2.8261
Covariate adjustment using the PS	0.2844	0.4244	-3.0313
Stabilized IPW	0.3339	0.4739	-3.3849
DR estimation	0.3120	0.4520	-3.2288

CHAPTER 3

Modelling non-linear relationships in epidemiological data: the application and interpretation of spline models

Noah A. Schuster
Judith J.M. Rijnhart
Jos W.R. Twisk
Martijn W. Heymans

Published in Frontiers in Epidemiology (2022)

Abstract

Objective

Traditional methods to deal with non-linearity in regression analysis often result in loss of information or compromised interpretability of the results. A recommended but underutilised method for modelling non-linear associations in regression models is spline functions. We explain spline functions in a non-mathematical way and illustrate the application and interpretation to an empirical data example.

Methods

Using data from the Amsterdam Growth and Health Longitudinal Study, we examined the non-linear relationship between the sum of four skinfolds and VO2max, which are measures of body fat and cardiorespiratory fitness, respectively. We compared traditional methods (i.e., quadratic regression and categorization) to spline methods (1- and 3-knot linear spline (LSP) models and a 3-knot restricted cubic spline (RCS) model) in terms of the interpretability of the results and their explained variance (r^2_{adj}).

Results

The spline models fitted the data better than the traditional methods. Increasing the number of knots in the LSP model increased the explained variance (from $r^2_{adj} = 0.578$ for the 1-knot model to $r^2_{adj} = 0.582$ for the 3-knot model). The RCS model fitted the data best ($r^2_{adj} = 0.591$), but results in regression coefficients that are harder to interpret.

Conclusion

Spline functions should be considered more often as they are flexible and can be applied in commonly used regression analysis. RCS regression is generally recommended for prediction research (i.e., to obtain the predicted outcome for a specific exposure value), whereas LSP regression is recommended if one is interested in the effects in a population.

Introduction

In epidemiological research, regression analysis is often used to examine the association between an outcome and an exposure (1). A principal assumption of regression analysis is that the continuous exposure is linearly related to the outcome. In other words, a one-unit difference in the exposure is associated with a fixed difference in the outcome, regardless of the values of the exposure (2). However, linearity should not be assumed without assessing that the association is indeed linear (3-5). If the linearity assumption is violated and associations are estimated as linear nonetheless, then the effect estimate might not be a good representation of the true underlying effect and bias might be introduced. In order to obtain unbiased effects, the non-linear association requires explicit modelling. Failing to estimate a truly non-linear relationship as non-linear may lead to over- or underestimation of the exposure effect. However, it is important to note that the estimation of complex models may come at cost of increase uncertainty, especially in small samples. Therefore, in practice, one may want to consider the balance between model complexity and model uncertainty when choosing an appropriate method to model non-linear relationships

There are different methods available to model non-linear associations. Simple methods such as polynomial regression (e.g., quadratic or cubic regression) and categorization of the exposure variable are widely used, largely due to historical precedent (6). With quadratic regression, for instance, the outcome is modelled as a quadratic function of the exposure (i.e., as a function of exposure $\boldsymbol{x}$ and the quadratic term $\boldsymbol{x^2}$) (2, 7, 8). Adding higher order terms (such as a quadratic term) to a basic linear function increases the flexibility of the model, but simultaneously complicates the interpretability of the results as the regression coefficients of the terms cannot be interpreted separately from each other.

With categorization, the exposure variable is grouped (e.g., based on percentile values) and subsequently analysed as a categorical variable with one of the groups as the reference category. However, categorization is associated with multiple issues, such as loss of information, discontinuity in the estimated average outcome value when moving from one category to the other, and difficulties with comparing results across studies as the cut-off points may be data dependent (2, 6, 8-13). Filardo et al. found that study findings were inconsistent under different exposure categorization schemes identified in the literature, which suggests that the way the exposure is categorized may impact

conclusions (14). This emphasizes the importance of correctly modelling non-linear relationships.

A different approach to model non-linear associations is the use of spline functions in the regression model (2, 3, 8, 11, 12, 15, 16). Spline functions are transformations of the continuous exposure variable and can be added to any regression analysis. They are available in different forms, such as simple linear spline (LSP) functions, more complex restricted cubic spline (RCS) functions and B-splines (2). Spline functions estimate exposure effects for specific intervals of the exposure variable and are subject to continuity restrictions (i.e., the interval functions meet at the common interval edges so that - in contrast to categorization – there are no jumps in the line at these points) (17). In this paper, we focus on LSP and RCS functions. LSP functions assume that the exposure effects within each interval follow a linear shape, but across the intervals the effect may be non-linear. Therefore, LSP functions are more flexible than simple linear regression and categorization. RCS functions assume that the exposure effects within each category are cubic functions, allowing for more flexibility than other methods.

Although spline functions are broadly accessible in the software packages commonly used by epidemiologists, they are not widely used (3, 18). Most papers published on spline functions present these as complex mathematical functions (15, 19, 20) and do not discuss their interpretation. This may be one of the reasons that researchers default to less optimal methods for estimating non-linear effects, such as quadratic terms and categorization.

The aim of this paper is to describe linear and restricted cubic spline functions in a step-by-step and non-mathematical manner, and to demonstrate the advantages of these methods over simple linear regression, quadratic regression and categorization using an empirical data example. First, we provide an introduction into spline regression and describe linear- and restricted cubic spline regression in the context of an empirical data example. Then, we illustrate the application of traditional methods and spline methods to model non-linear relationships to that same data example. Finally, we discuss the interpretation of the effect estimates from different methods and describe the context in which the use of LSP and RCS models may be relevant.

Methods

Example dataset

Spline functions will be explained by using an empirical data example from the Amsterdam Growth and Health Longitudinal Study (AGHLS). The AGHLS is an ongoing cohort study that was set up to examine the growth, health and lifestyle among teenagers (21). We use data from the third round of measurements, when the participants were 15 years old, because it contains a clear non-linear relationship.

Throughout this paper, we analyse the non-linear relationship between the sum of four skinfolds (SFS) and cardiorespiratory fitness (VO2max). SFS is an often used estimate of body fat and is calculated by summing the biceps-, triceps-, subscapular- and suprailiac skinfolds (in millimetres) (22). VO2max is defined as the absolute maximal oxygen uptake in centilitre per kilogram bodyweight (21). The relationship between SFS and VO2max in our data is shown in Figure 1. Only subjects with complete data on both variables were included in the analysis (n = 315, 6 subjects were excluded because of incomplete data).

Spline functions

Splines can be applied to any statistical model that linearly relates the exposure to the outcome, such as linear-, logistic- and Cox regression. With spline models, the continuous independent variable is divided into multiple intervals, and for each interval the relationship between the exposure and outcome is estimated separately. The relationship between the exposure and the outcome in each interval can, for example, be estimated with a linear function (resulting in linear spline regression) or with a cubic function (resulting in cubic spline regression). The use of so-called *spline basis functions*

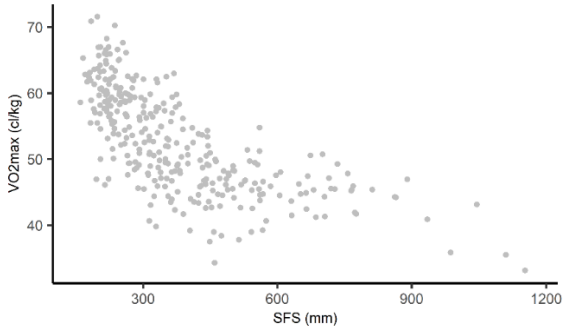

Figure 1 Non-linear relationship between SFS and VO2max in AGHLS data.
Abbreviations: SFS: sum of four skinfolds; AGHLS: Amsterdam Growth and Health Longitudinal Study

makes it possible to estimate the relationship between the exposure and the outcome for each of the intervals in the same model. The values of the exposure based on which the intervals are created are referred to as *knots*. Thus, each knot defines the end of one interval and the start of the next. In 3-knot models, the exposure is divided into four intervals. Subsequently, for each interval the exposure effect is estimated, resulting in four spline coefficients. Corresponding confidence intervals can, for example, be calculated with the standard errors or be obtained by bootstrapping (23).

In general, a small number of knots (i.e., 3 to 5) is sufficient to model a non-linear relationship. If the sample size is large and the relationship that is studied changes quickly, then more knots might be required (2, 24, 25). Increasing the number of knots generally improves the fit of the model, but may also lead to overfitting of the model to the data. If that is the case, the fitted function does not only follow the main features of the data but also small and random fluctuations (2, 7, 25). Wand presents an overview of statistical methods for establishing the number of knots (26).

Often, the locations of the knots are pre-specified based on the quantiles of the independent variable. For 3-knot models, Harrell recommends knots at the 10^{th}, 50^{th} and 90^{th} percentile. For 4-knot models, they are recommended at the 5^{th}, 35^{th}, 65^{th} and 95^{th} percentile (2). In some cases, knot locations are suggested by theory or by study design (e.g., an interrupted time series design). However, generally the fit of a spline model is more dependent on the number of knots than on the knot locations (25).

In this paper, for illustrational purposes, we demonstrate 1- and 3-knot linear spline models and a 3-knot restricted cubic spline model using the knot locations recommended by Harrell. Figure 2 shows the most important properties of a spline model. The grey points in Figure 2 represent the observed data, and the black line is the fitted 3-knot linear spline model. The vertical dotted lines represent the three knots (labelled as k1, k2 and k3) and the lines in between the knots represent the estimated exposure effect for the four intervals between the knots. Spline models are based on continuity restrictions, which ensures that the line is smooth at the knots. For example, the line for the first interval is smoothly connected to the line of the second interval, and the line of the second interval is smoothly connected to the line of the third interval, etcetera. An interactive visualization of LSP and RCS models and the influence of the continuity restrictions, number of knots and location of knots on the estimated line can be found elsewhere (27, 28).

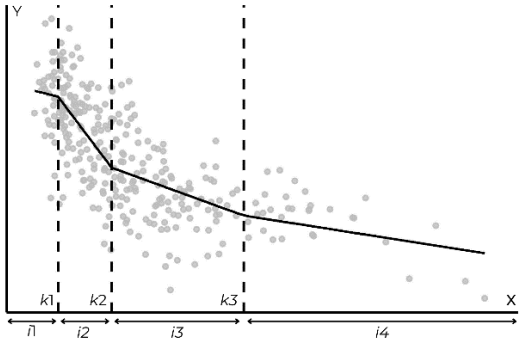

Figure 2 Graphical depiction of the important properties of a spline model. The grey points represent the observed data, and the black line is the fitted linear spline model. The vertical dotted lines represent the knots (located at *k1*, *k2* and *k3*). *i1*, *i2*, *i3* and *i4* represent the four intervals for which the exposure effect is estimated.

Linear spline models

In the 1-knot LSP model, the knot is located at the 50th percentile ($SFS = 330\ mm$). The corresponding linear spline model is

$$VO2\max = \beta_0 + \beta_1 * SFS + \beta_2^* * (SFS - 330)_+ + \varepsilon \quad (1)$$

where β_0 represents the intercept and ε represents an error term. To provide valid inference via e.g. confidence intervals for coefficients, it is assumed that the error terms for each observation are uncorrelated and follow a Gaussian distribution with expected value of zero. The term $(SFS - 330)_+$ represents the *spline basis function*. This function is assigned a value of zero when $SFS - 330 \leq 0$. Because of this, coefficient β_1 represents the exposure effect estimate for individuals whose SFS is equal to or less than 330 mm. Coefficient β_2^* represents the difference in the effect estimates between the individuals whose SFS is equal to or less than 330 mm and those whose SFS is greater than 330 mm. Thus, for individuals whose SFS is greater than 330 mm, their exposure effect estimate is represented by $\beta_1 + \beta_2^*$. The 95% confidence interval corresponding to β_2^* can be used to assess whether the slopes for the two intervals of SFS are statistically significantly different.

In the 3-knot LSP model, the knots are located at the 10th, 50th and 90th percentiles, i.e., at SFS = 212 mm, 330 mm and 621.4 mm, respectively. The corresponding LSP model is

$$VO2max = \beta_0 + \beta_1 * SFS + \beta_2^* * (SFS - 212)_+ + \beta_3^* * (SFS - 330)_+ + \beta_4^* * (SFS - 621.4)_+ + \varepsilon \quad (2)$$

In equation 2, spline coefficient β_2^* is only used whenever an individuals' SFS value is larger than 212, otherwise it is multiplied by zero and thus plays no role in the equation. For coefficient β_3^* this is for $SFS > 330$ and for coefficient β_4^* this is for $SFS > 621.4$, respectively. Thus, coefficient β_1 represents the exposure effect estimate for individuals whose SFS is equal to or less than 212 mm, while $\beta_1 + \beta_2^*$ represents the effect estimate for individuals whose SFS is greater than 212 and equal to or less than 330 mm. The exposure effect estimates for individuals in the third and fourth interval (i.e., individuals whose SFS is greater than 330 mm and equal to or less than 621.4 mm, and individuals whose SFS is greater than 621.4 mm) are represented by $\beta_1 + \beta_2^* + \beta_3^*$ and $\beta_1 + \beta_2^* + \beta_3^* + \beta_4^*$, respectively.

For both the 1- and 3-knot LSP models, fitting the spline models is straightforward once the spline basis functions have been established. Appendix A contains a step by step description of how to estimate these models, including R software code.

Restricted cubic spline models

Although LSP models can approximate many relationships, they do not draw smooth lines and do not fit highly curved relationships well. This can be resolved by fitting a cubic spline model, which joins smoothly at the knot locations because the slopes are restricted to be equal at the boundaries (8). To improve the performance of the spline model in the tails of the exposure variable, where little data is located, additional constraints are imposed in *restricted* cubic spline models. In RCS models, the spline functions are linear in the tails (i.e., before the first and after the last knot) (2, 29). Whereas in LSP models each interval is represented by a spline basis function, in RCS models $k - 2$ spline variables are fitted, where k is the number of knots. Thus, in a 3-knot restricted spline function, a single spline basis function is fitted (equation 3)

$$VO2max = \beta_0 + \beta_1 * SFS + \beta_2^\dagger * SFS_2^\dagger + \varepsilon \quad (3)$$

where $SFS_2^\dagger$ and $\beta_2^\dagger$ represent the spline basis function and corresponding cubic spline coefficient (2). Each participant's value for the spline basis function is estimated as a function of the observed exposure value and the knot locations (i.e., SFS = 212, 330 and 621.4, respectively). The exact formula with which spline basis function $SFS_2^\dagger$ is calculated is presented in Appendix B. Equation 3 can also be expressed as equation 4, which

contains the interval functions and has the same form as the 3-knot LSP model. The only difference between the LSP and RCS models is that for RCS regression all spline basis functions are raised to the power of three:

$$VO2max = \beta_0 + \beta_1 * SFS + \beta_2^* * (SFS - 212)_+^3 + \beta_3^* * (SFS - 330)_+^3 + \beta_4^* * (SFS - 621.4)_+^3 + \varepsilon \quad (4)$$

Equation 5 to 7 can be used to convert cubic spline coefficient $\beta_2^\dagger$ into regression coefficients for each of the intervals:

$$\beta_2^* = \frac{\beta_2^\dagger}{(621.4 - 212)^2} \quad (5)$$

$$\beta_3^* = \frac{\beta_2^* * (212 - 621.4)}{(621.4 - 330)} \quad (6)$$

$$\beta_4^* = \frac{\beta_2^* * (212 - 330)}{(330 - 621.4)} \quad (7)$$

In equation 5, β_2^* represents the coefficient for the interval between the first and the second knot and $\beta_2^\dagger$ is the cubic spline basis function coefficient from equation 3. In equation 6, β_3^* represents the coefficient for the interval between the second and third knot and β_2^* is the regression coefficient from equation 4. In equation 7, β_4^* represents the coefficient for the interval after the third and β_2^* is the regression coefficient from equation 4. Subsequently, coefficients β_2^*, β_3^* and β_4^* can be plugged into equation 4

Like in quadratic regression, the exposure effect estimates differ across exposure values, which makes it less straightforward to interpret the coefficients from an RCS model.

Results

We illustrate the interpretation and compare the performance of different methods to model non-linear relationships using the data example from the AGHLS. Table 1 presents the regression coefficients for each method. For the spline models, these regression coefficients are used to calculate the effects for each interval of SFS. These effects are presented under 'interval coefficient'. Table 2 presents the adjusted r^2 (i.e., the proportion of variance in VO2max explained by SFS) of each method (30).

Table 1 Regression- and interval coefficients for the relationship between VO2max and SFS derived from linear- and quadratic regression, categorization, 1- and 3-knot linear spline regression and 3-knot restricted cubic spline regression

Estimate	Regression coefficient	Interval coefficient
Linear regression		
β_0	64.0658	
β_1	-0.0304	
Quadratic regression		
β_0	73.2212	
β_1	-0.0746	
β_2	0.00004	
Categorization		
β_0	60.1339	
β_1	-4.9870	
β_2	-10.0695	
β_3	-15.1727	
1-knot linear spline regression		
β_0	77.5648	
β_1	-0.0810	$SFS \leq 330: -0.0810$
β_2^*	0.0632	$SFS > 330: -0.0810 + 0.0632 = -0.0178$
3-knot linear spline regression		
β_0	64.1788	
β_1	-0.0156	$SFS \leq 212: -0.0156$
β_2^*	-0.0671	$212 < SFS \leq 330: -0.0156 - 0.0671 = -0.0827$
β_3^*	0.0601	$330 < SFS \leq 621.4: -0.0156 - 0.0671 + 0.0601 = -0.0226$
β_4^*	0.0128	$SFS > 621.4: -0.0156 - 0.0671 + 0.0601 + 0.0128 = -0.0098$
3-knot restricted cubic spline		
β_0	75.9306	
β_1	-0.0738	
$\beta_2^\dagger$	0.0740	β_2^*: 0.0000004 β_3^*: -0.0000006 β_4^*: 0.0000002

β_2^*, β_3^* and β_4^* represent spline coefficients that correspond to spline basis functions. $\beta_2^\dagger$ represents the cubic spline coefficient that corresponds to spline variable $SFS_2^\dagger$

Table 2 Explained variance of each model

Model	Adjusted r^2
Linear regression	0.487
Quadratic regression	0.558
Categorization	0.537
1-knot linear spline regression	0.578
3-knot linear spline regression	0.582
3-knot restricted cubic spline regression	0.591

For illustrative purposes we first estimated a simple linear regression model. Linear regression fits a straight line to the data (Figure 3A) and assumes that the effect of the exposure on the outcome is the same for every value of the exposure. In our data, the exposure effect estimate was -0.0304, meaning that a 1 mm difference in SFS was associated with a 0.0304 cl/kg lower VO2max, regardless of the compared values of SFS. Naturally, this regression line was not a good representation of the relationship between SFS and VO2max, which was also reflected in the lowest explained variance ($r^2_{adj} = 0.487$) of all estimated models.

Quadratic regression

With quadratic regression, VO2max was estimated by SFS and the quadratic term SFS^2. As shown in Figure 3B and reflected in the explained variance ($r^2_{adj} = 0.558$), the quadratic model fitted the form of the relationship between SFS and VO2max quite well relative to the other models. However, the regression coefficients do not have a straightforward interpretation because the effect of SFS on VO2max is a function of both regression coefficients. That is, the effect of a one unit difference in SFS on VO2max differs across SFS. For example, the average difference in VO2max was -0.0506 cl/kg when SFS changed from 300 to 301 (i.e., $(-0.0746 * 301 + 0.00004 * 301^2) - (-0.0746 * 300 + 0.00004 * 300^2)$), while the average difference in VO2max was -0.0266 cl/kg when SFS changed from 600 to 601 (i.e., $(-0.0746 * 601 + 0.00004 * 601^2) - (-0.0746 * 600 + 0.00004 * 600^2)$). Compared to simple linear regression (Figure 3A), the confidence interval for the line estimated using quadratic regression becomes wider for higher values of SFS (Figure 3B). This reflects the additional uncertainty in the effect estimates from quadratic regression for higher SFS values. However, the wider confidence interval does not affect the conclusion that SFS is associated with VO2max.

Categorization
We divided SFS into four intervals based on quartiles. Because we used the lowest quartile as the reference category, the intercept represented the mean VO2max in cl/kg for individuals in that interval. The regression coefficients represented the mean difference in VO2max between individuals in the lowest quartile and the other quartiles. For example, -4.9870 was the mean difference in VO2max in cl/kg between subjects in the first and second quartile. The explained variance was slightly lower relative to the other models ($r^2_{adj} = 0.537$).

Figure 3C illustrates the assumed homogeneity within groups and the discontinuity in VO2max (i.e., the change in average VO2max value) when moving from one quartile to the next. For example, measures of SFS in the last quartile ranged between 458 and 1153 mm, but all individuals had the same estimated VO2max of 44.9612 cl/kg (i.e., $60.1339 - 15.1727$).

1-knot linear spline model
For individuals whose SFS was equal to or less than 330 mm, a 1 mm difference in SFS was associated with a 0.0810 cl/kg lower VO2max. The mean difference in the effect estimate between individuals in both intervals was 0.0632, meaning that for individuals whose SFS was greater than 330 mm, a 1 mm difference in SFS was associated with a 0.0178 cl/kg lower VO2max (i.e., $-0.0810 + 0.0632$). Thus, for individuals whose SFS was greater than 330 mm the association between SFS and VO2max was less strong than for individuals whose SFS was equal to or less than 330 mm. This is also illustrated in Figure 3D. The $r^2_{adj}$was 0.578. This indicates that the 1-knot linear spline model is a better fit to the data than both quadratic regression and categorization.

3-knot linear spline model
For individuals whose SFS was equal to or less than 212 mm, a 1 mm difference in SFS was associated with a 0.0156 cl/kg lower VO2max. For individuals whose SFS was between 213 and 330 mm, a 1 mm difference in SFS was associated with a 0.0827 cl/kg lower VO2max (i.e., $-0.0156 - 0.0671$). The interval coefficients for the other intervals can be found in Table 1.

Increasing the number of knots from 1 to 3 resulted in a slightly higher explained variance ($r^2_{adj} = 0.578$ versus $r^2_{adj} = 0.582$, respectively). Furthermore, compared to simple linear regression (Figure 3A) and the 1-knot model (Figure 3D), the confidence interval for the

line estimated using a 3-knot model becomes wider for higher values of SFS (Figure 3E). This reflects the additional uncertainty in the effect estimates from the 3-knot model for higher SFS values. However, the wider confidence interval based on the 3-knot model does not affect the conclusion that SFS is associated with VO2max.

3-knot restricted cubic spline regression

Like with quadratic regression, separate interpretation of the coefficients is of no practical value with RCS regression, as the effect of SFS on VO2max is a function of multiple regression coefficients. For example, the average decrease in VO2max was 0.0644 cl/kg

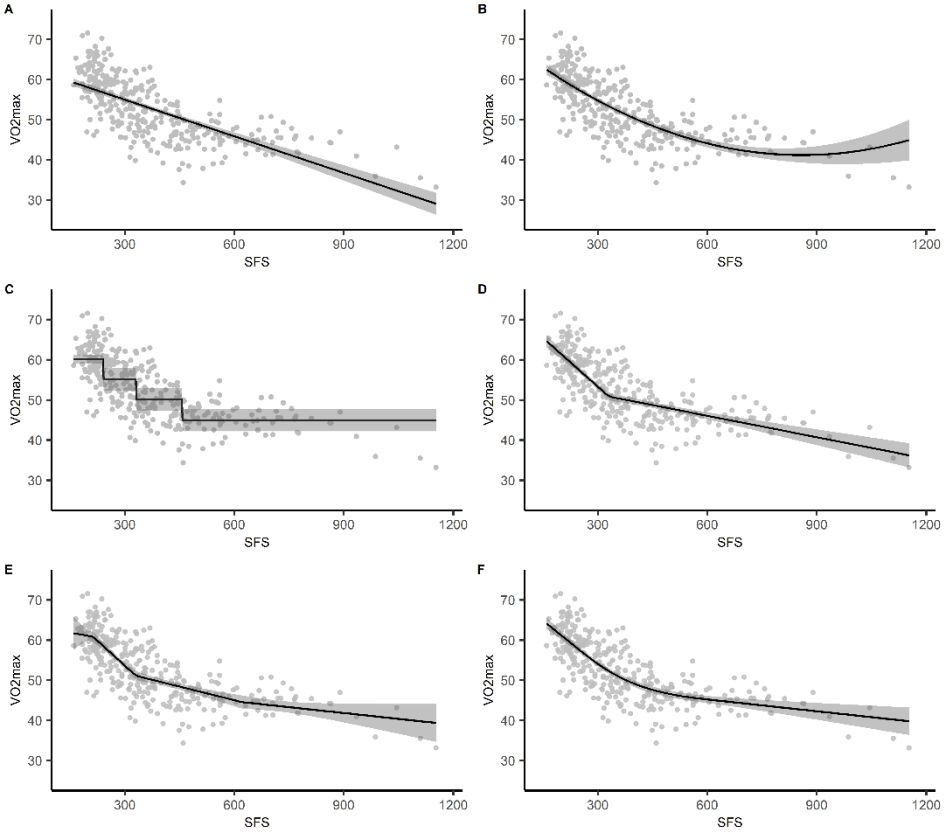

Figure 3 The estimated association between SFS and VO2max plotted against the observed values, with the shading representing the 95% confidence intervals based on standard errors. Panel A: simple linear regression, panel B: polynomial regression, panel C: categorization, panel D: 1-knot LSP regression, panel E: 3-knot LSP regression, panel F: 3-knot RCS regression. Abbreviations: SFS: sum or four skinfolds; LSP: linear spline; RCS: restricted cubic spline

when SFS changed from 300 to 301 mm (i.e., $(75.9306 - 0.0738 * 301 + 0.0000004 * (301 - 212)^3) - (75.9306 - 0.0738 * 300 + 0.0000004 * (300 - 212)^3)$), while the average decrease in VO2max was 0.0244 cl/kg when SFS changed from 600 to 601 mm (i.e., $(75.9306 - 0.0738 * 601 + 0.0000004 * (601 - 212)^3 - 0.0000006 * (601 - 330)^3) - (75.9306 - 0.0738 * 600 + 0.0000004 * (600 - 212)^3 - 0.0000006 * (600 - 330)^3)$.
Figure 3F illustrates the 'restrictions' (i.e., the function is linear for $SFS \leq 212$ and $SFS > 621.4$) and shows that the model fits the data quite well. This is also reflected in the explained variance ($r^2_{adj} = 0.591$).

Discussion

The aim of this paper was to explain linear and restricted cubic spline functions in a step-by-step and non-mathematical manner and to demonstrate the advantages of these methods over simple linear regression, quadratic terms and categorization using an empirical data example. Although spline regression is easy to implement with most statistical programs, epidemiologists still often apply traditional methods (e.g., quadratic regression and categorization) to model non-linear relationships.

In the data example, the spline models resulted in higher explained variance than the traditional methods. Both categorization and spline regression divided the continuous exposure variable into intervals. Categorization only allows for variation between categories, so that the estimated outcome is the same for each individual in an interval regardless of their individual exposure value. This explains the stepwise pattern in Figure 3C. Spline regression, on the other hand, allows for variation between and within intervals. As a result, the regression line shifts between knot locations, and regression lines meet at the knot locations. Although polynomial regression is easy to model, it suffers from a lack of smoothness and can lead to implausible curvatures, in particular at the edges. Splines provide a good alternative as they control for this curvature via the continuity restrictions. In addition, RCS models are linear before and after the last knot. LSP models provide a good balance between modelling the non-linear association and providing results that are relatively easy to interpret. Furthermore, RCS models provide a flexible method for modelling the non-linearity of an association, but come at the cost of regression coefficients that are less easy to interpret than LSP models. In our data example, the explained variance in the LSP model and the RCS model were comparable. If one is interested in reporting the association between sum of four skinfolds and VO2max, then LSP models provide easier interpretations than the RCS models.

For both quadratic regression and RCS models, the increased complexity of the interpretation of the regression coefficients makes it less straightforward to summarize the exposure effect at the population level, because the exposure effect estimates differs in magnitude across exposure values. However, this is not necessarily a problem when the aim of a study is to make individual-level predictions of the outcome, as it remains relatively straightforward to compute the predicted outcome value for a specific exposure value using equation 7 (8). Thus, in our data example, if one is interested in predicting VO2max based on specific values of the sum of four skinfolds, then RCS models may be preferred. Two things that might help with interpreting the results are the reporting of figures (such as Figure 3) and calculating the effect for a number of different exposure contrasts (i.e., the two exposure values that are being compared). The latter was done for the interpretation of the 3-knot RCS model, and showed that the decrease in VO2max was greater when SFS changed from 300 to 301 mm, then when it changed from 600 to 601 mm.

A strength of this paper are the non-mathematical explanations of LSP and RCS models. Although there are many other sources that describe spline models, most of these sources contain a high level of mathematical detail, which may discourage applied researchers from learning about these methods. In this paper, we tried to explained spline functions in a non-mathematical manner and in the context of an empirical data example. Furthermore, although we illustrated the application of spline models using cross-sectional data and within a linear regression context, the spline functions presented can be applied to all kinds of regression models, for example logistic and Cox regression. Further, they can also be used in longitudinal models such as generalized linear mixed models (GLMM) and generalized estimation equations (GEE).

Besides the methods discussed in the present paper, there are also other methods available that can be used to estimate non-linear associations. A method that we did not discuss is quadratic spline regression, in which the spline basis functions are quadratic functions. Although quadratic splines are often overlooked and not mentioned in known reference books (2), like cubic splines they result in smooth functions at the knot locations and can occur in restricted and unrestricted form. When the number of degrees of freedom are the same and the knots are located at comparable exposure values, restricted quadratic and cubic spline models might even yield similar results (31). SAS code for the estimation of restricted quadratic splines is provided by Howe et al. (31). Furthermore, we also did not discuss generalized additive models (GAMs), LOESS

smoothing, penalized splines and fractional polynomials (32, 33), which are all capable of capturing non-linear relationships. However, these methods are relatively complicated and therefore, not much used in practice.

In this paper, we explained spline models based on a single exposure. However, in practice, researchers may want to adjust their association model for potential confounders of the exposure-outcome association. Most researchers are unaware that, if these confounders are continuous, then the linearity assumption also applies to these variables (34). Failing to explicitly model a non-linear confounder-outcome association may result in an under- or overestimation of the true exposure effect. Therefore, the linearity assumption should be assessed for each continuous confounder in a regression model, and splines can be applied when necessary.

Spline regression is easy to implement with most statistical software programs often used by epidemiologists. Table 3 contains a (non-exhaustive) overview of packages and macros available in different software programs. The analyses in this paper were conducted using the R programming language version 4.0.3 (35) and the 'rms' package by Harrell (23). The R package 'splines' is part of the basic distribution of R (29). Other frequently downloaded packages include 'gss' (36) and 'polspline' (37). An overview of spline methods and other R packages that may be used to fit spline models is presented elsewhere (29). In STATA, spline functions can be fitted using, among others, the STATA package 'rmkspline' and the user-made package 'RCsplines' (38). In SPSS, spline functions have to be fitted by hand and can be applied using the REGRESSION procedure. In SAS, the 'effect' statement in 'proc glimmix' provides an automated implementation for fitting splines. Documentation including syntax commands are available from the IBM support page (39) and the SAS Help Center (40).

