Received: 28 September 2018 | Revised: 12 June 2019 | Accepted: 7 August 2019

DOI: 10.1002/sim.8359

RESEARCH ARTICLE

WILEY Statistics in Medicine

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

L. Nab[1] | **R.H.H. Groenwold**[1] | **P.M.J. Welsing**[2] | **M. van Smeden**[1]

[1]Department of Clinical Epidemiology, Leiden University Medical Center, Leiden, The Netherlands

[2]Department of Rheumatology and Clinical Immunology, University Medical Center Utrecht, Utrecht, The Netherlands

Correspondence
L. Nab, Department of Clinical Epidemiology, Leiden University Medical Center, Leiden, 2333 ZA, The Netherlands.
Email: L.Nab@lumc.nl

Funding information
Nederlandse Organisatie voor Wetenschappelijk Onderzoek, Grant/Award Number: 917.16.430

In randomised trials, continuous endpoints are often measured with some degree of error. This study 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 calibration sample). We show that, if measurement error in continuous endpoints is ignored, the treatment effect estimator is unbiased when measurement error is classical, 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 calibration 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) calibration sample is shown to improve inferences and should be considered in trials with error-prone endpoints. Implementation of the proposed correction methods is accommodated by a new software package for R.

KEYWORDS

bias, clinical trials, continuous endpoints, correction methods, measurement error

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 (eg, self-reported physical activity levels[1]), endpoints that are collected as part of routine care (eg, in pragmatic trials[2]), endpoints that are assessed without blinding the patient or assessor to treatment allocation (eg, 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 (eg, 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,[6] developments and applications of correction methods for error in endpoints are still rare.[7] Specifically, methodology that allows 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 errors in the applied medical literature is unusual.[14,15]

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.[16] To test the effect of the iron supplementation on maternal haemoglobin levels, haemoglobin concentrations were measured at delivery in venous blood.

This paper describes a taxonomy of measurement errors in trial endpoints, evaluates the effect of measurement errors on the analysis of trials and tests existing, and proposes new methods evaluating trials containing measurement errors. Implementation of the proposed measurement error correction methods (ie, the existing and novel methods) is supported by introducing a new R package `mecor`, available at www.github.com/LindaNab/mecor. This paper is structured as follows. In Section 2, we revisit the example trial introduced in the previous paragraph. Section 3 presents an exploration of the influence of measurement error structures and their impact on inferences of trials. In Section 4, measurement error correction methods are proposed. A simulation study investigating the efficacy of the correction methods is presented in Section 5. Conclusions and recommendations resulting from this study are provided in Section 6.

2 | ILLUSTRATIVE EXAMPLE: MEASUREMENT OF HAEMOGLOBIN LEVELS

Makrides et al[16] 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 CI (4.4; 9.3)). Haemoglobin concentrations were measured spectrophotometrically. Mean haemoglobin values were 137 (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 errors 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.[17] While easier to measure, capillary haemoglobin levels are less accurate than venous haemoglobin levels.[18] We now discuss how measurement errors in haemoglobin levels might affect trial inference, by assuming hypothetical differences between capillary and venous haemoglobin levels. Two more illustrative examples are discussed in Section 1 of the Supplementary Materials.

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 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 ordinary least squares (OLS) regression.

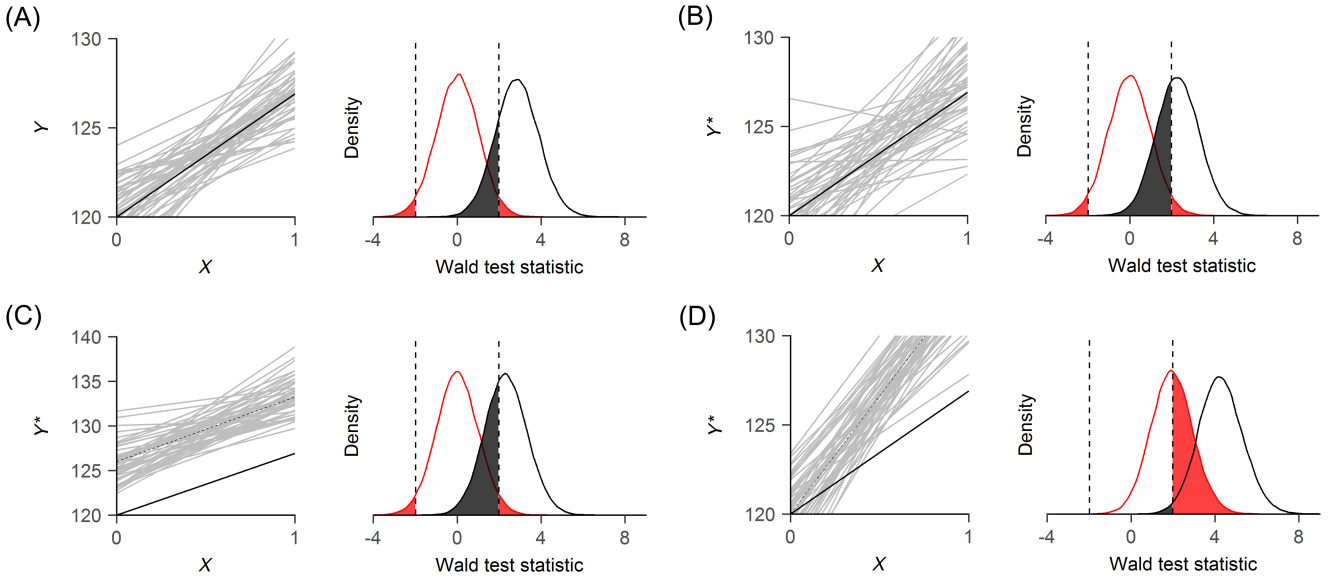

FIGURE 1 Illustration of impact of hypothetical measurement error in the example trial[16]: (A) No measurement error; (B) Classical measurement error; (C) Systematic measurement error;
(D) Differential measurement error. The left plots depict every thousandth estimated ordinary least squares regression line (grey lines), the average estimated treatment effect (dashed line), and the true effect (black line). The right plots depict the density distribution of the Wald test-statistic of the slope of the regression line [Colour figure can be viewed at wileyonlinelibrary.com]

2.1.1 | Classical measurement error in example trial

In the context of measurement of haemoglobin levels, random variability in the haemoglobin levels of capillary blood samples may be expected to vary more than haemoglobin levels in venous blood,[18] independently 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 of Figure 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 3.2). Type-II error increased (to 38%) (grey area in Figure 1, panel B) while Type-I error remained at the nominal level (at 5%, illustrated by the red area in Figure 1, panel B).

