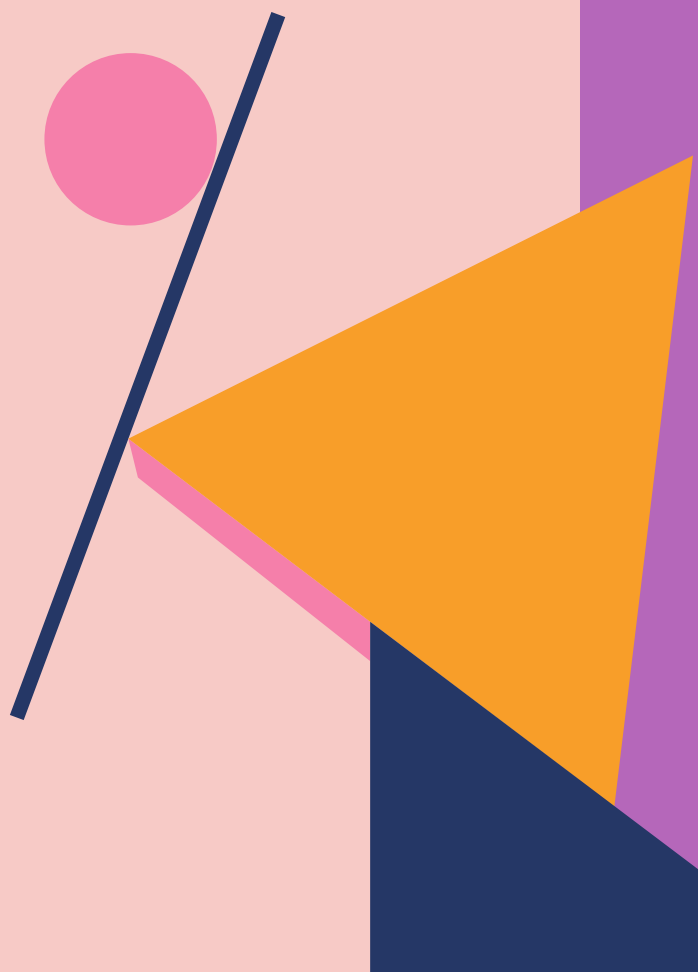
Correction methods for measurement error in epidemiologic research

Nab, L.

correction methods for measurement error in epidemiologic research

Linda Nab



CORRECTION METHODS FOR MEASUREMENT ERROR IN EPIDEMIOLOGIC RESEARCH

Art piece: Jesse Steenkist
Cover design: Lisa Tabak
Printing: Gildeprint

ISBN 978-94-6419-681-8

CORRECTION METHODS FOR MEASUREMENT ERROR IN EPIDEMIOLOGIC RESEARCH

Proefschrift

ter verkrijging van
de graad van doctor aan de Universiteit Leiden,
op gezag van rector magnificus prof. dr. ir. H. Bijl,
volgens besluit van het college voor promoties
te verdedigen op donderdag 26 januari 2023
klokke 13:45

door

Linda Nab

geboren te Zutphen
in 1991

Promotor: Prof. dr. R.H.H. Groenwold
Copromotor: Dr. M. van Smeden — Universitair Medisch Centrum Utrecht

Leden promotiecommissie:
Prof. dr. S.C. Cannegieter
Prof. dr. S. le Cessie
Prof. dr. D.L. Oberski — Utrecht Universiteit en Universitair Medisch Centrum Utrecht
Prof. dr. ir. H.C.W. de Vet — Amsterdam Universitair Medisch Centrum

Contents

1

General introduction and outline of the thesis

1.1. Introduction

Measurement error affects the validity of many epidemiologic studies, that often rely on imperfect data [1]. Epidemiologic studies may for example rely on data obtained from electronic health records. These records are retrieved for other purposes than epidemiologic precision and may therefore be more subject to measurement error than data retrieved for answering specific research questions. Another example of error-prone data includes data collection based on self-reports by study participants [2]. Self-reports may come with (selective) reporting and recollection biases [3]. The inability to accurately measure variables of interest in epidemiologic research studies may result in failure to observe associations between a certain exposure and health outcome [4], or oppositely, the observation of spurious associations [5].

Epidemiologic studies often rely on the salient assumption of no measurement error. This assumption may be satisfied for some variables (e.g., age in years) but much harder to justify for others, such as variables subject to natural variation (e.g., blood pressure) [6] or laboratory error (e.g., Inhibin B) [7]. As an example, Figure 1.1 illustrates the discrepancy between two consecutive measurements of systolic blood pressure in the National Health and Nutrition Examination Survey (NHANES) [8].

Other epidemiologic studies may rely on self-reported measures, such as self-reported length or weight [9], physical activity [10] or diet [11]. A self-reported measure tends to be prone to error and generally does not perfectly correlate with the phenomenon it aims to measure. In Figure 1.2 it is illustrated that in the NHANES [8] self-reported weight was not perfectly correlated with weight measured by trained health technicians with a calibrated weight scale.

When measurement error is not accounted for in the design or the analysis of an epidemiologic study, measurement error can lead to considerable bias in exposure-outcome associations. The consequences of measurement error in exposure and outcome variables have been well established in the scientific literature [12–15]. The triple whammy of

measurement error describes the three consequences of measurement error: i) it may lead to bias in statistical parameter estimation, ii) it may lead to a loss of power, and iii) it may mask the functional form of a relationship between two variables [13]. For the first whammy, a common misconception is that the bias due to measurement error always attenuates exposure-outcome associations. This general statement can be true in case of random measurement error in the exposure, also known as 'classical' measurement error. For other forms of measurement error, e.g., systematic or differential measurement error, this simple heuristic may not apply [16].

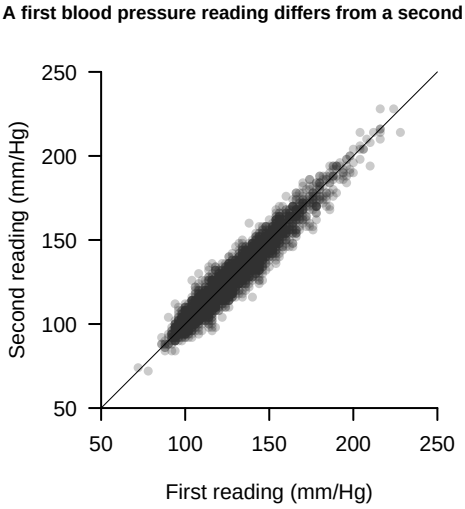


Figure 1.1: Discrepancy between two consecutive systolic blood pressure measurements in the National Health and Nutrition Examination Survey (NHANES) 2017-2018 cycle [8]

Various correction methods for measurement error exist. Examples include regression calibration [17, 18], simulation-extrapolation [19], moment reconstruction [20], non-parametric maximum likelihood estimation [21], imputation-based methods [22, 23], and Bayesian methods [15, 24]. Among these methods, regression calibration appears to be the one that is most commonly used in epidemiology [25, 26].

In spite of the abundance of literature on measurement error, and more specifically, on measurement error correction methods, correction for measurement error remains seldomly applied in epidemiologic research [25–27]. In most epidemiologic studies, the impact of measurement error is inadequately discussed [26] and often erroneously dismissed as leading to an underestimation of the exposure-outcome association [25, 26]. Importantly, this practice has not changed over the last decades [25–27]. This may, in part, be due to insufficient understanding of the impact of measurement error in settings that go beyond the classical example of attenuated exposure-outcome associations. An alternative explanation may be that researchers are unfamiliar with available measurement error correction methods and tools to quantitatively assess the impact of measurement error. In addition, researchers may not appreciate the added value of the collection of

(external) validation data for measurement error correction, hampering the inclusion of additional validation data within study designs when measurement error is suspected or anticipated.

The aim of this thesis is to improve the understanding of the impact of measurement error in epidemiologic studies, to facilitate the application of measurement error correction methods, to improve the design of epidemiologic studies when measurement error in a variable is suspected and to develop tools to quantitatively assess the impact of measurement error in epidemiologic studies.

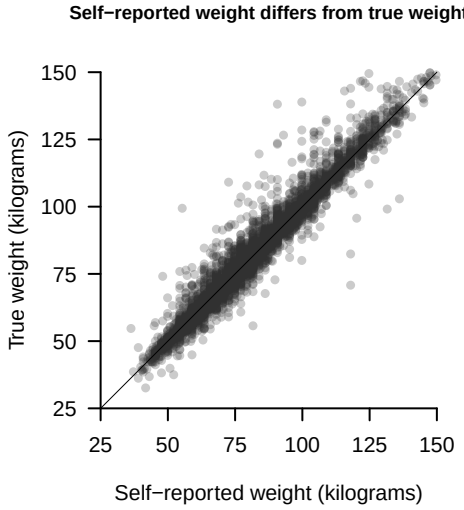


Figure 1.2: Discrepancy between weight in kilograms measured by health technicians using a calibrated weight scale (true weight) and self-reported weight in the National Health and Nutrition Examination Survey (NHANES) 2017-2018 cycle [8]

1.2. Outline

This thesis is organised as follows. To improve the understanding of the impact of measurement error, in Chapter 2, it is investigated how randomised controlled trials are affected by measurement error in a continuous endpoint. Three types of measurement error are distinguished, classical (or random) measurement error, systematic measurement error and differential measurement error.

To improve the application of measurement error correction methods, in Chapter 3 the `R` package `mecor` is described for measurement error correction in linear models with a continuous outcome. The `R` package `mecor` facilitates measurement error correction by means of regression calibration, method of moments and a maximum likelihood-based method. Information about the measurement error model and its parameters can be obtained from four types of validation studies: internal validation, replicates, calibration and external validation data. Each of these are discussed in detail. Chapter 4 provides

an exploration of the bias–variance trade off for the regression calibration estimator implemented in `mecor`, and an investigation of the performance of the estimator in settings where measurement error is relatively large.

To improve the design of epidemiologic studies when measurement error is suspected, guidance is provided for the collection of validation data needed for measurement error correction in Chapter 5. Here, sampling methods for validation data are studied and the assumptions required for the correct application of regression calibration for measurement error correction investigated. Deterministic and non-deterministic methods for validation data sampling are compared in terms of statistical efficiency. Next, in Chapter 6 reporting guidelines are proposed for studies on venous thromboembolism incidence in Corona disease patients. These studies on incidence report highly heterogeneous results. Different clinical and methodological sources of this heterogeneity are identified, including misclassification error in the diagnosis of venous thromboembolism and overall data quality. The proposed reporting guidelines guide future studies on venous thromboembolism incidence.

To quantitatively assess the impact of measurement error in the absence of validation data, sensitivity analysis or quantitative bias analysis could be used. In Chapter 7, two methods, regression calibration and simulation-extrapolation are compared for a sensitivity analysis for random exposure measurement error. In Chapter 8, a quantitative bias analysis for confounder misclassification is proposed. The quantitative bias analysis approach is described for traditional conditional regression and marginal structural models estimated using inverse probability weighting. This thesis ends with a general discussion including recommendations and directions for future research in Chapter 9.

References

[1] K. B. Michels, A renaissance for measurement error, International Journal of Epidemiology 30 (3) (2001) 421–422. doi:10.1093/ije/30.3.421.

[2] J. R. Marshall, Commentary: About that measurement problem, International Journal of Epidemiology 34 (6) (2005) 1376–1377. doi:10.1093/ije/dyi228.

[3] T. Johnson, M. Fendrich, Modeling sources of self-report bias in a survey of drug use epidemiology, Annals of Epidemiology 15 (5) (2005) 381–389. doi:10.1016/j.annepidem.2004.09.004.

[4] J. A. Hutcheon, A. Chiolero, J. A. Hanley, Random measurement error and regression dilution bias, BMJ 340 (c2289) (2010). doi:10.1136/bmj.c2289.

[5] J. R. Marshall, J. L. Hastrup, Mismeasurement and the resonance of strong confounders: Uncorrelated errors, American Journal of Epidemiology 143 (10) (1996) 1069–1078. doi:10.1093/oxfordjournals.aje.a008671.

[6] P. Burstyn, B. O'Donovan, I. Charlton, Blood pressure variability: The effects of repeated measurement, Postgraduate Medical Journal 57 (670) (1981) 488–491. doi:10.1136/pgmj.57.670.488.

[7] N. J. Perkins, J. Weck, S. L. Mumford, L. A. Sjaarda, E. M. Mitchell, A. Z. Pollack, E. F. Schisterman, Combining biomarker calibration data to reduce measurement error, Epidemiology 30 (2019) S3–S9. doi:10.1097/EDE.0000000000001094.

[8] Centers for Disease Control and Prevention (CDC). National Center for Health Statistics (NCHS), National health and nutrition examination survey data (2017). URL https://wwwn.cdc.gov/nchs/nhanes/continuousnhanes/default.aspx?BeginYear=2017

[9] D. Gunnell, L. Berney, P. Holland, M. Maynard, D. Blane, S. Frankel, G. D. Smith, How accurately are height, weight and leg length reported by the elderly, and how closely are they related to measurements recorded in childhood, International Journal of Epidemiology 29 (3) (2000) 456–464. doi:10.1093/intjepid/29.3.456.

[10] P. Ferrari, C. Friedenreich, C. E. Matthews, The role of measurement error in estimating levels of physical activity, American Journal of Epidemiology 166 (7) (2007) 832–840. doi:10.1093/aje/kwm148.

[11] W. Willett, Commentary: Dietary diaries versus food frequency questionnaires—A case of undigestible data, International Journal of Epidemiology 30 (2) (2001) 317–319. doi:10.1093/ije/30.2.317.

[12] J. Buonaccorsi, Measurement error: Models, methods, and applications, Chapman & Hall/CRC, Boca Raton, FL, 2010.

[13] R. Carroll, D. Ruppert, L. Stefanski, C. Crainiceanu, Measurement error in nonlinear models: A modern perspective, 2nd Edition, Chapman & Hall/CRC, Boca Raton, FL, 2006.

[14] W. Fuller, Measurement error models, John Wiley & Sons, New York, NY, 1987.

[15] P. Gustafson, Measurement error and misclassification in statistics and epidemiology: Impacts and Bayesian adjustments, Chapman & Hall/CRC, Boca Raton, FL, 2004.

[16] M. van Smeden, T. L. Lash, R. H. H. Groenwold, Reflection on modern methods: Five myths about measurement error in epidemiological research, International Journal of Epidemiology 49 (1) (2020) 338–347. doi:10.1093/ije/dyz251.

[17] R. J. Carroll, L. A. Stefanski, Approximate quasi-likelihood estimation in models with surrogate predictors, Journal of the American Statistical Association 85 (411) (1990) 652–663. doi:10.1080/01621459.1990.10474925.

[18] L. Gleser, Improvements of the naive approach to estimation in nonlinear errors-in-variables regression models, in: P. Brown, W. Fuller (Eds.), Statistical analysis of measurement error models, American Mathematics Society, Providence, 1990, pp. 99–114.

[19] J. R. Cook, L. A. Stefanski, Simulation-extrapolation estimation in parametric measurement error models, Journal of the American Statistical Association 89 (428) (1994) 1314–1328. doi:10.2307/2290994.

[20] L. S. Freedman, V. Fainberg, V. Kipnis, D. Midthune, R. J. Carroll, A new method for dealing with measurement error in explanatory variables of regression models, Biometrics 60 (1) (2004) 172–181. doi:10.1111/j.0006-341X.2004.00164.x.

[21] S. Rabe-Hesketh, A. Pickles, A. Skrondal, Correcting for covariate measurement error in logistic regression using nonparametric maximum likelihood estimation, Statistical Modelling 3 (3) (2003) 215–232. doi:10.1191/1471082X03st056oa.

[22] S. R. Cole, H. Chu, S. Greenland, Multiple-imputation for measurement-error correction, International Journal of Epidemiology 35 (4) (2006) 1074–1081. doi:10.1093/ije/dyl097.

[23] M. Blackwell, J. Honaker, G. King, A unified approach to measurement error and missing data: Overview and applications, Sociological Methods & Research 46 (3) (2017) 303–341. doi:10.1177/0049124115585360.

[24] J. W. Bartlett, R. H. Keogh, Bayesian correction for covariate measurement error: A frequentist evaluation and comparison with regression calibration, Statistical Methods in Medical Research 27 (6) (2018) 1695–1708. doi:10.1177/0962280216667764.

[25] T. B. Brakenhoff, M. Mitroiu, R. H. Keogh, K. G. M. Moons, R. H. H. Groenwold, M. van Smeden, Measurement error is often neglected in medical literature: A systematic review, Journal of Clinical Epidemiology 98 (2018) 89–97. doi:10.1016/j.jclinepi.2018.02.023.

[26] P. A. Shaw, V. Deffner, R. H. Keogh, J. A. Tooze, K. W. Dodd, H. Küchenhoff, V. Kipnis, L. S. Freedman, Epidemiologic analyses with error-prone exposures: Review of current practice and recommendations, Annals of Epidemiology 28 (11) (2018) 821–828. doi:10.1016/j.annepidem.2018.09.001.

[27] A. M. Jurek, G. Maldonado, S. Greenland, T. R. Church, Exposure-measurement error is frequently ignored when interpreting epidemiologic study results, European Journal of Epidemiology 21 (12) (2007) 871–876. doi:10.1007/s10654-006-9083-0.

2

Measurement error in continuous endpoints of randomised trials: Problems and solutions

In randomised trials, continuous endpoints are often measured with some degree of error. This chapter explores the impact of ignoring measurement error, and proposes methods to improve statistical inference in the presence of measurement error. Three main types of measurement error in continuous endpoints are considered: classical, systematic and differential. For each measurement error type, a corrected effect estimator is proposed. The corrected estimators and several methods for confidence interval estimation are tested in a simulation study. These methods combine information about error-prone and error-free measurements of the endpoint in individuals not included in the trial (external validation sample). We show that when classical measurement error in continuous endpoints is ignored, the treatment effect estimator is unbiased, while Type-II error is increased at a given sample size. Conversely, the estimator can be substantially biased when measurement error is systematic or differential. In those cases, bias can largely be prevented and inferences improved upon using information from an external validation sample, of which the required sample size increases as the strength of the association between the error-prone and error-free endpoint decreases. Measurement error correction using already a small (external) validation sample is shown to improve inferences and could be considered in trials with error-prone endpoints. Implementation of the proposed correction methods is accommodated by a new software package for R.

This chapter is based on: L. Nab, R.H.H. Groenwold, P.M.J. Welsing and M. van Smeden, Measurement error in continuous endpoints of randomised trials: Problems and solutions, Statistics in Medicine 38 (27) (2019) 5182–5196. doi:10.1002/sim.8359

2.1. Introduction

In randomised controlled trials, continuous endpoints are often measured with some degree of error. Examples include trial endpoints that are based on self-report (e.g. self-reported physical activity levels [1]), endpoints that are collected as part of routine care (e.g. in pragmatic trials [2]), endpoints that are assessed without blinding the patient or assessor to treatment allocation (e.g., in surgical [3] or dietary [4] interventions) and an alternative endpoint assessment that substitutes a gold-standard measurement because of monetary or time constraints or ethical considerations (e.g. food frequency questionnaire as substitute for doubly-labelled water to measure energy intake [5]). In these examples, the continuous endpoint measurements contain error in the sense that the recorded endpoints do not unequivocally reflect the endpoint one aims to measure.

Despite calls for attention to the issue of measurement error in endpoints (e.g., [6]), developments and applications of correction methods for error in endpoints are still rare [7]. Specifically, methodology that allow for correction of study estimates for the presence of measurement error have so far largely been focused on the setting of error in explanatory variables, which may give rise to inferential errors such as regression dilution bias [8–13]. In addition, the application of correction methods for measurement error in the applied medical literature is unusual [9, 14].

We provide an exploration of problems and solutions for measurement error in continuous trial endpoints. For illustration of the problems and solutions for measurement error in continuous endpoints we consider one published trial that examined the efficacy and tolerability of low-dose iron-supplements during pregnancy [15]. To test the effect of the iron supplementation on maternal haemoglobin levels, haemoglobin concentrations were measured at delivery in venous blood.

This chapter describes a taxonomy of measurement error in trial endpoints, evaluates the impact of measurement error on the analysis of trials and tests existing and proposes new methods evaluating trials containing measurement error. Implementation of the proposed measurement error correction methods (i.e., the existing and novel methods) are supported by introducing a new R package `mecor`, available at: `https://github.com/LindaNab/mecor`. This chapter is structured as follows. In section 2.2 we revisit the example trial introduced in the previous paragraph. Section 2.3 presents an exploration of measurement error structures and their impact on inferences of trials. In section 2.4 measurement error correction methods are proposed. A simulation study investigating the efficacy of the correction methods is presented in section 2.5. Conclusions and recommendations resulting from this study are provided in section 2.6.

2.2. Illustrative example: measurement of haemoglobin levels

Makrides et al. [15] tested the efficacy of a 20-mg daily iron supplement (ferrous sulfate) on maternal iron status in pregnant women in a randomised, two-arm, double-blind, placebo-controlled trial. Respectively, 216 and 214 women were randomised to the iron supplement and placebo arm. At delivery, a 5-mL venous blood sample was collected from the women to assess haemoglobin levels as a marker for their iron status. Haemoglobin levels of women in the iron supplement arm were significantly higher than haemoglobin levels of women in the placebo arm (mean difference 6.9, 95% confidence interval (CI) (4.4; 9.3)). Haemoglobin concentrations were measured spectrophotometrically. Mean

haemoglobin values were 137 (standard deviation (SD) 3.2) g/L when measured by certified measurements, compared to mean 135 (SD 0.96) g/L when measured using the equipment used in the trial to measure haemoglobin levels. This might indicate small measurement error in the measured haemoglobin levels of the women in the trial. The authors did not discuss if and how the remaining measurement error could have affected their results.

In this domain, similar trials have been conducted in which the endpoint was assessed with lower standards. For instance, in field trials testing the effectiveness of iron supplementation, capillary blood samples instead of venous blood samples are often used to measure haemoglobin levels (e.g., [16]). While easier to measure, capillary haemoglobin levels are less accurate than venous haemoglobin levels [17]. We now discuss how measurement error in haemoglobin levels might affect trial inference, by assuming hypothetical differences between capillary and venous haemoglobin levels. Two additional illustrative example trials are discussed in section S2.1 of the supplementary material.

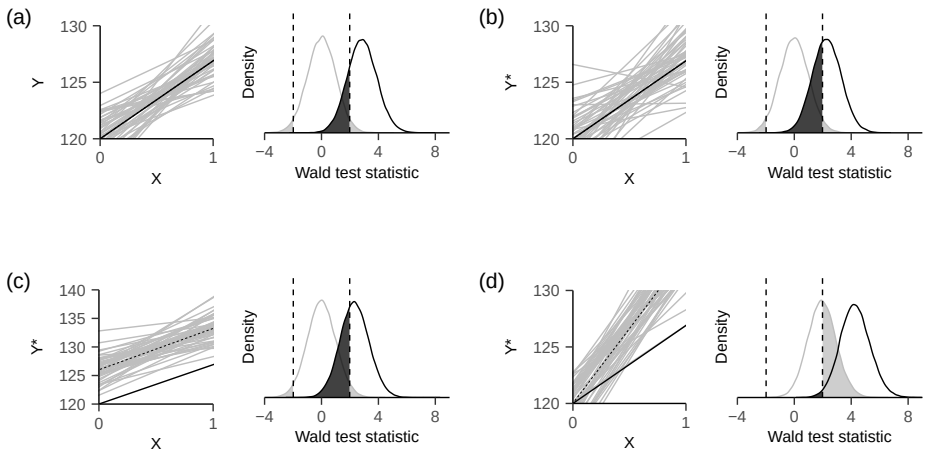


Figure 2.1: Illustration of impact of hypothetical measurement error in example trial 1 [15]: (a) no measurement error; (b) classical measurement error; (c) systematic measurement error; (d) differential measurement error. The left plots depict every thousandth estimated OLS regression line (grey lines), the average estimated treatment effect (dashed line) and the true effect (solid line). The right plots depict the density distribution of the Wald test-statistic of the slope of the regression line, under the null hypothesis of no effect (grey distribution) and the alternative hypothesis of any effect (black distribution).

2.2.1. Simulations based on example trial

We expand on the preceding example to hypothetical structures of error in measurement of the endpoints by simulation. These structures are only explained intuitively (explicit definitions are provided in section 2.3). For this example, we take the observed mean difference in haemoglobin levels in the two groups of the iron supplementation trials as a reference (6.9 g/L higher in the iron-supplemented group), and assume that haemoglobin levels are normally distributed with equal variance in both groups (SD 12.6 g/L). Fifty-thousand simulation samples were taken with 54 patients in each treatment arm. The number of patients differed from the 430 patients in the original trial to yield a Type-II

error of approximately 20% in the absence of measurement error at the usual alpha level (5%). Treatment effect for each simulation sample (mean difference in haemoglobin levels between the two arms) was estimated by OLS regression.

Classical measurement error in example trial. In the context of measurement of haemoglobin levels, random variability in the haemoglobin levels measured in capillary blood samples may be expected to vary more than haemoglobin levels measured in venous blood [17], independent of the true haemoglobin level and allocated treatment. Increased Type-II error is a well-known consequence of endpoints measured by the lower standard that are unbiased but more variable than the endpoints measured by the preferred measurement instruments [13]. This form of measurement error is commonly described as 'random measurement error' or 'classical measurement error' [10]. To simulate such independent variation, we arbitrarily increased the standard deviation of haemoglobin levels by 75% (from 12.6 to 22.05). This is equivalent to adding a term drawn from a normal distribution with mean 0 and standard deviation 18.1 to each endpoint. The impact of this imposed classical error was an increased between-replication variance of the estimated treatment effects of approximately 55% (left plot in panel b, Figure 2.1). The average estimated effect across simulations (depicted by the dashed line) is approximately equal to the true effect (depicted by the solid line), suggesting the classical measurement error did not introduce a bias in the estimated treatment effect (a formal proof is given in section Classical measurement error). Type-II error increased (to 38%) (grey area in Figure 2.1, panel b) while Type-I error remained at the nominal level (at 5%, illustrated by the red area in Figure 2.1, panel b).

Systematic measurement error in example trial. It may alternatively be assumed that capillary haemoglobin levels are systematically different from venous haemoglobin levels. This systematic difference can be either additive or multiplicative. For additive systematic measurement error, the capillary haemoglobin levels differ from venous haemoglobin levels with a certain constant, independently of venous haemoglobin levels. This implies that in both treatment groups mean haemoglobin level is higher, but that the difference between the two treatment groups is unbiased. The term systematic measurement error is often used to indicate multiplicative measurement error [18]. In that case, the expected capillary haemoglobin levels are equal to venous haemoglobin levels multiplied by a certain constant. Consequently, haemoglobin levels in capillary blood are more accurately measured in patients with low venous haemoglobin levels than in patients with high true haemoglobin levels (or vice versa). Under the assumption of a non-zero treatment effect, the expected difference between mean haemoglobin levels between the two treatment groups is biased; in the absence of a treatment effect, the expected difference between the two groups will remain unaffected. To simulate, we assumed that capillary haemoglobin levels are 1.05 times haemoglobin levels and we increased the standard deviation of haemoglobin levels by 75%, equivalent to the previous example. The impact of this imposed systematic measurement error structure is that the average treatment effect was biased, increasing from 6.9 to 7.2, and that there is an increased between-replication variance of the estimated treatment effect of approximately 66% (left plot in Figure 2.1, panel c). Type-II error increased (to 37%) (grey area in Figure 2.1, panel c) while Type-I error remained at rate close to nominal level (at 5%) (red area in Figure 2.1, panel c).

Differential measurement error in example trial. The measurement error structure may also differ between the treatment arms. In an extreme scenario, haemoglobin levels

in placebo group patients would be measured by venous blood samples while patients in active arm (iron supplemented) would be measured using capillary blood samples. To simulate such a scenario, we assume the same systematic error structure from the previous paragraph, now only applying to the active group. Additionally, we assume classical measurement error in the placebo group. This scenario classifies as differential measurement error [7]. The impact of this measurement error structure is that the average treatment effect was biased, increasing from 6.9 to 13.3, and that the between-replication variance of the estimated treatment effect is increased by approximately 62% (left plot in Figure 2.1, panel d). Type-II error decreased (to 0.1%) (grey area in Figure 2.1, panel d) and Type-I error rates increased (to 48%) (grey area in Figure 2.1, panel d).

2.3. Measurement error structures

Consider a two-arm randomised controlled trial that compares the effects of two treatments ($X \in \{0, 1\}$), where 0 may represent a placebo treatment or an active comparator. Let Y denote the true (or preferred) trial endpoint and Y^* an error prone operationalisation of Y. We will assume that both Y and Y^* are measured on a continuous scale. We assume a linear regression model for the endpoint Y:

$$Y = \alpha_Y + \beta_Y X + \varepsilon, \tag{2.1}$$

where ε is iid normally distributed with mean 0 and variance σ^2. Under these assumptions and assumptions about the model for Y^* (described below), simple formulas for the bias in the OLS estimator of the treatment effect can be derived. Details of these derivations can be found in the supplementary material, section S2.2.

2.3.1. Classical measurement error

There is classical measurement error in Y^* when Y^* is an unbiased proxy for Y [10]: $Y^* = Y + e$, where e has mean 0 and $\text{Var}(e) = \tau^2$ and e independent of Y, X, ε in (2.1). Using Y^* instead of Y in the linear model yields:

$$Y^* = \alpha_Y^* + \beta_Y^* X + \delta, \tag{2.2}$$

Where $\beta_Y^* = \beta_Y$ and the residuals δ have mean 0 and variance $\sigma_\delta^2 = \sigma^2 + \tau^2$. This leads to a larger variance in $\hat{\beta}_Y^*$ (the estimator for β_Y^*) compared to the variance in $\hat{\beta}_Y$ (the estimator for β_Y). Consequently, classical measurement error will not lead to bias in the effect estimator but will increase Type-II for a given sample size.

2.3.2. Heteroscedastic measurement error

In the above we assumed that the variance in e is equal in both arms. When this assumption is violated, there is so called heteroscedastic measurement error. Heteroscedastic error will not lead to bias in the effect estimator, but will invalidate the estimator of the variance of $\hat{\beta}_Y^*$ (proof is given in supplementary material section S2.2).

2.3.3. Systematic measurement error

There is systematic measurement error in Y^* if Y^* depends systematically on Y: $Y^* = \theta_0 + \theta_1 Y + e$, where e has mean 0 and $\text{Var}(e) = \tau^2$ and e independent of Y, X, ε in (2.1).

Throughout, we assume systematic measurement error if $\theta_0 \neq 0$ or $\theta_1 \neq 1$ (and of course, $\theta_1 \neq 0$ in all cases). We assume independence between e and Y, X, ε in (2.1). Using Y^* with systematic measurement error in the linear model yields in the model defined by (2.2) where $\beta_Y^* = \theta_1\beta_Y$ and the residuals δ have mean 0 and variance $\sigma_\delta = \theta_1^2\sigma^2 + \tau^2$. Depending on the value of θ_1, the variance of $\hat{\beta}_Y^*$ is larger or smaller than the variance of $\hat{\beta}_Y$. Hence, Type-II error will either decrease or increase under systematic measurement. Type-I error is unaffected since if $\beta_Y = 0$, $\beta_Y^* = 0$ (i.e., tests for null effects are still valid under systematic measurement error) (proof is given in supplementary material section S2.2).

2.3.4. Differential measurement error

There is differential measurement error in Y^* if Y^* depends systematically on Y varying for X: $Y^* = \theta_{00} + (\theta_{01} - \theta_{00})X + \theta_{10}Y + (\theta_{11} - \theta_{10})XY + e_X$, where e_X has mean 0 and $\text{Var}(e) = \tau_X^2$ and e_X independent of Y, and ε in (2.1) for $X = 0, 1$. Using Y^* with differential measurement error in the linear model yields in the model defined in (2.2) where $\beta_Y^* = \theta_{01} - \theta_{00} + (\theta_{11} - \theta_{10})\alpha_Y + \theta_{11}\beta_Y$ and the residuals δ have mean 0 and variance $\left[\theta_{10}^2 + (\theta_{11}^2 - \theta_{10}^2)X\right]\sigma^2 + \tau_X^2$ for $X = 0, 1$. Since the residual variance is not equal in both arms, the estimator of the variance of $\hat{\beta}_Y^*$ is invalid, and will underestimate the true variance. A heteroscedastic consistent estimator of the variance of $\hat{\beta}_Y^*$ is provided by the White estimator [19]. Assuming that the White estimator is used to estimate the variance of $\hat{\beta}_Y^*$, Type-I error is not expected the nominal level (α) and Type-II error will decrease or increase under the differential measurement error model (proof is given in supplementary material section S2.2).

2.4. Correction methods for measurement error in a continuous trial endpoint

In this section we describe several approaches to address measurement error in the trial endpoint. Throughout, we assume that Y^* is measured for all $i = 1, \ldots, N$ randomly allocated patients in the trial. We also assume that Y and Y^* are both measured for a smaller set of different individuals not included in the trial ($j = 1, \ldots, K, K < N$), hereinafter referred to as the external calibration sample. In all but one case, it is assumed that only Y^* and Y are measured in the external validation sample. In the case that the error in Y^* is different for the two treatment groups, it is assumed that the external validation sample is in the form of a small pilot study where both treatments are allocated (i.e., Y^* and Y are both measured after assignment of X). Instead of external validation data, we could use internal validation data to correct for measurement error (Y and Y^* are both measured in a small subset of the trial), which is not considered in this section as it was studied elsewhere [7].

A well-known consequence of classical measurement error in a continuous trial endpoint is that a larger sample size (as compared to the same situations without the measurement error) is needed to compensate for the reduced precision [13]. For example, the new sample size N^* may be calculated by N/R formula where R is the reliability coefficient and N the original sample size for the trial [20]. For solutions for heteroscedastic measurement error, we refer to standard theory of dealing with heteroscedastic errors in regression to find an unbiased estimator for the variance of $\hat{\beta}_{Y^*}$ (e.g., see [19] for an overview of different heteroscedasticity consistent covariance matrices).

Hereinafter we focus on measurement error in Y^* that is either systematic or differential, both of which have been shown to introduce bias in the effect estimator if measurement error is neglected (section 2.3). Consistent estimators for the intervention effects are introduced, and various methods for constructing CIs for these estimators are discussed. Section S2.3 in the supplementary material provides an explanation of the results stated in this section. Throughout, we assume that Y^* is measured for all $i = 1, \ldots, N$ patients in the trial. We also assume that Y and Y^* are both measured for a smaller set of different individuals not included in the trial ($j = 1, \ldots, K, K < N$), hereinafter referred to as the external validation sample. For an earlier exploration of the use of an internal validation set when there is systematic or differential measurement error in endpoints, see [7].

2.4.1. Systematic measurement error

From section Systematic measurement error it follows that natural estimators for α_Y and β_Y are

$$\hat{\alpha}_Y = (\hat{\alpha}_{Y^*} - \hat{\theta}_0)/\hat{\theta}_1 \quad \text{and} \quad \hat{\beta}_Y = \hat{\beta}_{Y^*}/\hat{\theta}_1, \tag{2.3}$$

Where $\hat{\theta}_0$ and $\hat{\theta}_1$ are the estimated error parameters from the validation data set using standard OLS regression. From equation (2.3), it becomes apparent that $\hat{\theta}_1$ needs to be assumed bounded away from zero for finite estimates of $\hat{\alpha}_Y$ and $\hat{\beta}_Y$ [8]. The estimators in (2.3) are consistent, see for a proof section S2.3 in the supplementary material.
The variance of the estimators defined in (2.3) can be approximated using the Delta method, the Fieller method, the Zero-variance method [21] and by bootstrap [22]. Further details are provided in section S2.3 of the supplementary material.

2.4.2. Differential measurement error

From section Differential measurement error it follows that natural estimators for α_Y and β_Y are,

$$\hat{\alpha}_Y = (\hat{\alpha}_{Y^*} - \hat{\theta}_{00})/\hat{\theta}_{10} \quad \text{and} \quad \hat{\beta}_Y = (\hat{\beta}_{Y^*} + \hat{\alpha}_{Y^*} - \hat{\theta}_{01})/\hat{\theta}_{11} - \hat{\alpha}_Y, \tag{2.4}$$

where $\hat{\theta}_{00}$, $\hat{\theta}_{10}$, $\hat{\theta}_{01}$ and $\hat{\theta}_{11}$ are estimated from the external validation set using standard OLS estimators. Here it is assumed that both $\hat{\theta}_{10}$ and $\hat{\theta}_{11}$ are bounded away from zero (for reasons similar to those mentioned in section 2.4.1). The estimators in (2.4) are consistent, see for a proof section S2.3 of the supplementary material. The variance of the estimators defined in (2.4) can be approximated using the Delta method [21], the Zero-variance method and by bootstrap [22]. Further details are provided in section S2.3 of the supplementary material.

2.5. Simulation study

The finite sample performance of the measurement error corrected estimators of the treatment effect was studied by simulation. We focussed on the setting of a two-arm trial in which the continuous surrogate endpoint Y^* was measured with systematic or differential measurement error, and in which an external validation set was available, which was varied in size. The results from example trial 1 are used to motivate our simulation study (see section 2.2).

2.5.1. Data generation

Data were generated for a sample of $N = 400$ individuals, approximately equal to the size of example trial 1 [15]. The individuals were equally divided in the two treatment arms. The true endpoints were generated according to model (2.1), assuming iid normal errors, and using the estimated characteristics found in example trial 1 ($\alpha_Y = 120$, $\beta_Y = 6.9$ and $\sigma = 12.6$). Surrogate endpoints Y^* were generated under models for systematic measurement error and differential measurement error described in section Systematic measurement error and Differential measurement error, respectively.

For systematic measurement error in Y^*, we set $\theta_0 = 0$ and $\theta_1 = 1.05$. Under the differential measurement error model we set $\theta_{00} = 0$, $\theta_{01} = 0$, $\theta_{10} = 1$, $\theta_{11} = 1.05$. We considered three scenarios based on the coefficient of determination between the Y^* and Y, $R^2_{Y^*,Y}$: (i) $R^2_{Y^*,Y} = 0.8$, (ii) $R^2_{Y^*,Y} = 0.5$ and (iii) $R^2_{Y^*,Y} = 0.2$. This large range in coefficient of determination values reflects the wide variation we anticipate in practice from very strong correlations between Y_* and Y ($R^2_{Y^*,Y} = 0.8$) to weak correlations ($R^2_{Y^*,Y} = 0.2$), as for example, one could expect in the context of trials with dietary intake as endpoints [7, 23]. For $R^2_{Y^*,Y} = 0.8$, $\tau = 6.6$ for systematic measurement error and $\tau_0 = 6.3$ and $\tau_1 = 6.6$ for differential measurement error. For $R^2_{Y^*,Y} = 0.5$, $\tau = 13.2$ for systematic measurement error and $\tau_0 = 12.6$ and $\tau_1 = 13.2$ for differential measurement error. For $R^2_{Y^*,Y} = 0.2$, $\tau = 26.5$ for systematic measurement error and $\tau_0 = 25.2$ and $\tau_1 = 26.5$ for differential measurement error. Additionally, we considered a scenario with greater systematic measurement error holding $\theta_0 = 0$ and $\theta_1 = 1.25$. Here, we only studied a high coefficient of determination $R^2_{Y^*,Y} = 0.8$, implying that $\tau = 7.9$.

For the scenarios with systematic measurement error induced, a separate validation set was generated of size K with the characteristics of the placebo arm for each simulated data set. For differential measurement error scenarios, a validation data set was generated of size K for each simulated data set, with $K_0 = K_1 = K/2$ subjects equally divided over the two treatment groups. The sample size of the external validation data set (K) was varied with $K \in \{5, 7, 10, 15, 20, 30, 40, 50\}$ for systematic measurement error and $K \in \{10, 20, 30, 40, 50\}$ for differential measurement error.

2.5.2. Computation

For each simulated data set the corrected treatment effect estimator (2.3) for systematic error and (2.4) for differential error were applied. In systematic measurement error scenarios, 95% CIs for the corrected estimator were constructed by using the Zero-variance method, the Delta method, the Fieller method, and bootstrap based on 999 replicates (as defined in section Systematic measurement error). In the case of differential measurement error, 95% CIs for the corrected estimator were constructed by using the Zero-Variance method, the Delta method and the bootstrap based on 999 replicates (as defined in section Differential measurement error). The HC3 heteroscedastic consistent variance estimator was used to accommodate for heteroscedastic error in the differential measurement error scenario [19]. Furthermore, for both the systematic and differential measurement error scenarios the naive analysis was performed (resulting in a naive effect estimate and naive CI), which is the 'regular' analysis which would be performed if measurement error was neglected.

We studied performance of the corrected treatment effect estimators in terms of

percentage bias [24], empirical standard error (EmpSE) and square root of the mean squared error (SqrtMSE) [25]. The performance of the methods for constructing the CIs was studied in terms of coverage and Type-II error [25].

In our simulations, the Fieller method resulted in undefined CIs if in an iteration $\hat{\theta}_1/\sqrt{t^2/S_{yy}^{(c)}} > t_{N-2}$. The percentage of iterations for which the Fieller method failed to construct CIs is reported. If the Fieller method resulted in undefined CIs in more than 5% of cases in one simulation scenario, the coverage and average CI width were not calculated as this would result in unfair comparisons between the different CI constructing methods. The bootstrap CIs were based on less than 999 estimates in case the sample drawn from the external validation set consisted of K equal replicates. These errors occurred more frequently for small values of K and low R-squared. All simulations were run in R version 3.4, using the `R` package `mecor` (version 0.1.0). The results of the simulation are available at doi:10.6084/m9.figshare.7068695 and the code is available at doi:10.6084/m9.figshare.7068773, together with the seed used for the simulation study.

2.5.3. Results of simulation study

Systematic measurement error. Table 2.1 shows percentage bias, EmpSE and SqrtMSE of the naive estimator and the corrected estimator when there is systematic measurement error. Naturally, the percentage of bias in the naive estimator is about 5% if $\theta_1 = 1.05$ and 25% if $\theta_1 = 1.25$. For the corrected estimator and $\theta_1 = 1.05$ or $\theta_1 = 1.25$ and $R^2_{Y^*,Y} = 0.8$, percentage bias, EmpSE and SqrtMSE of $\hat{\beta}_Y$ were reasonably small for $K \geq 10$. SqrtMSE of the corrected estimator was never lower than the SqrtMSE of the naive estimator because the bias in the naive estimator was small for $\theta_1 = 1.05$, . However, for settings where bias in the naive estimator was greater ($\theta_1 = 1.25$), SqrtMSE of the corrected estimator was smaller than SqrtMSE of the naive estimator for $K \geq 15$. For the corrected estimator and $\theta_1 = 1.05$ and $R^2_{Y^*,Y} = 0.5$, bias was reasonably small for $K \geq 30$. Nevertheless, SqrtMSE of the corrected estimator was always greater than SqrtMSE of the naive estimator. For the corrected estimator and $\theta_1 = 1.05$ and $R^2_{Y^*,Y} = 0.2$, the bias of $\hat{\beta}_Y$ fluctuated and EmpSE and SqrtMSE was large for all $K's$. Figure 2.2 shows the estimates of the intervention effect using the corrected estimator of each 10th iteration of our simulation, which provides a clear visualisation of the results formerly discussed. The larger the sample size of the external calibration set and the higher R-squared, the better the performance of the corrected estimator. The sampling distribution of $\hat{\theta}_1$ depicted in Figure 2.3 explains why there was so much variation in the corrected effect estimator for small sample sizes of the external validation set and low R-squared. Namely, for a number of iterations in our simulation, $\hat{\theta}_1$ was estimated close to zero, expanding the corrected estimator the same number of times resulting in large bias, EmpSE and MSE. Note that if $\hat{\theta}_1 < 0$, the sign of the corrected estimator changes, explaining why the corrected estimate of the intervention effect was sometimes below zero.

For $R^2_{Y^*,Y} = 0.8$ and both $\theta_1 = 1.05$ and $\theta_1 = 1.25$, the Fieller method failed to construct CIs in 15, 5, 1 and 0.1 % of simulated datasets for respectively $K = 5, 7, 10, 15$. Therefore, coverage and average CI width of the Fieller method was not evaluated for $K \in \{5, 7\}$. For $R^2_{Y^*,Y} = 0.5$, the Fieller method failed to construct CIs in 48, 36, 22, 8, 3, 0.3 % of simulated data sets for $K \in \{5, 7, 10, 15, 20, 30\}$, respectively. Consequently, coverage and average CI width was not evaluated for $K \in \{5, 7, 10, 15\}$. For $R^2_{Y^*,Y} = 0.2$, the Fieller method

failed to construct CIs in 74, 71, 64, 53, 43, 26, 15 and 8 % of simulated data sets for $K \in \{5, 7, 10, 15, 20, 30, 40, 50\}$, respectively. the Fieller method was therefore not evaluated for $R^2_{Y^*,Y} = 0.2$.

Table 2.2 shows coverage of the true intervention effect in the constructed CIs using the Zero-variance, Delta, Fieller method and bootstrap. Using Wald CIs for the naive effect estimator nearly yielded 95% coverage of the true treatment effect of 6.9, because for $\theta_1 = 1.05$ the bias percentage in the naive estimator was small (, 5%). Yet, as bias percentage increased in the naive estimator for $\theta_1 = 1.25$ (i.e., 25%) coverage dropped to 83.5%. Table 2.3 shows average CI width using the Zero-variance, Delta and bootstrap. The Zero-variance method yielded too narrow CIs for all scenario's, an intuitively clear result as the Zero-variance method neglects the variance in $\hat{\theta}_1$. For $R^2_{Y^*,Y} = 0.8$ the Delta, Fieller and bootstrap constructed correct CIs for $K \geq 15$. For $K \leq 10$ the Delta method and the Fieller method constructed too narrow CIs, and bootstrap too broad CIs. For $R^2_{Y^*,Y} = 0.5$ the Delta and bootstrap constructed correct CIs for $K \geq 30$. For $K \leq 20$ the Delta method constructed too narrow CIs, and bootstrap too broad CIs. Coverage of the Fieller method was about the desired 95% level for $K \geq 30$.

Using the naive effect estimator, Type-II error was 0.2%, 2.9% and 31.6% for $R^2_{Y^*,Y} = 0.8$ (both for $\theta_1 = 1.05$ and $\theta_1 = 1.25$), $R^2_{Y^*,Y} = 0.5$ and $R^2_{Y^*,Y} = 0.2$, respectively. Type-II error in the corrected estimator using the Zero-variance and Delta method and bootstrap was 0%. For the considered scenario's using the Fieller method, Type-II error was 0.02% for $R^2_{Y^*,Y} = 0.8$ and 2.9% for $R^2_{Y^*,Y} = 0.5$.

Table 2.1: Percentage bias, Empirical Standard Error (EmpSE) and Squared root of Mean Squared Error (SqrtMSE) of the naive estimator and the corrected estimator for systematic measurement error ($\theta_0 = 0$ and $\theta_1 = 1.25$ or $\theta_1 = 1.05$), R-squared equal to 0.8, 0.5 and 0.2 and, different sample sizes of the validation data set. Each scenario is based on 10,000 replicates, the value of the estimand is 6.9, based on example trial 1 by Makrides et al. [15]

Performance Measure[a]	θ_1	$R^2_{Y_*,Y}$	**Naive Estimator**	**Corrected Estimator Sample Size External Calibration Set**							
				5	**7**	**10**	**15**	**20**	**30**	**40**	**50**
Percentage bias (%)	1.25	0.8	24.9	88.9	29.0	3.7	2.0	1.6	0.9	0.7	0.4
	1.05	0.8	4.9	88.9	29.0	3.7	2.0	1.6	0.9	0.7	0.4
		0.5	4.9	55.3	57.5	−2.4	7.6	5.8	4.3	3.0	2.0
		0.2	4.9	168.2	−62.6	98.8	33.4	−142.2	−28.3	23.9	14.6
EmpSE	1.25	0.8	1.8	524.8	139.1	3.0	1.9	1.7	1.6	1.5	1.5
	1.05	0.8	1.5	524.8	139.1	3.0	1.9	1.7	1.6	1.5	1.5
		0.5	1.9	267.0	329.1	83.7	14.4	11.0	2.5	2.3	2.1
		0.2	3.0	1131.2	210.8	723.2	462.2	1044.4	225.5	70.5	24.8
SqrtMSE	1.25	0.8	2.5	524.8	139.1	3.1	1.9	1.7	1.6	1.5	1.5
	1.05	0.8	1.5	524.8	139.1	3.1	1.9	1.7	1.6	1.5	1.5
		0.5	1.9	267.0	329.1	83.7	14.4	11.0	2.5	2.3	2.1
		0.2	3.0	1131.2	210.8	723.1	462.2	1044.4	225.5	70.5	24.8

[a] Monte Carlo standard errors of Bias, EmpSE and MSE are EmpSE$\sqrt{1/10,000}$; EmpSE/$(2\sqrt{9,999})$; $\sqrt{\frac{\sum_{i=1}^{10,000}[(\hat{\beta}_1-6.9)^2-\text{MSE}]^2}{9,999\times 10,000}}$, respectively [25].

Table 2.2: Confidence interval (CI) coverage of the naive estimator and the corrected estimator for systematic measurement error ($\theta_0 = 0$ and $\theta_1 = 1.25$ or $\theta_1 = 1.05$), R-squared equal to 0.8, 0.5 and 0.2 and, different sample sizes of the validation data set. Each scenario is based on 10,000 replicates, the value of the estimand is 6.9, based on example trial 1 by Makrides et al. [15]

θ_1	$R^2_{Y_*,Y}$	**Naive Esti-mator**[a]	**Method for Construction of CI**	**Corrected Estimator Sample Size External Calibration Set**							
				5	**7**	**10**	**15**	**20**	**30**	**40**	**50**
1.25	0.8	83.5	Zero-Variance	70.3	74.0	77.4	80.3	82.8	84.4	85.3	86.3
			Delta	93.8	95.3	95.7	95.9	96.0	96.0	95.9	95.7
			Fieller[b]	-	-	94.5	94.7	95.0	95.3	95.2	95.0
			Bootstrap	95.9	96.1	95.5	94.9	94.8	95.0	95.1	94.8
1.05	0.8	94.6	Zero-Variance	77.8	81.3	84.4	87.1	89.2	90.9	92.0	92.2
			Delta	92.1	93.9	94.3	94.8	95.1	95.3	95.4	95.2
			Fieller[b]	-	-	94.5	94.7	95.0	95.3	95.2	95.0
			Bootstrap	95.9	96.1	95.5	94.9	94.8	95.0	95.1	94.8
	0.5	94.8	Zero-Variance	69.1	73.5	78.1	81.7	84.5	87.5	88.7	89.9
			Delta	89.7	92.0	92.9	93.9	94.3	95.2	95.4	95.3
			Fieller[b]	-	-	94.5	95.2	95.2	95.0	94.8	94.9
			Bootstrap	93.9	95.9	96.3	95.8	95.4	94.8	94.8	94.8
	0.2	95.1	Zero-Variance	57.1	64.5	71.0	76.8	80.3	84.3	86.0	87.6
			Delta	86.8	89.7	90.9	92.2	93.5	94.4	94.6	94.9
			Fieller[b]	-	-	89.8	93.2	94.9	95.8	95.8	95.7
			Bootstrap	88.9	93.8	95.5	96.4	96.7	96.8	96.8	96.1

Monte Carlo standard errors of Coverage are $\sqrt{[\text{Coverage} \times (1 - \text{Coverage})]/10,000}$ [25].
[a] Coverage of the true intervention effect using regular Wald CIs of the naive effect estimator.
[b] Results of the Fieller method are shown if less than 5% of cases resulted in undefined CIs (see section Computation).

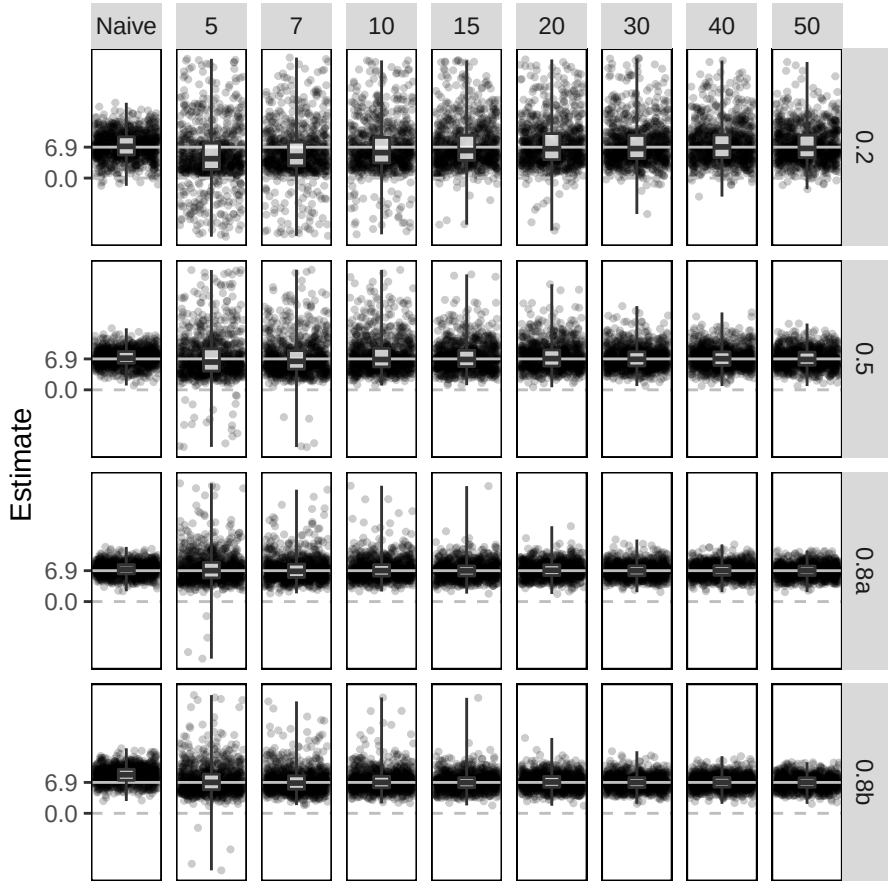


Figure 2.2: Estimates of the treatment effect using the naive estimator and corrected estimator for different values of R-squared (row grids) and different sample sizes of the external validation set (column grids) under systematic measurement error (θ_1 = 1.05 (0.2; 0.5; 0.8a) or θ_1 = 1.25 (0.8b)). Each grid is based on every 10th estimate of a simulation of 10,000 replicates, using an estimand of 6.9 (indicated by the solid supplementary material line), based on example trial 1 by Makrides et al. [15].

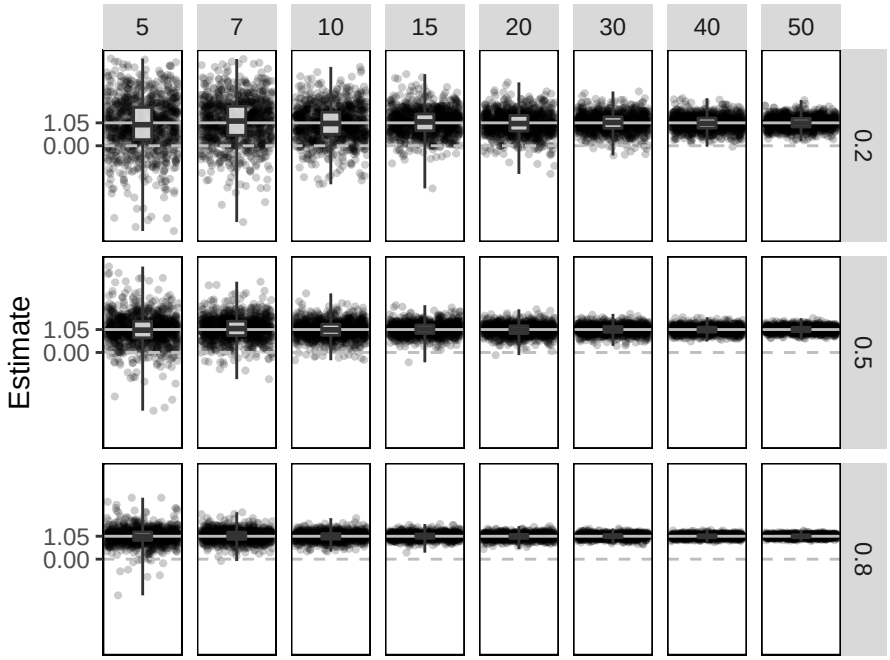


Figure 2.3: Estimates of θ_1 (i.e., slope of the systematic measurement error model) for different values of R-squared (row grids) and different sample sizes of the external validation set (column grids). Each grid is based on every 10th estimate of a simulation of 10,000 replicates, using an estimand of 1.05 (indicated by the solid supplementary material line).

Table 2.3: Average confidence interval (CI) width of the naive estimator and the corrected estimator for systematic measurement error ($\theta_0 = 0$ and $\theta_1 = 1.25$ or $\theta_1 = 1.05$), R-squared equal to 0.8, 0.5 and 0.2 and, different sample sizes of the validation data set. Each scenario is based on 10,000 replicates, the value of the estimand is 6.9, based on example trial 1 by Makrides et al. [15]

θ_1	$R^2_{Y^*,Y}$	Naive Estimator[a]	Method for Construction of CI	Corrected Estimator Sample Size External Calibration Set							
				5	7	10	15	20	30	40	50
1.25	0.8	6.9	Zero-Variance	30 333.0	1141.5	5.5	4.7	4.7	4.6	4.5	4.5
			Delta	40.7	13.6	8.7	7.5	7.0	6.5	6.3	6.1
			Fieller[b]	-	-	11.8	8.3	7.0	6.4	6.1	6.0
			Bootstrap	86.9	29.3	14.1	8.3	7.1	6.4	6.1	6.0
1.05	0.8	5.8	Zero-Variance	36 110.7	1359.0	6.5	5.6	5.5	5.4	5.4	5.4
			Delta	35.0	12.2	8.0	7.0	6.7	6.3	6.1	6.0
			Fieller[b]	-	-	11.8	8.3	7.0	6.4	6.1	6.0
			Bootstrap	86.9	29.3	14.1	8.3	7.1	6.4	6.1	6.0
	0.5	7.4	Zero-Variance	7228.9	9759.5	763.1	37.5	17.8	7.7	7.3	7.1
			Delta	58.1	43.2	21.2	12.6	11.0	9.3	8.7	8.4
			Fieller[b]	-	-	67.9	63.2	25.0	12.4	9.8	9.0
			Bootstrap	146.8	87.4	65.2	34.7	22.8	12.4	9.9	9.0
	0.2	11.6	Zero-Variance	126 830.3	11 677.5	87 123.4	30 709.4	324 870.7	12 430.8	774.6	126.8
			Delta	179.3	102.5	112.7	69.9	65.7	34.1	19.7	16.6
			Fieller[b]	-	-	92.6	95.1	72.1	82.2	60.6	59.2
			Bootstrap	176.0	121.9	126.2	118.7	107.7	77.6	54.8	39.7

[a] Average CI width using regular Wald CIs of the naive effect estimator.
[b] Results of the Fieller method are shown if less than 5% of cases resulted in undefined CIs (see section Computation).

Differential measurement error. Table 2.4 shows percentage bias, EmpSE and SqrtMSE of the naive estimator and the corrected estimator when there is differential measurement error. The percentage bias in the naive estimator was about 92%. For the corrected estimator and $R^2_{Y^*,Y} = 0.8$, percentage bias, EmpSE and SqrtMSE of $\hat{\beta}_Y$ were reasonably small for $K \geq 20$. For the naive estimator and $R^2_{Y^*,Y} = 0.5$, percentage bias, EmpSE and MSE of the corrected estimator were small for $K = 50$. For the naive estimator and $R^2_{Y^*,Y} = 0.2$, percentage bias, EmpSE and MSE of the corrected estimator was large for all K's. The estimates of the intervention effect using the corrected estimator of each 10th iteration of our simulation is shown in Figure 2.4, which provides a clear visualization of the results formerly discussed. the sample size of the external validation set and the higher R-squared, the better the performance of the corrected estimator.

Table 2.5 shows coverage of the true intervention effect in the constructed CIs and average CI width using the Zero-Variance and Delta method and bootstrap. Coverage of the true treatment effect of 6.9 using Wald CIs for the naive effect estimator were about 1%, 7% and 41% for $R^2_{Y^*,Y} = 0.8$, $R^2_{Y^*,Y} = 0.5$ and $R^2_{Y^*,Y} = 0.2$, respectively. In all cases, the Zero-Variance method yielded too narrow CIs; the Delta method yielded too broad CIs and the bootstrap yielded mostly too broad CIs, except for $R^2_{Y^*,Y} = 0.8$ and $K = 30$ and $K = 40$ (too narrow). For $R^2_{Y^*,Y} = 0.8$ and $K = 50$, coverage of the true intervention effect was 95%.

Type-II error in the naive effect estimator was 0%, 0% and 0.4% for $R^2_{Y^*,Y} = 0.8$, $R^2_{Y^*,Y} = 0.5$ and $R^2_{Y^*,Y} = 0.2$, respectively. Type-II error in the corrected effect estimator using the Zero-variance and Delta method and bootstrap was 0%.

Table 2.4: Percentage bias, Empirical Standard Error (EmpSE), Mean Squared Error (MSE), Squared root of Mean Squared Error (SqrtMSE) of the corrected estimator for differential measurement error ($\theta_{00} = 0$, $\theta_{10} = 1$, $\theta_{01} = 0$, $\theta_{11} = 1.05$) R-squared equal to 0.8, 0.5 and 0.2 and different sample sizes of the validation data set. Each scenario is based on 10,000 replicates, the value of the estimand is 6.9, based on example trial 1 by Makrides et al. [15]

Performance Measure[a]	$R^2_{Y^*,Y}$	**Naive Estimator**	**Corrected Estimator Sample Size External Calibration Set**				
			10	**20**	**30**	**40**	**50**
Percentage bias (%)	0.8	91.8	5.2	1.2	−0.4	−0.2	−0.1
	0.5	91.8	−9.7	33.0	154.2	−21.4	−0.1
	0.2	91.9	−319.4	152.9	193.1	−21.5	2.2
EmpSE	0.8	1.4	52.0	6.8	2.9	2.6	2.3
	0.5	1.8	949.1	369.1	1080.4	142.1	4.5
	0.2	2.9	2658.0	8425.8	1569.7	443.7	92.1
SqrtMSE	0.8	6.5	52.0	6.8	2.9	2.6	2.3
	0.5	6.6	949.1	369.1	1080.4	142.1	4.5
	0.2	7.0	2658.0	8425.4	1569.7	443.7	92.1

[a] Monte Carlo standard errors of Bias, EmpSE and MSE are EmpSE$\sqrt{1/10,000}$; EmpSE/$(2\sqrt{9,999})$; $\sqrt{\frac{\sum_{i=1}^{10,000}[(\hat{\beta}_i-6.9)^2-\text{MSE}]^2}{9,999\times 10,000}}$, respectively [25].

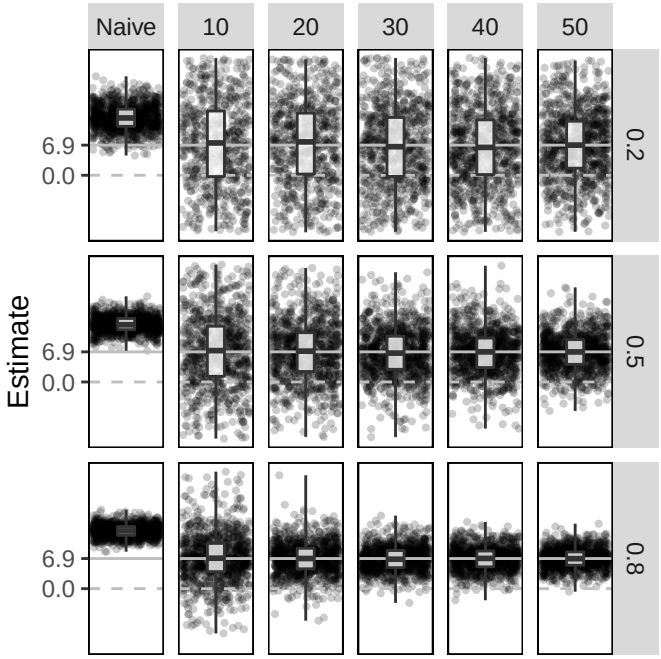


Figure 2.4: Estimates of the treatment effect using the naive estimator and corrected estimator for different values of R-squared (row grids) and different sample sizes of the external validation set (column grids) under differential measurement error ($\theta_{00} = 0$, $\theta_{10} = 1$, $\theta_{01} = 0$, $\theta_{11} = 1.05$). Each grid is based on every 10th estimate of a simulation of 10,000 replicates, using an estimand of 6.9 (indicated by the solid supplementary material line), based on example trial 1 by Makrides et al. [15].

Table 2.5: Coverage and average confidence interval (CI) width of the corrected estimator for differential measurement error (θ_{00} = 0, θ_{10} = 1, θ_{01} = 0, θ_{11} = 1.05), R-squared equal to 0.8, 0.5 and 0.2 and different sample sizes of the validation data set. Each scenario is based on 10,000 replicates, the value of the estimand is 6.9, based on example trial 1 by Makrides et al. [15]

Performance Measure[a]	$R^2_{Y^*,Y}$	**Naive Estimator**[b]	**Method for Construction of CI**	**Corrected Estimator Sample Size External Calibration Set**				
				10	**20**	**30**	**40**	**50**
Coverage (%)	0.8	0.7	Zero-Variance	43.8	59.9	67.9	72.7	76.8
			Delta	97.1	96.6	96.0	95.7	95.9
			Bootstrap	97.9	95.7	94.7	94.5	95.0
	0.5	6.7	Zero-Variance	30.3	43.3	50.2	55.5	61.0
			Delta	97.6	97.6	97.3	96.9	97.0
			Bootstrap	98.4	98.0	96.6	95.8	95.5
	0.2	41.1	Zero-Variance	25.7	35.0	41.9	46.6	52.2
			Delta	98.4	99.0	98.9	98.9	98.9
			Bootstrap	99.0	99.6	99.2	99.0	98.7
Average CI width	0.8	5.7	Zero-Variance	8.2	5.9	5.7	5.7	5.6
			Delta	2688.7	18.3	12.1	10.5	9.5
			Bootstrap	142.6	24.3	13.1	10.7	9.5
	0.5	7.2	Zero-Variance	33.0	17.9	30.3	10.6	7.5
			Delta	463 975.1	49 493.3	660 587.5	13 238.0	18.5
			Bootstrap	303.5	118.8	58.4	34.2	24.0
	0.2	11.4	Zero-Variance	64.6	150.5	53.1	43.1	26.8
			Delta	1 219 162.5	26 998 502.1	486 295.4	85 139.8	3407.5
			Bootstrap	562.9	353.8	283.3	221.4	170.2

[a] Monte Carlo standard errors of Coverage are $\sqrt{[\text{Cover.} \times (1 - \text{Cover.})]/10,000}$ [25]
[b] Coverage of the true intervention effect using regular Wald CIs of the naive effect estimator.

2.5.4. Measurement error dependent on a prognostic factor

Above, we focused on measurement error in endpoints that are either systematic (linearly dependent on true endpoint) or differential (linearly dependent on true endpoint and exposure). Yet, measurement error could depend on prognostic factors. For example, measurement error in haemoglobin levels measured in capillary blood may differ for women and men [17]. Moreover, haemoglobin levels are, on average, higher in men than in women. To illustrate the effect of measurement error that is dependent on a prognostic factor, we use example trial 1, here assuming that it was conducted in women and men. Data were generated for a sample of $N = 400$ individuals, equally divided in two treatment arms and with equal sex distribution in both arms. Let the proportion of women in the sample be 75% ($S = 1$ for men and $S = 0$ for women). Further, assume $Y = 120 + 6.9X + 10S + \varepsilon$, where ε has mean 0 and $\text{Var}(\varepsilon) = 158.8$. Additionally, assume additive systematic measurement error in Y^*, $Y^* = Y + 0.5S + e$ (additive systematic measurement error in men and random measurement error in women), where e has mean 0 and $\text{Var}(e) = 6.6$ and e independent of Y, X, S and ε. In a simulation of 10,000 replicates we estimated the effect of Y^* on X (naive analysis) and the effect of Y^* on X, conditional for S (conditional analysis). In section S2.4 of the supplementary material, we proof that both analyses will result in correct estimation of the treatment effect. The results of the simulation study show that the average treatment effect estimate of both analyses was 6.89, indicating that there is no bias in either of the analyses. Yet, the empirical variance of the effect estimate in the 10,000 replicates was somewhat lower for the conditional analysis compared to the naive analysis (2.01 vs. 2.22), indicating an efficiency gain in favour of the conditional analysis. By assuming that randomisation was well-performed, measurement error dependent on a prognostic factor does not introduce bias in the naive analysis other than the biases already discussed.

2.6. Discussion

This chapter outlined the ramifications for randomised trial inferences when a continuous endpoint is measured with error. Our study showed that when this measurement error is ignored, not only can trial results be hampered by a loss in precision of the treatment effect estimate (i.e., increased Type-II error for a given sample size), but trial inferences can be impacted through bias in the treatment effect estimator and a null-hypothesis significance test for the treatment effect can deviate substantially from the nominal level. In this chapter we proposed a number of regression calibration-like correction methods to reduce the bias in the treatment effect estimator and obtain CIs with nominal coverage. In our simulation studies, these methods were effective in improving trial inferences when an external validation dataset (containing information about error-prone and error-free measurements) with at least 15 subjects was available.

To anticipate the impact of measurement error on trial inferences, knowledge is needed on the mechanism and magnitude of the measurement error. Endpoints that are measured with purely homoscedastic classical measurement error are expected to reduce the precision of treatment effect estimates and increase Type-II error at a given sample size, proportional to the relative amount of variance that is due to the error. Heteroscedastic classical error and differential error also affect Type-I error. Under systematic measurement error, only Type-I errors for testing null effects are expected to be at the nominal level. The treatment

effect estimator itself is biased by systematic error and differential error. Heteroscedastic error can be addressed using standard robust standard error estimators (e.g., HC3 [19]). Systematic error and differential error in the endpoint can be addressed via regression calibration-like correction methods.

We considered regression calibration-like correction methods that rely on an external validation set that contains information about both error-prone and error-free measurements. We anticipate such an external validation set can be feasible as a planned pilot study phase of a trial. Our simulation study shows that the effectiveness of correction methods to adjust the trial results for endpoint measurement error are dependent on the size of the validation sample and the strength of the correlation between the error-free and error-prone measurement of the trial endpoint. For a weak relation (R^2 = 0.20) we found the correction methods to be generally ineffective in improving trial inference with reasonably sized validation sets (i.e., up to size N = 50). However, for medium (R^2 = 0.50) or strong (R^2 = 0.80) correlations, the regression calibration showed improvements with external validation samples as small as 15 observations. With the relatively small validation samples (up to 50 observations), our study showed that the bootstrap performed best in constructing CIs in terms of coverage. The use of percentiles might explain that CIs were slightly conservative (i.e., too broad) for small validation samples (10 observations), which may be improved by using bias-corrected and accelerated bootstrap intervals [26]. The proposed regression calibration-like correction methods rely on a linear regression framework and can thus easily be extended to incorporate covariables in the trial analysis [27].

The use of measurement error corrections is still rare in applied biomedical studies with measurement error problems usually reported as an afterthought [9, 14]. Indeed, to our knowledge, no measurement error correction methods have been used so far in the analysis of biomedical trials to correct for measurement error in the endpoint. This may in part be due to a common misconception that measurement error can only affect trial inference by reducing the precision of estimating the effect of treatment and increasing Type-II error, which can be improved by increasing the study sample size. Note that our study demonstrates that such an assumption is warranted only when strict classical homoscedastic error structure of the trial endpoint can be assumed. Such does not hold, for instance, when measurement error are more pronounced in the tails of the distribution, or when measurement error vary between treatment arms.

Instead of the use of external validation datasets, internal measurement correction approaches where both the preferred endpoint and the error contaminated endpoint are measured on a subset of trial participants may sometimes be more feasible. For internal validation, Keogh et al. [7] recently reviewed methods of moment estimation and maximum likelihood estimation approaches. There are also other approaches to correct for measurement error that we did not discuss in this chapter. For instance, Cole and colleagues suggested a multiple imputing approach based on an internal validation set [28]. We also only focused on continuous outcomes in this chapter. Problems and solutions for misclassified categorical outcomes can be found elsewhere [29]. Yet, to the best of our knowledge, none of these methods have been tested in the setting where trial endpoints are measured with error and thus need further study.

Lastly, we solely discuss parametric measurement error models, which might misspecify the measurement error model. The extent to which the distribution of the unmeasured outcome can be estimated without parametric assumptions is a question for further

research. In the context of measurement error in explanatory variables this is formerly described as deconvolution ([10], Chapter 12 and references therein). Further, the method of non-parametric maximum likelihood has been successfully applied for explanatory variables measured with error [30, 31] and this might be an avenue of future research.

In summary, the impact of measurement error in a continuous endpoint on trial inferences can be particularly non-ignorable when the measurement error is not strictly random, because Type-I error, Type-II and the effect estimates can be affected. To alleviate the detrimental effects of measurement error we proposed measurement error corrected estimators and a variety of methods to construct CIs for non-random measurement error. To facilitate the implementation of these measurement error correction estimators we have developed the R package `mecor`, available at: https://github.com/LindaNab/mecor.

References

[1] E. Cerin, K. Cain, A. Oyeyemi, N. Owen, T. Conway, T. Cochrane, D. Van Dyck, J. Schipperijn, J. Mitáš, M. Toftager, I. Aguinaga-Ontoso, J. Sallis, Correlates of agreement between accelerometry and self-reported physical activity, Medicine & Science in Sports & Exercise 48 (6) (2016) 1075–1084. doi:10.1249/MSS.0000000000000870.

[2] M. Lauer, R. D'Agostino, The randomized registry trial: The next disruptive technology in clinical research?, New England Journal of Medicine 369 (17) (2013) 1579–1581. doi:10.1056/NEJMp1310102.

[3] I. Boutron, F. Tubach, B. Giraudeau, P. Ravaud, Blinding was judged more difficult to achieve and maintain in nonpharmacologic than pharmacologic trials, Journal of Clinical Epidemiology 57 (2004) 543–550. doi:10.1016/j.jclinepi.2003.12.010.

[4] H. Staudacher, P. Irving, M. Lomer, K. Whelan, The challenges of control groups, placebos and blinding in clinical trials of dietary interventions, Proceedings of the Nutrition Society 76 (2017) 203–112. doi:10.1017/S0029665117000350.

[5] S. Mahabir, D. Baer, C. Giffen, A. Subar, W. Campbell, T. Hartman, B. Clevidence, D. Albanes, P. Taylor, Calorie intake misreporting by diet record and food frequency questionnaire compared to doubly labeled water among postmenopausal women., European Journal of Clinical Nutrition 60 (2006) 561–565. doi:10.1038/sj.ejcn.1602359.

[6] S. Senn, S. Julious, Measurement in clinical trials: A neglected issue for statisticians?, Statistics in Medicine 28 (2009) 3189–3209. doi:10.1002/sim.3603.

[7] R. H. Keogh, R. J. Carroll, J. A. Tooze, S. I. Kirkpatrick, L. S. Freedman, Statistical issues related to dietary intake as the response variable in intervention trials, Statistics in Medicine 35 (25) (2016) 4493–4508. doi:10.1002/sim.7011.

[8] J. Buonaccorsi, Measurement error: Models, methods, and applications, Chapman & Hall/CRC, Boca Raton, FL, 2010.

[9] T. B. Brakenhoff, M. Mitroiu, R. H. Keogh, K. G. M. Moons, R. H. H. Groenwold, M. van Smeden, Measurement error is often neglected in medical literature: A systematic review, Journal of Clinical Epidemiology 98 (2018) 89–97. doi:10.1016/j.jclinepi.2018.02.023.

[10] R. Carroll, D. Ruppert, L. Stefanski, C. Crainiceanu, Measurement error in nonlinear models: A modern perspective, 2nd Edition, Chapman & Hall/CRC, Boca Raton, FL, 2006.

[11] W. Fuller, Measurement error models, John Wiley & Sons, New York, NY, 1987.

[12] P. Gustafson, Measurement error and misclassification in statistics and epidemiology: Impacts and Bayesian adjustments, Chapman & Hall/CRC, Boca Raton, FL, 2004.

[13] J. A. Hutcheon, A. Chiolero, J. A. Hanley, Random measurement error and regression dilution bias, BMJ 340 (c2289) (2010). doi:10.1136/bmj.c2289.

[14] P. A. Shaw, V. Deffner, R. H. Keogh, J. A. Tooze, K. W. Dodd, H. Küchenhoff, V. Kipnis, L. S. Freedman, Epidemiologic analyses with error-prone exposures: Review of current practice and recommendations, Annals of Epidemiology 28 (11) (2018) 821–828. doi:10.1016/j.annepidem.2018.09.001.

[15] M. Makrides, C. Crowther, R. Gibson, R. Gibson, C. Skeaff, Efficacy and tolerability of low-dose iron supplements during pregnancy: A randomized controlled trial, American Journal of Clinical Nutrition 78 (1) (2003) 145–153. doi:10.1093/ajcn/78.1.145.

[16] S. Zlotkin, P. Arthur, K. Antwi, G. Yeung, Randomized, controlled trial of single versus 3-times-daily ferrous sulfate drops for treatment of anemia, Pediatrics 108 (3) (2001) 613–616. doi:10.1542/peds.108.3.613.

[17] A. Patel, R. Wesley, S. Leitman, B. Bryant, Capillary versus venous haemoglobin determination in the assessment of healthy blood donors, Vox Sanguinis 104 (4) (2013) 317–323. doi:10.1111/vox.12006.

[18] R. H. Keogh, I. R. White, A toolkit for measurement error correction, with a focus on nutritional epidemiology, Statistics in Medicine 33 (12) (2014) 2137–2155. doi:10.1002/sim.6095.

[19] J. S. Long, L. H. Ervin, Using heteroscedasticity consistent standard errors in the linear regression model, The American Statistician 54 (3) (2000) 217–224. doi:10.2307/2685594.

[20] G. Fitzmaurice, Measurement error and reliability, Nutrition 18 (1) (2002) 112–114. doi:10.1016/s0899-9007(01)00624-4.

[21] J. Buonaccorsi, Measurement errors, linear calibration and inferences for means, Computational Statistics & Data Analysis 11 (3) (1991) 239–257. doi:10.1016/0167-9473(91)90083-E.

[22] B. Efron, Bootstrap methods: Another look at the jackknife, Annals of Statistics 7 (1) (1979) 1–26. doi:10.1214/aos/1176344552.

[23] L. S. Freedman, D. Midthune, L. Arab, R. L. Prentice, A. F. Subar, W. Willett, M. L. Neuhouser, L. F. Tinker, V. Kipnis, Combining a food frequency questionnaire with 24-hour recalls to increase the precision of estimation of usual dietary intakes—Evidence from the validation studies pooling project, American Journal of Epidemiology 187 (10) (2018) 2227–2232. doi:10.1093/aje/kwy126.

[24] A. Burton, D. Altman, P. Royston, R. Holder, The design of simulation studies in medical statistics, Statistics in Medicine 25 (24) (2006) 4279–4292. doi:10.1002/sim.2673.

[25] T. P. Morris, I. R. White, M. J. Crowther, Using simulation studies to evaluate statistical methods, Statistics in Medicine 38 (11) (2019) 2074–2102. doi:10.1002/sim.8086.

[26] P. Hall, Rejoinder: Theoretical comparison of bootstrap confidence intervals, The Annals of Statistics 16 (3) (1988) 981–985. doi:10.1214/aos/1176350944.

[27] S. Senn, Covariate imbalance and random allocation in clinical trials, Statistics in Medicine 8 (1989) 467–475. doi:10.1002/sim.4780080410.

[28] S. R. Cole, H. Chu, S. Greenland, Multiple-imputation for measurement-error correction, International Journal of Epidemiology 35 (4) (2006) 1074–1081. doi:10.1093/ije/dyl097.

[29] D. Brooks, K. Getz, A. Brennan, A. Pollack, M. Fox, The impact of joint misclassification of exposures and outcomes on the results of epidemiologic research, Current Epidemiology Reports 5 (2) (2018) 166–174. doi:10.1007/s40471-018-0147-y.

[30] S. Rabe-Hesketh, A. Pickles, A. Skrondal, Correcting for covariate measurement error in logistic regression using nonparametric maximum likelihood estimation, Statistical Modelling 3 (3) (2003) 215–232. doi:10.1191/1471082X03st056oa.

[31] S. Rabe-Hesketh, A. Skrondal, A. Pickles, Maximum likelihood estimation of generalized linear models with covariate measurement error, The Stata Journal: Promoting communications on statistics and Stata 3 (4) (2003) 386–411. doi:10.1177/1536867X0400300408.

3

Mecor: An R package for measurement error correction in linear models with a continuous outcome

Measurement error in a covariate or the outcome of regression models is common, but is often ignored, even though measurement error can lead to substantial bias in the estimated covariate-outcome association. While several texts on measurement error correction methods are available, these methods remain seldomly applied. To improve the use of measurement error correction methodology, we developed `mecor`*, an* R *package that implements measurement error correction methods for regression models with continuous outcomes. Measurement error correction requires information about the measurement error model and its parameters. This information can be obtained from four types of studies, used to estimate the parameters of the measurement error model: an internal validation study, a replicates study, a calibration study and an external validation study. In the package* `mecor`*, regression calibration methods and a maximum likelihood method are implemented to correct for measurement error in a continuous covariate in regression analyses. Additionally, methods of moments methods are implemented to correct for measurement error in the continuous outcome in regression analyses. Variance estimation of the corrected estimators is provided in closed form and using the bootstrap.*

This chapter is based on: L. Nab, M. van Smeden, R.H. Keogh and R.H.H. Groenwold. Mecor: An R package for measurement error correction in linear regression models with a continuous outcome, Computer Methods and Programs in Biomedicine 208 (2021) 106238. doi:10.1016/j.cmpb.2021.106238

3.1. Introduction

Measurement error is common across research fields, affecting the measurement of outcomes as well as important covariates. When left uncorrected, this can lead to severely biased and inefficient estimates of associations between covariates and outcome variables. Several texts have been published describing the impact of measurement error, and measurement error correction methodology [1–4]. However, recent reviews by Brakenhoff et al. [5] and Shaw et al. [6] show that, in biomedical research, measurement error correction methods remain seldomly applied. Keogh et al. [7] suggest that one of the main barriers to the use of correction methods may be the lack of accessible software. Moreover, as exemplified in [8], measurement is not only common in biomedical research, but in bioinformatics, chemistry, astronomy and econometrics as well. Therefore, to facilitate and encourage the use of measurement error correction methodology, we developed `mecor`, an R package that provides measurement error correction methods for linear models with continuous outcomes.

Several approaches to measurement error correction have been developed in the past decade. Examples include, simulation-extrapolation (SIMEX) by Cook et al. [9], multiple imputation for measurement error by Cole et al. [10], Bayesian correction (e.g., [4, 11]), maximum likelihood-based methods (e.g., [12, 13]), method of moments (MM) (e.g., [1]), and regression calibration (RC) introduced by Gleser [14] and Carroll et al. [15]. Of all these measurement error correction methods, RC is among the most commonly applied in biomedical research [6], possibly because of its relative simplicity and the possibility to implement it in conjunction with a variety of analysis types, e.g., linear regression [14, 15], survival analysis [16]), logistic regression [17] and other generalized linear models [2, 18].

In R [19], covariate measurement error correction by means of SIMEX is implemented in the package `simex` by Lederer et al. [20]. The R package `simexaft` by He et al. [21] provides SIMEX covariate measurement error correction for accelerated failure time models. A special issue of the `Stata` [22] Journal was published in 2003 and dedicated to measurement error models [23]. Three different methods were introduced for correction of measurement error in covariates in a generalized linear model. The `rcal` and `eivreg` procedure were introduced for RC by Hardin et al. [24], the `simex` and `simexplot` procedure were introduced for SIMEX by Hardin et al. [25] and, the `cme` procedure was introduced by Rabe-Hesketh et al. [26] for measurement error correction using a maximum likelihood approach. In `SAS`, multiple macros have been developed for measurement error correction. These macros include `%blinplus`, implementing the method by Rosner et al. [17]), `%relibpls8`, implementing the method by Rosner et al. [27], and `%rrc`, implementing the method by Liao et al. [28]), and the National Cancer Institute method macros, implementing the methods by Kipnis et al. [29]. An overview of available software including useful web links can be found in Table 4 and 5 of the paper by Keogh et al. [7]. Although several measurement error correction methods are available in `Stata` and `SAS`, to date RC-like methods for measurement error correction in a covariate have not been implemented in an R package. Moreover, no method for measurement error correction in a continuous outcome has been implemented in R.