Although splines are easy to implement, they require certain choices to be made by the researcher. This concerns, for example, the number and location of the knots and the

Table 3 Spline regression options by software program

Software program	Packages/procedures
R	rms, splines, gss, polspline
STATA	mkspline, RCsplines
SPSS	REGRESSION
SAS	TRANSREG

type of basis function (2). In addition, not all non-linear relations are 'equally harmful' and the choice of spline model (e.g., linear or cubic) might depend on what's considered more important: LSP models might be used to model relations that only have a slight bend and that can be approximated by piecewise linear functions, whereas RCS might be used for maximum model accuracy. Another thing to consider is that some choices, such as increasing the number of knots, might introduce additional uncertainty to the model, especially in small samples. If the number of knots is too large, then the model overfits the data: it then describes the random error rather than the relationship between the variables. This affects the generalizability of the model outside of the data that it is based on (29). In our example, the confidence intervals were generally wider for more complex models, illustrating the additional model uncertainty introduced by more complex models. In some situations, the additional uncertainty might be a reason to use a more simple model.

Conclusion

Spline functions should be considered more often in the analysis of non-linear relationships as they allow for more flexibility in estimating non-linear associations than traditional methods such as quadratic regression and categorization and can be used in all kinds of regression analyses. With RCS models the exposure effect estimates differ across exposure values, making them more suitable for prediction (i.e., to obtain the predicted outcome for a specific exposure value). If one is interested in the effects in a population, then LSP models are more suitable due to the straightforward interpretation of the regression coefficients.

References

1. Lash TL, VanderWeele TJ, Haneuse S, Rothman KJ. Modern Epidemiology. 4 ed: Wolters Kluwer; 2021.
2. Harrell FE. Regression Modeling Strategies With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis: Springer International Publishing AG; 2015.
3. Marrie RA, Dawson NV, Garland A. Quantile regression and restricted cubic splines are useful for exploring relationships between continuous variables. Journal of Clinical Epidemiology. 2009;62(5):511-7.e1.
4. Philippe P, Mansi O. Nonlinearity in the Epidemiology of Complex Health and Disease Processes. Theoretical Medicine and Bioethics. 1998;19(6):591-607.
5. Rapoport J, Teres D, Lemeshow S, Avrunin JS, Haber R. Explaining Variability of Cost Using a Severity-of-Illness Measure for ICU Patients. Medical Care. 1990;28(4).
6. Bennette C, Vickers A. Against quantiles: categorization of continuous variables in epidemiologic research, and its discontents. BMC Medical Research Methodology. 2012;12(21).
7. Ruppert D, Wand MP, Carroll RJ. Semiparametric Regression. United States of America: Cambridge University Press; 2003.
8. Greenland S. Dose-response and trend analysis in epidemiology: alternatives to categorical analysis. Epidemiology. 1995;6(4):356-65.
9. Altman DG, Lausen B, Sauerbrei W, Schumacher M. Danger of using "optimal" cut points in the evaluation of prognostic factors. Journal of National Cancer Institute. 1994;86(11):829-35.
10. Gauthier J, Wu QV, Gooley TA. Cubic splines to model relationships between continuous variables and outcomes: a guide for clinicians. Bone Marrow Transplant. 2020;55(4):675-80.
11. Greenland S. Avoiding Power Loss Associated with Categorization and Ordinal Scores in Dose-Response and Trend Analysis. Epidemiology. 1995;6(4):450-4.
12. Royston P, Altman DG, Sauerbrei W. Dichotomizing continuous predictors in multiple regression: a bad idea. Stat Med. 2006;25(1):127-41.
13. Boucher KM, Slattery ML, Berry TD, Quesenberry C, Anderson K. Statistical Methods in Epidemiology: A Comparison of Statistical Methods to Analyze Dose–Response and Trend Analysis in Epidemiologic Studies. Journal of Clinical Epidemiology. 1998;51(12):1223-33.

14. Filardo G, Hamilton C, Hamman B, Ng HKT, Grayburn P. Categorizing BMI may lead to biased results in studies investigating in-hospital mortality after isolated CABG. Journal of Clinical Epidemiology. 2007;60(11):1132-9.
15. Durrleman S, Simon R. Flexible regression models with cubic splines. Statistics in Medicine. 1989;8(5):551-61.
16. James G, Witten D, Hastie T, Tibshirani R. An introduction to statistical learning: with applications in R. New York: Springer; 2013.
17. Marsh LC, Cormier DR. Spline Regression Models: Sage Publications, Inc.; 2002.
18. O'Brien SM. Cutpoint Selection for Categorizing a Continuous Predictor. Biometrics. 2004;60(2):504-9.
19. de Boor CR. A Practical Guide to Splines: Springer-Verlag New York; 1978.
20. Smith PL. Splines As a Useful and Convenient Statistical Tool. The American Statistician. 1979;33(2):57-62.
21. Wijnstok NJ, Hoekstra T, van Mechelen W, Kemper HCG, Twisk JWR. Cohort Profile: The Amsterdam Growth and Health Longitudinal Study. International Journal of Epidemiology. 2013;42(2):422-9.
22. Wijnstok NJ, Serné EH, Hoekstra T, Schouten F, Smulders YM, Twisk JWR. The relationship between 30-year developmental patterns of body fat and body fat distribution and its vascular properties: the Amsterdam Growth and Health Longitudinal Study. Nutrition & Diabetes. 2013;3(9):e90-e.
23. Harrell FE. rms: Regression Modeling Strategies. R package version 6.0-1 2020 [Available from: https://cran.r-project.org/package=rms.
24. Korn EL, Graubard bI. Analysis of Health Surveys. 1 ed: Wiley-Interscience; 1999.
25. Stone CJ. Additive splines in statistics. American Statistical Association Proceedings of the Statistical Computing Setting. 1985:45-8.
26. Wand MP. A comparison of regression spline smoothing procedures. Computational Statistics. 2000;15:443-62.
27. Lambert P. Spline Continuity n.d. [Available from: https://pclambert.net/interactivegraphs/spline_continuity/spline_continuity.
28. Lambert P. The Number and Location of Knots n.d. [Available from: https://pclambert.net/interactivegraphs/spline_eg/spline_eg.
29. Perperoglou A, Sauerbrei W, Abrahamowicz M, Schmid M. A review of spline function procedures in R. BMC Medical Research Methodology. 2019;19(1):46.
30. Ezekiel M. Methods of Correlation Analysis. New York: John Wiley and Sons; 1930.

31. Howe CJ, Cole SR, Westreich DJ, Greenland S, Napravnik S, Eron JJ, Jr. Splines for trend analysis and continuous confounder control. Epidemiology (Cambridge, Mass). 2011;22(6):874-5.
32. Eisen EA, Agalliu I, Thurston SW, Coull BA, Checkoway H. Smoothing in occupational cohort studies: an illustration based on penalised splines. Occup Environ Med. 2004;61(10):854-60.
33. Binder H, Sauerbrei W, Royston P. Comparison between splines and fractional polynomials for multivariable model building with continuous covariates: a simulation study with continuous response. Statistics in Medicine. 2013;32(13):2262-77.
34. Groenwold RHH, Klungel OH, Altman DG, van der Graaf Y, Hoes AW, Moons KGM, et al. Adjustment for continuous confounders: an example of how to prevent residual confounding. CMAJ. 2013;185(5):401-6.
35. R Core Team. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2019.
36. Gu C. Smoothing spline ANOVA models: R package gss. Journal of Statistical Software. 2014;58(5):1-25.
37. Kooperberg C. polspline: Polynomial Spline Routines. R package version 1.1.20 2022 [Available from: https://CRAN.R-project.org/package=polspline.
38. Cox NJ. RCSPLINE: Stata module for restriced cubic spline smoothing. Statistical Software Components S456884. 2007.
39. IBM SPSS Statistics. Spline regression with estimated knots in SPSS 2020 [Available from: https://www.ibm.com/support/pages/spline-regression-estimated-knots-spss#:~:text=Regression%20models%20in%20which%20the, with%20the%20SPSS%20REGRESSION%20procedure.
40. SAS Programming Documentation. The TRANSREG Procedure 2019 [Available from: https://documentation.sas.com/doc/en/pgmsascdc/9.4_3.4/statug/statug_transreg_details03.htm.

SUPPLEMENTARY MATERIALS FOR CHAPTER 3

Appendix A Fitting 1-knot and 3-knot linear spline models (by hand) including R software code

```
library(rms)
attach(data)
```

1-knot linear spline model

Determine knot location according to 50th percentile of SFS

```
t1 <- quantile(SFS, 0.50)
```

Compute SFS spline basis function **xt1** and recode so that $SFS - t_1 = 0$ if $SFS \leq t_1$

```
xt1 <- SFS – t1
xt1[xt1 <= 0] <- 0
```

Linear regression analysis including spline basis function **xt1**

```
ols(VO2max ~ SFS + xt1)
```

Automated implementation using the rms package

```
ols(VO2max ~ lsp(SFS, quantile(SFS, 0.5)), data = data)
```

3-knot linear spline model

Determine knot locations according to 10th, 50th and 90th percentile of SFS

```
t1 <- quantile(SFS, 0.10)
t2 <- quantile(SFS, 0.50)
t3 <- quantile(SFS, 0.90)
```

Compute SFS spline basis functions **xt1, xt2** and **xt3** and recode

```
xt1 <- SFS - t1
xt1[xt1 <= 0] <- 0

xt2 <- SFS - t2
xt2[xt2 <= 0] <- 0

xt3 <- SFS - t3
xt3[xt3 <= 0] <- 0
```

Linear regression analysis including spline basis functions **xt1, xt2** and **xt3**

```
ols(VO2max ~ SFS + xt1 + xt2 + xt3)
```

Automated implementation using the rms package

```
ols(VO2max ~ lsp(SFS, quantile(SFS, c(0.1, 0.5, 0.9))), data = data)
```

Appendix B Formula to calculate spline variable $\boldsymbol{SFS_2^{\dagger}}$

$$SFS_2^{\dagger} = \frac{(SFS - 212)_+^3 - (SFS - 330)_+^3 * \frac{(621.4 - 212)}{(621.4 - 330)} + (SFS - 621.4)_+^3 * \frac{(330 - 212)}{(621.4 - 212)}}{(621.4 - 212)^2}$$

where $\boldsymbol{SFS_2^{\dagger}}$ is the spline variable and $\boldsymbol{SFS}$ is the original exposure variable. Values 212, 330 and 621.4 represent the first, second and third knot location, respectively.

Appendix C Fitting a 3-knot restricted cubic spline model (by hand) including R software

```
library(rms)
attach(data)
```

Determine knot locations according to 10th, 50th and 90th percentile of SFS

```
t1 <- quantile(SFS, 0.10)
t2 <- quantile(SFS, 0.50)
t3 <- quantile(SFS, 0.90)
```

Compute SFS spline basis functions **xt1, xt2** and **xt3** and recode

```
xt1 <- SFS - t1
xt1[xt1 <= 0] <- 0

xt2 <- SFS - t2
xt2[xt2 <= 0] <- 0

xt3 <- SFS - t3
xt3[xt3 <= 0] <- 0
```

Compute spline variable **SFS_RCS** using equation B1

```
SFS_RCS <- (xt1^3 - xt2^3 * ((t3 - t1)/(t3 - t2)) + xt3^3 * ((t2 - t1)/(t3
– t1)))/(t3 - t1)^2
```

Linear regression analysis including spline variable **SFS_RCS**

```
fit <- ols(VO2max ~ SFS + SFS_RCS)
```

Assign cubic spline coefficient, corresponding to spline variable **SFS_RCS**, to object **csc**

```
b0 <- fit$coefficients[1]
b1 <- fit$coefficients[2]
csc <- fit$coefficients[3]
```

Transform cubic spline coefficient **csc** into regression coefficients **b2, b3** and **b4** corresponding to spline basis functions **xt1, xt2** and **xt3** using equations 4 to 6

```
b2 <- csc/(t3 - t1)^2
b3 <- (b2 * (t1 - t3))/(t3 - t2)
b4 <- (b2 * (t1 - t2))/(t2 - t3)
```

The complete regression formula becomes

VO2max = `b0 + b1 * SFS + b2 * xt1^3 + b3 * xt2^3 + b4 * xt3^3`

Automated implementation using the rms package

```
ols(VO2max ~ rcs(SFS, 3), data = data)
```

CHAPTER 4

Noncollapsibility and its role in quantifying confounding bias in logistic regression

Noah A. Schuster
Jos W.R. Twisk
Gerben ter Riet
Martijn W. Heymans
Judith J.M. Rijnhart

Published in BMC Medical Research Methodology (2021)

Abstract

Background

Confounding bias is a common concern in epidemiological research. Its presence is often determined by comparing exposure effects between univariable- and multivariable regression models, using an arbitrary threshold of a 10% difference to indicate confounding bias. However, many clinical researchers are not aware that the use of this change-in-estimate criterion may lead to wrong conclusions when applied to logistic regression coefficients. This is due to a statistical phenomenon called noncollapsibility, which manifests itself in logistic regression models. This paper aims to clarify the role of noncollapsibility in logistic regression and to provide guidance in determining the presence of confounding bias.

Methods

A Monte Carlo simulation study was designed to uncover patterns of confounding bias and noncollapsibility effects in logistic regression. An empirical data example was used to illustrate the inability of the change-in-estimate criterion to distinguish confounding bias from noncollapsibility effects.

Results

The simulation study showed that, depending on the sign and magnitude of the confounding bias and the noncollapsibility effect, the difference between the effect estimates from univariable- and multivariable regression models may underestimate or overestimate the magnitude of the confounding bias. Because of the noncollapsibility effect, multivariable regression analysis and inverse probability weighting provided different but valid estimates of the confounder-adjusted exposure effect. In our data example, confounding bias was underestimated by the change in estimate due to the presence of a noncollapsibility effect.

Conclusion

In logistic regression, the difference between the univariable- and multivariable effect estimate might not only reflect confounding bias but also a noncollapsibility effect. Ideally, the set of confounders is determined at the study design phase and based on subject matter knowledge. To quantify confounding bias, one could compare the unadjusted exposure effect estimate and the estimate from an inverse probability weighted model.

Background

In observational studies, the exposure levels are often influenced by characteristics of the study subjects. As a result, differences in background characteristics between exposed and unexposed individuals may exist. If these characteristics are also associated with the outcome, crude comparison of the average outcomes in both exposure groups does not yield an unbiased estimate of the exposure effect (1-5). Therefore, to obtain unbiased effects, adjustment for this imbalance in background characteristics is necessary. This is also called adjustment for confounding.

When selecting confounders for adjustment, researchers often use statistical methods to quantify the confounding bias. That is, oftentimes the confounding bias is quantified by comparing the exposure effect between a univariable- and a multivariable regression model, also called the change-in-estimate criterion (4, 6, 7). However, this method may lead to wrong conclusions about the presence and magnitude of confounding bias, as in logistic regression covariates may affect the effect estimate through two separate mechanisms: through confounding when covariates are associated with both the exposure and the outcome, and through noncollapsibility which is present when covariates are associated with the outcome (8). The total difference between the effect estimate from a univariable- and multivariable regression model may therefore be decomposed into an estimate of confounding bias and an estimate of the noncollapsibility effect (7, 9). Furthermore, even in the absence of confounding the exposure effect coefficients from both models might still differ. Thus, the change-in-estimate may misrepresent the true confounding bias (4).

Various rescaling methods have been proposed in the social sciences literature, which aim to equalize the scales of the effect estimates from a univariable and a multivariable regression model (10-13). However, when applied to effect estimates from a logistic regression, these rescaling measures are approximate rather than exact (10, 11, 14). Janes et al. (9) and Pang et al. (7) proposed an exact measure of confounding bias for logistic regression models. This measure is based on the comparison of the effect estimates from a univariable regression model and an inverse probability weighted (IPW) model. The latter is another popular method to adjust for confounding.

Noncollapsibility may not only affect the differences between the effect estimates from a univariable- and multivariable regression model, it also causes differences between the effect estimates from a multivariable regression model and an IPW model. Whereas

multivariable regression and IPW provide the same effect estimates in linear regression, this does not necessarily hold for logistic regression (7, 9, 15). That is, when a noncollapsibility effect is present, multivariable regression adjustment and IPW both yield valid estimates of the confounder-adjusted exposure effect, but their magnitude and interpretation differ (7, 16, 17). Therefore, the difference between the effect estimates from a multivariable regression model and IPW can be used to quantify the magnitude of noncollapsibility.

Because noncollapsibility is a relatively unknown mechanism among clinical researchers, many are unaware that the change-in-estimate criterion may lead to wrong conclusions about the presence and magnitude of confounding bias. Therefore, this paper aims to clarify the role of noncollapsibility in logistic regression and to provide guidance in determining the presence of confounding bias. First, we review the different confounder-adjustment methods and provide a detailed explanation of the noncollapsibility effect. Then, we use a Monte Carlo simulation study to uncover patterns of confounding bias and noncollapsibility effects in logistic regression. Subsequently, using an empirical data example, we demonstrate that the change-in-estimate criterion to determine confounding bias may be misleading. Finally, we provide guidance in determining the set of confounders and quantifying confounding bias.

Confounder adjustment and noncollapsibility

The presence and magnitude of confounding bias for models with a binary outcome is commonly determined by comparing the exposure effect estimates from a univariable- (equation 1) and multivariable (equation 2) logistic regression model:

$$logit\big(Pr(Y = 1|X)\big) = i_1 + \beta_1 X \quad (1)$$

$$logit\big(Pr(Y = 1|X, C_1, \dots, C_n)\big) = i_2 + \beta_1' X + \beta_2' C_1 + \cdots + \beta_{n+1}' C_n \quad (2)$$

where in both equations, Y and X represent the outcome and exposure variables and i_1 and i_2 represent the intercept terms, respectively. In equation 1, β_1 represents the *unadjusted* exposure effect estimate. In equation 2, β_1' represents the multivariable confounder-adjusted exposure effect estimate and β_2' to β_{n+1}' are the coefficients corresponding to observed background covariates C_1 to C_n. When C_1 to C_n are truly confounders, then β_1 will be a biased estimate of the causal exposure-outcome effect. Assuming that equation 2 contains all confounders of the exposure-outcome effect, β_1'

will have a causal interpretation. In practice researchers often determine the magnitude of confounding as the change in estimate, which is computed as the difference between β_1' and β_1. When using the change-in-estimate criterion to determine the presence of confounding bias typically a 10% difference between β_1' and β_1 is used in practice as an arbitrary threshold indicating confounding due to covariates C_1 to C_n in the association between X and Y (6, 18, 19).

When based on logistic regression, $\beta_1' - \beta_1$ may not only represent confounding bias but also a noncollapsibility effect. This noncollapsibility effect is sometimes also referred to as a form of the Simpson's paradox (16). The noncollapsibility effect is caused by a difference in the scale on which β_1 and β_1' are estimated. In linear regression, the total variance is the same for nested models: when the explained variance increases through adding a covariate to the model, the *un*explained variance decreases by the same amount. As a result, effect estimates from nested linear models are on the same scale and thus *collapsible.* In logistic regression, however, the unexplained variance has a fixed value of 3.29 (8). Adding covariates that are associated with the outcome (e.g., confounders) increases the explained variance and forces the total variance of *Y* to increase. When the total variance of *Y* increases, the scale of the estimated coefficients changes, causing negative exposure effects to become more negative and positive exposure effects more positive. This change in scales is called the *noncollapsibility* effect (5, 7, 8). Thus, to determine confounding bias, exposure effect estimates cannot be simply compared between nested logistic regression models as the difference might not only reflect confounding bias but also a noncollapsibility effect (8). The noncollapsibility effect also occurs when a covariate is associated with outcome *Y* but not with exposure *X* (i.e., when the covariate is not a confounder). The change in estimate then represents the noncollapsibility effect only, falsely indicating the presence of confounding bias. To preserve space in the main text, a hypothetical example illustrating how the change in estimate might be affected by the noncollapsibility effect in the absence of confounding is given in additional file A. An explanation of noncollapsibility based on a contingency table is provided by for example Pang et al. (7).

Recent studies by Janes et al. and Pang et al. presented an exact estimate of confounding bias unaffected by noncollapsibility based on logistic regression (7, 9), using the difference between the univariable exposure effect estimate and the effect estimate from an IPW model. With IPW, confounding bias is eliminated by creating a pseudo-population in which each covariate combination is balanced between both exposure groups (20-22).

When there is perfect covariate balance there is no longer an association between covariates C_1 to C_n and exposure status X. This pseudo-population can be created by weighting subjects so that for each combination of baseline covariates the sums of contributions for both exposure groups are equal (1, 20). These weights are the inverse of the probability that a subject was exposed, i.e. the inverse of a propensity score (23).

The propensity score is the predicted probability of endorsing the exposure, which can be estimated using equation 3:

$$PS = Pr(X = 1|C) = \frac{1}{1 + e^{-(i_3+\lambda_1 C_1+\cdots+\lambda_n C_n)}} \quad (3)$$

where *X* represents exposure, i_3 is the model intercept and λ_1 to λ_n are regression coefficients corresponding to covariates C_1 to C_n. The propensity score methodology can also be extended to continuous exposure variables using the Generalized Propensity Score (GPS), which has a similar balancing property to the classic propensity score. For more information on how to perform propensity score analysis with a continuous exposure variable, see Hirano (2004) and Imai (2004) (24, 25).

For exposed subjects, the weight is calculated as $\frac{1}{PS}$ and for unexposed subjects as $\frac{1}{1-PS}$ (1, 20, 22). Using these calculations, subjects with a propensity score close to 0 end up with large weights, and subjects with a propensity score close to 1 end up with small weights. Because in some situations these weights cause the IPW model to be unstable, stabilized weights have been proposed (26). For exposed subjects, the stabilized weight is calculated as $\frac{p}{PS}$ and for unexposed subjects as $\frac{1-p}{1-PS}$, where *p* is the probability of exposure without considering covariates C_1 to C_n (2, 26). Subsequently, a weighted regression analysis with exposure *X* as the only independent variable is carried out. We call the confounder-adjusted exposure effect estimate from the IPW model β_1^*.

Difference between IPW- and multivariable confounder-adjusted exposure effect estimates
Multivariable regression adjustment and IPW provide identical exposure effect estimates when based on linear regression, but not when based on logistic regression (15, 27). The difference between the IPW confounder-adjusted exposure effect estimate β_1^* and the multivariable confounder-adjusted exposure effect estimate β_1' is caused by noncollapsibility, and the difference between the unadjusted exposure effect estimate β_1 and β_1^* provides a measure of confounding bias (7, 9, 14). This is because in an IPW model

the total variance remains equal to the total variance of the unadjusted model, while in a multivariable regression model the addition of variables to the model leads to higher variance, changing the scale of the exposure effect estimate. This means that when there is confounding in a logistic regression model, multivariable regression analysis and IPW lead to different confounder-adjusted estimates of the exposure effect. Although β_1' and β_1^* are both valid estimates, they apply to different target populations and have their own respective interpretation (8, 27).

Simulation study

Simulation methods

A Monte Carlo simulation study was designed to investigate patterns of confounding bias and noncollapsibility effects in logistic regression. The R programming language version 4.0.2 (28) and STATA statistical software release 14 (29) were used to generate and analyze the data, respectively.

Three continuous covariates were generated from a standard normal distribution. The dichotomous exposure and outcome were generated from a binomial distribution conditional on the covariates and the covariates and exposure, respectively. Sample sizes were 250, 500, 750 and 1000. The parameter values for the exposure-outcome effect, confounder-exposure effect and the confounder-outcome effect were set to -1.42, -0.92, -0.38, 0, 0.38, 0.92 and 1.42. This way, the conditions reflected situations with combinations of zero effects, and positive and negative small (-0.38 and 0.38), medium (-0.92 and 0.92) and large (-1.42 and 1.42) effect sizes were mimicked (30). The total number of conditions was 1,372 with 1,000 repetitions per condition, resulting in 1,372,000 observations. Subsequently, we estimated the unadjusted exposure effect estimate β_1, the multivariable confounder-adjusted exposure effect estimate β_1' and the IPW confounder-adjusted exposure effect estimate β_1^* based on the simulated data. From these effect estimates we computed the change in estimate, the confounding bias and the noncollapsibility effect. The simulation code is available in additional file B.

Simulation scenarios

We expected to observe four scenarios based on the simulated data. In the first scenario (Figure 1A), the covariates are associated with both the exposure and the outcome. In this scenario there will be both confounding bias ($\beta_1 - \beta_1^*$) and a noncollapsibility effect ($\beta_1^* - \beta_1'$). Because the exposure-outcome effect is simulated to be positive and negative,

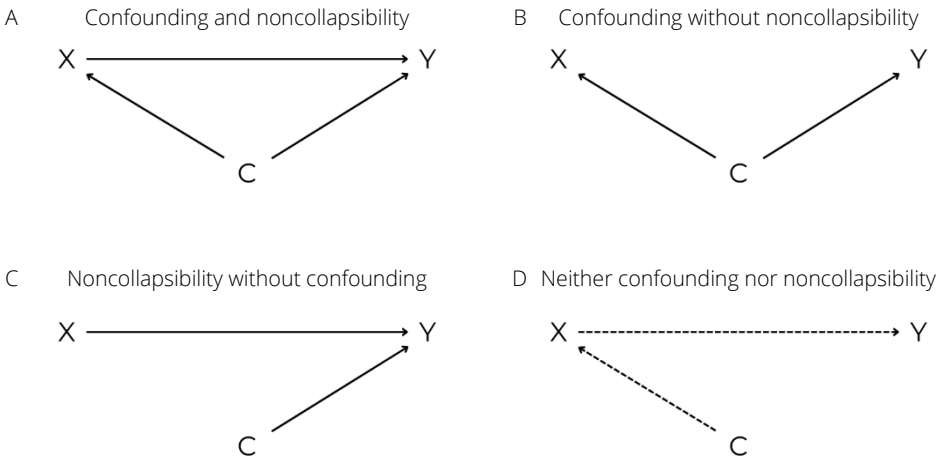

Figure 1 Directed acyclic graphs of the four possible scenarios into which each simulated condition can be classified. Panel A: both confounding and noncollapsibility. Panel B: confounding without noncollapsibility. Panel C: noncollapsibility without confounding. Panel D: neither confounding nor noncollapsibility. *C* represents three continuous covariates, *X* represents the dichotomous exposure and *Y* represents the dichotomous outcome. The dotted line in panel D between the covariates and the exposure and between the exposure and the outcome indicate there may or may not be an association.

we also expect to see positive and negative noncollapsibility effect estimates. This means that $\beta_1' - \beta_1$ might result in an under- or overestimation of the true confounding effect (8). In the second scenario (Figure 1B) the covariates are associated with both the exposure and outcome, but exposure and outcome are not associated with each other. In this scenario, any differences between β_1' and β_1 are fully explained by the covariates, so there is confounding bias without a noncollapsibility effect (8, 15). In the third scenario (Figure 1C), the covariates are only associated with the outcome. In this scenario there is a noncollapsibility effect but no confounding bias. In real-life situations with this structure, using the change-in-estimate criterion may lead one to conclude that the covariates are confounders in the relation between the exposure and the outcome although the difference between β_1' and β_1 is caused entirely by the noncollapsibility effect (7, 8). In the fourth scenario (Figure 1D), the covariates may be associated with the exposure, but not with the outcome. In this scenario, there is neither confounding bias nor a noncollapsibility effect and β_1, β_1' and β_1^* are identical. This scenario is also called *strict* collapsibility (15, 31, 32).

Simulation results

The difference between β_1' and β_1 can be negative, zero or positive, depending on the magnitude of the confounding bias and the noncollapsibility effect (Table 1). Only when there was no noncollapsibility effect (i.e., $\beta_1^* - \beta_1' = 0$), the change in estimate equaled the estimate of confounding bias. The noncollapsibility effect was zero when the exposure-outcome effect was zero and the confounder-exposure and confounder-outcome effects were both non-zero. When the exposure-outcome effect was also non-zero, the difference between β_1' and β_1 reflected both confounding bias and the noncollapsibility effect. In those situations, the change-in-estimate criterion could both under- and

Table 1 Difference between univariable- and multivariable exposure effects as combination of confounding bias and the noncollapsibility effect

Difference between multivariable- and univariable effect estimate ($\beta_1' - \beta_1$)	Confounding bias ($\beta_1 - \beta_1^*$)	Noncollapsibility effect ($\beta_1^* - \beta_1'$)
Negative	Negative value	Negative value
	Zero	Negative value
	Negative value	Zero
	Positive value	Greater negative value than the positive confounding bias value
	Greater negative value than the positive noncollapsibility effect value	Positive value
Zero	Zero	Zero
	Equal positive value as the negative noncollapsibility effect value	Equal negative value as the positive confounding bias value
	Equal negative value as the positive noncollapsibility effect value	Equal positive value as the negative confounding bias value
Positive	Positive value	Positive value
	Zero	Positive value
	Positive value	Zero
	Negative value	Greater positive value than the negative confounding bias value
	Greater positive value than the negative noncollapsibility effect value	Negative value

overestimate the true confounding bias. When the confounding bias and noncollapsibility effect had similar signs, i.e. both were positive or negative, $\beta_1' - \beta_1$ overestimated the true confounding bias. When the confounding bias and noncollapsibility effect had opposites signs, i.e. one was positive while the other was negative, the true confounding bias could be under- or overestimated by $\beta_1' - \beta_1$, depending on the magnitude of the confounding bias and noncollapsibility effect. Thus, when the exposure-outcome effect is non-zero, the change-in-estimate criterion might falsely indicate the presence of confounding or it might under- or overestimate the true confounding bias. Patterns of confounding bias and the noncollapsibility effect were similar across sample sizes and will be described below.

Confounding bias

Figure 2 plots confounding bias ($\beta_1 - \beta_1^*$) as a function of the confounder-outcome effect with the lines in panel A representing positive confounder-outcome effects of various magnitudes and the lines in panel B representing negative confounder-outcome effects of various magnitudes. Confounding bias was positive when the confounder-exposure effect and the confounder-outcome effect were both positive (panel A, first quadrant) and when they were both negative (panel B, second quadrant). When the effects had opposite signs, confounding bias was negative. The magnitude of confounding bias increased as the confounder-exposure or confounder-outcome effect increased in magnitude. There was no confounding bias when one or both effects equaled zero.

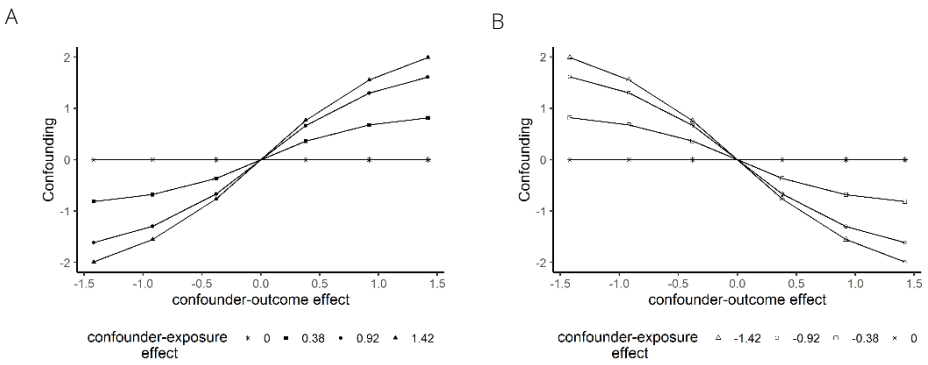

Figure 2 True confounding bias ($\beta_1 - \beta_1^*$) as a function of the confounder-outcome effect collapsed over all sample sizes. Panel A: each line represents a positive confounder-exposure effect. Panel B: each line represents a negative confounder-exposure effect.

