Correction methods for measurement error in epidemiologic research

Nab, L.

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.

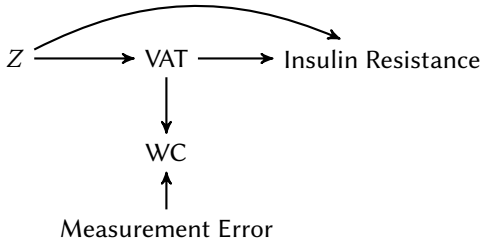

Figure 5.1: Assumptions of our motivating example. Error-prone waist circumference (WC) measures used as a substitute measurement to estimate the association between (VAT) and insulin resistance, confounded by Z (e.g., age, sex, total body fat).

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

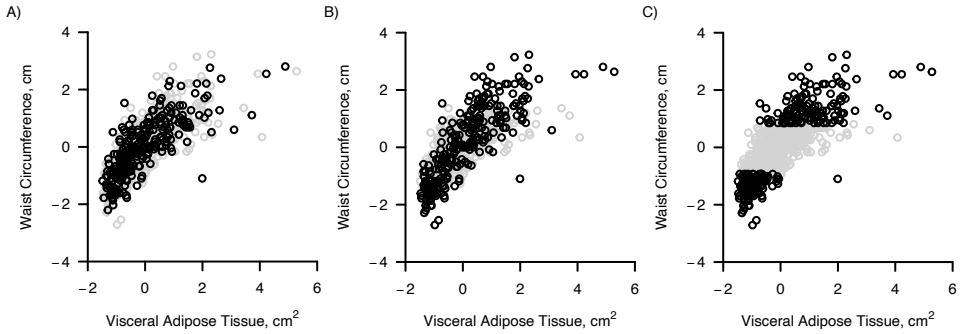

Figure 5.2: Visualisation of different internal validation sample sampling strategies in the Leiderdorp cohort of the Netherlands Epidemiology of Obesity. A) Visceral adipose tissue (VAT) measures are obtained at random (independent of waist circumference (WC)); B) VAT measures are obtained stratified 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.

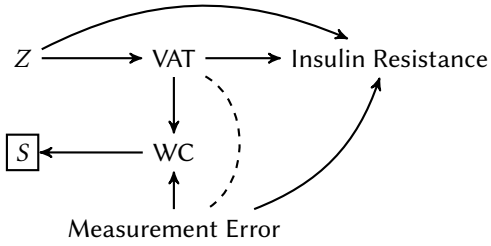

Figure 5.3: Collider stratification bias due to differential measurement error. Introduction of collider stratification bias when the data are observed (S) depending on the error-prone waist circumference (WC) measures with differential measurement error in a study estimating the association between (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.

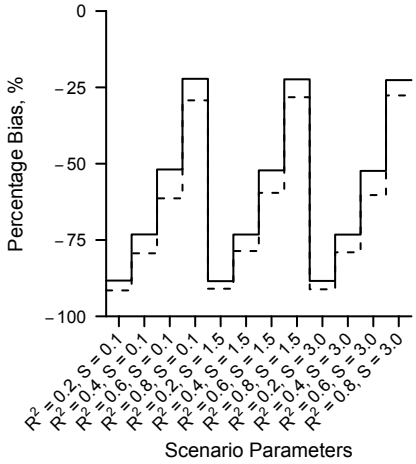

Figure 5.4: Nested loop plot of the percentage bias in the analysis ignoring measurement error. Solid line: Linear measurement error model; and dashed line: Non-linear measurement error model. Order from outer to inner loops: Skewness of the residual errors of the gold standard measure (S, 3 levels,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