In this paper we present and describe `mecor`, an R package for measurement error correction in linear regression models with a continuous outcome. Several methods (i.e., RC, MM and maximum likelihood) are implemented to correct covariate-outcome associations for measurement error in a covariate, or in the outcome. The package `mecor`

is flexible regarding the information that can be used to enable the measurement error correction, which can be of either of four types of measurement validation studies: an internal validation study, a replicates study, a calibration study and an external validation study. For each of these types of validation studies, standard RC, validation RC, efficient RC by Spiegelman et al. [30] and a maximum likelihood approach by Bartlett et al. [12] are implemented for measurement error correction in a covariate. For outcome measurement error correction, standard MM [1] and efficient MM [31] are available, for all different types of validation studies except replicates studies. The package `mecor` allows for random or systematic measurement error in a covariate, systematic measurement error in the outcome and, additionally, differential outcome measurement error in a univariable analysis. This broad spectrum of validation study types, measurement error models and correction methods in our easy-to-use software package should improve the application of measurement error corrections in research practice.

This paper is organized as follows. Section 3.2 introduces several measurement error models and the data structures of the four validation study types that can be used to estimate the parameters of the measurement error model. Section 3.3 outlines the measurement error correction methods. Section 3.4 introduces the functions in the package `mecor`. Section 3.5 demonstrates how the package `mecor` can be used in different settings using simulated example data.

3.2. Measurement error: notation, types and data structures

In this section, we introduce notation, derive expressions for the impact of measurement error on covariate-outcome associations and introduce the data structure of four different types of studies, that provide input for measurement error correction methods. Throughout, it is assumed that there is a continuous outcome Y, a continuous covariate X and a vector of k other covariates $\mathbf{Z} = (Z_1, Z_2, Z_3, \dots, Z_k)$. We consider measurement error in one variable at a time, i.e., in the covariate, X, or in the outcome, Y and assume that the other variables in the model are measured without error. Since our focus is on studies in which we aim to estimate the covariate-outcome association, the covariate X could be the main exposure of interest or a variable that confounds the relation between the main exposure and the outcome (one of the Z variables). The parameters of interest are $\boldsymbol{\beta} = (\beta_X, \beta_0, \boldsymbol{\beta}_Z)$ (with $\boldsymbol{\beta}_Z$ a $1 \times k$ matrix) from the linear model,

$$Y = \beta_X X + \beta_0 + \boldsymbol{\beta}_Z \mathbf{Z}' + e, \quad \text{Var}(e) = \sigma^2, \tag{3.1}$$

where we assume that $\text{E}(e) = 0$ and $\text{Cov}(e, X) = \text{Cov}(e, \mathbf{Z}) = 0$. This model will be referred to as the **outcome model**.

3.2.1. Types of measurement error and their impact

To quantify the impact of measurement error, we first define the assumed measurement error models. Subsequently, we outline the impact of measurement error in a covariate and the outcome on the estimates of the outcome model parameters, separately.

Covariate measurement error

Let X^* denote the error-prone substitute measure of the error-free reference measure X, following the measurement error model,

$$X^* = \theta_0 + \theta_1 X + U, \quad \text{Var}(U) = \tau^2, \tag{3.2}$$

and assume that $\text{E}(U) = 0$ and $\text{Cov}(U, X) = 0$. We assume non-differential covariate measurement error (i.e., $X^* \perp\!\!\!\perp Y | X, \mathbf{Z}$ or, equivalently, that the errors U are independent of the errors e in equation (3.1)). The measurement error is called 'classical' or 'random' if $\theta_0 = 0$ and $\theta_1 = 1$. The terms *classical measurement error* and *random measurement error* are used interchangeably in the literature. In this paper, we use the term random measurement error to refer to this type of measurement error. The measurement error is called 'systematic' for all other values of θ_0 and θ_1.

Suppose that there is one covariate $\mathbf{Z} = Z_1$ in the outcome model in (3.1), and that data on Y, X^* and Z_1 are available to fit the linear model,

$$\text{E}(Y | X^*, Z_1) = \beta_X^* X^* + \beta_0^* + \boldsymbol{\beta}_Z^* Z_1. \tag{3.3}$$

In this model, the least squares estimators $\hat{\boldsymbol{\beta}}^* = (\hat{\beta}_X^*, \hat{\beta}_0^*, \hat{\beta}_Z^*)$, are biased for $\boldsymbol{\beta}$, and consistent and unbiased estimators for $\boldsymbol{\beta}\boldsymbol{\Lambda}$ where $\boldsymbol{\Lambda}$ is the 3×3 **calibration model matrix**:

$$\boldsymbol{\Lambda} = \begin{pmatrix} \lambda_{X^*} & \lambda_0 & \lambda_{Z_1} \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

A well-known special case of the calibration model matrix is the attenuation factor. In particular, when there is random measurement error in the substitute error-prone measure X^*, we have $\beta_X^* = \lambda_{X^*}\beta$, where λ_{X^*} is called the attenuation factor [32] or regression dilution factor [33, 34]. When there is more than one $\mathbf{Z}$ covariate in the outcome model defined by equation (3.1), the **calibration model matrix** generalizes to the following $(2 + k) \times (2 + k)$ matrix:

$$\boldsymbol{\Lambda} = \begin{pmatrix} \lambda_{X^*} & \lambda_0 & \boldsymbol{\lambda}_\mathbf{Z} \\ \mathbf{0} & & I \end{pmatrix}, \tag{3.4}$$

where $\boldsymbol{\lambda}_\mathbf{Z}$ is a $1 \times k$ matrix, $\mathbf{0}$ is a $(1 + k) \times 1$ null matrix and $\mathbf{I}$ is a $(1 + k) \times (1 + k)$ identity matrix.

Outcome measurement error

Let Y^* denote the error-prone substitute measure of the error-free reference measure Y, following the measurement error model,

$$Y^* = \theta_0 + \theta_1 Y + U, \quad \text{Var}(U) = \tau^2, \tag{3.5}$$

and assume that $\text{E}(U) = 0$ and $\text{Cov}(U, Y) = 0$. We assume non-differential outcome measurement error (i.e., $Y^* \perp\!\!\!\perp X | Y, \mathbf{Z}$ or, equivalently, that the errors U are independent

of the errors e in equation (3.1)), unless specified otherwise. Random and systematic outcome measurement error are defined analogously to random and systematic covariate measurement error, respectively [35, 36].

Suppose, again, that there is one covariate $\mathbf{Z} = Z_1$ in the outcome model in (3.1) and that data on Y^*, X and Z_1 are available to fit the linear model,

$$\mathrm{E}[Y^*|X, Z_1] = \beta_X^* X + \beta_0^* + \beta_Z^* Z_1. \tag{3.6}$$

If the measurement error in Y^* is random, the least squares estimators $\hat{\boldsymbol{\beta}}^* = (\hat{\beta}_X^*, \hat{\beta}_0^*, \hat{\beta}_Z^*)$ are unbiased for $\boldsymbol{\beta}$. In contrast, if the error in Y^* is systematic, the least squares estimators $\hat{\boldsymbol{\beta}}^* = (\hat{\beta}_X^*, \hat{\beta}_0^*, \hat{\beta}_Z^*)$ are biased for $\boldsymbol{\beta}$ [1, 31, 36]. In order to identify consistent estimators for $\boldsymbol{\beta}$ by matrix multiplication, we add the integer 1 to the vector $\hat{\boldsymbol{\beta}}^*$. Then, $(\hat{\boldsymbol{\beta}}^*, 1)$ are consistent and unbiased estimators for $(\boldsymbol{\beta}, 1)\boldsymbol{\Theta}$ where $\boldsymbol{\Theta}$ is the 4×4 outcome **measurement error model matrix**:

$$\boldsymbol{\Theta} = \begin{pmatrix} \theta_1 & 0 & 0 & 0 \\ 0 & \theta_1 & 0 & 0 \\ 0 & 0 & \theta_1 & 0 \\ 0 & \theta_0 & 0 & 1 \end{pmatrix}.$$

When there is more than one $\mathbf{Z}$ covariate in the outcome model defined in equation (3.1), the calibration model matrix generalizes to the following $(2 + k + 1) \times (2 + k + 1)$ outcome **measurement error model matrix**:

$$\boldsymbol{\Theta} = \begin{pmatrix} \theta_1 & 0 & \dots & \dots & \dots & 0 \\ 0 & \ddots & \ddots & & & \vdots \\ \vdots & 0 & \ddots & \ddots & & \vdots \\ \vdots & \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & 0 & \dots & 0 & \theta_1 & 0 \\ 0 & \theta_0 & 0 & \dots & 0 & 1 \end{pmatrix}, \tag{3.7}$$

where $\hat{\boldsymbol{\Theta}}$ contains all zero's except on the diagonal and the $(2 + k + 1, 2)$th element.

Differential outcome measurement error in univariable analyses

We assume non-differential measurement error in the outcome in all but the following special case. Suppose exposure X is binary (e.g., in a two-arm controlled randomised trial) and that there are no other covariates $\mathbf{Z}$ in the outcome model defined by equation (3.1). Further, suppose that the measurement error in Y is differential such that the measurement error in the unexposed individuals (i.e., $X = 0$) is different from the measurement error in the exposed individuals (i.e., $X = 1$). Equivalently, let Y^* be the error-prone substitute measure of the error-free reference measure Y, with mean $\mathrm{E}(Y^*|Y, X) = \theta_{X0} + \theta_{X1} Y$ and variance τ^2, for $X = 0, 1$. Suppose now that data on Y^* and X are available to fit the linear model,

$$\mathrm{E}[Y^*|X] = \beta_X^* X + \beta_0^*.$$

In this model, the least squares estimators $\hat{\boldsymbol{\beta}}^* = (\hat{\beta}_X^*, \hat{\beta}_0^*)$ are biased for $\boldsymbol{\beta}$ [31, 36]. In order to identify consistent estimators for $\boldsymbol{\beta}$ by matrix multiplication, we again add the

integer 1 to the vector $\hat{\boldsymbol{\beta}}^*$. Then, $(\hat{\boldsymbol{\beta}}^*, 1)$ are consistent and unbiased estimators for $(\boldsymbol{\beta}, 1)\boldsymbol{\Theta}$ where, $\boldsymbol{\Theta}$ is the following 3×3 differential outcome **measurement error model matrix**:

$$\boldsymbol{\Theta} = \begin{pmatrix} \theta_{11} & 0 & 0 \\ \theta_{11} - \theta_{10} & \theta_{10} & 0 \\ \theta_{01} - \theta_{00} & \theta_{00} & 1 \end{pmatrix}. \tag{3.8}$$

3.2.2. Validation study data structures for measurement error correction

Four types of validation studies can be used to estimate the calibration model matrix or outcome measurement error model matrix defined in section Types of measurement error and their impact: an internal validation study, a replicates study, a calibration study or an external validation study [7, 37]. The first three validation studies make use of information internal to the study cohort, whereas the fourth makes use of information external to the study cohort.

Internal validation study

In an internal validation study, the error-free reference covariate values X or outcome values Y are observed in a subset of individuals (Table 3.1). Table 3.1a shows the structure of an internal validation study for covariate measurement error. In the main study, the outcome Y, the error-prone substitute covariate X^* and the covariates $\mathbf{Z}$ are measured in all n individuals. Additionally, in n_{sub} individuals ($n_{\text{sub}} < n$) the true covariate X is measured, assumed a random subset of the main study. As an example, suppose the true exposure of interest is visceral adipose tissue measurements (i.e., X) but that this is too expensive to obtain on all study participants and the error-prone substitute measure of waist circumference is instead collected for everyone (i.e, X^*) [38]. The same structure holds for an internal validation study for outcome measurement error, as shown in Table 3.1b.

Table 3.1: **Data structure of internal validation studies.** The true covariate or outcome is observed in a subset of the individuals from the main study. The superscript $*$ indicates that the variable was measured with error.

(a) Covariate-validation study

Y	X^*	$\mathbf{Z}$	X
y_1	x_1^*	$\mathbf{z_1}$	x_1
⋮	⋮	⋮	⋮
⋮	⋮	⋮	$x_{n_{\text{sub}}}$
⋮	⋮	⋮	-
⋮	⋮	⋮	⋮
y_n	x_n^*	$\mathbf{z_n}$	-

(b) Outcome-validation study

Y^*	X	$\mathbf{Z}$	Y
y_1^*	x_1	$\mathbf{z_1}$	y_1
⋮	⋮	⋮	⋮
⋮	⋮	⋮	$y_{n_{\text{sub}}}$
⋮	⋮	⋮	-
⋮	⋮	⋮	⋮
y_n^*	x_n	$\mathbf{z_n}$	-

Replicates study

A replicates study can be used if the measurement error in a covariate is random, denoted by X^{*r}. We will only use this type of study for covariate measurement error since random measurement error in an outcome does not result in biased association estimates (section Types of measurement error and their impact). In a replicates study, the error-prone

substitute covariate X^{*r} is repeatedly measured (i.e., m times, where $m \geq 2$) in all or in a random subset of individuals (Table 3.2). The repeated measures are denoted by $X_1^{*r}, \ldots, X_m^{*r}$. We assume that, in each individual, the same number of repeated measures was observed. Further, we assume that the measurement error in the replicates is jointly independent. Table 3.2a and 3.2b show the structure of a replicates study with full and partial replicates, respectively. In the main study, the outcome Y, the error-prone substitute covariate X_1^{*r} and the covariates $\mathbf{Z}$ are measured in all n individuals. Additionally, $n_{\text{sub}} \leq n$ individuals have m replicates of the error-prone substitute measure X_j^{*r} for $j = 2 \ldots m$. An example is the repeated measurement of several coronary risk factors in the Framingham Heart study, such as serum cholesterol, blood glucose, and systolic blood pressure [27].

Table 3.2: **Data structure of a covariate-replicates study for full or partial replicates.** The error-prone covariate is measured m times in all or a subset of individuals. The superscript $*_r$ indicates random measurement error.

(a) Full replicates study

Y	X_1^{*r}	$\mathbf{Z}$	X_2^{*r}	...	X_m^{*r}
y_1	x_{11}^{*r}	$\mathbf{z_1}$	x_{12}^{*r}	...	x_{1m}^{*r}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
y_n	x_{n1}^{*r}	$\mathbf{z_n}$	x_{n2}^{*r}	...	x_{nm}^{*r}

(b) Partial replicates study

Y	X_1^{*r}	$\mathbf{Z}$	X_2^{*r}	...	X_m^{*r}
y_1	x_{11}^{*r}	$\mathbf{z_1}$	x_{12}^{*r}	...	x_{1m}^{*r}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	...	$\vdots$
$\vdots$	$\vdots$	$\vdots$	$x_{n_{\text{sub}}2}^{*r}$	...	$x_{n_{\text{sub}}m}^{*r}$
$\vdots$	$\vdots$	$\vdots$	-	...	-
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
y_n	x_{n1}^{*r}	$\mathbf{z_n}$	-	...	-

Calibration study

A calibration study is a special type of sub-study where two types of error-prone substitute measurement methods are used to measure the covariate or outcome: substitute measurement prone to systematic measurement error and a substitute measurement prone to random measurement error (Table 3.3). Table 3.3a shows the structure of a calibration study for covariate measurement error. All n individuals in the main study have obtained measures of the outcome Y, the error-prone substitute covariate X^{*s} and the covariates $\mathbf{Z}$. The error-prone substitute covariate X^{*s} is systematically different from X, or, $\mathrm{E}(X^{*s}|X) \neq X$ (systematic measurement error). Additionally, a random subset of n_{sub} individuals ($n_{\text{sub}} < n$) have m replicates of the error-prone substitute measure X_j^{*r}, where $\mathrm{E}(X_j^{*r}|X) = X$ for $j = 1 \ldots m$ (random measurement error). The same structure holds for a calibration study for outcome measurement error, as shown in Table 3.3b. An example of an calibration study for outcome measurement error is a study of sodium intake measured by a 24-hour recall (assumed systematic measurement error) and urinary biomarkers (assumed random measurement error) [31].

External validation study

In an external validation study the error-free reference covariate values X or outcome values Y are observed in a small set of individuals not included in the main study (Table 3.4). Table 3.4a shows the structure of an external validation study for covariate measurement error. In all n individuals in the main study measures are obtained of outcome Y, the error-prone substitute covariate X^* and the covariates $\mathbf{Z}$. Additionally, there is an external data set

Table 3.3: **Data structure of calibration studies.** Two types of error-prone measurement methods are used to measure the covariate or outcome. The superscripts $*_r$ and $*_s$ indicate random and systematic measurement error, respectively.

(a) Covariate-calibration study

Y	X^{*_s}	$\mathbf{Z}$	$X_1^{*_r}$	...	$X_m^{*_r}$
y_1	$x_1^{*_s}$	$\mathbf{z_1}$	$x_{11}^{*_r}$	...	$x_{1m}^{*_r}$
⋮	⋮	⋮	⋮	...	⋮
⋮	⋮	⋮	$x_{n_{\text{sub}}1}^{*_r}$	...	$x_{n_{\text{sub}}m}^{*_r}$
⋮	⋮	⋮	-	...	-
⋮	⋮	⋮	⋮	⋮	⋮
y_n	$x_n^{*_s}$	$\mathbf{z_n}$	-	...	-

(b) Outcome-calibration study

Y^{*_s}	X	$\mathbf{Z}$	$Y_1^{*_r}$	...	$Y_m^{*_r}$
$y_1^{*_s}$	x_1	$\mathbf{z_1}$	$y_{11}^{*_r}$	...	$y_{1m}^{*_r}$
⋮	⋮	⋮	⋮	...	⋮
⋮	⋮	⋮	$y_{n_{\text{sub}}1}^{*_r}$	...	$y_{n_{\text{sub}}m}^{*_r}$
⋮	⋮	⋮	-	...	-
⋮	⋮	⋮	⋮	⋮	⋮
$y_n^{*_s}$	x_n	$\mathbf{z_n}$	-	...	-

comprising of individuals on whom measures are obtained of the error-free reference covariate X, the error-prone substitute covariate X^* and the other covariates $\mathbf{Z}$. Table 3.4b shows the structure of an external validation study for outcome measurement error. In this setting, there is an external data set comprising of individuals of whom measures are obtained of the error-free reference outcome Y and the error-prone substitute outcome Y^*. The external data set does not need to comprise measures of the covariates. An example of an external validation study for outcome measurement error is a trial designed to study the efficacy of iron supplementation in pregnant women where haemoglobin is measured in capillary blood samples (error-prone substitute measure) instead of in venous blood samples (error-free reference measure) [36].

Table 3.4: **Data structure of external validation studies.** An error-prone covariate or outcome is measured in the main study and the true covariate or outcome is measured in a small external set. The superscript $*$ indicates that there is random or systematic measurement error in the variables

(a) External covariate-validation study

Y	X^*	$\mathbf{Z}$	X	X^*	$\mathbf{Z}$
y_1	x_1^*	$\mathbf{z_1}$	x_1	x_1^*	$\mathbf{z_1}$
⋮	⋮	⋮	⋮	⋮	⋮
⋮	⋮	⋮	$x_{n_{\text{ex}}}$	$x_{n_{\text{ex}}}^*$	$\mathbf{z_{n_{ex}}}$
y_n	x_n^*	$\mathbf{z_n}$			
Main study			*External*		

(b) External outcome-validation study

Y^*	X	$\mathbf{Z}$	Y	Y^*
y_1^*	x_1	$\mathbf{z_1}$	y_1	y_1^*
⋮	⋮	⋮	⋮	⋮
⋮	⋮	⋮	$y_{n_{\text{ex}}}$	$y_{n_{\text{ex}}}^*$
y_n^*	x_n	$\mathbf{z_n}$		
Main study			*External*	

3.3. Measurement error correction

In section Types of measurement error and their impact, the calibration model matrix $\mathbf{\Lambda}$ and the measurement error model matrix $\mathbf{\Theta}$ were introduced. These matrices quantify the bias in the naive analysis, i.e., the analysis that does not take the measurement error in X^* or Y^* into account. In the following sections, measurement error correction methods are introduced that utilize the matrices $\mathbf{\Lambda}$ and $\mathbf{\Theta}$.

The standard method for covariate measurement error correction that uses the

calibration model matrix $\mathbf{\Lambda}$ is *standard regression calibration (RC)* [14, 15]. *Standard RC* can be applied in all four types of studies from the previous section. In addition, *validation RC*, an adapted version of *standard RC* for internal validation studies, is the standard covariate measurement error correction method for internal validation studies [2]. Further, the standard method for outcome measurement error correction that uses the measurement error model matrix $\mathbf{\Theta}$ is *standard method of moments (MM)* [1]. *Standard MM* can be applied in internal and external validation studies, and calibration studies.

Standard RC and *standard MM* do not make the most efficient use of the information available in internal validation studies and calibration studies [2]. More efficient methods for measurement error correction methods are therefore implemented in `mecor`. A more efficient RC estimator, called *efficient RC*, was introduced by Spiegelman et al. [30]. A more efficient MM estimator was introduced by Keogh et al. [31], which is called the Buonaccorsi approach using the method of moments. For simplicity, we will refer to this method as *efficient MM*.

Likewise, in replicates studies, *standard RC* does not make the most efficient use of the information available [33]. The *standard RC* method is sub-optimal in terms of efficiency, since the method depends on the ordering of the replicate measurements [33]. This can be intuitively understood as follows. The *standard RC* regresses the mean of all but the first replicate on the first replicate, but this could as easily be exchanged with the second replicate. Therefore, different approaches are possible (e.g., maximum likelihood) [33]. Bartlett et al. [12] showed how a standard random-intercepts model can be used to obtain *maximum likelihood* (ML) estimates that are more efficient than *standard RC*, at the cost of some additional parametric assumptions, discussed in section Maximum likelihood estimation for replicates studies.

Section Standard measurement error correction introduces *standard RC* and *validation RC* for covariate measurement error correction, and *standard MM* for outcome measurement error correction. *Efficient RC* and *efficient MM* are introduced in section More efficient measurement error correction and the maximum likelihood approach for replicates studies is introduced in section Maximum likelihood estimation for replicates studies. When no information is available to estimate the parameters of the measurement error model, a *sensitivity analysis* or *quantitative bias analysis* can be used to analyse the sensitivity of study results to measurement error [39, 40]. An approach for conducting *sensitivity analyses* is discussed in section Sensitivity analyses.

3.3.1. Standard measurement error correction

Covariate measurement error

In *standard RC*, the biased least squares estimator $\hat{\boldsymbol{\beta}}^*$ is multiplied by the inverse of an estimate of the calibration model matrix $\mathbf{\Lambda}$ to give a consistent and unbiased estimator of $\boldsymbol{\beta}$, denoted $\hat{\boldsymbol{\beta}}_{\text{RC}}$:

$$\hat{\boldsymbol{\beta}}_{\text{RC}} = \hat{\boldsymbol{\beta}}^* \hat{\mathbf{\Lambda}}^{-1} \tag{3.9}$$

Standard RC can be applied using all four types of validation studies (section Validation study data structures for measurement error correction).

To construct the calibration model matrix $\mathbf{\Lambda}$ (see equation (3.4)), we estimate its

components $\boldsymbol{\lambda} = (\lambda_{X^*}, \lambda_0, \boldsymbol{\lambda}_Z)$, from the linear calibration model:

$$\mathrm{E}(X|X^*, \mathbf{Z}) = \lambda_{X^*} X^* + \lambda_0 + \boldsymbol{\lambda}_\mathbf{Z} \mathbf{Z}', \tag{3.10}$$

using least squares. Here, $\boldsymbol{\lambda}_\mathbf{Z}$ is a $1 \times k$ matrix. Throughout, we assume that the calibration model matrix is correctly specified. To obtain estimates of the parameters of interest $\boldsymbol{\lambda}$ in an internal validation study (Table 3.1a) and external validation study (Table 3.4a), the error-free reference measure X is regressed on the error-prone substitute measure X^* and the other covariates $\mathbf{Z}$. To obtain estimates of the parameters of interest $\boldsymbol{\lambda}$ in a replicates study (Table 3.2a), the mean of all replicates except the first replicate (i.e., $X_2^{*r}, \ldots, X_m^{*r}$) is regressed on the first replicate X_1^* and the other covariates $\mathbf{Z}$. To obtain estimates of the parameters of interest $\boldsymbol{\lambda}$ in a calibration study (Table 3.3a), the mean of the replicates $X_1^{*r}, \ldots, X_m^{*r}$ with random measurement error is regressed on the measurement X^{*s} with systematic measurement error and the other covariates $\mathbf{Z}$.

An adapted version of *standard RC* in internal validation studies is *validation RC* [2]. In *validation RC*, the outcome Y is regressed on the calibrated values X_{cal} and $\mathbf{Z}$. The calibrated values X_{cal} are constructed as follows: if X is observed, $X_{\text{cal}} = X$, and if X is not observed, $X_{\text{cal}} = \mathrm{E}(X|X^*, \mathbf{Z})$. The parameters from the regression of Y on X_{cal} and $\mathbf{Z}$ are estimates of our parameters of interest $\boldsymbol{\beta}$ in equation (3.5). Note that *standard RC* described above is identical to using $X_{\text{cal}} = \mathrm{E}(X|X^*, \mathbf{Z})$ for all X [7].

Outcome measurement error

In *standard MM*, the biased least squares estimator $\hat{\boldsymbol{\beta}}^*$ is multiplied by the inverse of an estimate of the outcome measurement error model matrix $\boldsymbol{\Theta}$ to give a consistent and unbiased estimator of $\boldsymbol{\beta}$, denoted $\hat{\boldsymbol{\beta}}_{\text{MM}}$:

$$\hat{\boldsymbol{\beta}}_{\text{MM}} = (\hat{\boldsymbol{\beta}}^*, \mathbf{1})\hat{\boldsymbol{\Theta}}^{-1}. \tag{3.11}$$

Standard MM can be applied using internal and external validation studies, and calibration studies (section Validation study data structures for measurement error correction).

To construct the outcome measurement error model matrix $\boldsymbol{\Theta}$ (see equation (3.7)), we estimate its components $\boldsymbol{\theta} = (\theta_0, \theta_1)$ from the linear measurement error model $\mathrm{E}(Y^*|Y) = \theta_0 + \theta_1 Y$ using least squares. Throughout, we assume that the measurement error model matrix is correctly specified. To obtain estimates of the parameters of interest $\boldsymbol{\theta}$ in an internal validation study (Table 3.1b) and an external validation study (Table 3.4b), the error-prone substitute measurement Y^* is regressed on the error-free reference measurement Y. To obtain estimates of the parameters of interest $\boldsymbol{\theta}$ in a calibration study (Table 3.3b), the measurement Y^{*s} with systematic measurement error is regressed on the mean of the replicates $Y_1^{*r}, \ldots, Y_m^{*r}$ with random measurement error, thereby correcting for the bias in the estimated $\hat{\boldsymbol{\theta}}$ using standard RC (implying that $m > 1$).

Differential outcome measurement error in univariable analyses

For the special case of differential measurement error, the outcome measurement error model matrix $\boldsymbol{\Theta}$ (see equation (3.8)), can be constructed as follows. We estimate its components $\boldsymbol{\theta} = (\theta_{00}, \theta_{01}, \theta_{10}, \theta_{11})$ from the measurement error model $\mathrm{E}(Y^*|Y, X) = \theta_{00} + (\theta_{01} - \theta_{00})X + \theta_{10}Y + (\theta_{11} - \theta_{10})XY$. This model can be fitted directly in an internal validation study (Table 3.1b), provided that the random internal subset includes exposed

(i.e., $X = 1$) and non-exposed individuals (i.e., $X = 0$). The model can be fitted in an external validation study (Table 3.4b), provided that X is measured, and that exposed and non-exposed individuals are included in the external set. In a calibration study (Table 3.3b), the measurement with systematic measurement error is regressed on the mean of the replicates $Y_1^{*r}, \ldots, Y_m^{*r}$ with random measurement error and the covariate X (again, provided that the random subset with replicates with random measurement error includes exposed and non-exposed individuals).

Variance estimation

The variance of the *standard RC* estimator can be estimated using the multivariate delta method [17] or the zero-variance method [41]. Confidence intervals can then be obtained by constructing Wald-type confidence intervals using one of the former two methods. Additionally, confidence intervals can be obtained by the stratified bootstrap, by sampling the observations in the internal subset separately from the observations outside the internal subset. The variance of the *standard MM* estimator can also be estimated with the multivariate delta method, the zero-variance method or the stratified bootstrap. Additionally, for *standard RC*, confidence intervals for $\hat{\beta}_{X\,\mathrm{RC}}$ (the first element of the $\hat{\boldsymbol{\beta}}_{\mathrm{RC}}$) can be obtained by the Fieller method [33]. For *standard MM*, confidence intervals for $\hat{\beta}_{X\,\mathrm{MM}}$ and $\hat{\boldsymbol{\beta}}_{\mathbf{Z}\,\mathrm{MM}}$ (the first two elements of the $\hat{\boldsymbol{\beta}}_{\mathrm{MM}}$) can be obtained by the Fieller method [36]. Details of these procedures can be found in section S3.1 of the supplementary materials.

3.3.2. More efficient measurement error correction

Covariate measurement error

Efficient RC can be used in internal validation studies or calibration studies [30]. It pools the *standard RC* estimate with an internal estimate for $\boldsymbol{\beta}$ obtained in the internal validation study or calibration study.

In internal validation studies, the error-free reference covariate X is obtained in an internal subset of the main study (Table 3.1a). By regressing the outcome Y on X and the other covariates $\mathbf{Z}$ using least squares in the internal subset, one obtains an unbiased estimate for our parameters of interest $\boldsymbol{\beta}$. Denote this estimator by $\hat{\boldsymbol{\beta}}_{\mathrm{I}}$. This internal estimator $\hat{\boldsymbol{\beta}}_{\mathrm{I}}$ can then be combined with the standard RC estimator $\hat{\boldsymbol{\beta}}_{\mathrm{RC}}$ defined in equation (3.9), by taking the inverse variance weighted mean of the two estimates:

$$\hat{\boldsymbol{\beta}}_{\mathrm{ERC}} = [\hat{\Sigma}^{-1}_{\beta_{\mathrm{RC}}} + \hat{\Sigma}^{-1}_{\beta_{\mathrm{I}}}]^{-1}[\hat{\Sigma}^{-1}_{\beta_{\mathrm{RC}}}\hat{\beta_{\mathrm{RC}}} + \hat{\Sigma}^{-1}_{\beta_{\mathrm{I}}}\hat{\beta}_{\mathrm{I}}], \tag{3.12}$$

where $\hat{\Sigma}^{-1}_{\beta_{\mathrm{RC}}}$ is the variance–covariance matrix obtained from the multivariate delta method and $\hat{\Sigma}_{\beta_{\mathrm{I}}}$ is the standard variance–covariance matrix of a least squares estimator. The efficient RC estimator defined above is an unbiased, consistent and the most efficient estimator for $\boldsymbol{\beta}$ if sampling into the internal validation set is unbiased (e.g., if the validation study is a random subset of participants) [30].

In calibration studies, the covariate X is observed with random measurement error in an internal subset of the main study (Table 3.3a). If at least 2 replicates are available, an unbiased estimator for $\boldsymbol{\beta}$ can be obtained by using the standard RC estimator for a replicates study (see section Standard measurement error correction) in the internal subset. Again, denote this estimator by $\hat{\boldsymbol{\beta}}_{\mathrm{I}}$. Then, the estimate obtained from the internal subset can be

pooled with the standard RC estimate following equation (3.12). Alternatively, an unbiased estimator for $\boldsymbol{\beta}$ using the replicates in the internal subset can be obtained by using the ML estimation discussed in section Maximum likelihood estimation for replicates studies. Again, this estimate can then be pooled with the standard RC estimate following equation (3.12).

Outcome measurement error

Efficient MM can be used in internal validation studies or calibration studies [31]. It pools the *standard MM* estimate with an internal estimate for $\boldsymbol{\beta}$ obtained in the internal validation study or calibration study.

In internal validation studies, the error-free reference outcome Y is obtained in an internal subset of the main study (Table 3.1b). By regressing Y on the covariates X and $\mathbf{Z}$ using least squares in the internal subset, one obtains an unbiased estimator for $\boldsymbol{\beta}$. Denote this estimator by $\hat{\boldsymbol{\beta}}_{\text{I}}$. In calibration studies, the outcome is observed with random measurement error in an internal subset of the main study (Table 3.3b). The internal estimator $\hat{\boldsymbol{\beta}}_{\text{I}}$ is obtained by regressing the outcome $Y^{*,r}$ with random measurement error on the covariates X and $\mathbf{Z}$ using least squares in the internal subset. Using the outcome with random measurement error will lead to the unbiased estimation of the association under study since random outcome measurement error does not bias the association. A single measurement with random measurement error (i.e., $m = 1$ in Table 3.1b) is sufficient to obtain an internal estimate. However, if the outcome with random measurement error is observed more than once, the mean of the measures $Y_1^{*r}, \dots, Y_m^{*r}$ can be used and regressed on the covariates X and $\mathbf{Z}$. Subsequently, the estimate obtained from the internal subset in an internal validation study or calibration study can be pooled with the *standard MM* estimate following equation (3.12), by replacing the *standard RC* estimate with the *standard MM* estimate in the equation.

Differential outcome measurement error in univariable analyses

In internal validation studies, the internal estimator $\hat{\boldsymbol{\beta}}_{\text{I}}$ can be obtained by regressing Y on the covariates X and $\mathbf{Z}$ using least squares. In calibration studies, the internal estimator $\hat{\boldsymbol{\beta}}_{\text{I}}$ can be obtained by regressing the outcome $Y^{*,r}$ with random measurement error on the covariates X and $\mathbf{Z}$. A single measurement with random measurement error (i.e., $m = 1$ in Table 3.1b) is sufficient to obtain an internal estimate. However, if the outcome with random measurement error is observed more than once, the mean of the measures $Y_1^{*r}, \dots, Y_m^{*r}$ can be used and regressed on the covariates X and $\mathbf{Z}$. We assume that the internal subset is a random subset of the main study, and hence that exposed and unexposed are included in the internal subset. Subsequently, the estimate obtained from the internal subset in an internal validation study or calibration study can be pooled with the *standard MM* estimate following equation (3.12), by replacing the *standard RC* estimate with the *standard MM* estimate in the equation.

Variance estimation

The variance of the *efficient RC* estimator can be obtained from the following:

$$\hat{\Sigma}_{\beta_{\text{ERC}}} = [\hat{\Sigma}_{\beta}^{-1} + \hat{\Sigma}_{\beta_{\text{I}}}^{-1}]^{-1}.$$

The variance of the *efficient RC* estimator can also be obtained by stratified bootstrapping, by sampling the observations in the internal subset separately from the observations outside the internal subset. Confidence intervals can be obtained by constructing Wald-type confidence intervals using one of the former two variances or by stratified percentile bootstrap. The same applies for the *efficient MM* estimator.

3.3.3. Maximum likelihood estimation for replicates studies

The use of a standard random-intercepts model to obtain maximum likelihood (ML) estimates for $\boldsymbol{\beta}$ in replicates studies was introduced by Bartlett et al. [12]. To explain the *ML* method for replicates studies, we add the index $i = 1, \dots, n$ to our notation in the outcome model:

$$Y_i = \beta_X X_i + \beta_0 + \boldsymbol{\beta}_Z \mathbf{Z}_i' + e_i, \qquad \text{Var}(e_i) = \sigma^2,$$

where we again assume that $\text{E}(e_i) = 0$ and $\text{Cov}(e_i, X_i) = \text{Cov}(e_i, \mathbf{Z}_i) = 0$. Further, $\mathbf{Z}_i = (Z_{i1}, \dots, Z_{ik})$ and $\boldsymbol{\beta}_Z$ is again a $1 \times k$ matrix. On top of these assumptions, we also assume that the e_i are normal and independently distributed. Additionally, assume that X_i is normally distributed given $\mathbf{Z}_i$, with,

$$\text{E}(X_i|\mathbf{Z}_i) = \rho_0 + \boldsymbol{\rho}_\mathbf{Z} \mathbf{Z}_i' \quad \text{and} \quad \text{Var}(X_i|\mathbf{Z}_i) = \sigma^2_{X_i|\mathbf{Z}_i},$$

where $\boldsymbol{\rho}_Z$ is a $1 \times k$ matrix. In a replicates study, X_i is not observed. Instead, m replicates of the error-prone measurement $\mathbf{X}_i^{*r} = (X_{i1}^{*r}, \dots, X_{im}^{*r})$ are observed, for $i = 1, \dots, n$. In a full-replicates study (Table 3.2a), we assume that the number of replicate measurements $m \geq 2$ is constant for every individual. In a partial-replicates study (Table 3.2b), we assume that the number of replicates $m \geq 2$ is constant in the replicate sub-study and $m = 1$ in the main study. These measurements are assumed to follow the following random measurement error model:

$$X_{ij}^{*r} = X_i + U_{ij}, \qquad \text{Var}(U_{ij}) = \tau^2, \qquad j = 1, \dots, m,$$

where we again assume that $\text{E}(U_{ij}) = 0$, $\text{Cov}(U_{ij}, X_i) = 0$, and that the measurement error in non-differential, i.e., the errors U_{ij} are independent of the errors e_i in the outcome model described above. In addition, we also assume that the errors U_{ij} are normal and independently distributed.

We consider the likelihood function when only Y_i, $\mathbf{X}_i^{*r}$ and $\mathbf{Z}_i$ are observed. The log likelihood can be factorized as follows:

$$\ell(\boldsymbol{\theta}|Y_i, \mathbf{X}_i^{*r}, \mathbf{Z}_i) = \log(f(Y_i|\mathbf{Z}_i, \boldsymbol{\theta})) + \log(f(\mathbf{X}_i^{*r}|Y_i, \mathbf{Z}_i, \boldsymbol{\theta})), \tag{3.13}$$

where $\boldsymbol{\theta} = (\beta_X, \beta_0, \boldsymbol{\beta}_Z, \sigma^2, \rho_0, \boldsymbol{\rho}_\mathbf{Z}, \sigma^2_{X|\mathbf{Z}}, \tau^2)$. From the assumptions that $X_i|\mathbf{Z}_\mathbf{i}$ is normally distributed, the e_i are normally distributed and that $X_i|\mathbf{Z}_\mathbf{i}$ and e_i are independent, [12] show that Y_i given $\mathbf{Z}_i$ is normal with mean $\delta_0 + \boldsymbol{\delta}_\mathbf{Z} \mathbf{Z}_i$ and variance $\sigma^2_{Y|\mathbf{Z}}$, where $\boldsymbol{\delta}_Z$ is a $1 \times k$ matrix. Furthermore, since $X_i|\mathbf{Z}_\mathbf{i}$ and $Y_i|\mathbf{Z}_\mathbf{i}$ are jointly normal, $X_i|Y_i, \mathbf{Z}_i$ is also normal. [12] show that we can therefore write:

$$X_i = \kappa_0 + \kappa_Y Y_i + \boldsymbol{\kappa}_\mathbf{Z} \mathbf{Z}_i + b_i,$$

where $b_i \sim \text{N}(0, \sigma^2_{X|Y,\mathbf{Z}})$. Then, since $X_{ij}^* = X_i + U_{ij}$, it follows from the above equation that,

$$X_{ij}^* = \kappa_0 + \kappa_Y Y_i + \boldsymbol{\kappa}_\mathbf{Z} \mathbf{Z}_i + b_i + U_{ij},$$

where $U_{ij} \sim N(0, \tau^2)$ is independent of b_i [12] and $\boldsymbol{\kappa}_Z$ is a $1 \times k$ matrix. Hence, $\mathbf{X}_i^{*r}$ given Y_i and $\mathbf{Z}_i$ follows a random-intercepts model with fixed effects of Y_i and $\mathbf{Z}_i$, random intercepts variance $\sigma^2_{X|Y,\mathbf{Z}}$ and within subject variance τ^2.

The parameter vector $\boldsymbol{\zeta} = (\delta_0, \boldsymbol{\delta}_\mathbf{Z}, \sigma^2_{Y|\mathbf{Z}}, \kappa_0, \kappa_Y, \boldsymbol{\kappa}_\mathbf{Z}, \sigma^2_{X|Y,\mathbf{Z}}, \tau^2)$ is a one-to-one function of the original model parameter vector $\boldsymbol{\theta} = (\beta_X, \beta_0, \boldsymbol{\beta}_Z, \sigma^2, \rho_0, \boldsymbol{\rho}_\mathbf{Z}, \sigma^2_{X|\mathbf{Z}}, \tau^2)$. Accordingly, Bartlett et al. [12] show that the ML estimate for ζ can be obtained by maximizing the two likelihood components of equation (3.13) separately. The likelihood component corresponding to $f(Y_i|\mathbf{Z_i}, \boldsymbol{\zeta})$ in equation (3.13) can be maximized by fitting the least squares regression of Y_i on $\mathbf{Z}_i$. The likelihood component corresponding to $f(\mathbf{X}_i^{*r}|Y_i, \mathbf{Z}_i, \boldsymbol{\zeta})$ in equation (3.13) can be maximized by fitting a random-intercepts model for $\mathbf{X}_i^{*r}$ given Y_i and $\mathbf{Z}_i$.

An ML estimate for $\boldsymbol{\beta}$ can now be obtained by the following formulas:

$$\begin{aligned}
\beta_X &= \kappa_Y \times \frac{\sigma^2_{Y|\mathbf{Z}}}{\sigma^2_{X|Y,\mathbf{Z}} + \kappa_Y^2 \sigma^2_{Y|\mathbf{Z}}}, \\
\beta_0 &= \delta_0 - \beta_X \rho_0 = \delta_0 - \beta_X\{\kappa_0 + \kappa_Y \delta_0\}, \\
\boldsymbol{\beta}_\mathbf{Z} &= \boldsymbol{\delta}_\mathbf{Z} - \beta_X \boldsymbol{\rho}_\mathbf{Z} = \boldsymbol{\delta}_\mathbf{Z} - \beta_X\{\boldsymbol{\kappa}_\mathbf{Z} + \kappa_Y \boldsymbol{\delta}_\mathbf{Z}\}.
\end{aligned}$$

The estimator $\hat{\boldsymbol{\beta}}_{\mathsf{ML}} = (\hat{\beta}_{X_{\mathsf{ML}}}, \hat{\beta}_{0_{\mathsf{ML}}}, \hat{\boldsymbol{\beta}}_{Z_{\mathsf{ML}}})$ can be obtained by replacing the parameters from parameter vector $\boldsymbol{\zeta}$ by their estimates in the above equations.

Variance estimation

The variance of the maximum likelihood estimator can be estimated with the multivariate delta method [12]. Confidence intervals can then be obtained by constructing Wald-type confidence intervals. Confidence intervals can also be obtained by stratified bootstrap, by sampling the observations in the internal subset separately from the observations outside the internal subset. Details of these procedures can be found in the supplementary material section S3.2.

3.3.4. Sensitivity analyses

Information from a validation study may not always be available. In that case, a formal correction is not possible. Nevertheless, when measurement error in a covariate or the outcome is expected, one may check how sensitive study results are to that measurement error. Literature or expert knowledge can be used to inform this sensitivity analysis, e.g., by hypothesizing possible ranges for the parameter values of the measurement model.

When random covariate measurement error is expected, speculation is needed of the values of τ^2, i.e., the variance of the random measurement error. Additionally, when systematic covariate measurement error is suspected, speculation is needed about the parameter values of the calibration model described by equation (3.10). When systematic outcome measurement error is suspected, speculation is needed about the parameter values of the outcome measurement error model, described in equation (3.5).

3.4. The R package mecor

The R package mecor offers functionality to correct for measurement error in a continuous covariate or outcome in linear models with a continuous outcome. The main model fitting

function in mecor is mecor:

```
mecor(formula, data, method, B)
```

The function fits the linear model defined in formula, corrected for the measurement error in one of the variables. The arguments are as follows:

- formula a formula object, with the response on the left of a '~' operator and the terms, separated by + operators, on the right. This argument takes the form outcome ~ MeasError(substitute, reference, replicate, differential) + covariates for covariate measurement error, and MeasError(substitute, reference, replicate, differential) ~ covariates for outcome measurement error. The MeasError object can be used for measurement error correction in internal validation, replicates and calibration studies. For external validation studies or sensitivity analyses of systematic measurement error, the object MeasErrorExt(substitute, model) is used instead of a MeasError object. For sensitivity analyses of random measurement error, the object MeasErrorRandom(substitute, error) is used.
- data a data.frame containing the variables in the model specified by formula.
- method specifies the method used for measurement error correction. The options are "standard" for standard RC and standard MM, "valregcal" for validation RC, "efficient" for efficient RC and efficient MM, and "mle" for maximum likelihood estimation.
- B number of bootstrap samples used for standard error estimation. The default is set to 0.

An object of class mecor can be summarised using the summary function:

```
summary(object, alpha, zerovar, fieller)
```

The arguments are as follows:

- object an object of class mecor.
- alpha a numeric indicating the probability of obtaining a type II error. Defaults to 0.05.
- zerovar a boolean indicating whether confidence intervals using the zero-variance method [41] must be printed. Only available for mecor objects fitted with method equal to "standard". Defaults to FALSE.
- fieller a boolean indicating whether confidence intervals using the fieller method [33, 36] must be printed. Only available for mecor objects fitted with method equal to "standard". Defaults to FALSE.

The default summary object of an object of class mecor prints standard errors and confidence intervals obtained by the delta method. See the various 'Variance estimation' paragraphs in section 3.3 for a description of the methods for variance estimation.

The formula argument in mecor contains a MeasError object, a MeasErrorExt object or a MeasErrorRandom object. All three objects are described below.

3.4.1. The MeasError object

To correct for measurement error using an internal validation study, a replicates study or a calibration study, the `formula` argument in `mecor` contains a `MeasError` object on the right-hand side (covariate measurement error) or left-hand side (outcome measurement error). The `MeasError` object can be used for random and systematic measurement error correction, depending on the method used to correct for the measurement error in `mecor`:

```
MeasError(substitute, reference, replicate, differential)
```


with the arguments being described as follows:

- `substitute` the error-prone substitute measurement;
- `reference` the gold-standard reference measurement, to be used in case of an internal validation study, else `NULL`;
- `replicate` (a vector of) the replicate measurement of the error-prone substitute measurement, to be used in case of a replicates study or calibration study, else `NULL`;
- `differential` the binary exposure on which the outcome measurement error structure is dependent, to be used for differential outcome measurement error in univariable analyses, else `NULL`.

Depending on the type of validation study used, either argument `reference` (internal validation study) or `replicate` (replicates study or calibration study) can be used, but never both.

3.4.2. The MeasErrorExt object

To correct for measurement error using an external validation study, the `formula` object in `mecor` contains a `MeasErrorExt` object on the right-hand side (covariate measurement error) or left-hand side (outcome measurement error):

```
MeasErrorExt(substitute, model)
```

with the arguments being described as follows:

- `substitute` the error-prone measurement;
- `model` a fitted `lm` object of the calibration model in equation (3.10) (covariate measurement error) or the measurement error model in equation (3.5) (outcome measurement error). Or alternatively, a `list` with named arguments `coef` containing a vector of the coefficients of the calibration model or measurement error model and named argument `vcov` containing a matrix of the corresponding variance–covariance matrix. The argument `vcov` is not required.

The argument `model` is also used for conducting a sensitivity analysis by making informed guesses about the parameters of the calibration model (covariate measurement error) or measurement error model (outcome measurement error).

3.4.3. The MeasErrorRandom object

When random measurement error in a covariate is suspected but cannot be quantified, the MeasErrorRandom object can be used to conduct a sensitivity analysis:

```
MeasErrorRandom(substitute, variance)
```

with the arguments being described as follows:

- substitute the error-prone measurement;
- variance a numeric indicating the random measurement error variance in the substitute measurement, i.e., the parameter value of τ^2 in equation (3.2).

3.5. Examples

Six simulated datasets are included in the package mecor. These datasets mimic real datasets and represent the data structures described in section Validation study data structures for measurement error correction. There is an internal validation study with covariate measurement error (vat), an internal validation study with outcome measurement error (haemoglobin), a replicates study (bloodpressure) and a calibration study with outcome measurement error (sodium). The dataset vat_ext provides an external validation study for the vat dataset, and the dataset haemoglobin_ext provides an external validation study for the haemoglobin dataset. These datasets are described and analysed in the following sections.

3.5.1. Internal validation study

The dataset vat is a simulated dataset, representing the structure of the internal covariate-validation study shown in Table 3.1a. The dataset is inspired by the Netherlands Epidemiology of Obesity (NEO) study [42] and was used as the motivating example in a study investigating measurement error correction by Nab et al. [38]. The dataset represents a cross-sectional study of the association between visceral adipose tissue and insulin resistance. Visceral adipose tissue measures are expensive and therefore only available in 40% of the study population. Waist circumference measures however provide a simple proxy for visceral adipose tissue and are observed in the full study population. The dataset vat contains 650 observations of the natural logarithm of the outcome insulin resistance (ir_ln, fasting glucose (mmol/L) x fasting insulin (mU/L) / 22.5), the standardised error-prone exposure waist circumference (wc, cm), the covariates sex (sex, 0 = male, 1 = female), age (age, years), and standardised total body fat (tbf, %), and the standardised error-free measurement of the exposure visceral adipose tissue (vat, cm^2).

```
R> data("vat", package = "mecor")
R> head(vat)
```

```
        ir_ln         wc sex age        tbf       vat
1 -0.09341837 -1.3136816   1  48 -0.6571345        NA
2  0.16820894 -2.0336624   0  54 -1.5882163        NA
3  0.57299976 -0.2611214   0  46 -1.1033709        NA
4  0.63677178  0.8631987   0  55 -1.4785869 0.5083247
5  0.92908882 -1.2054861   1  61  0.9020136        NA
6 -0.72410039 -2.5032852   1  47 -0.9584166        NA
```

By ignoring the measurement error in wc, a linear model can be fitted to the data as follows:

```
R> lm(ir_ln ~ wc + sex + age + tbf, data = vat)

Call:
lm(formula = ir_ln ~ wc + sex + age + tbf, data = vat)

Coefficients:
(Intercept)           wc          sex          age          tbf
    0.50976      0.09697     -0.70953      0.01133      0.38783
```


The coefficients of this model will however be biased due to the measurement error in wc. The measurement error in wc can be corrected for using standard regression calibration (RC) as follows:

```
R> mecor(
+    ir_ln ~ MeasError(wc, reference = vat) + sex + age + tbf,
+    data = vat,
+    method = "standard")

Call:
mecor(formula = ir_ln ~ MeasError(wc, reference = vat) + sex +
 age + tbf, data = vat, method = "standard")

Coefficients Corrected Model:
(Intercept)          vat          sex          age          tbf
0.473398350  0.207598087 -0.438453038  0.009477677  0.270864391

Coefficients Uncorrected Model:
(Intercept)          wc         sex         age         tbf
 0.50976395  0.09697045 -0.70952736  0.01132712  0.38782671
```

Stratified percentile bootstrap confidence intervals of the coefficients of the corrected model can be obtained by using the argument B in the function mecor. To obtain standard errors and confidence intervals using the Fieller method or zero-variance method, the arguments zerovar and fieller of the summary object are set to TRUE:

```
R> set.seed(20210526)
R> mecor_fit <-
+    mecor(
+      ir_ln ~ MeasError(wc, reference = vat) + sex + age + tbf,
+      data = vat,
+      method = "standard",
+      B = 999)
R> summary(mecor_fit, zerovar = TRUE, fieller = TRUE)

Call:
mecor(formula = ir_ln ~ MeasError(wc, reference = vat) + sex +
```

```
    age + tbf, data = vat, method = "standard", B = 999)

95% Confidence Intervals:
             Estimate       LCI       UCI LCI (btstr)
(Intercept)  0.473398  0.185743  0.761054    0.212557
vat          0.207598  0.140549  0.274648    0.144636
sex         -0.438453 -0.594458 -0.282448   -0.577730
age          0.009478  0.004385  0.014570    0.005013
tbf          0.270864  0.199007  0.342721    0.200120
            UCI (btstr) LCI (zerovar) UCI (zerovar)
(Intercept)    0.721228      0.225140      0.721657
vat            0.281810      0.149712      0.265484
sex           -0.276988     -0.574231     -0.302675
age            0.014058      0.005096      0.013860
tbf            0.332815      0.208528      0.333201
            LCI (fieller) UCI (fieller)
(Intercept)            NA            NA
vat              0.145068      0.281464
sex                    NA            NA
age                    NA            NA
tbf                    NA            NA
Bootstrap Confidence Intervals are based on 999 bootstrap
 replicates using percentiles

The measurement error is corrected for by application of
 regression calibration

Coefficients Uncorrected Model:
             Estimate Std. Error  t value  Pr(>|t|)
(Intercept)  0.5097640  0.1264211   4.0323 6.185e-05
wc           0.0969705  0.0137957   7.0290 5.308e-12
sex         -0.7095274  0.0390086 -18.1890 < 2.2e-16
age          0.0113271  0.0022048   5.1374 3.695e-07
tbf          0.3878267  0.0201489  19.2481 < 2.2e-16

95% Confidence Intervals:
             Estimate       LCI       UCI
(Intercept)  0.509764  0.261517  0.758011
wc           0.096970  0.069881  0.124060
sex         -0.709527 -0.786127 -0.632928
age          0.011327  0.006998  0.015657
tbf          0.387827  0.348261  0.427392

Residual standard error: 0.3123469 on 645 degrees of
 freedom
```

In addition to standard RC, efficient RC (`method = "efficient"`) or validation RC (`method = "valregcal"`) can also be used to correct for the measurement error in the error-prone covariate `wc`.

The dataset `haemoglobin` is a simulated dataset, representing the structure of the internal outcome-validation study shown in Table 3.1b. The dataset is inspired by a trial investigating the efficacy of low-dose iron supplements [43] and was used as the motivating example for a study investigating measurement error correction in trial endpoints by Nab et al. [36]. The dataset represents a trial investigating the effect of low-dose iron supplements during pregnancy on haemoglobin levels at delivery. Haemoglobin levels were measured in venous blood in approximately 25% of the subjects (reference measure), and were measured in capillary blood in all subjects (substitute measure). The dataset `haemoglobin` contains 400 observations of the error-prone capillary haemoglobin levels (`capillary`, g/L), an indicator of whether the subject was randomised to receive the low-dose iron supplement (20 mg/d) (`supplement`, 0 = no, 1 = yes), and the error-free reference venous haemoglobin levels (`venous`, g/L).

```
R> data("haemoglobin", package = "mecor")
R> tail(haemoglobin)

    capillary supplement   venous
395  124.0489          1       NA
396  127.1005          0 127.9526
397  132.1858          1       NA
398  123.4427          0       NA
399  125.2438          1       NA
400  124.0738          0       NA
```

The measurement error in `capillary` can be accounted for by using standard method of moments (MM) as shown in the following:

```
R> mecor(
+    MeasError(capillary, reference = venous) ~
+    supplement,
+    data = haemoglobin,
+    method = "standard")

Call:
mecor(formula = MeasError(capillary, reference = venous) ~
 supplement, data = haemoglobin, method = "standard")

Coefficients Corrected Model:
(Intercept)  supplement
117.99341     6.97392

Coefficients Uncorrected Model:
(Intercept)  supplement
124.452261    7.764702
```

In addition to standard MM, efficient MM (`method = "efficient"`) can also be used to correct for the measurement error in the error-prone outcome `Y_star`.

When differential outcome measurement error in capillary haemoglobin measures is suspected, the argument `differential` of the `MeasError` object can be used to correct for differential measurement error as follows:

```
R> mecor(
+    MeasError(capillary,
+              reference = venous,
+              differential = supplement) ~ supplement,
+    data = haemoglobin,
+    method = "standard")

Call:
mecor(formula = MeasError(capillary, reference = venous,
 differential = supplement) ~ supplement,
 data = haemoglobin, method = "standard")

Coefficients Corrected Model:
(Intercept)  supplement
 118.386903    6.080729

Coefficients Uncorrected Model:
(Intercept)  supplement
124.452261    7.764702
```

Efficient MM (`method = "efficient"`) can also be used to correct for the differential measurement error in the error-prone outcome `Y_star`.

3.5.2. Replicates study

The dataset `bloodpressure` is a simulated dataset, representing the structure of the replicates study shown in Table 3.2a. The dataset represents a cross-sectional study of the association between blood pressure and creatinine in pregnant women [44]. Blood pressure measurements are prone to random measurement error. The dataset `bloodpressure` contains 450 observations of serum creatinine (`creatinine`, μmol/L), age (`age`, years), and systolic blood pressure (`sbp`, mm Hg). Systolic blood pressure is measured at 30, 60, 90 and 120 minutes.

```
R> data("bloodpressure", package = "mecor")
R> head(bloodpressure)

  creatinine age    sbp30    sbp60    sbp90   sbp120
1   53.75670  27 120.7987 113.2812 118.0705 124.2282
2   63.08498  36 121.7254 106.8143 118.9882 115.1341
3   60.04718  31 108.8798 119.6577 106.5588 117.5473
4   62.42976  43 116.5566 117.4964 126.3625 121.7148
5   61.31801  25 123.3018 116.4629 112.0310 109.8754
6   50.60952  35 124.9119 129.0927 129.0224 114.0828
```

In a study estimating the association between serum creatinine and systolic blood pressure, corrected for age, the random measurement error in the error-prone systolic blood pressure measurement at 30 minutes can be accounted for as follows:

```
R> mecor(
+    creatinine ~ MeasError(sbp30,
+                           replicate =
+                           cbind(sbp60,
+                                 sbp90,
+                                 sbp120)) + age,
+    data = bloodpressure,
+    method = "standard")

Call:
mecor(formula = creatinine ~ MeasError(sbp30, replicate
 = cbind(sbp60, sbp90, sbp120)) + age,
 data = bloodpressure, method = "standard")

Coefficients Corrected Model:
(Intercept)   cor_sbp30         age
 32.3796021   0.1877343   0.1743760

Coefficients Uncorrected Model:
(Intercept)       sbp30         age
 41.3050286   0.1165333   0.1650849
```

Maximum likelihood estimation (`method = "mle"`) can also be used to correct for the measurement error in the error-prone exposure `sbp30`. Note that, in this example dataset, the coefficients of the corrected model using standard RC will differ when `MeasError(sbp60, replicate = cbind(sbp30, sbp90, sbp120))` is used instead of `MeasError(sbp30, replicate = cbind(sbp60, sbp90, sbp120))`. In contrast, the corrected estimated coefficients obtained using maximum likelihood estimation will not change when the order of replicates is changed.

3.5.3. Calibration study

The dataset `sodium` is a simulated dataset, representing the structure of the outcome calibration study, shown in Table 3.3b. The dataset represents a randomised controlled trial designed to investigate whether a reduction in sodium intake results in satisfactory blood pressure control [45] and was used as the motivating example for a study investigating measurement error correction in dietary intake [31]. Sodium intake of the subjects was measured by a 24h recall and in urine. Sodium intake measured by a 24h recall is assumed prone to systematic measurement error and sodium intake measured in urine is assumed prone to random measurement error. The dataset `sodium` contains 1,000 observations of sodium intake measured by a 24h recall (`recall`, mg), an indicator of whether the subject was randomised to their usual diet or sodium-lowering diet (`diet`, 0 = usual, 1 = sodium-lowering), and two measures of urinary sodium (`urinary1`, `urinary2`, mg). The replicate urinary sodium are observed in approximately 50% of the subjects included in the trial.

```
R> data("sodium", package = "mecor")
R> tail(sodium)

       recall diet urinary1 urinary2
995  3.320633    1       NA       NA
996  3.496626    0       NA       NA
997  3.127590    1 3.818815 4.204880
998  4.363960    0       NA       NA
999  4.009316    1 4.719055 4.389111
1000 3.910490    0       NA       NA
```


The measurement error in the error-prone exposure `recall` can be accounted for as follows:

```
R> mecor(
+    MeasError(recall, replicate = cbind(urinary1,
+                                        urinary2)) ~ diet,
+    data = sodium,
+    method = "standard")

Call:
mecor(formula = MeasError(recall, replicate = cbind(
urinary1, urinary2)) ~ diet, data = sodium,
 method = "standard")

Coefficients Corrected Model:
(Intercept)        diet
  4.6075011  -0.4843495

Coefficients Uncorrected Model:
(Intercept)        diet
  3.8819732  -0.3051777
```

Efficient MM (`method = "efficient"`) can also be used to correct for the measurement error in the error-prone outcome `recall`.

3.5.4. External validation study

The dataset `vat_ext` is a simulated dataset, representing the structure of the external part of the external covariate-validation study shown in Table 3.4a. The dataset accompanies the dataset `vat` introduced in section Internal validation study. The dataset contains 100 observations of the error-free continuous exposure `vat`, the error-prone exposure `wc` and a covariates `sex`, `age` and `tbf`.

```
R> data("vat_ext", package = "mecor")
R> head(vat_ext)

           wc         vat sex age         tbf
1 -0.01357552 -1.69944962   1  50 -1.17103270
2  1.10201426  1.43889836   0  51 -0.99837467
```

```
3  1.23328072  1.24129099   0  54 -0.91030636
4 -0.07849380  0.05219091   0  55 -1.52766077
5 -0.47481715 -0.61165766   1  46  0.28706021
6 -1.33717429 -0.58193963   1  50  0.08718737
```

Suppose that in the dataset vat, the reference measure vat had not been observed. Using dataset vat_ext, we can correct for the measurement error in wc in dataset vat. The first step is to fit the calibration model in the external validation study as follows:


```
R> calmod_fit <- lm(vat ~ wc + sex + age + tbf,
                 data = vat_ext)
R> calmod_fit

Call:
lm(formula = vat ~ wc + sex + age + tbf, data = vat_ext)

Coefficients:
(Intercept)           wc          sex          age          tbf
   0.437466     0.571233    -0.984891     0.001111     0.488749
```

The second step is to use the calibration model calmod_fit in the MeasErrorExt object as follows:

```
R> data("vat", package = "mecor")
R> mecor(
+    ir_ln ~ MeasErrorExt(wc, calmod_fit) + sex + age + tbf,
+    data = vat,
+    method = "standard"
+  )

Call:
mecor(formula = ir_ln ~ MeasErrorExt(wc, calmod_fit) + sex +
 age + tbf, data = vat, method = "standard")

Coefficients Corrected Model:
(Intercept)      cor_wc         sex         age         tbf
 0.43550128  0.16975650 -0.54233566  0.01113844  0.30485834

Coefficients Uncorrected Model:
(Intercept)          wc         sex         age         tbf
 0.50976395  0.09697045 -0.70952736  0.01132712  0.38782671
```

Dataset haemoglobin_ext is a simulated dataset, representing the structure of the external part of the external outcome-validation study shown in Table 3.4b. The dataset accompanies the dataset haemoglobin introduced in section Internal validation study. The dataset contains 100 observations of the error-free outcome venous and the error-prone outcome capillary.

```
R> data("haemoglobin_ext", package = "mecor")
R> head(haemoglobin)

  capillary   venous
1  104.7269 115.3023
2  133.9946 119.7616
3  104.0304 108.0562
4  119.0214 121.1780
5  114.3891 111.7864
6  111.7754 112.8943
```


Suppose that in the dataset haemoglobin, the reference venous haemoglobin levels had not been observed. Using dataset haemoglobin_ext, we correct for the measurement error in capillary in haemoglobin, by fitting the measurement error model, as follows:

```
R> memod_fit <- lm(capillary ~ venous, data = haemoglobin_ext)
R> data("iovs", package = "mecor")
R> mecor(
+    MeasErrorExt(capillary, memod_fit) ~ supplement,
+    data = haemoglobin,
+    method = "standard")

Call:
mecor(formula = MeasErrorExt(capillary, memod_fit) ~
      supplement,
      data = haemoglobin,
      method = "standard")

Coefficients Corrected Model:
(Intercept)  supplement
 119.136649    7.227302

Coefficients Uncorrected Model:
(Intercept)  supplement
 124.452261    7.764702
```

3.5.5. Sensitivity analyses

Suppose that there is no error-free measure and no external validation study available for dataset vat. To investigate the sensitivity of study results to measurement error in variable vat, informed guesses of the coefficients of the calibration model are needed. Suppose one assumes that $E(\text{VAT}|\text{WC}, \text{sex}, \text{age}, \text{tbf}) = 0.4 + 0.6 \times \text{WC} - \text{sex} + 0 \times \text{age} + 0.5 \times \text{TBF}$. A sensitivity analysis could then be conducted as follows:

```
R> data("vat", package = "mecor")
R> mecor_fit_sens <-
+    mecor(
+      ir_ln ~ MeasErrorExt(wc, list(coef =
```

```
                                        c(0.4, 0.6, -1, 0, 0.5))) +
+      sex + age + tbf,
+      data = vat,
+      method = "standard")
R> mecor_fit_sens

Call:
mecor(formula = ir_ln ~ MeasErrorExt(wc, list(coef =
 c(0.4, 0.6, -1, 0, 0.5))) + sex + age + tbf,
 data = vat, method = "standard")

Coefficients Corrected Model:
(Intercept)      cor_wc         sex         age         tbf
 0.44511698  0.16161742 -0.54790994  0.01132712  0.30701800

Coefficients Uncorrected Model:
(Intercept)          wc         sex         age         tbf
 0.50976395  0.09697045 -0.70952736  0.01132712  0.38782671
```

The calibration model matrix used to correct for the measurement error in wc, is saved as matrix in the corfit object attached to mecor_fit_sens:

```
R> mecor_fit_sens$corfit$matrix

Lambda1 Lambda0 Lambda3 Lambda4 Lambda5
Lambda1     0.6     0.4      -1       0     0.5
Lambda0     0.0     1.0       0       0     0.0
Lambda3     0.0     0.0       1       0     0.0
Lambda4     0.0     0.0       0       1     0.0
Lambda5     0.0     0.0       0       0     1.0
```

In the dataset bloodpressure discussed in section Replicates study, random measurement error is suspected in systolic blood pressure. Suppose now that in the dataset bloodpressure, the three replicate measures sbp60, sbp90, sbp120 had not been observed. Suppose further that a measurement error variance of 30 mm Hg is assumed in the first systolic blood pressure measure sbp30. For measurement error correction, the MeasErrorRandom object could be used, here in combination with zerovariance estimation of standard errors (assuming that there is no uncertainty in the speculated value of the variance of the random measurement error sbp30):

```
R> mecor_fit_random <-
+    mecor(
+      creatinine ~ MeasErrorRandom(sbp30, variance = 30)
+      + age,
+      data = bloodpressure,
+      method = "standard")
R > summary(mecor_fit_random, zerovar = T)
```

```
Call:
mecor(formula = creatinine ~ MeasErrorRandom(sbp30,
 variance = 30) + age, data = bloodpressure,
 method = ”standard”)

Coefficients Corrected Model:
             Estimate SE (zerovar)
(Intercept) 33.568149     9.909771
cor_sbp30    0.182509     0.080298
age          0.159752     0.094837

95% Confidence Intervals:
             Estimate LCI (zerovar) UCI (zerovar)
(Intercept) 33.568149     14.145355     52.990943
cor_sbp30    0.182509      0.025127      0.339890
age          0.159752     -0.026125      0.345628

The measurement error is corrected for by application of
 regression calibration

Coefficients Uncorrected Model:
             Estimate Std. Error t value  Pr(>|t|)
(Intercept) 41.305029   6.758932  6.1112 2.155e-09
sbp30        0.116533   0.051271  2.2729   0.02351
age          0.165085   0.094705  1.7431   0.08200

95% Confidence Intervals:
             Estimate       LCI       UCI
(Intercept) 41.305029 28.021799 54.588258
sbp30        0.116533  0.015771  0.217296
age          0.165085 -0.021038  0.351208

Residual standard error: 9.897091 on 447 degrees of freedom
```

The calibration model matrix used to correct for the measurement error in `sbp30`, is again saved as `matrix` in the `corfit` object attached to `mecor_fit_random`:

```
R > mecor_fit_random$corfit$matrix

          Lambda1  Lambda0    Lambda3
Lambda1 0.6385083 42.39186 0.02922153
Lambda0 0.0000000  1.00000 0.00000000
Lambda3 0.0000000  0.00000 1.00000000
```

The sensitivity analyses could be expanded to ranges of possible coefficients of the calibration model or assumed variance of the random measurement error.

3.6. Conclusion

We demonstrated how measurement error correction methods can be applied using our R package mecor. These correction methods can be used in linear models with a continuous outcome when there is measurement error in the outcome or in a continuous covariate. The package accommodates measurement error correction methodology for a wide range of data structures: internal and external validation studies, replicates studies, and calibration studies. Various measurement error correction methods are implemented in the package: RC, MM and correction based on maximum likelihood estimation. For standard error estimation, the delta method and bootstrap are implemented for all methods. The package also facilitates sensitivity analysis or quantitative bias analysis when no data are available to estimate the parameters of the measurement error model, but the assumption of no measurement error is not warranted. A vast body of literature exists comparing the relative performance of the measurement error correction methods implemented in mecor [38, 46] and in comparison, with other methods e.g., simulation-extrapolation [47, 48], multiple imputation methods [49, 50] and Bayesian methods [11]. We focused on studies in which interest lies in estimating a covariate-outcome association. In other types of studies, e.g., prediction studies, considerations for measurement error correction are different and may not even require corrections [51, 52]. In future updates of the package, the measurement error correction methods may be extended to time-to-event [16] and binary outcomes, and multiple variables with measurement error [17, 27].

References

[1] J. Buonaccorsi, Measurement error: Models, methods, and applications, Chapman & Hall/CRC, Boca Raton, FL, 2010.

[2] R. Carroll, D. Ruppert, L. Stefanski, C. Crainiceanu, Measurement error in nonlinear models: A modern perspective, 2nd Edition, Chapman & Hall/CRC, Boca Raton, FL, 2006.

[3] W. Fuller, Measurement error models, John Wiley & Sons, New York, NY, 1987.

[4] P. Gustafson, Measurement error and misclassification in statistics and epidemiology: Impacts and Bayesian adjustments, Chapman & Hall/CRC, Boca Raton, FL, 2004.

[5] T. B. Brakenhoff, M. Mitroiu, R. H. Keogh, K. G. M. Moons, R. H. H. Groenwold, M. van Smeden, Measurement error is often neglected in medical literature: A systematic review, Journal of Clinical Epidemiology 98 (2018) 89–97. doi:10.1016/j.jclinepi.2018.02.023.

[6] P. A. Shaw, V. Deffner, R. H. Keogh, J. A. Tooze, K. W. Dodd, H. Küchenhoff, V. Kipnis, L. S. Freedman, Epidemiologic analyses with error-prone exposures: Review of current practice and recommendations, Annals of Epidemiology 28 (11) (2018) 821–828. doi:10.1016/j.annepidem.2018.09.001.

[7] R. H. Keogh, P. A. Shaw, P. Gustafson, R. J. Carroll, V. Deffner, K. W. Dodd, H. Küchenhoff, J. A. Tooze, M. P. Wallace, V. Kipnis, L. S. Freedman, STRATOS guidance document on measurement error and misclassification of variables in observational epidemiology: Part 1—Basic theory and simple methods of adjustment, Statistics in Medicine 39 (16) (2020) 2197–2231. doi:10.1002/sim.8532.

[8] X. Wang, B. Wang, Deconvolution estimation in measurement error models: The R package decon, Journal of Statistical Software 39 (10) (2011) 1–24. doi:10.18637/jss.v039.i10.

[9] J. R. Cook, L. A. Stefanski, Simulation-extrapolation estimation in parametric measurement error models, Journal of the American Statistical Association 89 (428) (1994) 1314–1328. doi:10.2307/2290994.

[10] S. R. Cole, H. Chu, S. Greenland, Multiple-imputation for measurement-error correction, International Journal of Epidemiology 35 (4) (2006) 1074–1081. doi:10.1093/ije/dyl097.

[11] J. W. Bartlett, R. H. Keogh, Bayesian correction for covariate measurement error: A frequentist evaluation and comparison with regression calibration, Statistical Methods in Medical Research 27 (6) (2018) 1695–1708. doi:10.1177/0962280216667764.

[12] J. W. Bartlett, B. L. De Stavola, C. Frost, Linear mixed models for replication data to efficiently allow for covariate measurement error, Statistics in Medicine 28 (25) (2009) 3158–3178. doi:10.1002/sim.3713.

[13] S. Rabe-Hesketh, A. Pickles, A. Skrondal, Correcting for covariate measurement error in logistic regression using nonparametric maximum likelihood estimation, Statistical Modelling 3 (3) (2003) 215–232. doi:10.1191/1471082X03st056oa.

[14] L. Gleser, Improvements of the naive approach to estimation in nonlinear errors-in-variables regression models, in: P. Brown, W. Fuller (Eds.), Statistical analysis of measurement error models, American Mathematics Society, Providence, 1990, pp. 99–114.

[15] R. J. Carroll, L. A. Stefanski, Approximate quasi-likelihood estimation in models with surrogate predictors, Journal of the American Statistical Association 85 (411) (1990) 652–663. doi:10.1080/01621459.1990.10474925.

[16] R. L. Prentice, Covariate measurement errors and parameter estimation in a failure time regression model, Biometrika 69 (2) (1982) 331–342. doi:10.2307/2335407.

[17] B. Rosner, D. Spiegelman, W. C. Willett, Correction of logistic regression relative risk estimates and confidence intervals for measurement error: The case of multiple covariates measured with error, American Journal of Epidemiology 132 (4) (1990) 734–745. doi:10.1093/oxfordjournals.aje.a115715.

[18] B. Armstrong, Measurement error in the generalised linear model, Communications in Statistics - Simulation and Computation 14 (3) (1985) 529–544. doi:10.1080/03610918508812457.

[19] R Core Team, R: A language and environment for statistical computing (2020). URL https://www.r-project.org/

[20] W. Lederer, H. Küchenhoff, The r journal: A short introduction to the simex and mcsimex, R News 6 (2006) 26–31, https://journal.r-project.org/articles/RN-2006-031/.

[21] W. He, J. Xiong, G. Yi, SIMEX R package for accelerated failure time models with covariate measurement error, Journal of Statistical Software 46 (1) (2012) 1–14. doi:10.18637/jss.v046.c01.

[22] StataCorp, Stata statistical software: Release 16 (2019). URL https://www.stata.com

[23] H. J. Newton, N. J. Cox, A special issue of the Stata Journal, The Stata Journal: Promoting communications on statistics and Stata 3 (4) (2003) 327–327. doi:10.1177/1536867X0400300401.

[24] J. W. Hardin, H. Schmiediche, R. J. Carroll, The regression-calibration method for fitting generalized linear models with additive measurement error, The Stata Journal 3 (4) (2003) 361–372. doi:10.1177/1536867X0400300406.

[25] J. W. Hardin, H. Schmiediche, R. J. Carroll, The simulation extrapolation method for fitting generalized linear models with additive measurement error, The Stata Journal 3 (4) (2003) 373–385. doi:10.1177/1536867X0400300407.

[26] S. Rabe-Hesketh, A. Skrondal, A. Pickles, Maximum likelihood estimation of generalized linear models with covariate measurement error, The Stata Journal: Promoting communications on statistics and Stata 3 (4) (2003) 386–411. doi:10.1177/1536867X0400300408.

[27] B. Rosner, D. Spiegelman, W. , Correction of logistic regression relative risk estimates and confidence intervals for random within-person measurement error, American Journal of Epidemiology 136 (11) (1992) 1400–1413. doi:10.1093/oxfordjournals.aje.a116453.

[28] X. Liao, D. M. Zucker, Y. Li, D. Spiegelman, Survival analysis with error-prone time-varying covariates: A risk set calibration approach, Biometrics 67 (1) (2011) 50–58. doi:10.1111/j.1541-0420.2010.01423.x.

[29] V. Kipnis, D. Midthune, D. W. Buckman, K. W. Dodd, P. M. Guenther, S. M. Krebs-Smith, A. F. Subar, J. A. Tooze, R. J. Carroll, L. S. Freedman, Modeling data with excess zeros and measurement error: Application to evaluating relationships between episodically consumed foods and health outcomes, Biometrics 65 (4) (2009) 1003–1010. doi:10.1111/j.1541-0420.2009.01223.x.

[30] D. Spiegelman, R. J. Carroll, V. Kipnis, Efficient regression calibration for logistic regression in main study/internal validation study designs with an imperfect reference instrument, Statistics in Medicine 20 (1) (2001) 139–160. doi:10.1002/1097-0258(20010115)20:1<139::AID-SIM644>3.0.CO;2-K.

[31] R. H. Keogh, R. J. Carroll, J. A. Tooze, S. I. Kirkpatrick, L. S. Freedman, Statistical issues related to dietary intake as the response variable in intervention trials, Statistics in Medicine 35 (25) (2016) 4493–4508. doi:10.1002/sim.7011.

[32] C. Spearman, The proof and measurement of association between two things, The American Journal of Psychology 15 (1) (1904) 72–101. doi:10.2307/1412159.

[33] C. Frost, S. G. Thompson, Correcting for regression dilution bias: comparison of methods for a single predictor variable, Journal of the Royal Statistical Society: Series A (Statistics in Society) 163 (2) (2000) 173–189. doi:10.1111/1467-985X.00164.

[34] J. A. Hutcheon, A. Chiolero, J. A. Hanley, Random measurement error and regression dilution bias, BMJ 340 (c2289) (2010). doi:10.1136/bmj.c2289.

[35] R. H. Keogh, I. R. White, A toolkit for measurement error correction, with a focus on nutritional epidemiology, Statistics in Medicine 33 (12) (2014) 2137–2155. doi:10.1002/sim.6095.

[36] L. Nab, R. H. H. Groenwold, P. M. J. Welsing, M. Smeden, Measurement error in continuous endpoints in randomised trials: Problems and solutions, Statistics in Medicine 38 (27) (2019) 5182–5196. doi:10.1002/sim.8359.

[37] R. H. Keogh, J. W. Bartlett, Measurement error as a missing data problem, in: G. Yi, A. Delaigle, P. Gustafson (Eds.), Handbook of measurement error models, 1st Edition, CRC Press, Boca Raton, FL, 2021, Ch. 20, pp. 429–452.

[38] L. Nab, M. van Smeden, R. de Mutsert, F. R. Rosendaal, R. H. H. Groenwold, Sampling strategies for internal validation samples for exposure measurement error correction: A study of visceral adipose tissue measures replaced by waist circumference measures, American Journal of Epidemiology 190 (9) (2021) 1935–1947. doi:10.1093/aje/kwab114.

[39] T. Lash, M. Fox, A. Fink, Applying quantitative bias analysis to epidemiologic data, Springer, New York, NY, 2009.

[40] L. Nab, R. H. H. Groenwold, M. van Smeden, R. H. Keogh, Quantitative bias analysis for a misclassified confounder: A comparison between marginal structural models and conditional models for point treatments, Epidemiology 31 (6) (2020) 796–805. doi:10.1097/EDE.0000000000001239.

[41] V. H. Franz, Ratios: A short guide to confidence limits and proper use (2007). arXiv:arXiv:0710.2024v1.

[42] R. de Mutsert, M. den Heijer, T. J. Rabelink, J. W. A. Smit, J. A. Romijn, J. W. Jukema, A. de Roos, C. M. Cobbaert, M. Kloppenburg, S. le Cessie, S. Middeldorp, F. R. Rosendaal, The Netherlands epidemiology of obesity (NEO) study: Study design and data collection, European Journal of Epidemiology 28 (6) (2013) 513–523. doi:10.1007/s10654-013-9801-3.

[43] M. Makrides, C. Crowther, R. Gibson, R. Gibson, C. Skeaff, Efficacy and tolerability of low-dose iron supplements during pregnancy: A randomized controlled trial, American Journal of Clinical Nutrition 78 (1) (2003) 145–153. doi:10.1093/ajcn/78.1.145.

[44] E. McCarthy, T. Carins, Y. Hannigan, N. Bardien, A. Shub, S. Walker, Data from: Effectiveness and safety of 1 vs 4 h blood pressure profile with clinical and laboratory assessment for the exclusion of gestational hypertension and pre-eclampsia: A retrospective study in a university affiliated maternity hospital, Dryad, Dataset (2015). doi:10.5061/dryad.0bq15.

[45] L. J. Appel, M. Espeland, P. K. Whelton, T. Dolecek, S. Kumanyika, W. B. Applegate, W. H. Ettinger, J. B. Kostis, A. C. Wilson, C. Lacy, S. T. Miller, Trial of nonpharmacologic intervention in the elderly (TONE), Annals of Epidemiology 5 (2) (1995) 119–129. doi:10.1016/1047-2797(94)00056-Y.

[46] S. W. Thurston, P. L. Williams, R. Hauser, H. Hu, M. Hernandez-Avila, D. Spiegelman, A comparison of regression calibration approaches for designs with internal validation data, Journal of Statistical Planning and Inference 131 (1) (2005) 175–190. doi:10.1016/j.jspi.2003.12.015.

[47] F. Perrier, L. Giorgis-Allemand, R. Slama, C. Philippat, Within-subject pooling of biological samples to reduce exposure misclassification in biomarker-based studies, Epidemiology 27 (3) (2016) 378–388. doi:10.1097/EDE.0000000000000460.

[48] E. Batistatou, R. McNamee, Performance of bias-correction methods for exposure measurement error using repeated measurements with and without missing data, Statistics in Medicine 31 (28) (2012) 3467–3480. doi:10.1002/sim.5422.

[49] L. S. Freedman, D. Midthune, R. J. Carroll, V. Kipnis, A comparison of regression calibration, moment reconstruction and imputation for adjusting for covariate measurement error in regression, Statistics in Medicine 27 (25) (2008) 5195–5216. doi:10.1002/sim.3361.

[50] K. Messer, L. Natarajan, Maximum likelihood, multiple imputation and regression calibration for measurement error adjustment, Statistics in Medicine 27 (30) (2008) 6332–6350. doi:10.1002/sim.3458.

[51] K. Luijken, L. Wynants, M. van Smeden, B. Van Calster, E. Steyerberg, R. Groenwold, Collaborators, Changing predictor measurement procedures affected the performance of prediction models in clinical examples, Journal of Clinical Epidemiology 119 (2020) (2019) 7–18. doi:10.1016/j.jclinepi.2019.11.001.

[52] K. Luijken, R. H. H. Groenwold, B. Van Calster, E. W. Steyerberg, M. Smeden, Impact of predictor measurement heterogeneity across settings on the performance of prediction models: A measurement error perspective, Statistics in Medicine 38 (18) (2019) 3444–3459. doi:10.1002/sim.8183.

4

Regression calibration for measurement error correction: The bias–variance trade off and finite sample performance

Correction of possible bias in exposure-outcome associations due to exposure measurement error using regression calibration may come at the cost of increased variance, referred to as the bias–variance trade off. Notably, in settings where measurement error is relatively large, the finite sample properties of regression calibration have not been investigated. We explore the bias–variance trade off for regression calibration and study the finite sample performance of regression calibration in settings where measurement error is relatively large using Monte Carlo simulation. The bias–variance trade off was of relevance in small samples (sample size <80) and was more pronounced in settings where measurement error was relatively large (reliability = 0.3) and residual error variance of the exposure-outcome association was relatively large (variance = 25). Particularly in settings where measurement error was relatively large (reliability <0.2) and sample size small (sample size <150), the performance of regression calibration was poor with percentage bias ranging from −99%–79% and mean squared error ranging from 6–25431. Application of regression calibration may not be useful in small sample size settings where measurement error is relatively large, because of the overall poor performance of the estimator in these settings.

This chapter has been submitted as: L. Nab and R.H.H. Groenwold, Regression calibration for measurement error correction: The bias–variance trade off and finite sample performance when measurement error is large

4.1. Introduction

Exposure measurement error is common in epidemiologic research but often neglected [1, 2]. When neglected, exposure measurement error can lead to bias in the exposure-outcome association [3], even when measurement error is random [4]. In the rare occasion of measurement error correction in epidemiologic research, regression calibration is among the methods used most often [1, 2]. Regression calibration relies on information about the measurement error model and its parameters, which can be estimated in extra data such as replicates data or internal validation data, or alternatively, informed by expert knowledge [5, 6].

When exposure measurement error is present, the estimator not correcting for this measurement error is typically biased. The application of regression calibration for measurement error correction is of particular interest when bias in the estimator not correcting for the exposure measurement error is relatively large. That is, when measurement error is relatively large, or equivalently, reliability of the error-prone measurement low. Regression calibration is a correction method designed to reduce this bias, at the price of an increased variance [7], a phenomenon referred to as the bias–variance trade off. We are unaware of reports of the finite sample performance of regression calibration in settings of highly unreliable measurements.

In this chapter we demonstrate settings in which the application of regression calibration can be useful, but importantly also when it may not. We report on settings where the estimator not correcting for exposure measurement error may be more efficient in terms of mean squared error than the regression calibration estimator (i.e., the bias–variance trade off). Additionally, we report on the performance of regression calibration in settings where the measurement error in the exposure is relatively large. Specific attention is paid to the performance of regression calibration in small samples. This is illustrated using an example of the association between active energy expenditure and lean body mass.

This chapter is organised as follows. In section 4.2, a study is introduced of the association between active energy expenditure and lean body mass. In section 4.3, the bias–variance trade off is illustrated for regression calibration. The finite sample performance of regression calibration when measurement error is relatively large is studied in section 4.4 by means of Monte Carlo simulation, focusing on settings where sample size is small. We conclude with a discussion of our results in section 4.5.