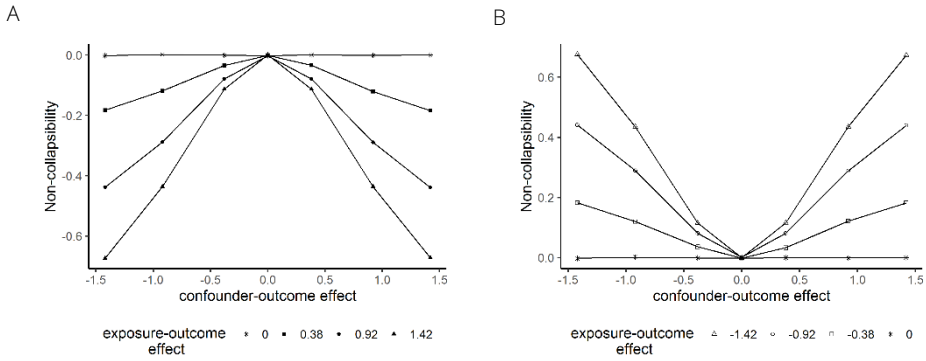

Figure 3 The noncollapsibility effect ($\beta_1^* - \beta_1'$) as a function of the confounder-outcome effect collapsed over all sample sizes. Panel A: each line represents a positive exposure-outcome effect. Panel B: each line represents a negative exposure-outcome effect.

The noncollapsibility effect

Figure 3 plots the noncollapsibility effect ($\beta_1^* - \beta_1'$) as a function of the confounder-outcome effect with the lines in panel A representing positive exposure-outcome effects of various magnitudes and the lines in panel B representing negative exposure-outcome effects of various magnitudes. The noncollapsibility effect and the exposure-outcome effect were inversely related: when the latter effect was positive, the noncollapsibility effect was negative, and vice versa. The noncollapsibility effect increased in magnitude as both the exposure-outcome effect and the confounder-outcome effect increased in magnitude. When either effect was zero, there was no noncollapsibility effect, regardless of the magnitude of the other effect.

Empirical data example

To illustrate how the noncollapsibility effect might affect conclusions about confounding bias in practice we use an example from the Amsterdam Growth and Health Longitudinal Study (AGHLS). The AGHLS started in 1976 with the aim was to examine growth and health among teenagers. Over the years, health and lifestyle measures, determinants of chronic diseases and parameters for the investigation of deterioration in health with age have been measured (33). The data in this example were collected in 2000, when the participants were in their late 30s. Using data from the AGHLS we investigated the association between hypercholesterolemia and hypertension, potentially confounded by physical activity. Using multivariable regression analysis and IPW we estimated the confounder-adjusted effect of hypercholesterolemia on hypertension in our sample, β_1'

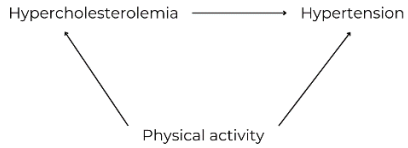

Figure 4 The assumed relations between hypercholesterolemia, hypertension and physical activity

and β_1^*, respectively. To quantify the magnitude of confounding bias and the noncollapsibility effect, we also estimated the unadjusted exposure effect β_1 using univariable regression analysis. Cut-offs for hypercholesterolemia and hypertension were based on guidelines from the U.S. National Institutes of Health (NIH) and NIH's National Heart, Lung and Blood Institute, respectively (34, 35). Physical activity was defined as the total hours per week spent on light, moderate or vigorous activities. Only subjects with complete data on the variables were considered in the analysis (n = 349). Figure 4 provides a graphical representation of the assumed relations among the variables.

Table 2 shows the effect estimates from univariable- and multivariable regression analysis and IPW. The unadjusted effect estimate β_1 was 0.90, corresponding to an odds ratio (OR) of 2.46. The multivariable confounder-adjusted exposure effect estimate β_1' was 0.93, corresponding to an OR of 2.53. The IPW confounder-adjusted exposure effect estimate β_1^* was 0.99, corresponding to an OR of 2.69. The difference between β_1' and β_1 was -0.03, or 3.3%. If one would use the change-in-estimate criterion with a cut-off of 10% to determine the presence of confounding, then physical activity would not be

Table 2 Relationship between hypercholesterolemia and hypertension estimated using univariable- and multivariable regression analysis and IPW

	β	$SE(\beta)$	95% CI	p
Univariable exposure effect				
Hypercholesterolemia	0.90	0.23	0.47; 1.35	< 0.01
Multivariable confounder-adjusted exposure effect				
Hypercholesterolemia	0.93	0.23	0.48; 1.38	< 0.01
Physical activity	0.01	0.01	-0.02; 0.03	0.60
IPW confounder-adjusted exposure effect				
Hypercholesterolemia	0.99	0.16	0.69; 1.30	< 0.01

Abbreviations: SE: standard error; CI: confidence interval

considered a confounder. Using the difference between β_1 and β_1^*, the estimate of confounding bias was 0.90 – 0.99 = -0.09. This corresponds to a 10% change in the exposure effect estimate. The noncollapsibility effect estimate was 0.99 – 0.93 = 0.06. Because of this noncollapsibility effect, the estimate of the true confounding bias of physical activity was considerably larger than it seemed based on the difference between β_1' and β_1. Thus, in our data example, the conventional method to determine the presence of confounding led to an underestimation of the true confounding bias of physical activity.

Discussion

This paper aimed to clarify the role of noncollapsibility in determining the magnitude of confounding bias in logistic regression. Because the difference between β_1' and β_1 reflects both confounding bias and a noncollapsibility effect, in logistic regression the change-in-estimate criterion should not be used to determine the presence of confounding. This was illustrated in our data example, in which confounding bias was underestimated because of the magnitude of the noncollapsibility effect. Our simulation study showed that confounding was mainly determined by the combination of the magnitude of the confounder-exposure and confounder-outcome effects, whereas noncollapsibility was mostly determined by the magnitude of the combination of the exposure-outcome and confounder-outcome effects. In situations in which confounding approached zero and noncollapsibility was non-zero, the change-in-estimate criterion wrongly indicated the presence of confounding bias, when in reality the difference between β_1' and β_1 was caused solely by the noncollapsibility effect.

Recommendations for practice

Rather than using an arbitrary statistical rule such as the 10% cut-off based on the change-in-estimate criterion, it is generally recommended to determine the confounder set based on subject matter knowledge. Directed acyclic graphs (DAGs) are helpful to determine which set of confounders should be adjusted for to eliminate confounding bias (36, 37). DAGs are causal diagrams in which the arrows represent the causal relations among variables. Therefore, DAGs contain information about the causal model that cannot be provided by statistical methods. For example, assuming the DAG is a correct representation of the causal relations among variables, it clarifies what the minimally sufficient set of confounders is to block any backdoor paths (i.e., confounding paths) from the exposure to the outcome. The amount of confounding bias could be quantified by looking at the difference between the unadjusted univariable exposure effect estimate

β_1 and the IPW confounder-adjusted exposure effect estimate β_1^* as proposed by Pang et al. (7) and Janes et al. (9). Bootstrap confidence intervals can be used to determine the statistical significance of the confounding bias.

Because of the noncollapsibility effect, multivariable regression analysis and IPW provide different estimates of the exposure effect. Multivariable regression analysis results in a *conditional* exposure effect estimate (16, 38), whereas IPW results in a *population-average* or *marginal* exposure effect estimate (16, 38-40). Marginal exposure effects can also be estimated with standardization using G-computation. A step-by-step demonstration of G-computation can be found elsewhere (41). It is often suggested that a population-average effect estimate should be reported when the target population is the entire study population, while the conditional exposure effect should be reported if the target population is a subset of the study population (7, 8, 16, 38, 39, 42, 43). Although this distinction is known from the literature, when it comes to the practical application, the exact differences between the two exposure effect estimates and their respective interpretations remain unclear.

In this study, we assume correct specification of both the confounder-exposure and the confounder-outcome effect. When these are not correctly specified, bias might be introduced and the difference between the unadjusted univariable exposure effect estimate and the IPW confounder-adjusted exposure effect estimate might not only reflect confounding bias but also the misspecification of the underlying models. Therefore, correct specification of all effects is necessary to estimate unbiased exposure effects and correctly quantify confounding bias.

Conclusion

To summarize, in this study we showed that in logistic regression the difference between univariable- and multivariable effect estimates may reflect both confounding bias and a noncollapsibility effect. To avoid wrong conclusions with respect to the magnitude and presence of confounding bias, confounders are ideally determined based on subject matter knowledge. To quantify confounding bias, one could look at the difference between the unadjusted univariable exposure effect estimate and the IPW confounder-adjusted exposure effect estimate.

References

1. 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.
2. Hernan MA, Robins JM. Causal Inference: What If. Boca Raton: Chapman & Hall/CRC; 2020.
3. Samuels ML. Matching and design efficiency in epidemiological studies. Biometrika. 1981;68(3):577-88.
4. Miettinen OS, Cook EF. Confounding: essence and detection. American Journal of Epidemiology. 1981;114(4):593-603.
5. Greenland S, Robins JM. Identifiability, exchangeability, and epidemiological confounding. Int J Epidemiol. 1986;15(3):413-9.
6. Kleinbaum DG, Sullivan KM, Barker ND. A Pocket Guide to Epidemiology: Springer Science + Business Media, LLC; 2007.
7. Pang M, Kaufman JS, Platt RW. Studying noncollapsibility of the odds ratio with marginal structural and logistic regression models. Stat Methods Med Res. 2016;25(5):1925-37.
8. Mood C. Logistic Regression: Why We Cannot Do What We Think We Can Do, and What We Can Do About It. European Sociological Review. 2009;26(1):67-82.
9. Janes H, Dominici F, Zeger S. On quantifying the magnitude of confounding. Biostatistics. 2010;11(3):572-82.
10. Cramer JS. Robustness of Logit Analysis: Unobserved Heterogeneity and Mis-specified Disturbances. Oxford Bulletin of Economics and Statistics. 2007;69(4):545-55.
11. MacKinnon DP, Luecken LJ. How and for whom? Mediation and moderation in health psychology. American Psychological Association; 2008. p. S99-S100.
12. Karlson KB, Holm A, Breen R. Comparing Regression Coefficients Between Same-sample Nested Models Using Logit and Probit: A New Method. Sociological Methodology. 2012;42(1):286-313.
13. Breen R, Karlson KB, Holm A. Total, Direct, and Indirect Effects in Logit and Probit Models. Sociological Methods & Research. 2013;42(2):164-91.
14. Rijnhart JJM, Valente MJ, MacKinnon DP. Total effect decomposition in mediation analysis in the presence of non-collapsibility. Submitted for publication.
15. Greenland S, Robins JM, Pearl J. Confounding and Collapsibility in Causal Inference. Statistical Science. 1999;14(1):29-46.

16. Hernan MA, Clayton D, Keiding N. The Simpson's paradox unraveled. Int J Epidemiol. 2011;40(3):780-5.
17. Neuhaus JM, Jewell NP. A geometric approach to assess bias due to omitted covariates in generalized linear models. Biometrika. 1993;80(4):807-15.
18. Lee PH. Is a cutoff of 10% appropriate for the change-in-estimate criterion of confounder identification? J Epidemiol. 2014;24(2):161-7.
19. Budtz-Jorgensen E, Keiding N, Grandjean P, Weihe P. Confounder selection in environmental epidemiology: assessment of health effects of prenatal mercury exposure. Ann Epidemiol. 2007;17(1):27-35.
20. Heinze G, Juni P. An overview of the objectives of and the approaches to propensity score analyses. Eur Heart J. 2011;32(14):1704-8.
21. Brookhart MA, Wyss R, Layton JB, Sturmer T. Propensity score methods for confounding control in nonexperimental research. Circ Cardiovasc Qual Outcomes. 2013;6(5):604-11.
22. Robins JM, Hernan MA, Brumback B. Marginal Structural Models and Causal Inference in Epidemiology. Epidemiology. 2000;11(5):550-60.
23. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. Biometrika. 1983;70:41-55.
24. Hirano K, Imbens GW. The propensity score with continuous treatments. In: Gelmanand A, Meng X-L, editors. Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives: John Wiley & Sons, Ltd; 2004. p. 73-84.
25. Imai K, van Dyk DA. Causal Inference With General Treatment Regimes. Journal of the American Statistical Association. 2004;99(467):854-66.
26. Xu S, Ross C, Raebel MA, Shetterly S, Blanchette C, Smith D. Use of stabilized inverse propensity scores as weights to directly estimate relative risk and its confidence intervals. Value in Health. 2010;13(2):273-7.
27. Martens EP, Pestman WR, de Boer A, Belitser SV, Klungel OH. Systematic differences in treatment effect estimates between propensity score methods and logistic regression. International Journal of Epidemiology. 2008;37(5):1142-7.
28. R Core Team. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2019.
29. StataCorp. Stata Statistical Software: Release 14. College Station, TX: StataCorp LP; 2015.

30. Chen H, Cohen P, Chen S. How Big is a Big Odds Ratio? Interpreting the Magnitudes of Odds Ratios in Epidemiological Studies. Communications in Statistics - Simulation and Computation. 2010;39(4):860-4.
31. Mansournia MA, Greenland S. The relation of collapsibility and confounding to faithfulness and stability. Epidemiology. 2015;26:466-72.
32. Whittemore AS. Collapsibility of multidimensional contingency tables. Royal Statistical Society. 1978;40(3):328-40.
33. Wijnstok NJ, Hoekstra T, van Mechelen W, Kemper HC, Twisk JW. Cohort profile: the Amsterdam Growth and Health Longitudinal Study. Int J Epidemiol. 2013;42(2):422-9.
34. National Institutes of Health's U.S. National Library of Medicine. Cholesterol levels: what do the results mean n.d. [Available from: https://medlineplus.gov/lab-tests/cholesterol-levels/.
35. NIH: National Heart L, and Blood Institute,. High blood pressure: confirming high blood pressure 2020 [Available from: https://www.nhlbi.nih.gov/health-topics/high-blood-pressure.
36. Shrier I, Platt RW. Reducing bias through directed acyclic graphs. BMC Med Res Methodol. 2008;8:70.
37. Greenland S, Pearl J, Robins JM. Causal diagrams for epidemiologic research. Epidemiology. 1999;10(1):37-48.
38. Daniel R, Zhang J, Farewell D. Making apples from oranges: Comparing noncollapsible effect estimators and their standard errors after adjustment for different covariate sets. Biom J. 2020.
39. Breen R, Karlson KB, Holm A. Interpreting and Understanding Logits, Probits, and Other Nonlinear Probability Models. Annual Review of Sociology. 2018;44(1):39-54.
40. Burgess S. Estimating and contextualizing the attenuation of odds ratios due to non collapsibility. Communications in Statistics - Theory and Methods. 2016;46(2):786-804.
41. Snowden JM, Rose S, Mortimer KM. Implementation of G-Computation on a Simulated Data Set: Demonstration of a Causal Inference Technique. American Journal of Epidemiology. 2011;173(7):731-8.
42. Zhang Z. Estimating a Marginal Causal Odds Ratio Subject to Confounding. Communications in Statistics - Theory and Methods. 2008;38(3):309-21.
43. Karlson KB, Popham F, Holm A. Marginal and Conditional Confounding Using Logits. Sociological Methods & Research. 2021:0049124121995548.

SUPPLEMENTARY MATERIALS FOR CHAPTER 4

Additional file A Hypothetical example

This hypothetical example involves three dichotomous variables: the exposure variable weight (not overweight vs. overweight), the event of interest diabetes and the potential confounding variable sex. The total population consists of 200 individuals. The data can be summarized according to Table A.1.

Of these 200 individuals, half are overweight and half are not overweight. A total of 80 individuals have diabetes and 120 individuals do not have diabetes. The unadjusted exposure effect, estimated with univariable regression analysis, is 1.299. This corresponds to an odds ratio of 3.667.

The sexes are evenly distributed over the weight groups: both groups consist of 50 males and 50 females. Because there are equal numbers of males and females in the overweight and not overweight group, weight status is not influenced by sex. In contrast to sex and weight, sex and diabetes are associated. Of the 100 women, 30 have diabetes and 70 do not, whereas for males half have diabetes and half do not. Because confounding requires the covariate to be associated with both the exposure and the outcome, sex is not a confounder in the relation between weight and diabetes, as it is not associated with weight.

Since sex is not a confounder of the exposure-outcome effect, adjustment for sex should not affect the exposure-outcome effect estimate. However, the adjusted exposure effect estimate, estimated with multivariable regression analysis, is 1.366, corresponding to an odds ratio of 3.921. This is different from the unadjusted effect estimate of 1.299. The results from all analyses are shown in Table A.2.

Although sex is not a confounder, the effect estimates from univariable- and multivariable regression analysis still differ. This difference of 0.067 is entirely caused by noncollapsibility. This example illustrates that, even in the absence of confounding, the univariable and multivariable exposure effect estimates might differ. Therefore, the change-in-estimate based on logistic regression coefficients may lead to wrong conclusions when used to determine the presence of confounding.

Table A.1 Hypothetical data example

Exposure	Event	Sex	*n*
Not overweight	No diabetes	Female	45
Not overweight	Diabetes	Female	5
Not overweight	No diabetes	Male	30
Not overweight	Diabetes	Male	20
Overweight	No diabetes	Female	25
Overweight	Diabetes	Female	25
Overweight	No diabetes	Male	20
Overweight	Diabetes	Male	30

Abbreviations: n: sample size

Additional file B Simulation code

Step 1 – generate data

```
generate_data <- function(seed, reps, n, ix, iy, cx, xy, cy){

  # define total number of rows required to store data
  rows <- n * reps

  # create data frame to store data in
  df <- as.data.frame(matrix(NA, nrow = rows, ncol = 14))
  colnames(df) <- c("ID",            # ID through entire data set
                    "repnr",         # for each repetition
                    "ID_repnr",      # ID through each repetition
                    "n",             # number of observations
                    "ix",            # intercept exposure
                    "iy",            # intercept outcome
                    "cx",            # CX effect
                    "xy",            # XY effect
                    "cy",            # CY effect
                    "C1",            # confounder 1
                    "C2",            # confounder 2
                    "C3",            # confounder 3
                    "X",             # dichotomous exposure
                    "Y")             # dichotomous outcome

  # define simulation parameters
  df[, "ID"] <- seq(1:rows)
  df[, "repnr"] <- rep(1:reps, each = n)
  df[, "ID_repnr"] <- rep(seq(1, n), reps)
  df[, "n"] <- n

  # define intercepts ix and iy
  df[, "ix"] <- ix
  df[, "iy"] <- iy

  # define coefficients a, b and c
  df[, "cx"] <- cx
  df[, "xy"] <- xy
  df[, "cy"] <- cy

  # generate confounders C1, C2 and C3
  df[, "C1"] <- rnorm(n = rows)
  df[, "C2"] <- rnorm(n = rows)
  df[, "C3"] <- rnorm(n = rows)

  # generate dichotomous exposure X
  lpx <- ix + cx * df[, "C1"] + cx * df[, "C2"] + cx * df[, "C3"]
  prx <- 1/(1 + exp(-lpx))
  df[, "X"] <- rbinom(n = rows, size = 1, prob = prx)
```

```
  # generate dichotomous outcome Y
  lpy <- iy + xy * df[, "X"] + cy * df[, "C1"] + cy * df[, "C2"] + cy *
  df[, "C3"]
  pry <- 1/(1 + exp(-lpy))
  df[, "Y"] <- rbinom(n = rows, size = 1, prob = pry)

  # transform X and Y into factor variables
  df[, "X"] <- factor(df[, "X"])
  df[, "Y"] <- factor(df[, "Y"])

  # return data frame
  return(df)

}

# define simulation parameters
seed <- 20200908
reps <- 1000
n <- c(250, 500, 750, 1000)
ix <- 0
iy <- 0
cx <- c(-1.42, -0.92, -0.38, 0, 0.38, 0.92, 1.42)
xy <- c(-1.42, -0.92, -0.38, 0, 0.38, 0.92, 1.42)
cy <- c(-1.42, -0.92, -0.38, 0, 0.38, 0.92, 1.42)

for(i in n){
  for(j in cx){
    for(k in xy){
      for(l in cy){

        df <- generate_data(seed = seed,
                            reps = reps,
                            n = i,
                            ix = ix,
                            iy = iy,
                            cx = j,
                            xy = k,
                            cy = l)

        # save each file in folder 'Step 1 - Generated datasets'
        save(df, file = paste0("Step 1 - Generated datasets\\",
        "n = ", i,", cx = ", j, ", xy = ", k, ", cy = ", l, ".RData"))

      }
    }
  }
}
```

Step 2 – generate models

```
library(dplyr)
```

```
generate_models <- function(data){

  # create data frame to store effect estimates in
  effects <- data.frame(matrix(NA, nrow = max(data$repnr), ncol = 8))
  colnames(effects) <- c("repnr",
                         "n",
                         "cx",
                         "xy",
                         "cy",
                         "coef_univar",
                         "coef_multivar",
                         "coef_ipw")

  # store simulation characteristics
  effects$repnr <- unique(data$repnr)
  effects$n <- unique(data$n)
  effects$cx <- unique(data$cx)
  effects$xy <- unique(data$xy)
  effects$cy <- unique(data$cy)

  # FIT MODELS
  # 1. univariable regression model
  estimates <- data %>%
    group_by(repnr) %>%
    do(model_univar = glm(Y ~ X, family = "binomial", data =
    .)$coefficients[2])

  effects$coef_univar <- unlist(estimates$model_univar)

  # 2. multivariable regression model
  estimates <- data %>%
    group_by(repnr) %>%
    do(model_multivar = glm(Y ~ X + C1 + C2 + C3, family = "binomial",
    data = .)$coefficients[2])

  effects$coef_multivar <- unlist(estimates$model_multivar)

  # 3. inverse probability weighting
  estimates <- data %>%
    group_by(repnr) %>%
    do(ps = predict(glm(X ~ C1 + C2 + C3, family = "binomial", data =
    .), type =
    "response"))

  data$ps <- unlist(estimates$ps)

  data$weights <- ifelse(data$X == 1, 1/data$ps, 1/(1 - data$ps))
  data$stab_weights <- data$weights/sum(data$weights)

  estimates <- data %>%
```

```
    group_by(repnr) %>%
    do(model_ipw = glm(Y ~ X, weights = stab_weights, family =
    "binomial", data =
    .)$coefficients[2])

  effects$coef_ipw <- unlist(estimates$model_ipw)

  # return data frame with simulation characteristics and treatment
  effects
  return(effects)

}

# save path
path <- "Step 1 - Generated datasets\\"

# save all file names in files
files <- list.files(path = path, pattern = "*.RData")

# loop through each file in the folder
for(i in files){

  # load the data into the environment
  load(paste0(path, i))

  # run function
  effects <- generate_models(df)

  # save each file in folder 'Step 2 - Generated models'
  save(effects, file = paste0("Step 2 - Generated models\\", i))

}
```

Step 3 – merge data

```
# save path
path <- "Step 2 - Generated models\\"

# save all file names in files
files <- list.files(path = path, pattern = "*.RData")

# load first dataset (effects) of files
load(paste0(path, files[1]))

# rename dataset (effects) to df
df <- effects
rm(effects)

# append all other files to current file df
for(i in paste0(path, files[-1])){
```

```
  load(i)
  df <- rbind(df, effects)
  rm(effects)

}

# add scenario number for each scenario (total = 1372)
df$scenario <- rep(seq(from = 1, to = nrow(df)/1000), each = 1000)

# change column order
df <- df[c("scenario", "repnr", "n", "cx", "xy", "cy", "coef_univar",
"coef_multivar", "coef_ipw")]

# save appended file df in folder 'Step 3 - Appended file'
save(df, file = "Step 3 - Appended file\\Appended file (all scenarios -
all effect measures).RData")
```

Step 4 – confounding decomposition

```
# df contains the treatment effects derived from the 3 methods,
# calculated for each repetition (n = 1000) within each scenario (n =
# 1372)
load("Step 3 - Appended file\\Appended file (all scenarios - all effect
measures).RData")

# calculate the difference between the unadjusted exposure effect and
the conditional exposure effect
df$diff <- df$coef_univar - df$coef_multivar

# calculate the true confounding effect
df$true_conf <- df$coef_univar - df$coef_ipw

# calculate the amount of non-collapsibility
df$non_collaps <- df$coef_ipw - df$coef_multivar

# save file containing the confounding decomposition in folder 'Step 4
- Confounding decomposition'
save(df, file = "Step 4 - Confounding decomposition\\Final
dataset.RData")
```

CHAPTER 5

Causal mediation analysis with a binary mediator: the influence of the estimation approach and causal contrast

Noah A. Schuster
Jos W.R. Twisk
Martijn W. Heymans
Judith J.M. Rijnhart

Accepted for publication in Structural Equation Modeling: A Multidisciplinary Journal

Abstract

Although causal mediation analysis clarifies causal effect estimation, little attention has been devoted to the differences between causal estimation approaches. This paper illustrates the difference between the causal estimation approaches for mediation models with a binary mediator. Using a Monte Carlo simulation study and an empirical data example we show that the regression- and simulation-based approaches provide indirect and total effect estimates that are dependent on the chosen causal contrast, while the imputation- and weighting-based approaches provide overall effect estimates. The results underline the importance of choosing an estimation approach that provides estimates of the causal effect of interest.

Introduction

Mediation analysis is popular in many fields, including medical and social sciences. With mediation analysis, the total effect of the exposure on the outcome can be decomposed into a direct and an indirect effect (1, 2). For example, cholesterol concentration may be a mediator of the effect of body mass index (BMI) on blood pressure as higher BMI is associated with higher cholesterol, and, in turn, higher cholesterol is associated with higher blood pressure. The positive association between BMI and blood pressure may then be (partially) explained by cholesterol. The direct effect is the effect of the exposure on the outcome after removing the influence of the mediator, i.e., the effect of BMI on blood pressure after removing the influence of cholesterol. The indirect effect is the effect of the exposure on the outcome through the mediator, i.e., the effect of BMI on blood pressure through cholesterol

In the past decade, causal mediation analysis methods gained in popularity and are now implemented in most software packages (3). Whereas traditional mediation analysis defines and estimates the direct and indirect effects in terms of regression coefficients, causal mediation analysis separates the causal effect definitions from effect estimation (4). It defines causal effects as the differences between two potential outcomes, i.e., the outcome that would be observed for a certain exposure (5, 6). For example, if we treat BMI as a binary exposure (i.e., overweight versus healthy weight), then two potential outcomes could be observed: a certain blood pressure if a person is overweight, and a certain blood pressure if the same person has a healthy weight. The difference between these two potential outcomes is the causal effect of weight status on blood pressure at the individual level. If we treat BMI as a continuous exposure, potential outcomes could be observed for any BMI value. To estimate the causal effect of BMI on blood pressure, we choose two BMI values for comparison. For example, we may want to compare someone's blood pressure value if BMI equals 21 to the same person's blood pressure if BMI equals 20. The two compared values for the exposure are also called the *causal contrast*.

However, in practice it is impossible to observe both potential outcomes for the same individual simultaneously. Therefore, the causal mediation effects are estimated on the population-average level as the expected differences between two population-average potential outcomes (5, 7, 8). Different estimation approaches can be used to estimate the population-average causal mediation effects (9, 10). In this paper we focus on four estimation approaches: regression, simulation, imputation and weighting. If the mediator

and outcome are both continuous, then all estimation approaches provide the same causal effect estimates (3). Furthermore, assuming that all pathways (i.e., the exposure-mediator, the mediator-outcome and the exposure-outcome associations) in the mediation model are linear, the mediation effect estimates will be the same for every one unit difference in a continuous exposure variable. Thus, the indirect effect of BMI on blood pressure through cholesterol will be the same when comparing a BMI of 21 and 20, or a BMI of 26 and 25, and so on.

The different estimation approaches do not necessarily provide the same indirect and total effect estimates when the exposure is continuous and the mediator is binary. In this situation, the causal indirect and total effect estimates from the regression-based and simulation-based approaches depend on the chosen causal contrast. That is, when the mediator is binary, such as hypercholesterolemia, the indirect effect based on a comparison of BMI 21 and 20 will differ in magnitude from the indirect effect based on BMI 26 and 25. This is not the case for the imputation- and weighting-based approaches, as these will still provide mediation effect estimates that are the same for every one unit difference in the continuous exposure variable.

Recent reviews showed that the uptake of causal mediation analysis remains low (11-13). Reasons for this may be the high level of technical detail in the seminal papers on causal mediation analysis and unfamiliarity with potential outcomes notation (11, 12, 14). Furthermore, previous studies showed that traditional and causal mediation analysis provide the same effect estimates for models with a continuous mediator and outcome (15, 16), but not necessarily for models with a binary mediator or outcome (17, 18). Furthermore, applied researchers may not be aware that the different causal estimation approaches provide different effect estimates when the outcome is binary. To stimulate the correct application of causal mediation methods, tutorial papers are needed that clarify causal effect estimation for mediation models commonly encountered in practice.

This paper demonstrates that four commonly-used causal estimation approaches provide different effect estimates with different interpretations for mediation models with binary mediators. First, we provide a brief introduction into causal mediation analysis and review the different estimation approaches and the role of the causal contrast in each approach. Then, using a Monte Carlo simulation study, we investigate the performance of the regression-, simulation-, imputation- and weighting-based approaches. Subsequently, using an empirical data example, we illustrate the

consequences of the used estimation approach and the selected causal contrast for the interpretation of the results. Finally, we discuss the interpretation of the effects for the different approaches and the practical implications of this paper.

Causal mediation analysis

In this section, we provide a brief introduction into causal mediation analysis, including the regression-, simulation-, imputation- and weighting-based estimation approaches. Throughout this section we use the running example of hypercholesterolemia as a binary mediator of the effect of BMI on blood pressure.

Causal mediation analysis uses a general notation for the potential outcomes: X denotes the exposure, and Y denotes the outcome (5, 6). The exposure levels of interest, i.e., the causal contrast, are represented by x and x^*. Based on these exposure levels, two potential outcomes could be observed: $Y(x)$ corresponding to exposure level x and $Y(x^*)$ corresponding to exposure level x^* (5, 6). Thus, $Y(x)$ represents the potential outcome for BMI value x, whereas $Y(x^*)$ represent the potential outcome for BMI value x^*. Additionally, we could observe two potential mediator values under the exposure levels of interest: $M(x)$ under exposure level x and $M(x^*)$ under exposure level x^* (8). Thus, $M(x)$ represents the risk of hypercholesterolemia under BMI value x, whereas $M(x^*)$ represents the risk of hypercholesterolemia under BMI value x^*. In a mediation model the potential outcome is a function of both the exposure and the mediator value. Based on the combination of exposure values x and x^* and potential mediator values $M(x)$ and $M(x^*)$, four nested potential outcomes can be defined: $Y(x, M(x))$, $Y(x, M(x^*))$, $Y(x^*, M(x))$ and $Y(x^*, M(x^*))$ (8, 19).