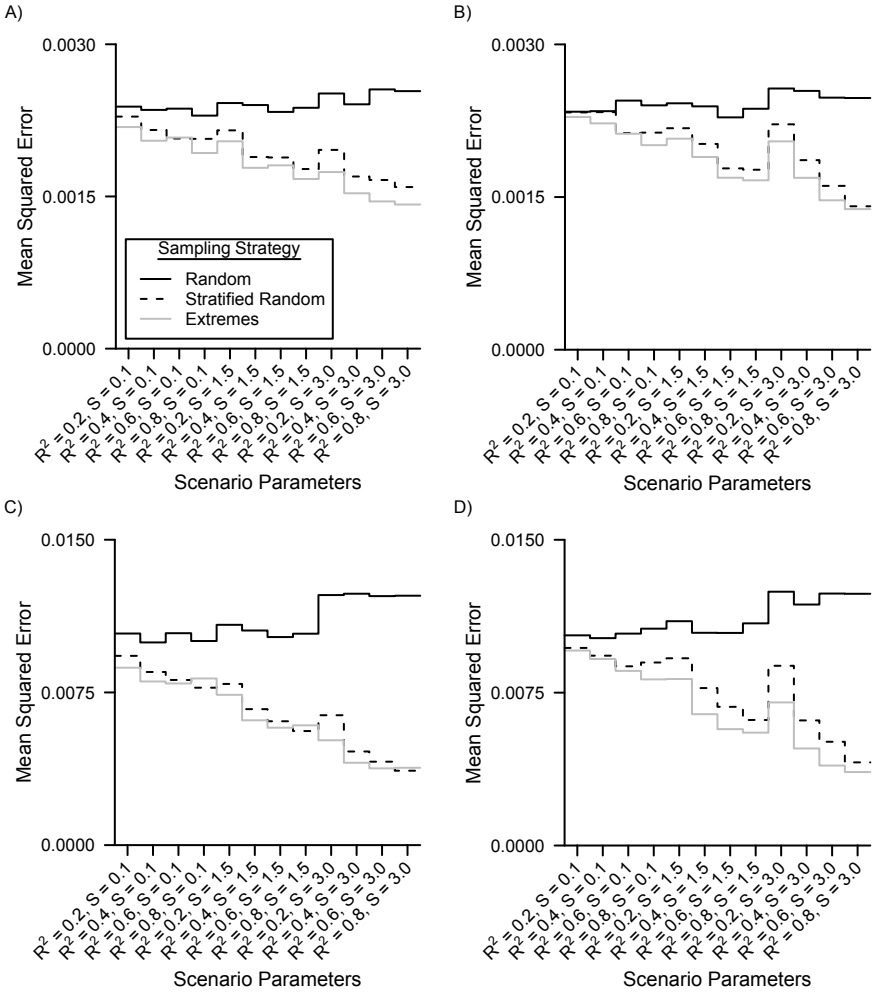

Figure 5.5: Nested loop plot of the mean squared errors in the analysis restricted to the internal validation sample for the three different sampling strategies. A) Linear measurement error model and an internal validation sample of 40% of the main study; B) Non-linear measurement error model and an 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).

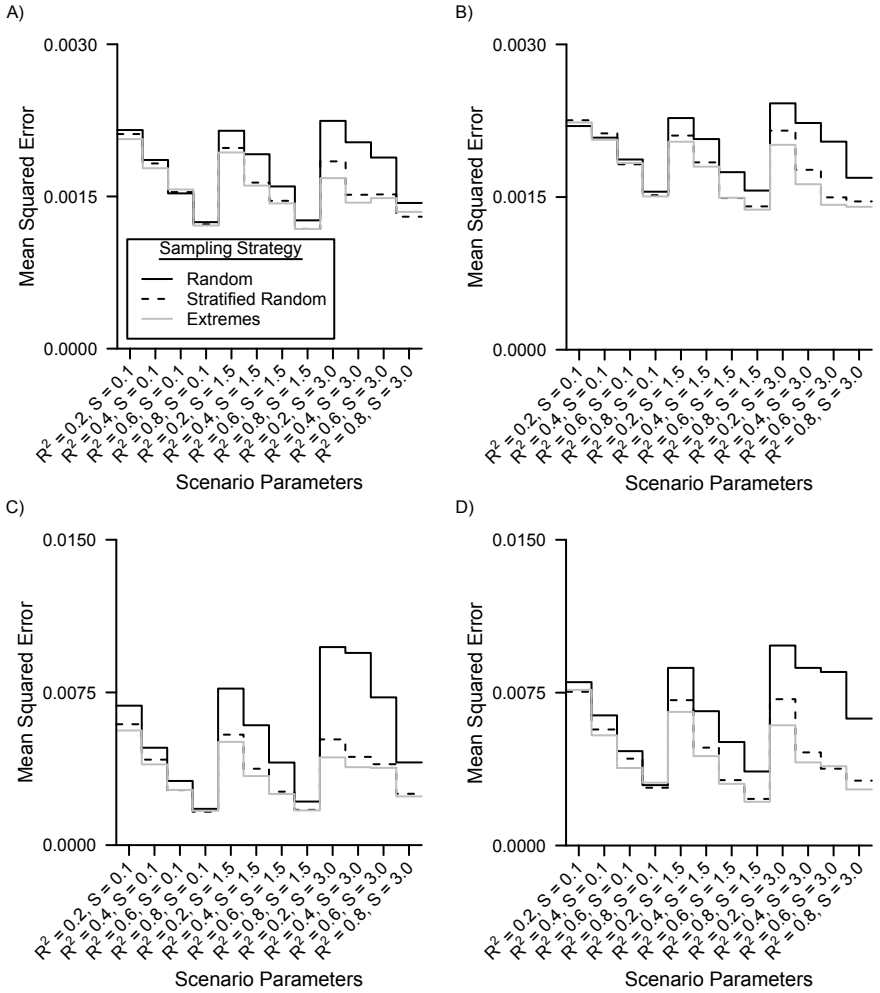

Figure 5.6: Nested loop plot of the mean squared errors in the analysis using validation regression calibration to correct for the measurement error for the three different sampling strategies. A) Linear measurement error model and an internal validation sample of 40% of the main study; B) Non-linear 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.