2.1.2 | 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.[19] 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 nonzero 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 panel C of Figure 1). Type-II error increased (to 37%) (grey area in Figure 1, panel C) while Type-I error remained at rate close to nominal level (at 5%) (red area in Figure 1, panel C).

2.1.3 | 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 panel D of Figure 1). Type-II error decreased (to 0.1%) (grey area in Figure 1, panel D) and Type-I error rates increased (to 48%) (red area in Figure 1, panel D).

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

$$Y = \alpha_Y + \beta_Y X + \varepsilon, \tag{1}$$

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 Section 2 of the Supplementary Materials.

3.1 | Classical measurement error

There is classical measurement error in Y^* if Y^* is an unbiased proxy for Y: $Y^* = Y+e$, where e has mean 0 and $\text{Var}(e) = \tau^2$ and e independent of Y, X, ε in (1). Using Y^* instead of Y in the linear model yields[10]:

$$Y^* = \alpha_Y^* + \beta_Y^* X + \delta, \tag{2}$$

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.

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 Section 2 of the Supplementary Materials).

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 (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 (1). Using Y^* with systematic measurement error in the linear model yields in the model defined by (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 $\beta_Y = 0$, $\beta_Y^* = 0$ (ie, tests for null effects are still valid under systematic measurement error) (proof is given in Section 2 of the Supplementary Materials).

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 (1) for $X = 0, 1$. Using Y^* with differential measurement error in the linear model yields in the model defined in (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 $[\theta_{10}^2 + (\theta_{11}^2 - \theta_{10}^2)X]\sigma^2 + \tau_X^2$ 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.[20] 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 Section 2 of the Supplementary Materials).

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 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 (ie, Y^* and Y are both measured after assignment of X). Instead of external calibration data, we could use internal calibration 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 paper 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.[21] 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 $\hat{\beta}_{Y^*}$ (eg, see the work of Long and Ervin[20] 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 3). Consistent estimators for the intervention effects are introduced, and various methods for constructing confidence intervals for these estimators are discussed. Section 3 of the Supplementary Materials 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 calibration sample. For an earlier exploration of the use of an internal calibration set when there is systematic or differential measurement error in endpoints, see the work of Keogh et al.[7]

4.1 | Systematic measurement error

From Section 3.3, it follows that natural estimators for α_Y and β_Y are

$$\hat{\alpha}_Y = (\hat{\alpha}_{Y^*} - \hat{\theta}_0)/\hat{\theta}_1 \quad \text{and} \quad \hat{\beta}_Y = \hat{\beta}_{Y^*}/\hat{\theta}_1, \tag{3}$$

where $\hat{\theta}_0$ and $\hat{\theta}_1$ are the estimated error parameters from the calibration data set using standard OLS regression. From Equation (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 (3) are consistent, see for a proof Section 3.1 of the Supplementary Materials.

The variance of the estimators defined in (3) can be approximated using the Delta method,[22] the Fieller method,[22] the Zero-variance method, and by bootstrap.[23] Further details are provided in Section 3.1 of the Supplementary Materials.

4.2 | Differential measurement error

From Section 3.4, it follows that natural estimators for α_Y and β_Y are

$$\hat{\alpha}_Y = (\hat{\alpha}_{Y^*} - \hat{\theta}_{00})/\hat{\theta}_{10} \quad \text{and} \quad \hat{\beta}_Y = (\hat{\beta}_{Y^*} + \hat{\alpha}_{Y^*} - \hat{\theta}_{01})/\hat{\theta}_{11} - \hat{\alpha}_Y, \tag{4}$$

where $\hat{\theta}_{00}$, $\hat{\theta}_{10}$, $\hat{\theta}_{01}$, and $\hat{\theta}_{11}$ are estimated from the external calibration 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 4.1). The estimators in (4) are consistent, see for a proof Section 3.1 of the Supplementary Materials. The variance of the estimators defined in (4) can be approximated using the Delta method,[22] the Zero-variance method, and the Bootstrap method.[23] Further details are provided in Section 3.2 of the Supplementary Materials.

5 | SIMULATION STUDY

The finite sample performance of the measurement error corrected estimators of the treatment effect was studied by simulation. We focused on the situation 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 calibration set was available, which was varied in size. The results from the example trial 1 are used to motivate our simulation study (see Section 2).

5.1 | Data generation

Data were generated for a sample of $N = 400$ individuals, approximately equal to the size of example trial 1.[16] The individuals were equally divided in the two treatment arms. The true endpoints were generated according to model (1), assuming iid normal errors and using the estimated characteristics found in the 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 Sections 3.3 and 3.4, 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.[7,24] 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 calibration set was generated of size K with the characteristics of the placebo arm for each simulated data set. For differential measurement error scenarios, a calibration 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 calibration 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.

5.2 | Computation

For each simulated data set, the corrected treatment effect estimators (3) for systematic error and (4) for differential error were applied. In systematic measurement error scenarios, confidence intervals for the corrected estimator for $\alpha = 0.05$

were constructed by using the Zero-variance method, the Delta method, the Fieller method, and the Bootstrap method based on 999 replicates (as defined in Section 4.1). In the case of differential measurement error, confidence intervals for the corrected estimator for $\alpha = 0.05$ were constructed by using the Zero-Variance method, the Delta method, and the Bootstrap method based on 999 replicates (as defined in Section 4.2). The HC3 heteroscedastic consistent variance estimator was used to accommodate for heteroscedastic error in the differential measurement error scenario.[20] Furthermore, for both the systematic and differential measurement error scenarios, the naive analysis was performed (resulting in a naive effect estimate and naive confidence interval), which is the "regular" analysis that would be performed if measurement errors were neglected.

