Approaches to addressing missing values, measurement error, and confounding in epidemiologic studies

Smeden, M. van; Vries, B.B.L.P. de; Nab, L.; Groenwold, R.H.H.

Journal of Clinical Epidemiology 131 (2021) 89–100

Journal of Clinical Epidemiology

ORIGINAL ARTICLE

Approaches to addressing missing values, measurement error, and confounding in epidemiologic studies

Maarten van Smeden[a,*], Bas B.L. Penning de Vries[a], Linda Nab[a], Rolf H.H. Groenwold[a,b]

[a]Department of Clinical Epidemiology, Leiden University Medical Center, Leiden, The Netherlands
[b]Department of Biomedical Data Sciences, Leiden University Medical Center, Leiden, The Netherlands

Accepted 4 November 2020; Published online 8 November 2020

Abstract

Objectives: Epidemiologic studies often suffer from incomplete data, measurement error (or misclassification), and confounding. Each of these can cause bias and imprecision in estimates of exposure–outcome relations. We describe and compare statistical approaches that aim to control all three sources of bias simultaneously.

Study Design and Setting: We illustrate four statistical approaches that address all three sources of bias, namely, multiple imputation for missing data and measurement error, multiple imputation combined with regression calibration, full information maximum likelihood within a structural equation modeling framework, and a Bayesian model. In a simulation study, we assess the performance of the four approaches compared with more commonly used approaches that do not account for measurement error, missing values, or confounding.

Results: The results demonstrate that the four approaches consistently outperform the alternative approaches on all performance metrics (bias, mean squared error, and confidence interval coverage). Even in simulated data of 100 subjects, these approaches perform well.

Conclusion: There can be a large benefit of addressing measurement error, missing values, and confounding to improve the estimation of exposure–outcome relations, even when the available sample size is relatively small.

Keywords: Data analysis; Confounding; Measurement error; Missing data; Simulation; Regression calibration; Imputation; Regression

1. Background

Researchers in epidemiology often aim to make inferences about causal relations between an exposure and a health outcome while analyzing observational data that are incomplete, contain measurement error (and misclassifications), and are subject to confounding, which are among the important analytical obstacles that prohibit making direct causal claims. These sources of bias will likely only become more prevalent and important in epidemiology with increasing use of data that were not collected primarily for scientific research, such as insurance claims databases and electronic health records [1,2].

Solutions to alleviate the consequences of incomplete (or missing) data, measurement error, and confounding have been proposed, some of which are widely implemented in epidemiologic literature. For instance, to account for missing values, approaches include likelihood-based methods [3,4], Bayesian models [5], and multiple imputation [6,7]. To account for measurement error or misclassification, likelihood-based methods [8], Bayesian models [9], multiple imputation [10,11], and regression calibration [12] have been proposed, among others. And to account for confounding, one could apply matching, stratification [13], propensity score methods [14], and multivariable regression adjustment [15,16], to name a few. Some of these methods can be applied sequentially to account for a combination of missing data, measurement error, and confounding. For instance, multiple imputation to account for missing data

Authors' contributions: M.v.S. contributed to conceptualization, methodology, software, and writing the article. B.B.L.P.d.V. contributed to methodology, software, and reviewing and editing the article. L.N. contributed to methodology, software, validation, and reviewing and editing the article. R.H.H.G. contributed to conceptualization, funding acquisition, methodology, and writing and editing the article.

Declarations of interest: none.

Funding: This work was supported by grants from the Netherlands Organization for Scientific Research (ZonMW-Vidi project 917.16.430) and the Leiden University Medical Center.

Simulation code available at: https://github.com/MvanSmeden/MiMeCo. Example data freely available from the National Health and Nutrition Examination Survey website.

Conflicts of interest: None.

* Corresponding author. Tel.: +31 88 75 509 40; fax: +31 88 75 680 99.
E-mail address: M.vanSmeden@umcutrecht.nl (M. van Smeden).

https://doi.org/10.1016/j.jclinepi.2020.11.006

What is new?

Key findings

- In a simulation study, four analysis strategies were shown to be able to effectively correct for the presence incomplete data, measurement error and confounding, even in settings where the dataset was relatively small (N = 100).
- The proposed methods are all effective in reducing bias and differ primarily in their statistical efficiency.

What this adds to what is known?

- Approaches that aim to correct for incomplete data, measurement error as well as confounding remain rarely applied. This article proposes, describes and evaluates four approaches that make use of an internal validation set to correct for measurement error.

What is the implication and what should change now?

- Incomplete data, measurement error and confounding are common sources of bias in epidemiologic research, which often require statistical adjustment to valid causal inference. Given that epidemiologic studies commonly suffer from all three issues, statistical solutions as discussed in this article should be considered more often in epidemiologic research practice.

may precede the development of a multivariable regression model to adjust for confounding or the development of a propensity score model to be used for marginal structural modeling [17]. As we will illustrate, Bayesian and frequentist methods exist that allow for the specification of separate models for the missing data, measurement error, and exposure−outcome models to be estimated simultaneously (i.e., iteratively) instead of in a sequential order.

Despite the availability of methodology, it appears that methods to adjust for missing data and measurement error have not yet found their way into epidemiologic research practice. Complete case analysis remains the most commonly used technique in epidemiology [4], ignoring potentially valuable observed information by excluding subjects with incomplete observations. Measurement error—while often discussed as an important study limitation—remains rarely accounted for in epidemiologic data analyses [18−20]. It is therefore unsurprising that the applications of combined methods to address all three aforementioned sources of bias are rare in the epidemiologic literature.

In this article, we illustrate and study the performance of combinations of analytical methods (i.e., analysis strategies) that aim to account for incomplete data, measurement error, and confounding when estimating the exposure−outcome relation either by applying multiple correcting methods consecutively and in a specific order or addressing all issues simultaneously. We will limit ourselves to analysis strategies that can be used when outcomes are measured on a continuous scale.

2. Illustrative example: effect of dietary fat intake on systolic blood pressure

We illustrate the possible impact of inadequate control of confounding, measurement error, and missing data using an example of a study of fat intake and blood pressure.