4.2. Example of lean body mass and energy expenditure

To motivate our study, we use an example of the association between energy expenditure and lean body mass. The association between active energy expenditure (mean active energy expenditure in kilo calories (kcal) per day) and percentage lean body mass (percentage of lean body mass of total body mass) was studied using publicly available data from the cross-sectional Karlsruhe Metabolomics and Nutrition study [8], detailed information on the study can be found here [9]. Body weight was measured in underwear and without shoes using a standardized scale. Lean body mass was measured in a standardized way by dual-energy X-ray absorptiometry and expressed as percentage of total body weight. The energy expenditure (in kcal per day) was measured by Actiheart® (CamNtech, Cambridge, United Kingdom). In addition, energy expenditure was measured

using the international physical activity questionnaire (IPAQ). This questionnaire provides a substitute measure of energy expenditure, based on physical activity and expressed in metabolic equivalent of task (MET)-minutes. This measure was then transformed to approximate subject's energy expenditure in kilocalories per day [10].

Throughout this example, we consider energy expenditure measured by Actiheart® the reference measure and energy expenditure measured by the IPAQ the (error-prone) substitute measure. Figure 4.1 shows the agreement between energy expenditure in kcal per day measured by Actiheart® and the IPAQ in the Karlsruhe Metabolomics and Nutrition study. The correlation between the two measures of energy expenditure was 0.10 (95% confidence interval (CI): -0.02;0.21).

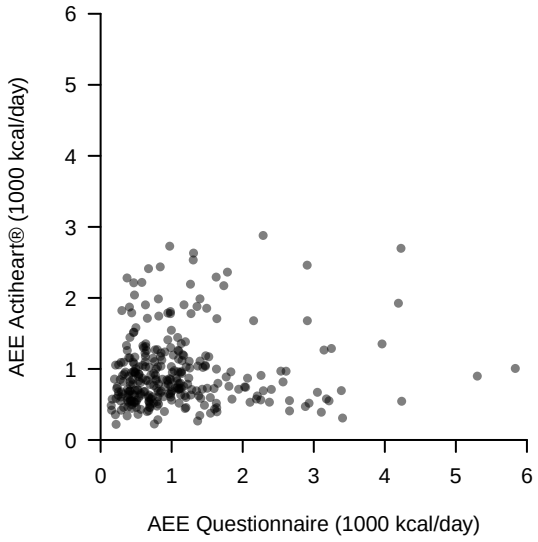


Figure 4.1: Agreement of mean active energy expenditure (AEE) measured by Actiheart® (reference measure) and AEE measured by the international physical activity questionnaire (substitute measure)

Table 4.1 gives an overview of four different estimates of the association between energy expenditure and lean body mass. Using the reference measure of energy expenditure measured by Actiheart®, we found that an increase in energy expenditure of 1,000 kcal per day was associated with a 3.2 increase in lean body mass (95%CI: 1.8;4.7). Using the substitute measure of energy expenditure measured by IPAQ instead, it was found that an increase in energy expenditure of 1,000 kcal per day was associated with a 0.7 *decrease* in lean body mass (95%CI: −1.5;0.1). This estimate was considered biased due to the measurement error in the questionnaire-based energy expenditure level. When this estimate was corrected by means of regression calibration (RC), informed by the relation between the substitute measure and reference measure, we found an increase in energy expenditure of 1,000 kcal per day was associated with an 18.7 *decrease* in lean body mass. Notably, there is a large discrepancy between the point estimate obtained by regressing the reference measure of energy expenditure (measured by Actiheart®) and lean body mass (i.e., 3.2) and the measurement error corrected point estimate (i.e., −18.7). Two different methods

for CI construction were available for the RC corrected estimate: the Delta method and the bootstrap percentile method, yielding 95% CIs of −56.6;19.2 and −270.0;217.2, respectively. The above estimates were all adjusted for sex. Covariate adjustment was restricted to sex for illustration purposes, the covariate adjustment set should potentially be expanded.

Table 4.1: Estimates of the association between an increase of 1,000 kcal/day in mean active energy expenditure and percentage lean body mass (adjusted for sex) and associated 95% confidence intervals (CIs) in the Karlsruhe Metabolomics and Nutrition study [8]

Method	**Point Estimate**	**95% CI**
Actiheart®	3.2	1.8;4.7
International Physical Activity Questionnaire	−0.7	−1.5;0.1
Regression Calibration and Delta for CI Construction	−18.7	−56.6;19.2
Regression Calibration and Bootstrap for CI Construction[a]	−18.7	−270.0;217.2

[a] Based on 999 replicates using percentiles

4.3. Bias–variance trade off for regression calibration

The estimator of the exposure-outcome association that does not account for measurement error in the exposure variable is typically biased. Nevertheless, a correction for this bias by means of RC may come at the price of an increased variance, sometimes referred to as the bias–variance trade off [7]. That is, the RC estimator is typically unbiased, yet it is more variable than the uncorrected estimator. Consequently, there may be circumstances where the uncorrected estimator is more efficient in terms of mean squared error (MSE) than the corrected estimator.

We illustrate this phenomenon here by graphical presentation of the MSE of the uncorrected estimator and the RC estimator in simple settings, by using the theoretical derivation of the MSE of the two estimators, described by Carroll et al. [7]. Since the theoretical derivation by Carroll et al. relies on the assumption that the correction factor used in RC is known, which is rare, we expand these results by means of Monte Carlo simulation to simple settings where the correction factor is not known and is estimated from the data.

The data generating mechanism used to generate sets of artificial data is described in Table 4.2. Parameters of the data generating mechanism were inspired by the motivating example of energy expenditure and body mass. For simplicity, we assume random measurement error in the error-prone AEE^* (i.e., AEE^* is distributed around AEE with independent error). In our artificial data, the reliability of AEE^* is equal to 0.25 / (0.25 + τ^2). This ratio is referred to as the 'reliability' in this chapter and in case of random measurement error as assumed here, the reliability is equal to the 'correction factor' mentioned above. There is an inverse relation between the measurement error variance (i.e., τ^2) and the reliability of the error-prone measure. When the measurement error variance is large, the reliability is low and vice versa.

Table 4.2: Data generating mechanism

Variable	Variable Name	Distribution
Active Energy Expenditure	AEE	$N(1, 0.25)$
Error-Prone Active Energy Expenditure	AEE^*	$N(AEE, \tau^2)$
Percentage Lean Body Mass	LBM	$N(80 + 3 \times AEE, \sigma^2)$

We refer to the estimator of the linear regression of *LBM* on the error-prone measurement of *AEE* (i.e., *AEE**) using ordinary least squares (OLS) as the OLS estimator. We refer to the corrected estimator by means of regression calibration (RC) as the 'RC estimator'. In this chapter, the RC estimator available in the package `mecor` [6] is used. This package adopts the RC estimator described by Rosner et al. in [11], which is for linear regression equivalent to the method of moments estimator [3]. The RC estimator divides the OLS estimator by a 'correction factor' which can be estimated in extra data.

4.3.1. Correction factor known

From the results from Carroll et al. [7] and the data generating mechanism in Table 4.2, the bias in the OLS estimator is equal to 1 minus the correction factor times the effect of *AEE* on *LBM* (i.e., 3 in Table 4.2). The variance of the OLS estimator is equal to the variance of the residual errors (i.e., σ^2 in Table 4.2) divided by the number of observations (i.e., n in Table 4.2) times the variance of AEE^* (i.e., $0.25 + \tau^2$ in the above). The MSE of the OLS estimator is equal to its bias squared plus its variance. The RC estimator is assumed unbiased, and its variance is equal to the variance of the OLS estimator divided by the correction factor squared. Figure 4.2 shows the MSE of the OLS estimator and the RC estimator for different scenarios of variance of the residual errors, sample size and reliability. It illustrates that when the variance of the residual errors is relatively large (25) and sample size is small (≤60), the OLS estimator may be more efficient than the RC estimator in terms of MSE. This gain in efficiency becomes smaller and ultimately turns around in favour of the RC estimator as reliability increases, sample size increases, or the variance of the residual errors decreases. Note that we fixed the variance of *AEE* (i.e., 0.25) and the effect of *AEE* on *LBM* (i.e., 3) throughout this illustration. Varying these will impact the graphical illustrations of the bias–variance trade off in Figure 4.2, but the phenomenon would still exist.

4.3.2. Correction factor not known

We compared the MSE of the OLS estimator and the RC estimator by means of Monte Carlo simulation investigating scenarios of finite samples where the correction factor is *not* known. The correction factor is estimated in an extra data set providing information about the reference measure *AEE* and the substitute measure *AEE**. We generated data using the data generating mechanism described in Table 4.2 and studied MSE for σ^2 equal to 5 or 25, reliability equal to 0.3, 0.6 or 0.9, and the number of observations 20, 40, 60, 80 or 100 in a full-factorial design ($2 \times 3 \times 5 = 30$ scenarios). We set the sample size of the set that is used to estimate the correction factor equal to the sample size of the study. For each scenario, 5000

datasets were generated. In each generated data set, the uncorrected effect was estimated by regressing the outcome percentage lean body mass on the error-prone active energy expenditure using standard software. Subsequently, the corrected effect was estimated by means of RC using the R package `mecor` [6]. The performance of these two estimators was evaluated in terms of MSE. Accompanying Monte Carlo standard errors (MCSE) were calculated [12], using the R package `rsimsum` [13]. All code used for the simulation study is publicly available via `https://github.com/LindaNab/woorc`. Figure 4.3 shows the MSE of the OLS estimator and the RC estimator. Overall, the same patterns were obtained as those described in section Correction factor known. An important difference is, however, that the MSE of the RC estimator was much larger than its theoretical derivation when sample size is small, which was most pronounced in the settings where reliability was 0.3 and the residual errors of the outcome model relatively high (i.e., σ^2 equal to 25) (Figure 4.3).

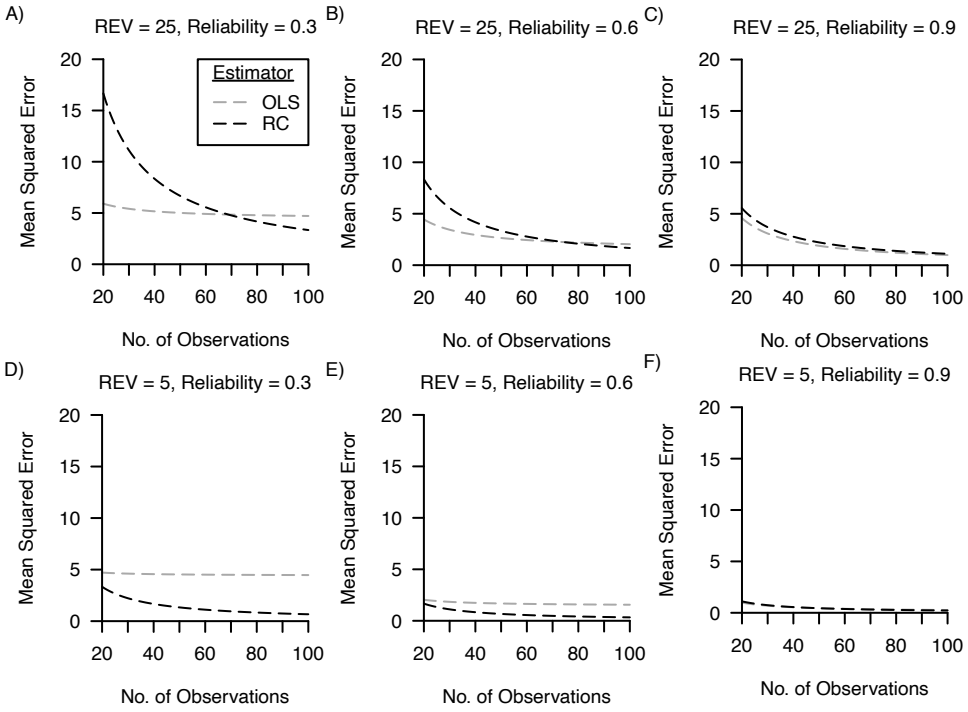


Figure 4.2: Theoretical mean squared error of the estimator not correcting for measurement error (OLS) (gray dashed line) and the regression calibration (RC) estimator (black dashed line), as derived by Carroll et al. [7], for varying sizes of the sample size (20-100, x axis) and for varying sizes of the residual error variance (REV) (25: panels A-C; 5: panels D-F) and for varying size of the reliability (0.3: panels A and D; 0.6: panels B and E; 0.9: panels C and F). In panel F, the lines overlap.

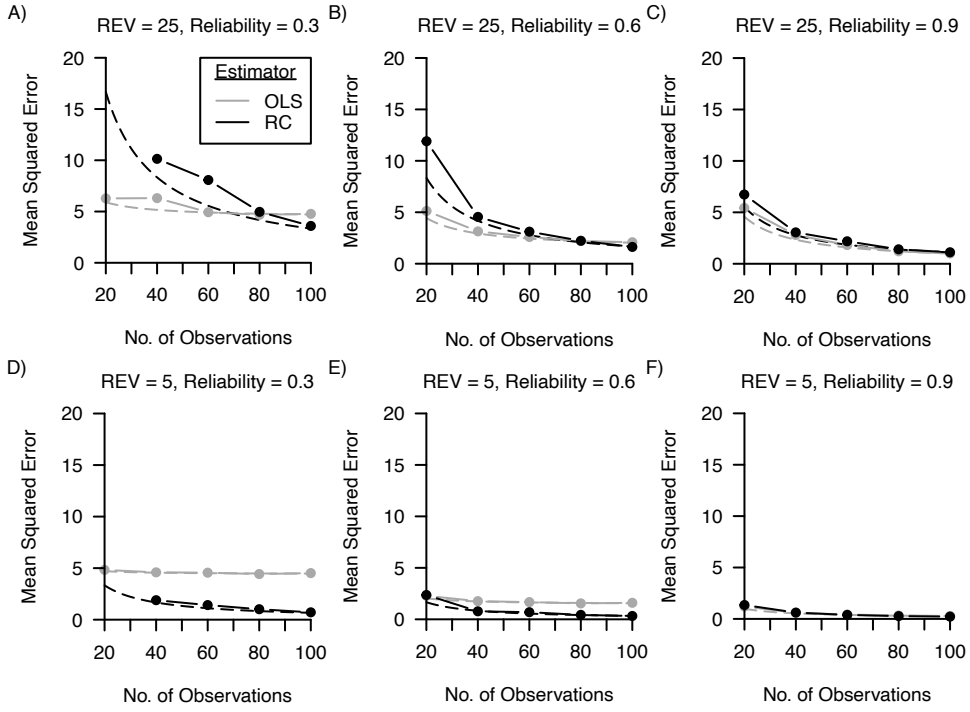


Figure 4.3: Results of a Monte Carlo simulation study of the mean squared error of the estimator not correcting for measurement error (OLS) (gray solid line with dots indicating the estimates) and the regression calibration (RC) estimator (black solid line with dots indicating the estimates) for varying sizes of the sample size (20-100, x-axis) and for varying sizes of the residual error variance (REV) (25: panels A-C; 5: panels D-F) and for varying size of the reliability (0.3: panels A and D; 0.6: panels B and E; 0.9: panels C and F). The dashed gray and black lines represent the theoretical mean squared error of the estimator not correcting for measurement error and the regression calibration estimator, respectively, as derived by Carroll et al. [7]. In panel A and D, the mean squared errors of the regression calibration estimator in the Monte Carlo simulation study fell outside the range of the graph when number of observations was 20, and were 211 (Monte Carlo standard error (MCSE) 142) and 133 (MCSE 89), respectively. In panel F, all lines overlap.

4.4. Finite sample properties of regression calibration

In case of exposure measurement error in a linear regression, RC provides consistent estimates if the correction factor is estimated consistently [3]. A consistent estimate of the correction factor can be obtained in extra data such as internal validation data or replicates data. However, earlier studies (e.g., [14]) suggested that the RC estimator is not necessarily unbiased, specifically in settings where the reliability of the error-prone measurement is low (i.e., 0.2). In addition, in our investigation of the efficiency of the RC estimator described in the previous section, we found that when the reliability was equal to 0.3 and sample size was 20, the MSE of the RC estimator was extremely large compared to the MSE of the OLS estimator (i.e., 211 vs 6 and 133 vs 5, for residual error variance equal to 25 and 5, respectively). Here, we aim to extend these results and investigate the finite sample performance of RC in settings where the measurement error is relatively large (i.e., reliability low), thereby focusing on small samples.

4.4.1. Data generating mechanism

Again, we used the generating mechanism described in Table 4.2. The number of observations (25, 50, 150, 300 and 600) and the reliability of the error-prone exposure *AEE** (0.99, 0.95, 0.9, 0.8, 0.7, 0.6, 0.5, 0.4, 0.3, 0.2, 0.1, 0.05 and 0.01) were varied in a full factorial design (5 × 13 = 65 scenarios). Although our specific interest was the performance of RC for an error-prone measure with low reliability, we studied the full range of the reliability for illustration purposes. We set the sample size of the set that is used to estimate the correction factor equal to the sample size of the study. For each scenario, 5,000 datasets were generated.

4.4.2. Assessment of performance

In each generated data set, the uncorrected effect was estimated by regressing the outcome variable *LBM* on the error-prone *AEE** using standard software. Subsequently, the corrected effect was estimated by means of RC using the R package `mecor` [6]. Ninety-five percent CIs of the uncorrected analysis were constructed using standard software, and for the RC analysis these were constructed using the Delta method and bootstrap resampling using 999 replicates and taking the 2.5% and 97.5% percentiles, both available in the R package `mecor`. Performance of the two different analyses was evaluated in terms of bias, MSE, confidence interval coverage (the proportion of 95% CIs that contained the true value of the true effect), empirical SE, and model based SE. Model based SE was estimated using the standard errors for the uncorrected analysis from standard software, and using the standard errors estimated by the Delta method or the standard deviation of the 999 replicates of the bootstrap samples for the RC analysis. Monte Carlo standard errors (MCSE) were calculated for all performance measures [12], using the R package `rsimsum` [13]. All code used for the simulation study is publicly available via `https://github.com/LindaNab/woorc`.

4.4.3. Results

Figures 4.4 and 4.5 show percentage bias, MSE and confidence interval coverage for varying levels of the reliability of the error-prone measure and number of observations. The OLS estimator was biased, with decreasing bias for increasing levels of reliability. Bias in the

OLS estimator was independent of sample size. Generally, the RC estimator was unbiased, except when reliability was 0.01 for all levels of the sample size. Specifically, for a sample size of 150 and reliability 0.01, the percentage bias was 659% (Monte Carlo SE (MCSE) of bias 23). In addition, the RC estimator was severely biased for a sample size of 50 and reliability equal to 0.01 and 0.05 (percentage bias was 78.7% (MCSE of bias 0.624) and -73.9% (MCSE of bias 2.255), respectively) and for a sample size of 25 and reliability 0.01, 0.05 and 0.1 (percentage bias was -9.4% (MCSE of bias 1.91), -83.7% (MCSE 0.514) and -34.2% (MCSE 0.273), respectively). MSE of the OLS estimator and the RC estimator decreased when reliability increased (Figures 4.4 and 4.5). Generally, the RC estimator was more efficient in terms of MSE than the OLS estimator, except for reliability equal to 0.01 for all sample sizes; reliability ≤ 0.2 for a sample size of 50; or reliability ≤ 0.3 for a sample size of 25 (Table 4.3). In addition, the RC estimator and OLS estimator show similar efficiency for high reliability (i.e., reliability ≥0.9).

Confidence interval coverage was around the nominal level of 95% for the CIs constructed using bootstrap resampling, independent of sample size or reliability. CI coverage was slightly above the nominal level of 95% for the CI constructed using the Delta method for reliability ≤ 0.8 (i.e., ranging between 96%–100%, MCSE <0.05) and at the nominal level for reliability greater or equal to 0.9, independent of sample size. Generally, the coverage of the CIs of the OLS estimator was lower than the nominal level of 95% and moved closer to the nominal level for increasing values of the reliability (ranging between 0%–97%, MCSE < 0.05).

Model based standard errors were equal to empirical standard error of the analysis ignoring measurement error for all studied settings (Figures 4.6 and 4.7 and Table 4.4). Generally, model based standard errors obtained by bootstrap resampling better approximated empirical standard errors of the RC analysis (Figures 4.6 and 4.7). Model based standard error of the RC analysis were equal to empirical standard error for reliability ranging between 0.1–1 for a sample size of 600; reliability ranging between 0.2–1 for a sample size of 300 or 150; and reliability ranging between 0.9–1 for a sample size of 50 or 25 (Figures 4.6 and 4.7 and Table 4.5). For all other studied simulation settings, model based standard errors differed from the empirical standard errors of the RC analysis (Table 4.5).

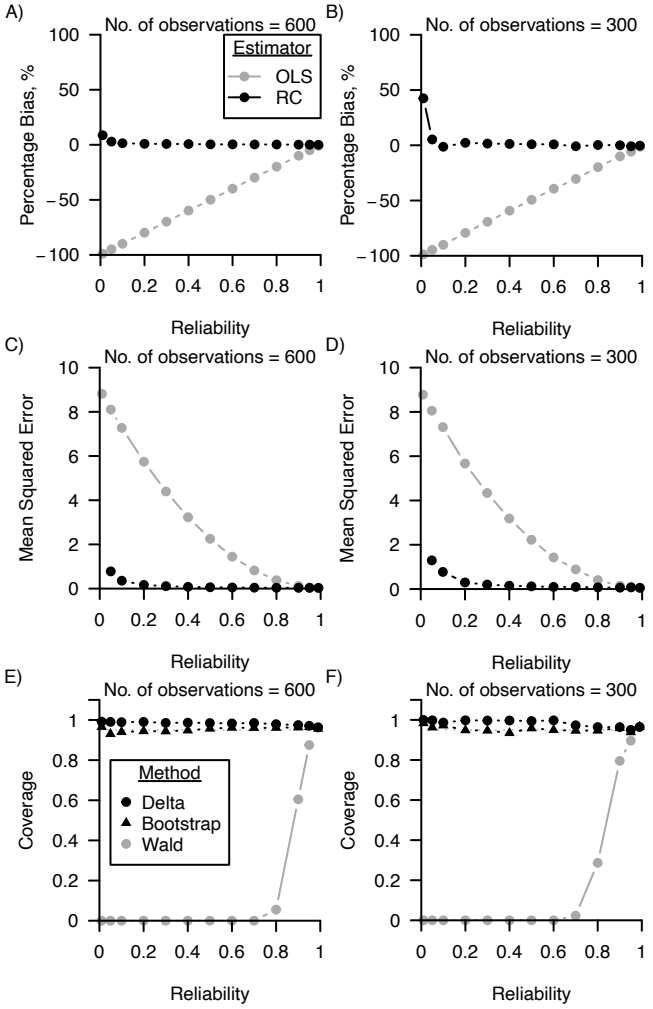


Figure 4.4: Performance of the analysis ignoring measurement error (OLS) and regression calibration (RC) in a setting with 600 (first column) and 300 (second column) observations, in terms of percentage bias (panels A and B); mean squared error (panels C and D) and coverage (panels E and F) for varying values of reliability of the error-prone exposure (x-axis). In panel C and D, the mean squared errors of the regression calibration estimator fell outside the range of the graph when reliability was 0.01, and were 12 (Monte Carlo standard error (MCSE) 2) and 80 (MCSE 29), respectively.

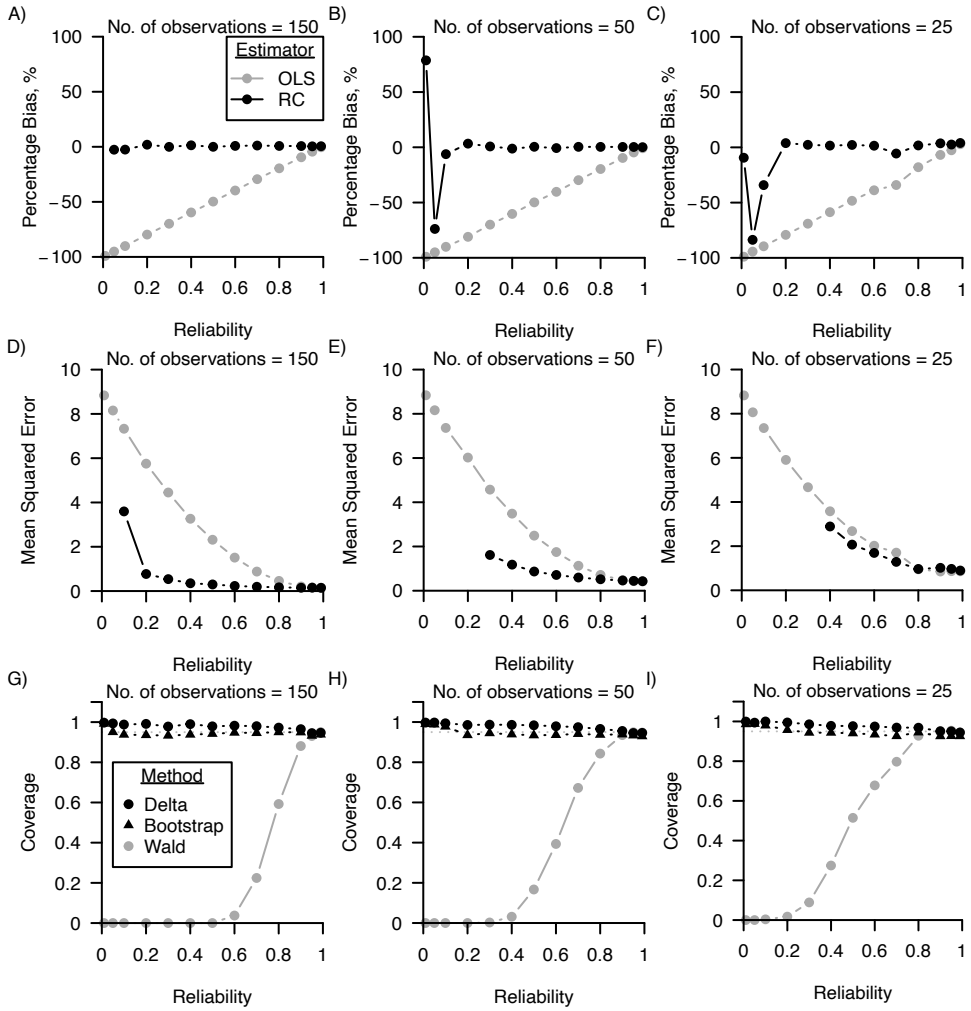


Figure 4.5: Performance of the analysis ignoring measurement error (OLS) and regression calibration (RC) in a setting with 150 (first column), 150 (second column) and 25 (third column) observations, in terms of percentage bias (panels A-C); mean squared error (panels D-F) and coverage (panels G-I) for varying values of reliability of the error-prone exposure (x-axis). In panel A, the percentage bias in the regression calibration estimator fell outside the range of the graph when reliablity was 0.01, and was 659% (Monte Carlo standard error (MCSE) of bias 23). The values that fell outside the range of panels D-F, can be found in Table 4.3

Table 4.3: Mean Squared error (MSE) of the regression calibration estimator in the settings which fell outside the plot range of the graphs in Panel D-F in Figure 4.5

n	Reliability	MSE	MCSE	Panel
150	0.01	2 536 515	2 307 534	D
	0.05	20	3	
50	0.01	1950	446	E
	0.05	25 431	9740	
	0.10	1025	325	
	0.20	20	5	
25	0.01	7092	1062	F
	0.05	1328	231	
	0.10	374	59	
	0.20	630	540	
	0.30	57	37	

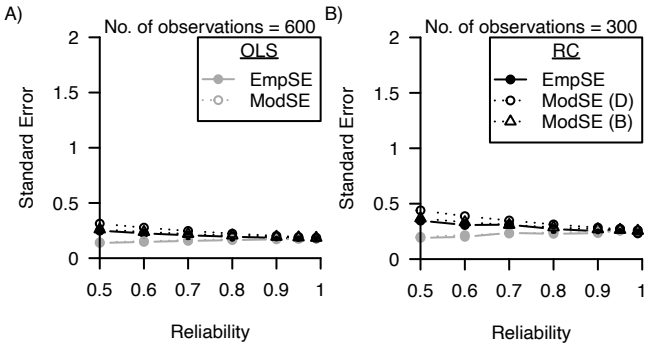


Figure 4.6: Empirical standard error (EmpSE) of the analysis ignoring measurement error (OLS) (solid gray lines with dots indicating the estimates) and regression calibration (RC) (solid black lines with dots indicating the estimates); and model based standard error (ModSE) of the analysis ignoring measurement error (OLS) (dotted gray lines with open dots indicating the estimates) and regression calibration (RC) using the Delta method (D) (dotted black lines with open dots indicating the estimates) or bootstrap resampling (B) (dotted black lines with open triangles indicating the estimates) in a setting with 600 (first column) and 300 (second column) observations for varying values of the reliability of the error-prone exposure (x-axis). The lines of the OLS estimator for the empirical standard error and model based standard error overlap. The lines of the RC estimator for the empirical standard error and model based standard error using bootstrap resampling overlap.

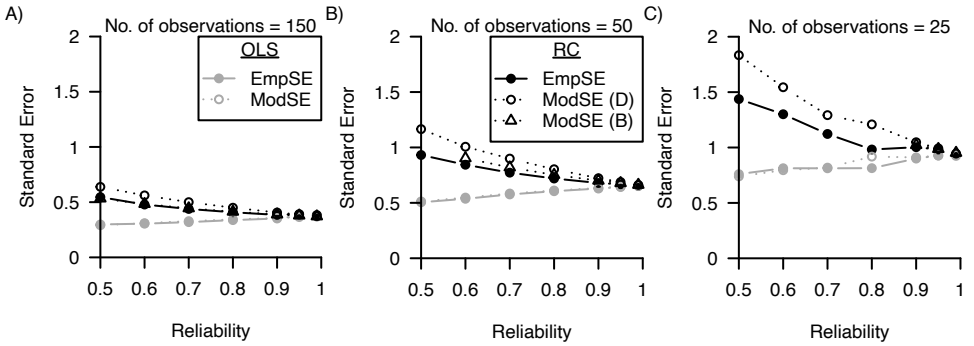


Figure 4.7: Empirical standard error (EmpSE) of the analysis ignoring measurement error (OLS) (solid gray lines with dots indicating the estimates) and regression calibration (RC) (solid black lines with dots indicating the estimates); and model based standard error (ModSE) of the analysis ignoring measurement error (OLS) (dotted gray lines with open dots indicating the estimates) and regression calibration (RC) using the Delta method (D) (dotted black lines with open dots indicating the estimates) or bootstrap resampling (B) (dotted black lines with open triangles indicating the estimates) in a setting with 150 (first column), 50 (second column) and 25 (third column) observations for varying values of the reliability of the error-prone exposure (x-axis). The lines of the OLS estimator for the empirical standard error and model based standard error overlap. The lines of the RC estimator for the empirical standard error and model based standard error using bootstrap resampling overlap in panel A. In panel B, the model based standard error of the RC estimator using bootstrap resampling fell outside the range of the graph for reliability equal to 0.5 and was 2.5 (Monte Carlo SE (MCSE) 0.411). In panel E, the model based standard error of the RC estimator using bootstrap resampling fell outside the range of the graph for reliability equal to 0.5, 0.6, 0.7, and 0.8, and was 211.9 (MCSE 25.489), 27.8 (MCSE 9.575), 6.4 (MCSE 0.551), 2.4 (MCSE 0.241), respectively.

Table 4.4: Empirical standard error (EmpSE) and model based standard error (ModSE) of the analysis ignoring measurement error for varying values of the sample size and reliability of the error-prone exposure measure

n	Reliability	EmpSE	MCSE	ModSE	MCSE
600	0.01	0.02	<0.001	0.02	<0.001
	0.05	0.05	<0.001	0.05	<0.001
	0.10	0.07	0.001	0.07	<0.001
	0.20	0.09	0.001	0.10	<0.001
	0.30	0.11	0.001	0.11	<0.001
	0.40	0.13	0.001	0.13	<0.001
	0.50	0.14	0.001	0.14	<0.001
300	0.01	0.03	<0.001	0.03	<0.001
	0.05	0.07	0.001	0.07	<0.001
	0.10	0.09	0.001	0.10	<0.001
	0.20	0.13	0.001	0.13	<0.001
	0.30	0.15	0.002	0.16	<0.001
	0.40	0.17	0.002	0.18	<0.001
	0.50	0.19	0.002	0.20	<0.001
150	0.01	0.04	<0.001	0.04	<0.001
	0.05	0.10	0.001	0.10	<0.001
	0.10	0.14	0.001	0.14	<0.001
	0.20	0.19	0.002	0.19	<0.001
	0.30	0.24	0.002	0.23	<0.001
	0.40	0.26	0.003	0.26	<0.001
	0.50	0.30	0.003	0.29	<0.001
50	0.01	0.08	0.001	0.08	<0.001
	0.05	0.17	0.002	0.17	<0.001
	0.10	0.24	0.002	0.25	0.001
	0.20	0.33	0.003	0.34	0.001
	0.30	0.40	0.004	0.41	0.001
	0.40	0.46	0.005	0.46	0.001
	0.50	0.51	0.005	0.51	0.001
25	0.01	0.12	0.001	0.12	<0.001
	0.05	0.22	0.002	0.25	0.001
	0.10	0.37	0.004	0.36	0.001
	0.20	0.51	0.005	0.50	0.002
	0.30	0.61	0.006	0.60	0.002
	0.40	0.70	0.007	0.68	0.002
	0.50	0.76	0.008	0.74	0.002

Table 4.5: Empirical standard error and model based standard error using the Delta method or bootstrap (btstrp) resampling of regression calibration and associated Monte Carlo standard errors (MCSE) for varying values of the sample size and reliability

n	Reliability	EmpSE	MCSE	ModSE Delta	MCSE	ModSE Btstrp	MCSE
600	0.01	3.44	0.034	28.23	10.792	719.18	73.283
	0.05	0.88	0.009	1.20	0.004	1.42	0.076
	0.10	0.60	0.006	0.81	0.002	0.62	0.001
	0.20	0.41	0.004	0.54	0.001	0.42	0.001
	0.30	0.33	0.003	0.43	< 0.001	0.34	< 0.001
	0.40	0.28	0.003	0.36	< 0.001	0.29	< 0.001
	0.50	0.25	0.002	0.31	< 0.001	0.26	< 0.001
300	0.01	8.84	0.088	17 882.74	8938.773	1006.92	173.947
	0.05	1.13	0.011	1.78	0.012	72.07	28.696
	0.10	0.88	0.009	1.17	0.004	1.08	0.131
	0.20	0.54	0.005	0.77	0.002	0.60	0.001
	0.30	0.45	0.004	0.61	0.001	0.48	0.001
	0.40	0.38	0.004	0.51	0.001	0.41	0.001
	0.50	0.35	0.003	0.44	0.001	0.37	0.001
150	0.01	1592.68	15.928	140 226.41	65 175.844	3727.84	1670.360
	0.05	4.49	0.045	102.07	17.746	1132.07	123.372
	0.10	1.89	0.019	3.35	0.438	97.39	14.841
	0.20	0.88	0.009	1.15	0.004	0.99	0.008
	0.30	0.73	0.007	0.89	0.002	0.72	0.003
	0.40	0.59	0.006	0.74	0.002	0.61	0.002
	0.50	0.55	0.005	0.64	0.001	0.53	0.001
50	0.01	44.09	0.441	12 448.39	2456.747	4439.06	756.806
	0.05	159.47	1.595	306 333.55	65 894.251	4958.66	631.884
	0.10	32.02	0.320	5841.30	1090.653	864.77	225.879
	0.20	4.42	0.044	140.83	22.227	178.76	13.801
	0.30	1.27	0.013	1.73	0.011	47.74	5.814
	0.40	1.09	0.011	1.37	0.007	13.04	1.355
	0.50	0.93	0.009	1.17	0.004	2.47	0.411
25	0.01	84.22	0.842	155 586.92	46 169.38	6699.71	1731.772
	0.05	36.35	0.364	2866.83	263.583	1173.37	48.557
	0.10	19.32	0.193	1048.99	125.869	2287.16	234.035
	0.20	25.09	0.251	4762.48	2369.783	527.13	100.104
	0.30	7.58	0.076	384.43	153.234	628.15	257.196
	0.40	1.70	0.017	2.44	0.112	207.84	78.895
	0.50	1.44	0.014	1.83	0.014	211.88	25.489

4.5. Discussion

This chapter studied settings in which application of regression calibration (RC) may not be appropriate for correcting bias induced by exposure measurement error. Particularly in small samples, the RC estimator may be less efficient in terms of MSE than an estimator not correcting for the exposure measurement error. This bias–variance trade off was most pronounced in settings where reliability was low and residual error variance high. In an investigation of the finite sample properties of RC, we showed that particularly when the measurement error is relatively large and sample size small, RC provided biased estimates, large MSEs and large empirical standard errors. Particularly, in these settings, the model based standard errors did not agree with the empirical standard errors and the RC estimator was instable as shown by large Monte Carlo standard errors.

In settings where the reliability of the error-prone measure was low (i.e., reliability <0.2) and sample size small (i.e., sample size <150), the performance of RC was poor. This is explained by the fact that by application of RC, the uncorrected estimate was divided by an estimate of the correction factor. This correction factor was equal to the reliability of the error-prone measurement in our study. In settings in which the correction factor was close to zero, it was more likely that in one of the replications in the simulation study the correction factor approached zero. Specifically when sample size was small. Consequently, the corrected estimate in that specific replication was large, affecting mean percentage bias, MSE, and the empirical standard error of the setting under study, since outliers affect these summary estimates. Bootstrapped confidence intervals were sensitive to this property as well. That is, independent of the original artificial data, one of the 999 replicates could provide a correction factor approaching zero, affecting the distribution of the estimates in the different bootstrap samples, and thus standard errors based on the standard deviation of that distribution. Taking the 2.5% and 97.5% bootstrap percentiles for CI construction was less sensitive to outliers, but when many of the bootstrap resamples provided a correction factor approaching zero, clearly the percentile-based CIs were affected too.

In our motivating example of active energy expenditure and lean body mass, RC provided an effect estimate that was large compared to the uncorrected estimate (-17.8 versus -0.7) accompanied with wide confidence intervals (-56.6;19.2 (Delta) and -270.0;217.2 (bootstrap)). The large width of the bootstrap confidence intervals can be explained by the fact that the correction factor was small and approached zero in some of the bootstrap resamples.

We only studied relatively simple settings, i.e., random measurement error and univariable models. However, the two phenomena explained here can be extended to settings where the measurement error is not random (e.g., in case of systematic measurement error) and in multivariable models. When differential measurement error is expected, the use of RC for measurement error correction is inappropriate [3, 15].

RC is not only suited for exposure measurement error correction in linear regression models but serves as a fair approximation in logistic regression and survival models as well [3]. In case of logistic regression and survival models, RC is only approximately consistent if 'measurement error is small' and the odds ratio or hazard ratio 'small to moderate'. See for a detailed discussion of RC for logistic regression, Kuha et al. [16] and for a detailed discussion of RC, Carroll et al. [17]. An investigation of the bias–variance trade off and finite sample performance of RC in logistic regression or survival models when measurement error is large is a topic for future research.

RC provides a valuable tool for exposure measurement error correction in epidemiologic studies but may not be particularly useful in settings where sample size is small and reliability of the error-prone exposure low. In those settings, it is advised to replace the substitute error-prone exposure by a more reliable measure of exposure and/or the collection of more data is needed.

References

[1] T. B. Brakenhoff, M. Mitroiu, R. H. Keogh, K. G. M. Moons, R. H. H. Groenwold, M. van Smeden, Measurement error is often neglected in medical literature: A systematic review, Journal of Clinical Epidemiology 98 (2018) 89–97. doi:10.1016/j.jclinepi.2018.02.023.

[2] P. A. Shaw, V. Deffner, R. H. Keogh, J. A. Tooze, K. W. Dodd, H. Küchenhoff, V. Kipnis, L. S. Freedman, Epidemiologic analyses with error-prone exposures: Review of current practice and recommendations, Annals of Epidemiology 28 (11) (2018) 821–828. doi:10.1016/j.annepidem.2018.09.001.

[3] R. H. Keogh, P. A. Shaw, P. Gustafson, R. J. Carroll, V. Deffner, K. W. Dodd, H. Küchenhoff, J. A. Tooze, M. P. Wallace, V. Kipnis, L. S. Freedman, STRATOS guidance document on measurement error and misclassification of variables in observational epidemiology: Part 1—Basic theory and simple methods of adjustment, Statistics in Medicine 39 (16) (2020) 2197–2231. doi:10.1002/sim.8532.

[4] J. A. Hutcheon, A. Chiolero, J. A. Hanley, Random measurement error and regression dilution bias, BMJ 340 (c2289) (2010). doi:10.1136/bmj.c2289.

[5] R. H. Keogh, J. W. Bartlett, Measurement error as a missing data problem, in: G. Yi, A. Delaigle, P. Gustafson (Eds.), Handbook of measurement error models, 1st Edition, CRC Press, Boca Raton, FL, 2021, Ch. 20, pp. 429–452.

[6] L. Nab, M. van Smeden, R. H. Keogh, R. H. H. Groenwold, Mecor: An R package for measurement error correction in linear regression models with a continuous outcome, Computer Methods and Programs in Biomedicine 208 (2021) 106238. doi:10.1016/j.cmpb.2021.106238.

[7] R. J. Carroll, D. Ruppert, L. A. Stefanski, C. M. Crainiceanu, Bias versus variance, in: Measurement error in nonlinear models, 2nd Edition, Chapman & Halll/CRC, Boca Raton, FL, 2006, Ch. 3, pp. 60–63.

[8] N. Biniaminov, Data from: Irisin, physical activity and fitness status in healthy humans: no association under resting conditions in a cross-sectional study, Dryad, Dataset (2019). doi:10.5061/dryad.ck501.

[9] N. Biniaminov, S. Bandt, A. Roth, S. Haertel, R. Neumann, A. Bub, Irisin, physical activity and fitness status in healthy humans: No association under resting conditions in a cross-sectional study, PLOS ONE 13 (1) (2018) e0189254. doi:10.1371/journal.pone.0189254.

[10] B. E. Ainsworth, W. L. Haskell, S. D. Herrmann, N. Meckes, D. R. Basset, C. Tudor-Locke, J. L. Greer, J. Vezina, M. C. Whitt-Glover, A. S. Leon, 2011 Compendium of physical activities, Medicine & Science in Sports & Exercise 43 (8) (2011) 1575–1581. doi:10.1249/MSS.0b013e31821ece12.

[11] B. Rosner, D. Spiegelman, W. C. Willett, Correction of logistic regression relative risk estimates and confidence intervals for measurement error: The case of multiple

covariates measured with error, American Journal of Epidemiology 132 (4) (1990) 734–745. doi:10.1093/oxfordjournals.aje.a115715.

[12] T. P. Morris, I. R. White, M. J. Crowther, Using simulation studies to evaluate statistical methods, Statistics in Medicine 38 (11) (2019) 2074–2102. doi:10.1002/sim.8086.

[13] A. Gasparini, Rsimsum: Summarise results from Monte Carlo simulation studies, Journal of Open Source Software 3 (26) (2018) 739. doi:10.21105/joss.00739.

[14] L. Nab, R. H. H. Groenwold, Sensitivity analysis for random measurement error using regression calibration and simulation-extrapolation, Global Epidemiology 3 (2021) 100067. doi:10.1016/j.gloepi.2021.100067.

[15] L. Nab, R. H. H. Groenwold, M. van Smeden, R. H. Keogh, Quantitative bias analysis for a misclassified confounder: A comparison between marginal structural models and conditional models for point treatments, Epidemiology 31 (6) (2020) 796–805. doi:10.1097/EDE.0000000000001239.

[16] J. Kuha, Corrections for exposure measurement error in logistic regression models with an application to nutritional data, Statistics in Medicine 13 (11) (1994) 1135–1148. doi:10.1002/sim.4780131105.

[17] R. J. Carroll, D. Ruppert, L. A. Stefanski, C. M. Crainiceanu, Regression calibration for survival analysis, in: Measurement error in nonlinear models, 2nd Edition, Chapman & Hall/CRC, Boca Raton, FL, 2006, Ch. 14, pp. 321–323.

5

Sampling strategies for internal validation samples for exposure measurement error correction

Statistical correction for measurement error in epidemiologic studies is possible, provided that information about the measurement error model and its parameters are available. Such information is commonly obtained from a randomly sampled internal validation sample. It is however unknown whether randomly sampling the internal validation sample is the optimal sampling strategy. We conducted a simulation study to investigate various internal validation sampling strategies in conjunction with regression calibration. Our simulation study showed that for an internal validation study sample of 40% of the main study's sample size, stratified random and extremes sampling had a small efficiency gain over random sampling (10% and 12% decrease on average over all scenarios, respectively). The efficiency gain was more pronounced in smaller validation samples of 10% of the main study's sample size, i.e., a 31% and 36% decrease on average over all scenarios, for stratified random and extremes sampling, respectively. To mitigate the bias due to measurement error in epidemiologic studies, small efficiency gains can be achieved for internal validation sampling strategies other than random, but only when measurement error is non-differential. For regression calibration, the gain in efficiency is, however, at the cost of a higher percentage bias and lower coverage.

This chapter is based on: L. Nab, M. van Smeden, R. de Mutsert, F.R. Rosendaal and R.H.H. Groenwold, Sampling strategies for internal validation samples for exposure measurement–error correction: A study of visceral adipose tissue measures replaced by waist circumference measures, American Journal of Epidemiology 190 (9) (2021) 1935–1947. doi:10.1093/aje/kwab114

5.1. Introduction

Preferred (or gold standard) measurements in large epidemiologic studies can be expensive, time consuming, invasive, or burdensome. These measures therefore are often replaced by simpler measures (less invasive, cheaper, faster), which are then assumed to highly correlate with the preferred measure. For example, consider studies of visceral adipose tissue (VAT), e.g. studies showing that higher values of VAT are associated with higher values of insulin resistance [1, 2]. Measurement of VAT involves magnetic resonance imaging (MRI) scans. Alternatively, measurement of waist circumference (WC), which requires only a measuring tape, can provide a proxy measure of VAT [3]. Nevertheless, the substitute measurements (e.g., WC) are not perfectly correlated with the gold standard (e.g., VAT) and, consequently, the substitute measurement can be viewed as an error-prone substitute for the gold standard.

Several methods have been developed to adjust for the bias in estimators of exposure-outcome associations when an exposure is measured with error [4–12]. Despite the abundance of literature on measurement error correction methodology, application of measurement error correction is still rare [13, 14]. Of the measurement error correction methods that are used, regression calibration is among the most commonly used in epidemiologic research [15], possibly because of its relative simplicity and the possibility to implement it in many situations [4, 7, 16, 17]. Regression calibration relies on information about the relation between the error-prone and the preferred (or gold standard) measurement, i.e., the measurement error model and its parameters. This relation can be estimated using an internal validation sample, a subset of the main study including individuals for whom both the error-prone substitute and gold standard measurement are available.

Several regression calibration methods have been proposed. In linear models, examples include standard and validation regression calibration (see e.g. [7]) as well as efficient regression calibration by Spiegelman et al. [18]. The efficiency of these different regression calibration methods has been compared in simulation studies (e.g., see [19]). Nonetheless, no studies have been conducted to investigate what internal validation sampling strategy (e.g., random, stratified random or extremes sampling) in conjunction with regression calibration provides the most efficient estimate of the corrected exposure-outcome association. The efficiency of regression calibration depends on the efficiency of the estimation of the calibration model, which may hypothetically be improved by sampling e.g. the extremes, assuming linear calibration models.

In the present study, we aim to compare different sampling strategies for the internal validation sample in combination with different regression calibration methods to correct for the bias in exposure-outcome associations caused by measurement error. First, we introduce the Netherlands Epidemiology of Obesity (NEO) study and illustrate three different internal validation sample sampling strategies. We then present a simulation study contrasting the finite sample properties of different sampling strategies of the internal validation sample in conjunction with regression calibration, motivated by the analysis of the NEO data. We conclude with a discussion of our results.

5.2. Case study: visceral adipose tissue measures as replacement for waist circumference measures

The NEO study is a large prospective observational cohort designed to investigate the pathways that lead to obesity-related diseases and conditions [20]. Men and women aged between 45 and 65 years with a self-reported body mass index of 27 or higher, living in the greater area of Leiden (in the West of the Netherlands) were eligible to participate in the NEO study. In addition, all inhabitants aged between 45 and 65 years from one municipality (Leiderdorp) were invited, irrespective of their body mass index, to represent the general population.

A cross-sectional analysis of the association between VAT and insulin resistance was conducted in the subset of individuals that originated from the Leiderdorp subcohort of the NEO study comprising of 1,670 individuals. VAT depots were quantified by means of MRI in a subsample of 668 (40%) individuals. These 668 individuals were randomly selected among the individuals who had no contraindication to undergo an MRI. WC was measured midway between the border of the lower costal margin and the iliac crest in all individuals. In this illustrative example we make two simplifying assumptions, 1) we consider WC measures as the error-prone substitute measure of the exposure of interest (i.e., VAT) and 2) we assume that WC is independent of insulin resistance given VAT and the confounding variables Z (i.e., non-differential measurement error). These two assumptions are summarized in the causal diagram in Figure 5.1. Violation of the non-differential measurement error assumption can lead to bias in both the regression calibration and internal validation analyses, under the circumstances explained in the 'Results' section below. For the assessment of insulin resistance, the homeostatic model assessment of insulin resistance was used as fasting glucose (in mmol/L) × fasting insulin (in mU/L)/22.5. Of the 668 selected individuals, 19 were excluded from analysis because they used glucose lowering therapy and, additionally, one patient was excluded because of a very low fasting glucose blood concentration. This resulted in including 648 individuals in our analysis. There were 22 missings in the selected variables for analysis, which were imputed once (single imputation), using multivariate imputation through chained equations by the package `mice` version 3.8.0 [21] with standard settings from the statistical software R [22]. The association between VAT and insulin resistance was adjusted for the potential confounding variables age, sex, ethnicity, educational level, smoking state, alcohol consumption, total body fat, physical activity, and additionally for hormonal use and menopausal state in women. We refer to [2] for further details on the assessment of all variables used in this study. Measures of VAT, WC and total body fat were standardized and measures of insulin resistance were log transformed. The effect sizes were derived from a linear regression analysis and expressed as percentages difference in outcome per standard deviation (SD) VAT.

After adjustment for confounding, insulin resistance was 27% higher (95% confidence interval (CI): 19%-35%) per SD VAT (54 cm^2). Alternatively, insulin resistance was 30% higher (95%CI: 18%-43%) per SD WC (12 cm), with adjustment for the same potential confounders as the association between VAT and insulin resistance. Under the assumptions depicted in Figure 5.1, the difference in these two estimates can be explained by the measurement error in WC as a measure of VAT.

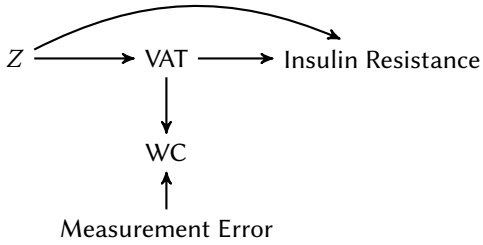


Figure 5.1: Assumptions of our motivating example. Error-prone waist circumference (WC) measures used as a substitute measurement to estimate the association between (VAT) and insulin resistance, confounded by Z (e.g., age, sex, total body fat).

5.2.1. Testing sampling strategies in a resampling study

To illustrate sampling strategies for an internal validation sample in combination with regression calibration to correct for measurement error, a resampling study was performed using data of the 648 individuals from the Leiderdorp cohort of whom both VAT and WC measures were taken. Five hundred new data sets were created by sampling from the 648 individuals with replacement. In each of the 500 resampled data sets, the association between VAT and insulin resistance was estimated (referred to as the reference analysis). In addition, WC measurements were considered as a proxy for VAT, and used to estimate the association between VAT and insulin resistance (referred to as the uncorrected analysis). Both analyses were adjusted for the same confounders as the original analysis presented above.

Next, 260 individuals (approximately 40% of 648) were included in the internal validation sample. This 40% was chosen to resemble the percentage of individuals of whom VAT depots were quantified of the whole Leiderdorp subcohort of the NEO study (i.e., in 668 individuals of the 1,670 individuals). The internal validation sample was sampled by using one of the following three sampling strategies: 1) random, 2) extremes or 3) stratified random (see next subsection). The VAT measures of all individuals who were not selected in the internal validation sample were removed. In each of these data sets, the association between VAT and insulin resistance was estimated by using only the information of the 40% of individuals included in the internal validation sample (internal validation sample restricted). Next, the VAT measurements available in the internal validation sample were used to correct for the measurement error in the association between WC and insulin resistance in three ways: 1) standard regression calibration, 2) validation regression calibration or 3) efficient regression calibration (see next subsection).

For each sampling strategy and each regression calibration method, the mean of the 500 effect estimates was calculated and corresponding 95% CIs were constructed based on the empirical standard errors. All analyses were adjusted for the above-mentioned potential confounders.

Sampling strategies and regression calibration methods. Figure 5.2 shows a visualisation of the three sampling strategies used in this study. The internal validation sample was sampled 1) randomly, 2) the 130 individuals with the lowest and 130 with the highest measured WC values were selected (extremes sampling) or 3) by grouping

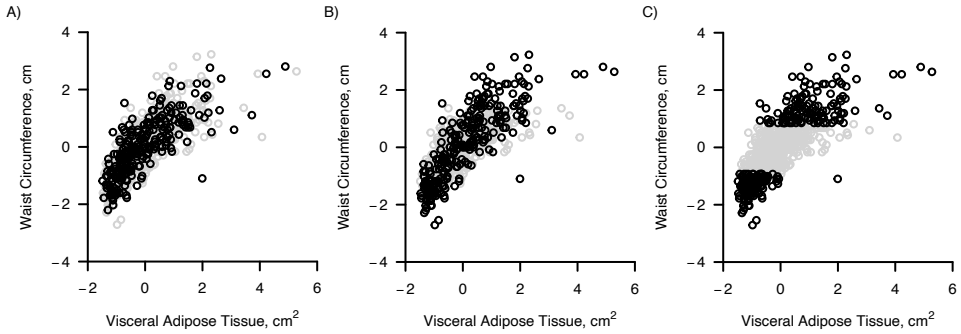


Figure 5.2: Visualisation of different internal validation sample sampling strategies in the Leiderdorp cohort of the Netherlands Epidemiology of Obesity. A) Visceral adipose tissue (VAT) measures are obtained at random (independent of waist circumference (WC)); B) VAT measures are obtained stratified randomly (stratified for strata of WC); and C) VAT measures are obtained in the individuals with the lowest and highest WC measures. The black points indicate the individuals included in the internal validation sample and the grey points the excluded individuals. The VAT measures and WC measures are standardized.

individuals according to tenths of the range of the measured WC values and sampling 26 individuals from each stratum (stratified random sampling). For stratified random sampling, when one of the strata contained less than 26 individuals, all individuals of this stratum were included in the internal validation sample. Subsequently, more than 26 individuals were sampled from the remaining strata, by equally distributing the shortage of individuals in the strata with less individuals among the strata with more individuals. We hypothesized that by sampling the extremes or by stratified random sampling, a linear relation between WC and VAT could be estimated more efficiently in the internal validation set. By increasing the efficiency of the estimation of the linear relation between WC and VAT, the efficiency of regression calibration was expected to increase simultaneously.

Three regression calibration methods were applied: 1) standard regression calibration, 2) validation regression calibration and 3) efficient regression calibration. Standard regression calibration and validation regression calibration are linear regressions where insulin resistance is regressed on a corrected version of the error-prone WC measures, and the confounding variables. Standard regression calibration replaces the error-prone WC measures with the predicted mean of VAT given WC and the confounding variables. Validation regression calibration replaces the error-prone WC measures with the predicted mean of VAT given WC and confounding variables for individuals not included in the internal validation sample. For the individuals included in the internal validation sample, the error-prone WC measurements are replaced by their VAT measurements. Efficient regression calibration takes the inverse variance weighted mean of the effect estimate of the internal validation sample restricted analysis (see above) and the standard regression calibration analysis. Further technical details (including standard error estimation) can be found in the supplementary material section S5.1.

Results. The results of the resampling study are shown in Table 5.1. In the uncorrected analysis, where WC was used to estimate the association between VAT and insulin resistance, the association between VAT and insulin resistance was overestimated compared with the reference analysis (30% vs 27%). When the internal validation sample was

Table 5.1: Estimated association between visceral adipose tissue and insulin resistance in the Leiderdorp cohort of the NEO study using different methods to correct for the measurement error when visceral adipose tissue measures were replaced by waist circumference measures

Method	Random		Stratified Random		Extremes	
	Effect Size (%)[a]	95% CI	Effect Size (%)[a]	95% CI	Effect Size (%)[a]	95% CI
IVS Restricted	26	14;40	20	9;33	18	7;31
Standard RC	67	24;126	60	25;105	59	24;104
Efficient RC	31	20;44	26	15;38	25	14;37
Validation RC	32	20;45	25	14;38	22	11;34

Abbreviations: CI = confidence interval; IVS = internal validation sample; and RC = regression calibration
[a] derived from β coefficients from linear regression analyses and expressed as percentages difference in outcome measure per standard deviation VAT; the effect size found in the reference analysis was 27% (95% CI 19%, 35%), the effect size found in the uncorrected analysis was 30% (18%,43%)

sampled randomly, the internal validation sample restricted analysis concurred with the reference analysis (26% vs 27%). However, the standard regression calibration approach overestimated the association between VAT and insulin resistance severely in comparison with the reference analysis (67% vs 27%). When the internal validation sample was sampled stratified randomly or by sampling the extremes, the internal validation restricted analysis underestimated the association between VAT and insulin resistance in comparison with the reference analysis (20% and 18%, respectively vs 27%). In comparison, the standard regression calibration analysis, again, severely overestimated the association between VAT and insulin resistance (60% and 59%, for stratified random and extremes sampling, respectively, vs 27%). Further, our results suggest that stratified random and extremes sampling improve the estimates of efficient regression calibration and validation regression calibration, since they appear to be closer to the reference analysis in comparison to random sampling, but this may be a chance finding due to cancellation of effects. Efficient and validation regression calibration are pooled averages of the underestimated association in the internal validation restricted analysis and the overestimated association in the standard regression calibration analysis. Specifically, the results of the standard regression calibration analysis are clearly biased for all sampling strategies, and we therefore expect the results of the efficient and validation regression calibration analyses to be biased as well.

The results of our empirical example seem to indicate that only the internal validation restricted analysis with a random sampling strategy concurs with the reference analysis. These results were not expected and can be explained by the fact that the measurement error in WC may depend on insulin resistance, since WC measures also provide a proxy for subcutaneous fat, which in turn is associated with insulin resistance. Consequently, the assumption of non-differential measurement error is violated. Particularly, to unbiasedly recover the exposure-outcome association under study, regression calibration relies on the assumption that the measurement error is non-differential. Furthermore, the internal validation sample restricted analysis is biased when the internal validation sample is obtained by sampling stratified randomly or extremes. In this case, sampling stratified randomly or the extremes introduced collider stratification bias, since inclusion in the internal validation sample is dependent on WC (depicted in the directed acyclic graph in

Figure 5.3). Consequently, the relation between VAT and insulin resistance is expected to be biased. Although sampling the internal validation sample other than randomly provides results that do not concur with the reference analysis here, general conclusions based on this empirical example are not warranted, which motivated our simulation study.

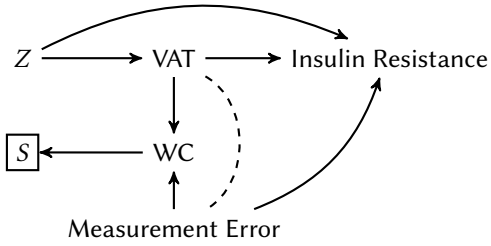


Figure 5.3: Collider stratification bias due to differential measurement error. Introduction of collider stratification bias when the data are observed (S) depending on the error-prone waist circumference (WC) measures with differential measurement error in a study estimating the association between (VAT) and insulin resistance, confounded by Z (e.g., age, sex, total body fat).

5.3. Simulation study

A simulation study was conducted to evaluate the finite-sample properties of the different internal validation sample sampling strategies combined with regression calibration. The sample size and the values of the parameters of the data generating mechanisms were similar to those estimated in the NEO subcohort mentioned in the previous section.

Generating data. Data sets were generated with a sample size of 650. The following data generating mechanisms were used to generate data on sex, age, total body fat (TBF), VAT, WC and insulin resistance (IR):

$$\text{sex} \sim \text{Bern}(0.5), \quad \text{age} \sim \text{Unif}(45, 65), \quad \text{TBF}|\text{sex, age} \sim \text{N}(-2 + \text{sex} + 0.01\times\text{age}, 0.5),$$
$$\text{VAT} = 0.4 - 2\times\text{sex} + 0.01\times\text{age} + 0.9\times\text{TBF} - (6\lambda\times\sqrt{\frac{0.5}{6\lambda}}) + \varepsilon, \quad \varepsilon \sim \text{Gamma}(6\lambda, \sqrt{\frac{0.5}{6\lambda}}),$$
$$\text{WC}|\text{VAT} \sim \text{N}(0.8\times\text{VAT}, \tau^2), \quad \text{and,}$$
$$\text{IR}|\text{VAT, sex, age, TBF} \sim \text{N}(0.5 + \beta\times\text{VAT} - 0.5\times\text{sex} + 0.01\times\text{age} + 0.3\times\text{TBF}, 0.3).$$

The estimand of this simulation study is the conditional effect of VAT on insulin resistance (i.e, β) and was set to 0.2. The parameters τ and λ were varied in different data generation scenarios of our simulation study. The variance of the measurement error (i.e., τ^2) was varied according to the explained variance of WC given VAT (hereafter referred to as R-squared). Values for R-squared were set to: 0.2, 0.4, 0.6, 0.8 and 0.9, corresponding values for τ can be found in Table S5.1a in the supplementary material section S5.2. For reference, the R-squared of the linear model of VAT and WC was approximately 0.6 in the NEO data. The above data generating mechanism for VAT allowed to change the skewness of the residual errors while keeping the mean and variance of the marginal distribution constant.

The skewness of the residual errors of VAT, ε, (hereafter referred to as skewness) were varied by changing λ. Values for the skewness were set to: 0.1, 1, 1.5 and 3, corresponding values for λ can be found in Table S5.1b in the supplementary material section S5.2. Additionally, we changed the distribution of WC|VAT by using the square root of VAT instead of VAT to generate WC, in what was called the non-linear scenario. R-squared, the skewness and linearity were varied in a full-factorial design (i.e., $5 \times 4 \times 2 = 40$ scenarios). For each scenario, 5000 datasets were generated.

Model estimation and performance measures. In each generated data set, we applied the three sampling strategies (i.e., random, extremes and stratified random sampling) and the five analyses (i.e., uncorrected, internal validation sample restricted and the three regression calibration analyses). Standard errors were calculated using standard software or by using the multivariate delta method, see for details supplementary material section S5.1. Subsequently, Wald based confidence intervals were constructed. Performance of the different analytical methods was evaluated in terms of the bias, mean squared error (MSE), the proportion of 95% CIs that contain the true value of the estimand (coverage), the empirical standard deviation of the estimated treatment effects and square root of mean model based variance of the estimated treatment effect. Monte Carlo standard errors (MCSE) were calculated for all performance measures [23], using the `R` package `rsimsum` version 0.9.0 [24]. All code used for the simulation study is publicly available at `https://github.com/LindaNab/me_neo`.

Sensitivity analyses. Two sensitivity analyses were conducted. First, to assess the sensitivity of our results to the size of the internal validation sample, we changed the percentage of individuals included to 10% and 25%. Second, in our empirical example in section 5.2, it was seen that the performance of the three regression calibration analyses was generally poor. We hypothesised that this is possibly due to differential measurement error in the WC measures. Differential measurement error occurs when WC depends on the outcome insulin resistance, conditional on VAT and the confounding variables (we refer to supplementary material section S5.1 for further details). To evaluate the impact of differential measurement error, one scenario was added by replacing the conditional distributions of WC and insulin resistance by:

$$\text{WC}|\text{VAT} \sim \text{N}(\theta \times \text{VAT} + \tau \times U, \tau^2) \quad \text{and,}$$

$$\text{IR}|\text{VAT, sex, age, TBF} \sim \text{N}(0.5 + \beta \times \text{VAT} - 0.5 \times \text{sex} + 0.01 \times \text{age} + 0.3 \times \text{TBF} + \sqrt{0.3} \times U, 0.3),$$

where U is a random variable with a Bernoulli distribution with mean 0.5. This scenario is an example of differential measurement error, since the distribution of the error-prone WC is dependent of the outcome insulin resistance via a third variable U, considered unmeasured. Here, τ was set equal to 0.44 (corresponding to an R-squared of 0.8 in the main study), the skewness of the residual errors of VAT was 0.1 and the estimand (β) was again 0.2.

5.3.1. Results

For brevity, here we do not show results of the scenarios where R-squared was equal to 0.9 or where skewness was equal to 1 (results are shown in Tables S5.2-S5.7 in the supplementary material section S5.3). The results of these parameter values did not contribute to the main comparisons made because the results of R-squared equal to 0.9 were similar to R-squared

equal to 0.8 and the results of skewness equal to 1 were similar to skewness equal to 1.5. Further, since the focus of this paper is the comparison between the three sampling strategies, we focus on the performance of the three sampling strategies in the internal validation restricted analysis and validation regression calibration. We chose to focus on validation regression calibration since this appears to be the standard method when applying regression calibration when there is an internal validation sample. The results of the sampling strategies using efficient regression calibration and standard regression calibration can be found in Figure S5.2-S5.3 and Tables S5.8-S5.18 in the supplementary material section S5.3.

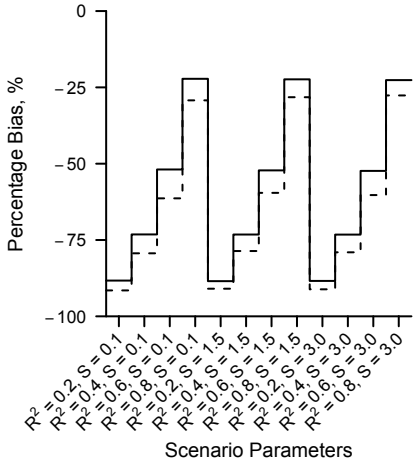


Figure 5.4: Nested loop plot of the percentage bias in the analysis ignoring measurement error. Solid line: Linear measurement error model; and dashed line: Non-linear measurement error model. Order from outer to inner loops: Skewness of the residual errors of the gold standard measure (S, 3 levels, increasing); R-squared of the measurement error model (R^2, 4 levels, increasing).

Figure 5.4 shows the percentage bias in the uncorrected analysis. In the uncorrected analysis, the association between VAT and insulin resistance was severely underestimated (bias ranging from -92% to -22%). The percentage bias decreased when R-squared increased and the bias in the uncorrected analysis was slightly higher when the measurement error model was non-linear compared to a linear model. The skewness of the residual errors of VAT had no bearing on bias.

Efficiency in terms of mean squared errors. Figure 5.5 shows the mean squared errors for the internal validation sample restricted analysis with an internal validation sample of 40% and 10% of the main study's sample size. Smaller mean squared errors were seen for stratified random and extremes sampling compared to random sampling for both samples sizes of the internal validation data. For the internal validation sample of 40% of the main study's sample size, the percentage decrease in mean squared error was 19% and 24% on average, for stratified and extremes sampling, respectively, MCSE < 0.0001. For the internal validation sample of 10% of the main study's sample size, the percentage decrease in mean squared error was 36% and 41% on average, for stratified and extremes sampling, respectively, MCSE < 0.0005. Most notably, mean squared errors decreased further for both stratified random and extremes sampling when the residuals error of VAT were more

skewed.

Figure 5.6 shows the mean squared errors for validation regression calibration with an internal validation sample of 40% and 10% of the main study's sample size. For the internal validation sample of 40% of the main study's sample size, mean squared errors were smaller for stratified random and extremes sampling compared to random sampling, with a 10% and 12% decrease on average, respectively, MCSE < 0.0001. For the internal validation sample of 10% of the main study's sample size, mean squared errors were found smaller for stratified random and extremes sampling compared to random sampling, with a 31% and 36% decrease on average, respectively, MCSE < 0.0005. The gain in efficiency was greatest for higher levels of skewness.

In a comparison between the internal validation restricted analysis and validation regression calibration, mean squared errors were generally smaller for validation regression calibration compared with the internal validation restricted analysis (compare Figure 5.5 and 5.6). The difference was more pronounced for high values of the R-squared and a validation sample of 10% of the main study's sample size.

The results for the internal validation restricted analysis and validation regression calibration with an internal validation sample comprising of 25% of the main study can be found in Figure S5.1 of supplementary material section S5.3.

Bias and coverage. Table 5.2 and 5.3 shows percentage bias and coverage of the internal validation restricted and the validation regression calibration analysis, respectively, with an internal validation sample of 40% of the main study's sample size. For the internal validation restricted analysis, all three different sampling strategies recovered the association between VAT and insulin resistance, with bias close to 0%. Additionally, coverage was close to the nominal level of 95% for all three sampling strategies. For the validation regression analysis and a randomly sampled internal validation sample, percentage bias was close to 0%. Contrary to random sampling, stratified random and extremes sampling introduced bias in the association under study. Which was greater for higher levels of the skewness and the R-squared. Coverage was close to the nominal level of 95% for random sampling. For stratified random and extremes sampling, coverage was close to the nominal level of 95% for all but the following three scenarios. There was undercoverage (91.5% and 91.9% (stratified) and, 90.1% and 90.1% (extremes)) in the linear setting when skewness was equal to 3.0 and R-squared was 0.6 or 0.8, respectively. Additionally, there was undercoverage (90.0% (stratified) and 91.3% (extremes)) in the non-linear setting when the skewness was equal to 3.0 and R-squared was 0.8.

Table 5.4 and 5.5 shows the percentage bias and coverage of the internal validation restricted and validation regression calibration analysis, respectively, with an internal validation sample of 10% of the main study's sample size. For the internal validation restricted analysis and all three sampling strategies, percentage bias and coverage were both close to levels of 0% and 95%, respectively. For validation regression calibration, the association between VAT and insulin resistance was biased in most scenarios. Percentages bias in the association under study ranged between −5.0% – 7.2% when skewness was equal to 0.1. When skewness was equal to 1.5 or 3.0, percentages bias ranged between −24.5% – 10.2%. Since the association under study was biased in almost all scenarios, the effect estimate was undercovered for most scenarios, and increasingly when residual errors were more skewed, since bias was greater in these settings. For random sampling, the association under study was undercovered with levels ranging between 82.7% – 92.9%. For

stratified random and extremes sampling, coverage was close to the nominal level of 95% when skewness was equal to 0.1 (ranging between 92.5%–95.4%). When skewness was equal to 1.5 or 3.0 the effect estimate was generally undercovered with levels ranging between 62.9% – 94.6%.

Differential measurement error. Table 5.6 shows that differential measurement error can cause bias in the association between VAT and insulin resistance. The internal validation sample restricted analysis using internal validation data that is sampled randomly recovers the association under study with percentage bias equal to 0%. The internal validation sample restricted analysis using stratified random or extremes sampling were both biased with percentage bias equal to 10% and 30%, respectively. The different regression calibration analyses were all biased, independent of how the internal validation sample was sampled.

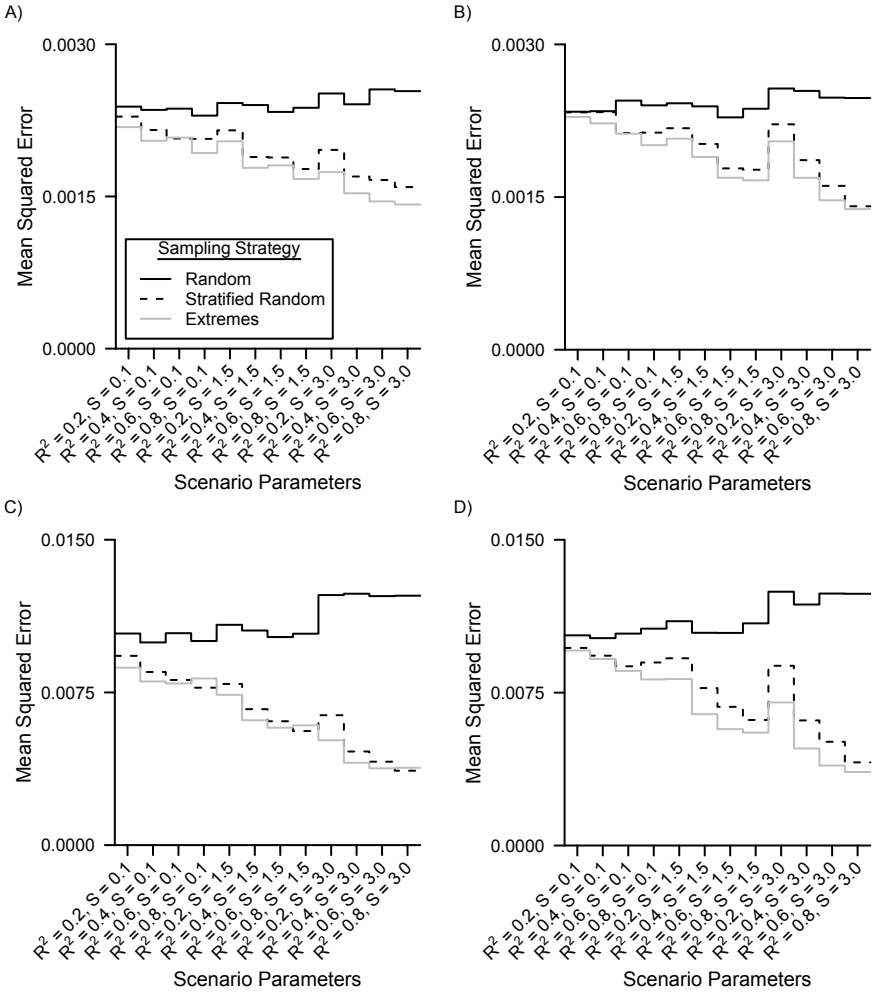


Figure 5.5: Nested loop plot of the mean squared errors in the analysis restricted to the internal validation sample for the three different sampling strategies. A) Linear measurement error model and an internal validation sample of 40% of the main study; B) Non-linear measurement error model and an internal validation sample of 40% of the main study; C) Linear measurement error model and an internal validation sample of 10% of the main study; and D) Non-linear measurement error model and an internal validation sample of 10% of the main study. Order from outer to inner loops: Skewness of the residual errors of the gold standard measure (S, 3 levels, increasing); R-squared of the measurement error model (R^2, 4 levels, increasing).

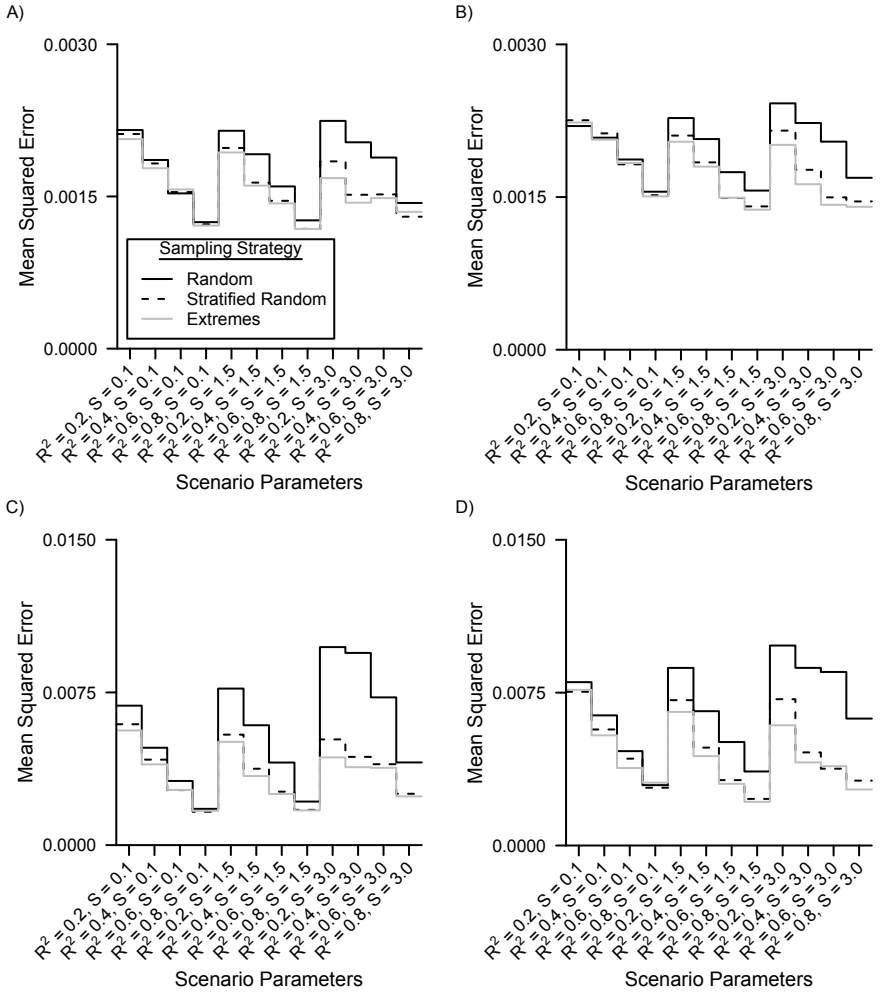


Figure 5.6: Nested loop plot of the mean squared errors in the analysis using validation regression calibration to correct for the measurement error for the three different sampling strategies. A) Linear measurement error model and an internal validation sample of 40% of the main study; B) Non-linear measurement error model and an internal validation sample of 40% of the main study; C) Linear measurement error model and an internal validation sample of 10% of the main study; and D) Non-linear measurement error model and an internal validation sample of 10% of the main study. Order from outer to inner loops: Skewness of the residual errors of the gold standard measure (S, 3 levels, increasing); R-squared of the measurement error model (R^2, 4 levels, increasing).

Table 5.2: Percentage bias and coverage in the estimated association between visceral adipose tissue and insulin resistance with an internal validation sample of 40% of the main study's sample size

Scenario			IVS Restricted Analysis					
Linear	R^2	Skewness	Percentage Bias (%)[a]			Coverage (%)[b]		
			R	SR	E	R	SR	E
Yes	0.2	0.1	−0.5	0.2	−0.1	94.9	94.8	95.1
		1.5	−0.1	−0.1	0.2	94.8	94.6	95.0
		3.0	−0.2	0.2	−0.1	94.7	94.4	94.7
	0.4	0.1	−0.1	0.4	0.1	95.0	95.3	94.9
		1.5	0.1	0.3	0.1	94.8	95.4	95.1
		3.0	0.3	0.0	0.2	95.3	94.9	94.9
	0.6	0.1	0.4	0.8	0.2	94.8	94.8	94.2
		1.5	0.1	−0.3	0.4	95.1	95.0	94.5
		3.0	0.0	−0.3	−0.1	94.8	94.8	94.6
	0.8	0.1	−0.3	0.1	0.1	94.9	94.7	95.3
		1.5	0.2	−0.2	−0.3	94.7	95.3	95.0
		3.0	0.0	−0.2	0.0	94.7	94.7	94.7
No	0.2	0.1	0.3	0.2	0.2	94.8	94.6	95.1
		1.5	−0.3	0.2	−0.2	94.6	95.0	95.4
		3.0	−0.2	0.2	0.1	94.3	94.5	94.3
	0.4	0.1	0.4	0.0	−0.1	95.3	94.4	94.9
		1.5	−0.6	−0.1	−0.2	94.8	95.4	95.0
		3.0	−0.2	−0.3	−0.4	94.6	94.3	94.4
	0.6	0.1	0.4	−0.4	−0.1	94.7	95.0	94.9
		1.5	0.2	0.4	0.4	95.1	95.3	95.4
		3.0	0.0	−0.1	0.0	94.6	94.8	94.5
	0.8	0.1	0.1	0.0	0.3	94.5	94.8	94.8
		1.5	0.0	−0.2	−0.2	94.9	94.4	94.8
		3.0	0.3	0.3	0.4	94.7	95.0	94.6

Abbreviations: IVS = internal validation sample; R = random; SR = stratified random; and E = extremes,
[a] Monte Carlo standard error (MCSE) <0.001, [b] MCSE <0.005

Table 5.3: Percentage bias and coverage in the estimated association between visceral adipose tissue and insulin resistance with an internal validation sample of 40% of the main study's sample size

Scenario			Validation Regression Calibration					
Linear	R^2	Skewness	Percentage Bias (%)[a]			Coverage (%)[b]		
			R	SR	E	R	SR	E
Yes	0.2	0.1	−0.5	0.2	−0.1	94.9	95.3	95.7
		1.5	−0.4	−0.7	−0.3	94.8	95.5	95.6
		3.0	−0.3	−1.2	−1.4	94.7	94.5	94.8
	0.4	0.1	−0.3	0.4	0.2	94.9	95.2	95.2
		1.5	0.1	−1.7	−1.9	94.8	95.0	95.2
		3.0	0.9	−4.1	−4.4	94.1	94.4	94.4
	0.6	0.1	0.6	0.9	0.6	95.2	94.9	94.8
		1.5	0.5	−3.3	−3.3	94.7	94.5	94.4
		3.0	0.9	−7.4	−8.9	93.2	91.5	90.8
	0.8	0.1	0.2	0.1	0.2	94.6	94.9	95.1
		1.5	0.4	−3.6	−4.2	94.9	94.7	94.0
		3.0	1.0	−7.7	−9.5	93.8	91.9	90.8
No	0.2	0.1	−0.2	0.1	0.1	95.3	94.9	95.4
		1.5	−0.7	−0.3	−0.5	94.7	95.2	95.5
		3.0	−0.5	−0.4	−0.4	94.7	94.7	94.7
	0.4	0.1	0.4	−0.1	−0.4	95.2	94.7	95.2
		1.5	−0.8	−1.3	−1.5	95.0	95.6	95.5
		3.0	−0.4	−2.7	−2.6	94.6	94.2	94.7
	0.6	0.1	0.1	−0.5	−1.0	94.8	95.2	94.8
		1.5	0.2	−2.2	−2.9	95.4	95.3	95.2
		3.0	0.2	−5.6	−5.6	94.0	93.5	93.5
	0.8	0.1	0.4	0.1	0.3	94.4	94.8	95.2
		1.5	−0.1	−5.7	−4.9	94.4	93.3	94.1
		3.0	1.0	−9.7	−9.1	94.0	90.0	91.3

Abbreviations: R = random; SR = stratified random; and E = extremes,
[a] Monte Carlo standard error (MCSE) <0.001, [b] MCSE <0.005

Table 5.4: Percentage bias and coverage in the estimated association between visceral adipose tissue and insulin resistance with an internal validation sample of 10% of the main study's sample size

Scenario			IVS Restricted Analysis					
Linear	R^2	Skewness	Percentage Bias (%)[a]			Coverage (%)[b]		
			R	SR	E	R	SR	E
Yes	0.2	0.1	−0.9	0.3	−0.6	94.2	94.5	94.1
		1.5	0.2	−0.4	0.0	94.2	95.0	94.0
		3.0	−0.2	0.2	0.2	94.5	94.5	94.6
	0.4	0.1	0.1	0.5	1.0	94.8	94.4	94.7
		1.5	−0.4	0.0	−0.3	95.1	94.8	94.4
		3.0	−0.2	−0.2	−0.1	94.6	94.8	94.4
	0.6	0.1	0.4	0.4	0.1	94.3	94.5	94.7
		1.5	−0.1	−0.4	0.2	95.3	94.3	94.5
		3.0	−0.2	−0.7	−0.2	94.3	94.0	94.4
	0.8	0.1	0.0	−0.6	−0.3	94.9	94.7	94.6
		1.5	−1.4	−0.5	−0.9	94.7	94.5	94.8
		3.0	−0.2	0.2	0.3	94.3	94.7	94.7
No	0.2	0.1	0.3	−0.7	1.1	94.3	94.0	94.3
		1.5	−0.1	0.1	−0.2	94.7	94.6	94.4
		3.0	−1.0	1.3	−0.2	94.2	94.0	94.5
	0.4	0.1	0.6	0.0	0.4	94.8	94.5	94.0
		1.5	−1.5	0.5	−1.0	94.3	94.5	94.5
		3.0	−0.4	−0.1	−0.1	94.9	94.4	95.0
	0.6	0.1	0.6	0.0	−0.1	94.7	94.7	94.2
		1.5	0.2	0.1	0.3	94.9	94.7	94.9
		3.0	−0.2	0.8	0.0	94.0	94.5	94.0
	0.8	0.1	−0.3	−0.2	0.4	93.7	94.2	94.2
		1.5	−0.8	−0.3	−0.5	94.3	94.0	94.2
		3.0	0.3	0.4	0.7	94.6	94.9	94.4

Abbreviations: IVS = internal validation sample; R = random; SR = stratified random; and E = extremes,
[a] Monte Carlo standard error (MCSE) <0.005, [b] MCSE <0.01

Table 5.5: Percentage bias and coverage in the estimated association between visceral adipose tissue and insulin resistance with an internal validation sample of 10% of the main study's sample size

Scenario			Validation Regression Calibration					
Linear	R^2	Skew-ness	Percentage Bias (%)[a]			Coverage (%)[b]		
			R	SR	E	R	SR	E
Yes	0.2	0.1	−1.4	0.1	−0.4	92.3	94.4	95.4
		1.5	−0.5	−4.0	−3.2	91.7	93.6	94.4
		3.0	1.4	−9.8	−8.1	89.1	89.5	91.3
	0.4	0.1	1.7	1.6	1.3	91.7	93.3	93.4
		1.5	3.9	−8.3	−8.2	89.7	89.3	91.5
		3.0	9.0	−20.1	−19.5	85.3	73.8	78.1
	0.6	0.1	2.9	2.0	1.8	91.2	92.7	93.3
		1.5	4.5	−10.9	−11.1	88.4	86.7	87.7
		3.0	10.2	−24.5	−24.5	82.7	62.9	65.5
	0.8	0.1	1.0	0.4	0.8	92.9	93.7	93.6
		1.5	2.5	−9.8	−8.9	91.1	88.7	89.0
		3.0	7.6	−19.0	−18.1	85.5	73.7	76.5
No	0.2	0.1	−5.0	−1.7	−0.5	92.9	94.2	94.9
		1.5	−3.7	−2.5	−3.0	92.2	94.2	94.6
		3.0	−3.7	−2.5	−3.8	91.9	93.5	94.2
	0.4	0.1	0.6	0.7	−1.7	92.3	93.9	93.9
		1.5	−0.4	−4.0	−8.7	91.4	93.0	92.7
		3.0	2.8	−10.2	−14.2	89.6	89.1	87.8
	0.6	0.1	1.2	2.3	−1.6	91.5	93.4	93.5
		1.5	3.5	−6.0	−10.8	90.5	92.0	91.2
		3.0	7.7	−16.4	−21.8	85.5	80.3	75.9
	0.8	0.1	2.0	4.1	7.2	91.6	92.6	92.5
		1.5	3.2	−8.6	−6.6	88.4	89.1	91.6
		3.0	8.8	−20.2	−18.3	83.5	71.2	77.7

Abbreviations: R = random; SR = stratified random; and E = extremes,
[a] Monte Carlo standard error (MCSE) <0.005, [b] MCSE <0.01

Table 5.6: Percentage bias in the estimated association between visceral adipose tissue and insulin resistance in case of differential measurement error

Method	Percentage Bias (%)[a]		
	Random	Stratified Random	Extremes
IVS Restricted	0	10	30
Standard RC	76	75	75
Efficient RC	42	45	46
Validation RC	35	36	36

Abbreviations: IVS = internal validation sample; and RC = regression calibration
[a] The percentage bias in the uncorrected analysis was 25%, Monte Carlo standard error < 0.001 for all analyses

5.4. Discussion

This study investigated three internal validation sampling strategies (i.e., random, stratified random and extremes sampling) in conjunction with regression calibration to correct for measurement error in a continuous exposure. Our simulation study showed a small efficiency gain in terms of mean squared error of stratified random and extremes sampling over a random sampling strategy for the internal validation restricted and regression calibration analyses, but only when measurement error was non-differential. For regression calibration, this gain in efficiency was at the cost of higher percentages bias and lower confidence interval coverage. We therefore recommend that, in general, regression calibration using randomly sampled validation samples are preferable over stratified or extremes sampled samples.