Based on the four nested potential outcomes, the natural direct effect (NDE), natural indirect effect (NIE) and total effect (TE) can be defined (8, 20). For the NDE, the mediator value is held constant at $M(x)$ while the exposure levels are changed from x to x^*, i.e., $Y(x, M(x)) - Y(x^*, M(x))$. Thus, we change BMI from x to x^*, while we hold every person's risk of hypercholesterolemia constant at the value that would have been observed had they had BMI value x. For the NIE, the mediator value is changed from $M(x)$ to $M(x^*)$ while the exposure is held constant at level x^*, i.e. $Y(x^*, M(x)) - Y(x^*, M(x^*))$. Thus, we hold BMI constant at level x^* while we change the person's risk of hypercholesterolemia under BMI value x to the person's risk of hypercholesterolemia under BMI value x^*. The total effect (TE) is defined as the effect of changing the exposure level from x to x^* and the mediator value from $M(x)$ to $M(x^*)$, i.e., $Y(x, M(x)) - Y(x^*, M(x^*))$. Thus, we change

BMI from x to x^* and the person's risk of hypercholesterolemia under BMI value x to the person's risk of hypercholesterolemia under BMI value x^*.

As mentioned before, causal mediation effects are estimated on the population-average level. The notation of these population-average causal effects is slightly different. For example, the population-average NIE is defined as $E[Y(x^*, M(x)) - Y(x^*, M(x^*))]$ (8).

For the NDE, NIE and TE to have a causal interpretation, it is necessary to control for any confounding variables. There are four no confounding assumptions that must be met (21):

1. No unmeasured confounding of the exposure-outcome relation.
2. No unmeasured confounding of the mediator-outcome relation.
3. No unmeasured confounding of the exposure-mediator relation.
4. No mediator-outcome confounders that are affected by the exposure.

Failing to adjust for confounding variables might result in bias, which means that the effects will not have a causal interpretation (21). Because these assumptions cannot be tested statistically, directed acyclic graphs (DAGs) may be used to determine the set of confounders that needs to be adjusted for (19).

Regression

The regression-based approach relies on two regression equations: a logistic regression model that relates the exposure to the mediator (equation 1), and a linear regression model that relates the exposure and the mediator to the outcome (equation 2):

$$logit(Pr(M = 1|X)) = i_1 + aX \quad (1)$$

$$E[Y|X, M] = i_2 + c'X + bM \quad (2)$$

where i_1 and i_2 represent the intercepts. The a coefficient in equation 1 represents the exposure-mediator effect, whereas the b coefficient in equation 2 represents the mediator-outcome effect *adjusted for the exposure*. The c' coefficient in equation 2 represents the exposure-outcome effect *adjusted for the mediator*.

For the regression-based approach, the regression coefficients from equation 1 and 2 are used to estimate the NDE, NIE and TE. The NDE, NIE and TE can be calculated using the equations below (21, 22):

$$NDE = c' * (x - x^*) \tag{3}$$

$$NIE = b * \left\{ \frac{\exp(i_1 + a * x)}{1 + \exp(i_1 + a * x)} - \frac{\exp(i_1 + a * x^*)}{1 + \exp(i_1 + a * x^*)} \right\} \tag{4}$$

$$TE = b * \left\{ \frac{\exp(i_1 + a * x)}{1 + \exp(i_1 + a * x)} - \frac{\exp(i_1 + a * x^*)}{1 + \exp(i_1 + a * x^*)} \right\} + c' \tag{5}$$

For the NDE, the exposure-outcome effect adjusted for the mediator (c') is multiplied by the causal contrast. This means that if the difference between x and x^* equals 1, then the NDE simplifies to c' (22). For the NIE, the term in between the accolades represents the difference in the risk of obtaining the mediator when changing from exposure level x to x^* (22). Thus, the risk difference depends on the exposure values chosen for x and x^*, i.e., on the selected causal contrast. Figure 1, based on the empirical data example, illustrates how the risk of hypercholesterolemia may depend on BMI values, and thus how the risk difference will depend on the chosen causal contrast. For example, the risk difference is 0.0348 for BMI values of 20 and 21, and 0.0484 for BMI values of 25 and 26. Since the NIE is estimated as the product of the risk difference based on the causal contrast for BMI and the b coefficient, the NIE will differ in magnitude across the chosen BMI values too. As can be seen from equation 3, the NDE is not dependent on x and x^* and therefore has the same magnitude across different causal contrasts. Since the TE is the sum of the NIE and the NDE, the TE also differs in magnitude across causal contrasts.

Simulation

The simulation-based approach consists of three steps. First, a large number of bootstrap samples are created. Next, for each bootstrap sample, the following steps are repeated:

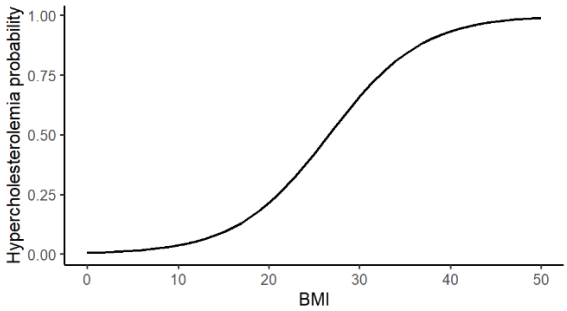

Figure 1 Risk of hypercholesterolemia for different BMI values

equations 1 and 2 are estimated, predicted values are calculated for $M(x)$ and $M(x^*)$ and each of the four potential outcomes for each observation, and subsequently the causal mediation effects (i.e., the NDE, NIE and TE) are calculated for each observation. Third, the average NDE, NIE and TE are estimated (9). Because the predicted values for the mediator depend on the values chosen for x and x^*, the NIE and TE depend on the selected causal contrast. In this respect, the regression- and simulation-based approaches correspond to each other.

In our running example, the risk of hypercholesterolemia for each individual is predicted twice: once based on BMI value x and once based on BMI value x^*. As shown in Figure 1, the magnitude of these risks depends on the values chosen for x and x^*. Next, the nested potential outcomes are simulated based on BMI values x and x^* and the potential mediator values. This way, the chosen causal contrast affects the size of the potential outcomes and thus the size of the NIE and TE estimates. As a result, the effects only apply to the two values selected for the causal contrast. The NDE is estimated as the difference between $Y\big(x, M(x)\big)$ and $Y(x^*, M(x))$, for which the mediator is held constant at the value $M(x)$ in the potential outcomes. Therefore, the causal contrast does not affect the magnitude of the NDE.

Imputation

The imputation-based approach uses a natural effect model to estimate the direct and indirect effects. Natural effect models were proposed by Lange et al. (23) and Vansteelandt et al. (24), and allow for the natural direct and indirect effect to be modelled simultaneously. A natural effect model has the form

$$E[Y\big(x, M(x^*)\big)] = i_3 + \beta_1 x + \beta_2 x^* \tag{6}$$

where β_1 represents the natural direct effect and β_2 represents the natural indirect effect, both corresponding to a one-unit increase in the exposure.

Because for each individual only one exposure value is observed (we denote this value as x^*), the data has to be expanded to include the unobserved exposure value (we denote this value as x). In our data example, if we treat BMI as a binary variable, this is easy: for overweight individuals x^* equals 1 (overweight) whereas x equals 0 (healthy weight). For individuals with a healthy weight, x^* equals 0 and x equals 1. However, if we treat weight as a continuous variable, x^* represents the observed BMI value but there is no clear value

for x. The value for x is then determined based on a set of random draws from the distribution of BMI values for each individuals (23, 25). In the imputation-based approach, the unobserved potential outcomes are treated as missing values, i.e., after the data is expanded the unobserved potential outcomes are imputed using the outcome model (equation 2) (25).

Finally, estimates of the natural direct and indirect effect can be obtained upon fitting a natural effect model (equation 6) to the imputed dataset. Because the potential outcomes are imputed directly based on equation 2, the potential outcomes are not dependent on the risk of the mediator under specific exposure values. Therefore, the indirect and total effect estimates based on the imputation-based approach are not dependent on the chosen causal contrast. That is, the size of the indirect and total effect estimates are the same for each one unit increase in the exposure. Thus, the imputation-based approach provides one overall indirect effect estimate and one overall total effect estimate that applies to every one unit increase in the exposure.

Weighting

The weighting-based approach generally follows the same steps as the imputation-based approach. However, instead of imputing the unobserved potential outcome values, the weighting-based approach weighs the observations in the expanded dataset. The weight is calculated as the probability to observe that particular value of the mediator for unobserved exposure value x divided by the probability to observe that particular value of the mediator for the observed exposure value x^* (equation 7) (3). The probabilities are predicted based on equation 1 (23).

For an overweight individual that suffers from hypercholesterolemia, weight w would be calculated as the probability of hypercholesterolemia if that individual had had a healthy weight divided by the probability of hypercholesterolemia for the actual weight status of that individual. This way, individuals whose observed mediator value is more typical for the unobserved exposure value x are up-weighted, whereas individuals whose observed mediator value is less typical for the unobserved exposure value x are down-weighted (25).

$$w = \frac{\Pr(M = 1|X = x)}{\Pr(M = 1|X = x^*)} \tag{7}$$

Finally, the natural effects model is estimated by regressing the outcome on the exposure values x and x^*, weighting each observation based on the created weights.

Although the weighting-based approach does estimate the mediation model, in contrast to the regression- and simulation-based approaches it is not used to estimate mediator values. Instead, it is used to construct the weights that indicate the probability that the observed mediator value is observed under the unobserved and observed exposure levels. Therefore, the potential outcomes are not dependent on the risk of the mediator under specific exposure values. This means that the causal contrast does not influence the magnitude of the effect estimates in the weighting-based approach, and the effect estimates represent overall effects.

Simulation study

Although in theory the regression- and simulation-based approaches aim to estimate the same indirect and total effect estimates, in practice researchers may observe differences in the effect estimates from these two methods. The same holds for the imputation- and weighting-based approaches. A Monte Carlo simulation study was designed to investigate the performance of the regression-, simulation-, imputation- and weighting-based approaches.

Simulation methods

The R programming language version 4.1.0 (26) was used to generate and analyze the data. To estimate the mediation effects, R packages *regmedint, mediation* and *medflex* were used (25, 27, 28). The continuous exposure and binary mediator were generated from a standard normal distribution and a binomial distribution conditional on the exposure, respectively. The continuous outcome was a function of the exposure and the mediator. Sample sizes were 200, 500, 1000 and 1500. Medium effect sizes were generated for all pathways in the mediation model: the exposure-mediator effect was set to 0.92, and the mediator-outcome effect and the exposure-outcome effect were set to 0.39 (29). Subsequently, we estimated the indirect, direct and total effect based on the simulated data using the regression-, simulation-, imputation- and weighting-based approaches. For the regression- and simulation-based approaches, the effects were estimated using three causal contrasts: $x^* = 0 \,\&\, x = 1$, $x^* = 1 \,\&\, x = 2$ and $x^* = 2 \,\&\, x = 3$.

The performance of each approach was compared based on the relative bias (RB) and mean squared error (MSE). RB was calculated by subtracting the true value from the estimated indirect, direct and total effect estimates, and then dividing this by the respective true value (30, 31). A value of less than 0.1 across replications is generally considered acceptable (32). MSE was calculated as the average squared difference between the effect estimates and the true values (30, 31). We treated the empirical true values based on a sample size of 500,000 as the true values. For all performance measures, a lower score corresponds to a better performance. The simulation code and the empirical true values are available in Appendix A and Appendix B, respectively.

Simulation results

Figure 2 shows the relative bias (panel A) and mean squared error (panel B) of the indirect effect estimated with the regression- and simulation-based approaches for the causal contrast $x^* = 0 \,\&\, x = 1$. Both in terms of RB and MSE, the regression-based approach performed slightly better than the simulation-based approach. Furthermore, as sample size increased, the RB approached 0.1, indicating an acceptable level of RB for both approaches. These patterns were also observed for the other causal contrasts and for the total effect estimates. The estimates of the direct effect were the same for the regression- and simulation-based approaches across all sample sizes, resulting in identical RB and MSE for these two approaches. As with the indirect effect, bias of the direct effect decreased as sample size increased. Figures of RB and MSE of the indirect, direct and total effects for all causal contrasts are available in Appendix C.

The mean estimated indirect, direct and total effects from the regression- and simulation-based approaches based on a sample size of 1,000 are available in Appendix D. For both

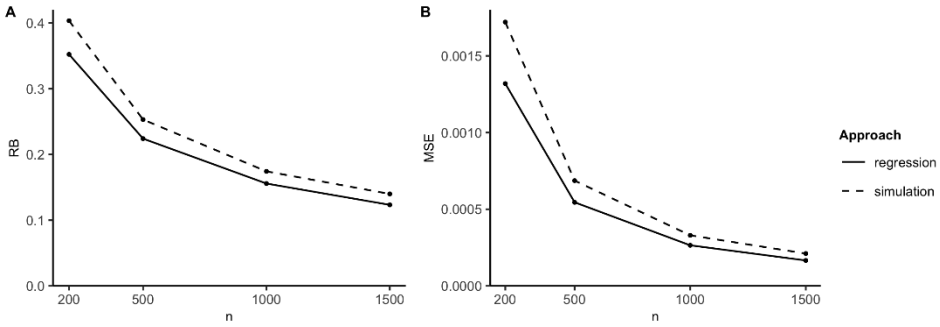

Figure 2 Relative bias (panel A) and mean squared error (panel B) of the indirect effect estimated with the regression- and simulation-based approaches for the causal contrast $x^* = 0 \,\&\, x = 1$

approaches, the indirect effect and total effect estimates depended on the chosen causal contrast, and the estimates only apply to the difference between the two exposure values selected for the causal contrast. For the regression-based approach, the indirect effect estimate for a causal contrast of $x^* = 0 \,\&\, x = 1$ was 0.08377, whereas the indirect effect estimate for a causal contrast of $x^* = 2 \,\&\, x = 3$ was 0.03014. For the simulation-based approach, the indirect effect estimate for a causal contrast of of $x^* = 0 \,\&\, x = 1$ was 0.08355, whereas the indirect effect estimate for a causal contrast of $x^* = 2 \,\&\, x = 3$ was 0.03033. The direct effects were the same across the chosen causal contrasts. Similar patterns were observed for sample sizes 200, 500 and 1500. This shows the impact of the chosen causal contrast on the effect estimates in causal mediation analysis with a binary mediator for the regression- and simulation-based approaches.

Figure 3 shows the relative bias (panel A) and mean squared error (panel B) of the indirect effect estimated with the imputation- and weighting-based approaches. Here too, RB and MSE followed identical patterns: for both approaches, RB and MSE decreased as sample size increased, and RB approaches 0.1. For each sample size, the estimate of the direct effect estimated with the imputation-based approach was identical to that of the regression- and simulation-based approaches. The RB and MSE of the direct effect estimated with the weighting-based approach were greater than those estimated with the other approaches. As sample size increased, bias in the direct effect estimates decreased. Figures of RB and MSE of the direct and total effect are available in Appendix E.

Thus, the effect estimates based on all estimation approaches were affected by finite sample bias: if sample size increased, bias decreased. In addition, the differences

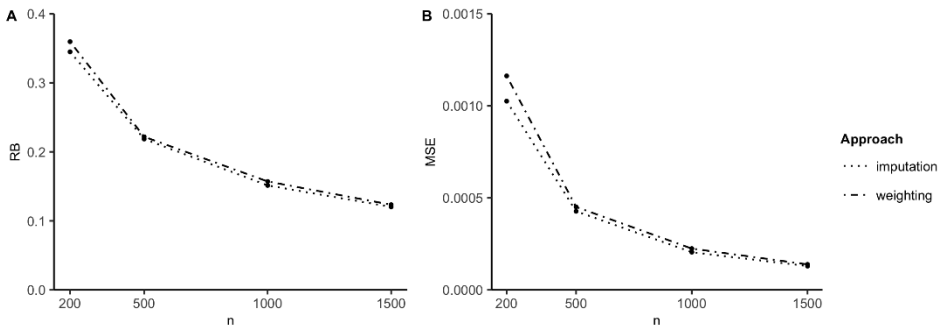

Figure 3 Relative bias (panel A) and mean squared error (panel B) of the indirect effect estimated with the imputation- and weighting-based approaches

between the regression- and simulation-based approach, and between the imputation- and weighting-based approach in terms of RB and MSE decreased when sample size increased.

Empirical data example

We use two empirical data examples from the Amsterdam Growth and Health Longitudinal Study (AGHLS) to illustrate the impact of the different estimation approaches and the causal contrast on the interpretation of the effect estimates for models with a binary mediator. The AGHLS is an ongoing cohort study that assesses the relations between the development of anthropometry, lifestyle and health from adolescence into adulthood (33). We used data collected in 2000, when the participants were in their late 30s. Only subjects with complete data on all variables in the mediation model were considered in the analysis (n = 378).

In both examples, we looked at BMI as the exposure, diastolic blood pressure as the continuous outcome and hypercholesterolemia as the binary mediator. The cut-off for hypercholesterolemia was based on guidelines from the U.S. National Institutes of Health (34). In example 1, we treated BMI as a binary variable, with a BMI ≥ 25 indicating overweight. In example 2, we treated BMI as a continuous exposure. In both examples, we estimated the indirect, direct and total effects using the regression-, simulation-, imputation- and weighting-based approaches.

In example 1, there are only two exposure values of substantive interest: 0, which denotes the subjects with healthy weight, and 1, which denotes the subjects with overweight. Therefore, the only causal contrast that we used for the estimation of the causal mediation effects using the regression-based and simulation-based approaches was $x^* = 0$ and $x = 1$. Table 1 shows the effect estimates with corresponding 95% percentile bootstrap confidence intervals for both examples. For example 1, the effect estimates from the regression- and imputation-based approaches were the same, and can be interpreted as: people who are overweight on average have a 0.926 mmHg higher diastolic blood pressure than people who are not overweight through an increase in the risk of hypercholesterolemia. The indirect effect estimates from the simulation- and weighting-based approaches were slightly larger and smaller, respectively, but have the same general interpretation as the indirect effect estimates based on the regression- and imputation-based approaches.

Table 1 Causal effect estimates for the association between BMI and diastolic blood pressure, mediated by hypercholesterolemia

	Indirect effect (95% CI)	Direct effect (95% CI)	Total effect (95% CI)
Binary exposure			
Regression	0.926 (0.311; 1.727)	4.114 (1.889; 6.201)	5.040 (2.855; 7.152)
Simulation	0.995 (0.273; 1.770)	4.114 (1.928; 6.190)	5.109 (2.889; 7.130)
Imputation	0.926 (0.273; 1.617)	4.114 (2.061; 6.251)	5.040 (3.010; 7.192)
Weighting	0.893 (0.278; 1.549)	4.239 (2.245; 6.312)	5.132 (3.143; 7.241)
Continuous exposure			
Regression			
$x^* = 0$ & $x = 1$	0.003 (0.000; 0.013)	1.152 (0.811; 1.456)	1.155 (0.816; 1.458)
$x^* = 20$ & $x = 21$	0.092 (0.017; 0.173)	1.152 (0.811; 1.456)	1.244 (0.913; 1.548)
$x^* = 25$ & $x = 26$	0.128 (0.024; 0.259)	1.152 (0.811; 1.456)	1.280 (0.942; 1.607)
Simulation			
$x^* = 0$ & $x = 1$	0.014 (-0.032; 0.050)	1.152 (0.815; 1.450)	1.166 (0.821; 1.470)
$x^* = 20$ & $x = 21$	0.077 (-0.069; 0.320)	1.152 (0.815; 1.450)	1.229 (0.895; 1.590)
$x^* = 25$ & $x = 26$	0.112 (-0.063; 0.420)	1.152 (0.815; 1.450)	1.264 (0.911; 1.680)
Imputation	0.115 (0.015; 0.219)	1.152 (0.844; 1.474)	1.267 (0.962; 1.591)
Weighting	0.119 (0.018; 0.223)	1.092 (0.788; 1.394)	1.210 (0.907; 1.516)

Abbreviations: CI: confidence interval

In example 2, we treated BMI as a continuous variable. Since the continuous BMI variable can take on various values, multiple causal contrasts can be defined for the regression- and simulation-based approaches. In our example, we estimated the causal mediation effects based on three different causal contrasts. First, we estimated the effects based on $x^* = 0$ and $x = 1$, corresponding to BMI values of 0 and 1, respectively. However, note that BMI values of 0 and 1 are clinically impossible. We included this contrast to demonstrate what might happen when researchers fail to specify a meaningful causal contrast in available causal mediation software programs, as 0 and 1 are default values in most programs (22, 28, 35). In addition, we estimated the effects based on two clinically relevant causal contrasts: $x^* = 20$ and $x = 21$, and $x^* = 25$ and $x = 26$. The effect estimates are estimated under both causal contrasts to demonstrate that the magnitude of the effect estimates differs across the chosen causal contrast.

In example 2, the indirect and total effects estimated with the regression- and simulation-based approaches were dependent on the chosen causal contrast. For example, the indirect effect estimated by the regression-based approach for the causal contrast of

$x^* = 20$ and $x = 21$ was 0.092, indicating that people with a BMI of 21 on average have a 0.092 mmHg higher diastolic blood pressure compared to people with a BMI of 20 through an increase in the risk of hypercholesterolemia. In comparison, the indirect effect estimated for the causal contrast of $x^* = 25$ and $x = 26$ was 0.128, indicating that people with a BMI of 26 on average have a 0.128 mmHg higher diastolic blood pressure compared to people with a BMI of 25 through an increased risk of hypercholesterolemia. Because the total effect is equal to the sum of the indirect and direct effect, the total effect estimates from the regression- and simulation-based approaches also differed in magnitude across the different causal contrasts. The indirect effect estimates based on the clinically implausible causal contrast of $x^* = 0$ and $x = 1$ was 0.003. In contrast with the indirect effect based on the clinically meaningful causal contrasts, this indirect effect estimate indicates that hypercholesterolemia is not a mediator of the relation between overweight and diastolic blood pressure. The indirect and total effect estimates based on the imputation- and weighting-based approaches did not depend on a specific causal contrast. Therefore these indirect effect estimates can be interpreted as the overall difference in diastolic blood pressure in mmHg for every one unit increase in BMI through an increased risk of hypercholesterolemia.

Discussion

This paper demonstrated that four commonly-used causal estimation approaches provide different effect estimates with different interpretations for mediation models with binary mediators. We focused on four approaches: regression, simulation, imputation and weighting. The regression- and simulation-based approaches require the selection of a causal contrast, whereas this is not required for the imputation- and weighting-based approaches. In our empirical data example we used two scenarios: a scenario in which the exposure and the mediator were both binary (example 1), and a scenario in which the exposure was continuous and the mediator was binary (example 2).

If the exposure is binary, then only two exposure values are of interest (i.e., 0 and 1). As a result, the regression- and simulation-based approaches provide the same effect estimates as the imputation- and weighting-based approaches. The indirect effect estimates across the four approaches can all be interpreted the same way, namely as the *average* difference in the outcome between the exposed and the unexposed individuals through an increase or decrease in the risk of the mediator. Thus, if both the exposure and mediator are binary, then researchers can choose between all four approaches.

If the exposure is continuous, then the magnitude of the effect estimates from the regression- and simulation-based approaches depend on the chosen causal contrast. For the imputation- and weighting-based approaches, no causal contrast has to be selected (25). These differences are also reflected in the interpretation of the results: the indirect effects from the regression- and simulation-based approaches only apply to the two values selected for the causal contrast, whereas the imputation- and weighting-based approaches return overall differences in the outcome for *every one unit difference* in the exposure through the mediator. Our empirical data example showed that the indirect effects estimated by the regression- and simulation-based approaches increased as the values selected for the causal contrast increased, and thus that the causal contrast should be based on substantive knowledge. Throughout this paper, we focused on a causal contrast that equals one, i.e., a one-unit difference between x and x^*. However, in practice, sometimes a larger causal contrast may be more interesting to investigate. In all approaches, x and x^* can be set to specific exposure values so that the causal contrast is larger than one. The difference in the interpretation of the indirect and total effect estimates between the regression- and simulation-based approaches and the imputation- and weighting-based approaches remains in this situation. Moreover, the models considered in this paper are relatively simple, i.e., they do not contain exposure-mediator interactions. However, these can be added easily in most software programs commonly used by epidemiologists and do not change findings. To our knowledge, this is the first study that studied patterns of finite sample bias for four commonly used estimation approaches for mediation models with a binary mediator and a continuous outcome. A previous simulation study reported that the imputation-based approach generally provides more precise estimates than the weighting-based approach (24). This is in line with the general finding that weighting-based methods, such as inverse probability weighting, are affected by finite sample bias, as the performance of these methods can be affected by extreme weights (36, 37). Our simulation study showed that all approaches were affected by finite sample bias to some extent, meaning that bias decreases if sample size increases. We also showed that the differences between the regression- and simulation-based approach, and between the imputation- and weighting-based approach decrease when sample size increases.

Traditional mediation analysis

Although causal mediation analysis clarifies causal effect estimation for mediation models, traditional mediation analysis is still most often used (11, 12). In traditional mediation analysis, the mediation effects are defined and estimated based on the effects

from three equations: the a coefficient from equation 1, the b and c' coefficients from equation 2, and the total exposure-outcome effect estimated based on a model that relates the exposure to the outcome (typically referred to as the c coefficient) (2). The direct effect is defined and estimated as the c' coefficient from equation 2. The indirect effect is defined and estimated using the product-of-coefficients method (i.e., $a * b$) or the difference-in-coefficients method (i.e., $c - c'$).

If the outcome and mediator are continuous, then the product-of-coefficients and difference-in-coefficients methods provide the same indirect effect estimates and equal the natural indirect effect (15, 16). However, this is not necessarily the case if the mediator is a binary variable. If the mediator is binary and equation 2 is estimated based on logistic regression, then the exposure-mediator effect is estimated on the log-odds scale and ranges from negative infinity to positive infinity. For the mediator-outcome effect, the mediator is dichotomous and ranges from 0 to 1. Multiplying these effects results in a mismatch of the scale on which the indirect effect is estimated (18, 38). Therefore, if the mediator is a binary variable and equation 2 is estimated with logistic regression, then the product-of-coefficients method should not be used to estimate the indirect effect. This mismatch in scales does not occur with the difference-in-coefficients method, as both the c and c' coefficients are estimated using a linear regression model. Therefore, the traditional difference-in-coefficients method provides indirect effect estimates similar to the imputation- and weighting-based approaches for mediation model with a binary mediator and a continuous outcome. Thus, for these models, the causal imputation- and weighting-based approaches as well as the traditional difference-in-coefficients method can be used to estimate the indirect effect.

If both the mediator and the outcome are binary, neither the product-of-coefficients method nor the difference-in-coefficients method provide estimates of the causal indirect effect (39). The indirect effect estimates based on the difference-in-coefficients method will be biased by noncollapsibility when the outcome is binary (40). This noncollapsibility effect stems from a change in scales that occurs in logistic regression when variables are added to the model (41). As a result, the difference between a univariable- (the c coefficient) and multivariable (the c' coefficient) exposure effect estimate not only represent the difference in coefficients but also a noncollapsibility effect. Therefore, for models with a binary mediator and a binary outcome, the causal estimation methods are preferred over traditional mediation analysis.

Recommendations for practice

Because the different causal estimation approaches return causal effect estimates with different interpretations if the exposure is continuous and the mediator is binary, researchers should inform their choice for an estimation approach based on whether they are interested in overall effects or in effects that correspond to specific causal contrasts. The imputation- and weighting-based approaches can be used to estimate overall causal mediation effects (25). In our running example, these approaches can be used to answer the question 'how does lowering the BMI of all individuals in a population by one point affect blood pressure overall?'. The regression- and simulation-based approaches can be used to estimate causal mediation effects that correspond to a specific causal contrast (28). In our running example, these approaches can be used to answer the question 'how does lowering the BMI of all individuals with a certain BMI, e.g., 25, by one point affect blood pressure on average?'. In practice, there are not many situations in which researchers are interested in a specific causal contrast, so often researchers will be forced to make arbitrary decisions on the causal contrast (25, 42). In most situations, the imputation- and weighting-based approaches will be more suited, as these provide mediation effect estimates with interpretations that align with the effects of interest in most studies. Furthermore, since the traditional difference-in-coefficients method provides the same indirect effect estimates as the imputation- and weighting-based approaches for mediation models with a binary mediator and a continuous outcome, it is also possible to estimate the average indirect effect using the traditional difference-in-coefficients method. Nevertheless, if one is interested in estimating effects that correspond to a specific causal contrast, then it is important to be aware that the default causal contrast in most software programs is $\boldsymbol{x^* = 0}$ and $\boldsymbol{x = 1}$. Failing to select the right causal contrast may lead to wrong conclusions regarding the presence of a mediated effect. If the exposure is binary, then the effect estimates from all approaches can be interpreted as the average difference in the outcome between the two compared groups.

Causal mediation analysis is implemented in most software programs commonly used by epidemiologists (i.e., SPSS, R, Stata, Mplus and SAS). A detailed overview of the implementation of the different estimation methods in these software programs (including software code) is provided elsewhere (3).

Conclusion

For mediation models with a binary mediator, causal estimation approaches provide different effect estimates with different interpretations. The regression- and simulation-based approaches require the selection of a causal contrast and result in effects that correspond to those specific exposure values, whereas the imputation- and weighting-based approaches result in overall causal mediation effects. For mediation models with a binary mediator and a continuous outcome, the traditional difference-in-coefficients method provides the same indirect effect estimate as the imputation- and weighting-based approaches. It is recommended that researchers inform their choice for an estimation method based on the type of effect that they are interested in.

References

1. Hernan MA, Robins JM. Causal Inference: What If. Boca Raton: Chapman & Hall/CRC; 2020.
2. MacKinnon DP. Introduction to Statistical Mediation Analysis. New York: Erlbaum; 2021.
3. Valente MJ, Rijnhart JJM, Smyth HL, Muniz FB, MacKinnon DP. Causal Mediation Programs in R, Mplus, SAS, SPSS, and Stata. Structural Equation Modeling: A Multidisciplinary Journal. 2020;27(6):975-84.
4. Pearl J. The Causal Mediation Formula—A Guide to the Assessment of Pathways and Mechanisms. Prevention Science. 2012;13(4):426-36.
5. Holland PW. Statistics and Causal Inference. Journal of the American Statistical Association. 1986;81(396):945-60.
6. Rubin DB. Estimating causal effects of treatments in randomized and nonrandomized studies. Journal of Educational Psychology. 1974;66(5):688-701.
7. Holland PW. Causal inference, path analysis and recursive structural equations models. ETS Research Report Series. 1988;1988(1):i-50.
8. Pearl J. Direct and indirect effects. Proceedings of the Seventeenth conference on Uncertainty in artificial intelligence; Seattle, Washington: Morgan Kaufmann Publishers Inc.; 2001. p. 411–20.
9. Imai K, Keele L, Tingley D. A general approach to causal mediation analysis. Psychological Methods. 2010;15(4):309-34.
10. Muthén BO, Muthén LK, Asparouhov T. Regression and Mediation Analysis using Mplus. Los Angeles, CA: Muthén & Muthén; 2017.
11. Vo T-T, Superchi C, Boutron I, Vansteelandt S. The conduct and reporting of mediation analysis in recently published randomized controlled trials: results from a methodological systematic review. Journal of Clinical Epidemiology. 2020;117:78-88.
12. Rijnhart JJM, Lamp SJ, Valente MJ, MacKinnon DP, Twisk JWR, Heymans MW. Mediation analysis methods used in observational research: a scoping review and recommendations. BMC Medical Research Methodology. 2021;21(1):226.
13. Rizzo RRN, Cashin AG, Bagg MK, Gustin SM, Lee H, McAuley JH. A Systematic Review of the Reporting Quality of Observational Studies That Use Mediation Analyses. Prevention Science. 2022.
14. Lipkovich I, Ratitch B, Mallinckrodt CH. Causal Inference and Estimands in Clinical Trials. Statistics in Biopharmaceutical Research. 2020;12(1):54-67.