We studied performance of the corrected treatment effect estimators in terms of percentage bias,[25] empirical standard error (EmpSE), and square root of the mean squared error (SqrtMSE).[26] The performance of the methods for constructing the confidence intervals was studied in terms of coverage and Type-II error.[26]

In our simulations, the Fieller method resulted in undefined confidence intervals 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 confidence intervals is reported. If the Fieller method resulted in undefined confidence intervals in more than 5% of cases in one simulation scenario, the coverage and average confidence interval width were not calculated as this would result in unfair comparisons between the different confidence interval constructing methods. The bootstrap confidence intervals were based on less than 999 estimates in case the sample drawn from the external calibration 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 library `mecor` (version 0.1.0). The results of the simulation are available at doi.org/10.6084/m9.figshare.7068695 and the code is available at doi.org/10.6084/m9.figshare.7068773, together with the seed used for the simulation study.

5.3 | Results of simulation study

5.3.1 | Systematic measurement error

Table 1 shows percentage bias, EmpSE, and SqrtMSE of the naive estimator and the corrected estimator for $\theta_1 = 1.25$ and $R^2_{Y^*,Y} = 0.8$ and $\theta_1 = 1.05$ and $R^2_{Y^*,Y} = 0.8$, $R^2_{Y^*,Y} = 0.5$ and $R^2_{Y^*,Y} = 0.2$ and $K \in \{5, 7, 10, 15, 20, 30, 40, 50\}$ 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 $\hat{\beta}_Y$ are reasonably small for $K \geq 10$. Yet, as the bias in the naive estimator is small when $\theta_1 = 1.05$, SqrtMSE of the corrected estimator is never lower than the SqrtMSE of the naive estimator. However, if bias in the naive estimator is greater ($\theta_1 = 1.25$), SqrtMSE of the corrected estimator is 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 is reasonably small for $K \geq 30$. Nevertheless, SqrtMSE of the corrected estimator is always greater than SqrtMSE of the naive estimator. For the corrected estimator and $\theta_1 = 1.05$ and $R^2_{Y^*,Y} = 0.2$, bias of $\hat{\beta}_Y$ fluctuates and EmpSE and SqrtMSE is 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, which provides a clear visualisation of the results formerly discussed. The bigger 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 3 explains why there is so much variation in the corrected effect estimator for small sample sizes of the external calibration 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 is 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 confidence intervals in 15%, 5%, 1%, 0.1% of simulated datasets for respectively $K = 5, 7, 10, 15$. Therefore, coverage and average confidence interval width of the Fieller method is not evaluated for $K \in \{5, 7\}$. For $R^2_{Y^*,Y} = 0.5$, the Fieller method failed to construct confidence intervals 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 confidence interval width is not evaluated for $K \in \{5, 7, 10, 15\}$. For $R^2_{Y^*,Y} = 0.2$, the Fieller method failed to construct confidence intervals in 74%, 71%, 64%, 53%, 43%, 26%, 15%, 8% of simulated data sets for $K \in \{5, 7, 10, 15, 20, 30, 40, 50\}$, respectively (ie, in every case more than 5%, thus the Fieller method is not evaluated for $R^2_{Y^*,Y} = 0.2$).

Table 1 shows coverage of the true intervention effect in the constructed confidence intervals and average confidence interval width using the Zero-variance, Delta, Fieller, and Bootstrap method. Using Wald confidence intervals for the

TABLE 1 Percentage bias, empirical standard error (EmpSE), square root of mean squared error (SqrtMSE), coverage, and average width of CIs of the naive estimator and the corrected estimator for systematic measurement error ($\theta_0 = 0$ and $\theta_1 = 1.05$ or $\theta_1 = 1.25$) for different values of R-squared and different sample sizes of the calibration 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[16]

Measure*	$R^2_{Y^*,Y}$	θ_1			**Sample size external calibration set**							
			Naive		**5**	**7**	**10**	**15**	**20**	**30**	**40**	**50**
Percentage bias (%)	0.8	1.25	24.9		88.9	29	3.7	2	1.6	0.9	0.7	0.4
	0.8	1.05	4.9		88.9	29	3.7	2	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	2
	0.2		4.9		168.2	−62.6	98.8	33.4	−142.2	−28.3	23.9	14.6
EmpSE	0.8	1.25	1.8		524.8	139.1	3	1.9	1.7	1.6	1.5	1.5
	0.8	1.05	1.5		524.8	139.1	3	1.9	1.7	1.6	1.5	1.5
	0.5		1.9		267	329.1	83.7	14.4	11	2.5	2.3	2.1
	0.2		3		1131.2	210.8	723.2	462.2	1044.4	225.5	70.5	24.8
SqrtMSE	0.8	1.25	2.5		524.8	139.1	3.1	1.9	1.7	1.6	1.5	1.5
	0.8	1.05	1.5		524.8	139.1	3.1	1.9	1.7	1.6	1.5	1.5
	0.5		1.9		267	329.1	83.7	14.4	11	2.5	2.3	2.1
	0.2		3		1131.2	210.8	723.1	462.2	1044.4	225.5	70.5	24.8
Coverage (%)	0.8	1.25	83.5‡	Zero-Variance	70.3	74	77.4	80.3	82.8	84.4	85.3	86.3
				Delta	93.8	95.3	95.7	95.9	96	96	95.9	95.7
				Fieller†	-	-	94.5	94.7	95	95.3	95.2	95
				Bootstrap	95.9	96.1	95.5	94.9	94.8	95	95.1	94.8
	0.8	1.05	94.6‡	Zero-Variance	77.8	81.3	84.4	87.1	89.2	90.9	92	92.2
				Delta	92.1	93.9	94.3	94.8	95.1	95.3	95.4	95.2
				Fieller†	-	-	94.5	94.7	95	95.3	95.2	95
				Bootstrap	95.9	96.1	95.5	94.9	94.8	95	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	92.9	93.9	94.3	95.2	95.4	95.3
				Fieller†	-	-	94.5	95.2	95.2	95	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	76.8	80.3	84.3	86	87.6
				Delta	86.8	89.7	90.9	92.2	93.5	94.4	94.6	94.9
				Fieller†	-	-	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
Av. CI width	0.8	1.25	6.9‡	Zero-Variance	30333	1141.5	5.5	4.7	4.7	4.6	4.5	4.5
				Delta	40.7	13.6	8.7	7.5	7	6.5	6.3	6.1
				Fieller†	-	-	11.8	8.3	7	6.4	6.1	6
				Bootstrap	86.9	29.3	14.1	8.3	7.1	6.4	6.1	6
	0.8	1.05	5.8‡	Zero-Variance	36110.7	1359	6.5	5.6	5.5	5.4	5.4	5.4
				Delta	35	12.2	8	7	6.7	6.3	6.1	6
				Fieller†	-	-	11.8	8.3	7	6.4	6.1	6
				Bootstrap	86.9	29.3	14.1	8.3	7.1	6.4	6.1	6
	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	9.3	8.7	8.4
				Fieller†	-	-	67.9	63.2	25	12.4	9.8	9
				Bootstrap	146.8	87.4	65.2	34.7	22.8	12.4	9.9	9
	0.2		11.6‡	Zero-Variance	126830.3	11677.5	87123.4	30709.4	324870.7	12430.8	774.6	126.8
				Delta	179.3	102.5	112.7	69.9	65.7	34.1	19.7	16.6
				Fieller†	-	-	92.6	95.1	72.1	82.2	60.6	59.2
				Bootstrap	176	121.9	126.2	118.7	107.7	77.6	54.8	39.7