Three different regression calibration methods (i.e., standard, efficient and validation) and an internal validation sample restricted analysis were tested in our simulation study. The internal validation sample restricted analysis and validation regression calibration showed the best overall performance in terms of percentage bias and confidence interval coverage of the true effect. Furthermore, validation regression calibration had the same efficiency as efficient regression calibration under strong correlations between the exposure and outcome. These findings are consistent with the work by Thurston et al. [19]. In addition, a slight undercoverage of the confidence intervals was found for the efficient regression calibration approach.

Our simulation study showed a gain in efficiency of validation regression calibration over the internal validation sample restricted analysis. The gain in efficiency was more pronounced when the R-squared of the measurement error model was high and for smaller validation samples (e.g., 10% of full sample). Intuitively, the validation sample restricted analysis uses information about the gold standard measurement, but only for those individuals in whom it was measured (i.e., the internal validation sample). For regression calibration, however, information about all individuals is used, which tends to increase the efficiency, compared to the restricted analysis. However, the efficiency is negatively affected by the uncertainty in the correction factor that needs to be estimated from the internal validation sample. The relative gain in efficiency for regression calibration compared to an analysis of the gold standard measurement only (restricted to the validation sample) depends on the correlation between the gold standard and the error-prone measurement [15], as well as the appropriateness of parametric assumptions made for regression calibration.

Related work on internal validation studies can be found in the field of psychology, often referred there as 'two-method designs' or 'planned missing data designs'. These terms were recently suggested by Rioux et al. for epidemiologic research [25]. Graham et al. studied the cost-effectiveness of two-method designs and concluded that, in comparison with an analysis restricted to the internal validation sample, the two-method design can yield lower standard errors for testing associations using structural equation modelling [26]. In particular, the benefit of the design can be enormous when there is a large cost difference between the error-prone and the gold standard measures and effect sizes are small.

Regression calibration is one approach to correct for measurement error. Other measurement error correction methods include multiple imputation for measurement error [8], simulation-extrapolation [9], Bayesian methods [5] and methods based on

maximum likelihood estimation [27]. Earlier simulation studies have been conducted comparing multiple imputation for measurement error and regression calibration. These studies showed that, in general, multiple imputation for measurement error produced less biased estimates than regression calibration, but can have larger variances [8, 28, 29]. Simulation-extrapolation was originally designed to correct for measurement error that is random, which the measurement error in our case study was not. Adaptations have been made to also allow for systematic measurement error [30].

In our motivating example, regression calibration performed poorly. This was likely caused by violation of the non-differential measurement error assumption that regression calibration relies on and it signifies the importance of this assumption. WC measures may contain differential measurement error, since WC measures also provide a proxy for subcutaneous fat, which in turn is associated with insulin resistance. In our simulation study, where measurement error was known to be non-differential or differential, regression calibration performed well (for non-differential measurement error) or poorly (for non-differential measurement error), which further adds to our suspicion that differential measurement error might have affected the results of the motivating example.

Non-differential measurement error is a strong assumption and may be unlikely in practice [31]. Our motivating example signifies the importance of this assumption for measurement error correction and illustrates that when measurement error is differential, 1) regression calibration is not an appropriate method for measurement error correction and 2) non-random internal validation sampling strategies introduce collider stratification bias (see Figure 5.3). In the case differential measurement error is assumed, alternative methods for measurement error correction can be used, for example multiple imputation for measurement error [8] and regular multiple imputation methods [32–34]. Future research could investigate if non-random validation sample strategies improve the efficiency of multiple imputation methods for measurement error correction.

Large epidemiologic studies could consider to use internal validation samples when a gold standard measurement is expensive, time consuming, or burdensome. Our publicly available code at `https://github.com/LindaNab/me_neo`, provides an opportunity for careful planning of a sampling strategy, including the size of the internal validation sample, and the choice between an analysis restricted to the internal validation sample or application of regression calibration. The code can be adapted to accommodate other situations than the ones studied here.

In summary, our study showed that there appears to be little added value of stratified random or extremes sampling in internal validation studies to correct for measurement error. Regression calibration, if non-differential measurement error can be assumed, was shown to be an effective approach to correct analyses for measurement error. When handled with care, application of regression calibration can improve efficiency of epidemiologic studies with internal validation samples.

References

[1] M. Zhang, T. Hu, S. Zhang, L. Zhou, Associations of different adipose tissue depots with insulin resistance: A systematic review and meta-analysis of observational studies, Scientific Reports 5 (1) (2015) 18495. doi:10.1038/srep18495.

[2] R. de Mutsert, K. Gast, R. Widya, E. de Koning, I. Jazet, H. Lamb, S. le Cessie, A. de Roos, J. Smit, F. Rosendaal, M. den Heijer, Associations of abdominal subcutaneous and visceral fat with insulin resistance and secretion differ between men and women: The Netherlands epidemiology of obesity study, Metabolic Syndrome and Related Disorders 16 (1) (2018) 54–63. doi:10.1089/met.2017.0128.

[3] Z. Ping, X. Pei, P. Xia, Y. Chen, R. Guo, C. Hu, M. U. Imam, Y. Chen, P. Sun, L. Liu, Anthropometric indices as surrogates for estimating abdominal visceral and subcutaneous adipose tissue: A meta-analysis with 16,129 participants, Diabetes Research and Clinical Practice 143 (2018) 310–319. doi:10.1016/j.diabres.2018.08.005.

[4] B. Armstrong, Measurement error in the generalised linear model, Communications in Statistics - Simulation and Computation 14 (3) (1985) 529–544. doi:10.1080/03610918508812457.

[5] J. W. Bartlett, R. H. Keogh, Bayesian correction for covariate measurement error: A frequentist evaluation and comparison with regression calibration, Statistical Methods in Medical Research 27 (6) (2018) 1695–1708. doi:10.1177/0962280216667764.

[6] J. Buonaccorsi, Measurement error: Models, methods, and applications, Chapman & Hall/CRC, Boca Raton, FL, 2010.

[7] R. Carroll, D. Ruppert, L. Stefanski, C. Crainiceanu, Measurement error in nonlinear models: A modern perspective, 2nd Edition, Chapman & Hall/CRC, Boca Raton, FL, 2006.

[8] S. R. Cole, H. Chu, S. Greenland, Multiple-imputation for measurement-error correction, International Journal of Epidemiology 35 (4) (2006) 1074–1081. doi:10.1093/ije/dyl097.

[9] J. R. Cook, L. A. Stefanski, Simulation-extrapolation estimation in parametric measurement error models, Journal of the American Statistical Association 89 (428) (1994) 1314–1328. doi:10.2307/2290994.

[10] P. Gustafson, Measurement error and misclassification in statistics and epidemiology: Impacts and Bayesian adjustments, Chapman & Hall/CRC, Boca Raton, FL, 2004.

[11] W. Fuller, Measurement error models, John Wiley & Sons, New York, NY, 1987.

[12] R. H. Keogh, I. R. White, A toolkit for measurement error correction, with a focus on nutritional epidemiology, Statistics in Medicine 33 (12) (2014) 2137–2155. doi:10.1002/sim.6095.

[13] T. B. Brakenhoff, M. Mitroiu, R. H. Keogh, K. G. M. Moons, R. H. H. Groenwold, M. van Smeden, Measurement error is often neglected in medical literature: A systematic review, Journal of Clinical Epidemiology 98 (2018) 89–97. doi:10.1016/j.jclinepi.2018.02.023.

[14] P. A. Shaw, V. Deffner, R. H. Keogh, J. A. Tooze, K. W. Dodd, H. Küchenhoff, V. Kipnis, L. S. Freedman, Epidemiologic analyses with error-prone exposures: Review of current practice and recommendations, Annals of Epidemiology 28 (11) (2018) 821–828. doi:10.1016/j.annepidem.2018.09.001.

[15] R. H. Keogh, P. A. Shaw, P. Gustafson, R. J. Carroll, V. Deffner, K. W. Dodd, H. Küchenhoff, J. A. Tooze, M. P. Wallace, V. Kipnis, L. S. Freedman, STRATOS guidance document on measurement error and misclassification of variables in observational epidemiology: Part 1—Basic theory and simple methods of adjustment, Statistics in Medicine 39 (16) (2020) 2197–2231. doi:10.1002/sim.8532.

[16] R. L. Prentice, Covariate measurement errors and parameter estimation in a failure time regression model, Biometrika 69 (2) (1982) 331–342. doi:10.2307/2335407.

[17] B. Rosner, D. Spiegelman, W. C. Willett, Correction of logistic regression relative risk estimates and confidence intervals for measurement error: The case of multiple covariates measured with error, American Journal of Epidemiology 132 (4) (1990) 734–745. doi:10.1093/oxfordjournals.aje.a115715.

[18] D. Spiegelman, R. J. Carroll, V. Kipnis, Efficient regression calibration for logistic regression in main study/internal validation study designs with an imperfect reference instrument, Statistics in Medicine 20 (1) (2001) 139–160. doi:10.1002/1097-0258(20010115)20:1<139::AID-SIM644>3.0.CO;2-K.

[19] S. W. Thurston, P. L. Williams, R. Hauser, H. Hu, M. Hernandez-Avila, D. Spiegelman, A comparison of regression calibration approaches for designs with internal validation data, Journal of Statistical Planning and Inference 131 (1) (2005) 175–190. doi:10.1016/j.jspi.2003.12.015.

[20] R. de Mutsert, M. den Heijer, T. J. Rabelink, J. W. A. Smit, J. A. Romijn, J. W. Jukema, A. de Roos, C. M. Cobbaert, M. Kloppenburg, S. le Cessie, S. Middeldorp, F. R. Rosendaal, The Netherlands epidemiology of obesity (NEO) study: Study design and data collection, European Journal of Epidemiology 28 (6) (2013) 513–523. doi:10.1007/s10654-013-9801-3.

[21] S. van Buuren, K. Groothuis-Oudshoorn, Mice : Multivariate imputation by chained equations in R, Journal of Statistical Software 45 (3) (2011) 1–67. doi:10.18637/jss.v045.i03.

[22] R Core Team, R: A language and environment for statistical computing (2020). URL https://www.r-project.org/

[23] T. P. Morris, I. R. White, M. J. Crowther, Using simulation studies to evaluate statistical methods, Statistics in Medicine 38 (11) (2019) 2074–2102. doi:10.1002/sim.8086.

[24] A. Gasparini, Rsimsum: Summarise results from Monte Carlo simulation studies, Journal of Open Source Software 3 (26) (2018) 739. doi:10.21105/joss.00739.

[25] C. Rioux, A. Lewin, O. A. Odejimi, T. D. Little, Reflection on modern methods: planned missing data designs for epidemiological research, International Journal of Epidemiology 49 (5) (2020) 1702–1711. doi:10.1093/ije/dyaa042.

[26] J. W. Graham, B. J. Taylor, A. E. Olchowski, P. E. Cumsille, Planned missing data designs in psychological research, Psychological Methods 11 (4) (2006) 323–343. doi:10.1037/1082-989X.11.4.323.

[27] S. Rabe-Hesketh, A. Pickles, A. Skrondal, Correcting for covariate measurement error in logistic regression using nonparametric maximum likelihood estimation, Statistical Modelling 3 (3) (2003) 215–232. doi:10.1191/1471082X03st056oa.

[28] L. S. Freedman, D. Midthune, R. J. Carroll, V. Kipnis, A comparison of regression calibration, moment reconstruction and imputation for adjusting for covariate measurement error in regression, Statistics in Medicine 27 (25) (2008) 5195–5216. doi:10.1002/sim.3361.

[29] K. Messer, L. Natarajan, Maximum likelihood, multiple imputation and regression calibration for measurement error adjustment, Statistics in Medicine 27 (30) (2008) 6332–6350. doi:10.1002/sim.3458.

[30] J. Pina-Sánchez, Adjustment of recall errors in duration data using SIMEX, Metodološki zvezki - Advances in Methodology and Statistics 13 (1) (2016) 27–58.

[31] M. van Smeden, T. L. Lash, R. H. H. Groenwold, Reflection on modern methods: Five myths about measurement error in epidemiological research, International Journal of Epidemiology 49 (1) (2020) 338–347. doi:10.1093/ije/dyz251.

[32] M. Blackwell, J. Honaker, G. King, A unified approach to measurement error and missing data: Overview and applications, Sociological Methods & Research 46 (3) (2017) 303–341. doi:10.1177/0049124115585360.

[33] M. Blackwell, J. Honaker, G. King, A unified approach to measurement error and missing data: Details and extensions, Sociological Methods & Research 46 (3) (2017) 342–369. doi:10.1177/0049124115589052.

[34] J. K. Edwards, S. R. Cole, D. Westreich, All your data are always missing: incorporating bias due to measurement error into the potential outcomes framework, International Journal of Epidemiology 44 (4) (2015) 1452–1459. doi:10.1093/ije/dyu272.

6

Guidance for reporting of studies on incidence of venous thromboembolism in COVID-19 patients

Coagulation abnormalities and coagulopathy are recognised as consequences of Coronavirus disease (COVID-19) and venous thromboembolism (VTE) has been reported as a frequent complication. By 27 May 2021, at least 93 original studies and 25 meta-analyses investigating VTE incidence in COVID-19 patients had been published, showing large heterogeneity in reported VTE incidence ranging from 0–85%. This large variation complicates interpretation of individual study results as well as comparisons across studies, e.g., to investigate changes in incidence over time, compare subgroups, and perform meta-analyses. We identified different sources of heterogeneity in VTE incidence studies and classified these as clinical sources and methodologic sources. Clinical sources of heterogeneity include differences between studies regarding patient characteristics which affect baseline VTE risk and protocols used for VTE testing. Methodologic sources of heterogeneity include differences in VTE inclusion types, data quality and the methods used for data analysis. Each of these issues is discussed and illustrated using examples of VTE incidence studies in COVID-19 patients. To appreciate reported estimates of VTE incidence in COVID-19 patients in relation to its aetiology, prevention, and treatment, researchers should unambiguously report about possible clinical and methodological sources of heterogeneity in those estimates. This chapter provides guidance for that.

This chapter is based on: L. Nab, R.H.H. Groenwold, F.A. Klok, B.S. Bhoelan, M.J.H.A. Kruip and S.C. Cannegieter, Estimating incidence of venous thromboembolism in COVID-19: Methodological considerations, Research and Practice in Thrombosis and Haemostasis 6 (6) (2022) e12776. doi:10.1002/rth2.1277

6.1. Introduction

Coronavirus disease (COVID-19), caused by the virus SARS-CoV-2, primarily affects the respiratory system, but coagulation abnormalities and coagulopathy are also recognised as consequences [1]. In particular, venous thromboembolism (VTE) has been reported as a major complication, with VTE incidences up to 50% in intensive care unit (ICU) admitted patients [2]. By 27 May 2021, already 25 systematic reviews investigating VTE incidence in COVID-19 patients had been published [3–27], of which the most recent meta-analysis by Kollias et al. identified 93 unique studies on VTE incidence [13].

Systematic reviews of VTE incidence in COVID-19 patients show large heterogeneity [13], [20], [12]. For instance, Nopp et al. described VTE incidences in hospital admitted patients ranging from 0% – 40.3% [20], while Jiménez et al. even reported VTE incidences between 0% – 85% [12]. Kollias et al. restricted their systematic review to studies that performed "screening/assessment in the total sample for DVT (lower limb ultrasonography) or were focused on patients with suspicion for PE (whole study population subjected to tomography pulmonary angiogram)", but despite this restriction, they found heterogenous VTE incidences ranging between 0% – 85% [13]. Possible explanations for this heterogeneity in VTE incidence include differences in design of the study, clinical setting, and local practice (e.g., thromboprophylaxis strategy) [20], differences in endpoint definition, testing strategies, and patients' characteristics [12].

The wide variation in VTE incidence not only raises questions about the interpretation of individual study results, but, more importantly, complicates comparisons between studies to investigate e.g., changes over time, difference among subgroups, and to perform meta-analyses. To appreciate reported VTE incidences and to diminish their heterogeneity, it is important to understand different sources of this heterogeneity across studies. Therefore, we provide an overview of such sources of heterogeneity in VTE incidence studies on COVID-19 and illustrate this using various examples. Conclusively, we add a list of essential information to report, in order to improve consistency and hence the relevance of studies on VTE incidence in COVID-19 patients.

6.2. Methods

The large heterogeneity in VTE incidence across studies found in the meta-analyses by Jiménez et al. [12] and Nopp et al. [20] incentivised this project. On 27 May 2021, a pragmatic search on PubMed using the search string "meta-analysis covid-19 venous thromboembolism" was performed resulting in a rough estimate of the number of meta-analyses published. Twenty-five meta-analyses were identified. The most recent meta-analysis was published on April 4, 2021 by Kollias et al. [13]. The individual VTE incidence studies included in the meta-analyses by Jiménez et al., Nopp et al. and Kollias et al. were screened, and an initial list of potential sources of heterogeneity was created through discussions by LN, RHHG and SCC. The initial list was discussed in meetings with FAK, BSB and MJHAK, and revised until consensus was reached. For educational purposes, an example was sought for each listed source of heterogeneity by identifying two heterogenous studies also showing heterogeneity in their estimated VTE incidence, without taking other explanations into account. For consistency, incidences of all VTE incidence studies reported in this study were calculated as "number of cases during the entire study follow up divided by the size of the study population" and accompanied by a

95% Wald based confidence interval (CI).

6.3. Sources of heterogeneity of VTE incidence studies

Figures 6.1–6.2 provide an overview of the identified sources of variation in VTE incidence studies. A distinction was made between clinical (Figure 6.1) and methodologic sources (Figure 6.2). Clinical sources of heterogeneity include differences related to study characteristics affecting VTE risk and VTE testing. Methodologic sources of heterogeneity refer to differences in VTE manifestation inclusion types (e.g., inclusion of DVT, PE or both), data quality and the analytical methods used, i.e., what method was used to estimate VTE incidence and how limitations of the data were handled. These clinical and methodologic sources are explained in more detail below.

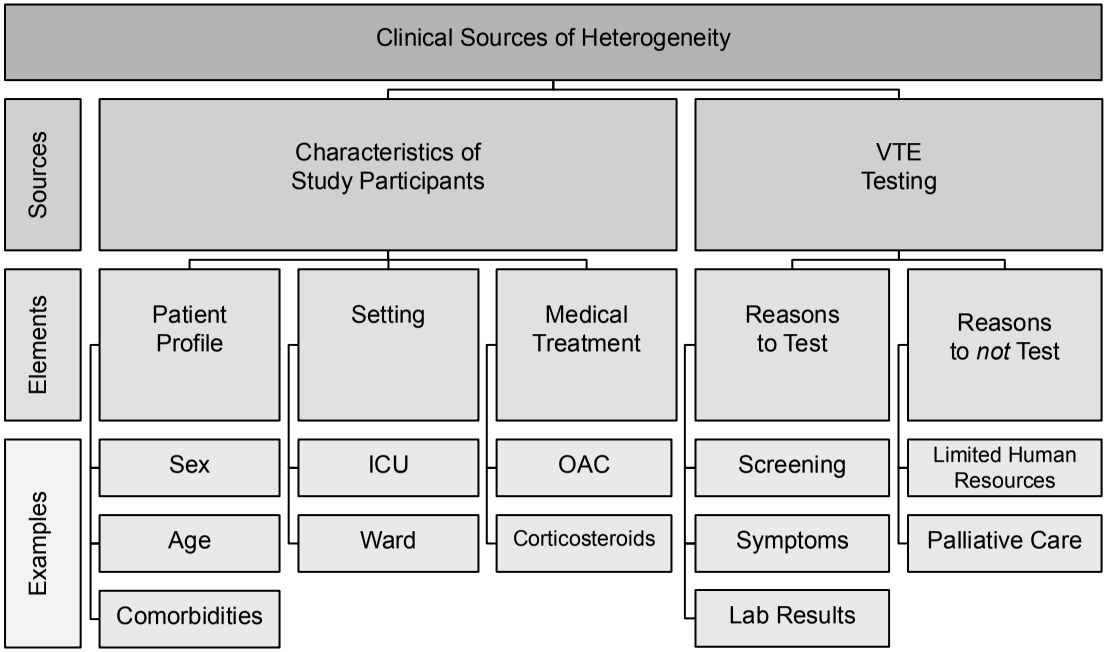


Figure 6.1: Clinical sources of heterogeneity in venous thromboembolism incidence studies that may explain observed heterogeneity across studies. ICU : Intensive Care Unit; OAC: oral anticoagulation.

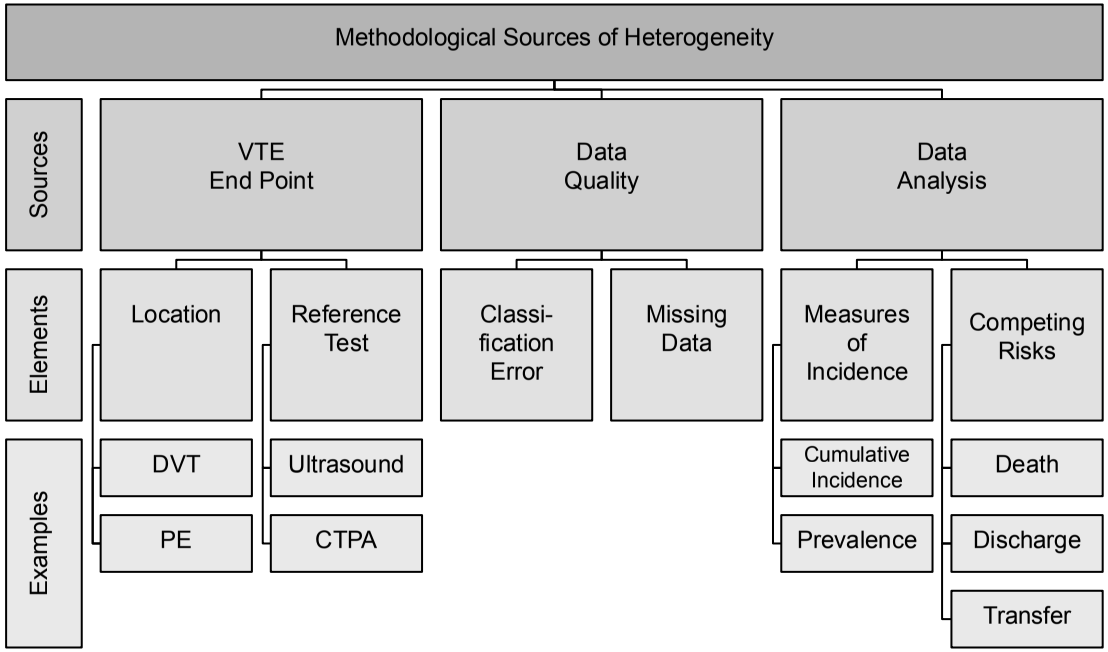


Figure 6.2: Methodological sources of heterogeneity in venous thromboembolism incidence studies that may explain observed heterogeneity across studies. DVT: deep vein thrombosis; PE: pulmonary embolism; CTPA: computed tomography pulmonary angiogram.

6.3.1. Clinical sources of heterogeneity

Patient characteristics

One potential source of heterogeneity in VTE incidence studies are differences in patient characteristics across studies. These patient characteristics are factors that increase or decrease the risk of VTE, i.e., established risk factors such as age and comorbidities [28] or ethnicity [29]. For example, Mei et al. performed a study among subjects with a mean age of 55.5 years (range 0.5-87), of whom 0.8% had a history of VTE. They reported a VTE incidence of 2.0% (95% CI 0.3;3.6).[30] In contrast, Middeldorp et al. reported a VTE incidence of 19.7% (95% CI 14.2;25.2) in a patient group that was older (mean age 61 years (standard deviation (SD) 4)) and in whom a history of VTE was more frequent (5.6%). [31] Hence, the underlying VTE risk may have been higher in the latter study, which may be one of the factors explaining the higher VTE incidence found in that study.

In addition to these patient profiles, also relevant are characteristics of the research setting, related to COVID-19 disease severity or VTE risk. For example, critically ill patients are at higher risk of developing VTE compared to non-ICU patients [32], so inclusion of patients from the ICU, the general ward or both, affects VTE incidence. In the study by Al-Samkari et al., patients from both the general ward and ICU were included. In their study, a VTE incidence was found of 3.1% in ward patients (95% CI 1.0;5.3) and of 7.6% in ICU patients (95% CI 3.3-12.0).[33] Of note, the case mix of COVID-19 ICU patients may differ between countries, due to national-level differences in accessibility of intensive care beds [34]. What is more, VTE incidence in COVID-19 outpatients is different from VTE incidence in hospitalised COVID-19 patients [35]. Limiting or not limiting the research setting to patients with e.g. an elevated D-dimer level may affect VTE incidence, because patients with an elevated D-dimer level are at high risk of developing VTE [36]. For example, Demelo-Rodríguez et al. only included patients with a D-dimer level >1000 ng/ml and reported a VTE incidence of 14.7% (95% CI 9.2;20.3) [37]. In comparison, Whyte et al. did not use a D-dimer level threshold to restrict patient inclusion and included all hospitalised COVID-19 patients and found a VTE incidence of 5.4% (95% CI 4.3;6.6) [38]. The research setting and VTE risk may also be affected by the way in which patient selection was performed. For example, in a study by Hill et al., a VTE incidence of 1.4% (95% CI 1.1;1.7) was found. In this study, patients were included retrospectively, by screening electronic health records and including all patients positive for Sars-CoV-2 on polymerase chain reaction-based testing. Patients were included after a visit to an emergency department and/or admission to an inpatient unit.[39] In comparison, in a study by Trimaille et al., a VTE incidence of 17.0% (95% CI 12.6;21.3) was found. In this study, all consecutive patients who were hospitalised for COVID-19 were included.[40] In these two studies, the clinical characteristics of the underlying populations where the study populations were sampled from were different, which may have affected baseline risk for VTE.

A third characteristic that requires attention is the local medical strategy, such as the use of anticoagulation treatment or COVID-19 treatment, which may influence the risk of VTE. A patient group that is treated with full dose anticoagulation may show a lower VTE incidence compared to patient groups receiving no or prophylactic anticoagulation. Cattaneao et al. reported a VTE incidence of 0.0% (95% CI 0.000;0.008). In this study, all patients had been treated with standard dose thromboprophylaxis.[41] In comparison, Zhang et al. reported a VTE incidence of 46.2% (95% CI 38.0;54.3). In this study, 90 patients out of 143 patients (63%) received no anticoagulation at all. [42] The difference in VTE

incidence between these two studies could be partly explained by the difference in use of anticoagulation. Additionally, as of 2 September 2020, the WHO recommended the use of systemic corticosteroids in patients with severe or critical COVID-19 [43–45]. For example, a meta-analysis showed an increased risk of 1.39 (95% CI 1.10-1.77) of VTE in COVID-19 patients when being administered corticosteroids [46]. The use of corticosteroids may therefore also be a source of heterogeneity in VTE incidence.

VTE testing

An additional clinical source of heterogeneity in VTE incidence is variation in VTE diagnostic practices. In particular, reasons to test for VTE and reasons to *not* test for VTE differed across studies. For example, diagnostic tests for PE or DVT may have been conducted at the occurrence of symptoms. Alternatively, a patient may have been tested for VTE independent of symptoms (i.e., screening), detecting both symptomatic and asymptomatic cases. For example, Cattaneao et al. found no symptomatic DVT cases in their study, resulting in a VTE incidence of 0.0% (95% CI 0.000;0.008) [41]. In comparison, Demelo-Rodríguez et al. screened for asymptomatic DVT and found an incidence of 14.7% (95% CI 9.2;20.3) [37]. Furthermore, studies may use a decision rule (e.g., based on lab results) before undertaking imaging. For example, Whyte et al. found a VTE incidence of 5.4% (95% CI 4.3;6.6). In this study, imaging was not undertaken "for those considered 'PE unlikely' by the Wells score (score <4) in conjunction with a D-dimer result below 500 ng/ml".[38] In comparison, Voicu et al. performed ultrasound imaging in all patients 3 days after intubation and found a VTE incidence of 35.7% (95% CI 23.2;48.3) [47]. Studies using a decision rule for VTE imaging, may miss cases of VTE [48], but are a closer resemblance of clinical practice. What is more, despite a VTE testing protocol in place, VTE testing may not be reasonable or feasible in some patients and studies may therefore deviate from their testing protocol in these cases. For example, VTE testing may not be reasonable in patients hospitalised for palliative care and VTE testing may not be feasible due to limitations in (human) resources in a health crisis setting, since performing a computed tomographic pulmonary angiogram (CTPA) in an ICU patient with mechanical ventilation can be laborious. For instance, Koleilat et al. reported a VTE incidence of 0.5% (95% CI 0.3;0.8). In this study, only "those patients with significant clinical concern for DVT or in those in whom the results were deemed to impact management were tested for DVT (e.g., patients who were mechanically ventilated and placed prone for persistently poor oxygenation were deemed too unstable, and those already on anticoagulation for other reasons such as cardiac arrhythmias or a prior history of thrombotic episodes requiring lifelong anticoagulation were unlikely to undergo venous duplex testing)".[49] In comparison, Ren et al. reported a (both a- and symptomatic) VTE incidence of 85.4% (95% CI 75.4;95.4). In the latter study, all patients were tested for DVT at least twice.[50]

6.3.2. Methodologic sources

VTE endpoint

Heterogeneity in VTE incidence studies may be caused by inconsistent inclusion of types of VTE across studies. For example, Mazzaccora et al. reported a VTE incidence of 65.6% (95% CI 49.2%;82.1%), where VTE included pulmonary embolism, diagnosed using a CT pulmonary angiogram, or DVT, which was diagnosed with an ultrasound of the veins of the upper and lower limbs.[51] In contrast, the study by Criel et al. reported a VTE incidence

of 7.3% (95% CI 1.7;13.0). Here, VTE included DVT only. [52] Furthermore, thrombosis in other venous compartments may be included (e.g., upper extremity, splanchnic veins). In the above mentioned study by Mazzaccaro et al. all patients underwent a duplex scan of the veins and arteries of the upper and lower limbs to investigate the presence of peripheral thrombosis.[51] In contrast, Santoliquido et al. reported a VTE incidence of 11.9% (95% CI 5.0;18.8).[53] Here, all patients were screened for DVT with lower-limb venous compression ultrasound. Another example is the study by Llitjos et al., that found a VTE incidence of 69.2% (95% CI 51.5;87.0) in which 4 of 18 (22%) reported DVTs were superficial. In comparison, the study by Desborough et al. did not include superficial DVTs and found a VTE incidence of 15.2% (95% CI 6.5;23.8). In addition, a distinction can be made between central, segmental and subsegmental PEs, based on the location of thrombi in the pulmonary vascular tree. Longchamp et al. "did not record subsegmental PEs" and reported a VTE incidence of 56.0% (95% CI 36.5;75.5) [54]. In comparison, Mazzaccaro et al. found a VTE incidence of 65.6% (95% CI 49.2;82.1), based on 21 cases of 'pulmonary vessel thrombosis', including 7 (33.3%) cases of subsegmental thrombi [51]. Additionally, DVT may be associated with indwelling lines and these DVTs could be included or not. For example, in the study by Desborough et al. [55], 10 of the 66 patients were diagnosed with VTE, resulting in a VTE incidence of 15.2% (95% CI 6.5;23.8). However, 6 of the DVTs were found to be associated with a line and one patient had a both a DVT and a PE. Consequently, 5 patients had a none line associated VTE, changing VTE incidence to 7.6% (95% CI 1.2;14.0).

VTE endpoint classification may also differ across studies in terms of the protocol that was used for the interpretation of the CT or ultrasound test by radiologists for a VTE diagnosis. Chen et al. reported a VTE incidence of 1% (95% CI 0.4;1.6). In this study, "all CT and CTPA image analyses were performed by 2 radiologists experienced in thoracic radiology [...], who were blinded to the clinical information. Disagreements were resolved through discussion until consensus was reached".[56] Conversely, Artifoni et al. reported a VTE incidence of 22.5% (95% CI 12.8;32.3) in a study where "chest angio-CT scan was performed in case of suspicion of pulmonary embolism" [57]. In the study by Chen et al., classification error in VTE diagnosis is less likely, while the study by Artifoni et al. more closely resembles clinical practice.

In addition, a potential source of heterogeneity in the VTE endpoint across studies is the data source used to classify the VTE endpoint. Data quality may differ between data sources, which is discussed in the subsequent subsection.

Data quality

A potential source of heterogeneity in VTE incidence studies is classification error or missing data in the VTE endpoint or in SARS-CoV-2 infection status. Thirty-four out of the 49 VTE incidence studies (69%) included in the meta-analysis by Jiménez et al. were identified as retrospective studies, 11 (22%) were identified as prospective studies, and 4 (8%) as cross-sectional studies [12]. In the 34 retrospective studies, data were often not primarily collected to study VTE incidence, which increases the potential risk for incorrectness of VTE endpoint classification and SARS-CoV-2 infection (i.e., a false-positive or false-negative diagnosis). Misclassification error occurs for example due to an incorrect interpretation of a radiologist or errors in data extraction and entering in databases. Specifically, in the study by Hill et al., reporting a VTE incidence of 3.1% (95% CI 2.5;3.8), electronic health records were queried to identify patients with diagnosis for VTE. Patients

were identified as cases when they received apixaban, rivaroxaban, or dabigatran.[39] Consequently, classification error in the VTE endpoint may be more likely in the study by Hill et al. than in a study using clinical radiology reports for VTE endpoint classification. Radiology reports were for example used in the study by Chen et al., reporting a VTE incidence of 1.0% (95% CI 0.4;1.6).[56]

The consequences of misclassification error and missing data in SARS-CoV-2 infection status is illustrated by the VTE incidence study conducted by Koleilat et al. [49]. In the meta-analysis conducted by Jiménez et al. [12], a VTE incidence of 0.5% (95% CI 0.3;0.8) was reported for the study by Koleilat et al.. This number was calculated by dividing the number of patients with a DVT diagnosis (18) by the number of patients admitted to the hospital with "confirmed COVID-19" (3404). It is unknown what was meant by "confirmed COVID-19" in the original article. The flowchart published in the paper by Koleilat et al. shows that of the 3404 patients with "confirmed COVID-19", 846 patients underwent lower extremity venous duplexes, of whom 145 patients were tested SARS-CoV-2 negative, 135 patients were tested SARS-CoV-2 positive, and 566 were not tested for SARS-CoV-2. Koleilat et al. did not report the SARS-CoV-2 status of the remaining 2558 patients that did not undergo lower extremity venous duplexes. Consequently, due to missing data in SARS-CoV-2 infection status, and potential misclassification error in the "confirmed COVID-19" cases, it is unknown what the correct estimate for VTE incidence was in this study. As an example, if VTE incidence calculation had been restricted to those patients who were known to be tested SARS-CoV-2 positive (135), VTE incidence would change dramatically to (18/135) 13.3% (95% CI 7.6;19.1).

Another example of the consequences of missing data in VTE diagnosis is the VTE incidence study conducted by Chen et al. In this study, 1,008 patients were hospitalised with COVID-19 associated pneumonia.[56] A VTE was diagnosed in 10 patients, and consequently, the meta-analysis by Jiménez et al. reported a VTE incidence of 1.0% (95% CI 0.4;1.6) [12]. In the analysis reported in the original paper, all patients were excluded who did not undergo CTPA examination, since these patients had missing data in VTE diagnosis (yes/no). Restricting the analysis to these 25 patients, would change VTE incidence dramatically to 40.0% (95% CI 20.8;59.2). Including or excluding the patients not undergoing VTE testing, changes the research setting of a study (see also section Patient characteristics). Particularly because, most likely, the odds of a VTE was a priori lower in the group not undergoing a CTPA, compared with the group undergoing a CTPA which was only performed in patients with "elevated D-dimer level or accompanying symptom(s), including chest pain, hemoptysis, and dyspnea" [56].

Data analysis

The frequency measures used to describe VTE incidence may also differ across studies, possibly contributing to heterogeneity in VTE incidence. Commonly used, yet distinct, measures include the cumulative incidence (or risk or incidence proportion), the prevalence and incidence rate.

The risk (or cumulative incidence) is the probability of getting a VTE in a certain period of time and is calculated by dividing the number of subjects who experienced the outcome in a certain time period by the total number of subjects that was observed during that time period. [58] Estimating this measure requires that patients are followed for the entire time period. In addition, interpretation of a risk is only warranted when the length of the time

period over which the risk applies is known, which should therefore be reported. When the time period is small, VTE incidence would approach 0, whereas as the time period becomes longer, VTE incidence will increase [59]. Whyte et al. reported a cumulative incidence at 24 hours of 2.1% (95% CI 1.4;2.8). In comparison, Middeldorp et al. reported cumulative incidences of VTE at 7 days, 14 days and 21 days of 16% (95% CI 10;22), 33% (95% CI 23;43) and 42% (95% CI, 30;54), respectively [31]. Since follow-up duration was highly variable in general, proper interpretation of the reported risks is a challenge and forms a possible explanation for the heterogeneity in VTE incidence.

Prevalence is defined as the proportion of COVID-19 patients with VTE at a particular moment in time [58]. In a cross-sectional study conducted by Criel et al., 82 patients were included and consecutively screened for VTE. In this study, the patients were not followed over time, and consequently, this study estimated VTE prevalence rather than incidence, which was found to be 7.3% (95% CI 1.7;13.0). [52] In comparison, for example, Desborough et al. followed patients for 28 days after admission to critical care (or until death) and, reported a VTE incidence of 15.2% (95% CI 6.5;23.8) [60]. These two estimates are heterogenous since VTE prevalence (Criel et al.) and VTE incidence (Desborough et al.) are incomparable.

Another measure of VTE incidence is the VTE incidence rate, which explicitly takes the duration of follow-up into account. It can be calculated by dividing the number of COVID-19 patients who developed VTE by the total amount of time those patients were followed. For example, Klok et al. reported a VTE incidence rate of 13/patient-year (95% CI 6;27) [2]. This method implicitly takes the length of the follow-up period, and variations between patients, into account.

Another source of heterogeneity across studies is how competing risks are handled in the analysis. In VTE studies of COVID-19, patients can develop a VTE, die, be discharged from the hospital, or be transferred to another hospital. It is often not known whether these patients developed a VTE (or not) after discharge or hospital transfer and it is impossible to know whether these patients would have developed a VTE if they had not died. Moreover, if a patient dies, autopsy should be performed to identify whether the cause of death was a VTE, which usually does not happen. In these patients, the VTE endpoint may, therefore, be misclassified (false-negative). For example, Middeldorp et al. reports "we did not adjudicate deaths to identify fatal PE because almost all deaths were due to hypoxemic respiratory failure, which can be indistinguishable from fatal PE, whereas autopsies were rarely performed in COVID-19 patients".[31] Adjusting or not adjusting for competing risks affects the reported cumulative incidence. Klok et al., reported a crude cumulative incidence of 57% (95% CI 47;67) and a cumulative incidence adjusted for the competing risk of death of 49% (95% CI 41;57). A cumulative incidence adjusted for the competing risk of death and hospital discharge may have decreased the cumulative incidence further since 43% of the patients included in the study were discharged alive.[2]

6.4. Discussion

Studies on VTE incidence in COVID-19 patients show highly heterogeneous results. We identified different sources of this phenomenon, notably, clinical and methodological sources, and illustrated these using various examples. The list of sources of heterogeneity in VTE incidence studies described here (characteristics of study participants, VTE testing,

VTE endpoint, data quality, and data analysis) is not exhaustive and more aspects may be needed to fully comprehend the heterogeneity across studies. Nevertheless, we consider these sources to be important explanations and, therefore, we feel that reporting of these aspects in future VTE incidence studies is required to appreciate and to properly interpret reported VTE incidences. A list of these recommendations is provided in Box 6.1.

We discussed individual sources of heterogeneity in VTE incidence, but obviously these could occur simultaneously. When two studies differ regarding multiple sources of heterogeneity, the difference in VTE incidence could increase but it could also lead to a cancellation of effects. We do not mean to suggest that the differences in VTE incidence in our examples are caused by the discussed sources of heterogeneity. We solely provide one of the many explanations for a difference in VTE incidence. The example studies referenced in this chapter are merely illustrations and do not reflect our view about their quality.

The heterogeneity in reports of VTE incidence not only complicates interpretation of VTE incidence but may also affect trials using VTE incidence as the primary endpoint or one of the secondary endpoints, such as trials comparing different thromboprophylaxis strategies in COVID-19 patients. Specifically, the sample size of these trials may be based on a reported VTE incidence which may not reflect VTE incidence in the research setting of the trial, leading to an under- or overpowered study. For example, the study by Connors et al., studying the effect of antithrombotic therapy on clinical outcomes in outpatients, "was terminated because of a control event rate lower than anticipated" [35]. It is unclear if sample size calculation was directly affected by studies reporting high VTE incidence in (ambulatory) COVID-19 patients. Connors et al. assumed a placebo event rate of 8% as "previous trials of anticoagulants for prevention of thrombotic events in ambulatory patients have noted similar event rates". What is more, most, if not all, aspects described in this paper translate to trial settings. For example, if no clear and unambiguous description is provided of VTE endpoint assessment in trials, estimates of the treatment effect (e.g., risk difference), or the number needed to treat derived from it, cannot be interpreted.

Standardising VTE research is not limited to COVID-19. Several efforts have already been made to improve the quality and consistency of VTE clinical research data and reporting practices. Examples include, amongst others, the VTE Common Data Elements project launched in November 2018 by the International Society on Thrombosis and Haemostasis [61] and recommendations for standardised reporting and analysis of VTE in oncology trials [62].

Standardising reports of VTE incidence studies is important and allows for comparisons of VTE incidence across groups (e.g., hospitals, countries, regions, sex, or over time), across diseases (e.g., influenza), and for better understanding and comparison of the results of trials on treatments aimed at reducing VTE incidence. Careful description of the elements affecting heterogeneity in future VTE incidence studies may better isolate important differences across groups, diseases and treatments and allow meta-analyses that provide summary results based on more homogeneous studies. Eventually such literature will contribute to improved management of VTE risk in COVID-19 patients.

Box 6.1: Recommendations for reporting of studies of incidence of venous thromboembolism (VTE) in Covid-19 patients

Clinical sources

Characteristics of study participants

- Describe the patient profiles (e.g., sex, age, comorbidities)
- Describe the research setting (e.g., intensive care unit, ward)
- Describe the patients' medical treatments (e.g., anticoagulation, steroids)

VTE testing

- Describe the VTE testing protocol (e.g., screening, symptoms, testing based on lab results)
- Describe the reasons to deviate from the VTE testing protocol (e.g., when testing was not reasonable (e.g., palliative care) or feasible (e.g., limited (human) resources) or when testing had no clinical consequences)

Methodological sources

VTE endpoint

- Describe the types of VTE that were included (e.g., pulmonary embolism, deep vein thrombosis)
- Describe the reference test used (e.g., ultrasound, computed tomographic pulmonary angiogram)

Data quality

- Describe the likelihood of classification error and its consequences (e.g., classification error in VTE diagnosis and SARS-CoV-2 infection)
- Describe missing data and its consequences (e.g., missing data in VTE diagnosis and SARS-CoV-2 infection)

Data analysis

- Describe the measure of incidence used and report its unit (if applicable) (e.g., cumulative incidence, prevalence)
- Describe competing risks (e.g., death, discharge, transfer)

References

[1] M. F. Osuchowski, M. S. Winkler, T. Skirecki, S. Cajander, M. Shankar-Hari, G. Lachmann, G. Monneret, F. Venet, M. Bauer, F. M. Brunkhorst, S. Weis, A. Garcia-Salido, M. Kox, J.-M. Cavaillon, F. Uhle, M. A. Weigand, S. B. Flohé, W. J. Wiersinga, R. Almansa, A. de la Fuente, I. Martin-Loeches, C. Meisel, T. Spinetti, J. C. Schefold, C. Cilloniz, A. Torres, E. J. Giamarellos-Bourboulis, R. Ferrer, M. Girardis, A. Cossarizza, M. G. Netea, T. van der Poll, J. F. Bermejo-Martín, I. Rubio, The COVID-19 puzzle: Deciphering pathophysiology and phenotypes of a new disease entity, The Lancet Respiratory Medicine 9 (6) (2021) 622–642. doi:10.1016/S2213-2600(21)00218-6.

[2] F. A. Klok, M. J. H. A. Kruip, N. J. M. van der Meer, M. S. Arbous, D. Gommers, K. M. Kant, F. H. J. Kaptein, J. van Paassen, M. A. M. Stals, M. V. Huisman, H. Endeman, Confirmation of the high cumulative incidence of thrombotic complications in critically ill ICU patients with COVID-19: An updated analysis, Thrombosis Research 191 (2020) 148–150. doi:10.1016/j.thromres.2020.04.041.

[3] F. Al-Ani, S. Chehade, A. Lazo-Langner, Thrombosis risk associated with COVID-19 infection. A scoping review, Thrombosis Research 192 (2020) 152–160. doi:10.1016/j.thromres.2020.05.039.

[4] S. Birocchi, M. Manzoni, G. M. Podda, G. Casazza, M. Cattaneo, High rates of pulmonary artery occlusions in COVID-19: A meta-analysis, European Journal of Clinical Investigation 51 (1) (2021) e13433. doi:10.1111/eci.13433.

[5] K. Boonyawat, P. Chantrathammachart, P. Numthavaj, N. Nanthatanti, S. Phusanti, A. Phuphuakrat, P. Niparuck, P. Angchaisuksiri, Incidence of thromboembolism in patients with COVID-19: A systematic review and meta-analysis, Thrombosis Journal 18 (34) (2020). doi:10.1186/s12959-020-00248-5.

[6] G. Chi, J. J. Lee, A. Jamil, V. Gunnam, H. Najafi, S. Memar Montazerin, F. Shojaei, J. Marszalek, Venous thromboembolism among hospitalized patients with COVID-19 undergoing thromboprophylaxis: A systematic review and meta-analysis, Journal of Clinical Medicine 9 (8) (2020) 2489. doi:10.3390/jcm9082489.

[7] A. Di Minno, P. Ambrosino, I. Calcaterra, M. N. D. Di Minno, COVID-19 and venous thromboembolism: A meta-analysis of literature studies, Seminars in Thrombosis and Hemostasis 46 (07) (2020) 763–771. doi:10.1055/s-0040-1715456.

[8] N. Gallastegui, J. Y. Zhou, A. von Drygalski, R. F. W. Barnes, T. M. Fernandes, T. A. Morris, Pulmonary embolism does not have an unusually high incidence among hospitalized COVID-19 patients, Clinical and Applied Thrombosis/Hemostasis 27 (2021). doi:10.1177/1076029621996471.

[9] M. Giustozzi, M. C. Vedovati, G. Agnelli, Venous thromboembolism and COVID-19: Mind the gap between clinical epidemiology and patient management, European Journal of Internal Medicine 82 (2020) 18–20. doi:10.1016/j.ejim.2020.10.018.

[10] J. Gratz, M. Wiegele, M. Maleczek, H. Herkner, H. Schöchl, E. Chwala, P. Knöbl, E. Schaden, Risk of clinically relevant venous thromboembolism in critically ill patients with COVID-19: A systematic review and meta-analysis, Frontiers in Medicine 8 (647917) (2021). doi:10.3389/fmed.2021.647917.

[11] S. S. Hasan, S. Radford, C. S. Kow, S. T. R. Zaidi, Venous thromboembolism in critically ill COVID-19 patients receiving prophylactic or therapeutic anticoagulation: A systematic review and meta-analysis, Journal of Thrombosis and Thrombolysis 50 (4) (2020) 814–821. doi:10.1007/s11239-020-02235-z.

[12] D. Jiménez, A. García-Sanchez, P. Rali, A. Muriel, B. Bikdeli, P. Ruiz-Artacho, R. Le Mao, C. Rodríguez, B. J. Hunt, M. Monreal, Incidence of VTE and bleeding among hospitalized patients with coronavirus disease 2019, Chest 159 (3) (2021) 1182–1196. doi:10.1016/j.chest.2020.11.005.

[13] A. Kollias, K. G. Kyriakoulis, S. Lagou, E. Kontopantelis, G. S. Stergiou, K. Syrigos, Venous thromboembolism in COVID-19: A systematic review and meta-analysis, Vascular Medicine 26 (4) (2021) 415–425. doi:10.1177/1358863X21995566.

[14] S. K. Kunutsor, J. A. Laukkanen, Incidence of venous and arterial thromboembolic complications in COVID-19: A systematic review and meta-analysis, Thrombosis Research 196 (2020) 27–30. doi:10.1016/j.thromres.2020.08.022.

[15] Y. Liu, J. Cai, C. Wang, J. Jin, L. Qu, Incidence, prognosis, and laboratory indicators of venous thromboembolism in hospitalized patients with coronavirus disease 2019: A systematic review and meta-analysis, Journal of Vascular Surgery: Venous and Lymphatic Disorders 9 (5) (2021) 1099–1111.E6. doi:10.1016/j.jvsv.2021.01.012.

[16] G. Longchamp, S. Manzocchi-Besson, A. Longchamp, M. Righini, H. Robert-Ebadi, M. Blondon, Proximal deep vein thrombosis and pulmonary embolism in COVID-19 patients: A systematic review and meta-analysis, Thrombosis Journal 19 (15) (2021). doi:10.1186/s12959-021-00266-x.

[17] Y. Lu, L. Pan, W. Zhang, F. Cheng, S. Hu, X. Zhang, H. Jiang, A meta-analysis of the incidence of venous thromboembolic events and impact of anticoagulation on mortality in patients with COVID-19, International Journal of Infectious Diseases 100 (2020) 34–41. doi:10.1016/j.ijid.2020.08.023.

[18] M. B. Malas, I. N. Naazie, N. Elsayed, A. Mathlouthi, R. Marmor, B. Clary, Thromboembolism risk of COVID-19 is high and associated with a higher risk of mortality: A systematic review and meta-analysis, EClinicalMedicine 29-30 (2020) 100639. doi:10.1016/j.eclinm.2020.100639.

[19] M. F. H. Mohamed, S. D. Al-Shokri, K. M. Shunnar, S. F. Mohamed, M. S. Najim, S. I. Ibrahim, H. Elewa, L. O. Abdalla, A. El-Bardissy, M. N. Elshafei, I. Y. Abubeker, M. Danjuma, K. M. Dousa, M. A. Yassin, Prevalence of venous thromboembolism in critically ill COVID-19 patients: Systematic review and meta-analysis, Frontiers in Cardiovascular Medicine 7 (598846) (2021). doi:10.3389/fcvm.2020.598846.

[20] S. Nopp, F. Moik, B. Jilma, I. Pabinger, C. Ay, Risk of venous thromboembolism in patients with COVID-19: A systematic review and meta-analysis, Research and Practice in Thrombosis and Haemostasis 4 (7) (2020) 1178–1191. doi:10.1002/rth2.12439.

[21] A. Porfidia, E. Valeriani, R. Pola, E. Porreca, A. W. Rutjes, M. Di Nisio, Venous thromboembolism in patients with COVID-19: Systematic review and meta-analysis, Thrombosis Research 196 (2020) 67–74. doi:10.1016/j.thromres.2020.08.020.

[22] G. K. Sridharan, R. Vegunta, V. R. P. Rokkam, V. Meyyur Aravamudan, R. Vegunta, S. R. Khan, S. Ponnada, U. Boregowda, K. Prudhvi, G. Chamarthi, B. P. Mohan, Venous thromboembolism in hospitalized COVID-19 patients, American Journal of Therapeutics 27 (6) (2020) e599–e610. doi:10.1097/MJT.0000000000001295.

[23] B. K. Tan, S. Mainbourg, A. Friggeri, L. Bertoletti, M. Douplat, Y. Dargaud, C. Grange, H. Lobbes, S. Provencher, J.-C. Lega, Arterial and venous thromboembolism in COVID-19: A study-level meta-analysis, Thorax 76 (10) (2021) 970–979. doi:10.1136/thoraxjnl-2020-215383.

[24] C. Wu, Y. Liu, X. Cai, W. Zhang, Y. Li, C. Fu, Prevalence of venous thromboembolism in critically ill patients with coronavirus disease 2019: A meta-analysis, Frontiers in Medicine 8 (603558) (2021). doi:10.3389/fmed.2021.603558.

[25] T. Wu, Z. Zuo, D. Yang, X. Luo, L. Jiang, Z. Xia, X. Xiao, J. Liu, M. Ye, M. Deng, Venous thromboembolic events in patients with COVID-19: A systematic review and meta-analysis, Age and Ageing 50 (2) (2021) 284–293. doi:10.1093/ageing/afaa259.

[26] C. Zhang, L. Shen, K.-J. Le, M.-M. Pan, L.-C. Kong, Z.-C. Gu, H. Xu, Z. Zhang, W.-H. Ge, H.-W. Lin, Incidence of venous thromboembolism in hospitalized coronavirus disease 2019 patients: A systematic review and meta-analysis, Frontiers in Cardiovascular Medicine 7 (151) (2020). doi:10.3389/fcvm.2020.00151.

[27] R. Zhang, L. Ni, X. Di, X. Wang, B. Ma, S. Niu, C. Liu, Systematic review and meta-analysis of the prevalence of venous thromboembolic events in novel coronavirus disease-2019 patients, Journal of Vascular Surgery: Venous and Lymphatic Disorders 9 (2) (2021) 289–298.e5. doi:10.1016/j.jvsv.2020.11.023.

[28] M. Cushman, Epidemiology and risk factors for venous thrombosis, Seminars in Hematology 44 (2) (2007) 62–69. doi:10.1053/j.seminhematol.2007.02.004.

[29] S. Liao, T. Woulfe, S. Hyder, E. Merriman, D. Simpson, S. Chunilal, Incidence of venous thromboembolism in different ethnic groups: A regional direct comparison study, Journal of Thrombosis and Haemostasis 12 (2) (2014) 214–219. doi:10.1111/jth.12464.

[30] F. Mei, J. Fan, J. Yuan, Z. Liang, K. Wang, J. Sun, W. Guan, M. Huang, Y. Li, W. W. Zhang, Comparison of venous thromboembolism risks between COVID-19 pneumonia and community-acquired pneumonia patients, Arteriosclerosis, Thrombosis, and Vascular Biology 40 (9) (2020) 2332–2337. doi:10.1161/ATVBAHA.120.314779.

[31] S. Middeldorp, M. Coppens, T. F. Haaps, M. Foppen, A. P. Vlaar, M. C. A. Müller, C. C. S. Bouman, L. F. M. Beenen, R. S. Kootte, J. Heijmans, L. P. Smits, P. I. Bonta, N. Es, Incidence of venous thromboembolism in hospitalized patients with COVID-19, Journal of Thrombosis and Haemostasis 18 (8) (2020) 1995–2002. doi:10.1111/jth.14888.

[32] C. Minet, L. Potton, A. Bonadona, R. Hamidfar-Roy, C. A. Somohano, M. Lugosi, J.-C. Cartier, G. Ferretti, C. Schwebel, J.-F. Timsit, Venous thromboembolism in the ICU: main characteristics, diagnosis and thromboprophylaxis, Critical Care 19 (287) (2015) 287. doi:10.1186/s13054-015-1003-9.

[33] H. Al-Samkari, R. S. Karp Leaf, W. H. Dzik, J. C. T. Carlson, A. E. Fogerty, A. Waheed, K. Goodarzi, P. K. Bendapudi, L. Bornikova, S. Gupta, D. E. Leaf, D. J. Kuter, R. P. Rosovsky, COVID-19 and coagulation: Bleeding and thrombotic manifestations of SARS-CoV-2 infection, Blood 136 (4) (2020) 489–500. doi:10.1182/blood.2020006520.

[34] J. Bauer, D. Brüggmann, D. Klingelhöfer, W. Maier, L. Schwettmann, D. J. Weiss, D. A. Groneberg, Access to intensive care in 14 European countries: A spatial analysis of intensive care need and capacity in the light of COVID-19, Intensive Care Medicine 46 (11) (2020) 2026–2034. doi:10.1007/s00134-020-06229-6.

[35] J. M. Connors, M. M. Brooks, F. C. Sciurba, J. A. Krishnan, J. R. Bledsoe, A. Kindzelski, A. L. Baucom, B.-A. Kirwan, H. Eng, D. Martin, E. Zaharris, B. Everett, L. Castro, N. L. Shapiro, J. Y. Lin, P. C. Hou, C. J. Pepine, E. Handberg, D. O. Haight, J. W. Wilson, S. Majercik, Z. Fu, Y. Zhong, V. Venugopal, S. Beach, S. Wisniewski, P. M. Ridker, S. Brakenridge, E. Leifer, A. Troxel, A. Maggioni, J. Jacobson, R. D. Lopes, R. Mentz, M. Sholzberg, J. H. Alexander, D. Schreiber, K. Yadav, A. Vecchiarelli, G. Oguchi, L. Merck, V. Altagracia, A. Gonzalez, D. Angiolillo, F. Blind, J. Cienki, R. Loy, E. Armas, J. Martinez, J. Ruiz Unger, V. Gulati, T. Robinson, C. Kroker-Bode, J. Wilson, D. Beiser, M. Hagan, J. Cohen, S. Goonewardena, J. Fletcher, R. Dolor, T. Jarrett, V. Patel, M.-L. Wang, N. Acquisto, A. Mehrle, K. Saangeeta, N. Hanna, H. Abouhouli, A. Weissman, R. Purighalla, N. Bennett, G. Jay, D. Barbham, T. Milling, D. Abdelsayed, D. McPherson, P. Salvato, C. Orgwu, A. Mundluru, P. Olusola, S. Majercik, W. Lewis, Effect of antithrombotic therapy on clinical outcomes in outpatients with clinically stable symptomatic COVID-19, JAMA 326 (17) (2021) 1703. doi:10.1001/jama.2021.17272.

[36] H. Zhan, H. Chen, C. Liu, L. Cheng, S. Yan, H. Li, Y. Li, Diagnostic value of D-Dimer in COVID-19: A meta-analysis and meta-regression, Clinical and Applied Thrombosis/Hemostasis 27 (2021). doi:10.1177/10760296211010976.

[37] P. Demelo-Rodríguez, E. Cervilla-Muñoz, L. Ordieres-Ortega, A. Parra-Virto, M. Toledano-Macías, N. Toledo-Samaniego, A. García-García, I. García-Fernández-Bravo, Z. Ji, J. De-Miguel-Diez, L. Álvarez-Sala-Walther, J. Del-Toro-Cervera, F. Galeano-Valle, Incidence of asymptomatic deep vein thrombosis in patients with COVID-19 pneumonia and elevated D-dimer levels, Thrombosis Research 192 (2020) 23–26. doi:10.1016/j.thromres.2020.05.018.

[38] M. B. Whyte, P. A. Kelly, E. Gonzalez, R. Arya, L. N. Roberts, Pulmonary embolism in hospitalised patients with COVID-19, Thrombosis Research 195 (2020) 95–99. doi:10.1016/j.thromres.2020.07.025.

[39] J. B. Hill, D. Garcia, M. Crowther, B. Savage, S. Peress, K. Chang, S. Deitelzweig, Frequency of venous thromboembolism in 6513 patients with COVID-19: A retrospective study, Blood Advances 4 (21) (2020) 5373–5377. doi:10.1182/bloodadvances.2020003083.

[40] A. Trimaille, A. Curtiaud, B. Marchandot, K. Matsushita, C. Sato, I. Leonard-Lorant, L. Sattler, L. Grunebaum, M. Ohana, J.-J. Von Hunolstein, E. Andres, B. Goichot, F. Danion, C. Kaeuffer, V. Poindron, P. Ohlmann, L. Jesel, O. Morel, Venous thromboembolism in non-critically ill patients with COVID-19 infection, Thrombosis Research 193 (2020) 166–169. doi:10.1016/j.thromres.2020.07.033.

[41] M. Cattaneo, E. M. Bertinato, S. Birocchi, C. Brizio, D. Malavolta, M. Manzoni, G. Muscarella, M. Orlandi, Pulmonary embolism or pulmonary thrombosis in COVID-19? Is the recommendation to use high-dose heparin for thromboprophylaxis justified?, Thrombosis and Haemostasis 120 (08) (2020) 1230–1232. doi:10.1055/s-0040-1712097.

[42] L. Zhang, X. Feng, D. Zhang, C. Jiang, H. Mei, J. Wang, C. Zhang, H. Li, X. Xia, S. Kong, J. Liao, H. Jia, X. Pang, Y. Song, Y. Tian, B. Wang, C. Wu, H. Yuan, Y. Zhang, Y. Li, W. Sun, Y. Zhang, S. Zhu, S. Wang, Y. Xie, S. Ge, L. Zhang, Y. Hu, M. Xie, Deep vein thrombosis in hospitalized patients with COVID-19 in Wuhan, China, Circulation 142 (2) (2020) 114–128. doi:10.1161/CIRCULATIONAHA.120.046702.

[43] The RECOVERY Collaborative Group, Dexamethasone in hospitalized patients with COVID-19, New England Journal of Medicine 384 (8) (2021) 693–704. doi:10.1056/NEJMoa2021436.

[44] A. Agarwal, B. Rochwerg, R. A. Siemieniuk, T. Agoritsas, F. Lamontagne, L. Askie, L. Lytvyn, Y.-S. Leo, H. Macdonald, L. Zeng, W. Amin, E. Burhan, F. J. Bausch, C. S. Calfee, M. Cecconi, D. Chanda, B. Du, H. Geduld, P. Gee, N. Harley, M. Hashimi, B. Hunt, S. K. Kabra, S. Kanda, Y. Kim, N. Kissoon, A. Kwizera, I. Mahaka, H. Manai, G. Mino, E. Nsutebu, J. Preller, N. Pshenichnaya, N. Qadir, S. Sabzwari, R. Sarin, M. Shankar-Hari, M. Sharland, Y. Shen, S. S. Ranganathan, J. P. Souza, M. Stegemann, A. De Sutter, S. Ugarte, S. Venkatapuram, V. Q. Dat, D. Vuyiseka, A. Wijewickrama, B. Maguire, D. Zeraatkar, J. J. Bartoszko, L. Ge, R. Brignardello-Petersen, A. Owen, G. Guyatt, J. Diaz, L. Kawano-Dourado, M. Jacobs, P. O. Vandvik, A living WHO guideline on drugs for COVID-19, BMJ 370 (2020) m3379. doi:10.1136/bmj.m3379.

[45] WHO, Therapeutics and COVID-19: Living guideline (2021).
URL https://www.who.int/publications/i/item/WHO-2019-nCoV-therapeutics-2021.2

[46] A. Sarfraz, Z. Sarfraz, A. A. Razzack, G. Patel, M. Sarfraz, Venous thromboembolism, corticosteroids and COVID-19: A systematic review and meta-analysis, Clinical and Applied Thrombosis/Hemostasis 27 (2021). doi:10.1177/1076029621993573.

[47] S. Voicu, P. Bonnin, A. Stépanian, B. G. Chousterman, A. Le Gall, I. Malissin, N. Deye, V. Siguret, A. Mebazaa, B. Mégarbane, High prevalence of deep vein thrombosis in mechanically ventilated COVID-19 patients, Journal of the American College of Cardiology 76 (4) (2020) 480–482. doi:10.1016/j.jacc.2020.05.053.

[48] M. A. M. Stals, F. H. J. Kaptein, R. H. H. Bemelmans, T. van Bemmel, I. C. Boukema, D. C. W. Braeken, S. J. E. Braken, C. Bresser, H. ten Cate, D. D. Deenstra, Y. P. A. van Dooren, L. M. Faber, M. J. J. H. Grootenboers, L. R. de Haan, C. Haazer, A. I. del Sol, S. Kelliher, T. Koster, L. J. M. Kroft, R. I. Meijer, F. Pals, E. R. E. van Thiel, P. E. Westerweel, M. ten Wolde, F. A. Klok, M. V. Huisman, Ruling out pulmonary embolism in patients with (suspected) COVID-19—A prospective cohort study, TH Open 05 (03) (2021) e387–e399. doi:10.1055/s-0041-1735155.

[49] I. Koleilat, B. Galen, K. Choinski, A. N. Hatch, D. B. Jones, H. Billett, J. Indes, E. Lipsitz, Clinical characteristics of acute lower extremity deep venous thrombosis diagnosed by duplex in patients hospitalized for coronavirus disease 2019, Journal of Vascular Surgery: Venous and Lymphatic Disorders 9 (1) (2021) 36–46. doi:10.1016/j.jvsv.2020.06.012.

[50] B. Ren, F. Yan, Z. Deng, S. Zhang, L. Xiao, M. Wu, L. Cai, Extremely high incidence of lower extremity deep venous thrombosis in 48 Patients with severe COVID-19 in Wuhan, Circulation 142 (2) (2020) 181–183. doi:10.1161/CIRCULATIONAHA.120.047407.

[51] D. Mazzaccaro, F. Giacomazzi, M. Giannetta, A. Varriale, R. Scaramuzzo, A. Modafferi, G. Malacrida, P. Righini, M. M. Marrocco-Trischitta, G. Nano, Non-overt coagulopathy in non-ICU patients with mild to moderate COVID-19 pneumonia, Journal of Clinical Medicine 9 (6) (2020) 1781. doi:10.3390/jcm9061781.

[52] M. Criel, M. Falter, J. Jaeken, M. Van Kerrebroeck, I. Lefere, L. Meylaerts, D. Mesotten, M. vander Laenen, T. Fivez, M. Thomeer, D. Ruttens, Venous thromboembolism in SARS-CoV-2 patients: Only a problem in ventilated ICU patients, or is there more to it?, European Respiratory Journal 56 (1) (2020). doi:10.1183/13993003.01201-2020.

[53] A. Santoliquido, A. Porfidia, A. Nesci, G. De Matteis, G. Marrone, E. Porceddu, G. Cammà, I. Giarretta, M. Fantoni, F. Landi, A. Gasbarrini, R. Pola, M. E. D'Alfonso, M. R. Lo Monaco, Incidence of deep vein thrombosis among non-ICU patients hospitalized for COVID-19 despite pharmacological thromboprophylaxis, Journal of Thrombosis and Haemostasis 18 (9) (2020) 2358–2363. doi:10.1111/jth.14992.

[54] A. Longchamp, J. Longchamp, S. Manzocchi-Besson, L. Whiting, C. Haller, S. Jeanneret, M. Godio, J. J. Garcia Martinez, T. Bonjour, M. Caillat, G. Maitre, J. M. Thaler, R. Pantet, V. Donner, A. Dumoulin, S. Emonet, G. Greub, R. Friolet, H. Robert-Ebadi, M. Righini, B. Sanchez, J. Delaloye, Venous thromboembolism in critically Ill patients with COVID-19: Results of a screening study for deep vein thrombosis, Research and Practice in Thrombosis and Haemostasis 4 (5) (2020) 842–847. doi:10.1002/rth2.12376.

[55] L. van Dam, L. Kroft, L. van der Wal, S. Cannegieter, J. Eikenboom, E. de Jonge, M. Huisman, F. Klok, Clinical and computed tomography characteristics of COVID-19 associated acute pulmonary embolism: A different phenotype of thrombotic disease?, Thrombosis Research 193 (2020) 86–89. doi:10.1016/j.thromres.2020.06.010.

[56] J. Chen, X. Wang, S. Zhang, B. Lin, X. Wu, Y. Wang, X. Wang, M. Yang, J. Sun, Y. Xie, Characteristics of acute pulmonary embolism in patients with COVID-19 associated pneumonia from the city of Wuhan, Clinical and Applied Thrombosis/Hemostasis 26 (2020). doi:10.1177/1076029620936772.

[57] M. Artifoni, G. Danic, G. Gautier, P. Gicquel, D. Boutoille, F. Raffi, A. Néel, R. Lecomte, Systematic assessment of venous thromboembolism in COVID-19 patients receiving thromboprophylaxis: incidence and role of D-dimer as predictive factors, Journal of Thrombosis and Thrombolysis 50 (1) (2020) 211–216. doi:10.1007/s11239-020-02146-z.

[58] K. J. Rothman, Epidemiology: an introduction, Oxford University Press, New York, NY, 2002.

[59] H. Morgenstern, D. G. Kleinbaum, L. L. Kupper, Measures of disease incidence used in epidemiologic research, International Journal of Epidemiology 9 (1) (1980) 97–104. doi:10.1093/ije/9.1.97.

[60] M. J. Desborough, A. J. Doyle, A. Griffiths, A. Retter, K. A. Breen, B. J. Hunt, Image-proven thromboembolism in patients with severe COVID-19 in a tertiary critical care unit in the United Kingdom, Thrombosis Research 193 (2020) 1–4. doi:10.1016/j.thromres.2020.05.049.

[61] G. Le Gal, M. Carrier, L. A. Castellucci, A. Cuker, J. Hansen, F. A. Klok, N. J. Langlois, J. H. Levy, S. Middeldorp, M. Righini, S. Walters, E. Klok, L. Bauman Kreuziger, S. Schulman, L. Skeith, N. Zakai, N. Riva, J. Douxfils, S. Kahn, A. Ten Cate-Hoek, S. Konstantinides, W. Ghanima, I. Lang, J. Galanaud, T. Moumneh, E. Gandara, P. Prandoni, C. Witmer, A. Lazo-Langner, H. Robert-Ebadi, F. Ní Áinle, C. Wu, T. Wang, J. Zwicker, M. Lauw, C. Ay, G. C. Maus, P. Angchaisuksiri, M. Steiner, R. Bartz, J. Connors, M. Samama, A. Spyropoulos, D. Faraoni, T. Iba, C. Kearon, S. Canniegeter, P. Morange, S. Brækkan, V. Morelli, F. Orsi, Development and implementation of common data elements for venous thromboembolism research: On behalf of SSC Subcommittee on official Communication from the SSC of the ISTH, Journal of Thrombosis and Haemostasis 19 (1) (2021) 297–303. doi:10.1111/jth.15138.

[62] M. Carrier, A. A. Khorana, J. I. Zwicker, G. H. Lyman, G. Le Gal, A. Y. Y. Lee, Venous thromboembolism in cancer clinical trials: Recommendation for standardized reporting and analysis, Journal of Thrombosis and Haemostasis 10 (12) (2012) 2599–2601. doi:10.1111/jth.12028.

7

Sensitivity analysis for random measurement error using regression calibration and simulation-extrapolation

Sensitivity analysis for random measurement error can be applied in the absence of validation data by means of regression calibration and simulation-extrapolation. These have not been compared for this purpose. A simulation study was conducted comparing the performance of regression calibration and simulation-extrapolation for linear and logistic regression. The performance of the two methods was evaluated in terms of bias, mean squared error (MSE) and confidence interval coverage, for various values of reliability of the error-prone measurement (0.05–0.91), sample size (125–4,000), number of replicates (2–10), and R-squared (0.03–0.75). It was assumed that no validation data were available about the error-free measures, while correct information about the measurement error variance was available. Regression calibration was unbiased while simulation-extrapolation was biased: median bias was 0.8% (interquartile range (IQR): −0.6;1.7%), and −19.0% (IQR: −46.4;−12.4%), respectively. A small gain in efficiency was observed for simulation-extrapolation (median MSE: 0.005, IQR: 0.004;0.006) versus regression calibration (median MSE: 0.006, IQR: 0.005;0.009). Confidence interval coverage was at the nominal level of 95% for regression calibration, and smaller than 95% for simulation-extrapolation (median coverage: 85%, IQR: 73;93%). The application of regression calibration and simulation-extrapolation for a sensitivity analysis was illustrated using an example of blood pressure and kidney function. Our results support the use of regression calibration over simulation-extrapolation for sensitivity analysis for random measurement error.

This chapter is based on: L. Nab and R.H.H. Groenwold, Sensitivity analysis for random measurement error using regression calibration and simulation-extrapolation, Global Epidemiology 3 (2021) 100067. doi:10.1016/j.gloepi.2021.100067

7.1. Introduction

Measurement error is common in biomedical research but often ignored [1, 2]. When ignored, measurement error can lead to considerable biases in exposure-outcome associations [3]. Random measurement error in the exposure variable, also known as 'classical' measurement error, occurs when the measured exposure is distributed around the true exposure with independent error and, is common in various domains of epidemiology [4–6]. Random measurement error in an exposure variable introduces bias in the exposure-outcome association, which is sometimes referred to as attenuation bias [7] or regression dilution bias [8, 9].

Various methods for measurement error correction are available [10–18], yet application of these methods is rare in biomedical research [3]. One possible barrier is the necessity of for instance validation data, which are often unavailable [5]. Validation data can be used to estimate the measurement error model and its parameters, and subsequently used for measurement error correction.

In the absence of validation data, regression calibration [19, 20] and simulation-extrapolation [21], among other, can be applied to correct for random exposure measurement error. Both methods only require assumptions about the variance of the random measurement error, for example based on literature or expert knowledge. Regression calibration in the absence of validation data is available in the R [22] package `mecor` for measurement error correction [23], that implements the regression calibration described by Rosner et al. [24]. Alternatively, simulation-extrapolation is easy to use due to its implementation in the R package `simex` [25] and the `simex` procedure [26] in Stata [27].

Simulation-extrapolation and regression calibration have been compared in simulation studies for scenarios where replicate measures of the error-prone exposure were available [5, 28, 29]. The studies by Perrier et al. [5], Batistatou et al. [28] and Fung et al. [29] were consistent and showed that, regression calibration and simulation-extrapolation reduced bias compared to when no measurement error correction was applied or when the replicate exposure measures were pooled. It was also shown that application of simulation-extrapolation generally produced more biased effect estimates than regression calibration, especially when the reliability of the error-prone measure was low.

Perrier et al. [5] and Batistatou et al. [28] studied a univariable linear regression in a limited number of scenarios, e.g., large sample sizes, and limited range of reliability of the error-prone measure. Fung et al. [29] studied a multivariable logistic regression in a limited number of scenarios, e.g., not varying sample size and a limited range of reliability. Further investigation is needed of the performance of regression calibration and simulation-extrapolation in more complex settings, as typically found in epidemiologic research (e.g., multivariable linear and logistic regression, varying sample size and levels of reliability). Moreover, since the previous simulation studies focused on settings where replicate measures were available, we aim to research how their results translate to settings where no replicate measures, but only an estimate of the measurement error variance is available. The quantification of the performance of the two methods in this broader range of settings is used as the input for a framework guiding the application of sensitivity analysis for random measurement error.

This chapter is structured as follows. Section 7.2 reviews and applies regression calibration and simulation-extrapolation by using two motivating examples of a

linear regression and logistic regression where the exposure is prone to error. In section 7.3, a simulation study is described comparing regression calibration and simulation-extrapolation for linear regression and logistic regression, and results from the simulation study are shown. Section 7.4 introduces a framework for conducting sensitivity analysis, also known as quantitative bias analysis [30], for random measurement error by means of regression calibration and simulation-extrapolation. We conclude with a discussion of our results and recommendations in section 7.5.

7.2. Review and motivating example

When an exposure variable is measured with random measurement error, the exposure-outcome association is biased. In a univariable model with a continuous outcome, under the assumption of random measurement error, the uncorrected effect estimate is biased by a factor equal to the variance of the true measure divided by the sum of the variance of the true measure and the measurement error variance. This is sometimes referred to as the 'attenuation factor' because the variance of the true measure plus the measurement error variance is always greater than the variance of the true measure alone [7]. For a linear regression, this can be expanded to the multivariable case by conditioning on the covariates in the multivariable model. For a logistic regression, the bias induced by random measurement error cannot be quantified exactly [13]. Kuha [31] shows that under the assumption that the effect of the exposure on the outcome is 'small to moderate' and/or the measurement error is 'small', the uncorrected effect estimate in a logistic regression is biased approximately by the attenuation factor. We refer to Kuha for a detailed discussion.

7.2.1. Review of regression calibration and simulation-extrapolation

In a linear regression or logistic regression, the random measurement error in an exposure can be corrected by application of regression calibration and simulation-extrapolation. Regression calibration starts by estimating the uncorrected effect of the error-prone measure on the outcome (in a multivariable model, given the covariates). Subsequently, the uncorrected estimate is multiplied by the inverse of the attenuation factor: the estimated variance of the error-prone exposure (given the covariates), divided by the estimated variance of the error-prone exposure (given the covariates) minus the measurement error variance. The measurement error variance can be estimated by using e.g., replicate measurements, by estimating the within individual variance and averaging over all individuals. Alternatively, the measurement error variance could be informed by e.g., external data or expert knowledge.

Simulation-extrapolation consists of two steps. In the simulation step, extra measurement error is added to the error-prone exposure. The size of this extra measurement error is typically 0.5, 1, 1.5 and 2 times the measurement error variance. Using these simulated exposure measurements with extra added measurement error, the outcome model is estimated. This is repeated 100 times for each value of the extra measurement error variance and the newly obtained estimates are averaged. Then, in the extrapolation step, a model (e.g., linear, quadratic) is fitted through the effect estimates for the varying sizes of the measurement error. The corrected effect estimate is then obtained by extrapolating the fitted model to the situation where the measurement error is equal to 0. For a visualisation of simulation-extrapolation, see e.g. Keogh et al. [3].

7.2.2. Motivating example

Hereafter, regression calibration and simulation-extrapolation to correct for random measurement error are demonstrated for linear regression and logistic regression, using an example about the association between systolic blood pressure and kidney function (serum creatinine) and an example about the association between sodium intake and hypertension, respectively.

Example 1: Linear regression of blood pressure and kidney function in pregnant women

For the first example, we used data of retrospective records of all women who attended a tertiary maternity hospital pregnancy day assessment clinic over a 6-month period in 2014 in Australia [32]. Care always included serial, manual blood pressure measurements every 30 min by registered midwives using aneroid sphygmomanometers [32]. Serum creatinine and demographic data were obtained using routinely collected data. One woman with a serum creatinine level lower than 10 μmol/L was excluded from the analysis.

First, the association between systolic blood pressure and serum creatinine was determined by only using the systolic blood pressure measurement obtained after 30 min. The association was adjusted for age. We found that an increase of 10 mmHg in systolic blood pressure was associated with a 1.18 μmol/L (CI: 0.14 - 2.23) increase in serum creatinine (Table 7.1). In this analysis, the random measurement error in the single systolic blood pressure measurement was not taken into account. Using the four consecutive blood pressure measurements (obtained after 30, 60, 90 and 120 minutes), it was found that the within individual variance of the systolic blood pressure measures was on average 48.3 mmHg, equal to a reliability of 0.6. The within individual variance of 48.3 mmHg was subsequently used to correct for the random measurement error in the single systolic blood pressure measurement using regression calibration and simulation-extrapolation, while adjusting for age. Using regression calibration, we found that an increase of 10 mmHg was associated with a 2.04 μmol/L (CI: 0.22;4.23) increase in serum creatinine (Table 7.1). Using simulation-extrapolation, we found that an increase of 10 mmHg was associated with a 1.67 μmol/L (CI: 0.15;3.29) increase in serum creatinine (Table 7.1).

Example 2: Logistic regression of sodium intake and hypertension in adults

For the second example, we used data of the 2015-2016 cycle of the National Health And Nutrition Examination Survey [33]. Given natural variation of sodium intake within individuals, a single measurement of sodium intake often does not reflect the true level of sodium intake. In the NHANES, two sodium intake measurements were taken using a 24-hour recall. The first dietary recall interview was collected in-person and the second interview was collected by telephone 3 to 10 days later. Participants' hypertension status was based on a combination of their self-reported history of any diagnosis of hypertension and self-reported use of prescribed hypertension medication. Demographic information was collected using the family and sample person demographics questionnaires in the home, by trained interviewers. Weight and height were measured by trained health professionals. For this analysis, participants between 18-80 years were included. Additionally, all participants with a body mass index (BMI) higher than 55 and a sodium intake of more than 10 gram per day were excluded from the analysis.

First, the association between sodium intake and hypertension was determined by only using the first sodium intake measurement. The association was adjusted for BMI and age. It was found that an increase of 1 gram in sodium intake was associated with

a 1.04 (95% CI: 1.00;1.09) times increase in the odds for hypertension. In the NHANES data, the within individual variation of sodium intake was on average 1.7 gram, which was obtained by using the two consecutive sodium intake measures, resulting in a reliability of 0.4. The within individual variance of 1.7 was subsequently used to correct for the random measurement error in the first sodium intake measure using regression calibration and simulation-extrapolation, while adjusting for age. Using regression calibration, we found that an increase of 1 gram in sodium intake was associated with a 1.12 (95% CI: 0.99;1.27) increase in the odds for hypertension (Table 7.1). Using simulation-extrapolation, we found that an increase of 1 gram in sodium intake was associated with a 1.07 (95% CI: 1.00;1.16) increase in the odds for hypertension (Table 7.1).

Table 7.1: Effect estimates (95% confidence intervals) of the association between blood pressure (systolic blood pressure, per 10 mmHg) and kidney function (serum creatinine, μmol/L) (linear regression, example 1) and the association between sodium intake (per gram) and hypertension (odds ratio obtained from a logistic regression, example 2). The uncorrected effect estimates are obtained by using the first measurement only, the corrected estimates are obtained by using the three consecutive blood pressure measurements (example 1) and the second consecutive sodium intake measurement (example 2).

Example	Uncorrected	Regression Calibration	Simulation-extrapolation
Systolic blood pressure and kidney function	1.18[b](0.14;2.23)	2.04 (0.22;4.23)	1.67 (0.15;3.29)
Sodium intake and hypertension[a]	1.04[c](1.00;1.09)	1.12 (0.99;1.27)	1.07 (1.00;1.16)

Estimates were obtained from the pregnancy day and assessment clinic study (systolic blood pressure and kidney function, reliability of the error-prone blood pressure measurement: 0.6) [32] and the national health and nutrition examination survey (sodium intake and hypertension, reliability of the error-prone sodium intake measurement: 0.4) [33]

[a] Odds ratio

[b] Estimate is corrected for age, but not for the measurement error in systolic blood pressure

[c] Estimate is corrected for age and body mass index, but not for the measurement error in sodium intake

7.3. Simulation study

To investigate the observed difference between the regression calibration corrected and simulation-extrapolation corrected analysis in our motivating examples above, a simulation study was conducted to study the relative performance of regression calibration and simulation-extrapolation in a linear regression model and a logistic regression model. The relative performance was studied in terms of bias, mean squared error, and confidence interval coverage of the true effect. Subsection Methods provides a general description and motivation of the scenarios studied, and an explanation of the specific parameters set in our simulation study. Subsection Results presents the results of our simulation study.