We used data from the 2003−2004 National Health and Nutrition Examination Survey on 2,075 subjects with 24-hour and 30-day recall measurements to obtain an estimate of the effect of dietary fat intake on systolic blood pressure. Twenty-four-hour recalled total fat intake measured in grams was taken as the exposure variable of choice. We anticipate the 24-hour recall of fat intake is subject to less recall bias, which we assume will be the major source of measurement error in 30-day recall of dietary fat intake. We assumed that the causal assumptions (details in Appendix) were met by conditioning on age in years, gender, diastolic blood pressure, and body mass index (BMI), which was computed from weight in kilograms and height in centimeters measured by a trained health technician. Systolic and diastolic blood pressures were computed as the average of three subsequent measurements (in mm Hg) administered by a trained examiner. Details for each of the measurements are given on the National Health and Nutrition Examination Survey website (https://www.cdc.gov/nchs/nhanes/). We emphasize that this illustrative example is simplified, and thorny issues such as BMI being a possible intermediate variable on the pathway from fat intake to the outcome and measurement error present in systolic and diastolic blood measurements are not further considered.

We first estimated the crude effect (i.e., unadjusted for potential confounders) of exposure, estimated by (ordinary least squares) linear regression with log-transformed total fat intake as the exposure variable of interest and systolic blood pressure as the outcome variable. Total fat intake measured by the 24-hour-recall was treated as the gold standard (GS) measurement for total fat intake. To evaluate the robustness of the results to measurement error in total fat intake, we replaced the GS measurement with error-prone measurements of 30-day self-reported total fat intake. Pearson's correlation between the GS 24-h recall and error-prone 30-day average total fat intake was 0.89 (95% confidence interval [CI]: 0.88−0.89). The effect of incomplete data was illustrated by re-estimating the regression on only the subjects with complete information on

Table 1. Estimates of the relation between dietary fat intake and blood pressure in National Health and Nutrition Examination Survey (2003–2004)

Total fat intake	Confounding adjustment	Complete case analysis	Estimate (in mm Hg)	95% Confidence interval		*N*
				Lower bound	Upper bound	
24-h recall	Crude effect	No	9.15	6.29	12.02	2,075
30-d average	Crude effect	No	8.01	5.21	10.80	2,075
30-d average	Crude effect	Yes	7.51	4.68	10.34	2,018
30-d average	Adjusted	Yes	4.21	1.95	6.47	2,018

exposure, outcome, and confounders (i.e., the complete cases). Multivariable regression adjustment was performed by adding the confounding variables (age, BMI, gender, and diastolic blood pressure) as linear effects in the regression of systolic blood pressure on total fat intake.

The estimated crude exposure effect corresponded to a 9.15 mm Hg (95% CI: 6.29–12.02) increase in blood pressure for every 1-unit increase in log-transformed total fat intake (Table 1). Replacing the exposure of interest with 30-day average total fat intake resulted in a decrease of the effect to 8.01 mm Hg (95% CI: 5.21–10.80) increase in blood pressure for every 1-unit increase. This effect further reduced (7.51 mm Hg, 95% CI: 4.68–10.34) when estimated using information about the 2,018 complete cases only. Multivariable regression adjustment for confounding using information about the complete cases and with exposure defined by 30-hour total fat intake resulted in a further reduction in the effect to 4.21 mm Hg (95% CI: 1.95–6.29) increase in blood pressure for every 1-unit increase in log-transformed total fat intake.

Although each of the previously mentioned analyses suffers from different degrees of bias, none should be considered a preferred method of analysis for these data. Instead, this example illustrates the sensitivity of estimated exposure effects to confounding, incompleteness of data, and measurement error. In settings where all three of these issues occur to a relevant degree, which we consider is often the case in epidemiologic research, solutions that handle only one or two of them are unlikely to provide unbiased results.

3. Simulation

3.1. Simulation setup

We conducted a simulation study to investigate to what extent confounding, incompleteness, and measurement error can be accounted for by various analysis strategies. The outcome data, Y, without measurement error and incompleteness, are assumed normally distributed, given exposure status (A) and P confounder values (L): $Y \mid A,L \sim \mathrm{N}(\beta_0 + \beta_e A + \beta_{L1}L_1 + \cdots + \beta_{LP}L_P, \sigma^2)$. For the exposure effect, β_e, a natural estimator is β_e in the ordinary least squares linear regression: $\mathrm{E}[Y|A,L] = \beta_0 + \beta_e A + \beta_{L1}L_1 + \cdots + \beta_{LP}L_P$, assuming linearity of exposure and confounder effects and absence of interactions. The exposure status of primary interest (A) is measured and available for a random validation subsample of size $N^* < N$. We assume that a surrogate measurement of exposure, A^*, with nondifferential measurement error is measured and available for every of N units (i.e., simulated subjects). Furthermore, we assume there are missing values on R data points in the confounding variable L_1. These are introduced such that the missingness is dependent on the surrogate (or error-prone) exposure status (A^*) and outcome status (Y) and a different confounding variable (L_2). This missing data mechanism satisfies the missing at random assumption.

For the simulation study, exposure and confounder data were sampled from a multivariate normal distribution with standard deviation 1 and equal pairwise correlations. The value of the effect of exposure, β_e, was arbitrarily set to 10. The correlations and number of confounding variables were varied between 0.05 and 0.40 and 2 and 6, respectively. The value of the residual variance (σ^2) was determined by the coefficient of determination (R^2), which was varied between 0.10 and 0.30, as detailed in Supplementary Material 1. The total sample size varied between 100 and 2,000, and the percentage of missing values on L_1 was set to 30%. The size of the random validation subset varied between 10% and 30% of the total sample size.

In total, 100,000 simulation data sets were created by randomly drawing values for the simulation factors, with equal probability for integer-valued factors (number of confounders and sample size) and for the other factors from a uniform distribution with limits as outlined previously and summarized in Table 2. To simulate the effects of ignored incompleteness, measurement error, and confounding, we varied the expected attenuation, i.e., the percentage bias in the estimator of exposure toward the null effect, for each of the effects between −1% and −50%, by adjusting the data-generating mechanisms accordingly (Supplementary Material 1). The expected biases were in the same direction by design to avoid they would cancel each other out. Simulation results are presented as averages over the simulated data sets. Relative bias, empirical standard errors, mean squared error, 95% CI (credible interval for the Bayesian model) coverage, and Monte Carlo standard errors, which provide estimates of the standard errors of these estimated simulation outcomes, were calculated as detailed in a study by Morris et al. [35].