15. MacKinnon DP, Valente MJ, Gonzalez O. The Correspondence Between Causal and Traditional Mediation Analysis: the Link Is the Mediator by Treatment Interaction. Prevention Science. 2020;21(2):147-57.
16. Rijnhart JJM, Twisk JWR, Chinapaw MJM, de Boer MR, Heymans MW. Comparison of methods for the analysis of relatively simple mediation models. Contemporary Clinical Trials Communications. 2017;7:130-5.
17. Rijnhart JJM, Valente MJ, MacKinnon DP, Twisk JWR, Heymans MW. The Use of Traditional and Causal Estimators for Mediation Models with a Binary Outcome and Exposure-Mediator Interaction. Structural Equation Modeling: A Multidisciplinary Journal. 2021;28(3):345-55.
18. Rijnhart JJM, Valente MJ, Smyth HL, MacKinnon DP. Statistical Mediation Analysis for Models with a Binary Mediator and a Binary Outcome: the Differences Between Causal and Traditional Mediation Analysis. Prevention Science. 2021.
19. Robins JM. Semantics of causal DAG models and the identification of direct and indirect effects. Oxford Statistical Science Series. 2003:70-82.
20. Robins JM, Greenland S. Identifiability and Exchangeability for Direct and Indirect Effects. Epidemiology. 1992;3(2):143-55.
21. VanderWeele TJ. Explanation in Causal Inference: Methods for Mediation and Interaction: Oxford University Press; 2015.
22. Valeri L, VanderWeele TJ. Mediation Analysis Allowing for Exposure–Mediator Interactions and Causal Interpretation: Theoretical Assumptions and Implementation With SAS and SPSS Macros. Psychological Methods. 2013;18(2):137-50.
23. Lange T, Vansteelandt S, Bekaert M. A Simple Unified Approach for Estimating Natural Direct and Indirect Effects. American Journal of Epidemiology. 2012;176(3):190-5.
24. Vansteelandt S, Bekaert M, Lange T. Imputation Strategies for the Estimation of Natural Direct and Indirect Effects. Epidemiologic Methods. 2012;1(1):131-58.
25. Steen J, Loeys T, Moerkerke B, Vansteelandt S. medflex: An R Package for Flexible Mediation Analysis using Natural Effect Models. Journal of Statistical Software. 2017;76(11).
26. R Core Team. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2019.
27. Kazuki Y, Yi L. regmedint: Regression-Based Causal Mediation Analysis with an Interaction Term 2020 [Available from: https://CRAN.R-project.org/package=regmedint.

28. Tingley D, Yamamoto T, Hirose K, Keele L, Imai K. mediation: R Package for Causal Mediation Analysis. Journal of Statistical Software. 2014;59(5):1-38.
29. Cohen J. Statistical power analysis for the behavioral sciences. Hillsdale, NJ: Lawrence Erlbaum Associates; 1988.
30. Burton A, Altman DG, Royston P, Holder RL. The design of simulation studies in medical statistics. Statistics in Medicine. 2006;25(24):4279-92.
31. Morris TP, White IR, Crowther MJ. Using simulation studies to evaluate statistical methods. Statistics in Medicine. 2019;38(11):2074-102.
32. Flora DB, Curran PJ. An Empirical Evaluation of Alternative Methods of Estimation for Confirmatory Factor Analysis With Ordinal Data. Psychological Methods. 2004;9(4):466-91.
33. Wijnstok NJ, Hoekstra T, van Mechelen W, Kemper HCG, Twisk JWR. Cohort Profile: The Amsterdam Growth and Health Longitudinal Study. International Journal of Epidemiology. 2013;42(2):422-9.
34. National Institutes of Health's U.S. National Library of Medicine. Cholesterol levels: what do the results mean 2020 [Available from: https://medlineplus.gov/lab-tests/cholesterol-levels/.
35. Emsley R, Liu H. PARAMED: Stata module to perform causal mediation analysis using parametric regression models. 2013.
36. Austin PC. An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies. Multivariate behavioral research. 2011;46(3):399-424.
37. Cole SR, Hernán MA. Constructing Inverse Probability Weights for Marginal Structural Models. American Journal of Epidemiology. 2008;168(6):656-64.
38. Li Y, Schneider JA, Bennett DA. Estimation of the mediation effect with a binary mediator. Statistics in Medicine. 2007;26(18):3398-414.
39. Rijnhart JJM, Twisk JWR, Eekhout I, Heymans MW. Comparison of logistic-regression based methods for simple mediation analysis with a dichotomous outcome variable. BMC Medical Research Methodology. 2019;19(1):19.
40. Mackinnon DP, Dwyer JH. Estimating Mediated Effects in Prevention Studies. Evaluation Review. 1993;17(2):144-58.
41. Schuster NA, Twisk JWR, ter Riet G, Heymans MW, Rijnhart JJM. Noncollapsibility and its role in quantifying confounding bias in logistic regression. BMC Medical Research Methodology. 2021;21(1):136.

42. Loeys T, Moerkerke B, De Smet O, Buysse A, Steen J, Vansteelandt S. Flexible mediation analysis in the presence of nonlinear relations: Beyond the mediation formula. Multivariate Behavioral Research. 2013;48(6):871-94.

SUPPLEMENTARY MATERIALS FOR CHAPTER 5

Appendix A Simulation code

Functions

```
library(regmedint)
library(mediation)
library(medflex)

generate_data <- function(reps, n, im, iy, xm, my, xy){
  ID <- seq(1:n)
  obs <- n * reps
  X <- rnorm(n = obs, mean = 0, sd = 1)
  lpm <- im + xm * X
  prm <- 1/(1 + exp(-lpm))
  M <- rbinom(n = obs, size = 1, prob = prm)
  Y <- iy + xy * X + my * M + rnorm(n = obs)

  return(list("ID" = rep(ID, reps),
              "n" = rep(n, obs),
              "repnr" = rep(seq(1:reps), each = n),
              "im" = rep(im, obs),
              "iy" = rep(iy, obs),
              "xm" = rep(xm, obs),
              "my" = rep(my, obs),
              "xy" = rep(xy, obs),
              "X" = X,
              "M" = M,
              "Y" = Y))
}

store_effects <- function(rows){
  df <- as.data.frame(matrix(NA, nrow = rows, ncol = 11))
  colnames(df) <- c("n",
                    "repnr",
                    "xm",
                    "my",
                    "xy",
                    "approach",
                    "x0",
                    "x1",
                    "indirect",
                    "direct",
                    "total")

  return(df)
}

regression_based <- function(df, x0, x1){
  effects <- regmedint(data = df,
                       yvar = "Y",
```

```
                        avar = "X",
                        mvar = "M",
                        cvar = NULL,
                        a0 = x0,
                        a1 = x1,
                        m_cde = 0,
                        c_cond = NULL,
                        mreg = "logistic",
                        yreg = "linear",
                        interaction = FALSE)

  indirect <- coef(summary(effects))[3]
  direct <- coef(summary(effects))[1]
  total <- coef(summary(effects))[6]

  return(list("n" = unique(df$n),
              "repnr" = unique(df$repnr),
              "xm" = unique(df$xm),
              "my" = unique(df$my),
              "xy" = unique(df$xy),
              "approach" = "regression",
              "x0" = x0,
              "x1" = x1,
              "indirect" = indirect,
              "direct" = direct,
              "total" = total))
}

simulation_based <- function(df, x0, x1){
  mediator_model <- glm(M ~ X, family = binomial, data = df)
  outcome_model <- glm(Y ~ X + M, data = df)
  effects <- mediate(mediator_model, outcome_model,
                     sims = 1000,
                     boot = TRUE,
                     treat = "X",
                     mediator = "M",
                     boot.ci.type = "perc",
                     treat.value = x1,
                     control.value = x0)
  indirect <- summary(effects)$d.avg
  direct <- summary(effects)$z.avg
  total <- summary(effects)$tau.coef

  return(list("n" = unique(df$n),
              "repnr" = unique(df$repnr),
              "xm" = unique(df$xm),
              "my" = unique(df$my),
              "xy" = unique(df$xy),
              "approach" = "simulation",
              "x0" = x0,
              "x1" = x1,
```

```
              "indirect" = indirect,
              "direct" = direct,
              "total" = total))
}

imputation_based <- function(df){
  outcome_model <- glm(Y ~ X + M, data = df)
  expData <- neImpute(outcome_model)
  model <- neModel(Y ~ X0 + X1, expData = expData)
  effects <- neEffdecomp(model)
  indirect <- coef(summary(effects))[2]
  direct <- coef(summary(effects))[1]
  total <- coef(summary(effects))[3]

  return(list("n" = unique(df$n),
              "repnr" = unique(df$repnr),
              "xm" = unique(df$xm),
              "my" = unique(df$my),
              "xy" = unique(df$xy),
              "approach" = "imputation",
              "x0" = 0, # causal contrast is 0 and 1 by default
              "x1" = 1,
              "indirect" = indirect,
              "direct" = direct,
              "total" = total))
}

weighting_based <- function(df){
  mediator_model <- glm(M ~ X, family = binomial, data = df)
  expData <- neWeight(mediator_model)
  model <- neModel(Y ~ X0 + X1, expData = expData)
  effects <- neEffdecomp(model)
  indirect <- coef(summary(effects))[2]
  direct <- coef(summary(effects))[1]
  total <- coef(summary(effects))[3]

  return(list("n" = unique(df$n),
              "repnr" = unique(df$repnr),
              "xm" = unique(df$xm),
              "my" = unique(df$my),
              "xy" = unique(df$xy),
              "approach" = "weighting",
              "x0" = 0, # causal contrast is 0 and 1 by default
              "x1" = 1,
              "indirect" = indirect,
              "direct" = direct,
              "total" = total))
}

store_performance <- function(rows){
  df <- as.data.frame(matrix(NA, nrow = rows, ncol = 16))
```

```
  colnames(df) <- c("n",
                    "xm",
                    "my",
                    "xy",
                    "approach",
                    "x0",
                    "x1",
                    "indirect_AB",
                    "indirect_RB",
                    "indirect_MSE",
                    "direct_AB",
                    "direct_RB",
                    "direct_MSE",
                    "total_AB",
                    "total_RB",
                    "total_MSE")

  return(df)
}

performance_measures <- function(df){
  AB <- function(estimate, true){
    return(abs(estimate - true))
  }
  RB <- function(estimate, true){
    return(abs(estimate - true)/true)
  }
  MSE <- function(estimate, true){
    return((estimate - true)^2)
  }

  return(list("n" = unique(df$n),
              "xm" = unique(df$xm),
              "my" = unique(df$my),
              "xy" = unique(df$xy),
              "approach" = unique(df$approach),
              "x0" = unique(df$x0),
              "x1" = unique(df$x1),
              "indirect_AB" = mean(AB(df$indirect,
               empirical_true$indirect)),
              "indirect_RB" = mean(RB(df$indirect,
               empirical_true$indirect)),
              "indirect_MSE" = mean(MSE(df$indirect,
               empirical_true$indirect)),
              "direct_AB" = mean(AB(df$direct, empirical_true$direct)),
              "direct_RB" = mean(RB(df$direct, empirical_true$direct)),
              "direct_MSE" = mean(MSE(df$direct,
               empirical_true$direct)),
              "total_AB" = mean(AB(df$total, empirical_true$total)),
              "total_RB" = mean(RB(df$total, empirical_true$total)),
              "total_MSE" = mean(MSE(df$total, empirical_true$total))))
```

```
}
```

Step 1 – Generate data for empirical true values

```
reps <- 1
n <- 500000
im <- 0
iy <- 0
xm <- 0.92
my <- 0.39
xy <- 0.39

df <- as.data.frame(generate_data(reps = reps,
                                  n = n,
                                  im = im,
                                  iy = iy,
                                  xm = xm,
                                  my = my,
                                  xy = xy))

save(df, file = "Data for empirical true values.RData")
```

Step 2 – Estimate empirical true values

```
load("Data for empirical true values.RData")

empirical_values <- store_effects(rows = 8)
empirical_values[1, ] <- regression_based(df = df, x0 = 0, x1 = 1)
empirical_values[2, ] <- regression_based(df = df, x0 = 1, x1 = 2)
empirical_values[3, ] <- regression_based(df = df, x0 = 2, x1 = 3)
empirical_values[4, ] <- simulation_based(df = df, x0 = 0, x1 = 1)
empirical_values[5, ] <- simulation_based(df = df, x0 = 1, x1 = 2)
empirical_values[6, ] <- simulation_based(df = df, x0 = 2, x1 = 3)
empirical_values[7, ] <- imputation_based(df = df)
empirical_values[8, ] <- weighting_based(df = df)

save(empirical_values, file = "Empirical true values.RData")
```

Step 3 – Generate actual data

```
reps <- 1000
n <- c(200, 500, 1000, 1500)
im <- 0
iy <- 0
xm <- 0.92
my <- 0.39
xy <- 0.39

for(i in n){
      df <- as.data.frame(generate_data(reps = reps,
                                        n = i,
                                        im = im,
```

```
                                        iy = iy,
                                        xm = xm,
                                        my = my,
                                        xy = xy))
      save(df, file = paste0("n = ", i, ".RData"))
}
```

Step 4 – Estimate effects

```
path <- "3. Actual data/"
files <- list.files(path = path,
                    pattern = "*.RData")
reps <- 1000

regression_effects <- store_effects(rows = 3 * reps)
for(i in files){
  load(paste0(path, i))
  CC01 <- 1
  CC12 <- 1001
  CC23 <- 2001

  for(j in unique(df$repnr)){
    regression_effects[CC01, ] <- regression_based(df = df[df$repnr ==
    j, ], x0 = 0, x1 = 1)
    regression_effects[CC12, ] <- regression_based(df = df[df$repnr ==
    j, ], x0 = 1, x1 = 2)
    regression_effects[CC23, ] <- regression_based(df = df[df$repnr ==
    j, ], x0 = 2, x1 = 3)

    CC01 <- CC01 + 1
    CC12 <- CC12 + 1
    CC23 <- CC23 + 1
  }
  save(regression_effects, file = paste0("regression, ", i, ".RData"))
}

rm(df, regression_effects, CC01, CC12, CC23, i, j)

simulation_effects <- store_effects(rows = 3 * reps)
for(i in files){
  load(paste0(path, i))
  CC01 <- 1
  CC12 <- 1001
  CC23 <- 2001

  for(j in unique(df$repnr)){
    simulation_effects[CC01, ] <- simulation_based(df = df[df$repnr ==
    j, ], x0 = 0, x1 = 1)
    simulation_effects[CC12, ] <- simulation_based(df = df[df$repnr ==
    j, ], x0 = 1, x1 = 2)
    simulation_effects[CC23, ] <- simulation_based(df = df[df$repnr ==
```

```
    j, ], x0 = 2, x1 = 3)

    CC01 <- CC01 + 1
    CC12 <- CC12 + 1
    CC23 <- CC23 + 1
  }
    save(simulation_effects, file = paste0("simulation, ", i,
    ".RData"))
}

rm(df, simulation_effects, CC01, CC12, CC23, i, j)

imputation_effects <- store_effects(rows = reps)
for(i in files){
  load(paste0(path, i))
  count <- 1
  for(j in unique(df$repnr)){
    imputation_effects[count, ] <- imputation_based(df = df[df$repnr ==
    j, ])
    count <- count + 1
  }
  save(imputation_effects, file = paste0("imputation, ", i, ".RData"))
}

rm(df, imputation_effects, count, i, j)

weighting_effects <- store_effects(rows = reps)
for(i in files){
  load(paste0(path, i))
  count <- 1
  for(j in unique(df$repnr)){
    weighting_effects[count, ] <- weighting_based(df = df[df$repnr ==
    j, ])
    count <- count + 1
  }
  save(weighting_effects, file = paste0("weighting, ", i, ".RData"))
}

rm(df, weighting_effects, count, i, j)
```

Step 5 – Calculate performance measures

```
load("Empirical true values.RData")

path <- "4. Effect estimates/"
reg <- list.files(path = path,
                  pattern = "regression")

performance <- store_performance(rows = 12)
count <- 0
for(i in reg){
```

```
  load(paste0(path, i))
  count <- count + 1
  empirical_true <- empirical_values[empirical_values$approach ==
  "regression" & empirical_values$x0 == 0, ]
  performance[count, ] <- performance_measures(df =
  regression_effects[regression_effects$x0 == 0, ])

  count <- count + 1
  empirical_true <- empirical_values[empirical_values$approach ==
  "regression" & empirical_values$x0 == 1, ]
  performance[count, ] <- performance_measures(df =
  regression_effects[regression_effects$x0 == 1, ])

  count <- count + 1
  empirical_true <- empirical_values[empirical_values$approach ==
  "regression" & empirical_values$x0 == 2, ]
  performance[count, ] <- performance_measures(df =
  regression_effects[regression_effects$x0 == 2, ])
}

save(performance, file = "Regression.RData")
rm(empirical_true, performance, regression_effects, count, reg, i)

sim <- list.files(path = path,
                  pattern = "simulation")
performance <- store_performance(rows = 12)
count <- 0
for(i in sim){
  load(paste0(path, i))
  count <- count + 1
  empirical_true <- empirical_values[empirical_values$approach ==
  "simulation" & empirical_values$x0 == 0, ]
  performance[count, ] <- performance_measures(df =
  simulation_effects[simulation_effects$x0 == 0, ])

  count <- count + 1
  empirical_true <- empirical_values[empirical_values$approach ==
  "simulation" & empirical_values$x0 == 1, ]
  performance[count, ] <- performance_measures(df =
  simulation_effects[simulation_effects$x0 == 1, ])

  count <- count + 1
  empirical_true <- empirical_values[empirical_values$approach ==
  "simulation" & empirical_values$x0 == 2, ]
  performance[count, ] <- performance_measures(df =
  simulation_effects[simulation_effects$x0 == 2, ])
}

save(performance, file = "Simulation.RData")
rm(empirical_true, performance, simulation_effects, count, sim, i)
```

```
imp <- list.files(path = path,
                  pattern = "imputation")
performance <- store_performance(rows = 4)
count <- 0
for(i in imp){
  load(paste0(path, i))
  count <- count + 1
  empirical_true <- empirical_values[empirical_values$approach ==
  "imputation", ]
  performance[count, ] <- performance_measures(df = imputation_effects)
}

save(performance, file = "Imputation.RData")
rm(empirical_true, performance, imputation_effects, count, imp, i)

weight <- list.files(path = path,
                     pattern = "weighting")
performance <- store_performance(rows = 4)
count <- 0
for(i in weight){
  load(paste0(path, i))
  count <- count + 1
  empirical_true <- empirical_values[empirical_values$approach ==
  "weighting", ]
  performance[count, ] <- performance_measures(df = weighting_effects)
}

save(performance, file = "Weighting.RData")
rm(empirical_true, performance, weighting_effects, count, weight, i)
```

Appendix B Empirical true values

Empirical true values of the indirect, direct and total effect estimated with the regression-, simulation-, imputation- and weighting-based approaches based on a sample size of 500,000. All simulated effects mimicked medium effect sizes: 0.92 for the exposure-mediator effect, and 0.39 for the mediator-outcome effect and the exposure-outcome effect.

Table B1 Empirical true values estimated with different approaches based on a sample size of 500,000

Approach	Indirect effect	Direct effect	Total effect
Regression			
$x^* = 0$ & $x = 1$	0.08286	0.39147	0.47433
$x^* = 1$ & $x = 2$	0.05704	0.39147	0.44851
$x^* = 2$ & $x = 3$	0.02990	0.39147	0.42137
Simulation			
$x^* = 0$ & $x = 1$	0.08282	0.39147	0.47429
$x^* = 1$ & $x = 2$	0.05726	0.39147	0.44873
$x^* = 2$ & $x = 3$	0.02953	0.39147	0.42101
Imputation	0.07496	0.39147	0.46643
Weighting	0.07583	0.39351	0.46934

Appendix C Relative bias and mean squared error of the indirect, direct and total effect estimated with the regression- and simulation-based approaches

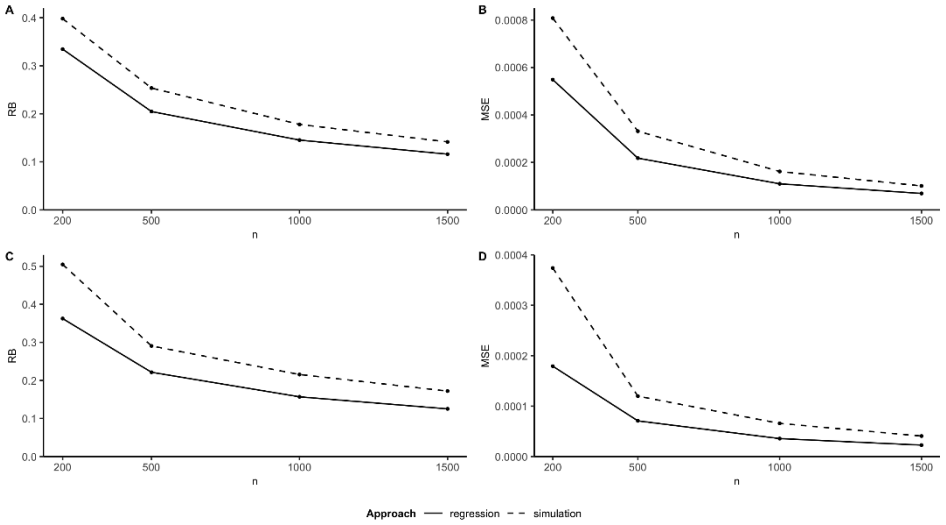

Figure C1 Relative bias (panels A and C) and mean squared error (panels B and D) of the indirect effect estimated with the regression- and simulation-based approaches for the causal contrasts $x^* = 1$ & $x = 2$ (top row) and $x^* = 2$ & $x = 3$ (bottom row)

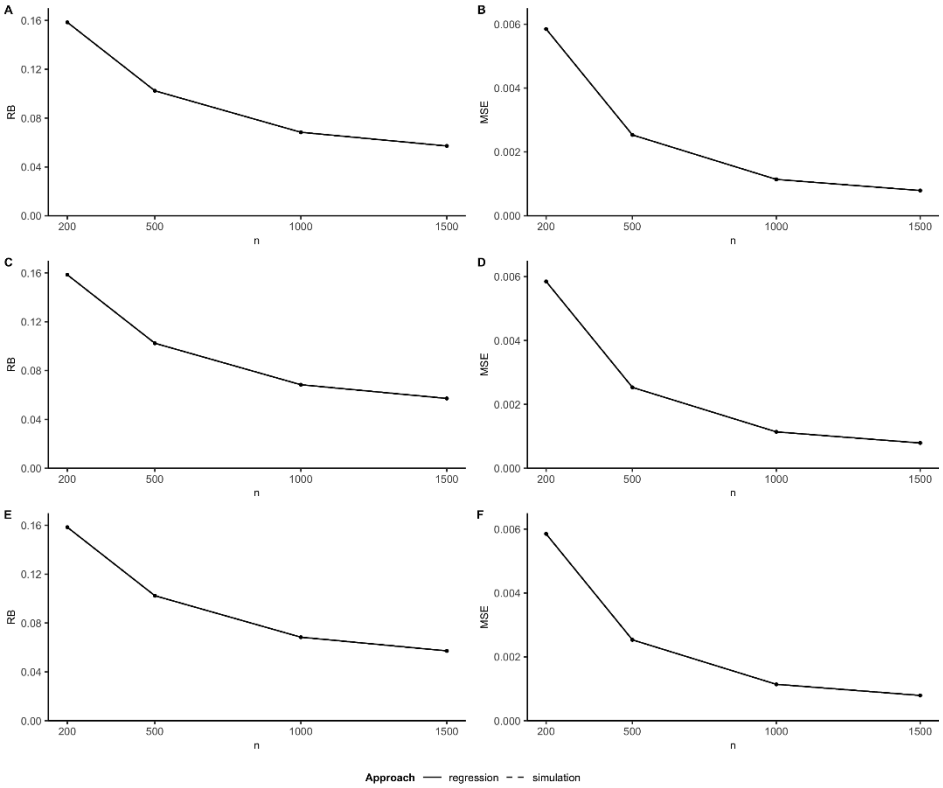

Figure C2 Relative bias (panels A, C and E) and mean squared error (panels B, D and F) of the direct effect estimated with the regression- and simulation-based approaches for the causal contrasts $x^* = 0$ & $x = 1$ (top row), $x^* = 1$ & $x = 2$ (middle row) and $x^* = 2$ & $x = 3$ (bottom row)

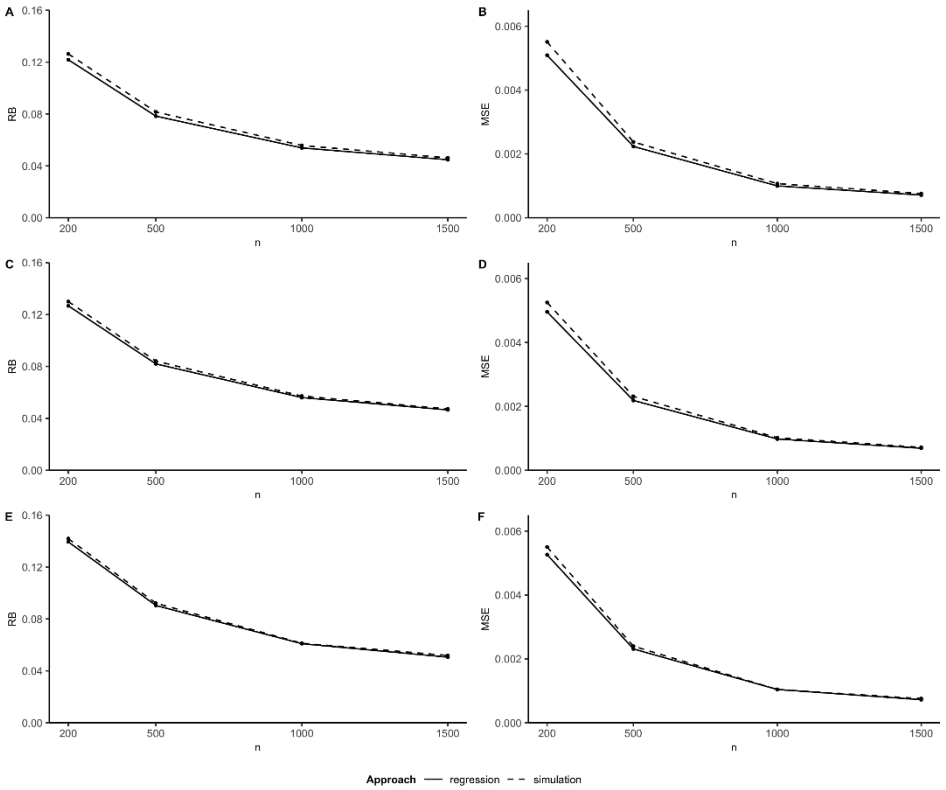

Figure C3 Relative bias (panels A, C and E) and mean squared error (panels B, D and F) of the total effect estimated with the regression- and simulation-based approaches for the causal contrasts $x^* = 0$ & $x = 1$ (top row), $x^* = 1$ & $x = 2$ (middle row) and $x^* = 2$ & $x = 3$ (bottom row)

Appendix D Mean estimated indirect, direct and total effects from the regression- and simulation-based approaches based on a sample size of 1,000

Table D1 Mean indirect, direct and total effect estimates from the regression- and simulation-based approaches based on a sample size of 1,000, estimated for different causal contrasts

Approach	Causal contrast	Indirect effect	Direct effect	Total effect
Regression	$x^* = 0$ & $x = 1$	0.08377	0.39024	0.47401
	$x^* = 1$ & $x = 2$	0.05742	0.39024	0.44767
	$x^* = 2$ & $x = 3$	0.03014	0.39024	0.42038
Simulation	$x^* = 0$ & $x = 1$	0.08355	0.39024	0.47379
	$x^* = 1$ & $x = 2$	0.05766	0.39024	0.44790
	$x^* = 2$ & $x = 3$	0.03033	0.39024	0.42058

Appendix E Relative bias and mean squared error of the direct and total effect estimated with the imputation- and weighting-based approaches

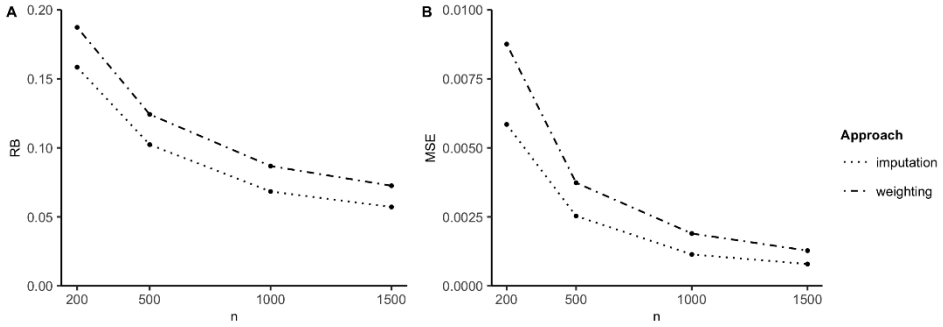

Figure E1 Relative bias (panel A) and mean squared error (panel B) of the direct effect estimated with the imputation- and weighting-based approaches

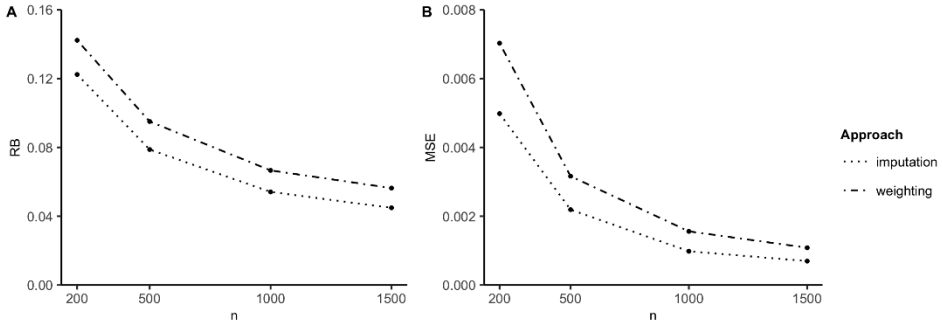

Figure E2 Relative bias (panel A) and mean squared error (panel B) of the total effect estimated with the imputation- and weighting-based approaches

CHAPTER 6

Ignoring competing events in the analysis of survival data may lead to biased results: a nonmathematical illustration of competing risk analysis

Noah A. Schuster
Emiel O. Hoogendijk
Almar A.L. Kok
Jos W.R. Twisk
Martijn W. Heymans

Published in Journal of Clinical Epidemiology (2020)

Abstract

Objective

Competing events are often ignored in epidemiological studies. Conventional methods for the analysis of survival data assume independent or noninformative censoring, which is violated when subjects that experience a competing event are censored. Because many survival studies do not apply competing risk analysis, we explain and illustrate in a nonmathematical way how to analyze and interpret survival data in the presence of competing events.

Study design and setting

Using data from the Longitudinal Aging Study Amsterdam, both marginal analyses (Kaplan–Meier method and Cox proportional-hazards regression) and competing risk analyses (cumulative incidence function [CIF], cause-specific and subdistribution hazard regression) were performed. We analyzed the association between sex and depressive symptoms, in which death before the onset of depression was a competing event.