Table 7.2: Simulation study settings linear regression

Scenarios	Parameters of Data Generating Mechanism[a]				
	τ^2	n	k	ϕ	γ
Base	30	500	3	100	0
Reliability[b]	200; 100; 50; 25; 20; 15; 10; 5	500	3	100	0
Sample Size	30	125; 250; 1,000; 10,000	3	100	0
Number of	30	500	2; 5; 10	100	0
R-squared[c]	30	500	3	20; 5; 1	0
Covariate Dependency[d]	30	500	3	100	1; 4; 8

[a] τ^2: measurement error variance of the error-prone blood pressure measurement; n: number of observations in the main study; k: number of replicate error-prone measurements; σ^2: residual variance of the outcome model; γ: association between blood pressure and age. The attenuation in the effect of blood pressure on creatinine due to random measurement error is equal to $50/(50 + \tau^2)$
[b] Reliability is equal to $(25\gamma^2 + 50)/(25\gamma^2 + 50 + \tau^2)$
[c] R-squared is equal to $1 - \sigma^2/(0.4 \times 50 + 10 + \sigma^2)$
[d] The effect of blood pressure on creatinine when age is not included in the model (crude model) is equal to $0.2 + 5\gamma/(25\gamma^2 + 50)$

7.3.1. Methods

Linear regression. The relative performance of regression calibration versus simulation-extrapolation for linear regression was studied before by Perrier et al. [5] and Batistatou et al. [28]. We aimed to extend these two former simulation studies by investigating the relative performance in scenarios other than those studied before. Perrier et al. and Batistatou et al. assumed relatively large sample sizes (i.e., 3,000 and 1,000, respectively), only four different values for the reliability of the error-prone exposure (i.e., 0.2 and 0.6 in the study by Perrier et al. and 0.2, 0.5 and 0.8 in the study by Batistatou et al.) and a small coefficient of determination for the exposure-outcome model (i.e., 0.004 and 0.0625, respectively). In addition, Perrier et al. studied the effect of increasing the number of replicate measures on the performance of regression calibration and simulation-extrapolation by pooling the replicate measurements. Moreover, Perrier et al. and Batistatou et al. only examined models with a single independent variable. Therefore, our simulation study focused on multivariable models, small sample sizes (i.e., smaller or equal to 1,000) and relatively large reliability of the error-prone measurement (i.e., greater or equal to 0.625). In addition, the effect of a change in the coefficient of determination of the outcome model was tested. Furthermore, increasing the number of replicate measurements available was studied, without having the advantage of pooling the replicate measurements in our analysis.

Data generating mechanism linear regression. Inspired by our example of blood pressure and kidney function in pregnant women [32] described in section Motivating example, we assumed the following data generating mechanisms for age, blood pressure (BP), error-prone blood pressure (BP*) and creatinine:

$$\text{Age} \sim \mathcal{N}(32, 25), \quad \text{BP}|\text{Age} \sim \mathcal{N}(120 + \gamma\text{Age}, 50), \quad \text{BP}^*|\text{BP} \sim \mathcal{N}(\text{BP}, \tau^2),$$

$$\text{and Creatinine}|\text{BP, Age} \sim \mathcal{N}(30 + 0.2\text{BP} + 0.2\text{Age}, \sigma^2).$$

The above defined data generating mechanism defines that the error-prone blood pressure (BP*) has random measurement error with measurement error variance equal to τ^2. In our simulation study, a 'base scenario' was assumed and in the consecutive scenarios studied, we changed one of the three parameters in the data generating mechanisms (i.e., γ, τ^2 or σ^2), the number of observations (i.e., n), or the number of replicate measures (i.e., k) (see Table 7.2). For each scenario, 5,000 datasets were generated. The parameters settings of the base scenario were inspired by our example of blood pressure and kidney function in pregnant women [32]. In the base scenario, $n = 500$, $\gamma = 0$, $\tau^2 = 30$ and $\sigma^2 = 100$ (Table 7.2). We assumed that three replicate measures of the error-prone blood pressure measure were obtained in all individuals. From the parameter settings in the base scenario, it follows that the reliability of the error-prone measure is 0.625. Further, in the base scenario, the R-squared of the outcome model is 0.03, and the attenuation due to measurement error of the effect of the error-prone blood pressure on creatinine (given age) is equal to the reliability, i.e., 0.625.

In each generated data set, the uncorrected effect was estimated using the first replicate measurement only. Subsequently, the corrected effect was estimated by application of regression calibration and simulation-extrapolation using the R package `mecor` [23] and `simex` [25], respectively. The measurement error variance was estimated using the replicate measures. Ninety-five percent CI's of the corrected effects were constructed using bootstrap resampling. Performance of the three different analyses were evaluated in terms of bias, mean squared error (MSE), and the proportion of 95% CIs that contained the true value of the estimand (coverage). Monte Carlo standard errors (MCSE) were calculated for all performance measures [34], using the R package `rsimsum` [35]. All code used for the simulation study is publicly available via `https://github.com/LindaNab/simexvsmecor`.

Logistic regression. The relative performance of regression calibration and simulation-extrapolation for logistic regression was studied before by Fung et al. [29]. Fung et al. assumed a relatively small sample size (i.e., 500) and relatively high reliability (i.e., 0.6 and 0.7). In our simulation study, we focus on parameters identical to the parameters varied in linear regression: reliability, sample size, number of replicates, pseudo R-squared (Nagelkerke) and covariate dependency.

Data generating mechanism logistic regression. Inspired by our example of sodium intake and hypertension in adults [33] described in section Motivating example, we assume the following data generating mechanisms for age, sodium intake (Na), error-prone sodium intake (Na*) and hypertension:

$$\text{Age} \sim \mathcal{U}_{[18,80]}, \quad \text{Na}|\text{Age} \sim \mathcal{N}(4 + \gamma\text{Age}, 1), \quad \text{Na}^*|\text{Na} \sim \mathcal{N}(\text{Na}, \tau^2),$$
$$\text{and Hypertension}|\text{Na, Age} \sim \mathcal{B}(1, 1/(1 + e^{-p})), \text{ where } p = -7 + 0.1\text{Na} + \phi\text{Age}.$$

The above defined data generating mechanism defines that the error-prone sodium intake (Na*) has random measurement error with measurement error variance equal to τ^2. In our simulation study, a 'base scenario' was assumed and in the consecutive scenarios studied, we changed one of the three parameters in the data generating mechanisms (i.e., τ^2, ϕ or γ), the number of observations (i.e., n), or the number of replicate measures (i.e., k)

Table 7.3: Simulation study settings logistic regression

Scenarios	Parameters of Data Generating Mechanism[a]				
	τ^2	n	k	ϕ	γ
Base	2	4,000	2	0.1	0
Reliability[b]	20; 10; 4; 1.5; 1; 0.5; 0.25; 0.1	4,000	2	0.1	0
Sample Size	2	500; 1,000; 2,000; 10,000	2	0.1	0
Number of Replicates	2	4,000	3; 5; 10	0.1	0
Pseudo R-squared[c]	2	4,000	2	0.06; 0.08; 0.2	0
Covariate Dependency[d]	2	4,000	2	0.1	0.01; 0.1; 0.2

[a] τ^2: measurement error variance of the error-prone sodium intake measurement; n: number of observations in the main study; k: number of replicate error-prone measurements; ϕ: association between hypertension and age (given sodium intake); γ: association between sodium intake and age. The attenuation in the effect of sodium intake on hypertension due to random measurement error is equal to $1/(1+\tau^2)$

[b] Reliability is equal to $(\gamma^2(1/12)(80-18)^2+1)/(\gamma^2(1/12)(80-18)^2+1+\tau^2)$

[c] Computational calculations show Nagelkerke R-squared is equal to 0.1, 0.3 and 0.7 for ϕ equal to 0.06, 0.08 and 0.2, respectively. In the base scenario, Nagelkerke R-squared is equal to 0.4.

[d] Computational calculations show that the effect of sodium intake on hypertension when age is not included in the model (crude model) is equal to 0.3, 0.8 and 0.6 for γ equal to 0.01, 0.1 and 0.2, respectively. In the base scenario, the effect of sodium intake on hypertension in the crude model is 0.06. Changing γ affects Nagelkerke R-squared, for γ equal to 0.1 and 0.2, Nagelkerke R-squared is 0.5. For γ equal to 0.01, Nagelkerke R-squared is comparable to the base scenario (0.4).

(see Table 7.3). For each scenario, 5,000 datasets were generated. The parameters settings of the base scenario were inspired by our example of sodium intake and hypertension in adults [33]. In the base scenario, sample size was 4000, $\tau^2 = 2$, $\phi = 0.1$ and $\gamma = 0$ (Table 7.3). Furthermore, we assumed that two replicate measures of the error-prone sodium intake measure were obtained in all individuals. From the parameter settings in the base scenario, it follows that the reliability of the error-prone measure is 0.33. Further, in the base scenario, the Nagelkerke pseudo R-squared of the outcome model was 0.4, and the attenuation due to measurement error of the effect of the error-prone sodium intake measure on hypertension (given age) was *approximately* equal to the reliability, i.e., 0.33.

In each generated data set, the uncorrected effect was estimated using the first replicate measurement only. Subsequently, the corrected effect was estimated by application of regression calibration and simulation-extrapolation using regression calibration for logistic regression as described by Rosner et al. [24] and by use of the `R` package `simex` [25], respectively. The measurement error variance was estimated using the replicate measures. Ninety-five percent CI's of the corrected effects were constructed using bootstrap resampling. Performance of the three different analyses were evaluated in terms of bias, mean squared error (MSE), and the proportion of 95% CIs that contained the true value of the estimand (coverage). Monte Carlo standard errors (MCSE) were calculated for all performance measures [34], using the R package `rsimsum` [35]. All code used for the simulation study is publicly available via `www.github.com/LindaNab/simexvsmecor`.

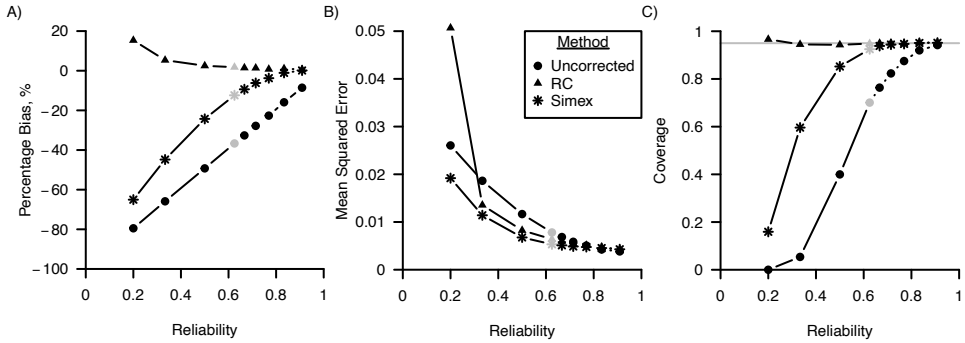


Figure 7.1: Performance in a linear regression model of regression calibration (RC), simulation-extrapolation (simex) and the analysis ignoring random measurement error for varying values of the reliability of the error-prone measure in terms of A) percentage bias; B) mean squared error and C) coverage. For all three performance measures, Monte Carlo standard errors were smaller than 0.01 in all scenarios. The grey points indicate the base scenario where reliability is assumed 0.625.

7.3.2. Results

Linear regression. Figure 7.1 shows the percentage bias, MSE and confidence interval coverage for varying values of the reliability of the error-prone measure. The uncorrected analysis was biased for all values of the reliability, and the percentage bias decreased when reliability increased. Regression calibration provided unbiased results when reliability was greater or equal to 0.33. Simulation-extrapolation provided biased results when reliability was smaller than 0.8. MSE was lower for simulation-extrapolation than for the uncorrected and regression calibration corrected analysis when reliability was equal to 0.2, and similar to MSE of regression calibration otherwise. Coverage of the 95% confidence intervals was at the nominal level for the regression calibration corrected analysis, and for the simulation-extrapolation corrected analysis when reliability was greater than or equal to 0.625.

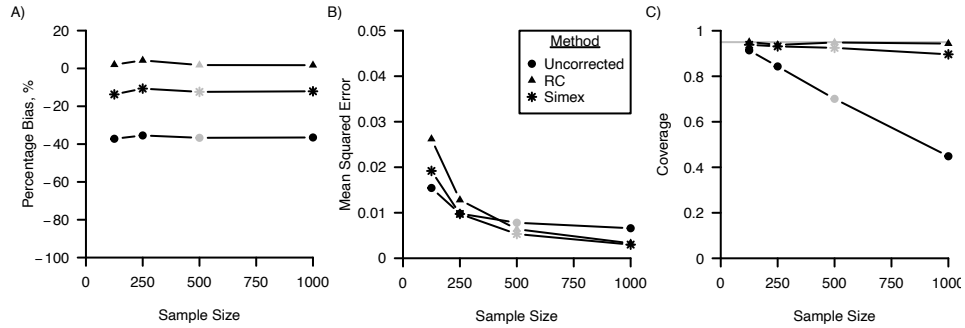


Figure 7.2: Performance in a linear regression model of regression calibration (RC), simulation-extrapolation (simex) and the analysis ignoring random measurement error for varying sample sizes of the error-prone measure in terms of A) percentage bias; B) mean squared error and C) coverage. For all three performance measures, Monte Carlo standard errors were < 0.01 in all scenarios. The grey points indicate the base scenario where sample size is assumed 500.

Figure 7.2 shows the percentage bias, MSE and confidence interval coverage for varying samples sizes of the main study. A sample size of 125, and 250 only increased percentage bias minimally compared to the base scenario were sample size was 500. MSE was greater for smaller sample sizes, and MSE of the uncorrected analysis with a sample size of 125 was smaller than the regression calibration and simulation-extrapolation corrected analysis (0.015 vs 0.026 and 0.019, respectively, MCSE < 0.005). Coverage was equal to the nominal level of 95% for regression calibration for all sample sizes, and the uncorrected analysis showed coverage levels that were subnominal, ranging between 45% and 91% (MCSE < 0.01). Coverage of the 95% confidence intervals of the simulation-extrapolation corrected analysis was close to the nominal level of 95% except when sample size was 1,000, in which case coverage was 90% (MCSE 0.004). A decline in confidence interval coverage for the simulation-extrapolation corrected analysis for larger sample sizes was confirmed by the scenario where sample size was 10,000, in which case coverage was 53% (MCSE 0.007)(not shown in the plots in Figure 7.2).

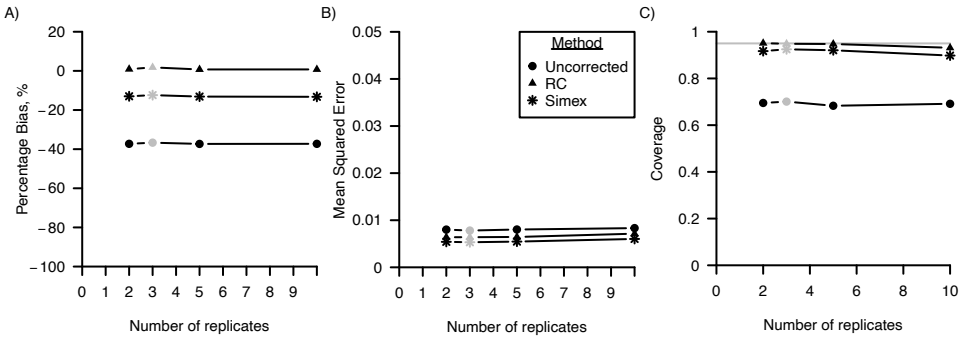


Figure 7.3: Performance in a linear regression model of regression calibration (RC), simulation-extrapolation (simex) and the analysis ignoring random measurement error for varying number of replicates of the error-prone measure in terms of A) percentage bias; B) mean squared error and C) coverage. For all three performance measures, Monte Carlo standard errors were smaller than 0.01 in all scenarios. The grey points indicate the base scenario where the number of replicates is assumed 3.

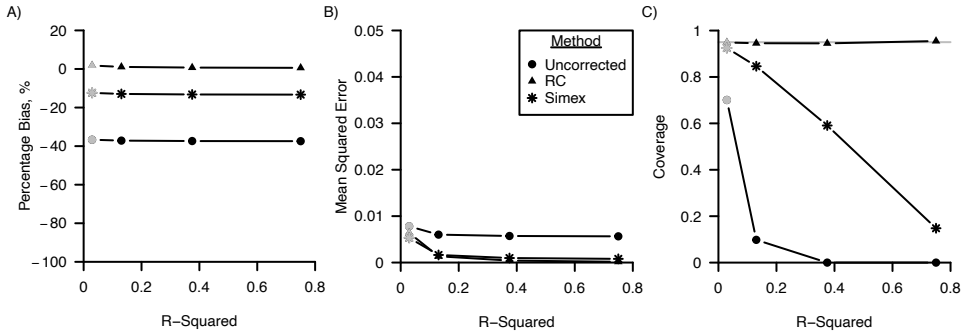


Figure 7.4: Performance in a linear regression model of regression calibration (RC), simulation-extrapolation (simex) and the analysis ignoring random measurement error for varying R-squared of the outcome model in terms of A) percentage bias; B) mean squared error and C) coverage. For all three performance measures, Monte Carlo standard errors were smaller than 0.01 in all scenarios. The grey points indicate the base scenario where R-squared is assumed 0.03.

Figure 7.3 shows that the number of replicates had no effect on the percentage bias, MSE and confidence interval coverage for varying number of replicates of the error-prone measure.
Figure 7.4 shows that R-squared had no effect on percentage bias, and only a minor decrease in MSE was found for increasing levels of R-squared. In addition, Figure 7.4 shows that 95% confidence interval coverage was around the nominal level for the regression calibration corrected analysis for all values of the R-squared. However, for the uncorrected and the simulation-extrapolation corrected analysis, confidence interval coverage decreased for increasing values of R-squared. For R-squared equal to 0.75, confidence interval coverage decreased to 15 % and 0 % (MCSE < 0.01) for the simulation-extrapolation corrected and the uncorrected analysis, respectively.

In the scenarios where a dependency between the covariate age and the exposure error-free blood pressure was introduced by changing parameter γ in the data generating mechanism, the reliability of the error-prone measure was respectively 0.71, 0.94 and 0.98. However, percentage bias, MSE and confidence interval coverage of the uncorrected and corrected analyses were equal to the base scenario (the values in the base scenario are shown in e.g. Figure 7.1). By introducing an effect of age on blood pressure, the total variance of the error-free blood pressure increased. Consequently, the extra variability in the error-prone blood pressure measurement due to measurement error was relatively smaller than in the base scenario. Hence, it seemed as if the error-prone variable was more reliable, though the attenuation due to random measurement error stayed constant at a rate of 0.625.

Logistic regression. Figure 7.5 shows the percentage bias, MSE and confidence interval coverage for varying values of the reliability of the error-prone measure. The uncorrected analysis was biased for all values of the reliability, and the percentage bias decreased when reliability increased. Regression calibration provided percentage bias close to null when reliability was greater or equal to 0.2. Simulation-extrapolation provided biased

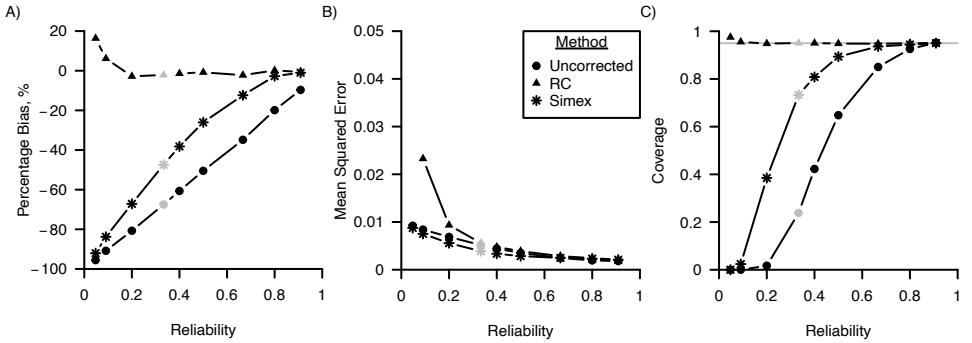


Figure 7.5: Performance in a logistic regression model of regression calibration (RC), simulation-extrapolation (simex) and the analysis ignoring random measurement error for varying values of the reliability of the error-prone measure in terms of A) percentage bias; B) mean squared error and C) coverage. For all three performance measures, Monte Carlo standard errors were smaller than 0.02 in all scenarios. The grey points indicate the base scenario where reliability is assumed 0.33. Mean squared error for RC and a reliability of 0.05 not shown (1.28, Monte Carlo standard error: 0.42).

results when reliability was smaller than 0.8 and bias was close to null otherwise. MSE was similar for simulation-extrapolation and the uncorrected analysis across the range of reliability. MSE was greater for regression calibration than for the uncorrected and simulation-extrapolation analysis when reliability was equal to or smaller than 0.2, and similar otherwise. Coverage of the 95% confidence intervals was at the nominal level for the regression calibration corrected analysis across the range of reliability, and for the simulation-extrapolation corrected analysis when reliability was greater than or equal to 0.66.

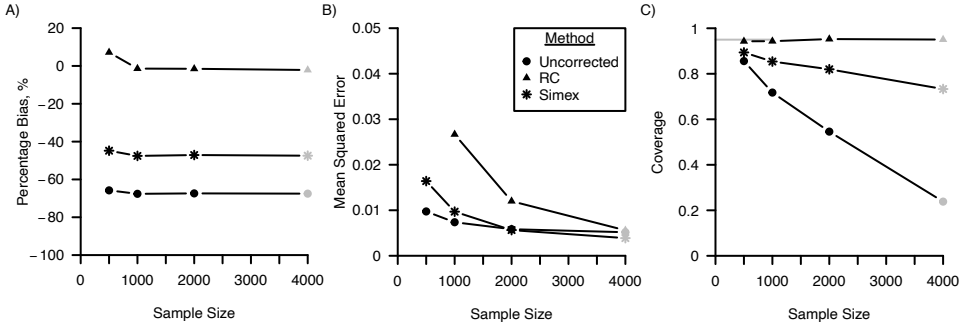


Figure 7.6: Performance in a logistic regression model of regression calibration (RC), simulation-extrapolation (simex) and the analysis ignoring random measurement error for varying sample sizes of the error-prone measure in terms of A) percentage bias; B) mean squared error and C) coverage. For all three performance measures, Monte Carlo standard errors were smaller than 0.01 in all scenarios. The grey points indicate the base scenario where sample size is assumed 4,000. Mean squared error for RC and a sample size of 500 not shown (0.06, Monte Carlo standard error: <0.01)

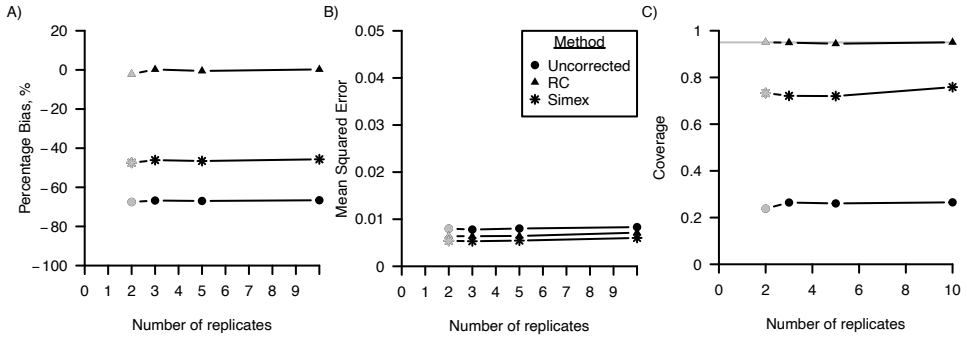


Figure 7.7: Performance in a logistic regression model of regression calibration (RC), simulation-extrapolation (simex) and the analysis ignoring random measurement error for varying number of replicates of the error-prone measure in terms of A) percentage bias; B) mean squared error and C) coverage. For all three performance measures, Monte Carlo standard errors were smaller than 0.01 in all scenarios. The grey points indicate the base scenario where the number of replicates is assumed 2.

Figure 7.6 shows the percentage bias, MSE and confidence interval coverage for varying samples sizes of the main study. Percentage bias was increased for regression calibration for a sample size of 500 compared to the base scenario where sample size was 4000 (7% vs -2%, MCSE < 0.01) and was similar otherwise. Percentage bias remained at a high level for simulation-extrapolation and the uncorrected analysis, ranging between -48% and -45% (MCSE < 0.01) and -68% and -66% (MCSE < 0.01), respectively. MSE was greater for smaller sample sizes. For a sample size of 4,000, MSE of regression calibration (ranging between 0.06 and 0.01, MCSE < 0.01) was greater than for simulation-extrapolation and the uncorrected analysis, ranging between 0.01 and 0.02, MCSE < 0.01 and 0.01, MCSE < 0.01, respectively. Coverage was equal to the nominal level of 95% for regression calibration for all sample sizes. The uncorrected analysis and simulation-extrapolation were undercovered with coverage levels decreasing for increasing size of the sample size, ranging between 24% and 86% (MCSE < 0.01) and 73% and 89% (MCSE < 0.01), respectively. A decline in confidence interval coverage for the simulation-extrapolation corrected analysis for larger sample sizes was confirmed by the scenario where sample size was 10,000, in which case coverage was 50% (MCSE 0.02) (not shown in the plots in Figure 7.6).

Figure 7.7 shows that the number of replicates had no effect on the percentage bias, MSE and confidence interval coverage for varying number of replicates of the error-prone measure.

Figure 7.8 shows that bias remains stable compared to the base scenario for varying values of Pseudo R-squared, except for Pseudo R-squared equal to 0.12. Mean squared error was higher for regression calibration than for the uncorrected or simulation-extrapolation corrected analysis for Pseudo R-squared equal to 0.12, 0.25 and 0.69. Coverage remained at the nominal level of 95% for regression calibration, and was subnominal for simulation-extrapolation and the uncorrected analysis with values ranging between 73%-85% and 24%-68%, respectively.

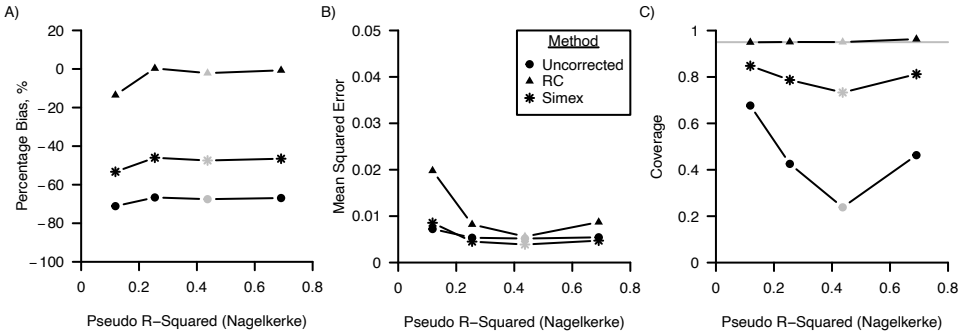


Figure 7.8: Performance in a logistic regression model of regression calibration (RC), simulation-extrapolation (simex) and the analysis ignoring random measurement error for varying R-squared (Nagelkerke) of the outcome model A) percentage bias; B) mean squared error and C) coverage. For all three performance measures, Monte Carlo standard errors were smaller than 0.01 in all scenarios. The grey points indicate the base scenario where R-squared is assumed 0.42

In the scenarios where a dependency between the covariate age and the error-free exposure sodium intake was introduced by changing parameter γ in the data generating mechanism, the reliability of the error-prone measure was respectively 0.34, 0.68, 0.87. Similar to what was seen for linear regression, percentage bias, MSE and confidence interval coverage of the uncorrected and corrected analyses were equal to the base scenario (the values in the base scenario are shown in e.g. Figure 7.5).

7

7.4. Sensitivity analysis in the absence of validation data

In the first example introduced in section Motivating example, replicate measurements of the error-prone systolic blood pressure were available. Nevertheless, validation data in the form of replicate measurements may not always be available. When random measurement error in a covariate is suspected in the absence of such validation data, a sensitivity analysis could be conducted using regression calibration or simulation-extrapolation. A general framework for conducting sensitivity analysis for random measurement error is described here, where we assume that the input of the sensitivity analysis, i.e., the measurement error variance and its uncertainty, are obtained from literature or expert knowledge. Next, a distribution for the measurement error variance is assumed, e.g., a uniform, triangular, or trapezidiol distribution [30]. Subsequently, regression calibration or simulation-extrapolation are applied to the data for measurement error correction, informed by the measurement error variance and its distribution. Finally, the results of the application of measurement error correction are presented, and conclusions drawn about the sensitivity of the results to measurement error.

7.4.1. Sensitivity analysis for measurement error in the example of blood pressure and kidney function in pregnant women

Suppose that in the example of the relation between blood pressure and kidney function in pregnant women discussed in section Motivating example (example 1), only the first systolic blood pressure measurement was available. A 10 mmHg increase in systolic blood

pressure was associated with a 1.18 μmol/L (95% CI 0.14;2.23) increase in serum creatinine. Random measurement error, however, could have been suspected in the single systolic blood pressure measurement and suppose the sensitivity of the results to the measurement error was studied. Suppose it was assumed that the variance of the measurement error in systolic blood pressure was equal to 48 mmHg, with a minimum of 37 mmHg and a maximum of 59 mmHg. Additionally, suppose a triangular distribution was assumed for the measurement error variance, meaning that most weight was put on 48 mmHg, and the weight was gradually reduced until it reached the assumed minimum and maximum level. The triangular distribution was sampled in accordance with Lash et al. [30].
Figure 7.9 shows the results of the application of regression calibration and simulation-extrapolation informed by the triangular distribution. For regression calibration, a clear pattern was obtained. The corrected effect estimates increased for larger values of the measurement error variance, with the effect estimates ranging from 1.75 - 2.38, with a median of 2.03. In addition, the associated lower limits of the confidence intervals consistently suggest an association between blood pressure and creatinine. In comparison, simulation-extrapolation did not show a clear pattern in the corrected effect estimates. The corrected effect estimates ranged from 1.43 - 1.88, with a median of 1.70. Figure 7.9 shows that the sampling variability that is inherent to simulation-extrapolation causes more variability in the effect estimates compared to the variability due to random measurement error. Nevertheless, the lower limits of the associated confidence intervals again consistently suggest an association between blood pressure and creatinine levels.

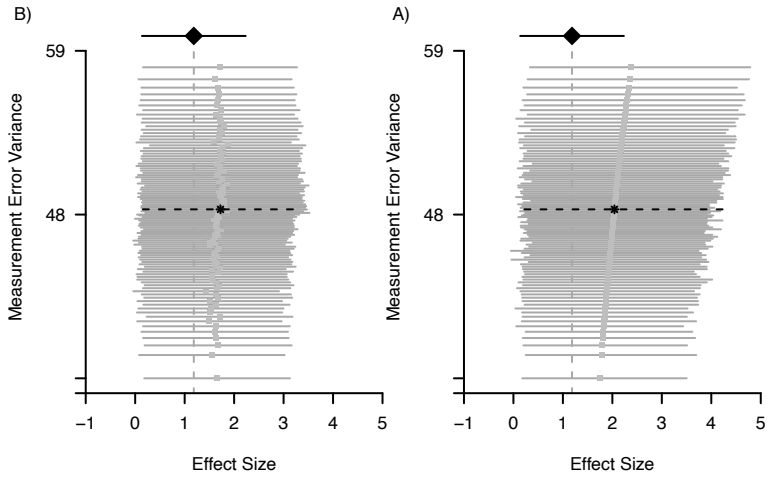


Figure 7.9: Sensitivity analysis for the association between blood pressure and kidney function in pregnant women (example 1, [32]) by application of regression calibration (panel A) and simulation-extrapolation (panel B). The uncorrected association and 95% confidence interval are depicted with a diamond and a solid black line, the measurement error corrected associations and 95% confidence intervals are depicted with a square and a solid gray line. The distribution of the measurement error variance is triangular. For reference, the measurement error corrected association and 95% confidence interval using the replicates data is depicted with a star and a dashed black line.

In the sensitivity analysis described here, results were presented graphically, and no summary estimate was shown. In the presented plots, the variability around the corrected

estimates was shown graphically. Alternatively, one can incorporate the variability around the corrected estimates in a so-called probabilistic bias analysis, by repeatedly sampling the corrected estimate from a distribution. Typically, it is assumed that the corrected estimate is normally distributed with mean equal to the point estimate and standard deviation equal to the standard error of the estimate. The sampled values can be presented by plotting the distribution of corrected estimates. There is a close resemblance between a probabilistic bias analysis and a Bayesian bias analysis with an uninformative prior for the association under study. We refer to MacLehose et al. for a thorough discussion of this issue [36]. More information about probabilistic bias analysis can be found in e.g., the book by Lash et al. [30] and more details about a Bayesian analysis for measurement error can be found in e.g., the book by Gustafson et al. [16].

7.5. Discussion

This chapter compared regression calibration and simulation-extrapolation for sensitivity analysis for random measurement error in an exposure variable. A simulation study showed that with correct assumptions about the measurement error variance, regression calibration was generally unbiased for linear and logistic regression when the reliability of the error-prone measurement was greater than 0.2. The bias in the regression calibration corrected analysis for linear regression for low reliability was unexpected, but may be explained by the instability of regression calibration when the correction factor is close to null. The bias in the regression calibration corrected analysis for logistic regression for low reliability was expected as Kuha showed that regression calibration for logistic regression is only approximately unbiased when the effect of the exposure on the outcome is 'small to moderate' and/or the measurement error variance 'small' [31]. Moreover, the uncorrected and simulation-extrapolation corrected analysis were generally biased, with higher bias for lower reliability of the error-prone exposure. Confidence interval coverage for regression calibration was generally close to the nominal level of 95% for linear and logistic regression. On the contrary, the confidence interval coverage of the simulation-extrapolation corrected and uncorrected analysis were subnominal.

The uncorrected analysis was shown more efficient than the corrected analyses for linear regression in settings where reliability was low (i.e., 0.2) or sample size small (i.e., 125). This observation is the result of the substantially smaller variance of the uncorrected analysis compared to the corrected analyses, which outweighs the larger bias for the uncorrected analysis. This is sometimes referred as the bias–variance trade off, we refer to chapter 3 of the book by Carroll et al. for a detailed discussion [13]. The same pattern was obtained for logistic regression. In addition, simulation-extrapolation showed a small gain in efficiency over regression calibration for linear regression in settings where reliability was low (i.e., 0.2) or sample size small (i.e., 250 and 125), and similar efficiency otherwise. The efficiency gain of simulation-extrapolation over regression calibration obtained for linear regression was not seen in logistic regression because of the large bias in the simulation-extrapolation analysis in most settings.

The results of our simulation study were in line with the results of three previous simulation studies: the corrected analyses showed lower percentages bias compared to the uncorrected analysis and the simulation-extrapolation corrected analysis showed higher percentage bias compared to regression calibration [5, 28, 29]. However, important

differences were observed. First, simulation-extrapolation showed a small gain in efficiency over regression calibration for linear regression in some settings, which was not found in the previous simulation studies. The sample sizes for which this gain in efficiency for simulation-extrapolation was observed (i.e., 125, 250 and 500) were smaller than those assumed by Perrier et al. [5] and Batistatou et al. [28] (i.e., 3,000 and 1,000, respectively), which may explain the found difference. Second, our simulation showed no effect of the number of replicates on bias. While the simulation study by Perrier et al. showed that an increasing number of replicates reduced bias in the corrected analyses [5]. This difference is explained by the fact that in the study by Perrier et al., the replicate measures were pooled before applying measurement error correction. By pooling the replicate measures with random measurement error, the measurement error variance is reduced. Therefore, bias decreased in the study by Perrier et al. with the availability of more replicate measures. This effect however is solely due to pooling the replicates measures and not due to a more precise estimate of the measurement error variance, as was shown by our results. Third, the simulation study by Fung et al. [29] showed that by increasing the correlation between the exposure and a covariate, the attenuation in the uncorrected analysis increased toward the null value. In comparison, our simulation study showed no effect of covariate dependency on bias in the uncorrected analysis. This is explained by the fact that in our simulation study, the variance of the exposure given the covariate was kept constant, while the total variance of the exposure was varied by introducing the covariate dependency (i.e., changing γ in the data generating mechanism). In comparison, in the simulation study by Fung et al., the variance of the exposure given the covariate was varied, resulting in an increase in the attenuation factor.

Our simulation study showed that percentage bias in the uncorrected analysis was equal to 1 minus the reliability of the error-prone measure times 100, in line with theory [8, 9]. The reliability of an error-prone measure equals the variance of the error-free measure divided by the variance of the error-prone measure. For example, in Figure 7.1, bias in the uncorrected analysis was equal to 80% for a reliability equal to 0.2. The uncorrected effect estimate is equal to 0.2 times the estimand 0.2, i.e., 0.04. From that, it follows that the bias is equal to $0.2-0.04 = 0.16$, which is 80% of 0.2. It is, however, important to note that the percentage bias is not equal to 1 minus the reliability of the error-prone measure when the total variance of the error-free measure depends on a covariate that is also included in the outcome model. For example, in our simulation study, the association between creatinine and systolic blood pressure given age was estimated. When a dependency between systolic blood pressure and age was introduced, the reliability increased to a maximum of 0.98 while the percentage bias in the uncorrected analysis was constant at 62.5%. A formula for the attenuation in the effect estimate due to random measurement error in multivariable models can be found in e.g. [13].

In our simulation study, the measurement error variance used to correct for the random measurement error was estimated using replicate measures. However, we assumed that these replicate measures were solely available to estimate the measurement error variance, to mimic a setting where such validation data are not available, yet unbiased information is available about the measurement error variance. In future studies, this work could be extended to settings where the measurement error is estimated with bias, and to settings where the measurement error model is misspecified. Also non-random measurement error, e.g., systematic measurement error and differential measurement error, which was not the

topic of this study, could be considered in future work. Simulation-extrapolation is not suited for the correction of measurement other than random measurement error, and for regression calibration, the full calibration model needs to be specified. We refer to Nab et al. [23] for a specification of the calibration model in case of systematic measurement error and what validation data can be used to estimate the calibration model. In addition, in our simulation study, models with one covariate and normal distributed measurement error were considered. The results of our study can be extended to settings with more covariates and measurement error with a skewed or heavy-tailed distribution. The covariate in our data generating mechanism can be viewed as a summary of a larger set of variables. What is more, transformations can turn a skewed or heavy-tailed measurement error distribution into a distribution that is closer to the normal distribution, as proposed by Carrol et al. [13]. Alternatively, adopted versions of regression calibration for heteroscedastic measurement error could be used [37].

Our study discussed measurement error correction methods for sensitivity analysis of random measurement error in a continuous exposure. For a categorical exposure, measurement error will lead to misclassification of the exposure. In this setting, different measurement error correction methods can be used. For example, the misclassification simulation-extrapolation [38], available in the R package `simex` [25]. Or alternatively, the probabilistic sensitivity analysis of misclassified binary variables described by Fox et al. [39].

Our study explored the performance of regression calibration and simulation-extrapolation for the correction of random measurement error in a linear regression model and a logistic regression model. For a survival model, the effects of random measurement error cannot be derived exactly as shown in chapter 14 of the book by Carroll et al. [13]. In particular, regression calibration gives approximately consistent estimates only in cases of a rare outcome, and for a hazard ratio of 'small to moderate' size or 'small' measurement error variance [40]. Xie et al. proposed a more flexible regression calibration approach for Cox regression that is referred to as 'risk set regression calibration' [41]. Alterations of the simulation-extrapolation method have been proposed for proportional hazard models [42] and accelerated failure time models [43]. For Poisson regression, regression calibration only provides estimates that are approximately unbiased, and usually works well, when the effect of the exposure on the outcome is 'small to moderate' or the measurement error variance 'small' [13]. Fung et al. compare regression calibration and simulation-extrapolation for Poisson regression and concluded that regression calibration performed best in all scenarios considered [44].

The Achilles heel of simulation-extrapolation is the extrapolation step [3]. Our simulation study uses a quadratic extrapolation. Lockwood et al. demonstrate the use of a quartic extrapolation, that may reduce bias in the simulation-extrapolation estimator [45].

In the example presented in section 7.4, the five steps of a sensitivity analysis for random exposure measurement error were described: 1) quantify the measurement error variance and its uncertainty; 2) specify the distribution of the measurement error variance; 3) perform measurement error correction by means of regression calibration or simulation-extrapolation; 4) visualise the results, and 5) draw conclusions. A sensitivity analysis using regression calibration showed that the higher the measurement error variance, the more the corrected effect estimate departs from the null, which is in line

with the literature [8, 9]. In the sensitivity analysis using simulation-extrapolation, the variability in the corrected effect estimates due to the sampling variability inherent to simulation-extrapolation exceeded the variability in the corrected effect estimates due to the assumed measurement error variance. Our simulation results showed that the regression calibration estimator is generally unbiased while the simulation-extrapolation estimator is. In contrast, simulation-extrapolation showed a small efficiency gain over regression calibration. Despite the efficiency gain for simulation-extrapolation, we recommend the use of regression calibration for sensitivity analysis. In a sensitivity analysis, focus is on the quantification of the impact of measurement error on the point estimate, and the confidence interval width may be of lesser importance.

In conclusion, regression calibration and simulation-extrapolation are suited for sensitivity analysis for random measurement error. It is difficult to say anything definite about the behavior of regression calibration and simulation-extrapolation based on a handful of simulation studies. We have, however, covered many aspects in our simulation study, i.e., reliability, sample size, number of replicates, explained variance of the outcome model and covariate dependency. The pattern is so pronounced and in accordance with findings of former simulation studies [5, 28, 29], that we think it is safe to say that regression calibration may be preferred over simulation-extrapolation. Nevertheless, if researchers want to compare simulation-extrapolation with regression calibration in simulation settings that are closer to their intended field of application, then we provided our simulation code, which can be modified easily to allow for investigation of such scenarios.

References

[1] T. B. Brakenhoff, M. Mitroiu, R. H. Keogh, K. G. M. Moons, R. H. H. Groenwold, M. van Smeden, Measurement error is often neglected in medical literature: A systematic review, Journal of Clinical Epidemiology 98 (2018) 89–97. doi:10.1016/j.jclinepi.2018.02.023.

[2] P. A. Shaw, V. Deffner, R. H. Keogh, J. A. Tooze, K. W. Dodd, H. Küchenhoff, V. Kipnis, L. S. Freedman, Epidemiologic analyses with error-prone exposures: Review of current practice and recommendations, Annals of Epidemiology 28 (11) (2018) 821–828. doi:10.1016/j.annepidem.2018.09.001.

[3] R. H. Keogh, P. A. Shaw, P. Gustafson, R. J. Carroll, V. Deffner, K. W. Dodd, H. Küchenhoff, J. A. Tooze, M. P. Wallace, V. Kipnis, L. S. Freedman, STRATOS guidance document on measurement error and misclassification of variables in observational epidemiology: Part 1—Basic theory and simple methods of adjustment, Statistics in Medicine 39 (16) (2020) 2197–2231. doi:10.1002/sim.8532.

[4] J. S. Frazer, G. E. Barnes, V. Woodcock, E. Flanagan, T. Littlewood, R. J. Stevens, S. Fleming, H. F. Ashdown, Variability in body temperature in healthy adults and in patients receiving chemotherapy: Prospective observational cohort study, Journal of Medical Engineering & Technology 43 (5) (2019) 323–333. doi:10.1080/03091902.2019.1667446.

[5] F. Perrier, L. Giorgis-Allemand, R. Slama, C. Philippat, Within-subject pooling of biological samples to reduce exposure misclassification in biomarker-based studies, Epidemiology 27 (3) (2016) 378–388. doi:10.1097/EDE.0000000000000460.

[6] B. Brunekreef, D. Noy, P. Clausing, Variability of exposure measurements in environmental epidemiology, American Journal of Epidemiology 125 (5) (1987) 892–898. doi:10.1093/oxfordjournals.aje.a114606.

[7] C. Spearman, The proof and measurement of association between two things, The American Journal of Psychology 15 (1) (1904) 72–101. doi:10.2307/1412159.

[8] C. Frost, S. G. Thompson, Correcting for regression dilution bias: comparison of methods for a single predictor variable, Journal of the Royal Statistical Society: Series A (Statistics in Society) 163 (2) (2000) 173–189. doi:10.1111/1467-985X.00164.

[9] J. A. Hutcheon, A. Chiolero, J. A. Hanley, Random measurement error and regression dilution bias, BMJ 340 (c2289) (2010). doi:10.1136/bmj.c2289.

[10] B. Armstrong, Measurement error in the generalised linear model, Communications in Statistics - Simulation and Computation 14 (3) (1985) 529–544. doi:10.1080/03610918508812457.

[11] J. W. Bartlett, R. H. Keogh, Bayesian correction for covariate measurement error: A frequentist evaluation and comparison with regression calibration, Statistical Methods in Medical Research 27 (6) (2018) 1695–1708. doi:10.1177/0962280216667764.

[12] J. Buonaccorsi, Measurement error: Models, methods, and applications, Chapman & Hall/CRC, Boca Raton, FL, 2010.

[13] R. Carroll, D. Ruppert, L. Stefanski, C. Crainiceanu, Measurement error in nonlinear models: A modern perspective, 2nd Edition, Chapman & Hall/CRC, Boca Raton, FL, 2006.

[14] S. R. Cole, H. Chu, S. Greenland, Multiple-imputation for measurement-error correction, International Journal of Epidemiology 35 (4) (2006) 1074–1081. doi:10.1093/ije/dyl097.

[15] J. R. Cook, L. A. Stefanski, Simulation-extrapolation estimation in parametric measurement error models, Journal of the American Statistical Association 89 (428) (1994) 1314–1328. doi:10.2307/2290994.

[16] P. Gustafson, Measurement error and misclassification in statistics and epidemiology: Impacts and Bayesian adjustments, Chapman & Hall/CRC, Boca Raton, FL, 2004.

[17] W. Fuller, Measurement error models, John Wiley & Sons, New York, NY, 1987.

[18] R. H. Keogh, I. R. White, A toolkit for measurement error correction, with a focus on nutritional epidemiology, Statistics in Medicine 33 (12) (2014) 2137–2155. doi:10.1002/sim.6095.

[19] L. Gleser, Improvements of the naive approach to estimation in nonlinear errors-in-variables regression models, in: P. Brown, W. Fuller (Eds.), Statistical analysis of measurement error models, American Mathematics Society, Providence, 1990, pp. 99–114.

[20] R. J. Carroll, L. A. Stefanski, Approximate quasi-likelihood estimation in models with surrogate predictors, Journal of the American Statistical Association 85 (411) (1990) 652–663. doi:10.1080/01621459.1990.10474925.

[21] L. Stefanski, J. Cook, Simulation-extrapolation: The measurement error jackknife, Journal of the Americal Statistical Association 90 (432) (1995) 1247. doi:10.2307/2291515.

[22] R Core Team, R: A language and environment for statistical computing (2020). URL https://www.r-project.org/

[23] L. Nab, M. van Smeden, R. H. Keogh, R. H. H. Groenwold, Mecor: An R package for measurement error correction in linear regression models with a continuous outcome, Computer Methods and Programs in Biomedicine 208 (2021) 106238. doi:10.1016/j.cmpb.2021.106238.

[24] B. Rosner, D. Spiegelman, W. , Correction of logistic regression relative risk estimates and confidence intervals for random within-person measurement error, American Journal of Epidemiology 136 (11) (1992) 1400–1413. doi:10.1093/oxfordjournals.aje.a116453.

[25] W. Lederer, H. Küchenhoff, The r journal: A short introduction to the simex and mcsimex, R News 6 (2006) 26–31, https://journal.r-project.org/articles/RN-2006-031/.

[26] J. W. Hardin, H. Schmiediche, R. J. Carroll, The simulation extrapolation method for fitting generalized linear models with additive measurement error, The Stata Journal 3 (4) (2003) 373–385. doi:10.1177/1536867X0400300407.

[27] StataCorp, Stata statistical software: Release 16 (2019).
URL https://www.stata.com

[28] E. Batistatou, R. McNamee, Performance of bias-correction methods for exposure measurement error using repeated measurements with and without missing data, Statistics in Medicine 31 (28) (2012) 3467–3480. doi:10.1002/sim.5422.

[29] K. Y. Fung, D. Krewski, Evaluation of regression calibration and SIMEX methods in logistic regression when one of the predictors is subject to additive measurement error, Journal of Epidemiology and Biostatistics 4 (2) (1999) 65–74.

[30] T. Lash, M. Fox, A. Fink, Applying quantitative bias analysis to epidemiologic data, Springer, New York, NY, 2009.

[31] J. Kuha, Corrections for exposure measurement error in logistic regression models with an application to nutritional data, Statistics in Medicine 13 (11) (1994) 1135–1148. doi:10.1002/sim.4780131105.

[32] E. McCarthy, T. Carins, Y. Hannigan, N. Bardien, A. Shub, S. Walker, Effectiveness and safety of 1 vs 4 h blood pressure profile with clinical and laboratory assessment for the exclusion of gestational hypertension and pre-eclampsia: a retrospective study in a university affiliated maternity hospital, BMJ Open 5 (11) (2015) e009492–e009492. doi:10.1136/bmjopen-2015-009492.

[33] Centers for Disease Control and Prevention (CDC). National Center for Health Statistics (NCHS), National health and nutrition examination survey data (2015).
URL https://wwwn.cdc.gov/nchs/nhanes/ContinuousNhanes/Default.aspx?BeginYear=2015

[34] T. P. Morris, I. R. White, M. J. Crowther, Using simulation studies to evaluate statistical methods, Statistics in Medicine 38 (11) (2019) 2074–2102. doi:10.1002/sim.8086.

[35] A. Gasparini, Rsimsum: Summarise results from Monte Carlo simulation studies, Journal of Open Source Software 3 (26) (2018) 739. doi:10.21105/joss.00739.

[36] R. F. MacLehose, P. Gustafson, Is probabilistic bias analysis approximately Bayesian?, Epidemiology 23 (1) (2012) 151–158. doi:10.1097/EDE.0b013e31823b539c.

[37] D. Spiegelman, R. Logan, D. Grove, Regression calibration with heteroscedastic error variance, The International Journal of Biostatistics 7 (1) (2011) 1–34. doi:10.2202/1557-4679.1259.

[38] H. Küchenhoff, S. M. Mwalili, E. Lesaffre, A general method for dealing with misclassification in regression: The misclassification SIMEX, Biometrics 62 (1) (2006) 85–96. doi:10.1111/j.1541-0420.2005.00396.x.

[39] M. P. Fox, T. L. Lash, S. Greenland, A method to automate probabilistic sensitivity analyses of misclassified binary variables, International Journal of Epidemiology 34 (6) (2005) 1370–1376. doi:10.1093/ije/dyi184.

[40] R. L. Prentice, Covariate measurement errors and parameter estimation in a failure time regression model, Biometrika 69 (2) (1982) 331–342. doi:10.2307/2335407.

[41] S. X. Xie, C. Y. Wang, R. L. Prentice, A risk set calibration method for failure time regression by using a covariate reliability sample, Journal of the Royal Statistical Society: Series B (Statistical Methodology) 63 (4) (2001) 855–870. doi:10.1111/1467-9868.00317.

[42] Y. Li, X. Lin, Functional inference in frailty measurement error models for clustered survival data using the SIMEX approach, Journal of the American Statistical Association 98 (461) (2003) 191–203. doi:10.1198/016214503388619210.

[43] W. He, G. Y. Yi, J. Xiong, Accelerated failure time models with covariates subject to measurement error, Statistics in Medicine 26 (26) (2007) 4817–4832. doi:10.1002/sim.2892.

[44] K. Y. Fung, D. Krewski, On measurement error adjustment methods in Poisson regression, Environmetrics 10 (2) (1999) 213–224. doi:10.1002/(SICI)1099-095X(199903/04)10:2<213::AID-ENV349>3.0.CO;2-B.

[45] J. Lockwood, D. F. McCaffrey, Simulation-extrapolation for estimating means and causal effects with mismeasured covariates, Observational Studies 1 (2) (2015) 241–290. doi:10.1353/obs.2015.0007.

8

Quantitative bias analysis for a misclassified confounder in marginal structural models

Observational data are increasingly used with the aim of estimating causal effects of treatments, through careful control for confounding. Marginal structural models estimated using inverse probability weighting (MSMs-IPW), like other methods to control for confounding, assume that confounding variables are measured without error. The average treatment effect estimator in a MSM-IPW may however be biased when a confounding variable is error-prone. Using the potential outcome framework, we derive expressions for the bias due to confounder misclassification in analyses that aim to estimate the average treatment effect using a MSM-IPW. We compare this bias with the bias due to confounder misclassification in analyses based on a conditional regression model. Focus is on a point-treatment study with a continuous outcome. Compared to bias in the average treatment effect estimator from a conditional model, the bias in MSM-IPW can be different in magnitude, but is equal in sign. Also, we use a simulation study to investigate the finite sample performance of MSM-IPW and conditional models when a confounding variable is misclassified. Simulation results indicate that confidence intervals of the treatment effect obtained from MSM-IPW are generally wider and coverage of the true treatment effect is higher compared to a conditional model, ranging from over-coverage if there is no confounder misclassification to under-coverage when there is confounder misclassification. We illustrate in a study of blood pressure lowering therapy, how the bias expressions can be used to inform a quantitative bias analysis to study the impact of confounder misclassification, supported by an online tool.

This chapter is based on: L. Nab, R.H.H. Groenwold, M. van Smeden and R.H. Keogh, Quantitative bias analysis for a misclassified confounder: A comparison between marginal structural models and conditional models for point treatments, Epidemiology, 31 (6) (2020) 796–805. doi:10.1097/EDE.0000000000001239

8.1. Introduction

The aim of many observational epidemiologic studies is to estimate a causal relation between an exposure and an outcome, through careful control for confounding. In the case of a point-treatment, that is estimating the effect of a treatment at a single time point on a subsequent outcome, many methods exist that aim to estimate average treatment effects. These include traditional conditional regression analysis as well as marginal structural models estimated using inverse probability weighting (MSMs-IPW) [1, 2]. Unlike conditional regression, MSMs extend to estimation of joint treatment effects over multiple time points in longitudinal settings with time-dependent confounding [1, 3].

To obtain valid inference, MSMs-IPW, like other methods to control for confounding, assume that confounding variables are measured without error, an assumption hardly ever warranted in observational epidemiologic research [4–7]. A type of measurement error is classification error, which occurs when categorical variables are misclassified. For instance, smoking status (smoker vs non-smoker) is prone to classification error, but has been used as a confounding variable in studies investigating dialysis on mortality [8] and iron supplement use during pregnancy on anemia at delivery [9]. Another example of the use of a potentially misclassified confounding variable is alcohol use during pregnancy (yes vs no) in studies investigating associations between exposure to triptans during fetal life and risk of externalizing and internalizing behaviors in children [10]. In all aforementioned examples, MSMs were used to estimate the exposure–outcome relation, but the assumption of error-free confounding variables is possibly violated and may lead to bias in the treatment effect estimator.

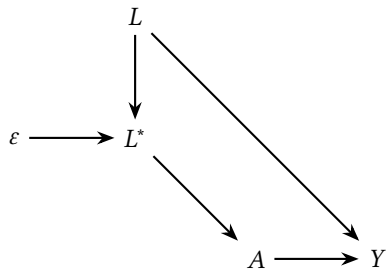
There is a substantial literature on bias due to measurement error in confounding variables in conditional regression analyses [11–15], but the impact of measurement error in confounding variables in causal inference methods, such as MSMs-IPW, has not received much attention. One exception is a study by Regier et al. that showed by means of a simulation study that measurement error in continuous confounding variables can introduce bias in the ATE in a point-treatment study [16]. McCaffrey et al. proposed a weighting method to restore the treatment effect estimator when covariates are measured with error [17].

We provide a discussion of measurement error in a confounding variable. In addition, we derive expressions that quantify the bias in the average treatment effect if a dichotomous confounding variable is misclassified, focusing on a point-treatment study with a continuous outcome. These expressions allow us 1) to quantify the bias due to classification error in a confounding variable in MSMs-IPW, and to compare this with the bias from a conditional regression analysis and 2) to inform quantitative bias analyses [18–20]. We use simulation results to study the finite sample performance of a MSM-IPW compared to that of conditional regression models if classification error in a confounding variable is present. We illustrate our quantitative bias analysis in a study of the effect of blood pressure lowering drugs on blood pressure.

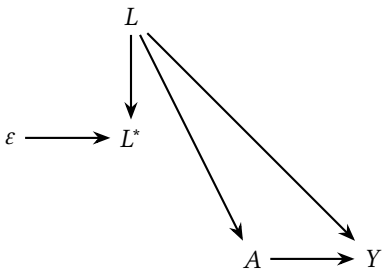
8.2. Settings and impact of measurement error, notation and assumptions

Let A denote the treatment indicator and Y the outcome. Let there be a variable L that confounds the association between treatment and outcome and suppose that, instead of confounding variable L, the error-prone confounding variable L^* is observed. We consider two settings in which measurement error in confounding variables may occur and discuss the impact of measurement error in both settings.

Settings and impact of measurement error. The directed acyclic graph (DAG) in Figure 8.1a illustrates **setting 1**. In this setting, treatment initiation is based on the error-prone confounding variable. Consider for example a study investigating the relation between the use of antidepressant drugs (A) and the risk of a hip fracture (Y) [21]. Benzodiazepine use may be a confounding variable, but is prone to classification error since only prescription information may be available and over-the-counter use is often unknown. The clinician initiating the antidepressant drugs might not know their patient's over-the-counter use and initiates treatment based on the observed error-prone benzodiazepine use (L^*) instead of actual use (L), as depicted in Figure 8.1a. Here, conditioning on the error-prone L^* will block the backdoor path from treatment A to outcome Y. Thus, it is sufficient to control for the error-prone confounding variable to estimate the causal effect of treatment on outcome. This means that measurement error in a confounding variable will not always lead to bias.



(a) **Setting 1**: treatment A is initiated based on the error-prone confounding variable L^*



(b) **Setting 2**: treatment A is initiated based on confounding variable L

Figure 8.1: Two settings of measurement error ε in variable L that confounds the association between treatment A and outcome Y illustrated in directed acyclic graphs

The DAG in Figure 8.1b illustrates **setting 2**, in which treatment initiation is based on L, but only a proxy of L is observed (L^*). An example here might be a study investigating the effect of influenza vaccination (A) on mortality (Y) in the elderly population [22]. Frailty (L) possibly confounds the association between influenza vaccination and mortality. Frailty is observed by a clinician, but only a proxy of frailty (L^*) may be available in electronic health records, as depicted in Figure 8.1b. Here, conditioning on L^* will not fully adjust for confounding by L, because conditioning on L^* does not block the backdoor path from A to Y via L.

Notation and assumptions. We will now continue investigating the impact of classification error in setting 2, by focusing on the setting where L is a dichotomous confounding variable and Y a continuous outcome. We use the potential outcomes framework [23, 24]. Let $Y^{a=0}$ denote the outcome that an individual would have had if treatment A was set to $a = 0$, and let $Y^{a=1}$ denote the outcome if treatment A was set to $a = 1$. We assume that L^* is non-differentially misclassified with respect to the outcome ($L^* \perp\!\!\!\perp Y|L$) and to the treatment ($L^* \perp\!\!\!\perp A|L$). Let p_1 denote the sensitivity of L^* and $1 - p_0$ the specificity of L^* (i.e., $P(L^*|L = l) = p_l$). We also denote the probability of treatment given the level of L by $P(A = 1|L = l) = \pi_l$ and the prevalence of L by $P(L = 1) = \lambda$. Here, we assume that $0 < \lambda < 1$ since we are not interested in populations where L is present or absent in everyone. Finally, we assume no measurement error in exposure and outcome.

We also assume that the following causal assumptions are satisfied to recover the causal effect of treatment on the outcome. Under the consistency assumption, we require that we observe $Y = Y^{a=0}$ if the individual is not exposed, or $Y = Y^{a=1}$ if the individual is exposed [25]. Further, we assume that the potential outcome Y^a for an individual does not depend on treatments received by other individuals and that there are not multiple versions of treatment, also referred to as Stable-Unit-Treatment-Value-Assumption [26]. Additionally, we assume conditional exchangeability, i.e., given any level of L, if the untreated group had in fact received treatment, then their expected outcome would have been the same as that in the treated, and vice versa [25]. In notation, $A \perp\!\!\!\perp Y^a|L$, for $a = 0, 1$. Finally, we assume $\pi_L > 0$ for $L = 0, 1$ (positivity) [27].

For causal contrasts, we compare expected potential outcomes (i.e., counterfactual outcomes) under the two different treatments. The average causal effect of the treatment on the outcome is $\beta = \mathrm{E}[Y^{a=1}] - \mathrm{E}[Y^{a=0}]$. Under the above defined assumptions, the conditional effect of treatment A on outcome Y can be defined through the following linear model:

$$\mathrm{E}[Y^a|L] = \mathrm{E}[Y|A = a, L] = \alpha + \beta a + \gamma L. \tag{8.1}$$

Estimates for β in the above model can be obtained by fitting a conditional regression model. Alternatively, the effect of treatment A on outcome Y may be estimated by fitting a MSM:

$$\mathrm{E}[Y^a] = \alpha_{\mathrm{msm}} + \beta a, \qquad \text{where} \quad \alpha_{\mathrm{msm}} = \alpha + \gamma \mathrm{E}[L]. \tag{8.2}$$

Estimates for β in the above model can be obtained by IPW estimation: by fitting a linear regression model for Y on A where the contribution of each individual is weighted by 1 over the probability of that individual's observed treatment given L [28], estimating the marginal treatment effect. Since our focus is on linear models and we make the simplifying assumption that the effect of A on Y does not vary between strata of L, the conditional and marginal treatment effects, denoted by β in model (8.1) and (8.2), respectively, are equal. This is not generally true for non-linear models due to non-collapsibility [28]. We assume that the effect of A on Y does not vary between strata of L, to derive bias expressions that are easier to use in practice and require fewer parameters [29].

8.3. Quantification of bias due to classification error in a confounding variable

Our aim is to study the effect of using the misclassified confounding variable L^* in place of the confounding variable L in the conditional regression model or in the model for the

weights used to fit the MSM on the average treatment effect estimator in the setting where L, not L^*, influences treatment initiation (setting 2 above).

Conditional model. By the law of total expectation, the expected value of the outcome Y given treatment A and L^* is (see S8.1 section Conditional model for further detail),

$$\begin{aligned}\mathsf{E}[Y|A=a,L^*] = \mathsf{E}_{L|A=a,L^*}\big[\mathsf{E}[Y|A=a,L^*,L]\big] &= \{\alpha+\gamma\phi_{00}+\delta u_0\}\\ &+ \{\beta+\gamma(\phi_{10}-\phi_{00})+\delta u_A\}a\\ &+ \{\gamma(\phi_{01}-\phi_{00})+\delta u_{L^*}\}L^*,\end{aligned}$$

where $\phi_{al^*} = P(L=1|A=a,L^*=l^*)$, $\delta = \mathsf{E}[Y|A=1,L^*=1] = \gamma(\phi_{11}-\phi_{10}-\phi_{01}+\phi_{00})$ and u_0, u_A, u_{L^*} represent the coefficients of the linear model $\mathsf{E}[AL^*|A,L^*] = u_0 + u_A A + u_{L^*}L^*$, modelling the mean of A times L^* (i.e., AL^*) given A and L^* (see next paragraph for an explanation of why these appear). The coefficient for treatment A in the above model is $\beta+\gamma(\phi_{10}-\phi_{00})+\delta u_A$, and is therefore biased for the parameter of interest (i.e., β). By rewriting u_A in terms of λ, π_0, π_1, p_0 and p_1 (see S8.1 section Conditional model), we find that the bias due to classification error in L^* in the average treatment effect in a conditional regression model is,

$$\begin{aligned}\text{Bias}_{\text{cm}}(\beta) = \gamma(\phi_{10}-\phi_{00})&\left(1-\ell\times\left\{\frac{\pi_1^*(1-\pi_1^*)}{\pi_1^*(1-\pi_1^*)\ell+\pi_0^*(1-\pi_0^*)(1-\ell)}\right\}\right)\\ +\gamma(\phi_{11}-\phi_{01})&\left(\ell\times\left\{\frac{\pi_1^*(1-\pi_1^*)}{\pi_1^*(1-\pi_1^*)\ell+\pi_0^*(1-\pi_0^*)(1-\ell)}\right\}\right),\end{aligned} \tag{8.3}$$

where $\pi_{l^*}^* = P(A=1|L^*=l^*)$, $\ell = P(L^*=1)$ (see S8.1 section Conditional model for a derivation).

We focused on a model for Y conditional on A and L^* which includes only main effects of A and L^*, as this is typically done in practice when replacing L with L^*. In fact, it can be shown that when the model for Y given A and L includes only main effects of A and L, the implied correctly specified model for Y given A and L^* also includes an interaction between A and L^*, explaining the appearance of u_0, u_A and u_L in the above since the interaction is not modeled. See S8.1 section Conditional model for the bias in case an interaction is modelled.

MSM-IPW. A MSM-IPW proceeds by fitting a linear regression for outcome Y on treatment A where the contribution of each individual is weighted by 1 over the probability of that individual's observed treatment given misclassified L^* [28]. An estimator for the average treatment effect β is,

$$\hat{\beta} = \frac{\sum_{i=1}^n \frac{1}{P(A_i|L_i^*)}(Y_i-\overline{Y}_w)(A_i-\overline{A}_w)}{\sum_{i=1}^n \frac{1}{P(A_i|L_i^*)}(A_i-\overline{A}_w)^2} \quad \text{where,} \quad \overline{Y}_w = \frac{\sum_{i=1}^n Y_i/P(A_i|L_i^*)}{\sum_{i=1}^n 1/P(A_i|L_i^*)}$$

$$\text{and,} \quad \overline{A}_w = \frac{\sum_{i=1}^n A_i/P(A_i|L_i^*)}{\sum_{i=1}^n 1/P(A_i|L_i^*)}.$$

It can be shown that $\mathsf{E}[\hat{\beta}] = \beta+\gamma(\phi_{10}-\phi_{00})(1-\ell)+\gamma(\phi_{11}-\phi_{01})\ell$. Consequently, the bias in the average treatment effect β in a MSM-IPW is,

$$\text{Bias}_{\text{msm}}(\beta) = \gamma(\phi_{10}-\phi_{00})(1-\ell)+\gamma(\phi_{11}-\phi_{01})\ell. \tag{8.4}$$

We refer to S8.1 section Marginal structural model estimated using inverse probability weighting for a derivation of the above formula.

8.3.1. Exploration of bias

To study the bias due to misclassification from the conditional model and MSM-IPW, we explore bias expressions (8.3) and (8.4).

Null-bias. To confirm the derived bias expressions, we consider three trivial conditions where bias in the average treatment effect is expected to be null, in line with general understanding of causal inference [30]. (1) If there is no classification error in L^*, i.e., specificity is 1 ($p_0 = 0$) and sensitivity is 1 ($p_1 = 1$), it follows that L corresponds to L^*, irrespective of treatment level (i.e., $\phi_{10} = 0$, $\phi_{00} = 0$, $\phi_{11} = 1$ and $\phi_{01} = 1$). (2) If the true relation between L and Y is null (i.e., γ is zero, thus there is no arrow from L to Y in Figure (8.1b)). (3) If L does not affect the probability of receiving treatment (i.e., $\pi_0 = \pi_1$, thus there is no arrow from L to A in Figure (8.1b)), the probability that L is 1 depends on the value of L^* but no longer on A (i.e., $\phi_{00} = \phi_{10}$ and $\phi_{01} = \phi_{11}$). Bias is null under these conditions for both models (MSM-IPW and conditional model). Since the bias expressions are strictly monotonic, the bias in a MSM-IPW cannot be negative if the bias in the conditional model is positive and vice versa (i.e., the bias will be in the same direction for both models).

Equal biases. The bias in the average treatment effect from the conditional regression analysis is equal to that from the MSM-IPW if bias in both models is null (see above). We also see that bias expressions (8.3) and (8.4) show that bias for the two methods is equal if the term between curly brackets in equation (8.3) is equal to 1, which is the case if: (i) $\ell = 1$; (ii) $\pi_0^* = \pi_1^*$; (iii) $\pi_0^* = 1 - \pi_1^*$. If conditions i and/or ii are met, there is no bias in a MSM-IPW nor in a conditional model. Under condition iii, bias is generally non-null (except if for example $\gamma = 0$, see null-bias).

Sign and magnitude of bias. Figures 8.2-8.4 illustrate the contributions to bias in the average treatment effect estimator due to misclassification components (sensitivity and specificity) and due to confounding components (prevalence of confounding variable, strength of association between confounding variable and treatment and outcome) in a conditional model and a MSM-IPW, obtained by using the bias expressions.

Figure 8.2 shows that: (1) the bias is positive if both the association between L and treatment and, L and outcome are positive (i.e., $\pi_1 > \pi_0$ and $\gamma = 2$, respectively), and (2) the bias is greater if the difference between π_1 and π_0 is greater (i.e., if the strength of the association between L and treatment is greater). In contrast, the bias is negative if $\pi_1 < \pi_0$, while γ is positive. In case $\gamma = -2$, Figure 8.2 is mirrored in $y = 0$ and consequently, bias is negative if $\pi_1 > \pi_0$ and positive if $\pi_1 < \pi_0$. An increment in γ will result in greater bias and steeper curves in Figure 8.2. Figure 8.3 shows that the magnitude of the bias depends on the prevalence of L. Further, it shows that bias is greater if the strength of association between L and treatment is greater. Figure 8.4 shows that, generally, the bias is greater if L^* has lower specificity and sensitivity. Moreover, for a fixed sensitivity, bias is minimal if specificity equals 1 and is maximal if 1 minus specificity equals sensitivity; by fixing specificity, bias is minimal if sensitivity equals 1 and is maximal if sensitivity equals 1 minus specificity. Figure 8.4 shows that the bias is greater if the strength of the association between L and treatment is greater. An increment in γ will result in greater bias and steeper curves in Figure 8.4. An online application can be used to obtain bias plots for other combinations of the parameters available at: https://lindanab.shinyapps.io/SensitivityAnalysis.

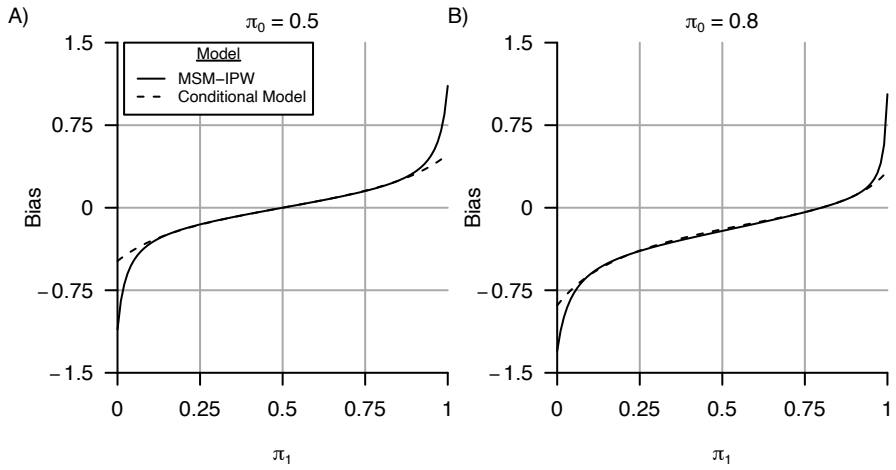


Figure 8.2: Visualisation of the direction and magnitude of the bias in the average treatment effect in relation to the prevalence of treatment among individuals with the confounding variable present. In this visualisation, the confounding variable L is misclassified with a sensitivity of 0.9 and specificity of 0.95. Consequently, the average treatment effect estimated in a MSM-IPW or conditional regression model is biased, independent of true average treatment effect. The prevalence of L is 50% (i.e., $P(L = 1) = 0.5$). The direction and magnitude of the bias depend on: (1) the strength and direction of the association between L and treatment (denoted by $\pi_1 = P(\text{treatment} = 1|L = 1)$ and $\pi_0 = P(\text{treatment} = 1|L = 0)$, here set at $\pi_0 = 0.5$ in the left-hand-side plot and $\pi_0 = 0.8$ in the right-hand-side plot); and (2) the strength and direction of the association between L and the outcome (denoted by γ in the text and here set at $\gamma = 2$). Larger values of γ will result in steeper curves; $\gamma = -2$ will mirror the graph in $y = 0$.

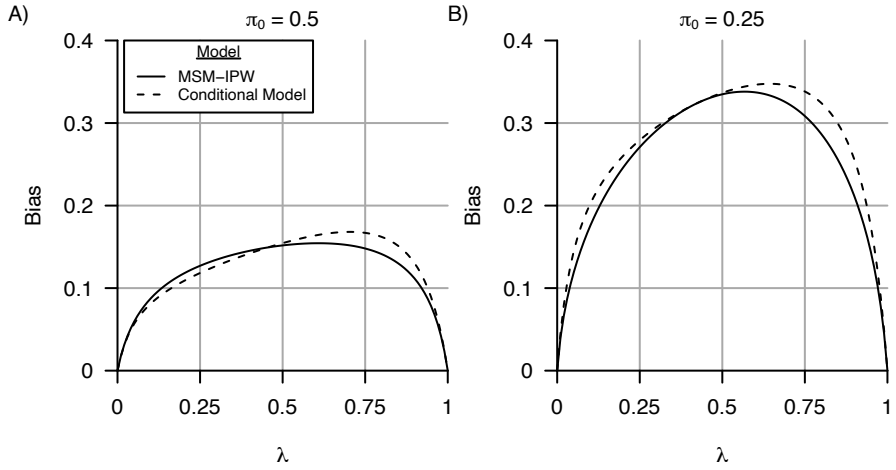


Figure 8.3: Visualisation of the magnitude of the bias in the average treatment effect in relation to the prevalence of a confounding variable. In this visualisation, the confounding variable L is misclassified with a sensitivity of 0.9 and specificity of 0.95. Consequently, the average treatment effect estimated in a MSM-IPW or conditional regression model is biased, independent of true average treatment effect. The confounding variable is positively associated with treatment (i.e., here $\pi_1 > \pi_0$, where $\pi_1 = P(\text{treatment} = 1|L = 1)$ and $\pi_0 = P(\text{treatment} = 1|L = 0)$), and outcome (denoted by γ in the text and here set at $\gamma = 2$). The magnitude of the bias depends on the prevalence of the confounding variable (i.e., $P(L = 1)$). Larger values of γ will result in steeper curves.

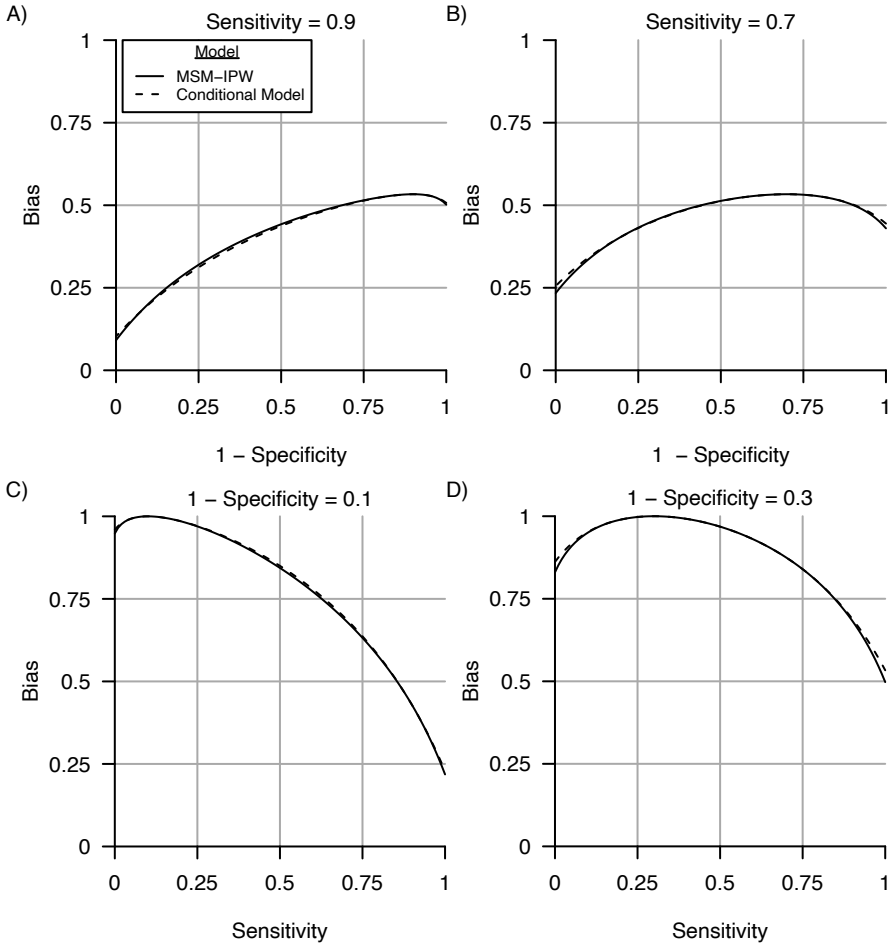


Figure 8.4: Visualisation of the magnitude of the bias in the average treatment effect in relation to specificity and sensitivity of a misclassified confounding variable. In this visualisation, the prevalence of the confounding variable L is 50% (i.e., $P(L = 1) = 0.5$), the association between L and treatment (denoted by $\pi_1 = P(\text{treatment} = 1|L = 1)$ and $\pi_0 = P(\text{treatment} = 1|L = 0)$) and outcome is positive (denoted by γ in the text and here set at $\gamma = 2$). Given these values, if L is misclassified, the average treatment effect estimated in a MSM-IPW or conditional regression model is biased, independent of true average treatment effect. The magnitude of the bias depends on the specificity and sensitivity of L and is maximal if sensitivity equals 1 minus specificity. The strength of the association between L and treatment is greater in the right-hand-side plot ($\pi_0 = 0.25, \pi_1 = 0.75$) compared to the left-hand-side plot ($\pi_0 = 0.5, \pi_1 = 0.75$) and consequently, bias is greater. Larger values of γ will result in steeper curves.

8.3.2. Simulation study

We conducted a simulation study to study the finite sample properties of MSMs-IPW and conditional models if there is classification error in the confounding variable. Five-thousand data sets were generated with sample sizes of 1,000 and 100, using the following data generating mechanisms:

$$\begin{aligned} L &\sim \text{Bern}(\lambda), \quad A|L \sim \text{Bern}\left(\pi_0^{(1-L)}\pi_1^{L}\right), \\ L^*|L &\sim \text{Bern}\left(p_0^{(1-L)}p_1^{L}\right) \quad \text{and,} \quad Y|A,L \sim \text{N}(1+\beta A+\gamma L, 1). \end{aligned}$$

We studied five different scenarios, of which the parameters values can be found in Table 8.1. In all scenarios, the average treatment effect β (estimand) is 1 and the association between the confounding variable L and outcome Y is 2 (i.e., $\gamma = 2$). In scenario 0, we assume no classification error. In scenarios 1-4, we assume that L^* has a specificity of 0.95 (i.e., $p_0 = 0.05$) and a sensitivity of 0.90 (i.e., $p_1 = 0.9$). In scenario 1, bias in the average treatment effect β is expected to be negative since the probability of receiving treatment given that L is not present is greater than receiving treatment given that L is present, and the association between L and Y is positive (i.e., $\pi_0 > \pi_1$ and $\gamma = 2$). In contrast, in scenario 2 and 3, bias in the average treatment effect is expected to be positive, since $\pi_0 < \pi_1$ and $\gamma = 2$. Further, after investigation of Figure 8.3, we expect that bias in the average treatment effect estimated in a conditional model is greater than that in a MSM-IPW in scenario 2 and 3. Finally, in scenario 4, we expect that bias in the average treatment effect from the conditional model is equal to that in a MSM-IPW.

Model estimation and performance measures. We obtained the average treatment effect β (estimand) by fitting a conditional model using conditional regression and by fitting a MSM-IPW, both using the misclassified L^* instead of L from the data generating mechanism. For the MSM-IPW analysis we used the R package `ipw` [31] [32]. Performance of both models was evaluated in terms of the bias, the mean squared error of the estimated treatment effect (MSE), the percentages of 95% confidence intervals that contain the true value of the estimand (coverage), the empirical standard deviation of the estimated treatment effects (empSE) and mean model based standard error of the estimated treatment effect. We estimated robust model based standard errors of the average treatment effect in a MSM-IPW using the R package `survey` [33]. We calculated Monte Carlo standard errors for all performance measures [34], using the R package `rsimsum` [35]. Additionally, we calculated the theoretical bias of the average treatment effect in both methods based on the bias expressions (8.3) and (8.4).

Table 8.1: Values of the parameters in the five different simulation scenarios

Scenario Number	**Parameters**						
	p_0	p_1	λ	π_0	π_1	β	γ
0	0	1	0.50	0.50	0.75	1	2
1	0.05	0.90	0.50	0.90	0.45	1	2
2	0.05	0.90	0.80	0.25	0.75	1	2
3	0.05	0.90	0.80	0.50	0.75	1	2
4	0.05	0.90	0.45	0.50	0.75	1	2

Table 8.2: Results of simulation study studying the finite-sample properties of a marginal structural models estimated using inverse probability weighting (MSM-IPW) and a conditional model (CM) if there is classification error in the confounding variable. Bias formula indicates the bias based on bias expressions (8.3) and (8.4) in the text, MSE mean squared error, EmpSE empirical SE and ModelSE model based SE.

Method	Sample Size	Scenario[a]	Bias Formula	Bias	MSE	Coverage	EmpSE	ModelSE
MSM-IPW	1,000	0	0.00	0.00 (0.001)	0.00 (0.000)	0.99 (0.001)	0.07 (0.001)	0.10 (0.000)
		1	-0.42	-0.42 (0.001)	0.18 (0.001)	0.03 (0.002)	0.10 (0.001)	0.11 (0.000)
		2	0.14	0.14 (0.001)	0.03 (0.000)	0.67 (0.007)	0.08 (0.001)	0.09 (0.000)
		3	0.29	0.29 (0.001)	0.09 (0.001)	0.08 (0.004)	0.08 (0.001)	0.09 (0.000)
		4	0.15	0.15 (0.001)	0.03 (0.000)	0.68 (0.007)	0.08 (0.001)	0.10 (0.000)
	100	**0**	0.00	0.00 (0.003)	0.05 (0.001)	0.99 (0.001)	0.22 (0.002)	0.31 (0.000)
		1	-0.42	-0.42 (0.005)	0.29 (0.005)	0.78 (0.006)	0.34 (0.003)	0.37 (0.001)
		2	0.14	0.14 (0.004)	0.08 (0.002)	0.94 (0.003)	0.25 (0.003)	0.29 (0.000)
		3	0.29	0.29 (0.004)	0.15 (0.002)	0.84 (0.005)	0.26 (0.003)	0.28 (0.000)
		4	0.15	0.15 (0.004)	0.08 (0.002)	0.95 (0.003)	0.25 (0.002)	0.31 (0.000)
CM	1,000	0	0.00	0.00 (0.001)	0.00 (0.000)	0.95 (0.003)	0.07 (0.001)	0.07 (0.000)
		1	-0.34	-0.34 (0.001)	0.12 (0.001)	0.02 (0.002)	0.09 (0.001)	0.08 (0.000)
		2	0.16	0.16 (0.001)	0.03 (0.000)	0.46 (0.007)	0.08 (0.001)	0.08 (0.000)
		3	0.32	0.32 (0.001)	0.11 (0.001)	0.02 (0.002)	0.08 (0.001)	0.08 (0.000)
		4	0.15	0.15 (0.001)	0.03 (0.000)	0.49 (0.007)	0.08 (0.001)	0.07 (0.000)
	100	0	0.00	0.00 (0.003)	0.05 (0.001)	0.95 (0.003)	0.22 (0.002)	0.22 (0.000)
		1	-0.34	-0.33 (0.004)	0.19 (0.003)	0.73 (0.006)	0.29 (0.003)	0.27 (0.000)
		2	0.16	0.16 (0.004)	0.09 (0.002)	0.90 (0.004)	0.25 (0.003)	0.25 (0.000)
		3	0.32	0.32 (0.004)	0.17 (0.003)	0.74 (0.006)	0.26 (0.003)	0.25 (0.000)
		4	0.15	0.15 (0.003)	0.08 (0.002)	0.90 (0.004)	0.24 (0.002)	0.24 (0.000)

[a] In all scenarios, the average treatment effect (estimand) is 1 ($\beta = 1$) and the effect of the confounding variable on the outcome is 2 ($\gamma = 2$). Five-thousand data sets were generated. Monte Carlo standard errors are shown between brackets. In scenario 0, there is no classification error (specificity and sensitivity of the misclassified confounding variable are 1, i.e., $p_0 = 0$ and $p_1 = 1$). In scenarios 1-4, the specificity of the misclassified confounding variable is 0.95 (i.e., $p_0 = 0.05$) and the sensitivity is 0.9 (i.e., $p_1 = 0.9$). The prevalence of the confounding variable (λ), and the probability of receiving treatment if the confounding is not present or present (π_0 and π_1, respectively) are set as follows in the scenarios: scenario 0: $\lambda = 0.5$, $\pi_0 = 0.5$, $\pi_1 = 0.75$; scenario 1: $\lambda = 0.5$, $\pi_0 = 0.9$, $\pi_1 = 0.45$; scenario 2: $\lambda = 0.8$, $\pi_0 = 0.25$, $\pi_1 = 0.75$; scenario 3: $\lambda = 0.8$, $\pi_0 = 0.5$, $\pi_1 = 0.75$; scenario 4: $\lambda = 0.45$, $\pi_0 = 0.5$, $\pi_1 = 0.75$.

