Received: 3 November 2021 | Revised: 2 August 2022 | Accepted: 28 August 2022

DOI: 10.1002/jrsm.1606

RESEARCH ARTICLE

Research Synthesis Methods WILEY

Adjusting for misclassification of an exposure in an individual participant data meta-analysis

Valentijn M. T. de Jong[1,2,3] | **Harlan Campbell**[4] | **Lauren Maxwell**[5] | **Thomas Jaenisch**[5,6,7] | **Paul Gustafson**[4] | **Thomas P. A. Debray**[1,2]

[1]Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht University, Utrecht, The Netherlands

[2]Cochrane Netherlands, University Medical Center Utrecht, Utrecht University, Utrecht, The Netherlands

[3]Data Analytics and Methods Task Force, European Medicines Agency, Amsterdam, The Netherlands

[4]Department of Statistics, University of British Columbia, Vancouver, British Columbia, Canada

[5]Heidelberg Institute of Global Health, Heidelberg Medical School, Heidelberg University, Heidelberg, Germany

[6]Center for Global Health, Colorado School of Public Health, Aurora, Colorado, USA

[7]Department of Epidemiology, Colorado School of Public Health, Aurora, Colorado, USA

Correspondence
Valentijn M. T. de Jong, Julius Center for Health Sciences and Primary Care, UMC Utrecht, 3508GA Utrecht, The Netherlands.
Email: valentijn.m.t.de.jong@gmail.com

Funding information
Horizon 2020 Framework Programme under ReCoDID grant agreement, Grant/Award Number: 825746; Canadian Institutes of Health Research, Institute of Genetics (CIHR-IG), Grant/Award Number: 01886-000

Abstract

A common problem in the analysis of multiple data sources, including individual participant data meta-analysis (IPD-MA), is the misclassification of binary variables. Misclassification may lead to biased estimators of model parameters, even when the misclassification is entirely random. We aimed to develop statistical methods that facilitate unbiased estimation of adjusted and unadjusted exposure-outcome associations and between-study heterogeneity in IPD-MA, where the extent and nature of exposure misclassification may vary across studies. We present Bayesian methods that allow misclassification of binary exposure variables to depend on study- and participant-level characteristics. In an example of the differential diagnosis of dengue using two variables, where the gold standard measurement for the exposure variable was unavailable for some studies which only measured a surrogate prone to misclassification, our methods yielded more accurate estimates than analyses naive with regard to misclassification or based on gold standard measurements alone. In a simulation study, the evaluated misclassification model yielded valid estimates of the exposure-outcome association, and was more accurate than analyses restricted to gold standard measurements. Our proposed framework can appropriately account for the presence of binary exposure misclassification in IPD-MA. It requires that some studies supply IPD for the surrogate and gold standard exposure, and allows misclassification to follow a random effects distribution across studies conditional on observed covariates (and outcome). The proposed methods are most beneficial when few large studies that measured the gold standard are available, and when misclassification is frequent.

KEYWORDS

individual participant data, measurement error, meta-analysis, misclassification

1 | INTRODUCTION

Individual participant data meta-analysis (IPD-MA) comprises the pooling and subsequent analysis of the participant-level data from multiple studies. As an IPD-MA summarizes the evidence through synthesis and analysis of all data available to answer a specific research question, it is generally seen as the highest standard of scientific evidence.[1] It is therefore unsurprising that IPD-MA have become increasingly common to summarize the evidence from experimental and observational studies, and that their results can substantially impact clinical practice. Although IPD-MA are frequently conducted to study the efficacy of therapeutic interventions, they can also be used to investigate etiologic, diagnostic, and prognostic variables. In observational research, data are commonly gathered using methods or instruments that are prone to measurement error (ME), but this may also occur in randomized controlled trials (RCTs).[2–4]

ME is any difference between the value that is observed for a variable and its true value. ME may arise due to a variety of random or systematic causes, such as errors in measurement instruments or their application, the reading of such instruments, poor recall, misunderstanding items on questionnaires, and data entry and management. The presence of ME may introduce (upward or downward) bias in estimators of parameters, even when the error is entirely random and independent of other variables.[5–7]

ME in categorical variables is referred to as misclassification. It is commonly believed that misclassification of the exposure leads to attenuation of exposure-outcome associations.[8] As a result, researchers often interpret estimates as conservative and dismiss the need for more advanced analyses that account for ME.[9] However, attenuation is only guaranteed to occur when the misclassification is non-differential (that is, misclassification is independent of the outcome given the measured covariates),[5,7,10–12] the exposure has no more than two categories[13,14] and all covariates are measured without error.[6] When a covariate is also measured with error, the bias introduced by including the mismeasured covariate in a multivariable regression analysis becomes much more difficult to quantify.[6] Further, extreme misclassification can reverse the sign of the observed association.[15]

In an individual participant data meta-analysis (IPD-MA), misclassification may be present in one or more studies. For instance, when the IPD from previously published studies are combined, a less accurate measurement instrument for a certain exposure variable may have been used in some studies. If one of these instruments is prone to misclassification, this will result in a biased estimator for the corresponding exposure's effect. Therefore, in IPD-MA it is generally recommended to standardize measurements, and where possible to adjust for misclassification to reduce bias.[16,17]