Results

The Kaplan–Meier method overestimated the cumulative incidence of depressive symptoms. Instead, the CIF should be used. As the subdistribution hazard model has a one-to-one relation with the CIF, it is recommended for prediction research, whereas the cause-specific hazard model is recommended for etiologic research.

Conclusion

When competing risks are present, the type of research question guides the choice of the analytical model to be used. In any case, results should be presented for all event types.

Introduction

Survival data are often encountered in epidemiologic studies. In this kind of data, the outcome of interest is time to the occurrence of a certain event. An important feature of survival data is censoring, which occurs when the exact survival time is unknown. This is the case, for example, when a subject is lost to follow-up, withdraws from the study, or does not experience the event of interest before the end of the study. Conventional methods used in the analysis of survival data like the Kaplan–Meier method and Cox proportional-hazards regression make the assumption of independent or non-informative censoring. This means that individuals who are censored have the same future risk of the event of interest as subjects under observation (1, 2). In other words, this kind of censoring does not change study outcome on disease prognosis or risk factor detection.

Another important but less well-known feature of survival data are competing risks. A competing risk is an event that prevents the event of interest from happening (3). Suppose we are interested in the onset of depression, then death before the onset of depression is a competing event. Censoring these subjects is problematic in two ways. First, the assumption of independence or noninformative censoring is violated, as a subject that experiences a competing event (death) is censored in an informative manner (4, 5). Second, the probability of experiencing the event of interest is estimated in a hypothetical setting in which the competing event cannot occur, which has very little clinical relevance (1, 2).

In epidemiological and medical research, competing risks are often ignored in the analysis of survival data. However, failing to account for competing risks generally leads to an overestimation of the cumulative incidence of the event of interest (1, 4, 6-8). In 2012, Koller et al. critically appraised 50 recently published articles in which competing risks were present from different biostatistical, clinical, and high-impact medical journals (9). In 70% of the included articles, they observed at least one competing risks issue. However, in only 20% of the studies, specific competing risks methodology was applied.

Although there is extensive literature on competing risks (1, 3, 7, 10, 11), articles that explain how to analyze survival data in the presence of competing risks in a nonmathematical way are scarce (5). In addition, there is a lack of articles that focus on the application of different methods in real-life data and subsequently on the interpretation of the results. Therefore, the aim of this study was to explain and illustrate

how to analyze and interpret survival data in the presence of a competing event. We will compare conventional methods of survival analysis with competing risk methods in the analysis of real-life data from an observational cohort study.

Description of the data

The application of methods is illustrated using data from the Longitudinal Aging Study Amsterdam (LASA), a prospective cohort study among older adults in the Netherlands (12, 13). In the present study, we included respondents that participated in the second measurement wave of LASA (1995–1996). Data on various domains of function were collected approximately every 3 years. More information on LASA and the measurements included in this study can be found elsewhere (12, 13).

The outcome of interest was incident depression, approximated by a score of ≥16 on the Center for Epidemiologic Studies Depression (CES-D) scale (14). Individuals who already suffered from depression at the start of the study were excluded, leaving a sample of 1,187 subjects.

Subjects that were not contacted for a new round of interviews, that were ineligible, or that refused were censored on the date of their last completed interview. Subjects that were still event free at the end of the study (01-07-2015) were also censored.

Statistical analyses

We analyzed the association between sex and the onset of depression. Because a comprehensive assessment of predictors of depression incidence was beyond the aim of this study, we limited our model to the inclusion of sex, baseline age, number of chronic diseases, and smoking. We performed both crude and adjusted analyses. Age was categorized into quartiles due to nonlinearity.

Marginal analyses

In a classic survival setting, the survivor function is estimated using the Kaplan–Meier (KM) method (15). The complement of the Kaplan–Meier estimate denotes the probability of experiencing the event of interest before a specified time. As this method can only handle one outcome and thus assumes independent or noninformative censoring, the cumulative incidence derived from this method is interpreted as the probability of depression in a world in which subjects cannot die before developing depressive symptoms (16, 17). Using the Kaplan–Meier method, censoring subjects at the time they

experience a competing event has no influence on the cumulative survival probability (18), which generally leads to an overestimation of the cumulative incidence (2, 4, 7).

Marginal multivariable survival analysis is performed using Cox PH regression. The marginal hazard derived from a Cox model denotes the instantaneous rate of occurrence of the event of interest in a setting in which subjects cannot experience the competing event. Just like the Kaplan–Meier method, Cox PH regression assumes independent or noninformative censoring. In the absence of competing risks, the hazard and cumulative incidence are directly related in such a way that an increased hazard has a one-to-one association with a shorter survival time (2, 3, 9, 19, 20). Then, by fitting a Cox PH regression model in our example dataset, inference can be made about the effect sex has on both the hazard function and on the prognosis or survival.

Competing risk analyses

The competing risk equivalent of the Kaplan–Meier method is the cumulative incidence function (CIF). The CIF denotes the probability of experiencing the event of interest before a specific time and before the occurrence of any other type of event (2), meaning that subjects experiencing the competing event are considered no longer to be at risk of developing the event of interest (16-18). As a result of this, the cumulative survival probability is lowered by the occurrence of a competing event because the number of persons at risk decreases more quickly over time (18). Thus, the CIF estimates the probability of depression in a clinically relevant setting in which subjects may also die (2, 21). In a scenario in which there are no competing events, the CIF yields the same cumulative incidence as the KM method.

The one-to-one relation between the hazard and cumulative incidence that is present in the multivariable marginal analysis does not automatically translate to a competing risk framework (22). Therefore, in the presence of competing risks, the hazard and cumulative incidence cannot be estimated from one single model and different models need to be applied to answer etiologic and prognostic epidemiologic research questions: the cause-specific hazard model (etiologic) or the subdistribution hazard model (prognostic) (3, 7, 9, 10, 23).

Cause-specific hazard regression

The cause-specific hazard denotes the instantaneous rate of occurrence of the event of interest in a setting in which subjects can also experience the competing event (1, 3). This

hazard is estimated by removing individuals from the risk set the moment they experience the competing event, meaning that competing events are treated as censored observations (3, 21). Thus, the estimation procedure is the same as the procedure for marginal survival analysis and the cause-specific hazard can be estimated by fitting a standard Cox PH model in which all events other than the event of interest are treated as censoring. Consequently, when censoring is noninformative, we quantify the effects on the marginal hazard, whereas in the case of informative censoring, we quantify the effects on the cause-specific hazard (1, 3, 23). Thus, hazard ratios derived from a cause-specific hazard model should be interpreted among subjects who did not (yet) experience the event of interest or a competing event (16). As the cause-specific hazard is directly quantified among subjects that are actually at risk of developing the event of interest, the cause-specific hazard model is considered more appropriate for etiologic research (16).

Whereas in the marginal analysis a model is fitted for the event of interest only, for the cause-specific hazard model, separate models are fitted for each type of event in which individuals that experience the competing event are censored (1, 3). Thus, in our study, we will fit two models: one for depression in which subjects that die are censored and one for death in which subjects that are diagnosed with depression are censored, and we interpret both hazard ratios at the same time.

Subdistribution hazard regression

The subdistribution hazard denotes the instantaneous risk of the event of interest in subjects that have not (yet) experienced the event of interest. This means that subjects who experience the competing event remain in the risk set (3, 10, 20). Thus, the risk set for the subdistribution hazard model contains not only subjects that are currently free of the event of interest but also subjects that have previously experienced the competing event. In our example, this means that the risk set consists of both individuals that have not (yet) developed depressive symptoms and individuals that died before the onset of depression. Although this feels unnatural—as subjects who have died are naturally no longer at risk of developing depressive symptoms—this is necessary to establish the one-to-one relation with the CIF. Because of the direct relation between the covariates and the CIF, the subdistribution hazard model is considered the right model for prediction research.

Because in the subdistribution hazard model individuals that experienced the competing event remain in the risk set, the hazard ratios derived from a subdistribution hazard model are not straightforward to interpret (7, 23). As a result of this, the subdistribution hazard model is not considered appropriate for etiologic research. However, in prediction research, the hazard ratios are used to calculate individual risks. Thus, the regression coefficients derived from the subdistribution hazard model can be used to compute the cumulative incidence of depression, taking competing risks into account (8, 20).

Like for the Cox model, both the cause-specific hazards and the subdistribution hazards are assumed to be proportional over time. This can be checked using Schoenfeld residuals (24).

Notation and reporting

In a classic survival setting, researchers often simply address the risk of an event without specifying whether risk denotes the hazard or the cumulative incidence of the event (2). In a competing risk framework, the use of clear terminology is required to avoid the misconception that the cause-specific and subdistribution hazard are essentially the same. Therefore, the cause-specific hazard and subdistribution hazard ratios will be reported as HR_{cs} and HR_{sd}, respectively. In addition, Latouche et al. have suggested to use both models and present the results for all causes for complete understanding (5, 21). Therefore, in competing risk analysis, in the example, both the hazard for depression and the hazard for death will be reported.

Software

All analyses were conducted using the R (version 3.5.3) statistical programming language (25) and the "cmprsk" package (version 2.2-7) for the competing risk analyses (26). Detailed information on how to perform competing risk analyses in R using the "cmprsk" package can be found elsewhere (2, 3, 27, 28).

Results

Descriptive statistical analyses

The population consisted of 625 males and 562 females (Table 1). Of all males, 16% developed clinically relevant depressive symptoms, whereas for women, this was 27%. Just over half of all women died without having had clinically relevant depressive symptoms during the study (50.36%), whereas for males, this percentage was much

Table 1 Characteristics of study population

	Males n = 625	Females n = 562
Status, n (%)		
Censored	109 (17.44)	125 (22.24)
Developed clinically relevant depressive symptoms	103 (16.48)	154 (27.40)
Deceased	413 (66.08)	283 (50.36)
Follow-up in days, median (IQR)	2441 (3621)	3220 (4159.25)

Abbreviations: n = number; IQR = inter quartile range

higher (66.08%). Median follow-up was longer for females (3,320 days) than for males (2,241 days).

Cumulative incidence

Figure 1 shows the cumulative incidence of depression (panel A) and both depression and death (panel B) for both males and females derived from the Kaplan–Meier method and the CIF, respectively. As anticipated, the Kaplan–Meier estimate of the incidence of clinically relevant depressive symptoms is larger than the corresponding estimate derived from the CIF. For instance, at 4,000 days, the cumulative incidence of clinically relevant depressive symptoms derived from the Kaplan–Meier method is 20.61% for males and 28.96% for females, whereas the cumulative incidence of depressive symptoms derived from the CIF is 14.55% for males and 24.21% for females. At 6,000 days, the difference in probabilities is even larger.

Modeling covariate effects

Table 2 shows the cause-specific and subdistribution hazard ratios for depression and death. The hazard ratio for depression derived from the Cox PH model is not included in Table 2 as this is equal to the cause-specific hazard ratio for depression.

Cause-specific hazard model

Female sex is associated with an increase in the rate of the development of clinically relevant depressive symptoms among those who are still alive and do not yet suffer from depressive symptoms (adjusted HR_{cs} 1.537, 95% CI 1.193–1.982), whereas it significantly decreases the rate of death before the onset of depression in the same group (adjusted HR_{cs} 0.684, 95% CI 0.586–0.797).

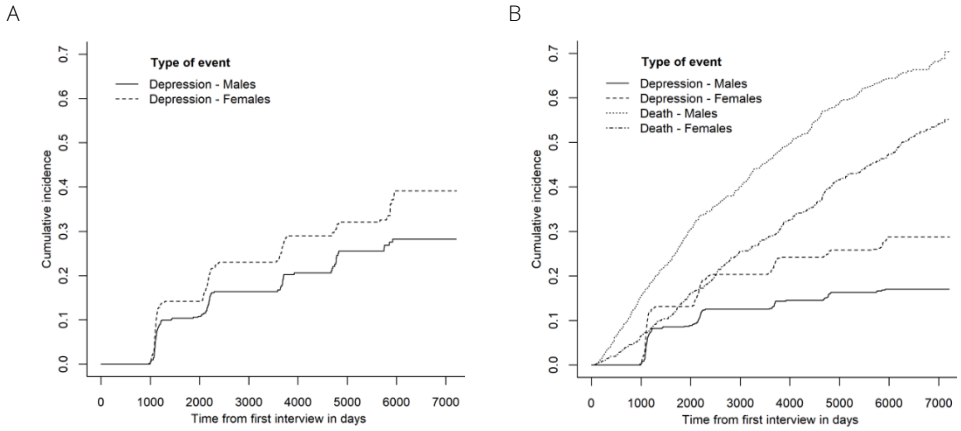

Figure 1 Cumulative incidence of depression derived from the Kaplan-Meier method (panel A) and the cumulative incidence function (panel B)

Subdistribution hazard model

As expected with a higher rate of depression for females associated with a reduced rate of death, we observe more females than males diagnosed with depression at any point during the study. Being female increases the probability of depression, resulting in an 84% higher relative incidence of clinically relevant depressive symptoms for females than for males (adjusted HR_{sd} 1.842, 95% CI 1.430–2.370), whereas it decreases the probability of dying before the onset of depression. The relative incidence of death was more than 35% lower for females than for males (adjusted HR_{sd} 0.639, 95% CI 0.547–0.746). Survival probabilities can be calculated for each individual by combining the subdistribution hazard ratios with their baseline hazard, just like one would do with the hazard ratios derived from a Cox model in a situation in which no competing risks are present.

In conclusion, sex has a more pronounced effect on the incidence of depression than on the cause-specific hazard of depression, as evidenced by the finding that the HR_{sd} (1.842) was larger than the HR_{cs} (1.537). The apparent increase in the absolute risk of depression for females might be explained via the effect sex has on death before the onset of depression.

Discussion

In epidemiologic research, competing risks are generally not considered in the analysis of survival data. In the presence of competing risks, cumulative incidence should be estimated using the cumulative incidence function instead of the Kaplan-Meier method.

Table 2 Cause-specific and subdistribution hazards for depression and death

	Cause-specific hazard models		Subdistribution hazard models	
	Depression	Death	Depression	Death
Crude				
Sex – female	1.453	0.664	1.780	0.618
	(1.132 – 1.865)	(0.571 – 0.722)	(1.390 – 2.290)	(0.534 – 0.718)
Adjusted				
Sex - female	1.537	0.684	1.842	0.639
	(1.193 – 1.982)	(0.586 – 0.797)	(1.430 – 2.370)	(0.547 – 0.746)

Death represents 'death prior to the onset of depression'. The cause-specific and subdistribution hazard model return the cause-specific (HR_{cs}) and subdistribution hazard (HR_{sd}) and their corresponding 95% confidence intervals, respectively. In the adjusted analyses we correct for age, number of chronic diseases and smoking.

Our illustration showed that failing to account for death before the onset of depression as a competing risk resulted in an overestimation of the cumulative incidence of clinically relevant depressive symptoms by 6.06 percentage point for males and 4.75 percentage point for females. For prediction research, the subdistribution hazard model should be used. In our illustration, the adjusted subdistribution hazard ratio for depression in females was greater than in the marginal analysis (HR_{sd} 1.842 [1.430–2.370] vs. HR 1.537 [1.193–1.982]), whereas the adjusted subdistribution hazard ratio for death in females was lesser (HR_{sd} 0.639 [0.547–0.746] vs. HR 0.684 [0.586–0.797]).

The extent to which the cumulative incidence is overestimated is related to the proportion of subjects experiencing the event of interest and the competing event. It is discussed in literature that specific competing risk analysis should be considered when the proportion of subjects that experience the competing event is equal to or greater than the proportion of subjects that experience the outcome of interest (6) or when the absolute percentage of competing events is greater than 10% (2). In our data example, the incidence of clinically relevant depressive symptoms is relatively low, whereas mortality is high. As a result, the cumulative incidence is greatly overestimated using marginal analysis methods, illustrating the importance of applying specific competing risk analysis (5). In a younger population in which the incidence of depression is higher (29, 30) and mortality naturally is lower, the estimates derived from marginal analyses and competing risk analyses will not differ to the same extent as what we found in our older study population.

Overestimation of the cumulative incidence of the outcome of interest has both practical and public health implications. An example of these implications is that treatment decisions by clinicians are often guided by risk prediction models. Ignoring competing risks in the development of these models could, among other things, lead to possible overtreatment in future patients.

Limitations

A limitation of the real-life data example is that in LASA, as in many cohort studies, disease information is collected at discrete follow-up visits, whereas the exact date of death is retrieved from municipality registers. It is therefore possible that we have missed some cases of incident depression (31). In addition, we could not distinguish between first-onset and recurrent depression. Because incidence of depression was based on a screening instrument (14), this does not necessarily indicate a clinical diagnosis, and there was no information on previous episodes. It is therefore possible that a part of the observed incidence of depression in our study represents recurrent episodes. Another limitation is that age was categorized into quartiles, which is associated with loss of information. Although there are better methods available to model nonlinear relationships (e.g., spline functions), in order not to divert attention from competing risk analysis we used categorization, which is still a widely used method in epidemiological research.

Prediction model performance in the presence of competing risks

The process of developing a prediction model in a competing risks framework is essentially the same as for other regression models, except that the subdistribution hazard model should be applied instead of regular Cox PH regression. The performance of a prediction model is usually assessed using the calibration and discrimination. A detailed proposal of how to assess calibration and discriminative capacity of a prediction model in a competing risks setting is described by Wolbers et al. (8).

Competing risks in randomized controlled trials

Whereas our paper focusses on competing risks in observational studies, competing risks also appear in the setting of randomized controlled trials (RCTs). A recent review of randomized controlled trials with survival outcomes that were published in four high-impact general medical journals showed that most of the studies were potentially susceptible to competing risks, but that this was not accounted for in the statistical analyses (32).

In RCTs with time-to-event outcomes, often additional effect measures that are derived from the KM survival curves, like the number needed to treat (NNT), are reported. Because the KM method overestimates the cumulative incidence in the presence of competing risks, the estimated NNT may also be biased. Therefore, to correctly estimate the NNT in the presence of competing risks, it is recommended to use a method based on the CIF (33). For multivariable analysis, the same applies for RCTs as for observational studies: the cause-specific hazard model should be used when one is interested in the effect of the intervention on the instantaneous rate of occurrence of the event of interest in subjects that are currently event free, whereas the subdistribution hazard model should be used when one is interested in the relative effect of the intervention on the cumulative incidence function (32).

Software

The cause-specific hazard model can be fitted with any software that can perform a Cox PH model. However, this is not the case for the CIF and the subdistribution hazard model. How to estimate the CIF in SPSS with the use of a macro is described elsewhere (18). In STATA, the subdistribution hazard model can be fitted using the *stcrreg* package (10). For SAS, macros for both the estimation of the CIF and the subdistribution hazard model are available (34, 35).

Conclusion

In conclusion, competing risks form an important issue in the analysis of survival data. Researchers should be aware of the potential problems associated with censoring subjects when they experience a competing event. Dealing with competing risks requires careful formulation of the research question, selection of the appropriate method for data analysis, and interpretation of the results.

References

1. Putter H, Fiocco M, Geskus RB. Tutorial in biostatistics: competing risks and multi-state models. Stat Med. 2007;26(11):2389-430.
2. Austin PC, Lee DS, Fine JP. Introduction to the analysis of survival data in the presence of competing risks. Circulation. 2016;133(6):601-9.
3. Geskus RB. Data Analysis with Competing Risks and Intermediate States. Boca Raton, FL: Taylor & Francis Group, LLC; 2016.
4. Satagopan JM, Ben-Porat L, Berwick M, Robson M, Kutler D, Auerbach AD. A note on competing risks in survival data analysis. Br J Cancer. 2004;91(7):1229-35.
5. Wolkewitz M, Cooper BS, Bonten MJM, Barnett AG, Schumacher M. Interpreting and comparing risks in the presence of competing events. BMJ : British Medical Journal. 2014;349:g5060.
6. Berry SD, Ngo L, Samelson EJ, Kiel DP. Competing risk of death: an important consideration in studies of older adults. J Am Geriatr Soc. 2010;58(4):783-7.
7. Lau B, Cole SR, Gange SJ. Competing risk regression models for epidemiologic data. Am J Epidemiol. 2009;170(2):244-56.
8. Wolbers M, Koller MT, Witteman JC, Steyerberg EW. Prognostic models with competing risks: methods and application to coronary risk prediction. Epidemiology. 2009;20(4):555-61.
9. Koller MT, Raatz H, Steyerberg EW, Wolbers M. Competing risks and the clinical community: irrelevance or ignorance? Stat Med. 2012;31(11-12):1089-97.
10. Fine JP, Gray RJ. A proportional hazards model for the subdistribution of a competing risk. Journal of the American Statistical Association. 1999;94(446):496-509.
11. Zhang M-J, Zhang X, Scheike TH. Modeling cumulative incidence function for competing risks data. Expert Rev Clin Pharmacol. 2008;1(3):391-400.
12. Hoogendijk EO, Deeg DJ, Poppelaars J, van der Horst M, Broese van Groenou MI, Comijs HC, et al. The Longitudinal Aging Study Amsterdam: cohort update 2016 and major findings. Eur J Epidemiol. 2016;31(9):927-45.
13. Hoogendijk EO, Deeg DJH, de Breij S, Klokgieters SS, Kok AAL, Stringa N, et al. The Longitudinal Aging Study Amsterdam: cohort update 2019 and additional data collections. Eur J Epidemiol. 2019.
14. Radloff LS. The CES-D scale: a self-report depression scale for research in the general population. Applied psychological measurement. 1977;1(3):385-401.
15. Bland JM, Altman DG. Survival probabilities (the Kaplan-Meier method). BMJ. 1998;317:1572.

16. Noordzij M, Leffondre K, van Stralen KJ, Zoccali C, Dekker FW, Jager KJ. When do we need competing risks methods for survival analysis in nephrology? Nephrol Dial Transplant. 2013;28(11):2670-7.
17. Southern DA, Faris PD, Brant R, Galbraith PD, Norris CM, Knudtson ML, et al. Kaplan-Meier methods yielded misleading results in competing risk scenarios. J Clin Epidemiol. 2006;59(10):1110-4.
18. Verduijn M, Grootendorst DC, Dekker FW, Jager KJ, le Cessie S. The analysis of competing events like cause-specific mortality--beware of the Kaplan-Meier method. Nephrol Dial Transplant. 2011;26(1):56-61.
19. Kleinbaum DG, Klein M. Survival Analysis: A Self-Learning Text. 3rd ed: Springer Science+Business Media; 2012.
20. Austin PC, Fine JP. Practical recommendations for reporting Fine-Gray model analyses for competing risk data. Stat Med. 2017;36(27):4391-400.
21. Latouche A, Allignol A, Beyersmann J, Labopin M, Fine JP. A competing risks analysis should report results on all cause-specific hazards and cumulative incidence functions. J Clin Epidemiol. 2013;66(6):648-53.
22. Gray RJ. A class of k-sample tests for comparing the cumulative incidence of a competing risk. The Annals of Statistics. 1988;16(3):1141-54.
23. Andersen PK, Geskus RB, de Witte T, Putter H. Competing risks in epidemiology: possibilities and pitfalls. Int J Epidemiol. 2012;41(3):861-70.
24. Haller B, Schmidt G, Ulm K. Applying competing risks regression models: an overview. Lifetime Data Anal. 2013;19(1):33-58.
25. R Core Team. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2019.
26. Gray RJ. cmprsk: Subdistribution Analysis of Competing Risks. 2014.
27. Scrucca L, Santucci A, Aversa F. Competing risk analysis using R: an easy guide for clinicians. Bone Marrow Transplant. 2007;40(4):381-7.
28. Scrucca L, Santucci A, Aversa F. Regression modeling of competing risk using R: an in depth guide for clinicians. Bone Marrow Transplant. 2010;45(9):1388-95.
29. Büchtemann D, Luppa M, Bramesfeld A, Riedel-Heller S. Incidence of late-life depression: A systematic review. Journal of Affective Disorders. 2012;142(1):172-9.
30. The ESEMeD Mhedea Investigators, Alonso J, Angermeyer MC, Bernert S, Bruffaerts R, Brugha TS, et al. Prevalence of mental disorders in Europe: results from the European Study of the Epidemiology of Mental Disorders (ESEMeD) project. Acta Psychiatrica Scandinavica. 2004;109(s420):21-7.

31. Binder N, Blümle A, Balmford J, Motschall E, Oeller P, Schumacher M. Cohort studies were found to be frequently biased by missing disease information due to death. Journal of Clinical Epidemiology. 2019;105:68-79.
32. Austin PC, Fine JP. Accounting for competing risks in randomized controlled trials: a review and recommendations for improvement. Stat Med. 2017;36(8):1203-9.
33. Gouskova NA, Kundu S, Imrey PB, Fine JP. Number needed to treat for time-to-event data with competing risks. Stat Med. 2014;33(2):181-92.
34. Kohl M, Plischke M, Leffondre K, Heinze G. PSHREG: a SAS macro for proportional and nonproportional subdistribution hazards regression. Comput Methods Programs Biomed. 2015;118(2):218-33.
35. Rosthøj S, Andersen PK, Abildstrom SZ. SAS macros for estimation of the cumulative incidence functions based on a Cox regression model for competing risks survival data. Computer Methods and Programs in Biomedicine. 2004;74(1):69-75.

Chapter 7

General discussion

Background

In epidemiological research, regression analysis is often used to examine the association between an exposure and an outcome, also called the exposure effect. There are many different types of regression techniques, and the distribution of the outcome determines which technique is most appropriate to estimate the exposure effect. Linear regression (for continuous outcomes), logistic regression (for binary outcomes) and Cox regression (for survival outcomes) are the most common techniques used in the field of epidemiology.

The main goal of regression analysis is to estimate the most accurate effect obtainable from the data. However, often the association between an exposure and an outcome is not entirely attributable to the exposure, i.e., the effect is *biased*. Bias can occur in all stages of a study and results in an underestimation or overestimation of the true effect. In some situations, it can even reverse the apparent direction of the effect. Failing to consider potential sources of bias may result in incorrect inference about the association between the exposure and the outcome. At the policy level, biased studies could influence policy development and ultimately lead to the implementation of ineffective public health policy (1). This, in turn, could lead to wrong conclusions about the harmful or beneficial effect of a certain treatment and thus to the decision to continue or stop treatment at the individual level (2).

In this thesis, I focused on the prevention of bias in the analysis stage of a study. The aim was to describe various situations in which bias can occur as a result of the incorrect application of linear-, logistic- and Cox regression models, and to propose solutions where possible. Four topics were covered: the estimation of non-linear effects, noncollapsibility, causal mediation analysis and competing risks. Although the mechanisms and methods described in this thesis are not new, existing literature contains a high level of technical and mathematical details, which may hamper the understanding and the application of correct methods. The chapters in this thesis were mainly written for applied researchers, meaning that the sources of bias and methods are described in a non-technical and non-mathematical way and that the emphasis is on the interpretation of the results. In addition, each chapter contains an empirical data example, and where possible we provide a detailed appendix including software code offering researchers all tools necessary to apply these methods to their own research. This chapter contains a discussion of the main findings of this thesis and provides recommendations for practice.

Non-linear effects

A principal assumption of linear-, logistic- and Cox regression is that the exposure is linearly related to the outcome, i.e., that the exposure effect is the same for all one-unit differences in the exposure values. If this assumption is not met, then the effect estimate is not a good representation of the true underlying effect, and bias is introduced. It is common practice to assess the linearity assumption for the exposure-outcome effect. However, when adjusting for a confounder, the linearity assumption no longer only applies to the exposure-outcome effect, but also to the confounder-exposure or confounder-outcome associations, depending on the confounder-adjustment method used. If the functional form (i.e., the shape) of these associations is misspecified (i.e., linearity is wrongly assumed), then bias might be introduced in an attempt to remove bias.

In chapter 2 of this thesis, we reviewed four confounder-adjustment methods: multivariable regression analysis, covariate adjustment using the propensity score (PS), inverse probability weighting (IPW) and double robust (DR) estimation. We used a Monte Carlo simulation study to assess and compare their performance when the functional form of the confounder-exposure and confounder-outcome associations were misspecified and correctly specified under multiple sample sizes. In order to estimate unbiased effects, for methods that use the propensity score (i.e., covariate adjustment using the PS and IPW) the confounder-exposure association needs to be correctly specified, whereas the confounder-outcome association or PS-outcome association needs to be correctly specified if the outcome is regressed on the confounder or the propensity score, respectively. For all methods, the amount of bias depends on the strength of the associations and the sample size. Our study showed that merely adjusting for confounding is not enough, but that correct specification of *all* effects in a model is crucial to obtain unbiased exposure effect estimates.

In our study we adjusted for one confounder, whereas in reality there might be multiple. Naturally, multiple confounders increase the likelihood of non-linear confounder-exposure or confounder-outcome associations. To obtain unbiased results, the functional form of the associations of each of the confounders needs to be assessed separately and non-linear associations need to be modelled if necessary. In addition, we assumed that associations were either correctly specified or misspecified, whereas in reality this might not be a clear dichotomy.

In a systematic review from 2013, 53 papers were identified in high-impact general medical journals that adjusted for the continuous confounder age (3). In 40 of those, age was included as a covariate in the regression model. Only 13 of those explicitly reported how age was included as the model (e.g., as a linear term, as a categorized variable or with the use of higher-order terms). For the other 27 studies it was unclear how the relation between age and the outcome was modelled. Like us, Groenwold et al. concluded that the impact of misspecification of the functional form depends on the strength of the association between the confounder and both the exposure and the outcome. In addition, they identified the distribution of the confounder, other confounders that are also adjusted for and the extent of departure from linearity as factors that influence the magnitude of bias. A cross-sectional survey from 2002 on the frequency and adequacy of adjustment for confounding found that 45% of the included papers did not explicitly report how multicategorical or continuous variables were adjusted for in the analysis (4). Failing to provide information on the assessment or modelling of continuous confounders complicates the assessment of the validity and the interpretation of the results.

To increase transparency on the risk of additional bias, researchers should report how the functional form of the confounder-exposure and confounder-outcome associations was assessed and taken into account. In 2007, the STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) initiative published a checklist, the aim of which was to improve the quality of reporting of observational studies (5, 6). The checklist contains 22 items, a number of which are about bias. Item number 9 emphasizes that researchers should asses the likelihood of relevant biases and should discuss, and if possible, estimate, the direction and magnitude of bias. Item numbers 12 and 16 address confounding: researchers should make clear which confounders were adjusted for and why they were included, and which statistical methods were used to control for confounding.

A systematic review from 2008 on the reporting of confounding in observational studies on medical interventions found that the quality was very poor (7). For example, only 10% of the articles reported reasons for the selection of potential confounders. A more recent study from 2016 assessed whether the reporting of confounding improved after publication of the STROBE checklist (8). They found that although the quality improved in certain aspects, the overall quality remained substandard. Transparent reporting also includes reporting the methods used to select confounders to adjust for. A review from

2019 found that 37% of the articles selected did not provide sufficient details to assess how variables were selected (9). Transparency on measures taken to reduce bias not only encompasses the reporting of bias but also careful interpretation of the results. A review from 2018 assessed whether authors of observational studies consider confounding bias when interpreting their results (10). They found that many studies lack satisfactory discussion of confounding bias, and that when it is mentioned authors are often confident that it is irrelevant to their results.