Table 2. Simulation factors and ranges

Simulation factors	Symbol	Values	
		Minimum	Maximum
Sample size	N	100	2,000
Model explained variance in Y	R^2	10%	30%
Pairwise correlations exposure and confounders	ρ	0.10	0.40
Confounding			
Bias due to ignored confounding	Δ	−1%	−50%
Number of confounding variables	P	2	6
Measurement error exposure variable			
Bias due to ignored measurement error	Λ	−1%	−50%
Size of internal validation subset	N^*	10% of N	30% of N
Missing values			
Bias due to ignored missing values	K	−1%	−50%

3.2. Analysis strategies

On each generated data set, we applied 10 different analysis strategies (Table 3). For reference, we applied a hypothetical GS, where the true exposure status and confounders are fully observed and the confounder adjusted exposure effect could be estimated by the linear regression, regressing the outcome on the exposure (A) and the confounders (L_1, …, L_p). For reference, we also considered the following linear regression models:

- CCUN: crude effect of error-prone exposure (A^*), no confounding adjustment, selection of complete cases only;
- CRUD: crude effect of error-prone exposure (A^*), no confounding adjustment;
- LADJ: adjustment for confounders, using error-prone exposure (A^*), selection of complete cases only;
- VALD: adjustment for confounders, using true exposure (A), selection of complete cases within the validation subsample.

Wald-based 95% CIs were constructed for the parameters of each of the models. Note that none of these models fully account for confounding, incomplete data, as well as measurement error.

We also implemented three approaches in which multiple imputation by chained equation was used to account for incomplete data [36]. We considered the following approaches:

- MIMD: imputing the missing values on L_1, followed by linear regression with adjustment for confounders, using error-prone exposure (A^*);
- MIME: imputing the missing values on L_1 and on A, followed by linear regression with adjustment for confounders, using true exposure (A);

Table 3. Analysis strategies for epidemiologic studies that face incomplete data, measurement error, and confounding

Description	Incomplete data	Measurement error	Confounding	Sequential or simultaneous adjustment	
GS	Not applicable	Not applicable	✓	Not applicable	Reference: theoretical gold standard
CRUD	Not applicable	X	X	Not applicable	Crude effect estimates of exposure
CCUN	X	X	X	Not applicable	Crude effect in confounder data complete cases
LADJ	X	X	✓	Not applicable	Confounder adjusted
VALD	X	✓	✓	Not applicable	Confounder adjusted in validation subset
MIMD	✓	X	✓	Sequential	Imputation of missing confounder data
MIME	✓	✓	✓	Sequential	Imputation of confounder and true exposure data
MIRC	✓	✓	✓	Sequential	Imputation confounder and regression calibration
FIML	✓	✓	✓	Simultaneous	Full information maximum likelihood
BAYES	✓	✓	✓	Simultaneous	Bayesian model

X = no, ✓ = yes.

Table 4. Simulation results: bias, empirical standard error, mean squared error, and 95% confidence interval coverage (coverage) with MCSE

		Relative bias		Empirical standard error		Mean squared error		95% Confidence interval coverage	
N	**Model**	**Est (%)**	**MCSE**	**Est**	**MCSE**	**Est**	**MCSE**	**Est (%)**	**MCSE**
100–700	GS	−0.18	0.01	0.00	0.00	2.35	0.01	95.15	0.07
	CRUD	−42.91	0.03	0.01	0.00	22.40	0.03	16.76	0.12
	LADJ	−45.87	0.03	0.02	0.00	26.09	0.04	23.72	0.13
	MIMD	−25.88	0.02	0.01	0.00	10.51	0.02	43.96	0.16
	MIME	1.61	0.01	0.01	0.00	7.13	0.04	94.34	0.07
	FIML	0.69	0.01	0.01	0.00	6.86	0.06	93.85	0.08
	MIRC	−0.23	0.01	0.01	0.00	4.16	0.02	92.20	0.08
	BAYES	0.19	0.01	0.01	0.00	3.85	0.02	95.06	0.07
700–1,400	GS	−0.03	0.01	0.00	0.00	0.71	0.00	95.10	0.07
	CRUD	−42.64	0.03	0.01	0.00	20.89	0.03	4.57	0.07
	LADJ	−45.73	0.03	0.02	0.00	24.20	0.03	8.28	0.09
	MIMD	−25.50	0.02	0.01	0.00	9.02	0.02	26.84	0.14
	MIME	1.07	0.01	0.00	0.00	1.80	0.01	93.94	0.08
	FIML	0.19	0.01	0.00	0.00	1.60	0.01	94.72	0.07
	MIRC	−0.09	0.01	0.00	0.00	1.18	0.00	92.19	0.08
	BAYES	0.02	0.01	0.00	0.00	1.08	0.00	94.93	0.07
1,400–2,000	GS	−0.01	0.00	0.00	0.00	0.42	0.00	95.03	0.07
	CRUD	−42.67	0.03	0.01	0.00	20.73	0.02	2.29	0.05
	LADJ	−45.86	0.03	0.02	0.00	24.03	0.03	4.88	0.07
	MIMD	−25.54	0.02	0.01	0.00	8.84	0.01	20.76	0.13
	MIME	0.82	0.01	0.00	0.00	1.03	0.00	94.41	0.07
	FIML	0.23	0.01	0.00	0.00	0.93	0.00	94.87	0.07
	MIRC	0.04	0.00	0.00	0.00	0.69	0.00	92.31	0.08
	BAYES	0.08	0.00	0.00	0.00	0.64	0.00	95.08	0.07

Abbreviations: Est = estimate; MCSE = Monte Carlo standard error.

- MIRC: imputing the missing values on L_1, followed by linear regression with adjustment for confounders, using error-prone exposure (A^*) in combination with regression calibration to correct for measurement error (using information from the internal validation sample). Regression calibration implements a linear regression where Y is regressed on a corrected version of the error-prone measures of A^* and the confounding variables. The error-prone measures of A^* is subsequently replaced by the predicted mean of A, given A^* and the confounding variables and standard errors are adjusted.

The latter two models attempt to fully account for confounding, incomplete data, and measurement error because missing values on L_1 and A (for individuals not in the validation subset) were imputed simultaneously (MIME) or the measurement error in A^* was corrected by means of regression calibration (MIRC). Imputation proceeded by predictive mean matching, generating 10 imputation data sets. Rubin's rules [37] were used to pool the results from the multiple imputed data sets.