In meta-analysis, methods must also account for the effects of clustering in individual studies[18] and should allow for heterogeneity of the effect of interest. Hence, methods that account for misclassification must do so as well. Further, it may occur that different measurement methods are used across studies. This directly implies that a gold standard measurement may be missing for entire studies. Applying a regression calibration to account for misclassification requires that the estimated probability of misclassification s transportable to other studies. This may be tenable when the measurement instruments, protocol, population, and setting are the same in the included studies, but this would be a rare occasion in the context of IPD-MA. Hence, a method that accounts for possible heterogeneity across studies in misclassification as well as outcome prevalence and the exposure-outcome association should then be applied.

In this article, we consider a binary exposure in an IPD-MA that is prone to misclassification error. We distinguish between measurements that are obtained (or defined) according to the gold standard, and measurements that are made using an instrument that is prone to error (further referred to as the surrogate exposure). We subsequently discuss how valid inferences (at least to a certain degree) can be made while the gold standard measurements for the exposure are missing in some studies, using information on the surrogate exposure and the observed participant characteristics. We adopt a Bayesian estimation framework that extends previously proposed methods[19–21] for addressing misclassification in single studies and in aggregate data meta-analysis (AD-MA).

In Section 2 we provide our motivating example of the diagnosis of the dengue virus. In Section 3 we discuss existing methods for dealing with misclassification, and provide our extensions thereof. We apply these methods in Section 4 and provide a discussion in Section 6.

2 | MOTIVATING EXAMPLE: DIAGNOSING DENGUE

An estimated 100 million infections of dengue occur globally each year.[22] Although dengue infection is often asymptomatic, it can also be fatal and patients can present with various clinical symptoms ranging from mild febrile illness to hemorrhagic fever, organ impairment and hypovolaemic shock.[22,23] In its early phase, dengue can be difficult to distinguish from other febrile illnesses (OFI) such as influenza, chikungunya, measles, leptospirosis, and typhoid due to the similarity of clinical symptoms, which include headache and rash. Therefore, the identification of laboratory and other clinical variables

that aid in the differential diagnosis of dengue is imperative.[22] In this motivating example we focus on the strength of the association between muscle pain and dengue versus OFI, conditional on the presence of joint pain.

To assess the added diagnostic value of muscle pain in the differential diagnosis of dengue versus OFI, a multivariable logistic prediction model can be developed. Suppose several studies have fit such models to data and that for some studies the presence of muscle pain data have been tainted by misclassification. In order to show the potential impact of the misclassification, we use simulated IPD for 10 studies (Figure 1), that are based on real data gathered in three cross-sectional studies of the IDAMS consortium (see Supporting Information A in Data S1) that aimed to improve the differential diagnosis of dengue.[22] The IPD were generated according to three scenarios with varying heterogeneity in the outcome model.

In the first scenario we defined the heterogeneity parameters such that all studies have the same true prevalence of the outcome (dengue infection) conditional on the exposure (muscle pain) and covariate (joint pain) and the same true exposure-outcome association, conditional on the covariate joint pain. In the second scenario we allowed for heterogeneity in the true prevalence of dengue conditional on the exposure and covariate but not in the true exposure-outcome association, conditional on the covariate. In the third scenario we allowed for the presence of heterogeneity in both the true prevalence of dengue conditional on the exposure and covariate as well as the true exposure-outcome association of muscle pain, conditional on the covariate. This third