Other checklists such as AGReMA (A Guideline for Reporting Mediation Analyses) also emphasize the transparent reporting of confounding (11). Items numbers 10 to 12, about the assumed causal model, causal assumptions and measurement of variables, encompasses possible confounders. Furthermore, item number 14 explicitly mentions that analytical strategies used to reduce confounding bias should be described.

To estimate unbiased effects it is important to examine the functional form of the confounder-exposure or confounder-outcome association depending on the confounder-adjustment method used and to adjust the model accordingly. The easiest way to assess linearity of the effects is by visual inspection: a scatterplot provides an indication of the nature of the relationship between the two variables. Non-visual ways to assess linearity include adding a non-linear term to the model and categorization of the continuous independent variable (12, 13).

If the linearity assumption does not hold, then the non-linear associations present in the data have to be modelled explicitly in order to obtain unbiased effects. There are different methods available to model non-linear associations, such as the use of higher-order terms, categorization of the exposure variable, linear spline regression and restricted cubic spline regression. In chapter 3 of this thesis we reviewed these methods and compared them in terms of their performance. We found that categorization of a continuous variable performed least well. Although this finding is not new and many argued against the categorization of continuous variables in the past (12, 14, 15), it remains a common technique in epidemiological research. Many non-linear associations can be modelled well using higher-order terms. However, this does not allow for straightforward interpretation of the effect estimates, which is problematic if the exposure-outcome effect is non-linear. This is not a hindrance when higher-order terms are used to model non-linear confounder-exposure and confounder-outcome associations, as the corresponding confounder-related coefficients are typically not

interpreted (3). Linear- and restricted cubic spline regression result in good approximations of the true effect. For restricted cubic spline regression, adding higher order terms further increases the flexibility of the model. However, again, this is at the expense of the interpretation of the coefficients.

Although spline regression is easy to implement with most statistical software programs, most papers on spline functions present these as complex mathematical functions (16-18). We presented spline functions in a step-by-step and non-mathematical way and focused on the application of the methods and on the interpretation of the results. In our study we illustrated the application of spline-models within a linear regression context. However, spline functions can be applied beyond standard linear regression models, for example in mediation models or in the analysis of longitudinal data.

Noncollapsibility

To determine which confounders to adjust for in the analysis, researchers often use the change-in-estimate criterion: they compare exposure effect estimates between a univariable- and a multivariable regression model and use an arbitrary cut-off value to determine the presence of relevant confounding (19-21). However, in logistic regression, the change-in-estimate might not only represent confounding bias but also a noncollapsibility effect. This noncollapsibility effect stems from a change in scales that occurs in logistic regression when variables are added to the model (20, 22, 23). As a result, negative effects become more negative, and positive effects become more positive. Thus, relying on the change-in-estimate might lead to wrong conclusions about the presence and magnitude of confounding bias (19).

Using a Monte Carlo simulation study, in chapter 4 of this thesis we found that depending on the sign and magnitude of the confounding bias and the noncollapsibility effect, the change-in-estimate may under- or overestimate the magnitude of the confounding bias. Because of the noncollapsibility effect, multivariable regression analysis and IPW – two often used confounder-adjustment methods – return different but both valid estimates of the confounder-adjusted exposure effect. Multivariable regression analysis results in a conditional exposure effect estimate (24, 25), whereas IPW results in a population-average exposure effect estimate (24-27). It is often suggested to report a population-average effect if the target population is the entire study population, while a conditional exposure effect should be reported if the target population is a subset of the study population (20, 22, 24-26, 28, 29). Although the exact differences between the effect

estimates and their respective interpretations remain unclear, it is important to consider these differences in the interpretation of the results as the populations about which inferences are made differ from each other. If one is unaware of the fact that multivariable regression analysis and IPW result in different exposure effect estimates with their own conclusions, conditional effects may be interpreted as population-average effects, and vice versa. Therefore, researchers should inform their choice for a confounder-adjustment method based on whether they are interested in conditional or population-average effects.

Noncollapsibility effects do not only occur in logistic regression: similar to the odds ratio, the hazard ratio also suffers from noncollapsibility. As a result, the change-in-estimate criterion should not be used to determine the presence of relevant confounding in Cox regression either, and the population-average hazard ratio differs from the conditional hazard ratio (30).

To quantify confounding bias, one could look at the difference between the unadjusted and IPW confounder-adjusted exposure effect estimates (20, 31). If noncollapsibility is not taken into account, this could lead to wrong conclusions about the magnitude and direction of confounding bias. Then, researchers may unnecessarily adjust for certain variables in the analysis, or fail to adjust for variables that explain part of the exposure effect, eventually leading to wrong conclusions about the magnitude and direction of the exposure effect.

To identify confounders it is generally recommended to determine the confounder set based on subject matter knowledge rather than on statistical methods. However, a recent review of studies in major epidemiological journals found that only 50% chose confounders based on prior knowledge, whereas 24% used data driven methods to select confounders (9). Thus, confounder selection based on the data is still common in epidemiological research. The same review found that the change-in-estimate criterion was the most popular data-driven method for confounder selection, which is attributed to the fact that the change-in-estimate criterion is recommended in many epidemiologic textbooks and articles (32).

Causal mediation analysis

Whereas a confounder does not lie in the causal pathway of the exposure on the outcome, a mediator does. With mediation analysis, the total effect of the exposure on

the outcome can be decomposed into an indirect effect through the mediator and a direct effect after removing the influence of the mediator. While traditional mediation analysis defines and estimates the mediation effects in terms of regression coefficients, causal mediation analysis separates the causal effect definitions from the effect estimation (33, 34). In our study, we reviewed the regression-, simulation-, imputation- and weighting-based approaches to perform causal mediation analysis. If the mediator and outcome are both continuous, then all estimation approaches provide the same causal effect estimates (35). This is not necessarily the case if the exposure is continuous and the mediator is binary. In this situation, the estimates from the regression- and simulation-based approaches depend on the chosen causal contrast (i.e., the two compared values for the exposure) (33, 36, 37). The imputation- and weighting-based approaches, on the other hand, still provide mediation effect estimates that are the same for every one unit difference in the continuous exposure variable (38). This is also reflected in the interpretation of the results: the indirect effects from the regression- and simulation-based approaches only apply to the two values selected for the causal contrast, whereas the imputation- and weighting-based approaches return average differences in the outcome for every one unit difference in the exposure through the mediator.

In chapter 5 of this thesis, we demonstrated that the differences between 1) the regression- and simulation-based approaches and 2) the imputation- and weighting-based approaches are explained by finite sample bias, meaning that bias decreases as sample size increases. The differences between the effect estimates obtained by the regression- and simulation-based approaches, and by the imputation- and weighting-based approaches in our empirical data example were thus explained by finite sample bias. The empirical data example also illustrated the importance of selecting the causal contrast based on substantive knowledge.

If researchers are unaware of the difference between the estimation approaches and the role of the causal contrast in the regression- and simulation-based approaches, then the mediation effect estimates may be interpreted incorrectly. It is therefore recommended that researchers inform their choice for an estimation method based on whether they are interested in average effects or in effects that correspond to specific exposure values. For the regression- and simulation-based approaches, failing to consider the correct causal contrast may lead to an over- or underestimation of the true indirect effect for an individual with certain exposure values.

Although in the past decade causal mediation analysis gained in popularity, a recent scoping review found that most studies (70.7%) still apply traditional mediation analysis (39-41). In traditional mediation analysis, the indirect effect is defined and estimated using the product-of-coefficients method or the difference-in-coefficients method (42). With the product-of-coefficients method, the indirect effect is calculated as the product of the exposure-mediator and mediator-outcome effect, while with the difference-in-coefficients method the indirect effect is calculated as the difference between the total exposure-outcome effect and the direct exposure-outcome effect adjusted for the mediator. These methods provide the same indirect effect estimates if the outcome and the mediator are both continuous (43, 44). However, if the mediator is a binary variable and the exposure-mediator effect is estimated using logistic regression, then this is no longer the case. In this situation, the product-of-coefficients method should not be used to estimate the indirect effect. This is due to a mismatch in the scales on which the effects are estimated (i.e., the exposure-mediator effect is estimated on the log-odds scale, whereas the mediator-outcome effect is estimated using a linear model) (45, 46). Because this mismatch in scales does not occur with the difference-in-coefficients method (i.e., the total exposure-outcome effect and the direct exposure-outcome effect adjusted for the mediator are estimated on the same scale), this method provides indirect effect estimates similar to the imputation- and weighting-based approaches. However, for models with a binary or time-to-event outcome that are analyzed using logistic- or Cox regression, the difference in coefficients may not only reflect the indirect effect but also a noncollapsibility effect (22, 47, 48). Like the change-in-estimate, the difference-in-coefficients is computed as the difference between nested regression models. Failing to take a possible noncollapsibility effect into account may result in biased conclusions about the magnitude of the indirect effect. Rijnhart et al. advised to use the potential outcomes framework or the product-of-coefficients method to estimate the indirect effect when mediation analysis is based on logistic regression analysis (49).

Of the studies included in the review, only 13.2% used causal mediation analysis, and of those studies most (more than 70%) used the regression- and simulation-based approach (39). It has been recommended that, to ensure a causal interpretation of the mediation effects, researchers apply causal mediation analysis. In addition, the uptake of causal mediation analysis could be enhanced through tutorial papers (39, 50). With our study on the influence of the estimation approaches and the chosen causal contrast on the mediation effect estimates and their interpretations we hope to have provided

researchers with such a tutorial. Valente et al. provide software code of causal mediation analysis in software programs commonly used by epidemiologists (35).

Competing risk analysis

Survival data is often encountered in epidemiologic studies. With survival data, the time till the occurrence of the event of interest is taken into account. Competing events (i.e., events that prevent the event of interest from happening) are an important feature of survival data (51), but are often ignored and individuals that experience a competing event get censored. Conventional methods used in the analysis of survival data such as Cox regression make the assumption of independent or noninformative censoring, meaning that individuals who are censored have the same future risk of the event of interest as the individuals that remain under observation (52, 53). Naturally, censoring individuals that experience a competing event violates this assumption, and failing to account for competing risks generally results in an overestimation of the true effect of the exposure on the outcome (52, 54-58). In chapter 6 of this thesis, we illustrated that, in the presence of competing risks, the cumulative incidence should be estimated using the cumulative incidence function (CIF) instead of the Kaplan-Meier method. To answer etiologic research questions, cause-specific hazard regression could be used, whereas subdistribution hazard regression could be used to answer prognostic research questions (51, 56, 58-60).

The extent to which the cumulative incidence is overestimated if competing risks are ignored is related to the proportion of individuals experiencing the event of interest and the competing event. In our study we illustrated the methods using a geriatric population, in which the proportion of individuals experiencing the competing event (i.e., death before the onset of depression) was high compared to the proportion of individuals experiencing the event of interest (i.e., incident depression). As a result, the cumulative incidence was greatly overestimated using marginal analysis methods. Because of the older age and comorbidities, the competing risk of death is especially high in geriatric study populations (55, 58). When mortality is high, such as in geriatric populations, the overestimation of the cumulative incidence of the event of interest may be substantial. As successful improvements in health care for older adults partly relies on accurate reporting of the incidence and predictors of disease (55), it is important that the competing risk of death is accounted for by applying specific competing risk analysis. Ignoring competing events could, for example, lead to overtreatment in future patients (58, 61). In 2012, Koller et al. examined how competing risk issues were treated in high-

impact medical journals (58). They selected 50 articles in which competing risks were present. In only 20% of the studies specific competing risk methodology was applied. This shows that a better recognition and understanding of competing events and the importance of applying competing risk analysis is needed.

Although there is a clear distinction between cause-specific hazard regression and subdistribution hazard regression, it is recommended to fit models for both the event of interest and the competing event, and to apply both regression techniques for complete understanding. In addition, it is advised to use clear terminology to avoid confusion about the hazard (cause-specific versus subdistribution) presented (62, 63).

Simulation studies

In chapters 2, 4 and 5 of this thesis, Monte Carlo simulation studies were used. Simulation studies allow for the assessment of the performance of a method in relation to the 'true' effect. This way, bias can be quantified and expressed, among other things, in terms of *absolute* and *relative* bias (64, 65). Other performance measures that are often used are accuracy and coverage. Collins et al. emphasized the importance of examining multiple performance measures, as results may vary across measures (66). Accuracy is often expressed in terms of the mean squared error, which incorporates both bias and variability. Coverage is the proportion of times the confidence interval contains the 'true' effect. For 95% confidence intervals, the simulated confidence intervals should contain the 'true' effect in approximately 95% of the samples. Over-coverage suggests that the results are conservative, whereas under-coverage leads to incorrect significant results (66). Because in simulation studies the 'true' effect is known, statistical methods can be compared to each other under different scenarios. Subsequently, statements can be made about which method is best to use under which circumstances.

In 2006, Burton et al. conducted a small review of articles that contained simulation studies (64). They concluded that the majority of the articles did not provide sufficient details to allow for exact replication of the simulation study. To enable the results to be reproduced, studies should include details of all simulation steps and procedures, including justification for the choices made. Most epidemiological journals actively encourage authors to make software code for the simulation study available and require the inclusion of a data availability statement in articles. The code for the simulation studies in this thesis are included in the appendices, which allows for the replication of

our studies. Detailed tutorials on the design, analysis, reporting and presentation of simulation studies can be found elsewhere (64, 65).

Directed Acyclic Graphs

In some chapters of this thesis, directed acyclic graphs (DAGs) are used to illustrate the assumed relations among variables. DAGs are causal diagrams: an arrow connecting two variables indicates that there is a causal relation. Using DAGs, researchers can determine how an exposure-outcome effect may change when adjusting for different covariates, and thus which variables to adjust for (67, 68). In addition, DAGs can be used to distinguish between a confounder, a mediator and a collider (69). Whereas confounding requires the application of confounder-adjustment methods to obtain unbiased results, adjusting for colliders introduces bias (69, 70) and adjusting for mediators results in direct exposure-outcome effect estimates (71). Moreover, DAGs are not bound by the data available, i.e., DAGs can also contain unmeasured variables. They therefore also provide insight into any residual confounding by confounders that are not included in the statistical model.

DAGs have been increasingly popular in health research but reporting is often inconsistent. Tennant et al. provide several recommendations to improve the transparency and utility of DAGs in future research (72).

Because the DAGs in this thesis only contain the variables that were included in the empirical data examples, they are simplified representations of the relations between the variables. In reality, the relations will be more complex, and the actual DAGs will contain more confounders, mediators or colliders.

Concluding remarks

Although regression models are commonly used in epidemiological research to estimate exposure effects, researchers often do not consider the many different ways in which bias can occur. In this thesis, we reviewed four different potential sources of bias in regression analysis, and we proposed solutions where possible. For each topic, the theory was illustrated using an empirical data example and, if applicable, simulation code was provided to reinforce understanding. To avoid bias, it is recommended that researchers consider the potential sources in the pre-analysis phase. This includes, for example, the type of effect they are interested in, the functional form of associations and the presence of competing risks in survival data. If necessary, researchers should adapt

their analysis, for example by explicitly modelling non-linear associations or by applying specific competing risk analysis. In addition, it is recommended to transparently report the measures taken to reduce bias and to carefully interpret the results, taking any remaining bias into consideration. Transparent reporting includes facilitating reproducibility by making software code available to readers and fellow researchers. Finally, I believe that the field of epidemiology would benefit from more non-technical and non-mathematical papers on advanced topics, as I aimed to contribute to with this thesis.

References

1. Brownson RC, Chriqui JF, Stamatakis KA. Understanding Evidence-Based Public Health Policy. American Journal of Public Health. 2009;99(9):1576-83.
2. Altman DG, Royston P. What do we mean by validating a prognostic model? Statistics in Medicine. 2000;19(4):453-73.
3. Groenwold RHH, Klungel OH, Altman DG, van der Graaf Y, Hoes AW, Moons KGM. Adjustment for continuous confounders: an example of how to prevent residual confounding. Canadian Medical Association Journal. 2013;185(5):401.
4. Müllner M, Matthews H, Altman DG. Reporting on Statistical Methods To Adjust for Confounding: A Cross-Sectional Survey. Annals of Internal Medicine. 2002;136(2):122-6.
5. Vandenbroucke JP, von Elm E, Altman DG, Gøtzsche PC, Mulrow CD, Pocock SJ, et al. Strengthening the Reporting of Observational Studies in Epidemiology (STROBE): Explanation and Elaboration. Epidemiology. 2007;18(6).
6. von Elm E, Altman Dg Fau - Egger M, Egger M Fau - Pocock SJ, Pocock Sj Fau - Gøtzsche PC, Gøtzsche Pc Fau - Vandenbroucke JP, Vandenbroucke JP. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. (0895-4356 (Print)).
7. Groenwold RHH, Van Deursen AMM, Hoes AW, Hak E. Poor Quality of Reporting Confounding Bias in Observational Intervention Studies: A Systematic Review. Annals of Epidemiology. 2008;18(10):746-51.
8. Pouwels KB, Widyakusuma NN, Groenwold RHH, Hak E. Quality of reporting of confounding remained suboptimal after the STROBE guideline. Journal of Clinical Epidemiology. 2016;69:217-24.
9. Talbot D, Massamba VK. A descriptive review of variable selection methods in four epidemiologic journals: there is still room for improvement. European Journal of Epidemiology. 2019;34(8):725-30.
10. Hemkens LG, Ewald H, Naudet F, Ladanie A, Shaw JG, Sajeev G, et al. Interpretation of epidemiologic studies very often lacked adequate consideration of confounding. Journal of Clinical Epidemiology. 2018;93:94-102.
11. Lee H, Cashin AG, Lamb SE, Hopewell S, Vansteelandt S, VanderWeele TJ, et al. A Guideline for Reporting Mediation Analyses of Randomized Trials and Observational Studies: The AGReMA Statement. JAMA. 2021;326(11):1045-56.
12. Harrel FE. Regression Modeling Strategies With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis: Springer International Publishing AG; 2015.

13. Lash TL, VanderWeele TJ, Haneuse S, Rothman KJ. Modern Epidemiology. 4 ed: Wolters Kluwer; 2020.
14. Bennette C, Vickers A. Against quantiles: categorization of continuous variables in epidemiologic research, and its discontents. BMC Medical Research Methodology. 2012;12(1):21.
15. Royston P, Ambler G, Sauerbrei W. The use of fractional polynomials to model continuous risk variables in epidemiology. International Journal of Epidemiology. 1999;28(5):964-74.
16. de Boor CR. A Practical Guide to Splines: Springer-Verlag New York; 1978.
17. Durrleman S, Simon R. Flexible regression models with cubic splines. Statistics in Medicine. 1989;8(5):551-61.
18. Smith PL. Splines As a Useful and Convenient Statistical Tool. The American Statistician. 1979;33(2):57-62.
19. Miettinen OS, Cook EF. Confounding: essence and detection. American Journal of Epidemiology. 1981;114(4):593-603.
20. Pang M, Kaufman JS, Platt RW. Studying noncollapsibility of the odds ratio with marginal structural and logistic regression models. Stat Methods Med Res. 2016;25(5):1925-37.
21. Kleinbaum DG, Sullivan KM, Barker ND. A Pocket Guide to Epidemiology: Springer Science + Business Media, LLC; 2007.
22. Mood C. Logistic Regression: Why We Cannot Do What We Think We Can Do, and What We Can Do About It. European Sociological Review. 2009;26(1):67-82.
23. Greenland S, Robins JM. Identifiability, exchangeability, and epidemiological confounding. Int J Epidemiol. 1986;15(3):413-9.
24. Daniel R, Zhang J, Farewell D. Making apples from oranges: Comparing noncollapsible effect estimators and their standard errors after adjustment for different covariate sets. Biom J. 2020.
25. Hernan MA, Clayton D, Keiding N. The Simpson's paradox unraveled. Int J Epidemiol. 2011;40(3):780-5.
26. Breen R, Karlson KB, Holm A. Total, Direct, and Indirect Effects in Logit and Probit Models. Sociological Methods & Research. 2013;42(2):164-91.
27. Burgess S. Estimating and contextualizing the attenuation of odds ratios due to non collapsibility. Communications in Statistics - Theory and Methods. 2016;46(2):786-804.
28. Karlson KB, Popham F, Holm A. Marginal and Conditional Confounding Using Logits. Sociological Methods & Research. 2021:0049124121995548.

29. Zhang Z. Estimating a Marginal Causal Odds Ratio Subject to Confounding. Communications in Statistics - Theory and Methods. 2008;38(3):309-21.
30. Martinussen T, Vansteelandt S. On collapsibility and confounding bias in Cox and Aalen regression models. Lifetime Data Analysis. 2013;19(3):279-96.
31. Janes H, Dominici F, Zeger S. On quantifying the magnitude of confounding. Biostatistics. 2010;11(3):572-82.
32. Talbot D, Diop A, Lavigne-Robichaud M, Brisson C. The change in estimate method for selecting confounders: A simulation study. Statistical Methods in Medical Research. 2021;30(9):2032-44.
33. Imai K, Keele L, Tingley D. A general approach to causal mediation analysis. Psychol Methods. 2010;15(4):309-34.
34. Muthén BO, Muthén LK, Asparouhov T. Regression and Mediation Analysis using Mplus. Los Angeles, CA: Muthén & Muthén; 2017.
35. Valente MJ, Rijnhart JJM, Smyth HL, Muniz FB, MacKinnon DP. Causal Mediation Programs in R, Mplus, SAS, SPSS, and Stata. Struct Equ Modeling. 2020;27(6):975-84.
36. VanderWeele TJ. Explanation in Causal Inference: Methods for Mediation and Interaction: Oxford University Press; 2015.
37. Valeri L, VanderWeele TJ. Mediation Analysis Allowing for Exposure–Mediator Interactions and Causal Interpretation: Theoretical Assumptions and Implementation With SAS and SPSS Macros. Psychological Methods. 2013;18(2):137-50.
38. Steen J, Loeys T, Moerkerke B, Vansteelandt S. medflex: An R Package for Flexible Mediation Analysis using Natural Effect Models. Journal of Statistical Software. 2017;76(11).
39. Rijnhart JJM, Lamp SJ, Valente MJ, MacKinnon DP, Twisk JWR, Heymans MW. Mediation analysis methods used in observational research: a scoping review and recommendations. BMC Medical Research Methodology. 2021;21(1):226.
40. Vo T-T, Superchi C, Boutron I, Vansteelandt S. The conduct and reporting of mediation analysis in recently published randomized controlled trials: results from a methodological systematic review. Journal of Clinical Epidemiology. 2020;117:78-88.
41. Rizzo RRN, Cashin AG, Bagg MK, Gustin SM, Lee H, McAuley JH. A Systematic Review of the Reporting Quality of Observational Studies That Use Mediation Analyses. Prevention Science. 2022.

42. VanderWeele TJ. Mediation Analysis: A Practitioner's Guide. Annual Review of Public Health. 2016;37(1):17-32.
43. MacKinnon DP, Valente MJ, Gonzalez O. The Correspondence Between Causal and Traditional Mediation Analysis: the Link Is the Mediator by Treatment Interaction. Prevention Science. 2020;21(2):147-57.
44. Rijnhart JJM, Twisk JWR, Chinapaw MJM, de Boer MR, Heymans MW. Comparison of methods for the analysis of relatively simple mediation models. Contemporary Clinical Trials Communications. 2017;7:130-5.
45. Rijnhart JJM, Valente MJ, Smyth HL, MacKinnon DP. Statistical Mediation Analysis for Models with a Binary Mediator and a Binary Outcome: the Differences Between Causal and Traditional Mediation Analysis. Prevention Science. 2021.
46. Li Y, Schneider JA, Bennett DA. Estimation of the mediation effect with a binary mediator. Statistics in Medicine. 2007;26(18):3398-414.
47. MacKinnon DP, Lockwood CM, Brown CH, Wang W, Hoffman JM. The intermediate endpoint effect in logistic and probit regression. Clinical Trials. 2007;4(5):499-513.
48. Jiang Z, VanderWeele TJ. When Is the Difference Method Conservative for Assessing Mediation? American Journal of Epidemiology. 2015;182(2):105-8.
49. Rijnhart JJM, Twisk JWR, Eekhout I, Heymans MW. Comparison of logistic-regression based methods for simple mediation analysis with a dichotomous outcome variable. BMC Medical Research Methodology. 2019;19(1):19.
50. Vo T-T, Cashin A, Superchi C, Tu PHT, Nguyen TB, Boutron I, et al. Quality assessment practice in systematic reviews of mediation studies: results from an overview of systematic reviews. Journal of Clinical Epidemiology. 2022;143:137-48.
51. Geskus RB. Data Analysis with Competing Risks and Intermediate States. Boca Raton, FL: Taylor & Francis Group, LLC; 2016.
52. Putter H, Fiocco M, Geskus RB. Tutorial in biostatistics: competing risks and multi-state models. Stat Med. 2007;26(11):2389-430.
53. Austin PC, Lee DS, Fine JP. Introduction to the analysis of survival data in the presence of competing risks. Circulation. 2016;133(6):601-9.
54. Satagopan JM, Ben-Porat L, Berwick M, Robson M, Kutler D, Auerbach AD. A note on competing risks in survival data analysis. Br J Cancer. 2004;91(7):1229-35.
55. Berry SD, Ngo L, Samelson EJ, Kiel DP. Competing risk of death: an important consideration in studies of older adults. J Am Geriatr Soc. 2010;58(4):783-7.

56. Lau B, Cole SR, Gange SJ. Competing risk regression models for epidemiologic data. Am J Epidemiol. 2009;170(2):244-56.
57. Wolbers M, Koller MT, Witteman JC, Steyerberg EW. Prognostic models with competing risks: methods and application to coronary risk prediction. Epidemiology. 2009;20(4):555-61.
58. Koller MT, Raatz H, Steyerberg EW, Wolbers M. Competing risks and the clinical community: irrelevance or ignorance? Stat Med. 2012;31(11-12):1089-97.
59. Andersen PK, Geskus RB, de Witte T, Putter H. Competing risks in epidemiology: possibilities and pitfalls. Int J Epidemiol. 2012;41(3):861-70.
60. Fine JP, Gray RJ. A proportional hazards model for the subdistribution of a competing risk. Journal of the American Statistical Association. 1999;94(446):496-509.
61. Steyerberg EW. Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating: Springer-Verlag New York; 2009.
62. Latouche A, Allignol A, Beyersmann J, Labopin M, Fine JP. A competing risks analysis should report results on all cause-specific hazards and cumulative incidence functions. J Clin Epidemiol. 2013;66(6):648-53.
63. Wolkewitz M, Cooper BS, Bonten MJM, Barnett AG, Schumacher M. Interpreting and comparing risks in the presence of competing events. BMJ : British Medical Journal. 2014;349:g5060.
64. Burton A, Altman DG, Royston P, Holder RL. The design of simulation studies in medical statistics. Statistics in Medicine. 2006;25(24):4279-92.
65. Morris TP, White IR, Crowther MJ. Using simulation studies to evaluate statistical methods. Statistics in Medicine. 2019;38(11):2074-102.
66. Collins LM, Schafer JL, Kam C-M. A comparison of inclusive and restrictive strategies in modern missing data procedures. Psychological Methods. 2001;6(4):330-51.
67. Shrier I, Platt RW. Reducing bias through directed acyclic graphs. BMC Medical Research Methodology. 2008;8(1):70.
68. Hernan MA, Robins JM. Causal Inference: What If. Boca Raton: Chapman & Hall/CRC; 2020.
69. Pearl J. Causality. 2 ed: Cambridge University Press; 2009.
70. 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. 2010;39(2):417-20.

71. Schisterman EF, Cole SR, Platt RW. Overadjustment bias and unnecessary adjustment in epidemiologic studies. Epidemiology (Cambridge, Mass). 2009;20(4):488-95.
72. Tennant PWG, Murray EJ, Arnold KF, Berrie L, Fox MP, Gadd SC, et al. Use of directed acyclic graphs (DAGs) to identify confounders in applied health research: review and recommendations. International journal of epidemiology. 2021;50(2):620-32.

English summary

Background

Epidemiologists are generally interested in the effect of an exposure on an outcome. This so-called exposure effect is often estimated using regression analysis, in which the outcome is regressed on the exposure. The distribution of the outcome determines which regression technique is most appropriate to estimate the exposure effect. In epidemiological research, linear- (for continuous outcomes), logistic- (for binary outcomes) and Cox regression (for survival outcomes) are most commonly applied.

In general, the aim is to isolate the true effect of the exposure on the outcome. However, often the association between an exposure and an outcome is not entirely attributable to the exposure, i.e., the effect is *biased*. If this bias is not accounted for, then the estimated effect is not a good representation of the true underlying effect. This could, for instance, result in under- or overtreatment in patients and influence clinical decision making.

Aim

In this thesis, I provide non-technical and non-mathematical descriptions of various situations in which bias can occur in regression analysis and propose solutions where possible. I focus on four potential sources of bias: the estimation of non-linear effects, noncollapsibility, causal mediation analysis and competing risks. In each chapter the theory is illustrated using an empirical data example from the Longitudinal Aging Study Amsterdam or the Amsterdam Growth and Health Longitudinal Study. Some chapters additionally contain a simulation study to evaluate model performance and compare methods.

Non-linear effects

A principal assumption of linear-, logistic- and Cox regression is that the exposure is linearly related to the outcome. If this assumption is violated, then the effect estimate is not a good representation of the true underlying effect and bias is introduced. Many researchers are unaware that, when adjusting for confounding, this linearity assumption no longer only applies to the exposure-outcome effect, but also to the confounder-exposure or confounder-outcome associations, depending on the confounder-adjustment method used. Chapter 2 shows that if the functional form (i.e., the shape) of these associations is misspecified, bias might be introduced in an attempt to remove bias. Four commonly used confounder-adjustment methods were reviewed: multivariable

regression analysis, covariate adjustment using the propensity score (PS), inverse probability weighting and double robust estimation.

Multivariable regression analysis requires correct specification of the confounder-outcome association, whereas inverse probability weighting requires correct specification of the confounder-exposure association. Covariate adjustment using the PS requires correct specification of both the confounder-exposure and PS-outcome association, while double robust estimation requires correct specification of only one of these associations. The amount of bias introduced if the functional form is not correctly specified depends on the method used, the strength of the confounder-exposure and confounder-outcome associations, and the sample size.

If the linearity assumption does not hold, then the non-linear associations present in the data have to be modelled explicitly in order to obtain unbiased effects. Chapter 3 compares general methods to deal with non-linearity such as the use of higher order terms and categorization of the exposure variable to spline-based methods such as linear spline (LSP) and restricted cubic spline (RCS) regression. Spline functions are transformations of the continuous independent variable: the variable is divided in multiple intervals, and for each interval the association between that variable and the outcome is estimated separately. With LSP, a linear relationship is modelled for each interval, whereas with RCS a third degree relationship is modelled. Compared with general methods, spline models are flexible and result mostly in greater explained variance. If one is interested in reporting the association between the exposure and the outcome, then LSP models provide easier interpretations than RCS models. If one is interested in predicting the outcome based on specific values of the exposure, then RCS models may be preferred. Spline functions can be applied in all kinds of regression models and are implemented in most software packages commonly used by epidemiologists.