*Monte Carlo standard errors of Bias, EmpSE, MSE, and Coverage are subsequently EmpSE$\sqrt{1/10,000}$, EmpSE/$(2\sqrt{9,999})$, $\sqrt{\frac{\sum_{i=1}^{10,000}[(\hat{\beta}_i-6.9)^2-\text{MSE}]^2}{9,999\times 10,000}}$, and $\sqrt{[\text{Cover.}\times(1-\text{Cover.})]/10,000}$.[26]

† Results of the Fieller method are shown if less than 5% of cases resulted in undefined confidence intervals (see last paragraph of Section 5.2).

‡Coverage of the true intervention effect and average confidence interval width using regular Wald confidence intervals of the naive effect estimator.

Type-II error using the naive effect estimator is 0.2%, 2.9%, and 31.6% for $R^2_{Y^*,Y} = 0.8$ (for both $\theta_1 = 1.05$ and $\theta_1 = 1.25$), $R^2_{Y^*,Y} = 0.5$ and $R^2_{Y^*,Y} = 0.2$, respectively. Type-II error using the corrected effect estimator using the Zero-Variance, Delta, and Bootstrap method was 0% in all scenarios. For the considered cases, Type-II error of the corrected effect estimator using the Fieller method was 0.2% and 2.9% for $R^2_{Y^*,Y} = 0.8$ (for both $\theta_1 = 1.05$ and $\theta_1 = 1.25$) and $R^2_{Y^*,Y} = 0.5$, respectively.