In addition, we specified a structural equation model (full information maximum likelihood [FIML]) that allowed for estimating the effect of exposure by simultaneous addressing incompleteness, measurement error, and confounding. In short, the approach entails maximizing the joint likelihood of the observed part of A^*, Y, A, and L_1, given $L_2,\ldots, L_P$, based on the following conditional models: $A^*\Big|Y,A,L\sim N(\mu_{A^*|Y,A,L},\sigma^2_{A^*|Y,A,L})$, $Y\Big|A,L\sim N(\mu_{Y|A,L},\sigma^2_{Y|A,L})$ $A|L\sim N\Big(\mu_{A|L},\sigma^2_{A|L}\Big)$, and $L_1\Big|L_2,\ldots,L_P\sim N(\mu_{L_1|L_2,\ldots,L_P},\sigma^2_{L_1|L_2,\ldots,L_P})$ [38].

Finally, a Bayesian model (BAYES) was fitted that was parameterized using the same equation as the FIML model. The model uses a Gibbs sampler to sample from the posterior distributions of the parameters. We used uninformative priors for all parameters (normal with mean zero and variance 1,000 or gamma with .001 shape and rate parameters for residual variances), 1,000 burn-in samples, after which 2,000 samples were drawn from the posterior to estimate the effect of exposure taking the mean. Convergence of the Bayesian approach was checked with the Gelman-Rubin statistic (cut-off 1.2).

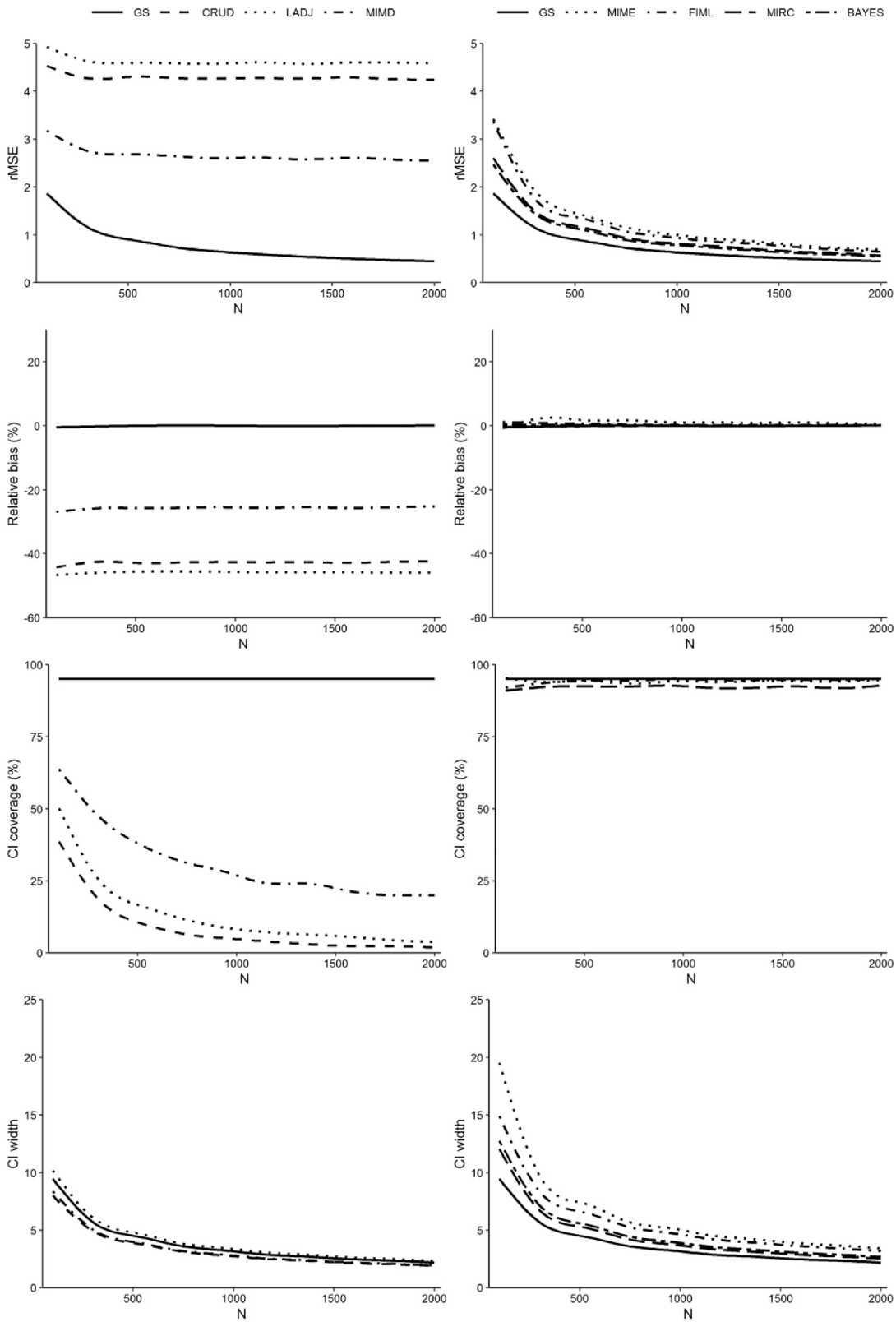

Fig. 1. Simulation results: average performance of analysis strategies. CI, confidence interval; rMSE: root mean squared error.