Results. Table 8.2 shows the results of the simulation study. Bias found in the simulation study corresponds to the theoretical bias derived from the bias expressions. The empirical standard deviation of the average treatment effect estimates (empSE) from the MSM-IPW is equal to or greater than that from the conditional model. Yet, in the scenarios where bias in the average treatment effect in the MSM-IPW was smaller than bias in the conditional model (scenarios 2 and 3), empSE of both methods was equal, and hence, MSE is smaller for one method if also bias is smaller. Furthermore, the (robust) model based standard errors of the average treatment effect in a MSM-IPW are conservative and greater than the empirical standard errors, since the uncertainty in estimating the treatment weights is not taken into account. Allowing for the estimation of the weights will shrink the standard errors [2, 28]. We chose not to use a less conservative standard error estimation for MSM-IPW, such as bootstrapping, since our goal was to frame this simulation as investigating the properties of the commonly used MSM-IPW estimation procedure. Consequently, confidence intervals of the treatment effect obtained in a MSM-IPW are generally wider and coverage of the true treatment effect is higher compared to a conditional model, ranging from over coverage if there is no classification error to smaller under coverage when there is classification error.

8.4. Illustration: quantitative bias analysis

Quantitative bias analysis provides a tool to incorporate uncertainty in study results due to systematic errors [18, 20]. Using an example study of blood pressure lowering therapy, we illustrate how the bias expressions (8.3) and (8.4) can be used to perform a quantitative bias analysis for misclassification of a confounding variable.

Application. For our illustration we use data of the National Health And Nutritional Examination Survey (NHANES) [36, 37], more details can be found in the supplementary material section S8.2. Specifically, we study the effect of diuretic use ($A = 1$) in comparison to beta blocker use ($A = 0$) on systolic blood pressure (Y) using two approaches: by inverse weighting with the propensity for diuretic or beta blocker use given self-reported categorical body mass index (BMI) (L^*), and using a conditional linear regression with adjustment for self-reported categorical BMI. For this illustration, we categorize self-reported BMI into two distinct categories: underweight/normal weight (BMI < 25 ($L^* = 0$)) and overweight/obese (BMI ≥ 25 ($L^* = 1$)). However, we stress that one should preferably not categorise BMI in most practical applications [38]. Moreover, we assume that dichotomizing self-reported BMI does not introduce differential misclassification [7].

We assume that blood pressure lowering therapy is initiated based on the true BMI (L) instead of the observed self-reported BMI (setting 2, Figure 8.1b). Further, we consider BMI the only confounding variable, and treatment and outcome to be measured without error, which is a simplification of reality. Additionally, we assume that the classification error in self-reported BMI category is non-differential for the subject's treatment or blood pressure (given true BMI category). Expert knowledge is needed to inform this assumption. To quantify how large the bias in the average treatment effect is expected to be due to classification error in self-reported BMI category, we perform a quantitative bias analysis using the bias expressions (8.3) and (8.4).

Average treatment effect. Table 8.3 shows the average treatment effect of diuretics use in comparison to beta blocker use on mean systolic blood pressure. In a MSM-IPW,

Table 8.3: Average treatment effect of diuretics use compared to beta blocker use on mean systolic blood pressure in NHANES [36, 37]. CI indicates confidence interval.

Model	Effect Size (95% CI)
Unadjusted	−4.03(−6.30; −1.76)
Marginal Structural Model[a]	−3.52(−5.74; −1.21)
Conditional Model[b]	−3.48(−5.76; −1.27)

[a] Estimated in a marginal structural model, by inverse weighting with the propensity for diuretic or beta blocker use given self-reported categorised body mass index (BMI).
[b] Estimated in a conditional regression model with adjustment for self-reported categorical BMI.

we estimated an average treatment effect (95 % CI) of −3.52 (−1.21; −5.74). In a conditional regression model, we estimated an average treatment effect (95 % CI) of −3.48 (−1.27; −5.76).

Quantitative bias analysis. To inform the quantitative bias analysis, we need to make assumptions on the sensitivity and specificity of the self-reported BMI as well as that classification errors are non-differential with respect to blood pressure and treatment. For the purpose of this illustration, we speculate ranges for the sensitivity and specificity of self-reported BMI category of 0.90 to 0.98. In practice, these parameters should be informed by reports in the literature and/or a researcher's expert experience. Researchers may also decide to investigate how extreme the misclassification (measured using sensitivity and specificity) would need to be to change the conclusions of their study. We refer to the Shiny application (introduced in the subsequent section) for other choices for the sensitivity and specificity of self-reported BMI category.

By uniformly sampling from the range of plausible values of p_0 and p_1 and using the bias expressions (8.3) and (8.4), a distribution of possible biases is obtained (see supplementary material section S8.2 for further details). The solid line in Figure 8.5 shows the distribution of bias in a MSM-IPW. Mean bias is -0.31 and median bias is -0.30 (interquartile range -0.40 to -0.20). We also considered sampling p_0 and p_1 from a trapezoidal (with modes at one third and two thirds between the minimum and maximum) or a symmetrical triangular distribution. Sampling from these distributions results in mean bias approximately equal to when uniform sampling is applied, but with less spread (panels B and C in Figure 8.5). This result suggests that the results in Table 8.3 are not affected much by the classification error in self-reported BMI category. In the NHANES, anthropometric measures were also taken by trained technicians. See S8.2 for the average treatment effect when BMI measures taken by trained technicians were used instead of self-reported BMI measures.

8.5. Shiny application: an online tool

We developed an online tool for creating bias plots (Figure 8.2-8.4) and performing quantitative bias analyses (illustrated in the previous section), available at `https://lindanab.shinyapps.io/SensitivityAnalysis`. The bias plots can be used to predict the implications of classification error in a confounding variable in specific study settings by varying: the strength of association between the confounding variable and treatment and between the confounding variable and outcome; prevalence of the confounding variable; specificity and sensitivity of the misclassified confounding variable. The quantitative bias analysis can be used for studying the impact of classification error

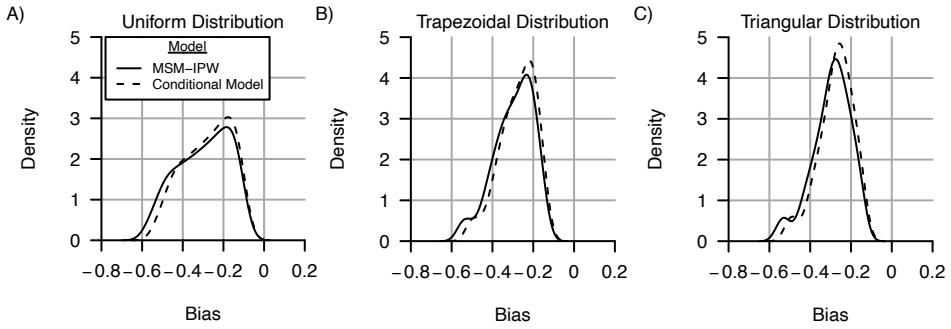


Figure 8.5: Density of predicted bias due to classification error in self-reported BMI category in NHANES [37]. Bias in the average treatment effect of diuretics use compared to beta blocker use on mean systolic blood pressure by inverse weighting with the propensity for diuretic or beta blocker use given self-reported categorical BMI (MSM-IPW), and using a conditional linear regression with adjustment for self-reported categorical BMI. The specificity and sensitivity of self-reported BMI category range from 0.90 to 0.98 and are sampled from a uniform distribution, trapezoidal (with modes on one-third and two-third), and symmetrical triangular distribution.

in a confounding variable at the analysis stage of a study, and to investigate how sensitive conclusions are to the assumption of no classification error. These bias plots can also be used to inform decisions about measurement methods or choice of variables to be extracted in the planning stage of studies.

8.6. Discussion

Inverse probability weighting and conditional models are both important and frequently used tools to adjust for confounding variables in observational studies. In this article, we derived expressions for the bias in the average treatment effect in a MSM-IPW and a conditional model. These expressions can inform quantitative bias analyses for bias due to a misclassified confounding variable.

Quantitative bias analysis of misclassified confounding variables is one example of quantitative bias analyses for observational epidemiologic studies. Several approaches exist to assess sensitivity of causal conclusions to unmeasured confounding [29, 39, 40]. These aim to quantify the impact of violations of the assumption of no unmeasured confounding, while our approach aims to quantify the impact of violations of the assumption that all confounding variables are measured without error.

Several methods have been proposed to adjust for measurement error in covariates in MSMs-IPW. Pearl developed a general framework for causal inference in the presence of error-prone covariates, which yields weighted estimators in the case of a dichotomous confounding variable measured with error [41]. The framework relies on a joint distribution of the outcome and the confounding variable. Conversely, the weighting method proposed by McCaffrey et al. does not require a model for the outcome [17]. Additionally, regression calibration [42], simulation-extrapolation [43, 44] and multiple imputation [45] have been proposed for correcting for measurement error in covariates of MSMs. These methods assume that the measurement error model is known, which may often be unrealistic. In this context it is also important to mention previous studies of the impact of measurement error in the exposure or the endpoint in MSMs, which has been studied by Babanezhad et

al. [46] and Shu et al. [47], respectively.

If treatment is allocated based on an error-prone confounding variable, the treatment effect will not be biased (see DAG in Figure 8.1a). However, investigators should be careful in concluding that covariate measurement error will not affect their analysis. Suppose that there is an unmeasured variable U that acts as a confounding variable between the error-prone covariate L^* and treatment A. Conditioning on L^* will then open a path between A and L via U and thus confound the relation between A and Y.

This article considered classification error in a dichotomous confounding variable in a point-treatment study with a continuous outcome. The same principles apply to measurement error in a categorical or continuous confounding variable or when multiple confounding variables are considered, although more elaborate assumptions should then be made [48]. Moreover, we assumed that the relation between exposure and outcome does not vary between strata of the confounding variable, i.e. that there is no treatment effect modification. Future research could extend our bias expressions by relaxing this simplifying assumption, therefore extending our results to more general settings.

MSMs-IPW are increasingly applied to longitudinal data to estimate the joint effects of treatment at multiple time points on a subsequent outcome, including time-dependent outcomes, addressing the problem of time-dependent confounding [1, 3]. There has been little work to understand or correct for the impact of misclassified or mismeasured confounding variables in this more complex setting. Our results extend directly to the time-dependent setting when the aim is to estimate the effect of a current treatment on a time-dependent outcome measured at the next time point [49]. An area for future work is to extend our results to the setting in which the aim is to estimate the joint effects of treatment at multiple time points. and to the time-dependent setting with time varying treatments and confounding variables. An additional factor to consider in the time-varying setting is the impact of stabilized vs unstabilized weights on the bias if both numerator and denominator of the stabilized weights involve conditioning on an error-prone covariate.

The bias expressions derived in this paper can be used to assess bias due to classification error in a dichotomous confounding variable. If classification error in confounding variables is suspected, a quantitative bias analysis provides an opportunity to quantitatively inform readers on the possible impact of such errors on causal conclusions.

References

[1] M. A. Hernán, B. A. Brumback, J. M. Robins, Estimating the causal effect of zidovudine on CD4 count with a marginal structural model for repeated measures, Statistics in Medicine 21 (12) (2002) 1689–1709. doi:10.1002/sim.1144.

[2] J. Robins, M. Hernán, B. Brumback, Marginal structural models and causal inference in Epidemiology, Epidemiology 11 (5) (2000) 550–560.

[3] R. Daniel, S. Cousens, B. De Stavola, M. G. Kenward, J. A. C. Sterne, Methods for dealing with time-dependent confounding, Statistics in Medicine 32 (9) (2013) 1584–1618. doi:10.1002/sim.5686.

[4] D. B. Rubin, For objective causal inference, design trumps analysis, The Annals of Applied Statistics 2 (3) (2008) 808–840. doi:10.1214/08-AOAS187.

[5] P. M. Steiner, T. D. Cook, W. R. Shadish, On the importance of reliable covariate measurement in selection bias adjustments using propensity scores, Journal of Educational and Behavioral Statistics 36 (2) (2011) 213–236. doi:10.3102/1076998610375835.

[6] K. B. Michels, A renaissance for measurement error, International Journal of Epidemiology 30 (3) (2001) 421–422. doi:10.1093/ije/30.3.421.

[7] M. van Smeden, T. L. Lash, R. H. H. Groenwold, Reflection on modern methods: Five myths about measurement error in epidemiological research, International Journal of Epidemiology 49 (1) (2020) 338–347. doi:10.1093/ije/dyz251.

[8] J. Kasza, K. R. Polkinghorne, M. R. Marshall, S. P. McDonald, R. Wolfe, Clustering and residual confounding in the application of marginal structural models: Dialysis modality, vascular access, and mortality, American Journal of Epidemiology 182 (6) (2015) 535–543. doi:10.1093/aje/kwv090.

[9] L. M. Bodnar, Marginal structural models for analyzing causal effects of time-dependent treatments: An application in perinatal epidemiology, American Journal of Epidemiology 159 (10) (2004) 926–934. doi:10.1093/aje/kwh131.

[10] M. E. Wood, K. Lapane, J. A. Frazier, E. Ystrom, E. O. Mick, H. Nordeng, Prenatal triptan exposure and internalising and externalising behaviour problems in 3-year-old children: Results from the Norwegian mother and child cohort study, Paediatric and Perinatal Epidemiology 30 (2) (2016) 190–200. doi:10.1111/ppe.12253.

[11] B. G. Armstrong, Effect of measurement error on epidemiological studies of environmental and occupational exposures, Occupational and Environmental Medicine 55 (10) (1998) 651–656. doi:10.1136/oem.55.10.651.

[12] J. Buonaccorsi, Measurement error: Models, methods, and applications, Chapman & Hall/CRC, Boca Raton, FL, 2010.

[13] R. Carroll, D. Ruppert, L. Stefanski, C. Crainiceanu, Measurement error in nonlinear models: A modern perspective, 2nd Edition, Chapman & Hall/CRC, Boca Raton, FL, 2006.

[14] P. Gustafson, Measurement error and misclassification in statistics and epidemiology: Impacts and Bayesian adjustments, Chapman & Hall/CRC, Boca Raton, FL, 2004.

[15] W. Fuller, Measurement error models, John Wiley & Sons, New York, NY, 1987.

[16] M. D. Regier, E. E. M. Moodie, R. W. Platt, The effect of error-in-confounders on the estimation of the causal parameter when using marginal structural models and inverse probability-of-treatment weights: A simulation study, The International Journal of Biostatistics 10 (1) (2014) 1–15. doi:10.1515/ijb-2012-0039.

[17] D. F. McCaffrey, J. R. Lockwood, C. M. Setodji, Inverse probability weighting with error-prone covariates, Biometrika 100 (3) (2013) 671–680. doi:10.1093/biomet/ast022.

[18] S. Greenland, Basic methods for sensitivity analysis of biases, International Journal of Epidemiology 25 (6) (1996) 1107–1116. doi:10.1093/ije/25.6.1107-a.

[19] T. Lash, M. Fox, A. Fink, Applying quantitative bias analysis to epidemiologic data, Springer, New York, NY, 2009.

[20] T. L. Lash, M. P. Fox, R. F. MacLehose, G. Maldonado, L. C. McCandless, S. Greenland, Good practices for quantitative bias analysis, International Journal of Epidemiology 43 (6) (2014) 1969–1985. doi:10.1093/ije/dyu149.

[21] M. S. Ali, R. H. H. Groenwold, S. V. Belitser, P. C. Souverein, E. Martín, N. M. Gatto, C. Huerta, H. Gardarsdottir, K. C. B. Roes, A. W. Hoes, A. de Boer, O. H. Klungel, Methodological comparison of marginal structural model, time-varying Cox regression, and propensity score methods: The example of antidepressant use and the risk of hip fracture, Pharmacoepidemiology and Drug Safety 25 (Suppl. 1) (2016) 114–121. doi:10.1002/pds.3864.

[22] L. A. Jackson, M. L. Jackson, J. C. Nelson, K. M. Neuzil, N. S. Weiss, Evidence of bias in estimates of influenza vaccine effectiveness in seniors, International Journal of Epidemiology 35 (2) (2006) 337–344. doi:10.1093/ije/dyi274.

[23] D. B. Rubin, Estimating causal effects of treatments in randomized and nonrandomized studies, Journal of Educational Psychology 66 (5) (1974) 688–701. doi:10.1037/h0037350.

[24] D. B. Rubin, Formal mode of statistical inference for causal effects, Journal of Statistical Planning and Inference 25 (3) (1990) 279–292. doi:10.1016/0378-3758(90)90077-8.

[25] M. Hernán, J. Robins, Causal inference: What if, Chapman & Hall/CRC, Boca Raton, 2020.

[26] D. B. Rubin, Randomization analysis of experimental data: The Fisher randomization test comment, Journal of the American Statistical Association 75 (371) (1980) 591. doi:10.2307/2287653.

[27] S. R. Cole, M. A. Hernan, Constructing inverse probability weights for marginal structural models, American Journal of Epidemiology 168 (6) (2008) 656–664. doi:10.1093/aje/kwn164.

[28] J. Robins, Marginal structural models versus structural nested models as tools for causal inference, in: E. Halloran, D. Berry (Eds.), Statistical models in epidemiology, the environment, and clinical trials, Springer-Verlag, New York, 2000, Ch. 2, pp. 95–133.

[29] T. VanderWeele, O. Arah, Bias formulas for sensitivity analysis of unmeasured confounding for general outcomes, treatments, and confounders, Epidemiology 22 (1) (2011) 42–52. doi:10.1097/EDE.0b013e3181f74493.

[30] P. M. Steiner, Y. Kim, The mechanics of omitted variable bias: Bias amplification and cancellation of offsetting biases, Journal of Causal Inference 4 (2) (2016). doi:10.1515/jci-2016-0009.

[31] R Core Team, R: A language and environment for statistical computing (2020). URL https://www.r-project.org/

[32] W. M. van der Wal, R. B. Geskus, Ipw : An R package for inverse probability weighting, Journal of Statistical Software 43 (13) (2011) 1–23. doi:10.18637/jss.v043.i13.

[33] T. Lumley, Analysis of complex survey samples, Journal of Statistical Software 9 (8) (2004) 1–19. doi:10.18637/jss.v009.i08.

[34] T. P. Morris, I. R. White, M. J. Crowther, Using simulation studies to evaluate statistical methods, Statistics in Medicine 38 (11) (2019) 2074–2102. doi:10.1002/sim.8086.

[35] A. Gasparini, Rsimsum: Summarise results from Monte Carlo simulation studies, Journal of Open Source Software 3 (26) (2018) 739. doi:10.21105/joss.00739.

[36] Centers for Disease Control and Prevention (CDC). National Center for Health Statistics (NCHS), National health and nutrition examination survey data (2011). URL https://wwwn.cdc.gov/nchs/nhanes/continuousnhanes/default.aspx?BeginYear=2011

[37] Centers for Disease Control and Prevention (CDC). National Center for Health Statistics (NCHS), National health and nutrition examination survey data (2013). URL https://wwwn.cdc.gov/nchs/nhanes/continuousnhanes/default.aspx?BeginYear=2013

[38] D. G. Altman, P. Royston, The cost of dichotomising continuous variables, BMJ 332 (7549) (2006) 1080. doi:10.1136/bmj.332.7549.1080.

[39] R. H. H. Groenwold, D. B. Nelson, K. L. Nichol, A. W. Hoes, E. Hak, Sensitivity analyses to estimate the potential impact of unmeasured confounding in causal research, International Journal of Epidemiology 39 (1) (2010) 107–117. doi:10.1093/ije/dyp332.

[40] P. Ding, T. J. VanderWeele, Sensitivity analysis without assumptions, Epidemiology 27 (3) (2016) 368–377. doi:10.1097/EDE.0000000000000457.

[41] J. Pearl, On measurement bias in causal inference, in: P. Grunwald P, Spirtes (Ed.), Proceedings of uncertainty in artificial intelligence, AUAI, AUAI, Corvallis, OR, 2010, pp. 425–432.

[42] S. R. Cole, L. P. Jacobson, P. C. Tien, L. Kingsley, J. S. Chmiel, K. Anastos, Using marginal structural measurement-error models to estimate the long-term effect of antiretroviral therapy on incident AIDS or death, American Journal of Epidemiology 171 (1) (2010) 113–122. doi:10.1093/aje/kwp329.

[43] R. Kyle, E. Moodie, M. Klein, M. Abrahamowicz, Correcting for measurement error in time-varying covariates in marginal structural models, American Journal of Epidemiology 184 (3) (2016) 249–258. doi:10.1093/aje/kww068.

[44] J. Lockwood, D. F. McCaffrey, Simulation-extrapolation for estimating means and causal effects with mismeasured covariates, Observational Studies 1 (2) (2015) 241–290. doi:10.1353/obs.2015.0007.

[45] Y. Webb-Vargas, K. Rudolph, D. Lenis, P. Murakami, E. Stuart, An imputation-based solution to using mismeasured covariates in propensity score analysis, Statistical Methods in Medical Research 26 (4) (2017) 1824–1837. doi:10.1177/0962280215588771.

[46] M. Babanezhad, S. Vansteelandt, E. Goetghebeur, Comparison of causal effect estimators under exposure misclassification, Journal of Statistical Planning and Inference 140 (5) (2010) 1306–1319. doi:10.1016/j.jspi.2009.11.015.

[47] D. Shu, G. Y. Yi, Causal inference with measurement error in outcomes: Bias analysis and estimation methods, Statistical Methods in Medical Research 28 (7) (2019) 2049–2068. doi:10.1177/0962280217743777.

[48] R. H. Keogh, I. R. White, A toolkit for measurement error correction, with a focus on nutritional epidemiology, Statistics in Medicine 33 (12) (2014) 2137–2155. doi:10.1002/sim.6095.

[49] R. H. Keogh, R. M. Daniel, T. J. VanderWeele, S. Vansteelandt, Analysis of longitudinal studies with repeated outcome measures: Adjusting for time-dependent confounding using conventional methods, American Journal of Epidemiology 187 (5) (2018) 1085–1092. doi:10.1093/aje/kwx311.

9

Summary and general discussion

9.1. Summary

Measurement error is common in epidemiologic research and may affect the validity of research results. It is therefore important to scrutinise the effects of measurement error in epidemiologic research. Even simple forms of measurement error, for instance random measurement error in an exposure, can introduce bias in exposure-outcome associations. And even though there are situations in which measurement error does not introduce bias in the exposure-outcome association, for instance in case of random measurement error in a continuous outcome, it nearly always affects the precision and power of a study. In addition, other forms of measurement error, for example systematic measurement error or differential measurement error in an exposure, covariate or outcome, can affect exposure-outcome associations in complex ways that may not easily be anticipated. Adjusting for measurement error using measurement error correction methods may thus be necessary to obtain reliable estimates of exposure-outcome associations.

To facilitate measurement error correction, information about the underlying measurement error mechanism (i.e., model) and its parameters is needed. The measurement error model can sometimes be estimated from internal or external validation data, replicates data or calibration data. Collection and the use of such measurement error mechanism data will likely improve the quality of epidemiologic analyses in the presence of measurement error. This can be done through the application of measurement error correction methods, which adjust the analyses taking into account the information from the measurement error mechanism. Alternatively, in the absence of concrete data about the mechanisms or the parameters of measurement error, sensitivity analysis for measurement error can be used, in which the impact on the epidemiologic analyses of one or a range of hypothesized measurement error mechanisms or their parameters can be investigated.

The studies described in the thesis were set out to improve the *understanding* of the impact of measurement error, to facilitate the *application* of measurement error correction methods, to improve the *design* of epidemiologic studies when measurement error in a variable is suspected and, to develop *tools* to quantitatively assess the impact of measurement error in epidemiologic research.

In Chapter 2, consequences were studied of measurement error in a continuous outcome in a randomized trial. Using an example of the efficacy of a low-dose iron supplement on haemoglobin levels in pregnant women, different forms of measurement error were discussed (i.e., random, systematic and differential measurement error). Using the example trial, it was shown that random measurement error in a trial outcome does not lead to bias in the effect estimator but can lead to a reduced precision and power. It was shown that systematic measurement error and differential measurement error in an outcome can lead to bias in the effect estimator and consequently, a null-hypothesis significance test for the treatment effect can deviate substantially from the nominal level. Subsequently, a regression calibration-like method was proposed to reduce bias in the treatment effect estimator and obtain confidence intervals with nominal coverage and tested in a Monte Carlo simulation study. The proposed method made use of external validation data to estimate the measurement error model and its parameters and four different methods for confidence interval construction were proposed. Different parameters for the measurement error model (i.e., systematic and differential measurement error) and explained variance of the measurement error model were tested. In our simulation study, it was shown that the regression calibration-like method was effective in improving trial inferences when an external validation dataset with at least 15 subjects was available.

In Chapter 3 the `R` package `mecor` for measurement error correction was introduced. The package facilitates measurement error correction in linear models with a continuous outcome if there is measurement error in the outcome or in a continuous covariate. The package accommodates measurement error correction methodology for a wide range of data structures: internal and external validation studies, replicates studies, and calibration studies. Various measurement error correction methods were implemented in the package: regression calibration, method of moments and correction based on maximum likelihood estimation. For standard error estimation and construction of confidence intervals, the delta method and bootstrap were implemented for all methods. The package also facilitates sensitivity analysis, when no data are available to estimate the parameters of the measurement error model. The package contains synthetic data based on examples from epidemiology following the structure of internal validation data, replicates data, calibration data and external validation data.

In Chapter 4 settings were studied in which application of regression calibration for exposure measurement error correction may not be appropriate. This was illustrated in a study of the association between active energy expenditure and lean body mass. A simulation study, based on the case study, showed that particularly in small samples the regression calibration estimator may be less efficient in terms of mean squared error than an estimator not correcting for the exposure measurement error. This phenomenon is an example of the commonly known bias–variance trade off. Particularly, when the measurement error is relatively large and sample sizes small, the simulation study showed that the performance of regression calibration was poor, indicated by biased estimates, large mean squared errors and large empirical standard errors in these settings.

In Chapter 5 three internal validation sampling strategies (i.e., random, stratified random and extremes sampling) were investigated in conjunction with regression calibration to correct for measurement error in a continuous covariate. This was illustrated in an example study of the investigation of the association between visceral adipose tissue and insulin resistance. The exposure measure visceral adipose tissue was only available in

40% of the population. Waist circumference was measured in all individuals and assumed an error-prone substitute measure of the reference measure visceral adipose tissue. In a setting where the reference measure is obtained in only 40% of the whole study, it was studied which individuals should be included in that subset and which not by means of Monte Carlo simulation. The simulation study showed a small efficiency gain in terms of mean squared error of stratified random and extremes sampling over a random sampling strategy for the internal validation restricted and regression calibration analyses, but only when measurement error was non-differential. For regression calibration, this gain in efficiency was at the cost of higher percentages bias and lower confidence interval coverage. It was therefore recommended that, in general, regression calibration using randomly sampled validation samples are preferred over stratified or extremes sampled samples.

The study described in Chapter 6 showed that studies on venous thromboembolism (VTE) incidence in Coronavirus disease 2019 (COVID-19) patients report highly heterogeneous results. Different sources of the observed heterogeneity were identified, notably, clinical and methodological sources, and illustrated using various examples. Clinical sources included the characteristics of study participants and testing for VTE. Methodological sources included inclusion types of the VTE endpoint, data quality and data analysis. Careful description was recommended of the elements that potentially affect VTE incidence and thus may cause heterogeneity in future VTE incidence studies and guidance was provided in the form of a list with reporting recommendations.

In Chapter 7 regression calibration and simulation-extrapolation were compared for sensitivity analysis for random measurement error in an exposure variable. These two random exposure measurement error correction methods were illustrated in two example studies. The first example study investigated the relation between the exposure blood pressure and , and the second example study investigated the relation between the exposure sodium intake and hypertension. These relations were modelled using linear and logistic regression, respectively. In both example studies the exposure variable was an error-prone version of an error-free exposure variable. Based on these two examples, a simulation study was conducted to study the relative performance of regression calibration and simulation-extrapolation in linear and logistic regression models. The simulation study showed that without extra data, but with correct assumptions about the variance of the measurement error, regression calibration was generally unbiased for linear and logistic regression, while simulation-extrapolation was biased. A small gain in efficiency in terms of mean squared error was seen for simulation-extrapolation in linear regression but not for logistic regression. The use of regression calibration for sensitivity analysis for random exposure measurement error was recommended and its use illustrated in the example study of the association between blood pressure and kidney function.

Inverse probability weighting and conditional models are both important and frequently used tools to adjust for confounding variables in observational studies. In Chapter 8, expressions were derived for the bias in the average treatment effect in a marginal structural model estimated using inverse probability weighting and a conditional model when a confounding variable is measured with error. Compared to bias in the average treatment effect estimator from a conditional model, the bias in a marginal structural model estimated using inverse probability weighting can be different in magnitude but is equal in sign. The derived bias expressions informed a quantitative bias analysis for bias due to a misclassified confounding variable. The use of a quantitative bias analysis was demonstrated in an

example study of the effect of using diuretics versus beta-blockers on blood pressure adjusted for the error-prone confounding variable self-reported body mass index category.

9.2. Discussion

This thesis provides an overview of correction methods for measurement error in epidemiologic research. The studies described in the thesis were set out to improve the *understanding* of the impact of measurement error and to facilitate the *application* of measurement error correction methods in epidemiologic studies. Guidance was provided to improve the *design* of epidemiologic studies when measurement error is suspected, and reporting guidelines proposed. All methods were demonstrated in case studies using empirical data (for an overview of case studies, see Table 9.1). Special attention was paid to sensitivity analysis for measurement error in settings where measurement error is suspected, but data about measurement error structure and its parameters, essential for measurement error correction methods, were not available. Here, we discuss the contribution of our work to this field and set out directions for future research.

9.2.1. Impact of measurement error in epidemiologic studies

The impact of measurement error often goes beyond the simple heuristic of 'attenuation to the null' [1]. This heuristic wrongfully suggests that estimates of effects in epidemiologic studies will only become smaller due to the measurement error. Unfortunately, this myth remains persistent despite a vast body of literature arguing against it [2–5]. Particularly, depending on the target of the analysis and the type of measurement error, the effects of measurement error can go in either direction and are therefore often unpredictable, as shown by Keogh et al. [6].

This thesis aimed at improving the *understanding* of the impact of measurement error in epidemiologic research. To evaluate the impact of measurement error in a specific study, four considerations are; i) what statistical model is used; ii) which of the variable(s) of the model is (are) error-prone and what is their role in the model; iii) what is the structure of the measurement error model; and iv) what are the parameters of the measurement error model (see Figure 9.1). All these components may affect *if* an epidemiologic study is affected by measurement error and if so, *how* an epidemiologic study is affected by measurement error. For example, random exposure measurement error introduces bias in the effect estimator of a linear regression model [4], and a logistic regression model [7] and leads to a so-called 'induced hazard function' for a Cox regression model [8]. In contrast, random measurement error in a continuous outcome does not introduce bias but reduces precision and power at a chosen sample size, and systematic and differential measurement error in such outcomes introduce bias in the effect estimator of a linear regression model that can go in either direction (Chapter 2). When exposure measurement error is suspected, restricting the analysis to the subset of individuals for whom the error-free exposure measurement is obtained, does not lead to biased inference. Yet, when that subset is sampled using information about an error-prone substitute exposure (e.g., when for all individuals exceeding a specific threshold of the substitute exposure, the error-free exposure is obtained), bias is introduced in the complete case analysis if the error in the substitute exposure is differential, but not if the error in the substitute exposure is non-differential (Chapter 5). When a confounding variable is misclassified,

marginal structural models estimated using inverse probability weighting were shown to be biased but affected differently than conditional models (Chapter 8). There are innumerable combinations of the considerations displayed in Figure 9.1 and, therefore, measurement error can affect estimated exposure-outcome associations in complex ways that may not easily be anticipated and need to be evaluated from one setting to another.

9.2.2. Measurement error correction methods in epidemiologic studies

There is an abundance of texts on measurement error correction methods [2–5]. Yet, correction methods remain seldomly applied in epidemiologic research [9–11]. Methods for measurement error corrections include, regression calibration [12, 13], simulation-extrapolation [14], moment reconstruction [15], non-parametric maximum likelihood estimation [16], imputation-based methods [17, 18] and Bayesian methods [5, 19]. Regression calibration is among the most commonly used methods in epidemiologic research [10, 11].

This thesis facilitated the *application* of measurement error correction in epidemiologic research with the development of the software package `mecor` for measurement error correction in linear models with a continuous outcome. In this software package for `R`, regression calibration [20], validation regression calibration, efficient regression calibration [21], method of moments [2] and maximum likelihood-based methods [22] were implemented for a wide range of validation data structures (Table 9.1). Notably, different methods for variance estimation of the corrected estimators were implemented in `mecor`. An informed choice for the variance estimation of the measurement error corrected estimators is important as was shown that the Zero Variance, Delta, Fieller and bootstrap methods had different performance in terms of coverage and average confidence interval width (Chapter 2 and 4). The methods implemented in `mecor` are consistent but not necessarily more statistically efficient than the uncorrected estimator nor unbiased. Particularly in small samples, the estimator not correcting for measurement error may be more efficient in terms of mean squared error compared to the regression calibration estimator (Chapter 4). A phenomenon referred to as the bias–variance trade off. Particularly when measurement error is relatively large, the performance of regression calibration can be poor in small samples, as was shown by high percentages bias and large mean squared errors in these settings. However, compared to regression calibration, the simulation-extrapolation estimator was even more prone to bias (Chapter 7). Regression calibration relies on the assumption of non-differential measurement error, and large biases can occur in the estimator if this assumption is not warranted, as was shown in Chapter 5. In conclusion, measurement error correction methods can correct for measurement error when extra data are available to estimate the measurement error model and its parameters provided sufficiently large sample size of the validation set and measurement error that is not extremely large. What constitutes 'sufficiently large' and 'not extremely large' will be study specific and can be informed by statistical simulation studies, as presented in Chapter 4.

9.2.3. Design of epidemiologic studies affected by measurement error

For measurement error correction, validation data are needed to estimate the measurement error model and its parameters. Collection of such data should preferably be included in the *design* of an epidemiologic study. Considerations include the data structure, size and the

Table 9.1: Empirical example case studies used for investigations of measurement error correction methods and sensitivity analysis in the thesis

Example (Chapter)	**Measurement Error**	**Data Structure**	**Availability in `mecor`**	**Solutions Investigated in the Thesis**
Low-dose iron suppletion and haemoglobin levels (Chapter 2)	The reference measure venous haemoglobin level was substituted by the error-prone measure of capillary haemoglobin level	External validation data	Available as dataset `haemoglobin` and `haemoglobin_ext`	• Method of moments (referred to as regression calibration in Chapter 2)
Dietary intervention and sodium intake (Chapter 3, based on an example from Keogh et al. [23])	Sodium intake measured in urine was assumed subject to random measurement error and repeatedly measured, and sodium intake measured by a questionnaire was assumed subject to systematic measurement error	Calibration data	Available as dataset `sodium`	• Method of moments • Efficient method of moments
Active energy expenditure and lean body mass (Chapter 4)	The reference measure active energy expenditure measured by Actiheart® was substituted by the error-prone measure based on the international physical activity questionnaire	Internal validation data	Not available	• Regression calibration
Visceral adipose tissue and insulin resistance (Chapter 5)	The reference measure visceral adipose tissue was substituted by the error-prone measure waist circumference in a subset of the study	Internal validation data	Available as dataset `vat`	• Regression calibration • Efficient regression calibration • Validation regression calibration • Internal validation restricted analysis
Blood pressure and kidney function (Chapter 7)	Blood pressure was assumed subject to random measurement error and measured repeatedly	Replicates data or sensitivity analysis	Available as dataset `bloodpressure`	• Regression calibration • Simulation-extrapolation Informed by the replicate measures or in a sensitivity
Sodium intake and Hypertension (Chapter 7)	Sodium intake was assumed subject to random measurement error and measured repeatedly	Replicates data or sensitivity analysis	Not available	• Regression calibration • Simulation-extrapolation Informed by the replicate measures or in a sensitivity analysis
Diuretics vs beta-blocker use and systolic blood pressure (Chapter 8)	The confounding variable self-reported body mass index was assumed subject to classification error	Sensitivity analysis	Not available	• Sensitivity analysis informed by the bias expressions derived in Chapter 8 visualised in a Shiny application

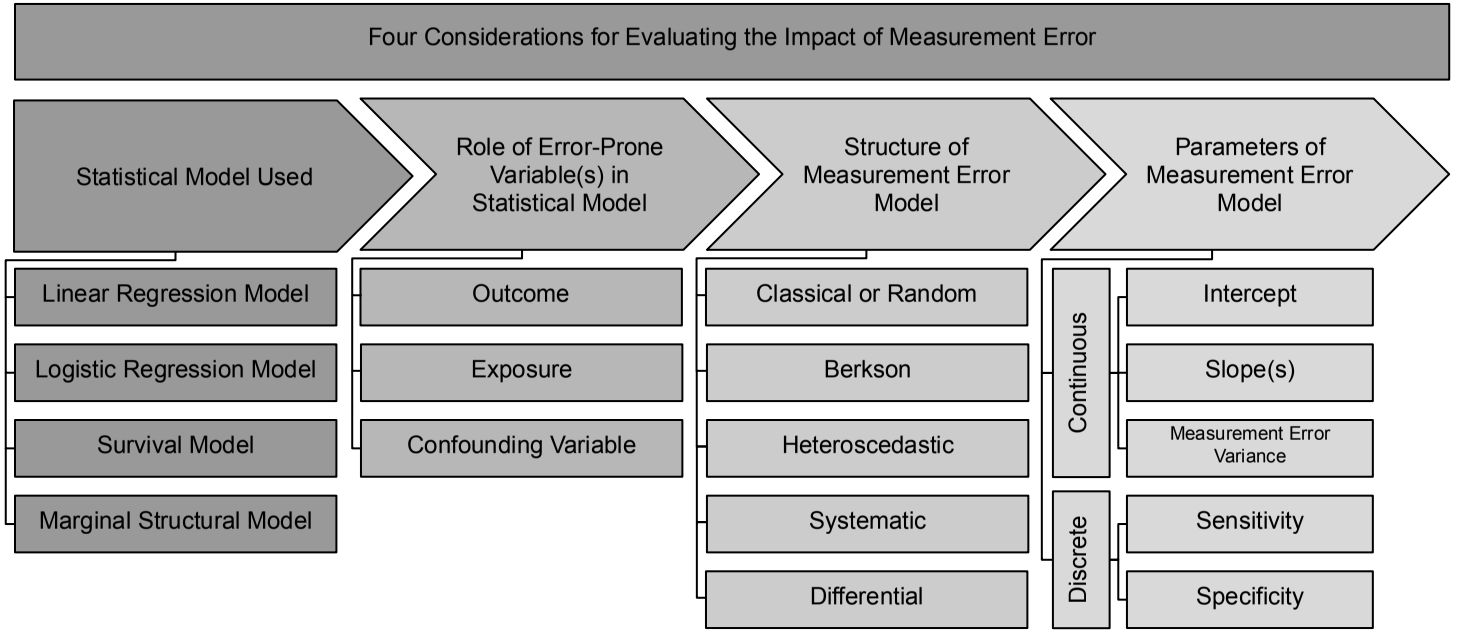


Figure 9.1: Four considerations for evaluating the impact of measurement error in an epidemiologic study. Focus lies on statistical models commonly used in epidemiologic research, measurement error model structures generally distinguished in the measurement error literature and, prevailing operationalisations of the parameters of the measurement error model.

sampling strategy of the validation data. For such considerations, the different components shown in Figure 9.1 needed to evaluate the impact of measurement error need to be taken into account. Particularly, certain structures of validation data (internal, external, replicates data or calibration data) are not suited for certain measurement error structures (e.g., replicates data can only be used if random measurement error is suspected). After making assumptions about the measurement error model structure and its parameters and deciding what type of validation data is suited, Monte Carlo simulation can be used to inform sample size and sampling strategy of the validation data. An example of a Monte Carlo simulation study to examine the optimal sampling strategy of an internal validation data set in the Netherlands Epidemiology of Obesity study [24] was described in Chapter 5. Here, sampling the extremes or stratified randomly showed a small gain in efficiency, but at the cost of bias and confidence interval coverage and should only be used when measurement error is strictly non-differential. A difficulty here is, however, that in studies like the Netherlands Epidemiology of Obesity study, the first two components that influence the impact of measurement error (described in first two columns in Figure 9.1) may differ across studies. Specifically, a variable can be an outcome in one study and an exposure in another study.

9.2.4. Sensitivity analysis for measurement error in epidemiologic studies

In epidemiologic research, it is commonly assumed (often implicitly) that all variables are measured without error; an assumption that is often not justified. Yet, when measurement error is suspected or anticipated, methods to correct for the measurement error rely on the availability of data on the measurement error mechanisms and parameters. Such data may not be available, maybe incomplete or be itself unreliable, in which case sensitivity analysis for measurement error can help to assess the sensitivity of research results to measurement error. In epidemiology, a sensitivity analysis may alternatively be referred to as quantitative bias analysis [25].

Sensitivity analysis for measurement error should be included in study protocols and valued independent of the outcome of the sensitivity analysis (i.e., results should not only be shown if the sensitivity analysis shows research results are *not* sensitive to the assumption of no measurement error). Sensitivity analysis can be informed by expert knowledge about the structure of the measurement error model and its parameters. Distributions of these parameters can be used to put more weight on the assumed most plausible values [25].

The sensitivity of research results to random exposure measurement error can be checked using regression calibration or simulation-extrapolation, of which regression calibration was shown most suited in Chapter 7. Graphical presentation of the results of a sensitivity analysis allows readers to judge the sensitivity of research results for the whole distribution of assumed parameters of the measurement error model, and may be preferred over a single summary number (see for example Figure 7.9 in Chapter 7). Alternatively, interactive *tools* may be designed to allow readers to test the sensitivity of research results to their own assumed parameters of the measurement error model, as was facilitated by the Shiny application demonstrated in Chapter 8.

9.2.5. Future research

The studies presented in the thesis aimed to improve the (application of) methods to limit the impact of measurement error in epidemiologic research. The application of

measurement error correction methods was facilitated through the development of the `R` package `mecor`. To aid the application of measurement error correction methods in epidemiology, numerous methods were illustrated in empirical data (Table 9.1). Extensive Monte Carlo studies were set up to study the performance of measurement error correction methods in epidemiologic studies based on the empirical data and have been made publicly available. The simulation code can easily be adapted by researchers to settings of intended use to improve the design and statistical analysis of epidemiologic studies when measurement error is suspected. However, we are not there yet. There are several topics that require future research to further develop the field of measurement error methodology.

First, the main focus of this thesis was on linear models with a continuous outcome and measurement error in one of the continuous variables of those models. In epidemiologic studies, measurement error may, however, be anticipated in more than one variable. In addition, other statistical models (e.g., logistic and survival analysis) are commonly used in epidemiologic research. For models with binary outcomes, the impact of covariate measurement error and classification error in the binary outcome has been studied by Carroll et al. in [7] and [26], respectively. Also, correction methods have been proposed for situations where one or multiple variables in a logistic regression model are measured with error [20]. For survival outcomes, the impact of covariate measurement error has been studied by Prentice et al. [8] and an investigation of measurement errors in the failure time outcome and correction methods for this setting were examined by Oh et al. [27]. Yet, the implications of a combination of complex forms of outcome measurement error and covariate measurement error need further study.

Second, this thesis only investigated the use of parametric measurement error models and it was generally assumed that the measurement error model was well specified. Future research may examine methods to test for the structure of the measurement error model in empirical data and study the impact of misspecification of the measurement error model structure on measurement error correction methods.

Third, the validation data structures discussed in the thesis that aid measurement error correction methods rely on certain assumptions. For an external data set, it is assumed that the measurement error model and its parameters are transportable from the main study to the external study. For a replicates study, it is assumed that measurement error in the subsequent replicate measurements is independent. Investigations are needed if information about the reliability of e.g., biomarkers can be transported to studies where these biomarkers are used and if the assumption of independent measurement error in such biomarkers is warranted.

Fourth, this thesis presents measurement error correction methods for measures of which a clear concept about the 'true' measure of a variable is needed and is in most instances assumed observable (except when random measurement is assumed in which case repeated measures of the error-prone measure are adequate). This assumption might be reasonable and applicable for measures such as an individual's weight in kilo grams or blood pressure, but may be difficult or even impossible to establish for constructs such as patient well-being or pain [28]. Future research may pay specific attention to the applicability of latent class analysis for the analysis of error-prone epidemiologic data, which does not rely on the assumption of observable 'true' measures. Instead, it is assumed that the true variable can be estimated by combining multiple imperfect measurements of the variable. These methods are widespread in psychology and the social sciences [29],

but received relatively little attention in epidemiologic research (exceptions include e.g., [30, 31]).

9.2.6. Conclusion

Measurement error in epidemiologic research is not uncommon and can hamper the validity of research results if ignored. The old saying "to prevent is better than to cure" also applies here, and therefore actions to improve the overall quality of measurement in epidemiologic analyses are likely to have a larger effect on the validity of epidemiologic studies than widespread application of measurement error correction methods. However, in settings where measurement error cannot be avoided, measurement error correction methods and sensitivity analysis for measurement error provide tools to correct for or quantitatively assess the impact of measurement error. In combination with reliable information about the measurement error model and its parameters, these methods can help to estimate relevant epidemiologic parameters that are more reliable than what would be obtained if estimated without taking account of possible measurement error.

References

[1] M. van Smeden, T. L. Lash, R. H. H. Groenwold, Reflection on modern methods: Five myths about measurement error in epidemiological research, International Journal of Epidemiology 49 (1) (2020) 338–347. doi:10.1093/ije/dyz251.

[2] J. Buonaccorsi, Measurement error: Models, methods, and applications, Chapman & Hall/CRC, Boca Raton, FL, 2010.

[3] R. Carroll, D. Ruppert, L. Stefanski, C. Crainiceanu, Measurement error in nonlinear models: A modern perspective, 2nd Edition, Chapman & Hall/CRC, Boca Raton, FL, 2006.

[4] W. Fuller, Measurement error models, John Wiley & Sons, New York, NY, 1987.

[5] P. Gustafson, Measurement error and misclassification in statistics and epidemiology: Impacts and Bayesian adjustments, Chapman & Hall/CRC, Boca Raton, FL, 2004.

[6] R. H. Keogh, P. A. Shaw, P. Gustafson, R. J. Carroll, V. Deffner, K. W. Dodd, H. Küchenhoff, J. A. Tooze, M. P. Wallace, V. Kipnis, L. S. Freedman, STRATOS guidance document on measurement error and misclassification of variables in observational epidemiology: Part 1—Basic theory and simple methods of adjustment, Statistics in Medicine 39 (16) (2020) 2197–2231. doi:10.1002/sim.8532.

[7] R. J. Carroll, C. H. Spiegelman, K. K. G. Lan, K. T. Bailey, R. D. Abbott, On errors-in-variables for binary regression models, Biometrika 71 (1) (1984) 19. doi:10.2307/2336392.

[8] R. L. Prentice, Covariate measurement errors and parameter estimation in a failure time regression model, Biometrika 69 (2) (1982) 331–342. doi:10.2307/2335407.

[9] A. M. Jurek, G. Maldonado, S. Greenland, T. R. Church, Exposure-measurement error is frequently ignored when interpreting epidemiologic study results, European Journal of Epidemiology 21 (12) (2007) 871–876. doi:10.1007/s10654-006-9083-0.

[10] T. B. Brakenhoff, M. Mitroiu, R. H. Keogh, K. G. M. Moons, R. H. H. Groenwold, M. van Smeden, Measurement error is often neglected in medical literature: A systematic review, Journal of Clinical Epidemiology 98 (2018) 89–97. doi:10.1016/j.jclinepi.2018.02.023.

[11] P. A. Shaw, V. Deffner, R. H. Keogh, J. A. Tooze, K. W. Dodd, H. Küchenhoff, V. Kipnis, L. S. Freedman, Epidemiologic analyses with error-prone exposures: Review of current practice and recommendations, Annals of Epidemiology 28 (11) (2018) 821–828. doi:10.1016/j.annepidem.2018.09.001.

[12] R. J. Carroll, L. A. Stefanski, Approximate quasi-likelihood estimation in models with surrogate predictors, Journal of the American Statistical Association 85 (411) (1990) 652–663. doi:10.1080/01621459.1990.10474925.

[13] L. Gleser, Improvements of the naive approach to estimation in nonlinear errors-in-variables regression models, in: P. Brown, W. Fuller (Eds.), Statistical analysis of measurement error models, American Mathematics Society, Providence, 1990, pp. 99–114.

[14] J. R. Cook, L. A. Stefanski, Simulation-extrapolation estimation in parametric measurement error models, Journal of the American Statistical Association 89 (428) (1994) 1314–1328. doi:10.2307/2290994.

[15] L. S. Freedman, V. Fainberg, V. Kipnis, D. Midthune, R. J. Carroll, A new method for dealing with measurement error in explanatory variables of regression models, Biometrics 60 (1) (2004) 172–181. doi:10.1111/j.0006-341X.2004.00164.x.

[16] S. Rabe-Hesketh, A. Pickles, A. Skrondal, Correcting for covariate measurement error in logistic regression using nonparametric maximum likelihood estimation, Statistical Modelling 3 (3) (2003) 215–232. doi:10.1191/1471082X03st056oa.

[17] S. R. Cole, H. Chu, S. Greenland, Multiple-imputation for measurement-error correction, International Journal of Epidemiology 35 (4) (2006) 1074–1081. doi:10.1093/ije/dyl097.

[18] M. Blackwell, J. Honaker, G. King, A unified approach to measurement error and missing data: Overview and applications, Sociological Methods & Research 46 (3) (2017) 303–341. doi:10.1177/0049124115585360.

[19] J. W. Bartlett, R. H. Keogh, Bayesian correction for covariate measurement error: A frequentist evaluation and comparison with regression calibration, Statistical Methods in Medical Research 27 (6) (2018) 1695–1708. doi:10.1177/0962280216667764.

[20] B. Rosner, D. Spiegelman, W. C. Willett, Correction of logistic regression relative risk estimates and confidence intervals for measurement error: The case of multiple covariates measured with error, American Journal of Epidemiology 132 (4) (1990) 734–745. doi:10.1093/oxfordjournals.aje.a115715.

[21] D. Spiegelman, R. J. Carroll, V. Kipnis, Efficient regression calibration for logistic regression in main study/internal validation study designs with an imperfect reference instrument, Statistics in Medicine 20 (1) (2001) 139–160. doi:10.1002/1097-0258(20010115)20:1<139::AID-SIM644>3.0.CO;2-K.

[22] J. W. Bartlett, B. L. De Stavola, C. Frost, Linear mixed models for replication data to efficiently allow for covariate measurement error, Statistics in Medicine 28 (25) (2009) 3158–3178. doi:10.1002/sim.3713.

[23] R. H. Keogh, R. J. Carroll, J. A. Tooze, S. I. Kirkpatrick, L. S. Freedman, Statistical issues related to dietary intake as the response variable in intervention trials, Statistics in Medicine 35 (25) (2016) 4493–4508. doi:10.1002/sim.7011.

[24] R. de Mutsert, M. den Heijer, T. J. Rabelink, J. W. A. Smit, J. A. Romijn, J. W. Jukema, A. de Roos, C. M. Cobbaert, M. Kloppenburg, S. le Cessie, S. Middeldorp,

F. R. Rosendaal, The Netherlands epidemiology of obesity (NEO) study: Study design and data collection, European Journal of Epidemiology 28 (6) (2013) 513–523. doi:10.1007/s10654-013-9801-3.

[25] T. Lash, M. Fox, A. Fink, Applying quantitative bias analysis to epidemiologic data, Springer, New York, NY, 2009.

[26] R. J. Carroll, D. Ruppert, L. A. Stefanski, C. M. Crainiceanu, Logistic regression with response error, in: Measurement error in nonlinear models, 2nd Edition, Chapman & Hall/CRC, Boca Raton, FL, 2006, Ch. 15, pp. 345–352.

[27] E. J. Oh, B. E. Shepherd, T. Lumley, P. A. Shaw, Considerations for analysis of time-to-event outcomes measured with error: Bias and correction with SIMEX, Statistics in Medicine 37 (8) (2018) 1276–1289. doi:10.1002/sim.7554.

[28] D. J. Hand, Measurement: A very short introduction, Oxford University Press, New York, NY, 2016.

[29] K. A. Bollen, Latent variables in psychology and the social sciences, Annual Review of Psychology 53 (1) (2002) 605–634. doi:10.1146/annurev.psych.53.100901.135239.

[30] J. Kaldor, D. Clayton, Latent class analysis in chronic disease epidemiology, Statistics in Medicine 4 (3) (1985) 327–335. doi:10.1002/sim.4780040312.

[31] M. van Smeden, C. A. Naaktgeboren, J. B. Reitsma, K. G. M. Moons, J. A. H. de Groot, Latent class models in diagnostic studies when there is no reference standard: A systematic review, American Journal of Epidemiology 179 (4) (2014) 423–431. doi:10.1093/aje/kwt286.

S2

Supplementary material Chapter 2

These are the supplementary materials accompanying Chapter 2. The supplementary materials are structured as follows. In section S2.1 we discuss two more example trials for illustration of measurement error in an endpoint. In section S2.2 we explain why and under which assumptions ignoring measurement error will lead to incorrect inference. Section S2.3 provides an explanation of corrected effect estimators (and why these are consistent) and explains the methods used for confidence interval estimation. In section S2.4 a proof is given that measurement error depending on prognostic factors does not introduce bias in the treatment effect estimator. In section S2.5 an approximation for the bias and variance of the corrected estimator is derived.

S2.1. Illustrative examples

We introduce here two additional example trials from literature, hypothesize that these trial could also have used endpoints measured with error to illustrate how the use of an endpoint that is contaminated with error would affect trial inference. We assume that the original endpoints used in our example trials are measurement error free.

S2.1.1. Example trial 2: energy expenditure

Poehlman and colleagues [1] studied the effects of endurance and resistance training on total daily energy expenditure in a randomised trial of young sedentary women. Participants were randomized to one of three six-month during exercise programmes: endurance training, resistance training or the control arm. Some controversy regarding the effect of exercise training on total energy expenditure (TEE) existed at the time of the start of the trial, partly because of the difficulty to assess daily energy expenditure [1]. Starting 72 hours after completion of the training program, TEE of the participants was measured by doubly labelled water during a ten day period, which is considered the gold standard in measuring energy expenditure in humans [2]. In short, the study found no evidence for an effect of resistance and endurance training (compared to placebo) on total energy expenditure. Post-trial, measured TEE was higher in the control arm than in the two intervention arms. Table 1 shows the decrease in TEE of the women exposed to the

existence training programme versus the placebo arm.

S2.1.2. Example trial 3: rheumatoid arthritis disease activity

The U-Act-Early trial tested the efficacy of a new treatment strategy for rheumatoid arthritis (RA) in patients with newly diagnosed RA [3] in a three-arm trial: tocilizumab plus methotrexate versus tocilizumab only versus methotrexate only, all as initial treatment. For endpoint assessment, this trial used a validated RA disease activity measure (the Disease Activity Score 28, DAS28) [4]) which is commonly used and recommended to measure endpoints in RA clinical trials [5, 6]. In short, the trial showed that immediate initiation of tocilizumab with or without methotrexate is more effective than methotrexate alone to achieve sustained remission in newly diagnosed RA patients. The difference in mean DAS28 score in the tocilizumab plus methotrexate versus methotrexate only group after 24 weeks is shown in Table S2.1. The sample size of the former groups reported in Table S2.1 is based on measurements available at 24 weeks of follow up.

A common alternative approach to measure energy expenditure (example trial 2) is by a accelerometer, that measures body movement via motion sensors to assess energy expenditure (e.g. [2]). As compared to double labelled water (example trial 2), the accelerometer is cheaper, but less accurate [2]. Lastly, instead of endpoint assessment by DAS28 (example trial 3), where assessment is done by trained medical staff [4], trials could alternatively use the patient-based RA disease activity score (PDAS), where endpoint assessment is done by the patient [7].

For the example trial in the paper and each of the aforementioned example trials here, in Table S2.1 we show to what extent the Type-II of a test for treatment effect changes when a hypothetical lower standard of endpoint measurement would have been used introducing classical measurement error. The table clearly shows the anticipated increase in Type-II error with increasing error at the same sample size.

S2.2. Measurement error structures

Consider a two-arm randomized controlled trial that compares the effects of two treatments ($X \in \{0, 1\}$), where 0 may represent a placebo treatment or an active comparator. Let Y denote the true (or preferred) trial endpoint and Y^* an error prone operationalisation of Y. We will assume that both Y and Y^* are measured on a continuous scale. Throughout, we assume that Y^* is measured for all $i = 1, \ldots, N$ randomly allocated patients in the trial. We assume that the effect of allocated treatment ($X \in \{0, 1\}$) on preferred endpoint Y is defined by the linear model

$$Y = \alpha_Y + \beta_Y X + \varepsilon, \tag{S2.1}$$

where β_Y defines the treatment effect on the endpoint, and ε has expected mean 0 and variance σ^2. Throughout, we assume that X is fixed. Further, we assume that model S2.1 is inestimable from the observed data because the endpoint Y^* instead of Y was measured. We will assume that the relation between Y and Y^* is given by a linear model,

$$Y^* = \theta_0 + \theta_1 Y + e, \tag{S2.2}$$

where e is a random variable whose distribution is independent of ε, Y and X. The parameters θ_0 and θ_1 define the relation between Y and Y^*, where it is assumed that θ_1

Table S2.1: Impact of classical measurement error on Type-II error in the three example trials. Effect estimates, standard errors and sample sizes are based on results in the papers by Makridis et al. [8] (trial 1), Poehlman et al. [1] (trial 2) and Bijlsma et al. [3] (trial 3)

Example	Effect Estimate	Standard Error	Sample Size	ρ[a]	Type-II Error[b]
Trial 1	6.9	1.27	393	0	-
		2.43	108	0	20%
		2.71	108	1/5	29%
		2.45	132	1/5	20%
Trial 2	−246.0	369.00	35	0	-
		88.70	600	0	20%
		109.00	600	1/3	38%
		88.70	900	1/3	20%
Trial 3	−1.4	0.08	198	0	-
		0.41	8	0	18%
		0.50	8	3/7	41%
		0.44	12	3/7	18%

[a] Proportion of observed variance in endpoints due to measurement error.
[b] Type-II error calculations are based on results provided in section 3.1.

does not equal 0. We assume that both parameters θ_0 and θ_1 are estimable only in the external calibration sample comprising individuals not included in the trial ($j = 1, \ldots, K$).

Simple OLS regression estimators for β_Y, α_Y and σ^2 (the variance of the errors ε) in (S2.1) are,

$$\hat{\beta}_{Y^*} = \frac{\sum_i (X_i - \bar{X})(Y_i^* - \bar{Y}^*)}{\sum_i (X_i - \bar{X})^2}, \tag{S2.3}$$

$$\hat{\alpha}_{Y^*} = \bar{Y}^* - \hat{\beta}_{Y^*}\bar{X}, \tag{S2.4}$$

$$\omega_i = Y_i^* - \hat{\alpha}_{Y^*} - \hat{\beta}_{Y^*}X_i, \tag{S2.5}$$

$$s^2 = \frac{1}{N-2}\sum_i \omega_i^2, \tag{S2.6}$$

respectively. In a two-arm trial, the interest is in making inferences about β_Y, which cannot be directly estimated because in the trial the endpoint of interest Y was replaced by Y^*. In the following we will show: a) that $\hat{\beta}_{Y^*}$ may be a poor estimator for β_Y (section 3.1-3.4), and b) how adjustments to $\hat{\beta}_{Y^*}$ using information from the calibration model described by (S2.2) can improve inference about the treatment effect (section 4). As a starting point, in the following section relevant and known properties are defined for the special case that $Y^* = Y$, which is then followed by the properties under different measurement error structures for Y^* in subsequent sections.

S2.2.1. No measurement error

Consider the hypothetical case that Y^* is a perfect proxy for Y, i.e. $Y^* = Y$. By using that $Y = \alpha_Y + \beta_Y X + \varepsilon$, as defined in (S2.1), it follows that:

$$Y^* = \alpha_Y + \beta_Y X + \varepsilon.$$

From standard regression theory (e.g. [9]), we know that if the errors ε satisfy the regular Gauss-Markov assumptions [9] and their variance is defined by σ^2, the OLS estimators $\hat{\beta}_Y^*$, $\hat{\alpha}_Y^*$, and s^2 (defined by S2.3, S2.4, and S2.6, respectively) are Best Linear Unbiased Estimators (BLUE) for β_Y, α_Y, and σ^2, respectively.

Moreover, if the ε are independently and identically (iid) normally distributed, the OLS estimators $\hat{\beta}_{Y^*}$ and $\hat{\alpha}_{Y^*}$ (defined in S2.3 and S2.4, respectively) are the Maximum Likelihood Estimators (MLE) of β_Y and α_Y, respectively. Note that the errors ε satisfy the Gauss-Markov assumptions if we assume that they are iid normally distributed with mean 0 and constant variance σ^2.

Hypotheses for the treatment effect β_Y, can be defined by:

$$\begin{aligned} H_0 &: \quad \beta_Y = \beta_0, \\ H_A &: \quad \beta_Y \neq \beta_0. \end{aligned}$$

Under normality of the error terms ε, the OLS estimator $\hat{\beta}_Y^*$ defined in (S2.3) is the MLE for β_Y and s^2 is an unbiased estimator for σ^2, the following is known for the Wald test:

$$T = \frac{\hat{\beta}_{Y^*} - \beta_0}{\sqrt{\widehat{\mathrm{Var}}(\hat{\beta}_{Y^*})}} \sim t_{N-2}, \tag{S2.7}$$

where,

$$\widehat{\mathrm{Var}}(\hat{\beta}_{Y^*}) = \frac{s^2}{\sum_i (X_i - \bar{X})^2}. \tag{S2.8}$$

Assuming no measurement error in Y and X, under H_0, T follows a Student's t distribution with $N-2$ degrees of freedom [9]. Under H_A, T follows a Student's t distribution with $N-2$ degrees of freedom and non-centrality parameter $(\beta_Y - \beta_0)/\sqrt{\widehat{\mathrm{Var}}(\hat{\beta}_{Y^*})}$.

S2.2.2. Classical measurement error

There is classical measurement error in Y^* if Y^* is an unbiased proxy for Y [10]:

$$Y^* = Y + e, \tag{S2.9}$$

where $\mathrm{E}[e] = 0$ and $\mathrm{Var}(e) = \tau^2$ and e mutually independent of Y, X, ε (in (S2.1)). By using that $Y = \alpha_Y + \beta_Y X + \varepsilon$ from (S2.1), it follows that:

$$Y^* = \alpha_Y + \beta_Y X + \varepsilon + e.$$

Given the aforementioned assumptions, the sum of e and ε, $\delta_1 = e + \varepsilon$, has variance $\mathrm{Var}(\delta_1) = \sigma^2 + \tau^2$. It follows that if the errors δ_1 satisfy the Gauss-Markov assumptions, $\hat{\beta}_{Y^*}$ in (S2.3)

remains a BLUE estimator for β_Y. Also, $\hat{\alpha}_{Y^*}$ in (S2.4) and s^2 in (S2.6) remain BLUE estimators for α_Y and the variance of δ_1, respectively.

Further, if δ_1 is iid normally distributed with mean 0 and variance $\sigma^2 + \tau^2$, then $\hat{\alpha}_{Y^*}$ is the MLE for α_Y and $\hat{\beta}_{Y^*}$ is the MLE for β_Y. Obviously, given that $\sigma^2 > 0$ and $\tau^2 > 0$, the variance of the OLS regression estimator $\hat{\beta}_{Y^*}$ is larger if there is classical measurement error in the outcome compared to the case when there is no measurement error. Under the null hypothesis, the Wald test-statistic T defined in (S2.7) still follows a Student's t distribution with $N - 2$ degrees of freedom. However, under the alternative hypothesis, the non-centrality parameter of T, $(\beta_Y - \beta_0)/\sqrt{\widehat{\text{Var}}(\hat{\beta}_{Y^*})}$, will be smaller in the presence of classical measurement error.

To summarize, in the presence of only classical measurement error, Type-II error for detecting any given treatment effect increases, Type-I error is unaffected and the treatment effect estimator is unbiased MLE under standard regularity conditions.

Heteroscedastic classical measurement error

In the preceding we assumed that the Gauss-Markov assumptions were met. But notably, in the case that the variance of the errors e in (S2.9) varies per treatment arm, the errors are no longer homoscedastic (as needed to satisfy the Gauss-Markov assumptions) but heteroscedastic. In the case of this type of heteroscedastic classical measurement error, it can be shown that the variance of β_{Y^*} will be underestimated by the default estimator of the variance of $\hat{\beta}_{Y^*}$ defined by (S2.8), affecting both Type-I and Type-II error.

S2.2.3. Systematic measurement error

There is systematic measurement error in Y^*, if Y^* systematically depends on Y. Assuming this dependence is linear, the relation between Y^* and Y can be defined as:

$$Y^* = \theta_0 + \theta_1 Y + e, \tag{S2.10}$$

where $\text{E}[e] = 0$ and $\text{Var}(e) = \tau^2$. Throughout, we assume systematic measurement error if $\theta_0 \neq 0$ or $\theta_1 \neq 1$ (and of course, $\theta_1 \neq 0$ in all cases). We assume mutual independence between e and Y, X, ε (in S2.1). Naturally, if $\theta_0 = 0$ and $\theta_1 = 1$ the measurement error is of the classical form.

By using that $Y = \alpha_Y + \beta_Y X + \varepsilon$ from (S2.1), it follows that:

$$Y^* = \theta_0 + \theta_1\alpha_Y + \theta_1\beta_Y X + \theta_1\varepsilon + e.$$

Given the aforementioned assumptions, $\delta_2 = \theta_1\varepsilon + e$ with expected variance $\theta_1^2\sigma^2 + \tau^2$. It follows that under the Gauss-Markov assumptions, $\hat{\beta}_{Y^*}$ defined in (S2.3) is BLUE for $\theta_1\beta_Y$, and $\hat{\alpha}_{Y^*}$ defined in (S2.4) is BLUE for $\theta_0 + \alpha_Y$ and s^2 defined in (S2.6) is BLUE for the variance of δ_2 (i.e. $\theta_1^2\tau^2 + \sigma^2$). Conversely, $\hat{\beta}_{Y^*}$ is no longer BLUE for β_Y. Note that in this case s^2 is BLUE for $\theta_1^2\sigma^2 + \tau^2$, that is, depending on θ_1, smaller or larger than σ^2 (the variance of the error terms if there is no measurement error).

If we further assume that δ_2 is iid normally distributed, we can conclude that $\hat{\alpha}_{Y^*}$ is the MLE for $\theta_0 + \alpha_Y$ and $\hat{\beta}_{Y^*}$ is the MLE for $\theta_1\beta_Y$. Conversely, $\hat{\beta}_{Y^*}$ is no longer the MLE for β_Y, if there is systematic measurement error in Y^*. In the absence of a treatment effect, as $\theta_1\beta_Y = 0$ if $\beta_Y = 0$, T defined in (S2.7) still follows a Student's t distribution with $N - 2$

degrees of freedom. In the presence of any given treatment effect, T follows a non-central Student's t distribution with $N - 2$ degrees of freedom and non-centrality parameter $(\theta_1\beta_Y - \beta_0)/\sqrt{\widehat{\text{Var}}(\hat{\beta}_{Y^*})}$. Depending on the value of θ_1, the non-centrality parameter will be smaller or larger than the non-centrality parameter in the absence of measurement error (see section 3.2).

In summary, if there is systematic measurement error in the endpoints, the Type-I error is unaffected under standard regularity conditions and hence testing whether there is no effect is still valid under the null hypothesis [11]). Type-II, however, is affected (it may increase or decrease) and the treatment effect estimator is a biased MLE.

S2.2.4. Differential measurement error

There is differential measurement error in Y^* when measurement error varies with X. Assuming a linear model for this variation, formally:

$$Y^* = \theta_{00} + (\theta_{01} - \theta_{00})X + \theta_{10}Y + (\theta_{11} - \theta_{10})XY + e_X, \tag{S2.11}$$

where $\text{E}[e_X] = 0$ and $\text{Var}(e_X) = \tau_X^2$ and e_X independent of the endpoint of interest Y, and ε in (S2.1). From the equations it becomes clear that systematic error (equation (S2.10)) can be seen as a special case of differential error, where $\theta_{00} = \theta_{01}$ and $\theta_{10} = \theta_{11}$.

By using that $Y = \alpha_Y + \beta_Y X + \varepsilon$ from (S2.1), it follows from equation (S2.11) that,

$$Y^* = \theta_{00} + \theta_{10}\alpha_Y + \left[\theta_{01} - \theta_{00} + (\theta_{11} - \theta_{10})\alpha_Y + \theta_{11}\beta_Y\right]X + \left[\theta_{10} + (\theta_{11} - \theta_{10})X\right]\varepsilon + e_X.$$

Let $\delta_{3X} = \left[\theta_{10} + (\theta_{11} - \theta_{10})X\right]\varepsilon + e_X$, with expected variance $\left[\theta_{10}^2 + (\theta_{11}^2 - \theta_{10}^2)X\right]\sigma^2 + \tau_X^2$. Since the the error term δ_{3X} is no longer homoscedastic, the OLS estimators defined in (S2.3) and (S2.4) are no longer BLUE. However, the OLS estimator $\hat{\beta}_{Y^*}$ in (S2.3) is consistent (although not efficient) for $\theta_{01} - \theta_{00} + (\theta_{11} - \theta_{10})\alpha_Y + \theta_{11}\beta_Y$. The OLS estimator $\hat{\alpha}_{Y^*}$ defined in (S2.4) is consistent (although not efficient) for $\theta_{00} + \theta_{10}\alpha_Y$. Nevertheless, the estimator for the variance of $\hat{\beta}_{Y^*}$ defined in (S2.8) is no longer valid.

By using the residuals ω_i defined in (S2.6), a heteroscedastic consistent estimator for the variance of $\hat{\beta}_{Y^*}$ is:

$$\widehat{\text{Var}}(\hat{\beta}_{Y^*}) = \frac{\sum_i \left[(X_i - \bar{X})^2\omega_i^2\right]}{\left[\sum_i (X_i - \bar{X})^2\right]^2},$$

which is known as the White estimator [12]. From standard regression theory, it is known that using the above defined estimator, T defined in (S2.7) is still valid. Yet, under differential measurement error no longer $\left[\theta_{01} - \theta_{00} + (\theta_{11} - \theta_{10})\alpha_Y + \theta_{11}\beta_Y\right] = 0$ if $\beta_Y = 0$. Thus, under the null hypothesis, T defined in (S2.7) follows a Student's t distribution with $N - 2$ degrees of freedom and non-centrality parameter $(\left[\theta_{01} - \theta_{00} + \theta_{11}\alpha_Y - \theta_{10}\alpha_Y + \theta_{11}\beta_0\right] - \beta_0)/\sqrt{\widehat{\text{Var}}(\hat{\beta}_{Y^*})}$. Consequently, Type-I error changes if there is differential measurement error in Y^* and test about contrast under the null hypothesis are invalid [11]. Moreover, under the alternative hypothesis, T follows a non-central Student's t distribution with $N - 2$ degrees of freedom and non-centrality parameter $(\left[\theta_{01} - \theta_{00} + (\theta_{11} - \theta_{10})\alpha_Y + \theta_{11}\beta_Y\right] - \beta_0)/\sqrt{\widehat{\text{Var}}(\hat{\beta}_{Y^*})}$. Depending on the values of the θ's and α_Y, the non-centrality parameters will be smaller or larger than 0 and the non-centrality parameter if there is no measurement error, respectively (see section 3.2). Hence, Type-I

error and Type-II error could increase or decrease if there is differential measurement error in Y^*.

To summarize, Type-I error is not expected nominal (α) if there is differential measurement error in Y^* (see also [11]). Also, similar to systematic error in Y^*, Type-II error is affected (may increase or decrease) and the treatment effect estimator is biased.

S2.3. Correction methods for measurement error in a continuous trial endpoint

To accommodate measurement error correction, we assume that Y and Y^* are both measured for a smaller set of different individuals not included in the trial ($j = 1, \dots, K, K < N$), hereinafter referred to as the external calibration sample. In all but one case, it is assumed that only Y^* and Y are measured in the external calibration sample. In the case that the error in Y^* is different for the two treatment groups, it is assumed that the external calibration sample is in the form of a small pilot study where both treatments are allocated (i.e., Y^* and Y are both measured after assignment of X).

S2.3.1. Systematic measurement error

Using an external calibration set and assuming that the errors e in (S2.10) are iid normal, the MLE of the measurement error parameters in (S2.10) are:

$$\begin{aligned} \hat{\theta}_1 &= \frac{\sum_j (Y_j^{(c)} - \bar{Y}^{(c)})(Y_j^{*(c)} - \bar{Y}^{*(c)})}{\sum (Y_j^{(c)} - \bar{Y}^{(c)})^2}, \\ \hat{\theta}_0 &= \bar{Y}^{*(c)} - \hat{\theta}_1 \bar{Y}^{(c)}, \\ t^2 &= \frac{1}{K-2} \sum_j (Y_j^{*(c)} - \hat{\theta}_0 - \hat{\theta}_1 Y_j^{(c)})^2. \end{aligned} \tag{S2.12}$$

The superscript (c) is used to indicate that the measurement is obtained in the calibration set. From section 3.4, under systematic measurement error and assuming that ε in (S2.1) and e in (S2.10) iid normal and independent, the estimator $\hat{\beta}_{Y^*}$ defined in (S2.3) is the MLE of $\theta_1 \beta_Y$ and, the estimator $\hat{\alpha}_{Y^*}$ defined in (S2.4) is the MLE of $\theta_0 + \theta_1 \alpha_Y$. Natural sample estimators for α_Y and β_Y are then

$$\hat{\alpha}_Y = (\hat{\alpha}_{Y^*} - \hat{\theta}_0)/\hat{\theta}_1 \quad \text{and} \quad \hat{\beta}_Y = \hat{\beta}_{Y^*}/\hat{\theta}_1, \tag{S2.13}$$

where $\hat{\theta}_0$ and $\hat{\theta}_1$ are the estimated error parameters from the calibration data set. From equation (S2.13), it becomes apparent that $\hat{\theta}_1$ needs to be assumed bounded away from zero for finite estimates of $\hat{\alpha}_Y$ and $\hat{\beta}_Y$ [13].

The first moment of estimators $\hat{\alpha}_Y$ and $\hat{\beta}_Y$ can be approximated by using multivariate Taylor expansions and assuming that ($\hat{\alpha}_{Y^*}$, $\hat{\beta}_{Y^*}$, $\hat{\theta}_0$, $\hat{\theta}_1$) are normally distributed [13],

$$\mathrm{E}[\hat{\alpha}_Y] \approx \alpha_Y + \frac{[\alpha_Y - \bar{y}^*]\tau^2}{\theta_1^2 S_{yy}^{(c)}} \quad \text{and} \quad \mathrm{E}[\hat{\beta}_Y] \approx \beta_Y + \frac{\beta_Y \tau^2}{\theta_1^2 S_{yy}^{(c)}},$$

where $S_{yy}^{(c)} = \sum(Y_j^{(c)} - \bar{Y}^{(c)})^2$, the total sum of squares of $Y^{(c)}$. In conclusion, the estimators $\hat{\alpha}_Y$ and $\hat{\beta}_Y$ are consistent. Formal derivations for the presented formulas are provided in section S2.5.

In the following we will focus on specifying confidence limits for the treatment effect estimator $\hat{\beta}_Y$ defined in (S2.13). We make use of the fact that this estimator is a ratio, which motivates the use of the Delta method, Fieller method and Zero-variance method [14]. We also present a non-parametric bootstrap method for specifying confidence limits [15].

Delta method

Assuming that $\hat{\beta}_{Y^*}$ and $\hat{\theta}_1$ are both normally distributed and applying the Delta method, the second moment of $\hat{\beta}_Y$ can be approximated [11]. Formal derivations of the presented formulas are provided in section S2.5. The Delta method variance of $\hat{\beta}_Y$ is given by:

$$\mathrm{Var}(\hat{\beta}_Y) \approx \frac{1}{\theta_1^2}\Big[\frac{\theta_1^2\sigma^2 + \tau^2}{S_{xx}} + \frac{\beta_Y^2\tau^2}{S_{yy}^{(c)}}\Big],$$

where $S_{xx} = \sum_i (X_i - \bar{X})^2$, the total sum of squares of X. An approximation of the above defined variance, denoted by $\widehat{\mathrm{Var}}(\hat{\beta}_Y)$, is provided by approximating θ_1, $\theta_1^2\sigma^2 + \tau^2$, τ^2 and β_Y respectively by $\hat{\theta}_1$, s^2, t^2 and $\hat{\beta}_Y$ [11].

An approximate confidence interval for the estimator $\hat{\beta}_Y$ is then given by

$$\hat{\beta}_Y \pm t_{(\alpha/2,n-2)}\sqrt{\widehat{\mathrm{Var}}(\hat{\beta}_Y)}. \tag{S2.14}$$

Fieller method

A second method to construct confidence intervals for the estimator $\hat{\beta}_Y$ in (S2.13), described by Buonaccorsi, is the Fieller method [11, 16]. In the case that $\hat{\theta}_1$ is significantly different from zero at a significance level of α (that is, $\hat{\theta}_1/\sqrt{t^2/S_{yy}^{(c)}} > t_{N-2}$), the $(1-\alpha)$ confidence intervals of $\hat{\beta}_Y$ are defined by the Fieller method by:

$$l_{upper,lower} = \frac{\hat{\beta}_{Y^*}\hat{\theta}_1 \pm \sqrt{\hat{\beta}_{Y^*}^2\hat{\theta}_1^2 - (\frac{t^2}{S_{yy}^{(c)}}t_q^2 - \hat{\theta}_1^2)(\frac{s^2}{S_{xx}}t_q^2 - \hat{\beta}_{Y^*}^2)}}{\frac{\tau^2}{S_{yy}^{(c)}}t_q^2 + \hat{\theta}_1^2}. \tag{S2.15}$$

A formal derivation can be found in section S2.5.

Zero-variance method

The zero-variance method adjusts the observed endpoints Y_i^* by

$$\hat{Y}_i = (Y_i^* - \hat{\theta}_0)/\hat{\theta}_1,$$

where $\hat{\theta}_0$ and $\hat{\theta}_1$ are derived from (S2.10). The adjusted endpoints are regressed on the treatment variable X, which yields,

$$\hat{\beta}_{\hat{Y}} = \frac{\sum_i (X_i - \bar{X})(\hat{Y}_i - \bar{\hat{Y}})}{\sum_i (X_i - \bar{X})^2} = \frac{\sum_i (X_i - \bar{X})(Y_i^* - \bar{Y}^*)/\hat{\theta}_1}{\sum_i (X_i - \bar{X})^2} = \hat{\beta}_{Y^*}/\hat{\theta}_1,$$

$$\hat{\alpha}_{\hat{Y}} = \bar{\hat{Y}} - \hat{\beta}_{\hat{Y}}\bar{X} = \frac{\bar{Y^*} - \hat{\beta}_{Y^*}\bar{X} - \hat{\theta}_0}{\hat{\theta}_1} = (\hat{\alpha}_{Y^*} - \hat{\theta}_0)/\hat{\theta}_1,$$
$$s^2_{\hat{Y}} = \frac{1}{N-2}\sum_i (\hat{Y}_i - \hat{\alpha}_{\hat{Y}} - \hat{\beta}_{\hat{Y}}X_i)^2 = \frac{1}{\hat{\theta}_1^2}s^2,$$

with $\hat{\beta}_{Y^*}$, $\hat{\alpha}_{Y^*}$ and s^2 as in equations (S2.3, S2.4 and S2.6), respectively. Thus, $\hat{\beta}_{\hat{Y}}$ equals $\hat{\beta}_Y$ and $\hat{\alpha}_{\hat{Y}}$ equals $\hat{\alpha}_Y$ defined in (S2.13).

When the value of $\hat{\theta}_1$ (i.e. θ_1) is known, the variance of the estimator $\hat{\beta}_{\hat{Y}}$ is equal to:

$$\text{Var}(\hat{\beta}_{\hat{Y}}) = \text{Var}(\hat{\beta}_{Y^*})/\theta_1^2 = \frac{\sigma^2 + \tau^2/\theta_1^2}{\sum_i (X_i - \bar{X})^2}.$$

Using the standard OLS regression framework the variance of $\hat{\beta}_{\hat{Y}}$ can be estimated by:

$$\widehat{\text{Var}}(\hat{\beta}_{\hat{Y}}) = \frac{s^2_{\hat{Y}}}{\sum_i (X_i - \bar{X})^2} = \frac{s^2/\hat{\theta}_1^2}{\sum_i (X_i - \bar{X})^2}. \tag{S2.16}$$

By replacing $\hat{\theta}_1$ by θ_1 in the above, the quantity in (S2.16) is in expectation equal to $\text{Var}(\hat{\beta}_{\hat{Y}})$ (defined above). The quantity in (S2.16) is used in the zero-variance method to construct confidence intervals for $\hat{\beta}_{\hat{Y}}$, by replacing $\widehat{\text{Var}}(\hat{\beta}_{\hat{Y}})$ for $\widehat{\text{Var}}(\hat{\beta}_Y)$ in equation S2.14. In conclusion, this zero-variance approach will provide confidence intervals for the treatment effect estimator while assuming there is no variance in $\hat{\theta}_1$ (giving it its name zero-variance method). Although the zero-variance approach wins in terms of simplicity, it may underestimate the variability of the ratio since the variance in $\hat{\theta}_1$ is assumed zero.

Bootstrap

An alternative for defining confidence intervals for the corrected treatment effect estimator $\hat{\beta}_Y$ is by using a non-parametric bootstrap [15]. We propose the following stepwise procedure:

1. Draw a random sample with replacement of size K of the calibration sample $(Y^{*(c)}, Y^{(c)})$ to estimate $\hat{\theta}_{1_B}$ defined in (S2.12).
2. Draw a random sample with replacement of size N of the trial data (Y^*, X) to calculate the corrected treatment effect estimate by $\hat{\beta}_{Y_B} = \beta_{Y_B^*}/\hat{\theta}_{1_B}$. Where $\beta_{Y_B^*}$ is defined in (S2.3).
3. Repeat step 1-2 B times, with B large (e.g. 999 times).
4. Approximate confidence intervals are given by the $(\alpha/2, 1-\alpha/2)$ percentile of the distribution of $\hat{\beta}_{Y_B}$.

S2.3.2. Differential measurement error

For corrections for endpoints that suffer from differential measurement error we will here assume the existence of a pilot trial, which serves as an external calibration set, where both

treatments are allocated at random that serves as an external calibration set to estimate the measurement error model in (S2.11). For notational convenience we rewrite the linear model in equation (S2.11) in matrix form as:

$$Y^* = X\theta + e, \tag{S2.17}$$

where $\mathrm{E}(e) = 0$ and $\mathrm{E}(ee') = \Sigma$, a positive definite matrix, with τ_X^2 on its diagonal. Further, $\theta = (\theta_1, \theta_2, \theta_3, \theta_4) = (\theta_{00}, \theta_{01} - \theta_{00}, \theta_{10}, \theta_{11} - \theta_{10})$. In the external calibration set, the measurement error parameters $\hat{\theta}$ can be estimated by,

$$\hat{\theta} = (X^{(c)\prime}X^{(c)})^{-1}X^{(c)\prime}Y^{(c)}, \tag{S2.18}$$

with variance,

$$\mathrm{Var}(\hat{\theta}) = (X^{(c)\prime}X^{(c)})^{-1}X^{(c)\prime}\Sigma X^{(c)}(X^{(c)\prime}X^{(c)})^{-1}.$$

See [12] for a discussion on different estimators for the above defined variance. From section 2.5 it follows that natural estimators for α_Y and β_Y are,

$$\hat{\alpha}_Y = (\hat{\alpha}_{Y^*} - \hat{\theta}_{00})/\hat{\theta}_{10} \quad \text{and} \quad \hat{\beta}_Y = (\hat{\beta}_{Y^*} + \hat{\alpha}_{Y^*} - \hat{\theta}_{01})/\hat{\theta}_{11} - \hat{\alpha}_Y, \tag{S2.19}$$

where $\hat{\theta}_{00}$, $\hat{\theta}_{10}$, $\hat{\theta}_{01}$ and $\hat{\theta}_{11}$ are estimated from the external calibration set. Here it is assumed that both $\hat{\theta}_{10}$ and $\hat{\theta}_{11}$ are bounded away from zero (for reasons similar to those mentioned in section 3.1).