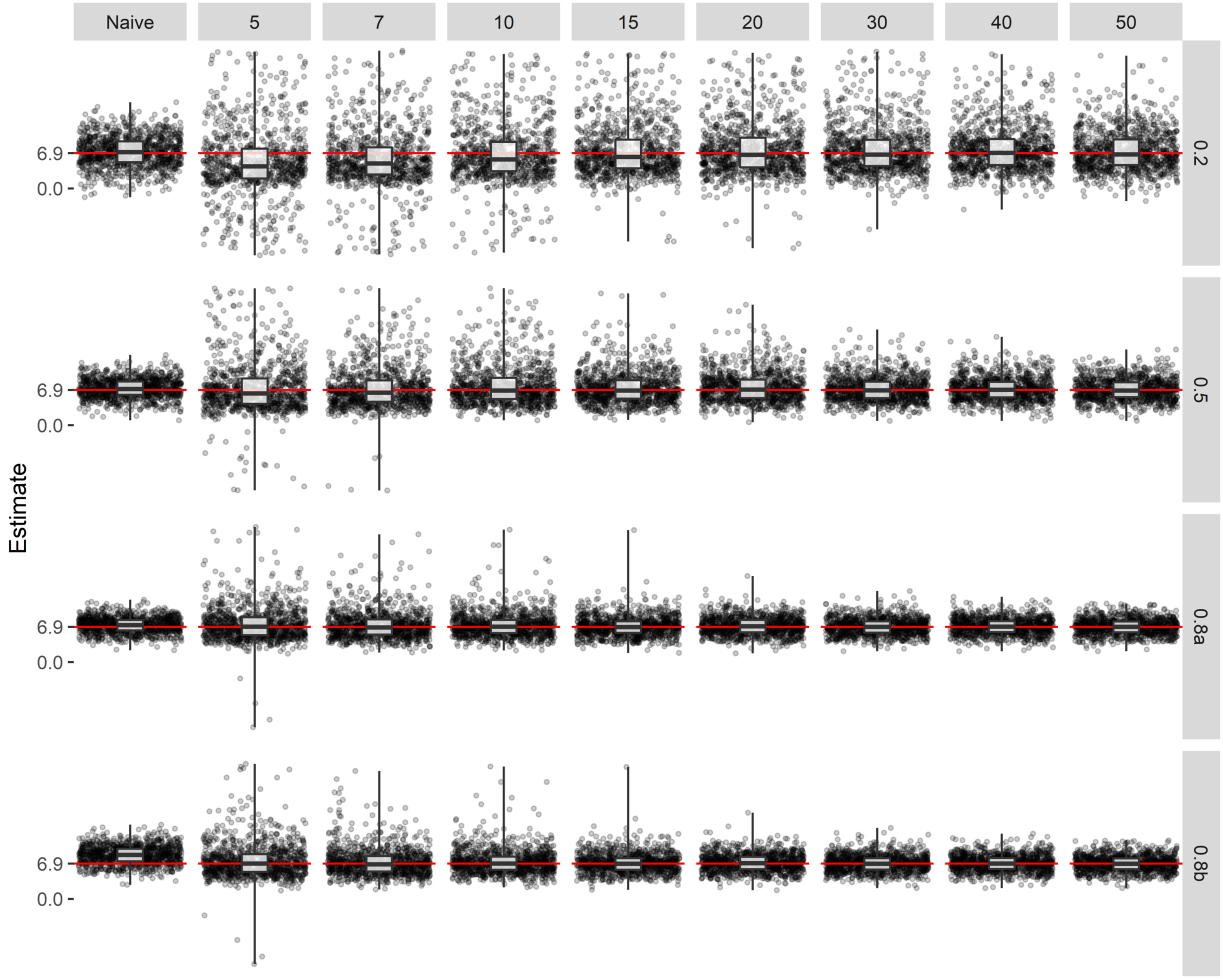

FIGURE 2 Estimates of the treatment effect using the naive estimator and corrected estimator for different values of R-squared (row grids) and different sample sizes of the external calibration set (column grids) under systematic measurement error ($\theta_1 = 1.05$ (0.2; 0.5; 0.8a) or $\theta_1 = 1.25$ (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 red line), based on the example trial 1 by Makrides et al[16] [Colour figure can be viewed at wileyonlinelibrary.com]

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 is small (ie, 5%). Yet, as bias percentage increased in the naive estimator for $\theta_1 = 1.25$ (ie, 25%), coverage dropped to 83.5%. The Zero-variance method yielded too narrow confidence intervals for all scenarios, 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 methods constructed correct confidence intervals for $K \geq 15$. For $K \leq 10$, the Delta method and the Fieller method constructed too narrow confidence intervals, and the Bootstrap method too broad confidence intervals. For $R^2_{Y^*,Y} = 0.5$, the Delta and Bootstrap methods constructed correct confidence intervals for $K \geq 30$. For $K \leq 20$, the Delta method constructed too narrow confidence intervals, and the Bootstrap method too broad confidence intervals. 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, Delta, and Bootstrap methods was 0%. For the considered scenarios using the Fieller method, Type-II error was 0.02% for $R^2_{Y^*,Y} = 0.8$ and 2.9% for $R^2_{Y^*,Y} = 0.5$.

5.3.2 | Differential measurement error

Table 2 shows percentage bias, EmpSE, and SqrtMSE of the naive estimator and the corrected estimator for $R^2_{Y^*,Y} = 0.8$, $R^2_{Y^*,Y} = 0.5$, and $R^2_{Y^*,Y} = 0.2$ and $K \in \{5, 7, 10, 15, 20, 30, 40, 50\}$ 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,