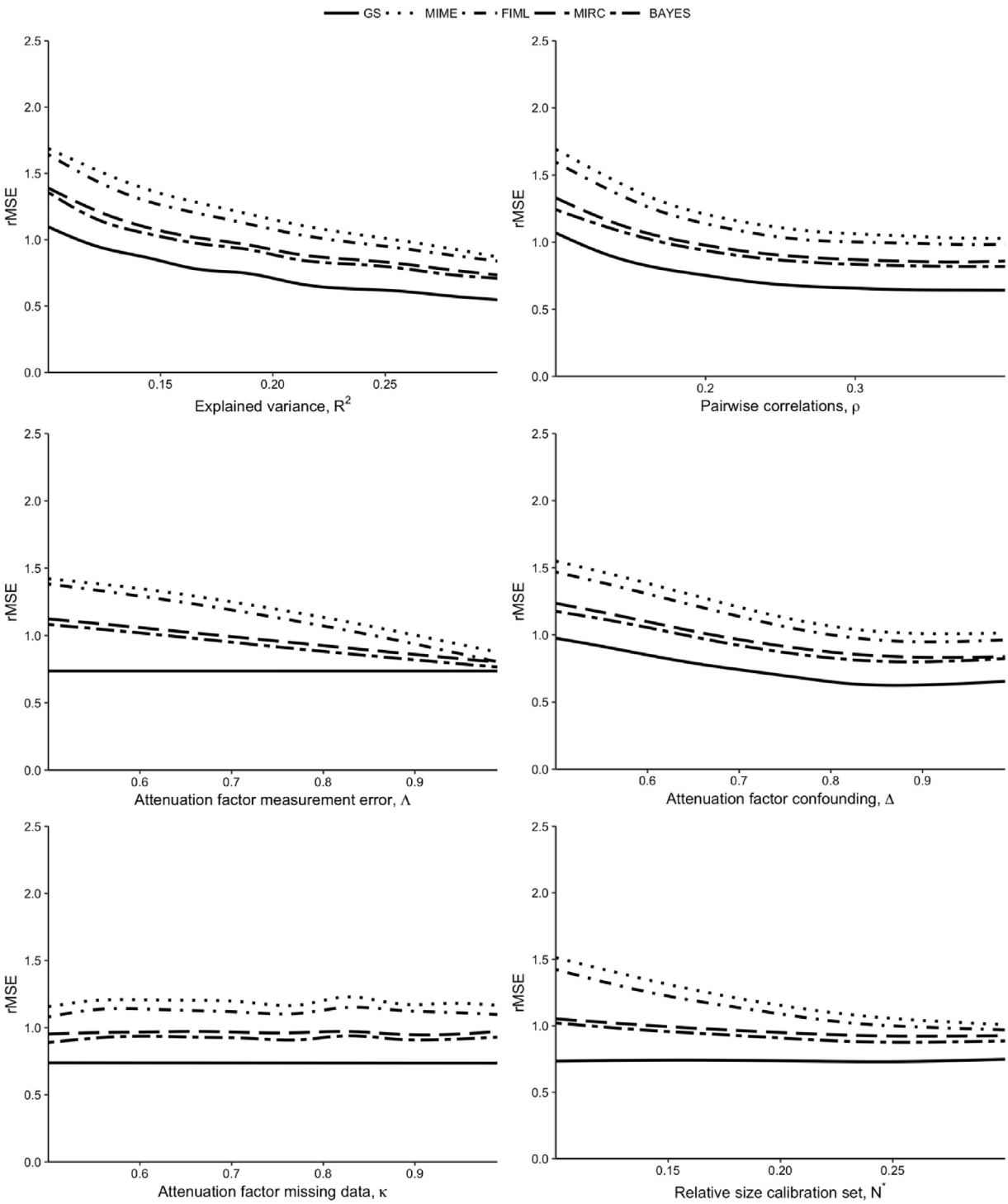

Fig. 2. Simulation results: average root mean squared error. CI, confidence interval; rMSE, root mean squared error.

Simulations were performed in *R* (version 3.5.0) (R Core Team) [39] with packages *mice* (version 3.4.0) [40] for multiple imputation, *mecor* (version 0.1.0) [41] for regression calibration, *lavaan* (version 0.6-4) [42] for the FIML model and *rjags* (version 4-8) [43] to estimate the Bayesian model. The full simulation code, including implementation of the different analysis strategies, is available at https://github.com/MvanSmeden/MiMeCo.

3.3. Results

Nonconvergence rates were close to 0% for almost all strategies (see Supplementary Material 2). For brevity, we