Noncollapsibility

To identify confounders, researchers often compare exposure effect estimates between univariable- and multivariable regression models, using an arbitrary threshold to indicate whether a variable is a confounder. Chapter 4 shows that when applied to logistic regression coefficients, this change-in-estimate criterion may lead to wrong conclusions due to a statistical phenomenon called noncollapsibility. This noncollapsibility effect stems from a change in scales that occurs when variables are added to the model. As a

result, the difference between univariable- and multivariable exposure effect estimates may not only represent confounding bias but also a noncollapsibility effect. Depending on the sign and magnitude of the confounding bias and the noncollapsibility effect, the change-in-estimate may under- or overestimate the magnitude of confounding bias. Because of the noncollapsibility effect, multivariable regression analysis and inverse probability weighting return different but valid estimates of the confounder-adjusted exposure effect, with their own respective interpretations. Ideally the set of confounders is determined in the study design phase and based on subject-matter knowledge. To quantify confounding bias, one could compare the unadjusted exposure effect estimate and the estimate from an inverse probability weighted model.

Causal mediation analysis

A mediator explains the effect of the exposure on the outcome, as the exposure causes the mediator, and the mediator in turn causes the outcome. Instead of adjusting for a mediator, mediation analysis can be used to decompose the total effect of the exposure on the outcome into an indirect effect through the mediator and a direct effect after removing the influence of the mediator.

With causal mediation analysis, the causal mediation effects can be estimated using different approaches, including regression, simulation, imputation and weighting. Chapter 5 shows that, if the exposure is continuous and the mediator is binary, then the different estimation approaches do not provide the same effect estimates. For these models, the regression- and simulation-based approaches require the selection of a causal contrast, i.e., the values chosen for the exposure. As a result, the regression- and simulation-based approaches return effects that correspond to specific exposure values, whereas the imputation- and weighting-based approaches return overall effects. If researchers are unaware of the differences between the approaches and the role of the causal contrast in the regression- and simulation-based approaches, then the mediation effect estimates may be interpreted incorrectly.

Competing risks

Conventional methods for the analysis of survival data make the assumption of independent or noninformative censoring, meaning that individuals who are censored have the same future risk of the event of interest as individuals that remain under observation. This assumption is not met if individuals who experience a competing event, i.e., an event that prevents the event of interest from happening, are censored. Therefore,

competing risk analysis should be applied to analyse survival data in the presence of competing risks.

Chapter 6 shows that, in the presence of competing risks, the cumulative incidence should be estimated using the cumulative incidence function (CIF). Using marginal methods such as the Kaplan-Meier method results in an overestimation of the cumulative incidence. The extent to which the cumulative incidence is overestimated is related to the proportion of individuals that experience the event of interest and the competing event. To answer etiologic and prognostic research questions, cause-specific hazard regression and subdistribution hazard regression can be used. In cause-specific hazard regression individuals that experience a competing event are removed from the risk set, whereas they remain in the risk set in subdistribution hazard regression. As a result, the cause-specific hazard is quantified among individuals that are at risk of developing the event of interest, but the subdistribution hazard has no straightforward interpretation. Therefore, the subdistribution hazard should only be used to estimate the incidence of the event of interest taking the competing risks into account. Dealing with competing risks requires careful formulation of the research question (etiologic vs. prognostic), selection of the appropriate method for data analysis and interpretation of the results. In addition, it is suggested to use both regression models and present the results for all causes for complete understanding.

Conclusion

Although regression models are commonly used in epidemiological research to estimate exposure effects, researchers do often not consider the many different ways in which bias can occur. In this thesis, I reviewed four different potential sources of bias in regression analysis, and proposed solutions where possible. To avoid bias, it is recommended that researchers consider the potential sources in the pre-analysis phase and adapt their analysis if necessary. In addition, it is recommended to transparently report the measures taken to reduce bias and to carefully interpret the results, taking any remaining bias into consideration. Finally, I believe that the field of epidemiology would benefit from more non-technical and non-mathematical papers on advanced topics, as I aimed to contribute to with this thesis.

Nederlandse samenvatting

Achtergrond

Epidemiologen zijn hoofdzakelijk geïnteresseerd in het effect van een determinant op een uitkomst. Dit zogenaamde determinant-uitkomst effect wordt vaak geschat met behulp van regressieanalyse, waarbij de determinant wordt gerelateerd aan de uitkomst. De verdeling van de uitkomst bepaalt welke regressietechniek het meest gepast is om het determinant-uitkomst effect zo nauwkeurig mogelijk te schatten. In epidemiologisch onderzoek worden lineaire (voor continue uitkomsten), logistische (voor dichotome uitkomsten) en Cox regressie (voor survival uitkomsten) het meest toegepast.

Het doel van onderzoek is om het werkelijke effect van de determinant op de uitkomst te isoleren, maar vaak is het verband tussen een determinant en een uitkomst niet volledig toe te schrijven aan de determinant, oftewel, het effect is vertekend. Deze vertekening wordt ook wel *bias* genoemd. Wanneer bias niet volledig geëlimineerd wordt is het geschatte effect geen goede weergave van het werkelijke onderliggende effect. Dit kan bijvoorbeeld leiden tot beïnvloeding van de klinische besluitvorming en onder- of overbehandeling van patiënten.

Doel

In dit proefschrift beschrijf ik op niet-technische en niet-wiskundige wijze verschillende situaties waarin bias kan optreden in regressieanalyse, en draag ik waar mogelijk oplossingen aan om deze bias te voorkomen. De focus ligt op vier mogelijke bronnen van bias: de schatting van niet-lineaire effecten, *noncollapsibility*, causale mediatie-analyse en *competing risks*. In elk hoofdstuk wordt de theorie geïllustreerd aan de hand van data van de Longitudinal Aging Study Amsterdam of van het Amsterdamse Groei en Gezondheids Onderzoek. Sommige hoofdstukken bevatten tevens een simulatiestudie om de prestaties van modellen te evalueren en methoden onderling te vergelijken.

Niet-lineaire effecten

Een belangrijke aanname van lineaire, logistische en Cox regressie is dat de determinant lineair gerelateerd is aan de uitkomst. Wanneer deze aanname wordt geschonden is de effectschatting geen goede weergave van het werkelijke onderliggende effect en wordt er bias geïntroduceerd. Een confounder is een variabele die gerelateerd is aan zowel de determinant als de uitkomst en die niet ligt in het causale pad tussen beiden. Hierdoor vertekent een confounder het determinant-uitkomst effect. Veel onderzoekers zijn zich er bij het corrigeren voor confounding niet van bewust dat deze lineariteitsaanname niet alleen van toepassing is op het determinant-uitkomst effect, maar ook op de confounder-

determinant en confounder-uitkomst associatie. Op welke van beide associaties de aanname van toepassing is, is afhankelijk van de methode die gebruikt wordt om te corrigeren voor confounding. Hoofdstuk 2 laat zien dat wanneer de functionele vorm van deze associaties verkeerd wordt gespecificeerd, bias geïntroduceerd kan worden in een poging om bias te verwijderen. Dit geldt voor vier veelgebruikte methoden om te corrigeren voor confounding: multivariabele regressieanalyse, *covariate adjustment using the propensity score* (PS), *inverse probability weighting* en *double robust estimation*.

Multivariabele regressieanalyse vereist een correcte specificatie van de confounder-uitkomst associatie, terwijl inverse probability weighting een correcte specificatie van de confounder-determinant associatie vereist. Covariate adjustment using the PS vereist een correcte specificatie van zowel de confounder-determinant als de PS-uitkomst associatie, terwijl double robust estimation de juiste specificatie van slechts één van beide associaties vereist. De hoeveelheid bias die wordt geïntroduceerd wanneer de functionele vorm niet correct is gespecificeerd hangt af van de gebruikte methode om te corrigeren voor confounding, de sterkte van de confounder-determinant en confounder-uitkomst associaties en de grootte van de steekproef.

Wanneer de aanname van een lineair verband niet opgaat dienen de niet-lineaire associaties expliciet gemodelleerd te worden om *unbiased* effecten te schatten. Hoofdstuk 3 vergelijkt conventionele methoden om met niet-lineariteit om te gaan, zoals het gebruik van kwadraattermen en het categoriseren van de determinant, met meer geavanceerde methoden zoals *linear spline* (LSP) en *restricted cubic spline* (RCS) regressie. Splinefuncties zijn transformaties van de continue onafhankelijke variabele: deze variabele wordt verdeeld in meerdere intervallen, en voor elk interval wordt de associatie tussen die variabele en de uitkomst afzonderlijk geschat. Bij LSP wordt voor elk interval een lineair verband gemodelleerd, terwijl bij RCS een derdegraads verband wordt gemodelleerd. In vergelijking met conventionele methoden zijn spline-modellen flexibel en resulteren ze doorgaans in een meer nauwkeurige schatting van de relatie tussen de determinant en de uitkomst. Als men geïnteresseerd is in het rapporteren van de associatie tussen de determinant en de uitkomst, dan bieden LSP modellen eenvoudigere interpretaties dan RCS modellen. Als men geïnteresseerd is in het voorspellen van de uitkomst op basis van specifieke waarden van de determinant, dan kunnen RCS modellen de voorkeur hebben. Splinefuncties kunnen worden toegepast in verschillende soorten regressiemodellen en zijn verwerkt in de meeste softwarepakketten die vaak door epidemiologen worden gebruikt.

Noncollapsibility

Om confounders te identificeren vergelijken onderzoekers vaak schattingen van determinant-uitkomstseffecten tussen univariabele en multivariabele regressiemodellen, waarbij een willekeurige drempelwaarde (doorgaans 10%) wordt gebruikt om aan te geven of een variabele een relevante confounder is. Hoofdstuk 4 laat zien dat het gebruik van dit criterium, ook wel het verandering-in-coëfficiënten criterium genoemd, kan leiden tot verkeerde conclusies wanneer het wordt toegepast op logistische regressiemodellen. Dit komt door *noncollapsbility*, een statistisch fenomeen dat voortkomt uit een verandering in de schaal waarop coëfficiënten worden geschat wanneer variabelen aan een logistisch model worden toegevoegd. Als gevolg hiervan is het verschil tussen de univariabele- en multivariabele schattingen van het determinant-uitkomst effect niet alleen een weergave van bias, maar ook van het noncollapsibility effect. Afhankelijk van de richting (m.a.w. positief of negatief) en de omvang van de confounding bias en de grootte van het noncollapsibility effect kan de verandering-in-coëfficiënten de werkelijke hoeveelheid confounding bias onder- of overschatten. Vanwege het noncollapsibility effect leveren multivariabele regressieanalyse en inverse probability weighting verschillende schattingen op van het gecorrigeerde determinant-uitkomst effect. Beide schattingen zijn correct, maar ze verschillen in hun interpretatie. Idealiter wordt de set van confounders bepaald bij het ontwerpen van het onderzoek en wordt deze set gebaseerd op vakinhoudelijke kennis. Om de confounding bias te kwantificeren kan de ongecorrigeerde schatting van het determinant-uitkomst effect vergeleken worden met de schatting van een inverse probability weighted model.

Causale mediatie-analyse

Een mediator is een variabele die het determinant-uitkomst effect kan verklaren, aangezien de determinant van invloed is op de mediator en de mediator op zijn beurt van invloed is op de uitkomst. Mediatie-analyse kan gebruikt worden om het totale effect van de determinant op de uitkomst op te splitsen in een indirect effect via de mediator en een direct effect waarin de invloed van de mediator is weggenomen.

Met causale mediatie-analyse kunnen de causale directe-, indirecte- en totale effecten worden geschat met behulp van verschillende methoden, waaronder regressie, simulatie, imputatie en weging. Hoofdstuk 5 laat zien dat, als de determinant continu is en de mediator dichotoom, deze verschillende methoden verschillende effectschattingen opleveren. Regressie en simulatie vereisen de selectie van een causaal contrast, d.w.z. specifieke determinantwaarden op basis waarvan de effecten geschat worden. Als gevolg

hiervan schatten regressie en simulatie effecten die horen bij deze determinantwaarden, terwijl imputatie en weging algemene effecten schatten. Als onderzoekers zich niet bewust zijn van de verschillen tussen deze methoden en van de rol van het causale contrast bij regressie en simulatie kunnen de mediatie-effecten onjuist worden geïnterpreteerd.

Competing risks

Conventionele methoden voor de analyse van survival data gaan uit van onafhankelijke of niet-informatieve *censoring*. Dit betekent dat personen die gecensored worden hetzelfde toekomstige risico op een bepaalde uitkomst hebben als personen die niet gecensored worden. Aan deze aanname wordt niet voldaan wanneer personen die een *competing event* meemaken, d.w.z. een event dat ervoor zorgt dat de uitkomst niet meer kan optreden, worden gecensored. Om bias te voorkomen kan survival data, in de aanwezigheid van competing risks, geanalyseerd worden met competing risk analyse.

Hoofdstuk 6 laat zien dat, in het geval van competing risks, de cumulatieve incidentie geschat moet worden met behulp van de cumulatieve incidentiefunctie (CIF). Het gebruik van conventionele methoden zoals de Kaplan-Meier methode resulteert in een overschatting van de cumulatieve incidentie. De mate van deze overschatting hangt af van de verhouding tussen individuen die over de tijd de uitkomst ontwikkelen en individuen die een competing event meemaken.

Om etiologische en prognostische onderzoeksvragen te beantwoorden, kunnen *cause-specific hazard* regressie en *subdistribution hazard* regressie gebruikt worden. Bij cause-specific hazard regressie worden de individuen die een competing event meemaken verwijderd uit de studie, terwijl deze bij subdistribution hazard regressie juist deel blijven uitmaken van de studie. Als gevolg hiervan berekent de cause-specific hazard het risico op de uitkomst voor individuen die hier nog gevaar voor lopen, maar heeft de subdistribution hazard geen eenduidige interpretatie. Daarom dient de subdistribution hazard enkel te worden gebruikt om, rekening houdend met de competing risks, de incidentie van de uitkomst te schatten. De aanwezigheid van competing risks in een studie vereist een zorgvuldige formulering van de onderzoeksvraag (etiologisch vs. prognostisch), selectie van de juiste methode om de data te analyseren en een juiste interpretatie van de resultaten. Bovendien wordt aanbevolen om beide regressiemodellen te gebruiken en voor de volledigheid de resultaten voor zowel de uitkomst als de competing events te presenteren.

Conclusie

Hoewel regressiemodellen vaak gebruikt worden in epidemiologisch onderzoek om determinant-uitkomst effecten te schatten houden onderzoekers vaak geen rekening met de vele verschillende manieren waarop bias kan optreden. In dit proefschrift heb ik vier mogelijke bronnen van bias in regressieanalyse beschreven en waar mogelijk oplossingen aangedragen. Om bias te voorkomen wordt aanbevolen dat onderzoekers kritisch nadenken over mogelijke bronnen van bias voordat zij hun data analyseren en zo nodig hun analyse aanpassen. Daarnaast wordt aanbevolen om transparant te rapporteren over de maatregelen die zijn genomen om bias te verminderen en om de resultaten zorgvuldig te interpreteren, daarbij rekening houdend met eventuele resterende bias. Ter afsluiting ben ik er van overtuigd dat de epidemiologie baat zou hebben bij meer niet-technische en niet-wiskundige artikelen over complexe onderwerpen, waar ik met dit proefschrift gepoogd heb aan bij te dragen.

PhD portfolio

Courses	Year	ECTs
Research Integrity, Amsterdam UMC Doctoral School	2021	2
Medische Basiskennis, EpidM	2021	8
Data Processing, University of Amsterdam	2020	6
Scientific Programming 2, University of Amsterdam	2019	3
Scientific Programming 1, University of Amsterdam	2019	3
Regressietechnieken, EpidM	2018	5
Longitudinale Data Analyse, EpidM	2018	3

Conferences and scientific meetings		
WEON 2021, online	2021	1
WEON 2019, Groningen	2019	1
rstudio::conf, Austin	2019	1.14
Intervision meetings, Amsterdam Public Health Research Institute	2018 - 2021	0.5

Supervision and teaching activities		
Supervision of Nine Droog, BSc Health and Life Sciences, "Has the publication of the GRoLTS-checklist improved the reporting of results of latent trajectory analyses?"	2021	1
Supervision of Rob Rekveld, BSc Health and Life Sciences, "The quality of reporting in latent trajectory studies through the years: associations with author-, journal- and study characteristics"	2021	1
Supervision of Sema Atmaca, BSc Health Sciences, "Associatie tussen fluctuaties in fysieke activiteit en lichaamsvetverdeling onder Nederlandse volwassenen"	2020	1
Supervision of Ewa Sillem, BSc Health Sciences, "Daily fluctuations in physical activity duration and its relationship with the need for recovery from work due to work-related fatigue in 42-year old adults"	2020	1
Supervision of Carolien de Visser, BSc Health Sciences, "Associaties fysieke activiteit en fluctuaties fysieke activiteit op slaapkwaliteit onder Nederlandse volwassenen"	2020	1

Teaching activities for EpidM	2018 - 2022	2.2
Teaching activities for the Department of Epidemiology and Data Science	2018 - 2022	4

Other activities

Building Tidy Tools, rstudio::conf	2019
Statistical consulting through E&B Xpert	2021 - 2022
Reviewer for various international journals	2020 - 2022
Member of the WEON 2021 organization committee	2019 - 2021
Member of the APH Methodology Junior Board	2019

List of publications

Schuster, N.A., Rijnhart, J.J.M., Twisk, J.W.R. & Heymans, M.W. (2022). Modelling non-linear relationships in epidemiological data: the application and interpretation of spline models. *Frontiers in Epidemiology*, 2.

Schuster, N.A., Twisk, J.W.R., Ter Riet, G., Heymans, M.W. & Rijnhart, J.J.M. (2021). Noncollapsibility and its role in quantifying confounding bias in logistic regression. *BMC Medical Research Methodology*, 21(1), 136.

Schuster, N.A., De Breij, S., Schaap, L.A., Van Schoor, N.M., Peters, M.J.L. De Jongh, R.T., Huisman, M. & Hoogendijk, E.O. (2021). Older adults report cancellation or avoidance of medical care during the COVID-19 pandemic: results from the Longitudinal Aging Study Amsterdam. *European Geriatric Medicine*, 12(5), 1075-1083.

De Breij, S., Van Hout, H.P.J., De Bruin, S.R., Schuster, N.A., Deeg, D.J.H., Huisman, M. & Hoogendijk, E.O. (2021). Predictors of frailty and vitality in older adults aged 75 years and over: results from the Longitudinal Aging Study Amsterdam. *Gerontology*, 67(1), 69-77.

De Breij, S., Rijnhart, J.J.M., Schuster, N.A., Rietman, M.L., Peters, M.J.L. & Hoogendijk, E.O. (2021). Explaining the association between frailty and mortality in older adults: the mediating role of lifestyle, social, psychological, cognitive, and physical factors. *Preventive Medicine Reports*, 24, 101589.

Twisk, J.W.R., Rijnhart, J.J.M., Hoekstra, T., Schuster, N.A., Ter Wee, M.M. & Heymans, M.W. (2020). Intention-to-treat analysis when only a baseline value is available. *Contemporary Clinical Trials Communications*, 20, 100684.

Schuster, N.A., Hoogendijk, E.O., Kok, A.A.L., Twisk, J.W.R. & Heymans, M.W. (2020). Ignoring competing events in the analysis of survival data may lead to biased results: a nonmathematical illustration of competing risk analysis. *Journal of Clinical Epidemiology*, 122, 42-48.

Pajouheshnia, R., Schuster, N.A., Groenwold, R.H.H., Rutten, F.H., Moons, K.G.M. & Peelen L.M. (2020). Accounting for time-dependent treatment use when developing a prognostic model from observational data: a review of methods. *Statistica Neerlandica*, 74(1), 38-51.

Hoogendijk, E.O., Smit, A.P., Van Dam, C., **Schuster, N.A.**, De Breij, S., Holwerda, T.J., Huisman, M., Dent, E. & Andrew, M.K. (2020). Frailty combined with loneliness or social isolation: an elevated risk for mortality in later life. *Journal of the American Geriatrics Society*, 68(11), 2587-2593.

Accepted for publication

Schuster, N.A., Twisk, J.W.R., Heymans, M.W. & Rijnhart, J.J.M. (2022). Causal mediation analysis with a binary mediator: the influence of the estimation approach and causal contrast. *Structural Equation Modeling: A Multidisciplinary Journal.*

Submitted for publication

Schuster, N.A., Rijnhart, J.J.M., Bosman, L.C., Twisk, J.W.R., Klausch, T. & Heymans, M.W. (2022). Misspecification of confounder-exposure and confounder-outcome associations leads to bias in effect estimates.

Hoogendijk, E.O., **Schuster, N.A.**, Van Tilburg, T.G., Schaap, L.A., Suanet, B., De Breij, S., Kok, A.A.L., Van Schoor, N.M., Timmermans, E.J., De Jongh, R.T., Visser, M. & Huisman, M. (2022). The Longitudinal Aging Study Amsterdam COVID-19 exposure index: a cross-sectional analysis of the impact of the pandemic on daily functioning of older adults.

About the author

Noah Alexandra Schuster was born on May 5th, 1992 in Amsterdam. After attending the Vossius Gymnasium, she went on to study for her bachelor's degree in Health Sciences at VU University in Amsterdam. She spent a semester abroad at Eötvös Loránd University in Budapest, Hungary, following courses from their master's program Health Policy, Planning and Financing. She wrote her bachelor's thesis on fluctuations in physical activity and physical fitness at the Department of Methodology and Applied Biostatistics under the supervision of dr. Trynke Hoekstra. During her studies, Noah rowed for A.A.S.R. Skøll, winning medals in different boat classes at both national and international regattas.

In 2016, Noah went to study Epidemiology at Utrecht University. She graduated with a double specialization in Medical Statistics and Pharmacoepidemiology. During her master's, she completed a 13-month research project under the supervision of dr. Linda Peelen and dr. Romin Pajouheshnia at the Department of Data Science and Biostatistics of the Julius Center for Health Sciences and Primary Care. This resulted in her thesis on approaches to account for time-varying treatment use in the development of prognostic models, which was published in *Statistica Neerlandica*. After this project, she completed another 5-month research project under the supervision of prof.dr. Michael Hauptmann at the Department of Psychosocial Research and Epidemiology of the Netherlands Cancer Institute. This resulted in a systematic review about diagnostic imaging among cancer patients.

In August 2018, Noah started her PhD research on bias in regression analysis at the former Department of Epidemiology and Biostatistics at the VU University Medical Center, now the Department of Epidemiology and Data Science at the Amsterdam University Medical Center, under the supervision of prof.dr. Jos Twisk, dr. Martijn Heymans and dr. Judith Rijnhart. Alongside her PhD, she tutored multiple EpidM courses, served as a statistical consultant and was a member of the WEON 2021 organization committee. As of September 2022, Noah works as a Senior Associate Consultant at Bain & Company.

Dankwoord

Zonder de onvoorwaardelijke en niet aflatende steun van vrienden, collega's en familie was dit proefschrift niet tot stand gekomen. Zij hebben mij de afgelopen jaren aangemoedigd, afgeleid en (tevergeefs) geprobeerd te laten ontspannen (JE MOET ONTSPANNEN!) wanneer dat nodig was. Onderstaande personen wil ik graag in het bijzonder bedanken.

Allereerst mijn promotor, Jos, en mijn copromotor, Martijn. Bedankt voor de vrijheid die jullie me hebben geboden om me te ontwikkelen als onderzoeker, en voor jullie enthousiasme: na overleg met jullie was ik er altijd van overtuigd dat het heus niet zo ingewikkeld was als ik het zelf had gemaakt. Soms weliswaar onterecht, maar voor de moraal deed het wonderen. Ook waren jullie overtuigd van een tijdige afronding van dit proefschrift voordat ik dat zelf was ("Je loopt goed op schema, en dit zeg ik niet om je gerust te stellen. Nee, dit zeg ik wél om je gerust te stellen, maar ik meen het ook echt").

Judith, ik kan niet in woorden uitdrukken hoe dankbaar ik je ben voor je tijd, inzet en betrokkenheid bij mijn proefschrift. Wat begon met een kop koffie in de bibliotheek in Zwolle heeft uiteindelijk geleid tot een heel fijne samenwerking en een aantal artikelen waar ik ontzettend trots op ben. Hoewel het soms even duurde voordat bij mij het kwartje viel (een change in scales?!) bleef jij altijd geduldig en vol vertrouwen. Ik had me de afgelopen vier jaar geen fijnere collega en vriendin, en uiteindelijk ook copromotor, kunnen wensen. Dankjewel! Matt, thank you for your help behind the scenes.

De leden van de leescommissie, prof.dr. Geert van der Heijden, prof.dr. Martijn Huisman, prof.dr. Frank van Lenthe, dr. Michiel de Boer en dr. Trynke Hoekstra wil ik graag hartelijk bedanken voor de tijd die zij hebben genomen om mijn proefschrift te lezen.

Sascha en Emiel, mijn lieve BBK'ers en Vrolikstraathuishouden in barre covid-tijden, zonder jullie hadden de afgelopen jaren er heel anders uitgezien. Sowieso een stuk nuchterder, maar daarmee ook een stuk minder plezierig. Sas, bedankt voor alle competing risk-overleggen, kaasplankjes, duinwandelingen en avonturen in Rome en Antwerpen ("ik heb uw nummer gerecupereerd"), maar vooral voor het feit dat je zo'n lieve en betrokken vriendin bent. Emiel O., dankzij jou zijn de afgelopen jaren tot in detail gedocumenteerd ([disclaimer]). Bedankt voor de bilateraaltjes en al je goede adviezen, zowel proefschriftinhoudelijk als op het gebied van olijfolie, stokbrood, pizza, aardbeien, gin, en natuurlijk Bella Blue (je bent zeer eclectisch!). Jij leerde me dat Goldstrike ook gewoon bij de lunch kan (met bijbehorend ademhalingsritueel) en introduceerde

#intervaldry. Ik ben ontzettend blij dat jullie mijn paranimfen zijn. Lisa, op donkere kelderdagen was jij altijd een lichtpuntje. Bedankt voor je eeuwige optimisme, je goede humeur, je openheid en alle gezelligheid op kantoor en daarbuiten. Silvia, Priyanta, Eva en de rest van de GT-groep, ik heb genoten van alle borrels. Ik ga het nog missen om 's avonds met mijn fiets door het gebouw te lopen als ik weer eens vergeten was dat het hek al eerder sluit. Niels, helaas is het ondanks de algoritmes nooit van caffè gekomen…

Ook de overige LASA collega's wil ik graag bedanken voor hun betrokkenheid en de gezellige lunches en borrels. Marjolein en Yvonne, bedankt voor de fijne samenwerking voor alles rondom EpidM.

Lisa, Emma and Zakile, whilst many people came and went in our Intervision group, you were always there to put my doubts into perspective. Thanks!

Trynke, mijn epidemiologiecarrière begon in 2016 toen ik onder jouw begeleiding mijn bachelorscriptie schreef, en zes jaar later sluit ik 'm af met jou in mijn oppositie. Bedankt voor de ontzettend fijne en leerzame samenwerking. Sema, Ewa, Carolien, Rob en Nine, ik hoop dat jullie net zo veel plezier hebben beleefd aan het schrijven van jullie scripties als ik destijds. Ik vond het bijzonder zo nauw betrokken te zijn en jullie met de week te zien groeien in het onderzoek.

Romin and Linda, I could not have wishes for better and more caring supervisors during my 13-month internship at the Julius Center. I can only hope that I've been able to offer my students the same level of support as you offered me at the time.

Mijn lieve oud-collega's van Chiever, ondanks dat ik de afgelopen jaren niks hoefde op te zoeken in het BBIE-register heb ik veel van wat ik in de zeven jaar bij jullie heb geleerd kunnen toepassen tijdens mijn PhD. Manon, ik hoop dat we in de toekomst weer kunnen samenwerken en dat ik van je kan blijven leren. Joke, ik vond het een genot al die jaren naast je te mogen werken.

Om in epidemiologie-termen te spreken: de relatie tussen mijn fysieke en mentale gezondheid is zeer significant ($p < 0.005$). Franc, bedankt voor alle (fysio-)therapiesessies en je (soms vergeefse) pogingen mij mijn aandacht te laten richten op de dingen die wél goed gaan. Marius & team, zonder jullie was ik gek geworden. Bedankt!

Lieve Viet, je bent natuurlijk maar een hobby-epidemioloog, maar zoals Whitney het zo mooi zegt: I learned from the best. Ik heb bij jou kunnen afkijken hoe het moet en had me de afgelopen jaren geen betere en lievere vriendin, huisgenoot (ondanks dat je vooral op vakantie was…), achtergrondzangeres en sous-chef kunnen wensen. Bedankt voor je eeuwige steun, zelfs vanaf de andere kant van de wereld. Ik ben blij dat ik je straks weer dichtbij me heb! Oscar, thank you for opening your home to me (and sharing Rio's love with me). L'chaim! Jabu, I can't wait to take you to Sin+Tax. Daphne, ondertussen heb ik je al een stuk langer wel in mijn leven dan niet, en daar ben ik ontzettend blij mee. We zijn weliswaar ouder geworden, maar wijs waren we altijd al. Ik prijs mezelf gelukkig dat ik altijd bij je terecht kan voor een luisterend oor, goede adviezen en classy boze brieven. Een eervolle vermelding voor Hans, die met zijn telefonische wiskundebijlessen onbewust de basis legde voor dit proefschrift. Je wordt gemist. Marjo, mijn lieve majoor. Door zon en regen, wind mee of tegen, je schreeuwt me al sinds 2010 naar de finish. Vroeger in de letterlijke zin van het woord, tegenwoordig (gelukkig) slechts nog figuurlijk. Van alle oudjaarsavonden in het buitenland tot ons favoriete plekje in de Marktkantine (rip), met jou is het altijd een feestje. Love you apie! Mariek, bedankt voor de fijne vakanties (met als absoluut hoogtepunt natuurlijk Plüderhausen), alle wijntjes en Aperol Spritz die ik voor je heb moeten opdrinken, alle voicememo's en vlogjes, de wekelijkse kooksessies, de Medische Basiskennisbootcamp, DJ Partyflock en alle circuskunstjes in de sportschool. Door jou realiseer ik me af en toe weer wat voor prestatie zo'n proefschrift eigenlijk is. Tom, bedankt voor de cocktails, je zangkunsten en de 40 heerlijke uren in Londen. Je weet wat ze zeggen he, als het niet goed is…. Lau, mijn IDFA-vriendin, ik houd van onze boekentips, filmbezoekjes, etentjes en fijne gesprekken. Bij jou kan ik altijd terecht als ik het even niet meer weet. Wil, waar die paar maanden aftrainen in Utrecht wel niet goed voor waren. Ik ben blij dat we elkaar nog altijd zien. Bedankt voor al je telefoontjes en goede adviezen. Anne, bedankt voor alle gezelligheid. Ik ben nog nooit zo goed op de hoogte geweest van het wel en wee van BN'ers (en hun aanwezigheid op het Scheldeplein) als in het afgelopen jaar.

John, tijgervader, ik weet dat je er trots op bent dat je niet langer de enige dr. Schuster bent. David, bedankt voor je inspirerend leiderschap (!) en goede adviezen. Van jou leer ik hoe je ondanks belachelijke successen tóch bescheiden kan blijven. Dear Paula, missy, thank you for all your support. Bel, bedankt dat ik altijd bij jou kon komen ontspannen in Rome en Lissabon, en voor alle ongevraagde knuffels. Mar, zoals je moeder altijd zei: je kan niet gek worden wanneer je wil. Ik heb het geprobeerd de afgelopen jaren. Bedankt voor je goede zorgen, niet alleen de afgelopen vier jaar maar ook alle jaren daarvoor.