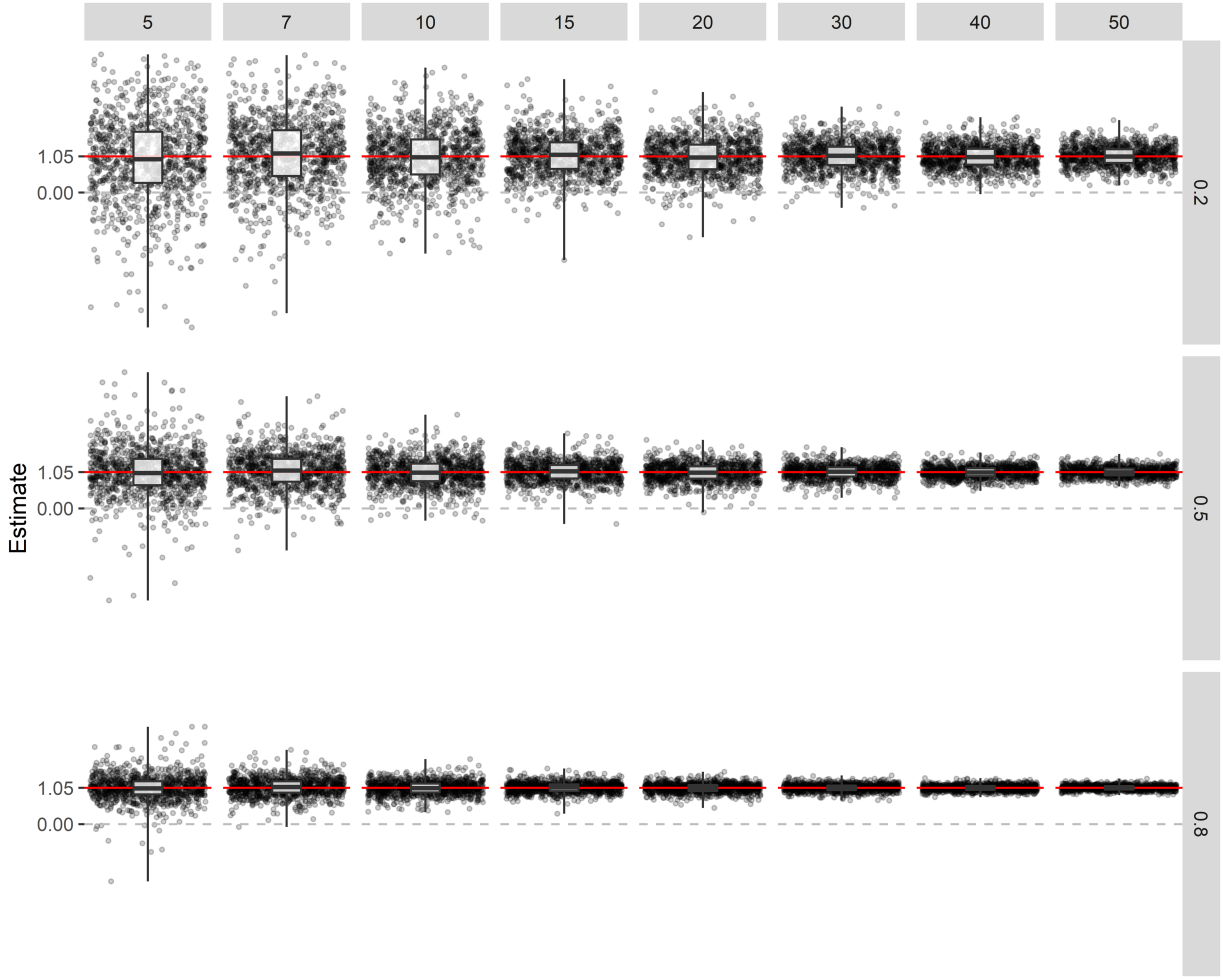

FIGURE 3 Estimates of θ_1 (ie, slope of the systematic measurement error model) for different values of R-squared (row grids) and different sample sizes of the external calibration set (column grids). Each grid is based on every 10th estimate of a simulation of 10 000 replicates, using an estimand of 1.05 (indicated by the red line) [Colour figure can be viewed at wileyonlinelibrary.com]

EmpSE, and SqrtMSE of $\hat{\beta}_Y$ are 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 are small for $K = 50$. For the naive estimator and $R^2_{Y^*,Y} = 0.2$, percentage bias, EmpSE, and MSE of the corrected estimator is 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 4, which provides a clear visualisation of the results formerly discussed. The bigger the sample size of the external calibration set and the higher R-squared, the better the performance of the corrected estimator.

Table 2 shows coverage of the true intervention effect in the constructed confidence intervals and average confidence interval width using the Zero-Variance, Delta, and Bootstrap methods. Coverage of the true treatment effect of 6.9 using Wald confidence intervals 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 confidence intervals; the Delta method yielded too broad confidence intervals, and the Bootstrap method yielded mostly too broad confidence intervals, 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, Delta, and Bootstrap methods was 0%.

5.4 | Measurement error dependent on a prognostic factor

In the above, we focused on measurement errors 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

TABLE 2 Percentage bias, empirical standard error (EmpSE), mean squared error (MSE), square root of mean squared error (SqrtMSE), coverage, and average width of CIs of the corrected estimator for differential measurement error ($\theta_{00} = 0, \theta_{10} = 1, \theta_{01} = 0, \theta_{11} = 1.05$) for different values of R-squared and different sample sizes of the calibration 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[16]

Measure*	$R^2_{Y^*,Y}$	Naive		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	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	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	8425.8	1569.7	443.7	92.1
SqrtMSE	0.8	6.5		52	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		2658	8425.4	1569.7	443.7	92.1
Coverage (%)	0.8	0.7‡	Zero-Variance	43.8	59.9	67.9	72.7	76.8
			Delta	97.1	96.6	96	95.7	95.9
			Bootstrap	97.9	95.7	94.7	94.5	95
	0.5	6.7‡	Zero-Variance	30.3	43.3	50.2	55.5	61
			Delta	97.6	97.6	97.3	96.9	97
			Bootstrap	98.4	98	96.6	95.8	95.5
	0.2	41.1‡	Zero-Variance	25.7	35	41.9	46.6	52.2
			Delta	98.4	99	98.9	98.9	98.9
			Bootstrap	99	99.6	99.2	99	98.7
Av. 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	17.9	30.3	10.6	7.5
			Delta	463975.1	49493.3	660587.5	13238	18.5
			Bootstrap	303.5	118.8	58.4	34.2	24
	0.2	11.4‡	Zero-Variance	64.6	150.5	53.1	43.1	26.8
			Delta	1219162.5	26998502.1	486295.4	85139.8	3407.5
			Bootstrap	562.9	353.8	283.3	221.4	170.2

*Monte Carlo standard errors of Bias, EmpSE, MSE, and Coverage are subsequently EmpSE$\sqrt{1/10,000}$, EmpSE/$(2\sqrt{9,999})$, $\sqrt{\frac{\sum_{i=1}^{10,000}[(\hat{\beta}_i-6.9)^2-\text{MSE}]^2}{9,999\times 10,000}}$, and $\sqrt{[\text{Cover.}\times(1-\text{Cover.})]/10,000}$.[26]

‡Coverage of the true intervention effect and average confidence interval width using regular Wald confidence intervals of the naive effect estimator.

Type-II error of 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 using the Zero-variance, Delta, and Bootstrap method was 0%.

prognostic factors. For example, measurement error in haemoglobin levels measured in capillary blood may differ for women and men.[18] Moreover, haemoglobin levels are, on average, higher in men than 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 $\text{Var}(\varepsilon) = 158.8$ (heamoglobin levels are, on average, higher in men). 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 4 of the Supplementary Materials, 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 favor of the conditional analysis. By assuming