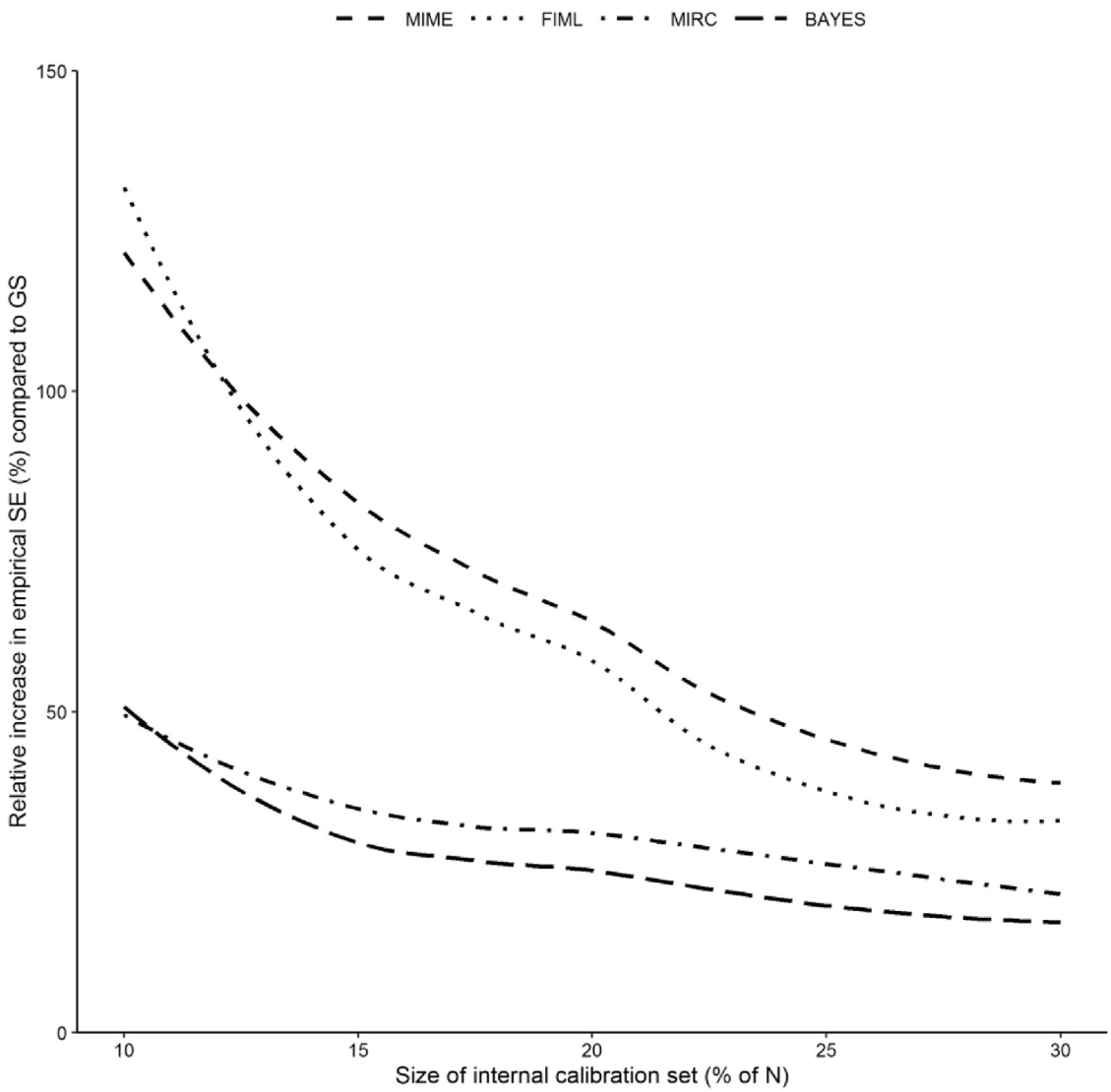

Fig. 3. Simulation results: average relative efficiency of analytical methods to correct for incomplete data, measurement error, and confounding.

removed two reference strategies (VALD and CCUN) from the main results because these performed poorly as expected and do not contribute to the main comparisons being made (results for these strategies are presented in Supplementary Material 3).

Table 4 shows the average performance of the remaining analysis strategies. As expected, the models that adjusted for confounding, incompleteness, and measurement error (MIME, FIML, MIRC, and BAYES) performed consistently better and were closer to the GS model than the approaches that did not account for all three sources of bias (CRUD, LADJ, and MIMD; Fig. 1). Relative bias was generally small for MIME, FIML, MIRC, and BAYES ($<2\%$), with the smallest average relative bias observed for BAYES. BAYES also showed lower average mean squared error (MSE) across sample sizes than MIME, FIML, and MIRC (Table 4). CI coverage was closer to the nominal level of 95% for BAYES, FIML, and MIME than MIRC, with MIRC showing slight undercoverage. The CI coverage of the CRUD, LADJ, and MIMD models was consistently poor and worsening with increasing sample size.

The MIME, FIML, MIRC, and BAYES models showed little variation in bias and 95% CI coverage across values of the coefficient of determination (R^2), correlation between the confounders and exposure variables (ρ), and the attenuating effects of unadjusted incompleteness, measurement error, and confounding (Fig. 2). All three incomplete adjustment approaches (CRUD, LADJ, and MIMD) showed decreasing MSE with decreasing magnitude of measurement error (Fig. 2). The order of average MSE remained consistent (MIME $>$ FIML $>$ MIRC $>$ BAYES) across the simulation factors.

Fig. 3 shows the efficiency of the adjustment approaches relative to the GS, that is, the relative increase in the

standard errors of the adjustment methods compared with the standard error of the GS strategy. Clearly, the efficiency of MIME, FIML, MIRC, and BAYES increased when the size of the internal validation subset increased. The differences between these approaches were substantial for the smaller validation subset. The BAYES had lowest variance on average across the range of internal validation subset size, followed by MIRC.

4. Discussion

In this article, we demonstrated several analysis strategies to estimate an exposure effect on a continuous outcome while accounting for three common sources of bias, namely, incompleteness of data, measurement error, and confounding. Statistical approaches that aim to account for all three sources of bias remain uncommon in epidemiology despite that these methods are generally well described, and software packages are widely available, making the implementation of these corrections relatively straightforward. Our simulations demonstrated for a large range of scenarios that these approaches can be beneficial relative to commonly used analysis approaches that address these issues only in part by eliminating bias and greatly reducing mean squared error, even in relatively small data sets (as low as $N = 100$).

As others have shown before [44], confounding, incomplete data, and measurement error can be conceptualized as special types of missing data. This study is one of the first to compare analysis strategies that aim to recover the exposure effect, when faced with incomplete data, measurement error, and confounding, by applying various correction approaches either sequentially (using multiple imputation and regression calibration) or simultaneously (FIML or Bayesian modeling). Among the analysis strategies that were considered, the Bayesian model showed the lowest variance while eliminating bias and reducing mean squared error across settings. Other approaches performed slightly less well, although all clearly outperformed the partial analysis strategies that did not account for the incompleteness of data, measurement error, or confounding.

The strategies that we implemented are flexible and can be extended to account for more complex multivariate missingness [34], differential measurement error, measurement error in the outcome or confounders, and a larger number of confounders than we have considered. Another advantage of these strategies is that it is relatively straightforward to apply them in standard software programs, such as R software (see Supplementary Material 4 for example code), SAS, and Stata. Nonetheless, a possible hurdle for implementing these approaches in applied epidemiologic research is the need for an internal validation subset to estimate the measurement error model. Although the sample size required for the internal validation subset may be only a small portion of the total sample size, a GS measurement is not always feasible to apply, for instance, because it does not always exist. Alternative approaches based on sensitivity analyses, latent variable models, and Bayesian models that address measurement error in the absence of a GS are discussed elsewhere in detail [45,46].

This study also has limitations. First, our simulation study was restricted to relatively simple confounding, measurement error, and missing data structures, to normally distributed outcome and exposure variables only, and relied on parametric estimation and adjustment approaches. Extensions of this study should shed light on the generalizability of our results to other types of data, models for different outcomes (e.g., binary logistic regression), different parameterizations of the models (e.g., by modifying the joint likelihood underlying the FIML and Bayesian modeling approaches), and nonparametric models. Second, we have not investigated the role of model misspecification. As the adjustments for confounding, measurement error, and missing data rely on different assumptions about modeling structure, we believe it is likely that models that account for all three sources of bias can be misspecified easily. Future studies that evaluate and compare the robustness of statistical models to realistic forms of model misspecification are much needed.

To conclude, confounding, incompleteness of data, and measurement error can each contribute to large biases in epidemiologic studies. Addressing all three in the data analysis is feasible, even when the sample size is relatively small. Given that epidemiologic studies commonly suffer from all three issues, statistical solutions as discussed in this article should be considered more often in epidemiologic research practice.

Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jclinepi.2020.11.006.