By multivariate Taylor expansions, the first moments of the estimators $\hat{\alpha}_Y$ and $\hat{\beta}_Y$ defined in (S2.19) can be approximated [11], in the same way as the estimators for systematic measurement error (section 4.1),

$$\begin{aligned}
\mathrm{E}\big[\hat{\alpha}_Y\big] &\approx \alpha_Y + \frac{1}{\theta_{10}^2}\Big[\alpha_Y \mathrm{Var}\big(\hat{\theta}_{10}\big) + \mathrm{Cov}\big(\hat{\theta}_{00}, \hat{\theta}_{10}\big)\Big], \\
\mathrm{E}\big[\hat{\beta}_Y\big] &\approx \beta_Y + \frac{1}{\theta_{11}^2}\Big[(\beta_Y + \alpha_Y)\mathrm{Var}\big(\hat{\theta}_{11}\big) + \mathrm{Cov}\big(\hat{\theta}_{01}, \hat{\theta}_{11}\big)\Big] \\
&\quad - \frac{1}{\theta_{10}^2}\Big[\alpha_Y \mathrm{Var}\big(\hat{\theta}_{10}\big) + \mathrm{Cov}\big(\hat{\theta}_{00}, \hat{\theta}_{10}\big)\Big].
\end{aligned}$$

From this, it is apparent that the estimators $\hat{\alpha}_Y$ and $\hat{\beta}_Y$ defined in (S2.19) are consistent (details are found in section S2.5). In the subsequent sections we review the Delta method, zero-variance and propose a bootstrap for specifying confidence limits for the estimator of the treatment effect under differential measurement error of the endpoints.

Delta method

The variance of the estimator $\hat{\beta}_Y$ defined in (S2.19) can be approximated by the Delta method [11]:

$$\begin{aligned}
\mathrm{Var}\big(\hat{\beta}_Y\big) \approx \frac{1}{\theta_{11}^2}\Big[&\big(\beta_Y + \alpha_Y\big)^2 \mathrm{Var}\big(\hat{\theta}_{11}\big) + \mathrm{Var}\big(\hat{\beta}_{Y^*}\big) + \mathrm{Var}\big(\hat{\alpha}_{Y^*}\big) + \\
&2\mathrm{Cov}\big(\hat{\alpha}_{Y^*}, \hat{\beta}_{Y^*}\big) + \mathrm{Var}\big(\hat{\theta}_{01}\big) + 2\big(\beta_Y + \alpha_Y\big)\mathrm{Cov}\big(\hat{\theta}_{11}, \hat{\theta}_{01}\big)\Big] +
\end{aligned}$$

$$\mathrm{Var}\big(\hat{\alpha}_Y\big),$$

where $\mathrm{Var}\big(\hat{\alpha}_Y\big)$ is approximated by:

$$\mathrm{Var}\Big(\frac{\hat{\alpha}_{Y^*} - \hat{\theta}_{00}}{\hat{\theta}_{10}}\Big) \approx \frac{1}{\theta_{10}^2}\Big[\mathrm{Var}\big(\hat{\alpha}_{Y^*}\big) + \alpha_Y^2\mathrm{Var}\big(\hat{\theta}_{10}\big) + \mathrm{Var}\big(\hat{\theta}_{00}\big) + 2\alpha_Y\mathrm{Cov}\big(\hat{\theta}_{00}, \hat{\theta}_{10}\big)\Big].$$

An approximate confidence interval for the estimator $\hat{\beta}_Y$ in (S2.19) is:

$$\hat{\beta}_Y \pm t_{(\alpha/2,n-2)}\sqrt{\mathrm{Var}\big(\hat{\beta}_Y\big)}. \tag{S2.20}$$

An approximation of θ_{11}, θ_{10}, $\theta_{11}^2\sigma^2 + \tau_1^2$, $\theta_{10}^2\sigma^2 + \tau_0^2$, τ_1^2, τ_0^2, β_Y and α_Y in the above is provided by: $\hat{\theta}_{11}$, $\hat{\theta}_{10}$, s_1^2, s_0^2, t_1^2, t_0^2, $\hat{\beta}_Y$ and $\hat{\alpha}_Y$ [11].

Zero-variance method

The zero-variance method adjusts the observed endpoints Y_i^* by

$$\hat{Y}_{ix} = (Y_{ix}^* - \hat{\theta}_{0x})/\hat{\theta}_{1x},$$

for $x \in \{0, 1\}$ and $\hat{\theta}_{0x}$) and $\hat{\theta}_{1x}$ derived from (S2.18). In the zero-variance method the above defined adjusted values are regressed on the treatment variable X, yielding in estimators $\hat{\alpha}_{\hat{Y}}$ and $\hat{\beta}_{\hat{Y}}$, which are, respectively, equal to the estimators $\hat{\alpha}_Y$ and $\hat{\beta}_Y$ defined in (S2.19). The variance of these estimators can be approximated with a heteroscedastic consistent covariance estimator (see [12] for an overview). Confidence intervals for $\hat{\beta}_{\hat{Y}}$ are subsequently constructed by using formula S2.20. Similar to what is described in section 4.1.3 discussing the zero-variance method for systematic measurement error, this way of constructing confidence intervals neglects the variance of the θ's from the calibration data set, and will thus often yield in confidence intervals that are too narrow.

Bootstrap

We here alternatively propose a non-parametric bootstrap procedure to specify confidence limits. This entails the following steps:

1. Draw a random sample with replacement of size K of the calibration sample and estimate $\hat{\theta}$ as defined in (S2.18).

2. Draw a random sample (with replacement) of size N of the study population and calculate the effect estimate by $\hat{\alpha}_{Y_B} = (\alpha_{Y_B^*} - \hat{\theta}_{00_B})/\hat{\theta}_{10_B}$ and $\hat{\beta}_{Y_B} = (\beta_{Y_B^*} + \alpha_{Y_B^*} - \hat{\theta}_{01_B})/\hat{\theta}_{11_B} - \hat{\alpha}_{Y_B}$. Where $\beta_{Y_B^*}$ and $\alpha_{Y_B^*}$ are defined in (S2.3) and (S2.4), respectively.

3. Repeat step 1-2 B times, with B large (e.g. 999 times).

4. Approximate confidence intervals are given by the $(\alpha/2, 1 - \alpha/2)$ percentile of the distribution of $\hat{\beta}_{Y_B}$.

S2.4. Measurement error depending on prognostic factors

Assume that, $\mathrm{E}[Y|X,S] = \alpha + \beta X + \gamma S$, $\mathrm{E}[Y^*|Y,S] = Y + \zeta S$, $Y^* \perp\!\!\!\perp X|Y$ (non-differential measurement error) and $S \perp\!\!\!\perp X$ (randomization is well-performed).

Suppose that we want to estimate the effect of Y on X (i.e., β), but instead of Y we have only measured the with measurement error contaminated Y^*. If one is aware that there is a prognostic factor that confounds the relation between Y^* and Y (and this factor is measured), one could decide to regress Y^* on X and S. The regression of Y^* on X and S equals,

$$\begin{aligned}
\mathrm{E}[Y^*|X,S] &= \mathrm{E}_{Y|X,S}\{\mathrm{E}_{Y^*|X,S,Y}[Y^*|X,S,Y]|X,S\} \\
&= \mathrm{E}_{Y|X,S}\{\mathrm{E}_{Y^*|S,Y}[Y^*|S,Y]|X,S\} \\
&= \mathrm{E}_{Y|X,S}\{Y + \boxtimes S|X,S\} \\
&= \alpha + \beta X + (\gamma + \zeta)S.
\end{aligned}$$

Thus, using the with measurement error contaminated endpoint Y^* instead of the preferred endpoint Y will provide an unbiased estimation of β.

However, if one is not aware of the prognostic factor, one might naively regress Y^* on X, which equals:

$$\begin{aligned}
\mathrm{E}[Y^*|X] &= \mathrm{E}_{S|X}\{\mathrm{E}_{Y|X,S}\{\mathrm{E}_{Y^*|X,S,Y}[Y^*|X,S,Y]|X,S\}|X\} \\
&= \mathrm{E}_{S|X}\{\alpha + \beta X + (\gamma + \zeta)S|X\} \\
&= \alpha + \beta X + (\gamma + \zeta)\mathrm{E}[S].
\end{aligned}$$

In conclusion, with ignoring the prognostic factor and using the with measurement error contaminated endpoint Y^* instead of the preferred endpoint Y, the regression of Y^* on X still results in an unbiased estimation of β.

S2.5. Approximation of bias and variance in corrected estimator

S2.5.1. Systematic measurement error

Obvious estimators for α_Y and β_Y are:

$$\hat{\alpha}_Y = (\hat{\alpha}_{Y^*} - \hat{\theta}_0)/\hat{\theta}_1 \quad \text{and} \quad \hat{\beta}_Y = \hat{\beta}_{Y^*}/\hat{\theta}_1.$$

These estimators can be approximated with a second order Taylor expansion by:

$$\begin{aligned}
\frac{\hat{\beta}_{Y^*}}{\hat{\theta}_1} &\approx \frac{\beta_{Y^*}}{\theta_1} - \frac{\beta_{Y^*}}{\theta_1^2}(\hat{\theta}_1 - \theta_1) + \frac{1}{\theta_1}(\hat{\beta}_{Y^*} - \beta_{Y^*}) \\
&\quad + \frac{1}{2!}\Big[\frac{2\beta_{Y^*}}{\theta_1^3}(\hat{\theta}_1 - \theta_1)^2 - \frac{2}{\theta_1^2}(\hat{\theta}_1 - \theta_1)(\hat{\beta}_{Y^*} - \beta_{Y^*})\Big], \\
\frac{\hat{\alpha}_{Y^*}}{\hat{\theta}_1} &\approx \frac{\alpha_{Y^*}}{\theta_1} - \frac{\alpha_{Y^*}}{\theta_1^2}(\hat{\theta}_1 - \theta_1) + \frac{1}{\theta_1}(\hat{\alpha}_{Y^*} - \alpha_{Y^*}) \\
&\quad + \frac{1}{2!}\Big[\frac{2\alpha_{Y^*}}{\theta_1^3}(\hat{\theta}_1 - \theta_1)^2 - \frac{2}{\theta_1^2}(\hat{\theta}_1 - \theta_1)(\hat{\alpha}_{Y^*} - \alpha_{Y^*})\Big],
\end{aligned}$$

$$\frac{\hat{\theta}_0}{\hat{\theta}_1} \approx \frac{\theta_0}{\theta_1} - \frac{\theta_0}{\theta_1^2}(\hat{\theta}_1 - \theta_1) + \frac{1}{\theta_1}(\hat{\theta}_0 - \theta_0) + \frac{1}{2!}\Big[\frac{2\theta_0}{\theta_1^3}(\hat{\theta}_1 - \theta_1)^2 - \frac{2}{\theta_1^2}(\hat{\theta}_1 - \theta_1)(\hat{\theta}_0 - \theta_0)\Big].$$

Simplifying these terms and substraction of the latter two, will lead to the following approximations for $\hat{\alpha}_Y$ and $\hat{\beta}_Y$:

$$\frac{\hat{\beta}_{Y^*}}{\hat{\theta}_1} \approx \frac{\beta_{Y^*}}{\theta_1} + \frac{1}{\theta_1}\Big[-\frac{\beta_{Y^*}}{\theta_1}(\hat{\theta}_1 - \theta_1) + (\hat{\beta}_{Y^*} - \beta_{Y^*})\Big] + \frac{1}{\theta_1^2}\Big[\frac{\beta_{Y^*}}{\theta_1}(\hat{\theta}_1 - \theta_1)^2 - (\hat{\beta}_{Y^*} - \beta_{Y^*})(\hat{\theta}_1 - \theta_1)\Big],$$

$$\frac{\hat{\alpha}_{Y^*} - \hat{\theta}_0}{\hat{\theta}_1} \approx \frac{\alpha_{Y^*} - \theta_0}{\theta_1} + \frac{1}{\theta_1}\Big[-\frac{\alpha_{Y^*} - \theta_0}{\theta_1}(\hat{\theta}_1 - \theta_1) + (\hat{\alpha}_{Y^*} - \alpha_{Y^*}) - (\hat{\theta}_0 - \theta_0)\Big] + \frac{1}{\theta_1^2}\Big[\frac{\alpha_{Y^*} - \theta_0}{\theta_1}(\hat{\theta}_1 - \theta_1)^2 - (\hat{\alpha}_{Y^*} - \alpha_{Y^*})(\hat{\theta}_1 - \theta_1) + (\hat{\theta}_0 - \theta_0)(\hat{\theta}_1 - \theta_1)\Big].$$

Since $\mathrm{E}[\hat{\theta}_1 - \theta_1] = 0$, $\mathrm{E}[\hat{\theta}_0 - \theta_0] = 0$, $\mathrm{E}[\hat{\alpha}_{Y^*} - \alpha_{Y^*}] = 0$ and $\mathrm{E}[\hat{\beta}_{Y^*} - \beta_{Y^*}] = 0$ an approximation of the expected value of the estimator $\hat{\alpha}_Y$ is given by:

$$\begin{aligned}\mathrm{E}\Big[\frac{\hat{\alpha}_{Y^*} - \hat{\theta}_0}{\hat{\theta}_1}\Big] &\approx \frac{\alpha_{Y^*} - \theta_0}{\theta_1} + \frac{1}{\theta_1^2}\Big[\frac{\alpha_{Y^*} - \theta_0}{\theta_1}\mathrm{E}\big[(\hat{\theta}_1 - \theta_1)^2\big] \\ &\quad - \mathrm{E}\big[(\hat{\alpha}_{Y^*} - \alpha_{Y^*})(\hat{\theta}_1 - \theta_1)\big] + \mathrm{E}\big[(\hat{\theta}_0 - \theta_0)(\hat{\theta}_1 - \theta_1)\big]\Big] = \\ &= \frac{\alpha_{Y^*} - \theta_0}{\theta_1} + \frac{1}{\theta_1^2}\Big[\frac{\alpha_{Y^*} - \theta_0}{\theta_1}\mathrm{Var}\big(\hat{\theta}_1\big) - \mathrm{Cov}\big(\hat{\alpha}_{Y^*}, \hat{\theta}_1\big) + \mathrm{Cov}\big(\hat{\theta}_0, \hat{\theta}_1\big)\Big] = \\ &= \alpha_Y + \frac{1}{\theta_1^2}\Big[\frac{\tau^2[\alpha_Y - \bar{Y}^{(c)}]}{\sum(Y_j^{(c)} - \bar{Y}^{(c)})^2}\Big].\end{aligned}$$

Congruently, an approximation of the expected value of the estimator $\hat{\beta}_Y$ is given by:

$$\begin{aligned}\mathrm{E}\Big[\frac{\hat{\beta}_{Y^*}}{\hat{\theta}_1}\Big] &\approx \frac{\beta_{Y^*}}{\theta_1} + \frac{1}{\theta_1^2}\Big[\frac{\beta_{Y^*}}{\theta_1}\mathrm{E}\big[(\hat{\theta}_1 - \theta_1)^2\big] - \mathrm{E}\big[(\hat{\beta}_{Y^*} - \beta_{Y^*})(\hat{\theta}_1 - \theta_1)\big]\Big] = \\ &= \frac{\beta_{Y^*}}{\theta_1} + \frac{1}{\theta_1^2}\Big[\frac{\beta_{Y^*}}{\theta_1}\mathrm{Var}\big(\hat{\theta}_1\big)\Big] = \\ &= \beta_Y + \frac{1}{\theta_1^2}\Big[\frac{\tau^2\beta_Y}{\sum(Y_j^{(c)} - \bar{Y}^{(c)})^2}\Big].\end{aligned}$$

Only using the first order Taylor expansion of the estimators, approximations of the variance of $\hat{\alpha}_Y$ and $\hat{\beta}_Y$ are respectively:

$$\mathrm{Var}\Big(\frac{\hat{\alpha}_{Y^*} - \hat{\theta}_0}{\hat{\theta}_1}\Big) \approx \frac{1}{\theta_1^2}\Big[\alpha_Y^2\mathrm{Var}\big(\hat{\theta}_1\big) + \mathrm{Var}\big(\hat{\alpha}_{Y^*} - \hat{\theta}_0\big) - 2\alpha_Y\mathrm{Cov}\big(\hat{\theta}_1, \hat{\alpha}_{Y^*} - \hat{\theta}_0\big)\Big] =$$

S2

$$
\begin{aligned}
&= \frac{1}{\theta_1^2}\Big[\alpha_Y^2 \mathrm{Var}\big(\hat{\theta}_1\big) + \mathrm{Var}\big(\hat{\alpha}_{Y^*}\big) + \mathrm{Var}\big(\hat{\theta}_0\big) - 2\mathrm{Cov}\big(\hat{\alpha}_{Y^*}, \hat{\theta}_0\big) \\
&- 2\alpha_Y \mathrm{Cov}\big(\hat{\theta}_1, \hat{\alpha}_{Y^*}\big) + 2\alpha_Y \mathrm{Cov}\big(\hat{\theta}_1, \hat{\theta}_0\big)\Big] = \\
&= \frac{1}{\theta_1^2}\Big[\frac{(\theta_1^2\sigma^2 + \tau^2)\sum X_i^2}{N\sum (X_i - \bar{X})^2} + \alpha_Y^2 \frac{\tau^2}{\sum (Y_j^{(c)} - \bar{Y}^{(c)})^2} \\
&+ \frac{\tau^2 \sum (Y_j^{(c)})^2}{K\sum (Y_j^{(c)} - \bar{Y}^{(c)})^2} \\
&+ 2\alpha_Y \frac{-\tau^2 \bar{Y}^{(c)}}{\sum (Y_j^{(c)} - \bar{Y}^{(c)})^2}\Big] = \\
&= \frac{1}{\theta_1^2}\Big[\frac{(\theta_1^2\sigma^2 + \tau^2)\sum X_i^2}{N\sum (X_i - \bar{x})^2} + \alpha_Y^2 \frac{\tau^2}{\sum (y_j^{(c)} - \bar{y}^{(c)})^2} \\
&+ \frac{\tau^2(\sum (y_j^{(c)} - \bar{y}^{(c)})^2 + K(\bar{y}^{(c)})^2)}{K\sum (y_j^{(c)} - \bar{y}^{(c)})^2} \\
&- 2\alpha_Y \frac{\tau^2 \bar{y}^{(c)}}{\sum (y_j^{(c)} - \bar{y}^{(c)})^2}\Big] = \\
&= \frac{1}{\theta_1^2}\Big[\frac{(\theta_1^2\sigma^2 + \tau^2)\sum x_i^2}{N\sum (x_i - \bar{x})^2} + \tau^2\Big(\frac{1}{K} + \frac{(\bar{y}^{(c)} - \alpha_Y)^2}{\sum (y_j^{(c)} - \bar{y}^{(c)})^2}\Big)\Big],
\end{aligned}
$$

$$
\mathrm{Var}\Big(\frac{\hat{\beta}_{Y^*}}{\hat{\theta}_1}\Big) \approx \frac{1}{\theta_1^2}\Big[\frac{\theta_1^2\sigma^2 + \tau^2}{\sum (x_i - \bar{x})^2} + \frac{\beta_Y^2 \tau^2}{\sum (y_j^{(c)} - \bar{y}^{(c)})^2}\Big].
$$

Fieller method

Assume that $\hat{\beta}_{Y^*}$ and $\hat{\theta}_1$ are normally distributed (note that this assumption is satisfied with large study samples (N) and large calibration samples (K)). The sum of two normally distributed variables is normally distributed, hence, $\hat{\beta}_{Y^*} - \beta_Y \hat{\theta}_1$ is normally distributed. Furthermore, we have,

$$
\mathrm{Var}(\hat{\beta}_{Y^*} - \beta_Y \hat{\theta}_1) = \mathrm{Var}(\hat{\beta}_{Y^*}) + \beta_Y^2 \mathrm{Var}(\hat{\theta}_1).
$$

Where,

$$
\begin{aligned}
\mathrm{Var}(\hat{\beta}_{Y^*}) &= \frac{\theta_1^2\sigma^2 + \tau^2}{\sum (x_i - \bar{x})^2} \\
\mathrm{Var}(\hat{\theta}_1) &= \frac{\tau^2}{\sum (y_j^{(c)} - \bar{y}^{(c)})^2}
\end{aligned}
$$

If we now divide the term $\hat{\beta}_{Y^*} - \beta_Y \hat{\theta}_1$ by its standard deviation, we get:

$$T_0 = \frac{\hat{\beta}_{Y^*} - \beta_Y \hat{\theta}_1}{\sqrt{\frac{\theta_1^2\sigma^2+\tau^2}{\sum(x_i-\bar{x})^2} + \frac{\tau^2}{\sum(y_j^{(c)}-\bar{y}^{(c)})^2}\beta_Y^2}} \tag{S2.21}$$

We are interested to find the set of β_Y values for which the corresponding T_0 values lie within the $(1-\alpha)$ quantiles of the t-distribution with $N-2$ degrees of freedom (this only holds approximately, see for details [14]). Let us denote these values by t_q, from (S2.21) we have,

$$(\frac{\tau^2}{\sum(y_j^{(c)} - \bar{y}^{(c)})^2} t_q^2 - \hat{\theta}_1^2)\beta_Y^2 + 2\hat{\beta}_{Y^*}\hat{\theta}_1\beta_Y + (\frac{\theta_1^2\sigma^2 + \tau^2}{\sum(x_i - \bar{x})^2} t_q^2 - \hat{\beta}_{Y^*}^2) = 0.$$

In the case that $\hat{\theta}_1$ is significantly different from zero at a significance level of α (that is, $\hat{\theta}_1 / \sqrt{\frac{\tau^2}{\sum(y_j^{(c)}-\bar{y}^{(c)})^2}} > t_q$), solving this for β_Y results in the following $(1-\alpha)$ confidence intervals:

$$\beta_Y = \frac{-\hat{\beta}_{Y^*}\hat{\theta}_1 \pm \sqrt{\hat{\beta}_{Y^*}^2\hat{\theta}_1^2 - (\frac{\tau^2}{\sum(y_j^{(c)}-\bar{y}^{(c)})^2} t_q^2 - \hat{\theta}_1^2)(\frac{\theta_1^2\sigma^2+\tau^2}{\sum(x_i-\bar{x})^2} t_q^2 - \hat{\beta}_{Y^*}^2)}}{\frac{\tau^2}{\sum(y_j^{(c)}-\bar{y}^{(c)})^2} t_q^2 - \hat{\theta}_1^2}.$$

In the other case, the confidence intervals are unbounded, see for more details [14].

S2.5.2. Differential measurement error

Obvious estimators for α_Y and β_Y are:

$$\hat{\alpha}_Y = (\hat{\alpha}_{Y^*} - \hat{\theta}_{00})/\hat{\theta}_{10} \quad \text{and} \quad \hat{\beta}_Y = (\hat{\beta}_{Y^*} + \hat{\alpha}_{Y^*} - \hat{\theta}_{01})/\hat{\theta}_{11} - \hat{\alpha}_Y.$$

These estimators can be approximated with a second order Taylor expansion by:

$$\begin{aligned}
\frac{\hat{\alpha}_{Y^*} - \hat{\theta}_{00}}{\hat{\theta}_{10}} &\approx \frac{\alpha_{Y^*} - \theta_{00}}{\theta_{10}} + \frac{1}{\theta_{10}}\Big[- \frac{\alpha_{Y^*} - \theta_{00}}{\theta_{10}}(\hat{\theta}_{10} - \theta_{10}) + (\hat{\alpha}_{Y^*} - \alpha_{Y^*}) - (\hat{\theta}_{00} - \theta_{00})\Big] \\
&\quad + \frac{1}{\theta_{11}^2}\Big[\frac{\alpha_{Y^*} - \theta_{00}}{\theta_{10}}(\hat{\theta}_{10} - \theta_{10})^2 - (\hat{\alpha}_{Y^*} - \alpha_{Y^*})(\hat{\theta}_{10} - \theta_{10}) \\
&\quad + (\hat{\theta}_{00} - \theta_{00})(\hat{\theta}_{10} - \theta_{10})\Big], \\
\frac{\hat{\beta}_{Y^*} - \hat{\theta}_{01}}{\hat{\theta}_{11}} &\approx \frac{\beta_{Y^*} - \theta_{01}}{\theta_{11}} + \frac{1}{\theta_{11}}\Big[- \frac{\beta_{Y^*} - \theta_{01}}{\theta_{11}}(\hat{\theta}_{11} - \theta_{11}) + (\hat{\beta}_{Y^*} - \beta_{Y^*}) - (\hat{\theta}_{01} - \theta_{01})\Big] \\
&\quad + \frac{1}{\theta_{11}^2}\Big[\frac{\beta_{Y^*} - \theta_{01}}{\theta_{11}}(\hat{\theta}_{11} - \theta_{11})^2 - (\hat{\beta}_{Y^*} - \beta_{Y^*})(\hat{\theta}_{11} - \theta_{11}) \\
&\quad + (\hat{\theta}_{01} - \theta_{01})(\hat{\theta}_{11} - \theta_{11})\Big], \\
\frac{\hat{\alpha}_{Y^*}}{\hat{\theta}_{11}} &\approx \frac{\alpha_{Y^*}}{\theta_{11}} + \frac{1}{\theta_{11}}\Big[- \frac{\alpha_{Y^*}}{\theta_{11}}(\hat{\theta}_{11} - \theta_{11}) + (\hat{\alpha}_{Y^*} - \alpha_{Y^*}))\Big]
\end{aligned}$$

$$+ \frac{1}{\theta_{11}^2}\Big[\frac{\alpha_{Y^*}}{\theta_{11}}(\hat{\theta}_{11} - \theta_{11})^2 - (\hat{\alpha}_{Y^*} - \alpha_{Y^*})(\hat{\theta}_{11} - \theta_{11})\Big].$$

S2

Congruent to the results for the estimators under systematic measurement error, we can conclude:

$$\mathrm{E}\Big[\frac{\hat{\alpha}_{Y^*} - \hat{\theta}_{00}}{\hat{\theta}_{10}}\Big] \approx \alpha_Y + \frac{1}{\theta_{10}^2}\Big[\alpha_Y \mathrm{Var}(\hat{\theta}_{10}) + \mathrm{Cov}(\hat{\theta}_{00}, \hat{\theta}_{10})\Big].$$

Congruently, an approximation of the expected value of the estimator $\hat{\beta}_Y$ is given by:

$$\begin{aligned}\mathrm{E}\Big[\frac{\hat{\beta}_{Y^*} + \hat{\alpha}_{Y^*} - \hat{\theta}_{01}}{\hat{\theta}_{11}} - \hat{\alpha}_Y\Big] \approx \beta_Y &+ \frac{1}{\theta_{11}^2}\Big[(\beta_Y + \alpha_Y)\mathrm{Var}(\hat{\theta}_{11}) + \mathrm{Cov}(\hat{\theta}_{01}, \hat{\theta}_{11})\Big] \\ &- \frac{1}{\theta_{10}^2}\Big[\alpha_Y \mathrm{Var}(\hat{\theta}_{10}) + \mathrm{Cov}(\hat{\theta}_{00}, \hat{\theta}_{10})\Big].\end{aligned}$$

And the variance of the estimators is approximated by:

$$\begin{aligned}\mathrm{Var}\Big(\frac{\hat{\alpha}_{Y^*} - \hat{\theta}_{00}}{\hat{\theta}_{10}}\Big) \approx\ & \frac{1}{\theta_{10}^2}\Big[\mathrm{Var}(\hat{\alpha}_{Y^*}) \\ &+ \alpha_Y^2 \mathrm{Var}(\hat{\theta}_{10}) + \mathrm{Var}(\hat{\theta}_{00}) + 2\alpha_Y \mathrm{Cov}(\hat{\theta}_{00}, \hat{\theta}_{10})\Big], \\ \mathrm{Var}\Big(\frac{\hat{\beta}_{Y^*} + \hat{\alpha}_{Y^*} - \hat{\theta}_{01}}{\hat{\theta}_{11}} - \hat{\alpha}_Y\Big) \approx\ & \frac{1}{\theta_{11}^2}\Big[(\beta_Y + \alpha_Y)^2 \mathrm{Var}(\hat{\theta}_{11}) + \mathrm{Var}(\hat{\beta}_{Y^*}) + \mathrm{Var}(\hat{\alpha}_{Y^*}) \\ &+ 2\mathrm{Cov}(\hat{\alpha}_{Y^*}, \hat{\beta}_{Y^*}) + \mathrm{Var}(\hat{\theta}_{01}) \\ &+ 2(\beta_Y + \alpha_Y)\mathrm{Cov}(\hat{\theta}_{11}, \hat{\theta}_{01})\Big] \\ &+ \mathrm{Var}(\hat{\alpha}_Y).\end{aligned}$$

Note that in the case of differential measurement error, we assume that $\mathrm{Cov}(\hat{\theta}_{11}, \hat{\theta}_{00}) = 0$, $\mathrm{Cov}(\hat{\theta}_{11}, \hat{\theta}_{10}) = 0$, $\mathrm{Cov}(\hat{\theta}_{01}, \hat{\theta}_{00}) = 0$ and $\mathrm{Cov}(\hat{\theta}_{01}, \hat{\theta}_{10}) = 0$.

References

[1] E. T. Poehlman, Effects of endurance and resistance training on total daily energy expenditure in young women: A controlled randomized trial, Journal of Clinical Endocrinology & Metabolism 87 (3) (2002) 1004–1009. doi:10.1210/jc.87.3.1004.

[2] G. Plasqui, K. R. Westerterp, Physical activity assessment with accelerometers: An evaluation against doubly labeled water, Obesity 15 (10) (2007) 2371–2379. doi:10.1038/oby.2007.281.

[3] J. W. J. Bijlsma, P. M. J. Welsing, T. G. Woodworth, L. M. Middelink, A. Pethö-Schramm, C. Bernasconi, M. E. A. Borm, C. H. Wortel, E. J. ter Borg, Z. N. Jahangier, W. H. van der Laan, G. A. W. Bruyn, P. Baudoin, S. Wijngaarden, P. A. J. M. Vos, R. Bos, M. J. F. Starmans, E. N. Griep, J. R. M. Griep-Wentink, C. F. Allaart, A. H. M. Heurkens, X. M. Teitsma, J. Tekstra, A. C. A. Marijnissen, F. P. J. Lafeber, J. W. G. Jacobs, Early rheumatoid arthritis treated with tocilizumab, methotrexate, or their combination (U-Act-Early): a multicentre, randomised, double-blind, double-dummy, strategy trial, The Lancet 388 (10042) (2016) 343–355. doi:10.1016/S0140-6736(16)30363-4.

[4] M. L. L. Prevoo, M. A. Van'T Hof, H. H. Kuper, M. A. Van Leeuwen, L. B. A. Van De Putte, P. L. C. M. Van Riel, Modified disease activity scores that include twenty-eight-joint counts development and validation in a prospective longitudinal study of patients with rheumatoid arthritis, Arthritis & Rheumatism 38 (1) (1995) 44–48. doi:10.1002/art.1780380107.

[5] T. Pincus, Y. Yazici, T. Sokka, Complexities in assessment of rheumatoid arthritis: Absence of a single gold standard measure, Rheumatic Disease Clinics of North America 35 (4) (2009) 687–697. doi:10.1016/j.rdc.2009.10.002.

[6] J. Anderson, L. Caplan, J. Yazdany, M. L. Robbins, T. Neogi, K. Michaud, K. G. Saag, J. R. O'dell, S. Kazi, Rheumatoid arthritis disease activity measures: American College of Rheumatology recommendations for use in clinical practice, Arthritis Care & Research 64 (5) (2012) 640–647. arXiv:NIHMS150003, doi:10.1002/acr.21649.

[7] E. H. Choy, B. Khoshaba, D. Cooper, A. MacGregor, D. L. Scott, Development and validation of a patient-based disease activity score in rheumatoid arthritis that can be used in clinical trials and routine practice, Arthritis & Rheumatism 59 (2) (2008) 192–199. doi:10.1002/art.23342.

[8] M. Makrides, C. Crowther, R. Gibson, R. Gibson, C. Skeaff, Efficacy and tolerability of low-dose iron supplements during pregnancy: A randomized controlled trial, American Journal of Clinical Nutrition 78 (1) (2003) 145–153. doi:10.1093/ajcn/78.1.145.

[9] R. Davidson, J. MacKinnon, Econometric Theory and Methods, Oxford University Press, New York, NY, 2004.

[10] R. Carroll, D. Ruppert, L. Stefanski, C. Crainiceanu, Measurement error in nonlinear models: A modern perspective, 2nd Edition, Chapman & Hall/CRC, Boca Raton, FL, 2006.

[11] J. Buonaccorsi, Measurement errors, linear calibration and inferences for means, Computational Statistics & Data Analysis 11 (3) (1991) 239–257. doi:10.1016/0167-9473(91)90083-E.

[12] J. S. Long, L. H. Ervin, Using heteroscedasticity consistent standard errors in the linear regression model, The American Statistician 54 (3) (2000) 217–224. doi:10.2307/2685594.

[13] J. Buonaccorsi, Measurement error: Models, methods, and applications, Chapman & Hall/CRC, Boca Raton, FL, 2010.

[14] V. H. Franz, Ratios: A short guide to confidence limits and proper use (2007). arXiv:arXiv:0710.2024v1.

[15] B. Efron, Bootstrap methods: Another look at the jackknife, Annals of Statistics 7 (1) (1979) 1–26. doi:10.1214/aos/1176344552.

[16] E. C. Fieller, The biological standardization of insulin, Supplement to the Journal of the Royal Statistical Society 7 (1) (1940) 1–64. doi:10.2307/2983630.

S3

Supplementary material Chapter 3

These are the supplementary materials accompanying Chapter 3. The supplementary materials are structured as follows. In section S3.1, the variance of the regression calibration estimator is derived. In section S3.2, the variance of the maximum likelihood estimator for replicates study is derived.

S3.1. Variance estimation: standard regression calibration

Covariate measurement error. The variance–covariance matrix for the standard regression estimator $\hat{\beta}_{\text{RC}}$ can be approximated by using the multivariate delta method as described by [1], given by

$$\hat{\Sigma}_{\beta_{\text{RC}}}(j_1, j_2) = (\hat{A}' \hat{\Sigma}_{\beta^*} \hat{A})_{j_1, j_2} + \hat{\beta}^* \hat{\Sigma}_{A, j_1, j_2} \hat{\beta}^{*\prime}, \qquad j_1, j_2 = 1, \dots, (k+2), \tag{S3.1}$$

where $\hat{A}$ is the inverse of the calibration model matrix $\hat{\Lambda}$. Further, $\hat{\Sigma}_{\beta^*}$ is the variance–covariance matrix obtained from the naive regression defined in equation (3.2) in the main chapter and $\hat{\Sigma}_{A, j_1, j_2}$ is the $(k+2) \times (k+2)$ matrix relating the j_1th and j_2th column of A (we refer to Appendix of [1] for a derivation). Additionally, the so-called zero-variance variance–covariance matrix for $\hat{\beta}$ can be estimated by $\hat{A}' \Sigma_{\beta^*} \hat{A}$ (i.e., by omitting the variance in the calibration model matrix).

A $100(1 - \alpha)$ percent confidence interval for the jth element of $\hat{\beta}_{\text{RC}}$ is then

$$\hat{\beta}_{\text{RC}\,j} \pm \sqrt{\text{Var}(\hat{\beta}_{\text{RC}\,j})}, \tag{S3.2}$$

where $\text{Var}(\hat{\beta}_{\text{RC}\,j})$ is the jth element on the diagonal of $\hat{\Sigma}_{\beta_{\text{RC}}}$. The variance–covariance matrix $\hat{\Sigma}_{\beta_{\text{RC}}}$ can be obtained by either using the delta variance–covariance matrix or zero-variance variance–covariance matrix. In general, the zero-variance variance–covariance matrix will underestimate the true variance–covariance matrix and thus lead to too narrow confidence intervals.

Other methods to construct confidence intervals include stratified bootstrap [2] and the Fieller method [3–6]. In case of covariate measurement error, the Fieller method can

only be applied to construct a $100(1-\alpha)$ percent confidence interval for the first element of $\hat{\beta}_{\text{RC}}$, i.e., $\hat{\phi}_{\text{RC}}$. From [6] we obtain:

$$\{f_1 \pm \sqrt{f_1^2 - f_0 f_2}/f_2\}, \tag{S3.3}$$

where $f_0 = z^2_{\alpha/2}\text{Var}(\hat{\phi}^*) - \hat{\phi}^*$, $f_1 = z^2_{\alpha/2}\text{Cov}(\hat{\phi}^*, \hat{\lambda}_1) - \hat{\phi}^*\hat{\lambda}_1$, $f_2 = z^2_{\alpha/2}\text{Var}(\hat{\lambda}_1) - \hat{\lambda}_1^2$. Where it is assumed that $\text{Cov}(\hat{\phi}^*, \hat{\lambda}_1)$ is null. If the $(1-\alpha)\times 100\%$ confidence interval of $\hat{\lambda}_1$ includes 0, the Fieller method does not lead to bounded confidence intervals. Bootstrap confidence intervals are obtained by sampling the people in the validation set separately from the people not included in the validation set [2] and taking the $(100-\alpha)$ percentiles of the obtained distribution.

Outcome measurement error. The variance–covariance matrix for the standard regression estimator $(\hat{\beta_{\text{RC}}}, 1)$ can be approximated by applying the multivariate delta method similar to the variance obtained for the corrected estimator for covariate maesurement error,

$$\hat{\Sigma}_{(\beta_{\text{RC}},1)}(j_1, j_2) = (B'\hat{\Sigma}_{(\beta^*,1)}B)_{j_1,j_2} + (\hat{\beta}^*, 1)\hat{\Sigma}_{B,j_1,j_2}(\hat{\beta}^*, 1)', \qquad j_1, j_2 = 1, \ldots, (k+3),$$

where $\hat{B}$ is the inverse of the measurement error model matrix $\hat{\Theta}$. $\hat{\Sigma}_{(\beta^*,1)}$ is a $(k+3)\times(k+3)$ matrix where the upper $(k+2)\times(k+2)$ comprises the variance–covariance matrix obtained from the uncorrected regression defined by model (3.6) and the last row and column contain zeros. Further, $\hat{\Sigma}_{B,j_1,j_2}$ is the $(k+3)\times(k+3)$ matrix relating the j_1th and j_2th column of B (similar to [1]). The so-called zero-variance variance–covariance matrix for $\hat{\beta}$ can be estimated by $B'\hat{\Sigma}_{(\beta^*,1)}B$.

A $100(1-\alpha)$ percent confidence interval can be obtained from equation (S3.2). Further, $100(1-\alpha)$ percent confidence intervals for $\hat{\phi}$ and $\hat{\gamma}$ can be approximated by the Fieller method as defined in model S3.3, where $f_0 = \hat{\phi}^* - z^2_{\alpha/2}\text{Var}(\hat{\phi}^*)$, $f_1 = \hat{\phi}^*/\hat{\theta}_1 - z^2_{\alpha/2}\text{Cov}(\hat{\phi}^*, 1/\hat{\theta}_1)$, $f_2 = 1/\hat{\lambda}_1^2 - z^2_{\alpha/2}\text{Var}(1/\hat{\lambda}_1)$ and idem for $\hat{\gamma}$. Additionally, bootstrap can be used to construct confidence intervals for $\hat{\beta_{\text{RC}}}$. Bootstrap confidence intervals are obtained by sampling the individuals in the internal adjustment set separately from the individuals not included in the internal adjustment set and taking the $(100-\alpha)$ percentiles of the obtained distribution.

Differential outcome measurement error in univariable analyses. The varia-nce–covariance matrix for the standard regression estimator $(\hat{\beta_{\text{RC}}}, 1)$ can be estimated similar to non-differential outcome measurement error defined above (by using the measurement error matrices for differential outcome measurement error). Confidence intervals can then be obtained from equation (S3.2). Bootstrap confidence intervals are obtained by sampling the individuals in the internal adjustment set separately from the individuals not included in the internal adjustment set and taking the $(100-\alpha)$ percentiles of the obtained distribution.

S3.2. Variance estimation: maximum likelihood for replicates studies

The variance–covariance matrix for the maximum likelihood estimator $\hat{\beta}_{\text{MLE}}$ can be approximated by the multivariate delta method [7]. Denote $\zeta^* = (\delta_0, \delta_Z, \sigma^2_{Y|Z}, \kappa_0, \kappa_Y, \kappa_Z, \sigma^2_{X|Y,Z})$, leaving the τ^2 from ζ in the main chapter (see section 3.3.3) out as this parameter is not needed for the estimation of $\beta = (\alpha, \phi, \gamma)$. A standard result from linear mixed models is that the estimators of fixed parameters are asymptotically uncorrelated with the estimators of the variance component parameters [7]. If one further assumes that the estimators from the linear model of Y given Z are uncorrelated with the parameters estimated in the linear mixed model, it follows for large samples that $\hat{\zeta}^*$ is multivariate normal with mean ζ and variance covariance matrix $\text{Var}(\hat{\zeta})$ equal to:

$$\begin{pmatrix}
\text{Var}(\hat{\delta}_0) & \text{Cov}(\hat{\delta}_0, \hat{\delta}_Z) & 0 & 0 & 0 & 0 & 0 \\
\text{Cov}(\hat{\delta}_Z, \hat{\delta}_0) & \text{Var}(\hat{\delta}_Z) & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & \text{Var}(\hat{\sigma}^2_{Y|Z}) & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & \text{Var}(\hat{\kappa}_0) & \text{Cov}(\hat{\kappa}_0, \hat{\kappa}_Y) & \text{Cov}(\hat{\kappa}_0, \hat{\kappa}_Z) & 0 \\
0 & 0 & 0 & \text{Cov}(\hat{\kappa}_Y, \hat{\kappa}_0) & \text{Var}(\hat{\kappa}_Y) & \text{Cov}(\hat{\kappa}_Y, \hat{\kappa}_Z) & 0 \\
0 & 0 & 0 & \text{Cov}(\hat{\kappa}_Z, \hat{\kappa}_0) & \text{Cov}(\hat{\kappa}_Z, \hat{\kappa}_Y) & \text{Var}(\hat{\kappa}_Z) & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & \text{Var}(\hat{\sigma}^2_{X|Y,Z})
\end{pmatrix}$$

If $g : \mathbb{R}^{5+2k} \longrightarrow \mathbb{R}^{2+k}$ is the function that transforms ζ^* to $\beta_{\text{ML}} = (\alpha_{\text{ML}}, \phi_{\text{ML}}, \gamma_{\text{ML}})$, as defined in the main chapter, then by the multivariate delta method it follows that in large samples:

$$\hat{\beta}_{\text{ML}} \sim N(\beta_{\text{ML}}, Jg\text{Var}(\hat{\zeta})(Jg)'),$$

Where J is the Jacobian matrix of g:

$$Jg = \begin{pmatrix}
\frac{\partial \phi}{\partial \delta_0} & \frac{\partial \phi}{\partial \delta_Z} & \frac{\partial \phi}{\partial \sigma^2_{Y|Z}} & \cdots & \frac{\partial \phi}{\partial \sigma^2_{X|Y,Z}} \\
\frac{\partial \alpha}{\partial \delta_0} & \frac{\partial \alpha}{\partial \delta_Z} & \frac{\partial \alpha}{\partial \sigma^2_{Y|Z}} & \cdots & \frac{\partial \alpha}{\partial \sigma^2_{X|Y,Z}} \\
\frac{\partial \gamma}{\partial \delta_0} & \frac{\partial \gamma}{\partial \delta_Z} & \frac{\partial \gamma}{\partial \sigma^2_{Y|Z}} & \cdots & \frac{\partial \gamma}{\partial \sigma^2_{X|Y,Z}}
\end{pmatrix}.$$

Confidence intervals can then be obtained from equation (S3.2). Bootstrap confidence intervals are obtained by sampling the individuals in the internal adjustment set separately from the individuals not included in the internal adjustment set and taking the $(100 - \alpha)$ percentiles of the obtained distribution.

References

[1] B. Rosner, D. Spiegelman, W. C. Willett, Correction of logistic regression relative risk estimates and confidence intervals for measurement error: The case of multiple covariates measured with error, American Journal of Epidemiology 132 (4) (1990) 734–745. doi:10.1093/oxfordjournals.aje.a115715.

[2] R. Carroll, D. Ruppert, L. Stefanski, C. Crainiceanu, Measurement error in nonlinear models: A modern perspective, 2nd Edition, Chapman & Hall/CRC, Boca Raton, FL, 2006.

[3] J. Buonaccorsi, Measurement error: Models, methods, and applications, Chapman & Hall/CRC, Boca Raton, FL, 2010.

[4] V. H. Franz, Ratios: A short guide to confidence limits and proper use (2007). arXiv:arXiv:0710.2024v1.

[5] C. Frost, S. G. Thompson, Correcting for regression dilution bias: comparison of methods for a single predictor variable, Journal of the Royal Statistical Society: Series A (Statistics in Society) 163 (2) (2000) 173–189. doi:10.1111/1467-985X.00164.

[6] L. Nab, R. H. H. Groenwold, P. M. J. Welsing, M. Smeden, Measurement error in continuous endpoints in randomised trials: Problems and solutions, Statistics in Medicine 38 (27) (2019) 5182–5196. doi:10.1002/sim.8359.

[7] J. W. Bartlett, B. L. De Stavola, C. Frost, Linear mixed models for replication data to efficiently allow for covariate measurement error, Statistics in Medicine 28 (25) (2009) 3158–3178. doi:10.1002/sim.3713.

S5

Supplementary material Chapter 5

These are the supplementary materials accompanying Chapter 5. The supplementary materials are structured as follows. In section S5.1 we introduce notation, describe the implications of exposure measurement error and describe the different analyses used in the main chapter. In section S5.2 the parameters of the simulation study from the main chapter are presented. Section S5.3 contains the additional results from the simulation study in the main chapter that were left out for brevity there.

S5.1. Notation, impact of measurement error and different analysis strategies

Throughout the main paper, our interest is the causal effect of the exposure VAT on the outcome insulin resistance IR, adjusted for a predefined set of k confounders, jointly written as Z (e.g., age, sex and total body fat). We assume a linear model for the outcome without interaction between exposure and covariates:

$$\text{IR} = \text{intercept} + \beta\text{VAT} + \gamma^{'} Z + \varepsilon. \tag{S5.1}$$

Here, we assume that the residuals errors ε are independent of VAT and confounders Z, with mean 0 and variance σ^2. Additionally, γ is assumed a $k \times 1$ vector of regression coefficients. The parameter β in equation (S5.1) is the parameter of interest. We consider the setting that instead of the exposure of interest, VAT, WC is measured. The variable WC is the error-prone substitute measure for VAT, where we assume that $\text{WC} = \theta_1\text{VAT} + U$, where U is a random variable, with mean 0 and variance τ^2, and U is assumed independent of VAT. The factor θ is a scalar, used to scale VAT to the same scale as WC. We also assume *non-differential* measurement error, i.e., $\text{WC}|\text{VAT} \perp\!\!\!\perp Y$. This form of measurement error is referred to as random (or sometimes classical) measurement error if $\theta = 1$ and systematic (or sometimes linear) measurement error otherwise [1, 2]. Since the substitute measure is often measured on a different scale than the true measure, measurement error will often be of the systematic form. Using WC instead of VAT in the linear model yields:

$$\text{E}[\text{IR}|\text{WC}, Z] = \text{intercept}^{*} + \beta^{*}\text{WC} + \gamma^{*'} Z. \tag{S5.2}$$

Under this model, by the law of total expectation, we have $E[IR|WC, Z]$ = intercept + $\beta \times E[VAT|WC, Z] + \gamma^* Z$, which relies on the assumption that the measurement error is non-differential [3]. It follows that,

$$\beta^* = \alpha\beta \quad \text{with} \quad \alpha = \frac{\text{Var}(WC, VAT|Z)}{\text{Var}(WC|Z)}. \tag{S5.3}$$

In conclusion, the ordinary least squared estimator for β^* is biased for β by a factor α. This factor is sometimes referred to as the attenuation factor in case of random measurement error, because in that case $\text{Var}(WC, VAT|Z) < \text{Var}(WC|Z)$ and hence, $\alpha < 1$.

S5.1.1. The different analyses with internal validation samples

When a study contains an internal validation sample for which information is available on both WC and VAT, different analyses can be conducted. Five different estimators are explained below. The variance of these estimators can be obtained from standard output of statistical software when no further details on variance estimation are provided below. The internal validation sample restricted analysis relies on the assumption that the VAT measures in the main study are completely missing at random and the regression calibration methods rely on the assumption that measurement error in WC is non-differential.

Uncorrected analysis. The measurement error is ignored and the relation between VAT and IR is estimated using the error-prone substitute measure WC. Under the assumptions in section S5.1, as shown in equation (S5.3), this estimator is biased by a factor α.

Internal validation sample restricted analysis. The association between VAT and IR is determined using only the data from the internal validation sample (in which a direct measure of VAT is available). This approach will naturally yield unbiased estimates if measures of VAT are missing completely at random in the main study, but power of the study will substantially decrease as only a part of the data available in the main study is used.

Standard regression calibration. The basis of regression calibration is the replacement of WC by a corrected version of WC, based on the regression of VAT on WC and the confounders Z. In this way, the induced measurement error in the uncorrected analysis is corrected by regressing the outcome IR on the confounders Z and $E[VAT|WC, Z]$ instead of WC (i.e., by using the predicted values from regressing VAT on WC and Z, instead of WC). This method is identical to dividing the least squares estimator β^* in equation (S5.2) by the correction factor α defined in equation (S5.3) [2]. The variance of this estimator can be estimated by applying the Delta method described by Rosner et al. [4].

Efficient regression calibration. This analysis pools the estimator of the internal validation sample restricted analysis with the regression calibration estimator, by using weights equal to the inverse of the variance of the two estimates, and was described by Spiegelman et al. [5]. This approach is called efficient regression calibration since it makes use of the fact that in the individuals included in the internal validation sample, VAT is actually known and does not neglect this information. The variance of this estimator can be estimated by taking the inverse of: the sum of the inverse of the variance of the internal validation sample restricted estimator and the inverse of the variance of the regression calibration estimator, as described by Spiegelman et al. [5].

Validation regression calibration. This analysis uses the predicted values from regressing VAT on WC and Z for individuals outside the internal validation sample and VAT otherwise. We call this approach validation regression calibration approach since this is the standard regression calibration approach in internal validation studies [1]. Validation regression calibration treats the predicted values as if they were known and therefore neglects their uncertainty.

S5.2. Simulation study parameters

In the simulation study presented in the main chapter, the measurement error variance τ and the parameter λ in the gamma distribution of the residual errors of VAT were varied according to the R-squared of the measurement error model and skewness of the residuals errors, respectively. The corresponding values for τ and λ in the data generating mechanism found in the main chapter can be found in Table S5.1.

Table S5.1: Values of the parameters R-squared and skewness varied in the simulation study in a full factorial design.The values for τ and λ present the values for that parameter in the data generating mechanism that corresponds to the given R-squared and skewness, respectively.

(a) R-squared and corresponding τ

R-Squared	τ
0.2	1.8
0.4	1.1
0.6	0.7
0.8	0.4
0.9	0.3

(b) Skewness and corresponding λ

Skewness	λ
0.1	65.6
1.0	0.7
1.5	0.3
3.0	0.1

S5.3. Simulation study results

The results of the simulation study that were left out the main chapter for brevity are shown in the following subsections. Full results of the simulation study can also be found on the online repository at `https://github.com/LindaNab/me_neo`. Specifically, `Rds` summary files are available at `https://github.com/LindaNab/me_neo/results/summaries`. These summary files contain more detailed information on e.g. model based standard errors, empirical standard errors and Monte Carlo standard errors. Additionally, output of each single run of the simulation study can be found at `https://github.com/LindaNab/me_neo/data/output` and subsequent folders.

S5.3.1. Internal validation restricted analysis

The main results of the internal validation restricted analysis were shown in the main chapter. Panels A and B in Figure S5.1 show the mean squared error of the association between visceral adipose tissue and insulin resistance using an internal validation sample of 25% of the main study's sample size. Table S5.2 shows the mean squared error of the association under study in the scenarios where R-squared was equal to 0.9 or skewness was equal to 1.0, that were left out the main chapter for brevity. Tables S5.3 and S5.4 show the percentage bias and coverage, respectively, of the association under study in the scenarios

where R-squared was equal to 0.9 or skewness was equal to 1.0.

S5.3.2. Validation regression calibration

The main results of validation regression calibration were shown in the main chapter. Panels C and D in Figure S5.1 show the mean squared error of the association between visceral adipose tissue and insulin resistance using an internal validation sample of 25% of the main study's sample size. Table S5.5 shows the mean squared error of the association under study in the scenarios where R-squared was equal to 0.9 or skewness was equal to 1.0, that were left out the main chapter for brevity. Tables S5.6 and S5.7 show the percentage bias and coverage, respectively, show the percentage bias and coverage of the association under study in the scenarios where R-squared was equal to 0.9 or skewness was equal to 1.0.

S5.3.3. Efficient regression calibration

The results of the application of efficient regression calibration for measurement error correction were as follows. Figure S5.2 shows the mean squared error of the association between visceral adipose tissue and insulin resistance using an internal validation sample of 10%, or 40% of the main study's sample size. Figure S5.3 shows the mean squared error of the association between visceral adipose tissue and insulin resistance using an internal validation sample of 25% of the main study's sample size. Table S5.8 shows the mean squared error of the association under study in the scenarios where R-squared was equal to 0.9 or skewness was equal to 1.0, that were left out Figure S5.2 and S5.3 for comparability with Figure 5.5 and 5.6 in the main chapter. Table S5.9 and S5.10 show the percentage bias in the association between visceral adipose tissue and insulin resistance using an internal validation sample of 10%, 25% or 40% of the main study's sample size for a linear and non-linear measurement error model, respectively. Table S5.11 and S5.12 show the coverage of the association between visceral adipose tissue and insulin resistance using an internal validation sample of 10%, 25% or 40% of the main study's sample size for a linear and non-linear measurement error model, respectively.

S5.3.4. Standard regression calibration

The results of the application of standard regression calibration for measurement error correction were as follows. Table S5.13 and S5.14 show the mean squared error of the association between visceral adipose tissue and insulin resistance using an internal validation sample of 10%, 25% or 40% of the main study's sample size for a linear and non-linear measurement error model, respectively. Table S5.15 and S5.16 show the percentage bias in the association between visceral adipose tissue and insulin resistance using an internal validation sample of 10%, 25% or 40% of the main study's sample size for a linear and non-linear measurement error model, respectively. Table S5.17 and S5.18 show the coverage of the association between visceral adipose tissue and insulin resistance using an internal validation sample of 10%, 25% or 40% of the main study's sample size for a linear and non-linear measurement error model, respectively.

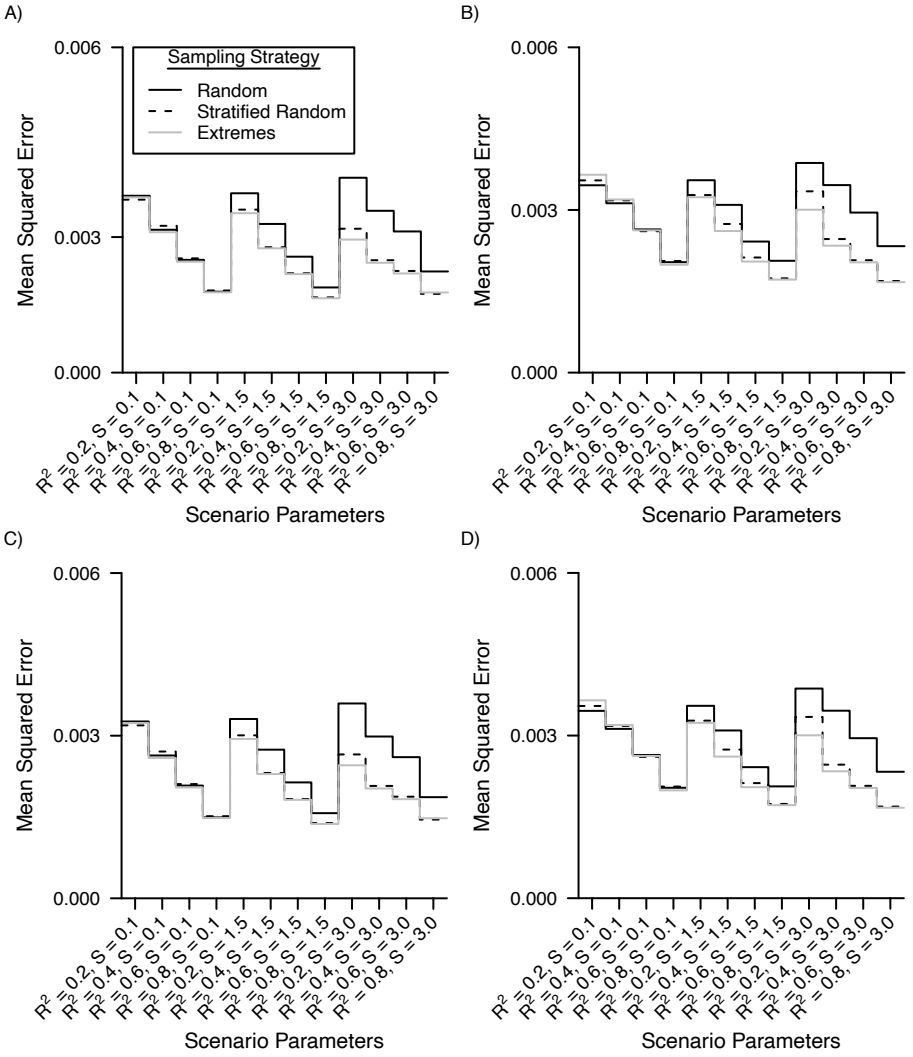


Figure S5.1: Nested loop plot of the mean squared errors in the analysis restricted to the internal validation sample (panels A and B) and the validation regression analysis (panels C and D) for the three different sampling strategies. A and C) Linear measurement error model and an internal validation sample of 25% of the main study; and B and D) Non-linear measurement error model and an internal validation sample of 25% of the main study. Order from outer to inner loops: Skewness of the residual errors of the gold standard measure (S, 3 levels, increasing); R-squared of the measurement error model (R^2, 4 levels, increasing).

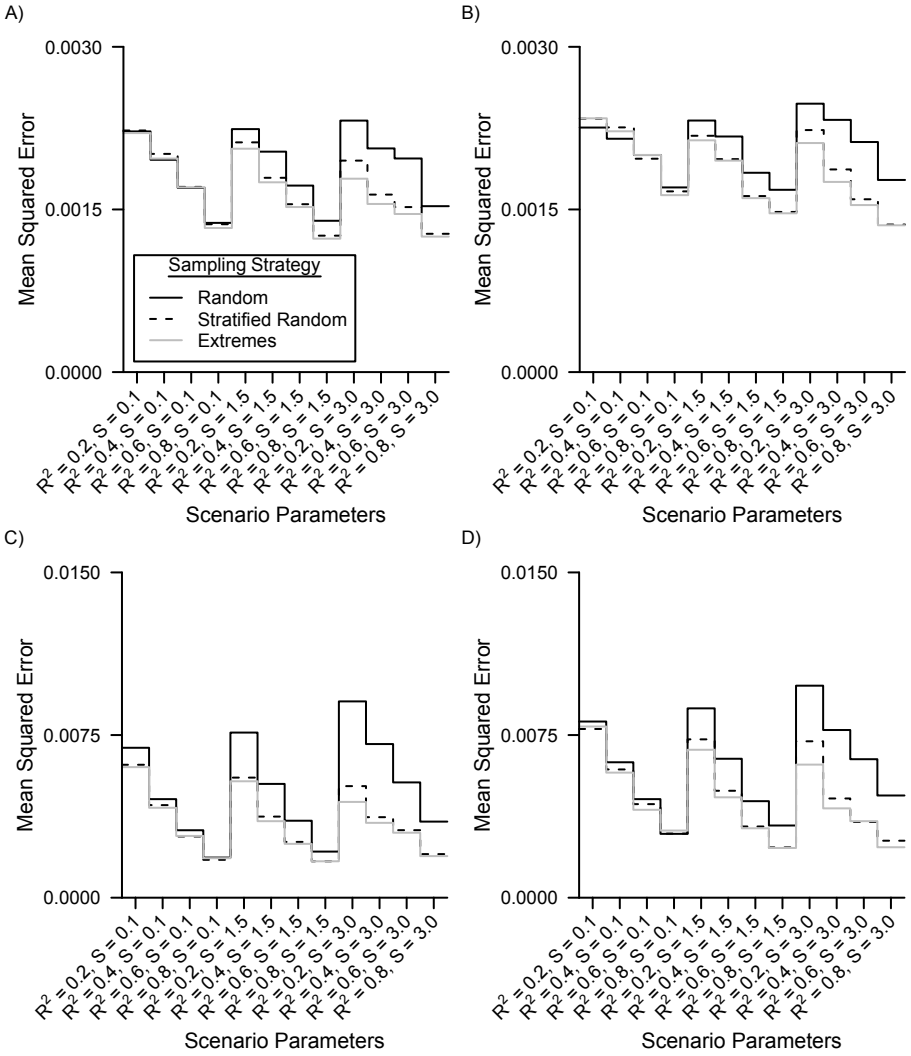


Figure S5.2: Nested loop plot of the mean squared errors in the analysis using efficient regression calibration to correct for the measurement error for the three different sampling strategies. A) Linear measurement error model and an internal validation sample of 40% of the main study; B) Non-linear measurement error model and an internal validation sample of 40% of the main study; C) Linear measurement error model and an internal validation sample of 10% of the main study; and D) Non-linear measurement error model and an internal validation sample of 10% of the main study. Order from outer to inner loops: Skewness of the residual errors of the gold standard measure (S, 3 levels, increasing); R-squared of the measurement error model (R^2, 4 levels, increasing).

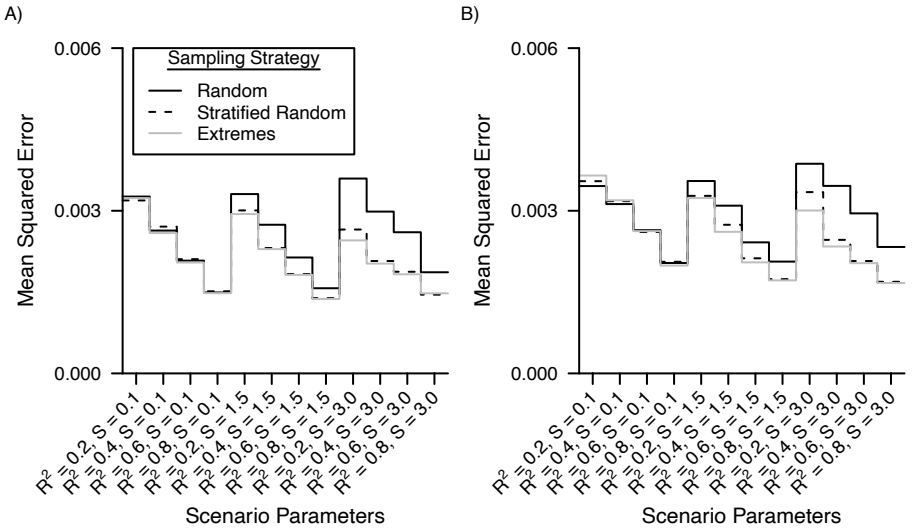


Figure S5.3: Nested loop plot of the mean squared errors in the analysis using efficient regression calibration to correct for the measurement error for the three different sampling strategies. A) Linear measurement error model and an internal validation sample of 25% of the main study; and B) Non-linear measurement error model and an internal validation sample of 25% of the main study. Order from outer to inner loops: Skewness of the residual errors of the gold standard measure (S, 3 levels, increasing); R-squared of the measurement error model (R^2, 4 levels, increasing).

Table S5.2: Mean squared error of the estimated association between visceral adipose tissue and insulin resistance in the analysis restricted to the internal validation sample

Scenario			IVS 40% of Main Study[a]			IVS 25% of Main Study[a]			IVS 10% of Main Study[a]		
Linear	Skew-ness	R^2	Mean Squared Error[b]			Mean Squared Error[b]			Mean Squared Error[b]		
			R	SR	E	R	SR	E	R	SR	E
Yes	0.1	0.9	0.0023	0.0020	0.0019	0.0039	0.0031	0.0031	0.0100	0.0080	0.0084
	1.0	0.2	0.0024	0.0022	0.0022	0.0038	0.0034	0.0033	0.0104	0.0086	0.0080
		0.4	0.0024	0.0021	0.0020	0.0038	0.0031	0.0031	0.0103	0.0077	0.0069
		0.6	0.0024	0.0019	0.0018	0.0039	0.0029	0.0027	0.0100	0.0068	0.0067
		0.8	0.0023	0.0019	0.0018	0.0038	0.0029	0.0027	0.0106	0.0066	0.0070
		0.9	0.0024	0.0019	0.0018	0.0039	0.0028	0.0029	0.0103	0.0064	0.0071
	1.5	0.9	0.0024	0.0018	0.0017	0.0039	0.0025	0.0026	0.0109	0.0052	0.0059
	3.0	0.9	0.0024	0.0015	0.0013	0.0040	0.0020	0.0019	0.0120	0.0036	0.0040
No	0.1	0.9	0.0023	0.0020	0.0019	0.0038	0.0032	0.0031	0.0100	0.0089	0.0081
	1.0	0.2	0.0024	0.0022	0.0022	0.0040	0.0036	0.0034	0.0104	0.0094	0.0085
		0.4	0.0024	0.0021	0.0020	0.0038	0.0033	0.0031	0.0107	0.0084	0.0074
		0.6	0.0024	0.0019	0.0019	0.0038	0.0029	0.0028	0.0102	0.0074	0.0068
		0.8	0.0024	0.0019	0.0018	0.0039	0.0029	0.0027	0.0103	0.0068	0.0064
		0.9	0.0023	0.0018	0.0017	0.0039	0.0027	0.0026	0.0103	0.0069	0.0064
	1.5	0.9	0.0024	0.0016	0.0016	0.0039	0.0024	0.0023	0.0108	0.0060	0.0053
	3.0	0.9	0.0026	0.0015	0.0014	0.0042	0.0019	0.0019	0.0123	0.0038	0.0036

[a] Internal validation sample
[b] For varying sampling strategies of the internal validation sample, R: random, SR: stratified random, E: extremes

Table S5.3: Percentage bias in the estimated association between visceral adipose tissue and insulin resistance in the analysis restricted to the internal validation sample

Scenario			IVS 40% of Main Study[a]			IVS 25% of Main Study[a]			IVS 10% of Main Study[a]		
Linear	Skewness	R^2	Percentage Bias (%)[b]			Percentage Bias (%)[b]			Percentage Bias (%)[b]		
			R	SR	E	R	SR	E	R	SR	E
Yes	0.1	0.9	−0.3	0.0	0.5	−0.5	−0.3	0.5	−0.7	−0.3	0.1
	1.0	0.2	0.3	−0.1	0.1	0.1	−0.4	0.0	0.9	−0.9	0.2
		0.4	−0.1	−0.6	−0.3	0.2	−0.7	−0.7	−0.6	−0.4	−0.3
		0.6	0.0	−0.2	−0.4	−0.2	−0.6	−0.2	0.8	−0.3	−0.3
		0.8	0.5	−0.2	0.1	0.2	−0.3	0.1	0.3	−0.4	0.3
		0.9	−0.2	−0.2	−0.2	−0.3	−0.4	−0.3	0.2	−0.6	−0.8
	1.5	0.9	−0.6	−0.5	−0.3	−1.1	−0.8	−0.4	−1.8	0.1	−0.5
	3.0	0.9	0.4	−0.2	−0.2	1.0	−0.1	−0.3	2.4	−0.1	0.0
No	0.1	0.9	−0.4	−0.1	−0.3	−0.4	−0.5	0.0	−0.7	−0.2	−0.7
	1.0	0.2	0.0	0.3	0.0	−0.6	0.0	−0.4	0.3	0.1	−0.5
		0.4	−0.4	−0.1	0.2	−0.5	0.1	−0.1	−0.7	0.0	0.4
		0.6	0.4	0.2	0.1	0.1	0.5	0.0	−0.1	0.5	0.8
		0.8	0.3	0.2	−0.1	−0.3	0.2	0.0	−0.3	−0.3	0.1
		0.9	0.2	0.2	0.5	0.3	0.4	0.2	0.7	1.0	0.4
	1.5	0.9	−0.5	0.2	0.2	−0.3	−0.2	0.3	0.4	−0.2	0.6
	3.0	0.9	0.1	0.2	0.0	0.7	0.2	0.2	1.7	0.2	0.6

[a] Internal validation sample
[b] For varying sampling strategies of the internal validation sample, R: random, SR: stratified random, E: extremes

Table S5.4: Coverage of the estimated association between visceral adipose tissue and insulin resistance in the analysis restricted to the internal validation sample

Scenario			IVS 40% of Main Study[a] Coverage (%)[b]			IVS 25% of Main Study[a] Coverage (%)[b]			IVS 10% of Main Study[a] Coverage (%)[b]		
Linear	Skewness	R^2	R	SR	E	R	SR	E	R	SR	E
Yes	0.1	0.9	95.1	94.9	95.1	94.9	94.8	94.9	94.6	93.8	94.7
	1.0	0.2	94.9	94.6	94.8	95.2	94.2	94.2	94.5	94.7	94.3
		0.4	94.7	94.9	94.5	94.7	94.9	94.1	94.7	94.5	94.3
		0.6	95.1	94.6	95.1	94.6	94.4	95.3	94.8	94.7	94.3
		0.8	94.9	95.3	95.0	94.9	94.7	95.4	94.5	94.4	94.5
		0.9	94.9	94.5	95.1	94.5	94.6	94.7	94.4	94.1	94.5
	1.5	0.9	95.3	94.5	94.9	95.3	94.4	94.6	94.3	94.6	94.5
	3.0	0.9	95.2	95.5	95.3	95.2	94.6	95.1	94.7	93.9	94.3
No	0.1	0.9	95.3	95.1	94.9	94.9	94.9	94.6	95.4	94.1	94.3
	1.0	0.2	95.0	95.3	94.6	94.2	95.1	94.6	94.5	94.4	94.1
		0.4	94.8	94.7	94.7	95.1	94.4	94.7	94.3	94.7	94.7
		0.6	94.7	95.1	95.0	94.9	95.0	94.7	94.8	95.2	94.7
		0.8	94.7	94.6	95.2	95.1	94.2	94.9	94.5	95.0	94.9
		0.9	94.9	95.4	95.4	94.7	94.7	95.3	94.5	94.7	94.7
	1.5	0.9	95.0	95.2	94.8	94.6	95.3	94.7	94.5	94.3	95.1
	3.0	0.9	94.3	94.8	94.7	94.6	94.5	94.9	94.0	94.6	94.7

[a] Internal validation sample
[b] For varying sampling strategies of the internal validation sample, R: random, SR: stratified random, E: extremes

Table S5.5: Mean squared error of the estimated association between visceral adipose tissue and insulin resistance in the validation regression calibration analysis

Scenario			IVS 40% of Main Study[a] Mean Squared Error[b]			IVS 25% of Main Study[a] Mean Squared Error[b]			IVS 10% of Main Study[a] Mean Squared Error[b]		
Linear	Skewness	R^2	R	SR	E	R	SR	E	R	SR	E
Yes	0.1	0.9	0.0011	0.0011	0.0011	0.0012	0.0012	0.0012	0.0014	0.0013	0.0013
	1.0	0.2	0.0022	0.0021	0.0020	0.0033	0.0030	0.0029	0.0073	0.0058	0.0057
		0.4	0.0019	0.0018	0.0017	0.0025	0.0023	0.0023	0.0052	0.0038	0.0037
		0.6	0.0016	0.0014	0.0014	0.0021	0.0017	0.0017	0.0035	0.0025	0.0025
		0.8	0.0012	0.0012	0.0012	0.0014	0.0013	0.0013	0.0019	0.0016	0.0016
		0.9	0.0011	0.0011	0.0011	0.0012	0.0011	0.0011	0.0014	0.0012	0.0013
	1.5	0.9	0.0011	0.0010	0.0010	0.0012	0.0011	0.0011	0.0015	0.0013	0.0013
	3.0	0.9	0.0012	0.0011	0.0011	0.0014	0.0013	0.0013	0.0022	0.0016	0.0015
No	0.1	0.9	0.0014	0.0013	0.0013	0.0016	0.0015	0.0016	0.0021	0.0020	0.0024
	1.0	0.2	0.0022	0.0021	0.0022	0.0034	0.0032	0.0032	0.0081	0.0070	0.0066
		0.4	0.0021	0.0019	0.0019	0.0030	0.0027	0.0027	0.0067	0.0052	0.0047
		0.6	0.0018	0.0016	0.0016	0.0025	0.0021	0.0021	0.0048	0.0036	0.0032
		0.8	0.0015	0.0014	0.0014	0.0019	0.0016	0.0016	0.0031	0.0023	0.0023
		0.9	0.0013	0.0012	0.0012	0.0015	0.0014	0.0014	0.0021	0.0017	0.0018
	1.5	0.9	0.0013	0.0013	0.0012	0.0016	0.0014	0.0013	0.0024	0.0019	0.0017
	3.0	0.9	0.0016	0.0015	0.0014	0.0020	0.0020	0.0016	0.0043	0.0031	0.0024

[a] Internal validation sample
[b] For varying sampling strategies of the internal validation sample, R: random, SR: stratified random, E: extremes

Table S5.6: Percentage bias in the estimated association between visceral adipose tissue and insulin resistance in the validation regression calibration analysis

Scenario			IVS 40% of Main Study[a]			IVS 25% of Main Study[a]			IVS 10% of Main Study[a]		
Linear	Skewness	R^2	Percentage Bias (%)[b]			Percentage Bias (%)[b]			Percentage Bias (%)[b]		
			R	SR	E	R	SR	E	R	SR	E
Yes	0.1	0.9	−0.1	0.0	0.0	0.0	0.0	0.0	0.3	0.3	0.3
	1.0	0.2	0.0	−0.4	−0.2	−0.5	−1.0	−0.6	−1.2	−2.2	−1.6
		0.4	−0.1	−1.5	−1.2	0.3	−2.5	−1.9	2.2	−3.9	−3.7
		0.6	0.0	−2.1	−2.3	0.2	−3.7	−3.1	2.3	−6.2	−6.4
		0.8	0.1	−2.2	−2.3	0.3	−3.5	−3.0	1.4	−5.6	−4.9
		0.9	0.0	−1.5	−1.6	0.1	−2.3	−1.9	0.6	−3.7	−2.9
	1.5	0.9	0.0	−3.1	−3.4	0.2	−4.6	−4.1	1.0	−6.8	−5.7
	3.0	0.9	0.6	−5.5	−6.8	1.3	−8.2	−8.3	4.1	−11.9	−10.6
No	0.1	0.9	−0.1	−1.6	0.1	−0.1	−1.1	2.3	0.7	0.0	6.4
	1.0	0.2	−0.5	−0.1	−0.3	−1.5	−0.7	−0.9	−3.8	−1.7	−2.7
		0.4	−0.6	−1.0	−0.9	−0.9	−1.5	−2.4	−0.2	−2.6	−6.1
		0.6	0.5	−1.2	−2.0	0.8	−1.8	−4.4	3.1	−1.4	−6.2
		0.8	0.3	−3.2	−2.7	0.7	−3.5	−3.0	2.5	−3.1	−0.4
		0.9	0.2	−4.8	−2.1	0.5	−5.4	−1.3	1.8	−6.5	−0.6
	1.5	0.9	0.2	−6.9	−4.3	0.7	−8.6	−4.4	2.5	−11.3	−5.6
	3.0	0.9	0.7	−11.6	−9.6	1.8	−15.6	−11.6	6.7	−22.1	−17.4

[a] Internal validation sample
[b] For varying sampling strategies of the internal validation sample, R: random, SR: stratified random, E: extremes

Table S5.7: Coverage of the estimated association between visceral adipose tissue and insulin resistance in the validation regression calibration analysis

Scenario			IVS 40% of Main Study[a] Coverage (%)[b]			IVS 25% of Main Study[a] Coverage (%)[b]			IVS 10% of Main Study[a] Coverage (%)[b]		
Linear	Skewness	R^2	R	SR	E	R	SR	E	R	SR	E
Yes	0.1	0.9	94.4	95.0	94.9	94.2	94.6	94.4	93.2	93.4	93.4
	1.0	0.2	95.1	94.7	95.1	94.6	94.5	94.9	91.5	93.5	94.4
		0.4	94.9	94.6	94.9	94.6	94.0	95.0	90.3	91.8	92.6
		0.6	94.4	94.9	95.0	93.9	94.1	94.6	90.6	90.4	91.2
		0.8	94.8	95.0	95.1	94.5	94.4	94.6	92.4	92.3	92.7
		0.9	94.5	94.5	94.3	93.9	94.5	94.1	92.7	93.4	93.5
	1.5	0.9	95.3	94.6	94.7	94.6	94.2	94.4	92.9	91.9	92.4
	3.0	0.9	94.5	93.2	92.4	93.4	90.8	91.2	90.2	86.4	88.1
No	0.1	0.9	94.8	94.7	95.3	94.1	94.4	94.8	92.3	93.0	92.3
	1.0	0.2	95.3	95.7	94.8	94.6	95.4	95.5	92.3	94.5	94.8
		0.4	94.9	94.9	95.2	94.6	94.9	94.9	91.0	93.1	93.5
		0.6	94.7	95.2	94.8	94.2	94.8	94.6	91.2	92.5	93.4
		0.8	94.7	94.5	94.2	94.1	94.3	94.1	90.2	91.8	93.2
		0.9	95.0	94.1	94.5	94.4	93.4	94.9	91.3	90.9	92.9
	1.5	0.9	94.1	92.9	93.8	93.7	91.3	93.7	90.6	86.5	91.1
	3.0	0.9	93.1	87.4	89.7	91.6	81.4	87.4	84.6	64.2	76.9

[a] Internal validation sample
[b] For varying sampling strategies of the internal validation sample, R: random, SR: stratified random, E: extremes

Table S5.8: Mean squared error of the estimated association between visceral adipose tissue and insulin resistance in the efficient regression calibration analysis

Scenario			IVS 40% of Main Study[a] Mean Squared Error[b]			IVS 25% of Main Study[a] Mean Squared Error[b]			IVS 10% of Main Study[a] Mean Squared Error[b]		
Linear	Skewness	R^2	R	SR	E	R	SR	E	R	SR	E
Yes	0.1	0.9	0.0012	0.0012	0.0012	0.0013	0.0013	0.0013	0.0014	0.0014	0.0014
	1.0	0.2	0.0023	0.0022	0.0022	0.0033	0.0032	0.0032	0.0072	0.0059	0.0060
		0.4	0.0020	0.0019	0.0019	0.0026	0.0025	0.0025	0.0049	0.0038	0.0038
		0.6	0.0018	0.0016	0.0016	0.0022	0.0019	0.0019	0.0033	0.0026	0.0026
		0.8	0.0014	0.0013	0.0012	0.0015	0.0014	0.0014	0.0020	0.0017	0.0016
		0.9	0.0012	0.0012	0.0012	0.0013	0.0012	0.0012	0.0015	0.0013	0.0014
	1.5	0.9	0.0012	0.0011	0.0011	0.0013	0.0012	0.0012	0.0015	0.0013	0.0013
	3.0	0.9	0.0013	0.0011	0.0011	0.0014	0.0012	0.0012	0.0021	0.0014	0.0013
No	0.1	0.9	0.0015	0.0014	0.0014	0.0017	0.0017	0.0017	0.0022	0.0021	0.0024
	1.0	0.2	0.0023	0.0022	0.0022	0.0035	0.0033	0.0033	0.0082	0.0073	0.0069
		0.4	0.0022	0.0021	0.0020	0.0031	0.0029	0.0029	0.0064	0.0053	0.0050
		0.6	0.0019	0.0018	0.0018	0.0025	0.0023	0.0023	0.0043	0.0037	0.0034
		0.8	0.0017	0.0015	0.0015	0.0020	0.0018	0.0018	0.0030	0.0024	0.0024
		0.9	0.0014	0.0013	0.0013	0.0017	0.0015	0.0015	0.0022	0.0018	0.0019
	1.5	0.9	0.0015	0.0013	0.0013	0.0017	0.0014	0.0014	0.0023	0.0019	0.0017
	3.0	0.9	0.0016	0.0014	0.0013	0.0000	0.0016	0.0015	0.0037	0.0025	0.0020

[a] Internal validation sample
[b] For varying sampling strategies of the internal validation sample, R: random, SR: stratified random, E: extremes

Table S5.9: Percentage bias in the estimated association between visceral adipose tissue and insulin resistance in the efficient regression calibration analysis for a linear measurement error model

Scenario			IVS 40% of Main Study[a] Percentage Bias (%)[b]			IVS 25% of Main Study[a] Percentage Bias (%)[b]			IVS 10% of Main Study[a] Percentage Bias (%)[b]		
Linear	Skew-ness	R^2	R	SR	E	R	SR	E	R	SR	E
Yes	0.1	0.2	−1.2	−0.3	−0.4	−2.2	−0.6	−0.7	−6.3	−2.5	−2.3
		0.4	−0.6	0.0	−0.1	−0.9	−0.2	−0.3	−1.9	0.0	0.0
		0.6	0.6	0.8	0.5	0.6	0.9	0.7	1.4	1.3	1.2
		0.8	0.0	0.1	0.2	0.0	0.1	0.4	0.8	0.3	0.7
		0.9	−0.1	0.0	0.1	−0.1	−0.1	0.2	0.2	0.2	0.3
	1.0	0.2	−0.6	−0.8	−0.5	−1.9	−1.7	−1.1	−5.8	−4.5	−3.1
		0.4	−0.5	−1.5	−1.2	−0.6	−2.4	−2.2	−1.7	−4.5	−4.2
		0.6	−0.2	−1.6	−1.9	−0.3	−3.0	−2.7	0.7	−5.3	−5.6
		0.8	0.1	−1.6	−1.6	0.1	−2.6	−2.2	1.0	−4.7	−3.9
		0.9	−0.1	−1.1	−1.1	0.0	−1.7	−1.4	0.6	−3.1	−2.3
	1.5	0.2	−1.1	−1.1	−0.7	−2.2	−2.0	−1.7	−5.8	−5.6	−4.7
		0.4	−0.2	−1.5	−1.6	−0.4	−3.1	−2.7	−1.3	−7.4	−7.4
		0.6	0.3	−2.4	−2.4	0.7	−4.5	−3.9	2.1	−8.9	−8.9
		0.8	0.3	−2.4	−2.9	0.4	−4.3	−3.9	1.9	−7.9	−7.0
		0.9	−0.2	−2.2	−2.3	−0.2	−3.4	−3.1	0.6	−5.5	−4.4
	3.0	0.2	−1.0	−1.4	−1.5	−2.1	−3.2	−2.6	−5.4	−9.5	−8.0
		0.4	0.2	−3.3	−3.7	0.4	−7.1	−5.9	0.8	−16.2	−16.0
		0.6	0.7	−5.3	−6.5	1.5	−9.9	−9.7	5.3	−19.2	−19.2
		0.8	0.9	−5.1	−6.4	1.8	−8.4	−8.9	6.2	−14.3	−13.3
		0.9	0.6	−3.6	−4.4	1.3	−5.5	−5.7	4.0	−8.9	−7.5

[a] Internal validation sample
[b] For varying sampling strategies of the internal validation sample, R: random, SR: stratified random, E: extremes

Table S5.10: Percentage bias in the estimated association between visceral adipose tissue and insulin resistance in the efficient regression calibration analysis for a non-linear measurement error model