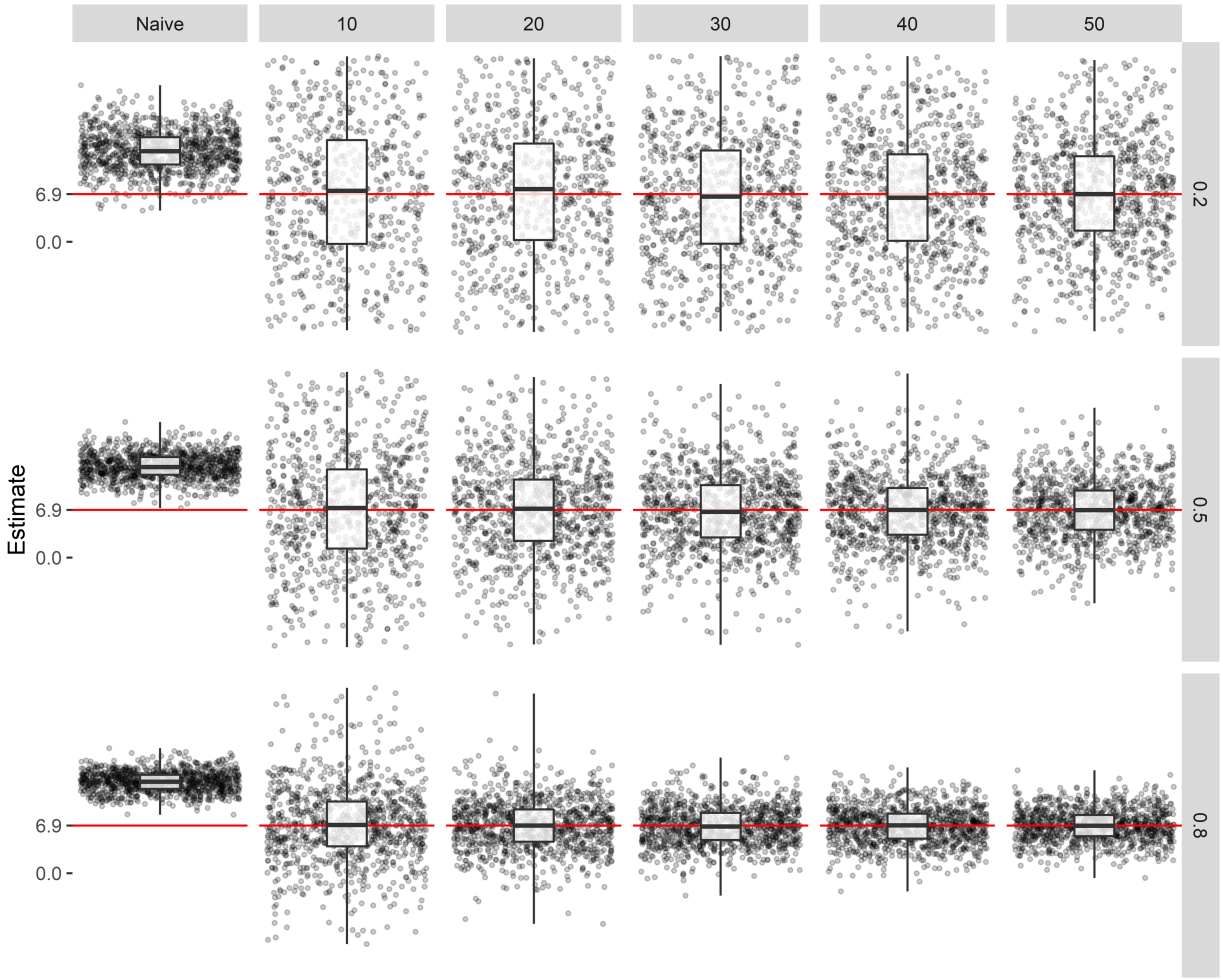

FIGURE 4 Estimates of the treatment effect using the naive estimator and corrected estimator for different values of R-squared (row grids) and different sample sizes of the external calibration set (column grids) under differential measurement error ($\theta_{00} = 0, \theta_{10} = 1, \theta_{01} = 0, \theta_{11} = \%1.05$). Each grid is based on every 10th estimate of a simulation of 10 000 replicates, using an estimand of 6.9 (indicated by the red line), based on example trial 1 by Makrides et al[16] [Colour figure can be viewed at wileyonlinelibrary.com]

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.

6 | DISCUSSION

This paper outlined the ramifications for randomised trial inferences when a continuous endpoint is measured with error. Our study showed that, when measurement error is ignored, not only can trial results be hampered by a loss in precision of the treatment effect estimate (ie, 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 article, we proposed a number of regression calibration-like correction methods to reduce the bias in the treatment effect estimator and obtain confidence intervals with nominal coverage. In our simulation studies, these methods were effective in improving trial inferences when an external calibration 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, the mechanism and magnitude of the measurement error should be considered. 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, propor-

tional 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 (eg, HC3; see the work of Long and Ervin[20]). Systematic error and differential error in the endpoint can be addressed via regression calibration.

We considered regression calibration-like correction methods that rely on an external calibration set that contains information about both error-prone and error-free measurements. We anticipate such an external calibration 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 calibration 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 calibration sets (ie, 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 calibration samples as small as 15 observations. With the relatively small calibration samples (up to 50 observations), our study showed that the Bootstrap method performed best in constructing confidence intervals in terms of coverage. The use of percentiles might explain that confidence intervals were slightly conservative (ie, too broad) for small calibration samples (10 observations) and might be improved by using bias-corrected and accelerated bootstrap intervals.[27] The proposed calibration correction methods rely on a linear regression framework and can thus easily be extended to incorporate covariables in the trial analysis.[28]

The use of measurement error corrections is still rare in applied biomedical studies despite an abundance of measurement error problems usually reported as an afterthought to a study.[14,15] 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 errors are more pronounced in the tails of the distribution or when measurement errors vary between treatment arms.

Instead of the use of external calibration 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 calibration, 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 paper. For instance, Cole et al suggested a multiple imputing approach based on an internal calibration set.[29] We also focused only on continuous outcomes in this paper. Problems and solutions for misclassified categorical outcomes can be found elsewhere.[30] 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 (see chapter 12 in the work of Carroll et al[10] and the references therein). Further, the method of nonparametric maximum likelihood has been successfully applied for explanatory variables measured with error[31,32] 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 nonignorable 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 confidence intervals for nonrandom measurement error. To facilitate the implementation of these measurement error correction estimators, we have developed the R package `mecor`, available at www.github.com/LindaNab/mecor.

ACKNOWLEDGEMENT

This work was supported by the Netherlands Organisation for Scientific Research (NWO, project 917.16.430).

DATA AVAILABILITY STATEMENT

The data and code used for the simulation study have been made publicly. The data is available at doi.org/10.6084/m9.figshare.7068695 and the code is available at doi.org/10.6084/m9.figshare.7068773.