References

[1] Roubinian N, Brambilla D, Murphy EL. Statistical caution in big data approaches to transfusion medicine research. JAMA Intern Med 2017;177(6):860–1.
[2] Ehrenstein V, Nielsen H, Pedersen AB, Johnsen SP, Pedersen L. Clinical epidemiology in the era of big data: new opportunities, familiar challenges. Clin Epidemiol 2017;9:245–50.
[3] Ibrahim JG, Chen M-H, Lipsitz SR, Herring AH. Missing-data methods for generalized linear models. J Am Stat Assoc 2005;100:332–46.
[4] Eekhout I, de Boer RM, Twisk JWR, de Vet HCW, Heymans MW. Missing data: a systematic review of how they are reported and handled. Epidemiology 2012;23:729–32.
[5] Erler NS, Rizopoulos D, van Rosmalen J, Jadoe VWV, Franco OH, Lesaffre EMEH. Dealing with missing covariates in epidemiologic

studies: a comparison between multiple imputation and a full Bayesian approach. Stat Med 2016;35:2955−74.
[6] Groenwold RHH, Donders ART, Roes KCB, Harrell Jr FE, Moons KGM. Dealing with missing outcome data in randomized trials and observational studies. Am J Epidemiol 2012;175:210−7.
[7] Harel O, Mitchell EM, Perkins NJ, Cole SR, Tchetgen Tchetgen EJ, Sun B, et al. Multiple imputation for incomplete data in epidemiologic studies. Am J Epidemiol 2018;187:576−84.
[8] Carroll RJ, Ruppert D, Stefanski LA, Crainiceanu CM. Measurement error in nonlinear models: a modern perspective. Boca Raton, FL: Chapmann & Hall/CRC; 2006.
[9] Gustafson P. Measurement error and misclassification in statistics and epidemiology: impacts and Bayesian adjustments. Boca Raton, FL: Chapmann & Hall/CRC; 2003.
[10] Freedman LS, Midthune D, Carroll RJ, Kipnis V. A comparison of regression calibration, moment reconstruction and imputation for adjusting for covariate measurement error in regression. Stat Med 2008; 27:5195−216.
[11] Cole SR, Chu H, Greenland S. Multiple-imputation for measurement-error correction. Int J Epidemiol 2006;35:1074−81.
[12] Keogh RH, White IR. A toolkit for measurement error correction, with a focus on nutritional epidemiology. Stat Med 2014;33: 2137−55.
[13] Hernan MA, Hernandez-Diaz S, Werler MM, Mitchell AA. Causal knowledge as a prerequisite for confounding evaluation: an application to birth defects epidemiology. Am J Epidemiol 2002;155:176−84.
[14] Austin PC. An introduction to propensity score methods for reducing the effects of confounding in observational studies. Multivariate Behav Res 2011;46(3):399−424.
[15] Rothman KJ, Greenland S, Lash T. Modern epidemiology. 3rd ed. Philadelphia, PA: Lippincott Williams and Wilkins; 2014.
[16] Groenwold RHH, Klungel OH, Altman DG, van der Graaf Y, Hoes AW, Moons KGM. Adjustment for continuous confounders: an example of how to prevent residual confounding. Can Med Assoc J 2013;185(5):401−6.
[17] Penning de Vries BBL, Groenwold RHH. A comparison of two approaches to implementing propensity score methods following multiple imputation. Epidemiol Biostat Public Heal 2013e12630−2.
[18] Jurek AM, Maldonado G, Greenland S, Church TR. Exposure-measurement error is frequently ignored when interpreting epidemiologic study results. Eur J Epidemiol 2007;21(12):871−6.
[19] 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.
[20] Shaw PA, Deffner V, Keogh RH, Tooze JA, Dodd KW, Kuchenhoff H, et al. Epidemiologic analyses with error-prone exposures: review of current practice and recommendations. Ann Epidemiol 2018;28:821−8.
[21] Neyman J, Iwaszkiewicz K, Kolodziejczyk S. Statistical problems in agricultural experimentation. *Suppl. To.* J R Stat Soc 1935;2(2):107−80.
[22] Rubin DB. Estimating causal effects of treatments in randomized and nonrandomized studies. J Educ Psychol 1974;66(5):688−701.
[23] Hernan MA, Robins JM. Estimating causal effects from epidemiological data. J Epidemiol Community Heal 2006;60(7):578−86.
[24] Tchetgen EJT, VanderWeele TJ. On causal inference in the presence of interference. Stat Methods Med Res 2012;21(1):55−75.
[25] Cole SR, Frangakis CE. The consistency statement in causal inference. Epidemiology 2009;20:3−5.
[26] Hernan MA, Robins JM. Causal inference: What if. Boca Raton, FL: Chapmann & Hall/CRC; 2020.
[27] Lesko CR, Buchanan AL, Westreich D, Edwards JK, Hudgens MG, Cole SR. Generalizing study results: a potential outcomes perspective. Epidemiology 2017;28:553−61.
[28] Sterne JAC, White IR, Carlin JB, Spratt M, Royston P, Kenward MG, et al. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. BMJ 2009; 338:b2393.
[29] Rubin DB. Inference and missing data. Biometrika 1976;63(3): 581−92.
[30] Little RJA, Rubin DB. Statistical analysis with missing data. Hoboken, NJ: John Wiley & Sons, Inc.; 2002.
[31] Bartlett JW, Harel O, Carpenter JR. Asymptotically unbiased estimation of exposure odds ratios in complete records logistic regression. Am J Epidemiol 2015;182:730−6.
[32] Hughes RA, Heron J, Sterne JAC, Tilling K. Accounting for missing data in statistical analyses: multiple imputation is not always the answer. Int J Epidemiol 2019;48:1294−304.
[33] Michels KB. A renaissance for measurement error. Int J Epidemiol 2001;30:421−2.
[34] Van Smeden M, Lash TL, Groenwold RHH. Reflection on modern methods: five myths about measurement error in epidemiologic research. Int J Epidemiol 2019dyz251.
[35] Morris TP, White IR, Crowther MJ. Using simulation studies to evaluate statistical methods. Stat Med 2019;38:2074−102.
[36] Buuren S. Flexible imputation of missing data. Boca Raton, FL: CRC/Chapman & Hall; 2018.
[37] Rubin DB. Multiple imputation for nonresponse in surveys. Hoboken, NJ: John Wiley & Sons, Inc.; 1987.
[38] Enders CK. A primer on maximum likelihood algorithms available for use with missing data. Struct Equ Model A Multidiscip J 2001;8(1): 128−41.
[39] R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. R Foundation for Statistical Computing; 2020.
[40] Buuren S van, Groothuis-Oudshoorn K. Mice : multivariate imputation by chained equations in R. J Stat Softw 2011;45(3).
[41] Nab L, Groenwold RHH, Welsing PMJ, van Smeden M. Measurement error in continuous endpoints in randomised trials: problems and solutions. Stat Med 2019;38:5182−96.
[42] Rosseel Y. Lavaan: an R package for structural equation modeling. J Stat Softw 2012;48(2).
[43] Plummer M. rjags: Bayesian Graphical Models using MCMC. CRAN; 2019.
[44] Edwards JK, Cole SR, Westreich D. All your data are always missing: incorporating bias due to measurement error into the potential outcomes framework. Int J Epidemiol 2015;44:1452−9.
[45] Rutjes AWS, Reitsma JB, Coomarasamy A, Khan KS, Bossuyt PMM. Evaluation of diagnostic tests when there is no gold standard. A review of methods. Health Technol Assess (Rockv) 2007;11(50). ix−51.
[46] Lash TL, Fox MP, Fink AK. Applying quantitative bias analysis to epidemiologic data. New York, NY, USA: Springer; 2009.