Scenario			IVS 40% of Main Study[a] Percentage Bias (%)[b]			IVS 25% of Main Study[a] Percentage Bias (%)[b]			IVS 10% of Main Study[a] Percentage Bias (%)[b]		
Linear	Skewness	R^2	R	SR	E	R	SR	E	R	SR	E
No	0.1	0.2	−0.9	−0.4	−0.3	−2.0	−1.2	−0.7	−7.9	−4.8	−2.8
		0.4	−0.3	−0.3	−0.7	−0.9	−0.6	−1.4	−4.2	−1.9	−3.4
		0.6	−0.3	−0.7	−1.2	−0.4	−0.4	−2.5	−2.2	0.3	−3.0
		0.8	0.2	−0.1	0.3	0.4	0.6	1.0	0.7	2.7	4.9
		0.9	−0.2	−1.5	−0.1	−0.2	−1.3	1.3	0.3	−0.3	4.5
	1.0	0.2	−1.1	−0.4	−0.6	−2.8	−1.4	−1.6	−6.9	−4.2	−4.5
		0.4	−1.4	−1.2	−1.2	−2.3	−1.9	−2.9	−5.4	−4.1	−6.8
		0.6	0.2	−0.9	−1.7	−0.1	−1.4	−3.8	−0.7	−1.9	−5.9
		0.8	0.3	−2.2	−2.1	0.3	−2.6	−2.5	1.0	−3.0	−0.9
		0.9	0.3	−3.4	−1.4	0.5	−4.1	−1.2	1.4	−5.4	−0.5
	1.5	0.2	−1.3	−0.6	−0.9	−2.3	−1.6	−1.6	−7.3	−4.8	−4.5
		0.4	−1.4	−1.5	−1.8	−2.1	−2.5	−3.5	−5.5	−5.0	−9.0
		0.6	−0.1	−1.9	−2.6	−0.2	−3.2	−5.5	−0.4	−5.5	−10.0
		0.8	−0.1	−4.0	−3.7	0.2	−5.3	−5.1	1.5	−7.3	−5.6
		0.9	0.0	−4.9	−3.1	0.4	−6.5	−3.4	1.9	−9.5	−4.3
	3.0	0.2	−1.3	−0.8	−0.7	−2.7	−1.6	−1.5	−8.1	−4.6	−5.6
		0.4	−0.9	−2.5	−2.7	−1.7	−4.4	−5.0	−4.2	−9.5	−12.8
		0.6	−0.2	−4.4	−4.8	−0.1	−7.5	−9.3	1.1	−13.3	−18.0
		0.8	0.8	−6.6	−6.4	1.4	−9.8	−9.5	5.3	−15.8	−13.7
		0.9	0.6	−8.0	−6.8	1.6	−11.3	−8.6	5.6	−17.6	−12.8

[a] Internal validation sample
[b] For varying sampling strategies of the internal validation sample, R: random, SR: stratified random, E: extremes

Table S5.11: Coverage of the estimated association between visceral adipose tissue and insulin resistance in the efficient regression calibration analysis for a linear measurement error model

Scenario			IVS 40% of Main Study[a] Coverage (%)[b]			IVS 25% of Main Study[a] Coverage (%)[b]			IVS 10% of Main Study[a] Coverage (%)[b]		
Linear	Skewness	R^2	R	SR	E	R	SR	E	R	SR	E
Yes	0.1	0.2	93.4	92.7	92.6	93.9	93.0	92.4	92.9	93.2	92.2
		0.4	92.1	91.2	90.3	93.4	90.8	91.1	93.6	92.0	91.5
		0.6	90.9	89.7	88.7	92.7	90.4	89.6	93.8	92.3	91.5
		0.8	89.6	88.8	88.5	91.9	89.5	89.6	94.2	92.6	91.5
		0.9	88.5	87.1	87.0	90.6	89.0	88.1	93.2	91.4	91.1
	1.0	0.2	93.5	92.3	91.9	93.5	92.1	91.2	92.3	91.9	91.1
		0.4	92.0	90.2	90.0	92.9	90.7	89.7	92.3	91.5	90.2
		0.6	91.0	89.0	88.5	91.6	89.7	89.4	93.4	89.6	88.7
		0.8	90.5	88.8	88.7	91.7	89.6	90.2	93.5	90.4	90.2
		0.9	88.7	87.5	86.7	90.4	88.8	88.5	92.7	90.2	90.1
	1.5	0.2	93.7	92.7	92.5	93.4	92.5	92.0	91.9	92.5	91.4
		0.4	91.9	91.0	91.0	92.4	90.9	90.4	92.3	88.9	88.8
		0.6	90.8	89.4	88.2	92.1	88.9	88.4	92.7	86.5	85.7
		0.8	89.3	87.7	87.1	91.7	88.2	87.6	92.7	87.3	86.8
		0.9	89.4	86.7	86.6	90.6	87.9	87.6	92.5	89.0	88.5
	3.0	0.2	93.0	92.1	92.5	93.0	92.2	92.4	90.7	88.5	88.8
		0.4	92.0	90.9	89.7	92.3	88.4	88.3	90.9	77.8	79.0
		0.6	89.8	86.3	85.3	90.2	83.0	83.5	90.0	69.2	70.2
		0.8	89.2	85.2	83.4	90.2	83.1	81.9	90.2	76.2	77.4
		0.9	89.1	85.8	84.8	89.8	85.3	84.9	91.4	83.5	84.6

[a] Internal validation sample
[b] For varying sampling strategies of the internal validation sample, R: random, SR: stratified random, E: extremes

Table S5.12: Coverage of the estimated association between visceral adipose tissue and insulin resistance in the efficient regression calibration analysis for a non-linear measurement error model

Scenario			IVS 40% of Main Study[a] Coverage (%)[b]			IVS 25% of Main Study[a] Coverage (%)[b]			IVS 10% of Main Study[a] Coverage (%)[b]		
Linear	Skewness	R^2	R	SR	E	R	SR	E	R	SR	E
No	0.1	0.2	94.2	93.7	93.4	93.6	93.3	92.7	93.0	92.7	92.9
		0.4	93.6	91.8	91.4	93	92.2	91.6	93.2	92.7	90.8
		0.6	92.2	91.3	90.0	92.5	91.6	89.8	93.2	92.5	90.7
		0.8	91.3	89.4	89.6	92.4	90.5	90.3	93.8	92.6	92.2
		0.9	90.4	89.1	88.7	91.3	89.6	89.1	93.7	92.2	92.3
	1.0	0.2	94.3	93.8	92.8	93.6	93.4	93.1	92.0	93.2	92.6
		0.4	93.3	92.2	91.7	93.7	91.9	91.1	92.1	91.8	89.9
		0.6	92.1	90.7	89.9	92.9	91.1	89.3	93.3	91.5	89.5
		0.8	90.3	88.5	88.0	91.8	89.4	89.2	93.0	91.0	90.8
		0.9	90.0	88.1	87.7	91.3	88.4	89.1	93.2	90.1	91.0
	1.5	0.2	94.1	93.5	93.6	94.3	93.9	93.3	92.4	92.9	92.3
		0.4	93.1	91.9	91.5	93.0	91.7	91.2	92.5	91.1	88.8
		0.6	92.7	91.4	90.6	93.5	90.9	89.1	93.3	91.1	87.3
		0.8	90.6	87.3	87.0	91.4	88.1	87.2	91.9	88.9	88.2
		0.9	89.7	86.5	86.9	90.8	87.0	87.7	92.8	85.5	88.9
	3.0	0.2	93.5	92.7	92.8	93.7	93.0	92.9	91.8	92.1	92.2
		0.4	92.6	91.1	90.9	92.5	91.0	89.8	91.1	88.0	85.1
		0.6	91.4	89.3	88.8	91.7	87.4	85.3	90.5	82.1	76.0
		0.8	90.1	85.8	86.0	90.8	83.5	83.2	90.2	76.1	79.0
		0.9	88.6	82.1	83.5	89.6	78.8	82.1	89.6	69.9	78.9

[a] Internal validation sample
[b] For varying sampling strategies of the internal validation sample, R: random, SR: stratified random, E: extremes

Table S5.13: Mean squared error of the estimated association between visceral adipose tissue and insulin resistance in the standard regression calibration analysis for a linear measurement error model

Scenario			IVS 40% of Main Study[a] Mean Squared Error[b]			IVS 25% of Main Study[a] Mean Squared Error[b]			IVS 10% of Main Study[a] Mean Squared Error[b]		
Linear	Skewness	R^2	R	SR	E	R	SR	E	R	SR	E
Yes	0.1	0.2	0.013	0.011	0.011	0.020	0.012	0.011	2.144	0.042	0.015
		0.4	0.005	0.005	0.005	0.006	0.005	0.005	0.013	0.007	0.006
		0.6	0.003	0.003	0.003	0.003	0.003	0.003	0.004	0.003	0.003
		0.8	0.002	0.002	0.002	0.002	0.002	0.002	0.002	0.002	0.002
		0.9	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001
	1.0	0.2	0.014	0.011	0.011	0.094	0.012	0.011	5.839	0.107	0.016
		0.4	0.005	0.005	0.004	0.006	0.005	0.005	0.013	0.006	0.006
		0.6	0.003	0.003	0.002	0.003	0.003	0.003	0.005	0.003	0.003
		0.8	0.002	0.001	0.001	0.002	0.002	0.001	0.002	0.002	0.002
		0.9	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001
	1.5	0.2	0.013	0.011	0.011	0.019	0.012	0.011	2.667	0.067	0.013
		0.4	0.005	0.004	0.004	0.006	0.004	0.004	0.641	0.006	0.005
		0.6	0.003	0.002	0.002	0.003	0.003	0.002	0.006	0.003	0.003
		0.8	0.002	0.001	0.001	0.002	0.002	0.002	0.002	0.002	0.002
		0.9	0.001	0.001	0.001	0.001	0.001	0.001	0.002	0.001	0.001
	3.0	0.2	0.017	0.011	0.010	0.030	0.011	0.011	20.259	1.602	0.016
		0.4	0.007	0.005	0.004	0.011	0.005	0.005	11.778	0.007	0.006
		0.6	0.004	0.003	0.003	0.005	0.003	0.003	0.012	0.005	0.005
		0.8	0.002	0.002	0.002	0.002	0.002	0.002	0.005	0.003	0.003
		0.9	0.001	0.001	0.001	0.001	0.001	0.001	0.002	0.002	0.002

[a] Internal validation sample
[b] For varying sampling strategies of the internal validation sample, R: random, SR: stratified random, E: extremes

Table S5.14: Mean squared error of the estimated association between visceral adipose tissue and insulin resistance in the standard regression calibration analysis for a non linear measurement error model

Scenario			IVS 40% of Main Study[a] Mean Squared Error[b]			IVS 25% of Main Study[a] Mean Squared Error[b]			IVS 10% of Main Study[a] Mean Squared Error[b]		
Linear	Skewness	R^2	R	SR	E	R	SR	E	R	SR	E
No	0.1	0.2	0.167	0.025	0.023	0.355	0.030	0.025	6.808	0.247	0.045
		0.4	0.010	0.009	0.008	0.012	0.009	0.009	0.143	0.014	0.010
		0.6	0.005	0.005	0.004	0.005	0.005	0.004	0.011	0.007	0.005
		0.8	0.003	0.002	0.002	0.003	0.003	0.003	0.004	0.003	0.004
		0.9	0.002	0.002	0.002	0.002	0.002	0.002	0.002	0.002	0.003
	1.0	0.2	0.036	0.020	0.019	46.762	0.023	0.021	19.508	8.612	1.576
		0.4	0.009	0.008	0.007	0.012	0.008	0.008	1.074	0.012	0.008
		0.6	0.004	0.004	0.004	0.005	0.004	0.004	0.014	0.005	0.004
		0.8	0.002	0.002	0.002	0.003	0.002	0.002	0.004	0.003	0.003
		0.9	0.002	0.002	0.002	0.002	0.002	0.002	0.002	0.002	0.002
	1.5	0.2	0.244	0.021	0.020	1.050	0.023	0.021	106.246	9.679	0.116
		0.4	0.009	0.008	0.007	0.011	0.008	0.007	10.060	0.012	0.008
		0.6	0.004	0.004	0.004	0.005	0.004	0.004	0.012	0.005	0.004
		0.8	0.003	0.002	0.002	0.003	0.002	0.002	0.005	0.003	0.003
		0.9	0.002	0.002	0.002	0.002	0.002	0.002	0.003	0.002	0.002
	3.0	0.2	0.036	0.022	0.020	0.402	0.029	0.022	66.440	0.601	0.085
		0.4	0.011	0.007	0.007	0.019	0.008	0.007	2.702	0.021	0.008
		0.6	0.006	0.004	0.004	0.007	0.004	0.005	0.082	0.006	0.006
		0.8	0.003	0.003	0.003	0.004	0.003	0.003	0.012	0.004	0.003
		0.9	0.002	0.002	0.002	0.002	0.003	0.002	0.005	0.004	0.003

[a] Internal validation sample
[b] For varying sampling strategies of the internal validation sample, R: random, SR: stratified random, E: extremes

Table S5.15: Percentage bias in the estimated association between visceral adipose tissue and insulin resistance in the standard regression calibration analysis for a linear measurement error model

Scenario			IVS 40% of Main Study[a] Percentage Bias (%)[b]			IVS 25% of Main Study[a] Percentage Bias (%)[b]			IVS 10% of Main Study[a] Percentage Bias (%)[b]		
Linear	Skewness	R^2	R	SR	E	R	SR	E	R	SR	E
Yes	0.1	0.2	3.5	1.2	0.9	7.4	2.1	1.6	26.9	7.3	4.4
		0.4	0.7	0.4	0.2	1.8	0.8	0.7	6.9	3.4	2.5
		0.6	1.5	1.4	1.3	2.0	1.8	1.6	4.3	2.8	2.6
		0.8	0.3	0.2	0.3	0.4	0.3	0.5	1.2	0.6	1.1
		0.9	0.1	0.0	0.0	0.1	0.1	0.1	0.4	0.3	0.3
	1.0	0.2	3.4	−0.2	−0.7	9.1	−0.1	0.1	10.6	2.9	1.9
		0.4	1.0	−2.2	−2.4	2.3	−3.2	−2.5	9.2	−3.5	−3.6
		0.6	0.1	−3.3	−3.6	0.7	−4.8	−4.3	3.5	−6.7	−7.1
		0.8	−0.1	−2.7	−3.0	0.2	−4.0	−3.5	1.5	−5.8	−5.2
		0.9	0.0	−1.7	−1.8	0.1	−2.5	−2.1	0.7	−3.7	−2.9
	1.5	0.2	2.2	−2.2	−2.3	6.7	−2.6	−1.9	30.8	1.2	−2.5
		0.4	1.7	−4.4	−4.5	3.5	−6.8	−5.3	20.5	−9.6	−9.8
		0.6	1.4	−4.9	−5.5	2.3	−7.8	−6.9	6.8	−12.0	−12.3
		0.8	0.5	−4.4	−5.0	1.1	−6.7	−6.1	2.9	−10.2	−9.3
		0.9	−0.1	−3.4	−3.7	0.1	−4.8	−4.3	1.0	−7.0	−5.8
	3.0	0.2	6.4	−5.6	−6.0	14.1	−9.2	−5.3	108.3	−22.6	−12.2
		0.4	4.1	−9.8	−11.4	8.6	−16.4	−13.2	51.9	−25.6	−25.6
		0.6	2.6	−11.6	−14.2	5.2	−18.1	−17.3	15.4	−27.5	−27.7
		0.8	1.6	−9.5	−11.7	2.8	−13.9	−14.0	8.6	−20.0	−19.0
		0.9	0.7	−6.1	−7.5	1.4	−8.7	−8.8	4.2	−12.1	−10.9

[a] Internal validation sample
[b] For varying sampling strategies of the internal validation sample, R: random, SR: stratified random, E: extremes

Table S5.16: Percentage bias in the estimated association between visceral adipose tissue and insulin resistance in the standard regression calibration analysis for a non-linear measurement error model

Scenario			IVS 40% of Main Study[a]			IVS 25% of Main Study[a]			IVS 10% of Main Study[a]		
Linear	Skewness	R^2	Percentage Bias (%)[b]			Percentage Bias (%)[b]			Percentage Bias (%)[b]		
			R	SR	E	R	SR	E	R	SR	E
No	0.1	0.2	5.6	1.9	0.0	16.4	4.1	1.3	45.8	16.7	7.1
		0.4	3.0	1.2	−0.4	5.3	2.1	−0.9	16.4	6.7	−0.4
		0.6	0.3	−0.2	−2.1	1.3	1.0	−3.3	6.5	5.2	−1.1
		0.8	0.9	0.3	0.5	1.5	1.9	2.2	3.2	5.2	8.9
		0.9	0.1	−2.0	0.4	0.2	−1.2	2.9	1.1	0.3	7.3
	1.0	0.2	7.9	0.7	−0.3	−29.6	2.3	0.6	2.3	−5.5	−4.9
		0.4	1.0	−2.8	−4.2	3.4	−2.9	−5.8	21.2	0.7	−7.9
		0.6	1.9	−2.4	−3.9	2.9	−2.8	−6.8	9.0	−0.6	−7.3
		0.8	0.8	−4.7	−4.0	1.5	−4.5	−4.0	3.7	−3.2	−0.2
		0.9	0.6	−6.0	−2.7	0.8	−6.3	−1.5	2.1	−6.9	−0.5
	1.5	0.2	9.8	0.3	−0.5	14.3	1.9	0.4	−79.8	39.5	4.5
		0.4	1.0	−4.4	−5.4	3.7	−5.1	−7.8	−16.0	−3.0	−11.3
		0.6	1.2	−5.7	−7.2	2.4	−7.1	−10.9	9.5	−6.9	−13.5
		0.8	0.2	−8.3	−7.5	1.2	−9.1	−8.5	4.4	−9.5	−7.3
		0.9	0.5	−8.8	−5.6	0.9	−10.0	−5.3	2.9	−11.9	−6.1
	3.0	0.2	8.9	−1.5	−2.9	26.9	−0.5	−2.9	115.6	1.6	−1.5
		0.4	3.3	−9.1	−9.2	8.8	−12.6	−13.4	29.2	−13.6	−22.0
		0.6	2.3	−13.1	−13.9	5.1	−17.6	−20.0	20.1	−21.1	−28.0
		0.8	1.9	−14.8	−14.1	3.7	−18.6	−17.5	12.3	−22.8	−21.2
		0.9	1.0	−14.9	−12.7	2.3	−18.5	−14.1	7.6	−23.7	−19.0

[a] Internal validation sample
[b] For varying sampling strategies of the internal validation sample, R: random, SR: stratified random, E: extremes

Table S5.17: Coverage of the estimated association between visceral adipose tissue and insulin resistance in the standard regression calibration analysis for a linear measurement error model

Scenario			IVS 40% of Main Study[a] Coverage (%)[b]			IVS 25% of Main Study[a] Coverage (%)[b]			IVS 10% of Main Study[a] Coverage (%)[b]		
Linear	Skewness	R^2	R	SR	E	R	SR	E	R	SR	E
Yes	0.1	0.2	96.9	96.9	96.8	96.8	97.0	97.0	94.9	96.7	97.0
		0.4	95.9	95.9	95.8	96.4	95.9	95.9	95.6	96.0	95.7
		0.6	95.9	96.0	96.1	95.9	96.0	96.2	96.3	96.3	96.7
		0.8	95.8	95.5	95.7	95.6	95.6	95.9	95.9	95.6	95.8
		0.9	94.7	94.7	94.9	94.9	95.0	95.2	95.3	94.9	95.2
	1.0	0.2	96.5	96.4	96.3	96.5	96.3	96.5	94.5	95.6	95.9
		0.4	95.9	95.8	96.1	96.2	95.3	95.7	94.7	94.2	94.1
		0.6	95.7	95.3	95.4	95.3	94.7	95.1	95.3	92.2	93.0
		0.8	95.7	95.5	95.6	95.8	95.0	95.4	95.7	93.8	94.3
		0.9	95.2	95.0	95.0	94.9	94.8	94.9	94.8	94.6	94.9
	1.5	0.2	96.9	96.7	96.9	96.3	96.4	97.0	94.0	94.6	95.5
		0.4	96.2	95.8	96.2	96.2	95.1	95.7	94.6	90.5	92.1
		0.6	96.1	95.2	95.1	95.6	93.5	94.3	94.7	88.5	89.7
		0.8	95.9	95.0	94.9	95.8	93.7	93.9	95.6	90.2	90.8
		0.9	95.5	95.1	95.2	95.2	94.8	95.0	95.4	93.0	93.5
	3.0	0.2	96.4	95.9	96.6	95.8	94.7	96.2	92.5	88.3	91.7
		0.4	95.2	93.5	93.8	95.0	89.6	92.3	92.4	74.3	77.7
		0.6	94.3	91.3	90.0	93.4	84.2	86.3	92.0	64.6	66.8
		0.8	94.4	91.8	90.0	93.2	86.7	87.1	91.3	74.8	78.1
		0.9	94.6	93.2	92.5	94.2	91.0	90.9	93.4	86.9	88.7

[a] Internal validation sample
[b] For varying sampling strategies of the internal validation sample, R: random, SR: stratified random, E: extremes

Table S5.18: Coverage of the estimated association between visceral adipose tissue and insulin resistance in the standard regression calibration analysis for a non-linear measurement error model

Scenario			IVS 40% of Main Study[a] Coverage (%)[b]			IVS 25% of Main Study[a] Coverage (%)[b]			IVS 10% of Main Study[a] Coverage (%)[b]		
Linear	Skewness	R^2	R	SR	E	R	SR	E	R	SR	E
No	0.1	0.2	97.3	97.2	97.2	96.9	97.3	97.3	94.7	96.6	97.0
		0.4	97.2	96.9	96.6	97.4	97.2	96.9	95.6	97.1	96.5
		0.6	96.7	96.3	96.1	96.6	96.3	96.0	95.4	96.8	96.3
		0.8	95.8	95.9	96.1	95.9	96.1	96.1	96.4	96.4	96.6
		0.9	95.6	95.1	95.5	95.9	95.1	95.5	96.1	96.0	96.3
	1.0	0.2	97.4	97.3	97.2	97.0	97.3	97.1	94.8	96.7	96.7
		0.4	96.4	96.3	96.1	96.5	96.2	95.7	94.3	95.4	94.3
		0.6	96.5	96.2	96.0	96.5	95.9	95.0	95.7	95.4	94.1
		0.8	95.7	94.8	95.1	95.5	95.2	94.9	95.3	94.4	95.5
		0.9	95.6	94.1	95.1	95.7	94.2	95.5	95.9	92.9	95.5
	1.5	0.2	97.4	97.3	97.5	96.4	97.0	97.3	94.4	95.9	96.4
		0.4	96.5	96.2	96.2	96.0	95.9	95.5	94.1	94.0	92.9
		0.6	96.4	95.7	95.4	96.5	95.3	94.2	95.0	93.9	91.5
		0.8	94.9	93.6	93.7	94.9	92.7	93.4	94.2	91.1	93.2
		0.9	95.2	92.5	94.6	95.1	91.5	94.5	94.8	88.4	93.1
	3.0	0.2	97.2	97.2	97.1	96.6	96.6	97.0	93.9	94.9	95.4
		0.4	96.5	95.1	95.7	95.9	93.6	94.1	92.9	88.9	86.1
		0.6	95.2	92.9	92.5	94.5	88.9	87.4	92.8	80.5	75.1
		0.8	94.6	89.2	90.1	93.6	84.1	85.9	91.6	73.5	79.5
		0.9	94.1	86.1	89.3	93.3	80.0	87.1	90.4	66.1	78.7

[a] Internal validation sample
[b] For varying sampling strategies of the internal validation sample, R: random, SR: stratified random, E: extremes

References

[1] R. Carroll, D. Ruppert, L. Stefanski, C. Crainiceanu, Measurement error in nonlinear models: A modern perspective, 2nd Edition, Chapman & Hall/CRC, Boca Raton, FL, 2006.

[2] R. H. Keogh, P. A. Shaw, P. Gustafson, R. J. Carroll, V. Deffner, K. W. Dodd, H. Küchenhoff, J. A. Tooze, M. P. Wallace, V. Kipnis, L. S. Freedman, STRATOS guidance document on measurement error and misclassification of variables in observational epidemiology: Part 1—Basic theory and simple methods of adjustment, Statistics in Medicine 39 (16) (2020) 2197–2231. doi:10.1002/sim.8532.

[3] R. H. Keogh, I. R. White, A toolkit for measurement error correction, with a focus on nutritional epidemiology, Statistics in Medicine 33 (12) (2014) 2137–2155. doi:10.1002/sim.6095.

[4] B. Rosner, D. Spiegelman, W. C. Willett, Correction of logistic regression relative risk estimates and confidence intervals for measurement error: The case of multiple covariates measured with error, American Journal of Epidemiology 132 (4) (1990) 734–745. doi:10.1093/oxfordjournals.aje.a115715.

[5] D. Spiegelman, R. J. Carroll, V. Kipnis, Efficient regression calibration for logistic regression in main study/internal validation study designs with an imperfect reference instrument, Statistics in Medicine 20 (1) (2001) 139–160. doi:10.1002/1097-0258(20010115)20:1<139::AID-SIM644>3.0.CO;2-K.

S8

Supplementary material Chapter 8

These are the supplementary materials accompanying Chapter 8. The supplementary materials are structured as follows. In section S8.1, the bias formulas of a conditional model and marginal structural model estimated using inverse probability weighting are derived. Section S8.2 illustrates the application of the bias formulas in a quantitative bias analysis.

S8.1. Quantification of bias due to classification error in a confounding variable

S8.1.1. Conditional model

Under the assumptions and notation described in section 8.2 of the main chapter and by the law of total expectation, the expected value of the outcome Y given the covariates A and L^* is,

$$\begin{aligned}
\mathsf{E}[Y|A,L^*] = \mathsf{E}_{L|A,L^*}[\mathsf{E}[Y|A,L^*,L]] &= \mathsf{E}_{L|A,L^*}[\alpha + \beta A + \gamma L] \\
&= \alpha + \beta A + \gamma \mathsf{E}[L|A,L^*] \\
&= \alpha + \beta A + \gamma \phi_{aL^*} \\
&= \{\alpha + \gamma\phi_{00}\} + \{\beta + \gamma(\phi_{10} - \phi_{00})\}A \\
&+ \{\gamma(\phi_{01} - \phi_{00})\}L^* \\
&+ \gamma(\phi_{11} - \phi_{10} - \phi_{01} + \phi_{00})AL^*,
\end{aligned}$$

which relies on the assumption that L^* is non-differentially misclassified with respect to the outcome (i.e., $L^* \perp\!\!\!\perp Y|L$) and includes an interaction between A and L^*. Further, ϕ_{al^*} is the probability that confounding variable L is one, given that treatment A is a and that misclassified confounding variable L^* is l^*, or,

$$\begin{aligned}
\phi_{al^*} &= P(L = 1|A = a, L^* = l^*) \\
&= \frac{P(A|L = 1, L^* = l^*)P(L = 1|L^* = l^*)}{P(A = a|L^* = l^*)}
\end{aligned}$$

$$
\begin{aligned}
&= \frac{P(A=a|L=1)P(L=1|L^*=l^*)}{P(A=a|L^*=l^*)} \\
&= \frac{P(A=a|L=1)\frac{P(L^*=l^*|L=1)P(L=1)}{P(L^*=l^*)}}{P(A=a|L^*=l^*)} \\
&= \frac{P(A=a|L=1)P(L^*=l^*|L=1)P(L=1)}{P(A=a|L^*=l^*)P(L^*=l^*)} \\
&= \frac{\lambda(1-\pi_1)^{(1-a)}\pi_1^a(1-p_1)^{(1-l^*)}p_1^{l^*}}{(1-\pi_{l^*}^*)^{(1-a)}\pi_{l^*}^{*\,a}(1-\ell)^{(1-l^*)}\ell^{l^*}}.
\end{aligned}
$$

Here $\ell = P(L^* = l^*) = p_0(1-\lambda) + p_1\lambda$ and $\pi_{l^*}^*$ is the probability of receiving treatment A given that the misclassified confounding variable $L^* = l^*$. Note that the above is only defined if $0 < \ell < 1$ and $0 < \pi_{l^*}^* < 1$. To satisfy that $0 < \ell < 1$, we use our assumption that $0 < \lambda < 1$, and additionally, we assume that if $p_0 = 1$ then $p_1 \neq 1$, and if $p_0 = 0$ then $p_1 \neq 0$ (and vice versa). Under the assumption that $0 < \ell < 1$, it follows that,

$$
\begin{aligned}
\pi_{l^*}^* &= P(A=1|L^*=l^*) \\
&= \Sigma_l P(A=1|L^*=l^*, L=l)P(L=l|L^*=l^*) \\
&= \Sigma_l P(A=1|L=l)P(L=l|L^*=l^*) \\
&= \Sigma_l P(A=1|L=l)\frac{P(L^*=l^*|L=l)P(L=l)}{P(L^*=l^*)} \\
&= \Sigma_l \pi_l \frac{(1-p_l)^{(1-l^*)}p_l^{l^*}(1-\lambda)^{1-l}\lambda^l}{(1-\ell)^{1-l^*}\ell^{l^*}},
\end{aligned}
$$

we find that $0 < \pi_{l^*}^* < 1$, if, again, $0 < \lambda < 1$, and if $p_0 = 1$ then $p_1 \neq 1$, and if $p_0 = 0$ then $p_1 \neq 0$ (and vice versa) and $0 < \pi_l < 1$ (positivity assumption).

S8

The bias in the regression based estimator of the effect of A is $\gamma(\phi_{10} - \phi_{00})$ if the interaction between A and L^* is included in the model. However, in this model, the coefficient for A now represents the treatment effect given that L^* is null. Typically, only main effects of A and L^* are included in a regression model of Y conditional on A and L^*:

$$
\begin{aligned}
\mathsf{E}_{AL^*|A,L^*}\{\mathsf{E}[Y|A,L^*]\} &= \{\alpha + \gamma\phi_{00}\} + \{\beta + \gamma(\phi_{10} - \phi_{00})\}A + \{\gamma(\phi_{01} - \phi_{00})\}L^* \\
&+ \gamma(\phi_{11} - \phi_{10} - \phi_{01} + \phi_{00})\mathsf{E}[AL^*|A,L] \\
&= \{\alpha + \gamma\phi_{00} + \delta u_0\} + \{\beta + \gamma(\phi_{10} - \phi_{00}) + \delta u_A\}A \\
&+ \{\gamma(\phi_{01} - \phi_{00}) + \delta u_{L^*}\}L^*,
\end{aligned}
$$

where u_0, u_A, and u_{L^*} are the coefficients of the linear model $\mathsf{E}[AL^*|A,L^*] = u_0 + u_A A + u_{L^*}L^*$ and $\delta = \gamma(\phi_{11} - \phi_{10} - \phi_{01} + \phi_{00})$. Here,

$$
\begin{aligned}
u_A &= \frac{\mathrm{Var}(L^*)\mathrm{Cov}(A,AL^*) - \mathrm{Cov}(A,L^*)\mathrm{Cov}(L^*,AL^*)}{\mathrm{Var}(L^*)\mathrm{Var}(A) - \mathrm{Cov}(A,L^*)^2}, \\
u_{L^*} &= \frac{\mathrm{Var}(A)\mathrm{Cov}(L^*,AL^*) - \mathrm{Cov}(A,L^*)\mathrm{Cov}(A,AL^*)}{\mathrm{Var}(L^*)\mathrm{Var}(A) - \mathrm{Cov}(A,L^*)^2}, \\
u_0 &= \overline{AL^*} - u_A\overline{A} - u_{L^*}\overline{L^*},
\end{aligned}
$$

where $\overline{AL^*}$, $\overline{A}$, and $\overline{L^*}$ denote the mean of A times L^*, A, and L^*, respectively.

If we want to express u_A and u_{L^*} in terms of λ, π_0, π_1, p_0, and p_1, we can write a linear model for A conditional on L^* denoting that $P(A = 1|L^* = l^*) = \pi^*_{l^*}$ and using standard regression theory to get an expression for $\text{Cov}(A, L^*)$:

$$\mathsf{E}[A|L^*] = \pi^*_0 + (\pi^*_1 - \pi^*_0)L^*, \quad \pi^*_1 - \pi^*_0 = \frac{\text{Cov}(A, L^*)}{\text{Var}(L^*)},$$
$$\text{thus} \quad \text{Cov}(A, L^*) = (\pi^*_1 - \pi^*_0)\text{Var}(L^*),$$

where $\text{Var}(L^*) = \ell(1 - \ell)$. Since $\mathsf{E}[AL^*|L^* = 0] = 0$ and $\mathsf{E}[AL^*|L^* = 1] = \mathsf{E}[A|L^* = 1] = \pi^*_1$, it follows,

$$\mathsf{E}[AL^*|L^*] = \pi^*_1 L^*, \quad \pi^*_1 = \frac{\text{Cov}(AL^*, L^*)}{\text{Var}(L^*)}, \quad \text{thus} \quad \text{Cov}(AL^*, L^*) = \pi^*_1\text{Var}(L^*).$$

Equivalently, since $\mathsf{E}[AL^*|A = 0] = 0$ and $\mathsf{E}[AL^*|A = 1] = \mathsf{E}[L^*|A = 1]$, it follows that,

$$\mathsf{E}[AL^*|A] = \mathsf{E}[L^*|A = 1]A = \frac{P(A = 1|L^* = 1)P(L^* = 1)}{P(A = 1)}A,$$
$$\mathsf{E}[L^*|A = 1] = \frac{\pi^*_1\ell}{\omega}, \quad \frac{\pi^*_1\ell}{a} = \frac{\text{Cov}(AL^*, A)}{\text{Var}(A)}, \quad \text{thus} \quad \text{Cov}(AL^*, A) = \frac{\pi^*_1\ell}{\omega}\text{Var}(A).$$

Here, $\text{Var}(A) = \omega(1-\omega)$, and $\text{Var}(L^*) = \ell(1-\ell)$. Denoting that $\omega = P(A = 1) = \pi^*_0(1-\ell)+\pi^*_1\ell$. Combining the different expressions gives,

$$\begin{aligned}
u_A &= \frac{\pi^*_1\ell/\omega\text{Var}(A)\text{Var}(L^*) - \pi^*_1(\pi^*_1 - \pi^*_0)\text{Var}(L^*)^2}{\text{Var}(A)\text{Var}(L^*) - (\pi_1 - \pi_0)^2\text{Var}(L^*)^2} \\
&= \frac{\pi^*_1\ell/\omega\text{Var}(A) - \pi^*_1(\pi^*_1 - \pi^*_0)\text{Var}(L^*)}{\text{Var}(A) - (\pi_1 - \pi_0)^2\text{Var}(L^*)} \\
&= \ell \times \frac{\pi^*_1(1 - \omega) - \pi^*_1(\pi^*_1 - \pi^*_0)(1 - \ell)}{\omega(1 - \omega) - (\pi^*_1 - \pi^*_0)^2\ell(1 - \ell)} \\
&= \ell \times \frac{\pi^*_1 - \pi^{*^2}_1}{(\pi^*_1 - \pi^{*^2}_1)\ell + (\pi^*_0 - \pi^{*^2}_0)(1 - \ell)}, \\
u_{L^*} &= \frac{\pi^*_1\text{Var}(A)\text{Var}(L^*) - \pi^*_1\ell/\omega(\pi^*_1 - \pi^*_0)\text{Var}(A)\text{Var}(L^*)}{\text{Var}(L^*)\text{Var}(A) - ((\pi^*_1 - \pi^*_0)\text{Var}(L^*))^2} \\
&= \frac{\pi^*_1\omega - \pi^*_1\ell(\pi^*_1 - \pi^*_0)}{\omega - (\pi^*_1 - \pi^*_0)^2\text{Var}(L^*)/(1 - \omega)} \\
&= \frac{\pi^*_1\pi^*_0(1 - \pi^{*^2}_1)\ell + \pi^*_1\pi^*_0(1 - \pi^{*^2}_0)(1 - \ell)}{(\pi^*_1 - \pi^{*^2}_1)\ell + (\pi^*_0 - \pi^{*^2}_0)(1 - \ell)}, \\
u_0 &= \overline{AL^*} - u_A\overline{A} - u_{L^*}\overline{L^*}.
\end{aligned}$$

The intercept, the coefficient for A and the coefficient for L^* of the conditional regression model for Y given A and L^* which includes only main effects of A and L^* are, respectively:

$$\alpha + \gamma\phi_{00} + \delta u_0,$$

$$\beta + \gamma(\phi_{10} - \phi_{00})\Big(1 - \ell \times \{\frac{\pi_1^* - \pi_1^{*^2}}{(\pi_1^* - \pi_1^{*^2})\ell + (\pi_0^* - \pi_0^{*^2})(1-\ell)}\}\Big)$$
$$+\gamma(\phi_{11} - \phi_{01})\Big(\ell \times \{\frac{\pi_1^* - \pi_1^{*^2}}{(\pi_1^* - \pi_1^{*^2})\ell + (\pi_0^* - \pi_0^{*^2})(1-\ell)}\}\Big),$$
$$\text{and} \quad \gamma(\phi_{01} - \phi_{00}) + \delta u_{L^*}.$$

S8.1.2. Marginal structural model estimated using inverse probability weighting

Under the assumptions described in section 8.2 of the main chapter, an MSM-IPW under model (8.2) is estimated by fitting a linear regression model for A on Y, where each subject i is weighted by 1 over the probability of that subject's observed exposure given the misclassified confounding variable L^*. Hence, an MSM-IPW proceeds by solving the weighted regression model,

$$\sum_{i=1}^{n} \frac{1}{P(A_i|L_i^*)}(Y_i - \alpha_{\text{msm}} - \beta A_i) = 0 \quad \text{and} \quad \sum_{i=1}^{n} \frac{A_i}{P(A_i|L_i^*)}(Y_i - \alpha_{\text{msm}} - \beta A_i) = 0.$$

Solving these equations for α_{msm} and β result in the following estimators:

$$\hat{\alpha}_{\text{msm}} = \overline{Y}_{w^*} - \hat{\beta}_{\text{msm}}\overline{A}_{w^*} \quad \text{and} \quad \hat{\beta} = \frac{\sum_{i=1}^{n} \frac{1}{P(A_i|L_i)}(Y_i - \overline{Y}_{w^*})(A_i - \overline{A}_{w^*})}{\sum_{i=1}^{n} \frac{1}{P(A_i|L_i)}(A_i - \overline{A}_{w^*})^2},$$

where,

$$\overline{Y}_{w^*} = \frac{\sum_{i=1}^{n} Y_i/P(A_i|L_i^*)}{\sum_{i=1}^{n} 1/P(A_i|L_i^*)} \quad \text{and} \quad \overline{A}_{w^*} = \frac{\sum_{i=1}^{n} A_i/P(A_i|L_i^*)}{\sum_{i=1}^{n} 1/P(A_i|L_i^*)}.$$

S8

Let n_{al}^* be the number of subjects with $A = a$ and $L^* = l^*$ and n_{al} be the number of subjects with $A = a$ and $L = l$. In a population of n subjects,

$$\begin{aligned}
n_{00}^* &= nP(A = 0, L^* = 0) = nP(A = 0|L^* = 0)P(L^* = 0) \\
&= n\sum^{l} P(A = 0|L = l, L^* = 0)P(L = l|L^* = 0)P(L^* = 0) \\
&= n\sum^{l} P(A = 0|L = l)P(L = l|L^* = 0)P(L^* = 0) \\
&= n\sum^{l} P(A = 0|L = l)P(L = l)P(L^* = 0|L = l) \\
&= n_{00}(1 - p_0) + n_{01}(1 - p_1),
\end{aligned}$$

which relies on the assumption that L^* is non-differentially misclassified with respect tot the exposure (i.e., $L^* \perp\!\!\!\perp A|L$). Equivalently,

$$n_{01}^* = n_{00}p_0 + n_{01}p_1, \quad n_{10}^* = n_{10}(1 - p_0) + n_{11}(1 - p_1),$$

$$\text{and} \quad n_{11}^* = n_{10}p_0 + n_{11}p_1.$$

Hence,

$$\begin{aligned}
\sum_{i=1}^{n} 1/P(A_i|L_i^*) &= \sum_{i=1}^{n} \frac{1}{\sum_l [P(A_i|L_i^*, L=l)P(L=l|L_i^*)]} \\
&= \sum_{i=1}^{n} \frac{1}{\sum_l [P(A_i|L=l)P(L=l|L_i^*)]} \\
&= \sum^{n_{00}^*} \frac{1}{\sum_l [(1-\pi_l)P(L=l|L^*=0)]} \\
&+ \sum^{n_{01}^*} \frac{1}{\sum_l [(1-\pi_l)P(L=l|L^*=1)]} \\
&+ \sum^{n_{10}^*} \frac{1}{\sum_l [\pi_l P(L=l|L^*=0)]} \\
&+ \sum^{n_{11}^*} \frac{1}{\sum_l [\pi_l P(L=l|L^*=1)]}.
\end{aligned}$$

Here,

$$\begin{aligned}
\sum^{n_{00}^*} \frac{1}{\sum_l [(1-\pi_l)P(L=l|L^*=0)]} &= \\
\frac{n_{00}(1-p_0) + n_{01}(1-p_1)}{(1-\pi_0)P(L=0|L^*=0) + (1-\pi_1)P(L=1|L^*=0)} &= \\
\frac{n_{00}(1-p_0) + n_{01}(1-p_1)}{(1-\pi_0)\frac{P(L^*=0|L=0)(1-\lambda)}{P(L^*=0)} + (1-\pi_1)\frac{P(L^*=0|L=1)\lambda}{P(L^*=0)}} &= \\
\frac{n_{00}(1-p_0) + n_{01}(1-p_1)}{\frac{n_{00}}{nP(L^*=0)}(1-p_0) + \frac{n_{01}}{nP(L^*=0)}(1-p_1)} &= \\
\frac{1}{1/(nP(L^*=0))} &= \\
nP(L^*=0) &= n(1-\ell), \\
\sum^{n_{01}^*} \frac{1}{\sum_l [(1-\pi_l)P(L=l|L^*=1)]} &= \\
nP(L^*=1) &= n\ell, \\
\sum^{n_{10}^*} \frac{1}{\sum_l [\pi_l P(L=l|L^*=0)]} &= \\
nP(L^*=0) &= n(1-\ell), \\
\sum^{n_{11}^*} \frac{1}{\sum_l [\pi_l P(L=l|L^*=1)]} &=
\end{aligned}$$

$$nP(L^* = 1) = n\ell.$$

From these expressions it follows that,

$$\sum_{i=1}^{n} 1/P(A_i|L_i^*) = 2n(1-\ell) + 2n\ell = 2n.$$

Further,

$$\begin{aligned}
\sum_{i=1}^{n} \mathsf{E}[Y_i]/P(A_i|L_i^*) &= \sum^{n_{00}^*} \frac{\mathsf{E}[Y_i]}{\sum_l[(1-\pi_l)P(L=l|L^*=0)]} \\
&+ \sum^{n_{01}^*} \frac{\mathsf{E}[Y_i]}{\sum_l[(1-\pi_l)P(L=l|L^*=1)]} \\
&+ \sum^{n_{10}^*} \frac{\mathsf{E}[Y_i]}{\sum_l[\pi_l P(L=l|L^*=0)]} \\
&+ \sum^{n_{11}^*} \frac{\mathsf{E}[Y_i]}{\sum_l[\pi_l P(L=l|L^*=1)]} \\
&= \sum^{n_{00}^*} \frac{\alpha + \gamma P(L=1|A=0,L^*=0)}{\sum_l[(1-\pi_l)P(L=l|L^*=0)]} \\
&+ \sum^{n_{01}^*} \frac{\alpha + \gamma P(L=1|A=0,L^*=1)}{\sum_l[(1-\pi_l)P(L=l|L^*=1)]} \\
&+ \sum^{n_{10}^*} \frac{\alpha + \beta + \gamma P(L=1|A=1,L^*=0)}{\sum_l[\pi_l P(L=l|L^*=0)]} \\
&+ \sum^{n_{11}^*} \frac{\alpha + \beta + \gamma P(L=1|A=1,L^*=1)}{\sum_l[\pi_l P(L=l|L^*=1)]} \\
&= n\alpha(1-\ell) + n\gamma(1-\ell)\phi_{00} + n\alpha\ell + n\gamma\phi_{01} \\
&+ n(\alpha+\beta)(1-\ell) + n\gamma(1-\ell)\phi_{10} \\
&+ n(\alpha+\beta)\ell + n\gamma\phi_{11} \\
&= 2n\alpha + n\beta + n\gamma(1-\ell)(\phi_{00}+\phi_{10}) + n\gamma\ell(\phi_{01}+\phi_{11}),
\end{aligned}$$

and,

$$\begin{aligned}
\sum_{i=1}^{n} A_i/P(A_i|L_i) &= \sum^{n_{10}^*} \frac{1}{\sum_l[\pi_l P(L=l|L^*=0)]} + \sum^{n_{11}^*} \frac{1}{\sum_l[\pi_l P(L=l|L^*=1)]} \\
&= n(1-p_0)(1-\lambda) + n(1-p_1)\lambda + np_0(1-\lambda) + np_1\lambda = n.
\end{aligned}$$

Combining these expressions leads to,

$$\mathsf{E}[\overline{Y}_{w^*}] = \alpha + \beta/2 + \gamma/2(1-\ell)(\phi_{00}+\phi_{10}) + \gamma/2\ell(\phi_{01}+\phi_{11})$$

$$\text{and} \quad \overline{A}_{w^*} = n/2n = 1/2, \quad \text{and}$$

$$\begin{aligned}
\sum_{i=1}^{n} \frac{(A_i - \overline{A}_{w^*})^2}{P(A_i|L_i^*)} &= \sum^{n_{00}^*} \frac{(-1/2)^2}{\sum_l[(1-\pi_l)P(L=l|L^*=0)]} \\
&+ \sum^{n_{01}^*} \frac{(-1/2)^2}{\sum_l[(1-\pi_l)P(L=l|L^*=1)]} \\
&+ \sum^{n_{10}^*} \frac{(1-1/2)^2}{\sum_l[\pi_l P(L=l|L^*=0)]} \\
&+ \sum^{n_{11}^*} \frac{(1-1/2)^2}{\sum_l[\pi_l P(L=l|L^*=1)]} \\
&= 1/4 \times \sum_{i=1}^{n} 1/P(A_i|L_i^*) = n/2.
\end{aligned}$$

Further,

$$\begin{aligned}
\sum_{i=1}^{n} \frac{\mathsf{E}[(Y_i - \overline{Y}_{w^*})](A_i - \overline{A}_{\tilde{w}})}{P(A_i|L_i^*)} &= \\
\sum^{n_{00}^*} \frac{\beta/4 - \gamma/2\phi_{00} + \gamma/4(1-\ell)(\phi_{00}+\phi_{10}) + \gamma/4\ell(\phi_{01}+\phi_{11})}{\sum_l[(1-\pi_l)P(L=l|L^*=0)]} &+ \\
\sum^{n_{01}^*} \frac{\beta/4 - \gamma/2\phi_{01} + \gamma/4(1-\ell)(\phi_{00}+\phi_{10}) + \gamma/4\ell(\phi_{01}+\phi_{11})}{\sum_l[(1-\pi_l)P(L=l|L^*=1)]} &+ \\
\sum^{n_{10}^*} \frac{\beta/4 + \gamma/2\phi_{10} - \gamma/4(1-\ell)(\phi_{00}+\phi_{10}) - \gamma/4\ell(\phi_{01}+\phi_{11})}{\sum_l[\pi_l P(L=l|L^*=0)]} &+ \\
\sum^{n_{11}^*} \frac{\beta/4 + \gamma/2\phi_{11} - \gamma/4(1-\ell)(\phi_{00}+\phi_{10}) - \gamma/4\ell(\phi_{01}+\phi_{11})}{\sum_l[\pi_l P(L=l|L^*=0)]} &= \\
n(1-\ell)(\beta/4 - \gamma/2\phi_{00} + \gamma/4(1-\ell)(\phi_{00}+\phi_{10}) + \gamma/4\ell(\phi_{01}+\phi_{11})) &+ \\
n\ell(\beta/4 - \gamma/2\phi_{01} + \gamma/4(1-\ell)(\phi_{00}+\phi_{10}) + \gamma/4\ell(\phi_{01}+\phi_{11})) &+ \\
n(1-\ell)(\beta/4 + \gamma/2\phi_{10} - \gamma/4(1-\ell)(\phi_{00}+\phi_{10}) - \gamma/4\ell(\phi_{01}+\phi_{11})) &+ \\
n\ell(\beta/4 + \gamma/2\phi_{11} - \gamma/4(1-\ell)(\phi_{00}+\phi_{10}) - \gamma/4\ell(\phi_{01}+\phi_{11})) &= \\
n/2(\beta(1-\ell) + \beta\ell - \gamma(1-\ell)\phi_{00} - \gamma\ell\phi_{01} + \gamma(1-\ell)\phi_{10} + \gamma\ell\phi_{11}) &= \\
n/2(\beta + \gamma(1-\ell)(\phi_{10}-\phi_{00}) + \gamma\ell(\phi_{11}-\phi_{01}).
\end{aligned}$$

The above mentioned leads to the following expression for the expected estimated value of the effect of A, based on the MSM-IPW,

$$\mathsf{E}[\hat{\beta}] = \beta + \gamma(\phi_{10}-\phi_{00})(1-\ell) + \gamma(\phi_{11}-\phi_{01})\ell \quad \text{and}$$
$$\mathsf{E}[\hat{\alpha}_{\text{msm}}] = \alpha + \gamma/2 \times [2(1-\ell)\phi_{00} + 2\ell\phi_{01})] = \alpha + \gamma\phi_{00}(1-\ell) + \gamma\phi_{01}\ell.$$

S8.2. Illustration: quantitative bias analysis

Using an example study of blood pressure lowering therapy, we illustrate how the bias expressions in section 8.3 of the main chapter can be used to perform a quantitative bias analysis for misclassification of a confounding variable. For our illustration we use data of the National Health And Nutritional Examination Survey (NHANES) [1, 2]. Specifically, we study the average treatment effect of diuretic use ($A = 1$) in comparison to beta blocker use ($A = 0$) on systolic blood pressure (Y) using two approaches: by inverse weighting with the propensity for diuretic or beta blocker use given self-reported categorical body mass index (BMI) (L^*), and using a conditional linear regression with adjustment for self-reported categorical BMI. This supplement comprises background material that complements the motivating example in the main chapter. Additionally, equations are derived to inform the quantitative bias analysis.

NHANES. The NHANES survey consists of questionnaires, followed by a standardized health examination in specially equipped mobile examination centers. In the 2011-2014 sample 19,151 participants were physically examined. Of the 19,151 physically examined people, 12,185 participants aged over 16 were asked to fill out a questionnaire, including questions on self-reported weight and height, used to calculate self-reported BMI. For this illustration, we used complete data on 585 users of diuretics and 824 users of beta blockers (excluding non-users and people using both).

Parameters estimated in NHANES. In the NHANES data, it was found that the prevalence of self-reported overweight/obese was 0.77 (ℓ), the probability of receiving treatment given that one self-reports to be underweight/normal weight is 0.32 (π_0^*), the probability of receiving treatment given that one self-reports to be overweight/obese is 0.44 (π_1^*). Finally, the association between L^* and Y, given that $A = 0$ estimated in a conditional regression model including an interaction between A and L* was -6.63.

BMI measured by trained technicians. In the NHANES, anthropometric measures were also taken by trained health technicians. By using these measures to calculate BMI category, we found that the specificity of self-reported BMI category was 0.94 (p_1), and the sensitivity was 0.92 ($p_0 = 0.08$). The average treatment effect (95 % CI) of diuretics use in comparison to beta blocker use on mean blood pressure was -3.59 (-5.84; -1.35) estimated using MSM-IPW (by inverse weighting with the propensity for diuretic or beta blocker use given categorical BMI). Given that a subject is not overweight/obese, the fitted weights were 1.48 and 3.09 for beta blocker and diuretics use, respectively. Given that a subject is overweight/obese, the fitted weights were 1.77 and 2.30, respectively. In comparison, if self-reported categorical BMI was used, the fitted weights slightly differed: 1.46, 3.17, 1.79 and 2.26, respectively. Consequently, estimates of the average treatment effect differed, depending on the BMI measure that was used to calculate the inverse probability weights (-3.59 using categorical BMI versus -3.52 using categorical self-reported BMI (Table 8.3, main chapter)).

Performing a quantitative bias analysis. To inform a quantitative bias analysis, one needs to specify the bias parameters for sensitivity (p_1) and specificity ($1-p_0$) using external validation data, internal validation data, or an educated guess. From the data, one can estimate the prevalence of misclassified confounding variable L^* (i.e., ℓ), the probability of receiving treatment given that L^* is null (i.e., π_0^*) and the probability of receiving treatment given that L^* is one (i.e., π_1^*). We calculate the probability of receiving treatment given that L is null or one (i..e, π_0, and π_1, respectively) using the data and the assumed values of p_0

and p_1. Since,

$$\pi_0^* = \frac{\pi_0(1-p_0)(1-\lambda) + \pi_1(1-p_1)\lambda}{(1-\ell)}, \quad \text{and} \quad \pi_1^* = \frac{\pi_0 p_0(1-\lambda) + \pi_1 p_1 \lambda}{\ell},$$

it follows that if $p_0 = 1$, $\pi_1 = \pi_0^*$ and if $p_1 = 0$, $\pi_0 = \pi_1^*$ (using that $0 < \ell < 1$, as used in S8.1 section S8.1.1). Further, if $p_0 = 1$ and $0 < p_1 < 1$, we obtain,

$$\pi_1 = \pi_0^*, \quad \text{and} \quad \pi_0 = \frac{\pi_0^* p_1 \lambda - \pi_1^* \ell}{(1-\lambda)}.$$

Additionally, if $p_1 = 0$ and $0 < p_0 < 1$, we obtain

$$\pi_0 = \pi_1^*, \quad \text{and} \quad \pi_1 = \frac{\pi_0^*(1-\ell) - \pi_1^*(1-p_0)(1-\lambda)}{\lambda}.$$

If we assume that $p_0 \neq 1$ and $p_1 \neq 0$ and use our assumption that $0 < \lambda < 1$, it follows that,

$$\pi_0 = \frac{\pi_0^*(1-\ell) - \pi_1(1-p_1)\lambda}{(1-p_0)(1-\lambda)}, \quad \pi_1 = \frac{\pi_1^* \ell - \pi_0 p_0(1-\lambda)}{p_1 \lambda}. \tag{S8.1}$$

By rewriting the expression for π_1 using the expression for π_0, it follows that,

$$\begin{aligned}
\pi_1 &= \frac{\pi_1^* \ell - \pi_0 p_0(1-\lambda)}{p_1 \lambda} \\
&= \frac{\pi_1^* \ell - \frac{\pi_0^*(1-\ell) - \pi_1(1-p_1)\lambda}{(1-p_0)(1-\lambda)} p_0(1-\lambda)}{p_1 \lambda} \\
&= \frac{\pi_1^* \ell - (\pi_0^*(1-\ell) - \pi_1(1-p_1)\lambda)\frac{p_0}{(1-p_0)}}{p_1 \lambda} \\
&= \frac{\pi_1^* \ell - \pi_0^*(1-\ell)\frac{p_0}{(1-p_0)} + \frac{(1-p_1)p_0}{(1-p_0)}\lambda \pi_1}{p_1 \lambda} \\
&= \frac{\pi_1^* \ell - \pi_0^*(1-\ell)\frac{p_0}{(1-p_0)}}{p_1 \lambda} + \frac{(1-p_1)p_0}{(1-p_0)p_1}\pi_1 \\
&= \frac{\pi_1^* \ell - \pi_0^*(1-\ell)\frac{p_0}{(1-p_0)}}{p_1 \lambda} + \frac{(1-p_1)p_0}{(1-p_0)p_1}\pi_1.
\end{aligned}$$

Consequently,

$$\begin{aligned}
(1 - \frac{(1-p_1)p_0}{(1-p_0)p_1})\pi_1 &= \frac{\pi_1^* \ell - \pi_0^*(1-\ell)\frac{p_0}{(1-p_0)}}{p_1 \lambda}, \\
\pi_1 &= \frac{\frac{\pi_1^* \ell - \pi_0^*(1-\ell)\frac{p_0}{(1-p_0)}}{p_1 \lambda}}{\frac{(1-p_0)p_1 - (1-p_1)p_0}{(1-p_0)p_1}}
\end{aligned}$$

$$= \frac{\pi_1^*\ell - \pi_0^*(1-\ell)\frac{p_0}{(1-p_0)}}{p_1\lambda} \times \frac{(1-p_0)p_1}{(1-p_0)p_1 - (1-p_1)p_0}. \tag{S8.2}$$

From expression (S8.2) we now obtain a value for π_1, which we use to obtain a value for π_0 from expression (S8.1). We calculate the prevalence of L (i.e., λ) by,

$$\lambda = p_0, \quad \text{if } p_0 = p_1 \quad \text{and} \quad \lambda = \frac{\ell - p_0}{p_1 - p_0} \quad \text{otherwise.}$$

Subsequently, the expressions for π_0, π_1 and λ can be used to obtain estimates for ϕ_{al^*} using the expression in section Conditional model. Lastly, an estimate for γ can be obtained by fitting a conditional regression model on Y given A and L^*, including an interaction between A and L^*. The coefficient for L^* from this model is then divided by $(\phi_{01} - \phi_{00})$ to get an estimate for γ, holding that $\phi_{01} \neq \phi_{00}$. The inequality $\phi_{01} \neq \phi_{00}$ holds if $p_0 \neq p_1$, in the case that $p_0 = p_1$, γ is not identifiable from the data (and thus, bias is not identifiable). The bias expressions (8.3) and (8.4) in the main chapter of the article can subsequently be used to calculate bias in the average treatment effect estimator.

References

[1] Centers for Disease Control and Prevention (CDC). National Center for Health Statistics (NCHS), National health and nutrition examination survey data (2011).
URL https://wwwn.cdc.gov/nchs/nhanes/continuousnhanes/default.aspx?BeginYear=2011

[2] Centers for Disease Control and Prevention (CDC). National Center for Health Statistics (NCHS), National health and nutrition examination survey data (2013).
URL https://wwwn.cdc.gov/nchs/nhanes/continuousnhanes/default.aspx?BeginYear=2013

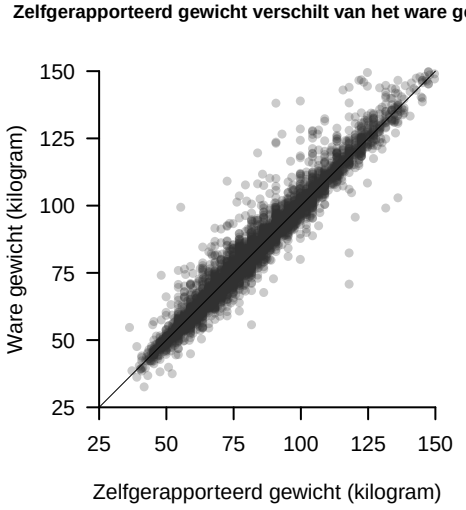
A

Dutch summary

Meetfouten komen vaak voor in epidemiologisch onderzoek. Neem bijvoorbeeld een epidemiologisch onderzoek waarin het verband tussen een bepaalde blootstelling en uitkomst wordt onderzocht. Om dit verband te onderzoeken, zijn gegevens nodig: in een groep mensen (de studiepopulatie) wordt dan bijvoorbeeld eerst de blootstelling gemeten en vervolgens de uitkomst. Vaak worden er meetfouten gemaakt in deze metingen, bijvoorbeeld omdat gegevens uit elektronische patiënten dossiers worden gebruikt voor het epidemiologisch onderzoek. Deze dossiers zijn in eerste instantie niet bedoeld voor het doen van epidemiologisch onderzoek en mogelijk zijn daarom de metingen niet met de gewenste wetenschappelijke precisie gedaan. Zo zou het kunnen zijn dat het lichaamsgewicht dat geregistreerd staat in deze dossiers gebaseerd is op de vraag "hoeveel weeg jij?" en niet op een nauwkeurige meting van het lichaamsgewicht met een weegschaal. Figuur A.1 illustreert de discrepantie tussen zelfgerapporteerd lichaamsgewicht en lichaamsgewicht gemeten met een nauwkeurige (gevalideerde) weegschaal in een Amerikaanse studie. Als er geen meetfout zou zitten in het zelfgerapporteerde lichaamsgewicht in deze studie dan hadden de puntjes in de grafiek allemaal op de 45 graden lijn gelegen.

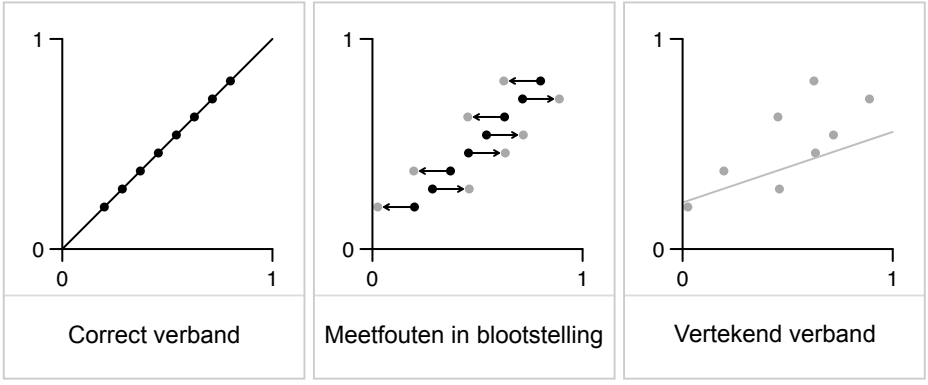
In dit proefschrift wordt verslag gedaan van de invloed van meetfouten in de gegevens (ook wel, data) die gebruikt worden in epidemiologisch onderzoek. Daarnaast wordt beschreven hoe we voor deze meetfouten kunnen corrigeren in epidemiologisch onderzoek. Voor het corrigeren van meetfouten is vaak informatie nodig over het 'meetfoutmodel': het verband tussen de meting die gedaan wordt (de meting met meetfout) en de 'ware' waarde van datgene dat de meting beoogd te meten.

Om in een epidemiologisch onderzoek een verband (of associatie) te schatten tussen een blootstelling en een uitkomst, wordt meestal een statistisch model gebruikt. Parameters van een statistisch model kunnen worden geschat bijvoorbeeld met lineaire of logistische regressie-analyse. In een regressie-analyse wordt een regressiecoëfficiënt geschat: dit is een numerieke waarde die het verband (of associatie) tussen de blootstelling en uitkomst uitdrukt. Meetfouten kunnen bepaalde parameters van dit model beïnvloeden, zoals wordt geïllustreerd in figuur A.2. Een statistisch model bestaat uit afhankelijke variabelen (de uitkomst) en onafhankelijke variabelen (covariaten, waaronder de blootstelling en soms ook zogenaamde 'confounders'). In alle soorten variabelen kunnen meetfouten voorkomen.



Figuur A.1: Discrepantie tussen lichaamsgewicht in kilogram gemeten met een gekalibreerde weegschaal (ware gewicht) and zelfgerapporteerd gewicht in de Amerikaanse 'National Health and Nutrition Examination Survey' studie in 2017-2018.

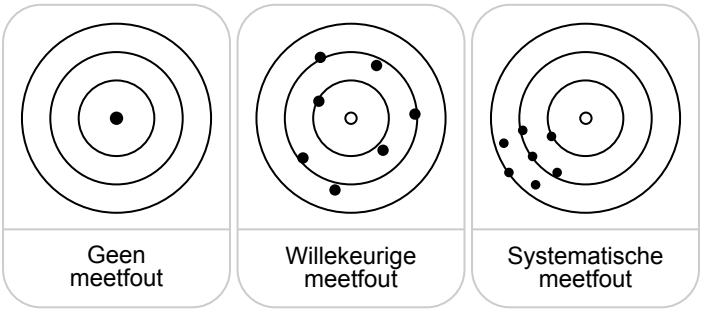
Afhankelijk van het 'type' meetfout, de grootte van de meetfout en in welke variabele(n) de meetfout zit (de afhankelijke of onafhankelijke variabele(n)), kan een meetfout de af te schatten regressiecoëfficiënt vertekenen: het verband (of associatie) tussen de blootstelling en uitkomst kan bijvoorbeeld sterker of juist zwakker lijken dan dat deze daadwerkelijk is.



Figuur A.2: Vereenvoudigde weergave van de invloed van meetfouten in de blootstelling op een lineaire regressie-analyse die het verband tussen een blootstelling (horizontale as) en uitkomst (verticale as) schat.

De verschillende typen meetfouten die aan bod komen in dit proefschrift zijn geïllustreerd in figuur A.3. Onderscheid wordt gemaakt tussen willekeurige meetfouten, systematische meetfouten en differentiële meetfouten. Een willekeurige meetfout wordt

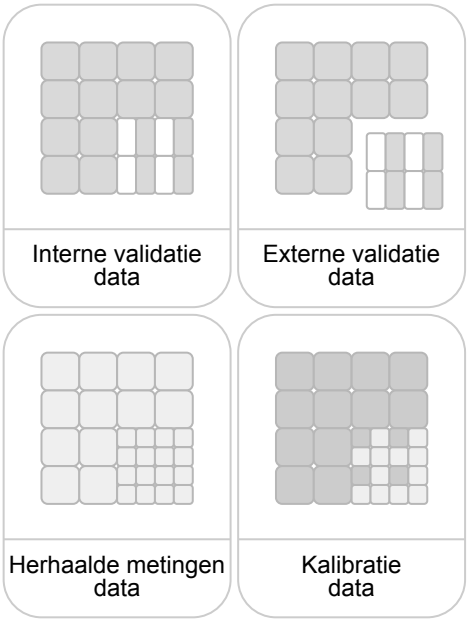
gemaakt als de meting soms een beetje hoger en soms een beetje lager is dan de 'ware' waarde, waarbij de afwijking tussen de meting die gedaan wordt en de 'ware' waarde willekeurig is. Een systematische meetfout wordt gemaakt als de afwijking tussen de meting die wordt gedaan en de 'ware' waarde systematisch is: de meting is bijvoorbeeld altijd lager. Een differentiële meetfout wordt gemaakt als de meetfout afhangt van een andere factor, bijvoorbeeld als een uitkomst wordt gemeten met een meetfout en die meetfout afhangt van de waarde van de blootstelling.



Figuur A.3: Illustratie van verschillende soorten meetfouten. *Willekeurige meetfout*: de meting is soms een beetje hoger en soms een beetje lager dan de 'ware' waarde; *systematische meetfout*: de meting is altijd lager dan de 'ware' waarde.

In figuur A.4 wordt een overzicht gegeven van de verschillende soorten data die gebruikt kunnen worden voor het afschatten van het 'meetfoutmodel'. Interne validatie data zijn data waarbij in een deel van de studiepopulatie zowel de meting met meetfout als de meting zonder meetfout wordt gedaan. In externe validatie data worden deze twee metingen gedaan in een groep mensen die geen onderdeel uitmaken van de studiepopulatie. Naast interne en externe validatie data, zijn er ook nog kalibratie data en data van herhaalde metingen. In deze laatste twee soorten data is de 'ware' waarde voor niemand uit de studiepopulatie bekend. Herhaalde metingen data kunnen gebruikt worden om het 'meetfoutmodel' af te schatten van een meting met een willekeurige meetfout. In iedereen of in een deel van de studiepopulatie wordt dan de meting met een willekeurige meetfout minstens twee keer gedaan. Kalibratie data kunnen gebruikt worden om het 'meetfoutmodel' te schatten van een meting met een systematische meetfout. In een deel van de studiepopulatie wordt dan naast de meting met systematische meetfout, een andere meting gedaan (met een ander meetinstrument) met een willekeurige meetfout.

De meetfout correctiemethoden die het meest aan bod komen in dit proefschrift zijn regressie kalibratie en simulatie-extrapolatie. Figuur A.5 en A.6 geven een schematische weergave van hoe regressie kalibratie en hoe simulatie-extrapolatie werken, respectievelijk, als er zich meetfouten in de blootstelling voordoen. Bij regressie kalibratie wordt de regressie voor de te schatten associatie tussen een blootstelling en uitkomst gekalibreerd naar de regressie wanneer er geen meetfout is. Deze kalibratie wordt gedaan door niet de regressie van de blootstelling op de uitkomst te schatten, maar in plaats daarvan de regressie van een afgeleide van de blootstelling op de uitkomst te schatten. Simulatie-extrapolatie bestaat uit twee stappen. In de eerste stap, de 'simulatie' stap, wordt er een twee keer zo grote meetfout als de initiële meetfout aan de blootstelling toegevoegd.

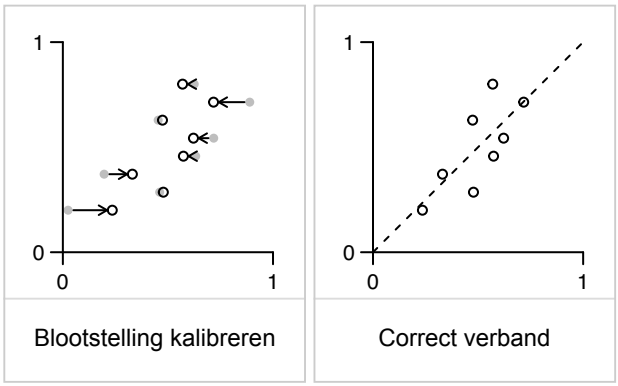


Figuur A.4: Schematische weergave van verschillende soorten validatie data. *Interne validatie data*: in een deel van de studiepopulatie wordt zowel een meting met meetfout als een meting zonder meetfout gedaan; *externe validatie data*: in een groep mensen die geen onderdeel zijn van de studiepopulatie wordt zowel een meting met meetfout als een meting zonder meetfout gedaan; *herhaalde metingen data*: in een deel van de studiepopulatie wordt de meting met een willekeurige meetfout vier keer gedaan; *kalibratie data*: in een deel van de studiepopulatie wordt de meting met systematische meetfout gedaan en drie keer een meting met willekeurige meetfout.

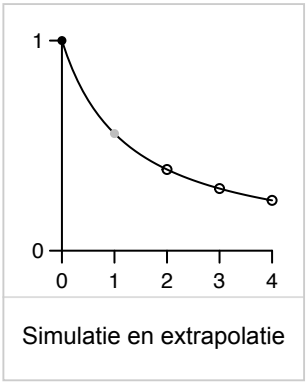
Dit wordt dan herhaald voor een drie keer zo grote meetfout, enzovoort. In de tweede stap, de 'extrapolatie' stap, wordt er een lijn geschat door de nieuwe schattingen om vervolgens met behulp van deze lijn terug te extrapoleren naar de situatie zonder meetfout.

Een simulatie studie is een experiment met kunstmatige, door een computer gecreëerde, data. Dit heeft als voordeel dat bijvoorbeeld in simulatiestudies naar methoden voor correctie van meetfouten, de 'ware' regressiecoëfficiënt bekend is bij de onderzoekers en dus dat op die manier kan worden onderzocht op welke manier meetfouten de regressiecoëfficiënt 'vertekenen'. Ook kan bijvoorbeeld interne of externe validatie data eenvoudig worden gecreëerd. In het onderzoek beschreven in dit proefschrift is veelvuldig gebruik gemaakt van simulatiestudies.

In hoofdstuk 2 worden de consequenties van meetfouten in eindpunten van gerandomiseerde studies en een daarbij horende oplossingen beschreven. Dit wordt geïllustreerd aan de hand van een gerandomiseerde studie naar het effect van de inname van ijzertabletten op het hemoglobinegehalte in bloed in zwangere vrouwen. Het hemoglobinegehalte kan in dit voorbeeld op verschillende manieren gemeten worden: aan de hand van een bepaling in veneus bloed of op basis van een vingerprik. Hierbij wordt aangenomen dat de meting in veneus bloed de 'ware' hemoglobinewaarde is en de meting in bloed na een vingerprik de waarde met meetfout. Willekeurige meetfouten in het eindpunt hebben invloed op de statistische 'power' en 'precisie' van een studie,



Figuur A.5: Schematische weergave van regressie kalibratie. De lineaire regressie-analyse voor het te schatten verband tussen de blootstelling met meetfout (horizontale as) en uitkomst (verticale as) wordt gekalibreerd naar de regressie wanneer er geen meetfout is in Figuur A.2.



Figuur A.6: Schematische weergave van simulatie-extrapolatie. In de simulatie stap wordt de regressie coëfficiënten geschat als er twee, drie of vier keer zo grote meetfout op de blootstelling zit (horizontale as), vervolgens wordt er een lijn geschat door de nieuwe schattingen en wordt er in de extrapolatie stap met behulp van deze lijn terug geëxtrapoleerd naar de regressiecoëfficiënt (verticale as) wanneer er geen meetfout is in Figuur A.2.

terwijl systematische en differentiële fouten ook de grootte van de associatie tussen de blootstelling en het eindpunt kunnen vertekenen. Een manier om te corrigeren voor deze vertekening is regressie kalibratie in combinatie met externe validatie data. We laten in een simulatiestudie zien dat indien het 'meetfoutmodel' nauwkeurig kan worden geschat in deze externe data, kan op basis van externe data van minimaal 15 metingen al de door meetfout verstoorde associatie volledig worden gecorrigeerd.

In hoofdstuk 3 wordt het softwarepakket ('package') `mecor` geïntroduceerd, geschreven in de programmeertaal `R`. Met behulp van `mecor` kan worden gecorrigeerd voor meetfouten in continue covariaten en continue uitkomsten. Mecor faciliteert correctie voor willekeurige, systematische en differentiële meetfouten en kan gebruik maken van de vier soorten data (interne, externe, kalibratie en herhaalde metingen data). Ook kan

mecor meetfouten corrigeren in een zogeheten sensitiviteitsanalyse als er geen extra data beschikbaar is om het 'meetfoutmodel' of te schatten.

In hoofdstuk 4 worden situaties beschreven waarin het gebruik van regressie kalibratie niet goed werkt. We illustreren dit aan de hand van een studie naar de associatie tussen actieve energie en de vetvrije massa van een persoon. In een simulatiestudie wordt gedemonstreerd dat in het bijzonder in kleine studies het statistisch efficiënter kan zijn om niet voor de meetfout te corrigeren dan wel. Het corrigeren van meetfouten introduceert namelijk meer onnauwkeurigheid in de associatie die wordt geschat in de data. Soms kan het dan statistisch voordeliger zijn om niet te corrigeren voor de meetfout. Verder blijkt dat in kleine studies waar de meetfout relatief groot is, regressie kalibratie niet goed werkt.

In hoofdstuk 5 worden verschillende manieren voor het includeren van een deel van een studiepopulatie in interne validatie data beschreven. De manieren die onderzocht worden zijn het includeren op een volledige willekeurige manier, gestratificeerd willekeurig of het includeren van de extremen. Dit wordt beschreven aan de hand van een voorbeeld uit de Nederlandse Epidemiologie van Obesitas studie. In deze studie werd gekeken naar het verband tussen visceraal vet en het insulinegehalte in bloed. Visceraal vet kan worden gemeten door middel van een MRI scan, maar een grove meting kan ook worden gedaan met behulp van een meting van de omtrek van de middel van een persoon. Omdat een meting met een MRI scan heel kostbaar is, is de precieze meting van visceraal vet alleen in een deel van de studiepopulatie gedaan. In het hoofdstuk wordt onderzocht of de grove meting van de middelomtrek kan worden gebruikt voor het afschatten van de associatie tussen visceraal vet en insulinegehalte in bloed. Hierbij wordt gebruik gemaakt van de associatie tussen de middelomtrek en visceraal vet in een subgroep waarin beide metingen beschikbaar zijn en onderzocht in welke mensen het beste beide metingen kunnen worden gedaan. Een simulatiestudie laat zien dat het willekeurig kiezen van de mensen waarin alle twee de metingen worden gedaan het beste blijkt te werken.

In hoofdstuk 6 wordt de heterogeniteit in studies naar de incidentie van veneuze tromboses in COVID-19 patiënten beschreven. Verschillende oorzaken van deze heterogeniteit worden onderzocht, waarbij onderscheid wordt gemaakt tussen klinische bronnen en methodologische bronnen van heterogeniteit. Deze oorzaken worden geïllustreerd met voorbeelden uit gepubliceerde incidentie studies. De klinische bronnen die beschreven worden zijn verschillen in patiëntkarakteristieken en verschillen in de redenen om een veneuze trombose uit te (kunnen) sluiten of niet. De methodologische bronnen die beschreven worden zijn verschillen in de definitie voor de veneuze trombose (wat wordt geteld als veneuze trombose en wat niet), de kwaliteit van de data en de statistische analyse die gebruikt worden. We raden aan om in toekomstige studies naar de incidentie van veneuze trombose in COVID-19 patiënten, deze elementen nauwkeurig te beschrijven zodat dergelijke studies in de toekomst beter onderling kunnen worden vergeleken en samengevat.

In hoofdstuk 7 wordt het gebruik van regressie kalibratie en simulatie-extrapolatie vergeleken voor het doen van sensitiviteitsanalyses. Een simulatiestudie laat zien dat zonder extra data om het 'meetfoutmodel' af te schatten, maar met de correcte assumpties over het 'meetfoutmodel', regressie kalibratie de vertekening in de associatie verhelpt terwijl simulatie-extrapolatie dat niet doet. Slechts in enkele gevallen was het gebruik van simulatie-extrapolatie statistisch efficiënter.

Om causale verbanden te schatten in de epidemiologie, worden steeds vaker statistische

modellen gebruikt die gewogen worden aan de hand van gewichten die de kans op een blootstelling voorspellen. Omdat het nog relatief onbekend is wat het effect is van misclassificatie in een 'confounder' in deze eerstgenoemde modellen, wordt in hoofdstuk 8 de invloed van meetfouten in deze modellen vergeleken met de invloed van meetfouten in een meer traditioneel model. Ten opzichte van de invloed van een meetfout in het traditionele model, is de vertekening in de associatie in dezelfde richting (door de meetfout lijkt de associatie sterker of zwakker dan dat hij werkelijk is, ongeacht het statistische model dat gebruikt wordt) maar niet altijd even groot.

Concluderend, meetfouten kunnen de resultaten van epidemiologisch onderzoek vertekenen indien deze meetfouten worden genegeerd. Het gezegde "voorkomen is beter dan genezen" gaat in deze situatie ook op. Het verbeteren van de kwaliteit van de metingen die worden gebruikt voor epidemiologisch onderzoek is mogelijk beter dan het inzetten van meetfoutcorrectiemethoden. Dat gezegd hebbende, in situaties waarin meetfouten niet kunnen worden voorkomen kunnen de correctiemethoden uitkomst bieden en zou men meetfouten niet moeten negeren.

B

List of publications

10. **L. Nab**, R.H.H. Groenwold, F.A. Klok, B.S. Bhoelan, M.J.H.A. Kruip and S.C. Cannegieter, Estimating incidence of venous thromboembolism in COVID-19: Methodological considerations, Research and Practice in Thrombosis and Haemostasis 6 (6) (2022) e12776. doi:10.1002/rth2.1277

9. **L. Nab** and R.H.H. Groenwold, Sensitivity analysis for random measurement error using regression calibration and simulation-extrapolation, Global Epidemiology 3 (2021) 100067. doi:10.1016/j.gloepi.2021.100067

8. **L. Nab**, M. van Smeden, R.H. Keogh and R.H.H. Groenwold, Mecor: An R package for measurement error correction in linear regression models with a continuous outcome, Computer Methods and Programs in Biomedicine 208 (2021) 106238. doi:10.1016/j.cmpb.2021.106238

7. **L. Nab**, M. van Smeden, R. de Mutsert, F.R. Rosendaal and R.H.H. Groenwold, Sampling strategies for internal validation samples for exposure measurement–error correction: A study of visceral adipose tissue measures replaced by waist circumference measures, American Journal of Epidemiology 190 (9) (2021) 1935–1947. doi:10.1093/aje/kwab114

6. M. van Smeden, B.B.L. Penning de Vries, **L. Nab**, and R.H.H. Groenwold, Approaches to addressing missing values, measurement error, and confounding in epidemiologic studies. Journal of Clinical Epidemiology 131 (2021) 89–100. doi:10.1016/j.jclinepi.2020.11.006

5. M. Linschoten, **L. Nab**, I.C.C. van der Horst, R. Tieleman and F.W. Asselbergs, Response to "Early hydroxychloroquine but not chloroquine use reduces ICU admission in COVID-19 patients", International Journal of Infectious Diseases 103 (2020) 560–561. doi:10.1016/j.ijid.2020.12.006

4. **L. Nab**, R.H.H. Groenwold, M. van Smeden and R.H. Keogh, Quantitative bias analysis for a misclassified confounder: A comparison between marginal structural models and conditional models for point treatments, Epidemiology, 31 (6) (2020) 796–805. doi:10.1097/EDE.0000000000001239

3. K. Hettne, R. Proppert, **L. Nab**, L.P. Rojas-Saunero and D. Gawehns, ReprohackNL 2019: How libraries can promote research reproducibility through community engagement. IASSIST Quarterly, 44 (1-2) (2020) 1–10. doi:10.29173/iq977

2. **L. Nab**, R.H.H. Groenwold, P.M.J. Welsing and M. van Smeden, Measurement error in continuous endpoints of randomised trials: Problems and solutions, Statistics in Medicine 38 (27) (2019) 5182–5196. doi:10.1002/sim.8359

1. R.H.H. Groenwold, **L. Nab** and M. van Smeden, Groot, groter grootst: Big data in medisch onderzoek. Nederlands tijdschrift voor Geneeskunde 162 (2018) D3108.

C

Curriculum Vitæ

Linda Nab was born on May 17th, 1991, in Zutphen, the Netherlands. In 2009, she graduated from her secondary education (gymnasium) at Stedelijk Daltoncollege in Zutphen. She obtained her BSc degree in Mathematics at Utrecht University in 2013. In the 2 years that followed, she enhanced her programming skills while working as an IT consultant. In 2017, she graduated from the 2-year MSc programme Nutrition and Health, with a specialisation in epidemiology at Wageningen University & Research (cum laude). In the final year of her MSc programme, she undertook an internship at the Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, under supervision of Prof R.H.H. Groenwold and Dr M. van Smeden. She continued her research under the same supervision as a PhD candidate at the department of Clinical Epidemiology at Leiden University Medical Center. In 2022, she started working as an epidemiologist at the Bennett Institute for Applied Data Science at the University of Oxford.

D

Acknowledgements

Een promotietraject doe je niet alleen. Daarom zou ik graag een aantal mensen willen bedanken.

Prof. Groenwold, beste Rolf, bedankt dat je zo'n fijne mentor was. Je kennis, relativeringsvermogen, humor en enthousiasme motiveerden mij keer op keer, ook als ik het even niet meer zag zitten.

Dr. van Smeden, beste Maarten, bedankt dat ik altijd even bij je binnen kon lopen voor advies. Onze gesprekken hielpen mij om te groeien in mijn zelfvertrouwen.

Prof Keogh, dear Ruth, thank you for giving me the opportunity to visit your group at the London School of Hygiene and Tropical Medicine in 2019. The knowledge I gained and skills I learned during my stay are invaluable.

Beste Kim en Bas, bedankt voor alle interessante discussies die we hebben gevoerd in C7-92 en tijdens de COVID-19 pandemie over Zoom. Ik heb onwijs veel van jullie geleerd en met jullie gelachen.

Beste Esmee, bedankt voor de tijd die we doorbrachten als kamergenootjes in het raamloze C7-104. Ik heb altijd erg genoten van onze gesprekken.

Dear Jungyeon, thank you for your kindness. I will always keep very fond memories of our time in Leuven at the ISCB40.

Dear 'Epi Promovendi', thank you for our many coffee and lunch breaks, outings and 'borrels'. It was very unfortunate that the COVID-19 pandemic made this all impossible for a long time, making the ones that did happen even more valuable.

Dear 'ReproHack' ladies, thank you for your companionship and teaching me so much about reproducible science.

Lieve ‘Vordense Kuikens’, bedankt voor de kopjes koffie en thee in Vorden. Lieve ‘Praathoek’, bedankt voor de momenten dat we weer even lachten om de dingen die we deden toen we zestien waren.

Lieve ‘Ouwe Bokken’, bedankt voor alle goede gesprekken tijdens onze etentjes. Onwijs fijn dat jullie altijd zonder oordeel luisterden naar de uitdagingen die ik tegenkwam als beginnend onderzoeker. Lieve Anouk, bedankt voor al je steun en goede adviezen. Ik leer super veel van jouw kijk op het leven.

Lieve familie Heemskerk, bedankt voor alle te gekke vakanties en prachtige reizen die mij mijn proefschrift even deden vergeten. Lieve Dick en Wilma, bedankt voor jullie lieve zorgen en luisterend oor.

D

Lief gezin Nab, bedankt voor jullie Achterhoekse nuchterheid en werkmentaliteit (“niet lullen maar poetsen”). Lieve pap en mam, bedankt voor de zelfstandigheid en zelfbeschikking die jullie mij hebben meegegeven in jullie opvoeding. Zonder dat zou ik niet staan waar ik nu stond.

Lieve Jorg, bedankt voor je onschatbare steun tijdens de afgelopen vijf jaar. Je liet mij schaterlachen op de momenten dat ik het even niet meer zag zitten en nam mij mee op eindeloos veel avonturen. Je leerde me skiën en keihard fietsen. Jouw liefde geeft mij de kracht om mijn dromen waar te maken.