ORCID

L. Nab https://orcid.org/0000-0002-1821-7246
R.H.H. Groenwold https://orcid.org/0000-0001-9238-6999
M. van Smeden https://orcid.org/0000-0002-5529-1541

REFERENCES

1. Cerin E, Cain KL, Oyeyemi AL, et al. Correlates of agreement between accelerometry and self-reported physical activity. *Med Sci Sports Exerc*. 2016;48(6):1075-1084.
2. Lauer MS, D'Agostino RB. The randomized registry trial–the next disruptive technology in clinical research? *N Engl J Med*. 2013;369(17):1579-1581.
3. Boutron I, Tubach F, Giraudeau B, Ravaud P. Blinding was judged more difficult to achieve and maintain in nonpharmacologic than pharmacologic trials. *J Clin Epidemiol*. 2004;57(6):543-550.
4. Staudacher HM, Irving PM, Lomer MCE, Whelan K. The challenges of control groups, placebos and blinding in clinical trials of dietary interventions. *Proc Nutr Soc*. 2017;76(3):203-112.
5. Mahabir S, Baer DJ, Giffen C, et al. Calorie intake misreporting by diet record and food frequency questionnaire compared to doubly labeled water among postmenopausal women. *Eur J Clin Nutr*. 2006;60(4):561-565.
6. Senn S, Julious S. Measurement in clinical trials: a neglected issue for statisticians? *Statist Med*. 2009;28(26):3189-3209.
7. Keogh RH, Carroll RJ, Tooze JA, Kirkpatrick SI, Freedman LS. Statistical issues related to dietary intake as the response variable in intervention trials. *Statist Med*. 2016;35(25):4493-4508.
8. Buonaccorsi JP. *Measurement Error: Models, Methods, and Applications*. Boca Raton, FL: Chapman & Hall/CRC; 2010.
9. Brakenhoff TB, van Smeden M, Visseren FLJ, Groenwold RHH. Random measurement error: why worry? an example of cardiovascular risk factors. *PLOS One*. 2018;13(2):1-8.
10. Carroll RJ, Ruppert D, Stefanski LA, Crainiceanu CM. *Measurement Error in Nonlinear Models: A Modern Perspective*. 2nd ed. Boca Raton, FL: Chapman & Hall/CRC; 2006.
11. Fuller WA. *Measurement Error Models*. New York, NY: John Wiley & Sons; 1987.
12. Gustafson P. *Measurement Error and Misclassification in Statistics and Epidemiology: Impacts and Bayesian Adjustments*. Boca Raton, FL: Chapman & Hall/CRC; 2004.
13. Hutcheon JA, Chiolero A, Hanley JA. Random measurement error and regression dilution bias. *BMJ*. 2010;340:c2289.
14. Brakenhoff TB, Mitroiu M, Keogh RH, Moons KGM, Groenwold RHH, van Smeden M. Measurement error is often neglected in medical literature: a systematic review. *J Clin Epidemiol*. 2018;98:89-97.
15. Shaw PA, Deffner V, Keogh RH, et al. Epidemiologic analyses with error-prone exposures: review of current practice and recommendations. *Ann Epidemiol*. 2018;28(11):821-828.
16. Makrides M, Crowther CA, Gibson RA, Gibson RS, Skeaff CM. Efficacy and tolerability of low-dose iron supplements during pregnancy: a randomized controlled trial. *Am J Clin Nutr*. 2003;78(1):145-153.
17. Zlotkin S, Arthur P, Antwi KY, Yeung G. Randomized, controlled trial of single versus 3-times-daily ferrous sulfate drops for treatment of anemia. *Pediatrics*. 2001;108(3):613-616.
18. Patel AJ, Wesley R, Leitman SF, Bryant BJ. Capillary versus venous haemoglobin determination in the assessment of healthy blood donors. *Vox Sanguinis*. 2013;104(4):317-323.
19. Keogh RH, White IR. A toolkit for measurement error correction, with a focus on nutritional epidemiology. *Statist Med*. 2014;33(12):2137-2155.
20. Long JS, Ervin LH. Using heteroscedasticity consistent standard errors in the linear regression model. *Am Stat*. 2000;54(3):217-224.
21. Fitzmaurice G. Measurement error and reliability. *Nutrition*. 2002;18(1):112-114.
22. Buonaccorsi JP. Measurement errors, linear calibration and inferences for means. *Comput Stat Data Anal*. 1991;11(3):239-257.
23. Efron B. Bootstrap methods: another look at the jackknife. *Ann Statist*. 1979;7(1):1-26.
24. Freedman LS, Midthune D, Arab L, et al. 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. *Am J Epidemiol*. 2018;187(10):2227-2232.
25. Burton A, Altman DG, Royston P, Holder RL. The design of simulation studies in medical statistics. *Statist Med*. 2006;25(24):4279-4292.
26. Morris TP, White IR, Crowther MJ. Using simulation studies to evaluate statistical methods. *Statist Med*. 2019;38(11):2074-2102.
27. Hall P. Theoretical comparison of bootstrap confidence intervals. *Ann Statist*. 1988;16(3):927-953.
28. Senn SJ. Covariate imbalance and random allocation in clinical trials. *Statist Med*. 1989;8(4):467-475.

29. Cole SR, Chu H, Greenland S. Multiple-imputation for measurement-error correction. *Int J Epidemiol*. 2006;35(4):1074-1081.
30. Brooks DR, Getz KD, Brennan AT, Pollack AZ, Fox MP. The impact of joint misclassification of exposures and outcomes on the results of epidemiologic research. *Curr Epidemiol Rep*. 2018;5(2):166-174.
31. Rabe-Hesketh S, Pickles A, Skrondal A. Correcting for covariate measurement error in logistic regression using nonparametric maximum likelihood estimation. *Stat Modelling*. 2003;3(3):215-232.
32. Rabe-Hesketh S, Skrondal A, Pickles A. Maximum likelihood estimation of generalized linear models with covariate measurement error. *Stata J*. 2001;3(4):386-411.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

How to cite this article: Nab L, Groenwold RHH, Welsing PMJ, van Smeden M. Measurement error in continuous endpoints in randomised trials: Problems and solutions. *Statistics in Medicine*. 2019;38:5182–5196. https://doi.org/10.1002/sim.8359