Appendix.
Theory

We suppose the interest of a particular epidemiologic study is in estimating an average causal effect of a single not experimentally assigned factor, e.g., a treatment, called the exposure denoted by A, on a single continuous outcome variable denoted by Y. We adopt the widely used potential outcomes view on causality [21−23], which assumes that for any given individual at any particular point in time there is a value of a potential outcome for each level of exposure. The average causal effect is then defined by the expected difference in potential outcomes for a single unit increase in level of exposure. Since for any given individual in the study, the potential outcomes for all possible exposure levels other than the observed level are missing, the observed data is invariably insufficient to directly calculate the average causal effect. The computation of this effect relies on the unobservable average of within-person contrasts in potential outcomes for all levels of exposure.

To recover the average causal effect, in particular in settings where the exposure status is not experimentally assigned, several causal assumptions need to be made. The following common assumptions are known to be sufficient to recover the causal effect: i) exposure of one individual does not affect the outcomes of other individuals, also known as *no interference* [24]; ii) the observed outcome under the exposure level a equals the counterfactual outcome under treatment a, also known as *consistency* [25]; iii) within levels of covariates, L, the potential outcomes are equally distributed across individuals with different observed exposure statuses, also known as *conditional exchangeability* [26]; iv) there is a nonzero probability of acquiring each level of the exposure for each combination of levels of L, also known as *positivity* [27].

Confounding

Of all causal assumptions (implicitly or explicitly), the *conditional exchangeability* assumption (assumption iii above) is usually most in the forefront of epidemiologic data analyses. The aim is to select a set of covariates, L, that represent a sufficient set of confounding variables to ensure the *conditional exchangeability* assumption is met. In combination with the aforementioned other causal assumptions, by conditioning on the observed confounding variables in the analysis of the exposure−outcome relation (e.g., by multivariable regression or matching).

Incomplete data

Incomplete data (or records) in the form of missing values in a combination of A, Y or L are common and often unavoidable in epidemiologic data sets. By default, most statistical analyses are restricted to use only the data from individuals who have no missing values, so-called complete case analysis. Missingness in more than a few individuals is often enough to cause concerns, as it may result in a substantial loss in statistical efficiency, resulting in lower statistical power and wider confidence intervals, and biased estimates of the exposure−outcome relation [28].

Assumptions about the missing data mechanism have to be made to understand the possible impact of missing data and to improve statistical efficiency and reduce bias in the exposure-effect estimate in the analyses with incomplete data. Rubin's well-known taxonomy distinguishes three major missing data mechanisms [29,30]: Missing Completely At Random (MCAR), Missing At Random (MAR), and Missing Not At Random (MNAR). If missing data are MCAR, meaning records are missing for reasons that are unrelated to characteristics or responses for the subjects, including the values of the missing records were they be known, analysis of only complete records is generally less efficient but (large sample) unbiased. Complete case analysis of the incomplete data are generally biased when data are not MCAR, although there are exceptions that are described elsewhere [31,32].

Imputation methods for missing data are a popular group of statistical approaches to account for incomplete data in epidemiologic data analysis, applied in particular when the missing data mechanisms can be assumed MCAR or MAR. In brief, imputation aims to predict the missing values from the data that is observed, and to use these predictions as well as the uncertainty of the predictions to obtain a consistent and efficient estimator of the exposure−outcome relation.

Measurement error

Measurement error in the form of mistaken classifications and inaccurate recordings are often unavoidable in epidemiologic data collection [33]. One way to conceptualize measurement error in A^*, Y^*, or L^*, where the *-notation indicates the observed variable contains some form of measurement error, are imperfect measurements of the unobserved random variables A, Y and, L. The true values on A, Y, or L can thus be considered missing [11]. Measurement error generally leads to a loss in statistical efficiency and bias in estimates of the exposure−outcome relation [34].

Assumptions about the measurement error mechanism are often made with respect to the measurement error model and the (in)dependence of the error in the measurement [12]. The measurement error model is assumed *classical* if the error-prone measurement randomly fluctuates around its true value and the model is assumed *systematic* if the error-prone measurement is systematically different from its true value. If the error in measurement occurs in exposure or outcome (A^* or Y^*), the error is described as *nondifferential* when the error in A^* or Y^* is independent of the true values of Y or A respectively, and *differential* otherwise [15]. When two or more variables are measured with error, errors are said to be *independent* when the errors are statistically independent and

dependent otherwise. (Note that nondifferential and dependent measurement error in other literature sometimes refers to broader definitions that involve independence assumptions that are conditional on other modeled covariates.)

Approaches to ‘‘correct’’ for measurement error in the analyses can be broadly categorized in three groups: 1) approaches that rely on measurements with and without error in an internal subgroup or external group; 2) approaches that rely on repeated measurements; 3) sensitivity or (quantitative) bias analyses. A variety of statistical approaches that accommodate for data in the form of 1) or 2), including regression calibration and imputation for measurement error, have been developed to recover an unbiased and more efficient estimator of the exposure−outcome relation [34].