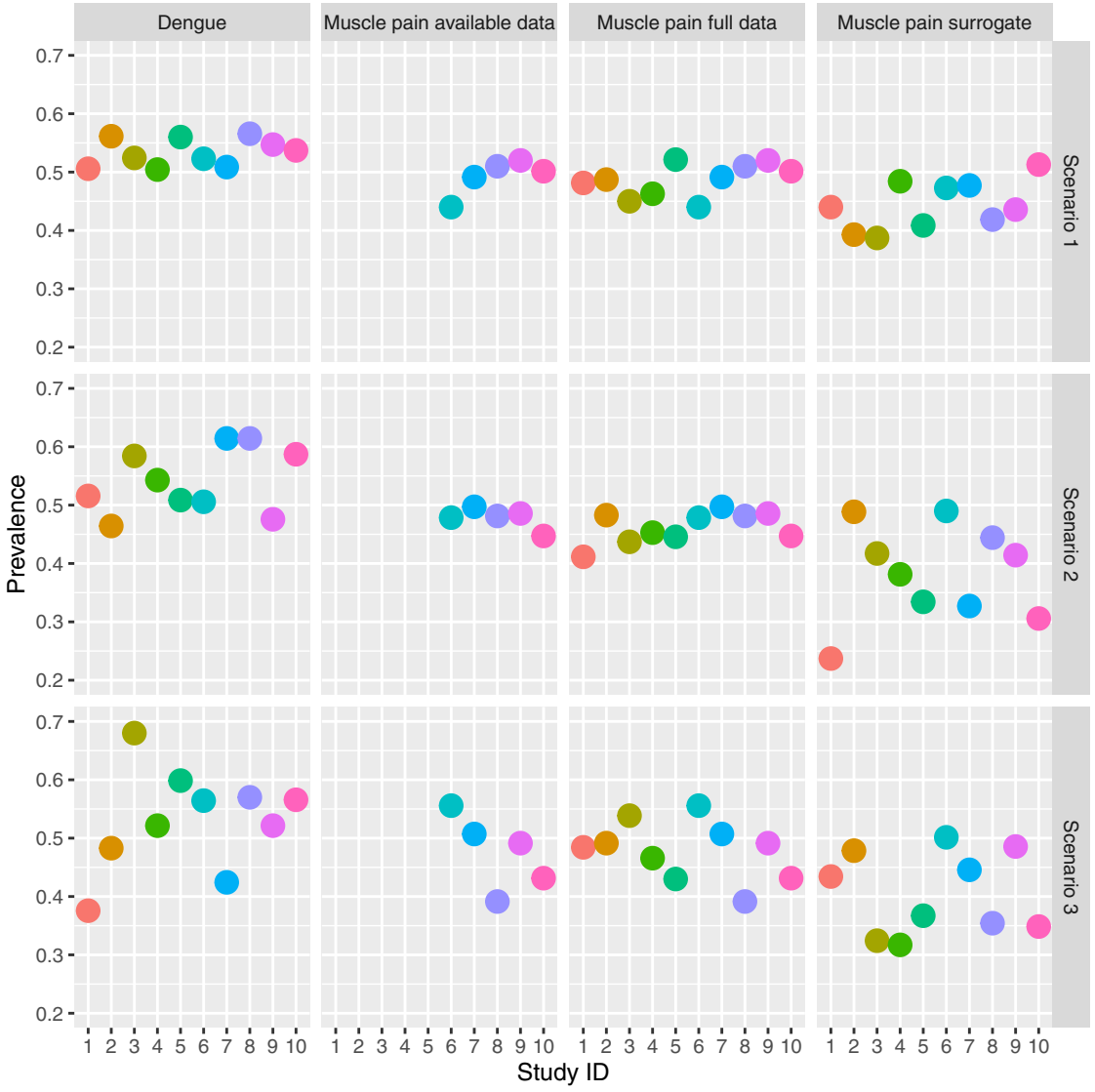

FIGURE 1 Prevalence of dengue and muscle pain measurements in the motivating example. Muscle pain was not observed in Studies 1 to 5. The values for Studies 1 to 5 under "Muscle pain full data" indicate the true values of the prevalence that were not observed [Colour figure can be viewed at wileyonlinelibrary.com]

scenario resembles the real-world scenario where the diagnosis of dengue is made more difficult by the possible presence of chikungunya, which is also a febrile illness.

As chikungunya is associated with (often even worse) muscle pain,[24] the association between muscle pain and dengue may be smaller (or entirely absent) in studies where chikungunya is prevalent, compared with studies where it is not present. In all scenarios, we allowed the true prevalence of muscle pain and the true probability of misclassification to vary across studies. The challenge is to account for this rate of misclassification that is heterogeneous across studies and depends on patient covariates, while simultaneously accounting for heterogeneity in the prevalence of dengue, and heterogeneity in the muscle pain-dengue association. In the following sections we first provide a short overview of methods for accounting for misclassification in single studies and in AD-MA before we move on to accounting for these sources of heterogeneity in IPD, such in this IPD-MA of the muscle pain-dengue association.

3 | METHODS

Many methods have been developed to adjust for misclassification of exposures in the analysis of a single study. These include regression calibration and multiple imputation-based methods. Methods for adjusting meta-analyses of aggregate data for misclassification have also been proposed. We start by briefly summarizing these methods and their characteristics. More detailed information is available from Keogh et al.[25]

3.1 | Adjusting for misclassification in a single study

In regression calibration, the outcome is regressed on the expected value of the exposure, given the surrogate exposure and covariates. The expected value of the exposure can be estimated by regressing the exposure on the surrogate exposure and covariates for participants for whom all these variables have been measured. When modeling a continuous outcome with linear regression using this approach the estimator for the exposure-outcome association is unbiased, provided that the error is nondifferential, the ME is constant across studies or validation data are available for each study where the outcome is measured, and no other biases (such as confounding or selection bias) are present.[6] However, regression calibration has been demonstrated to yield (somewhat) biased estimators when applied to logistic regression.[6,26,25] As regression calibration does not use the observed outcome for estimating the expected value of the exposure, it cannot account for differential misclassification.

Alternatively, one may apply multiple imputation for measurement error (MIME), which treats the gold standard (and the surrogate measurement) as just another variable to be imputed using all other variables. MIME models typically include the outcome as covariate, which naturally accounts for differential error if the imputation model is correctly specified. However, it overestimates the uncertainty in the imputation of the true exposure,[26] and has not been investigated for IPD. Before discussing IPD, we turn to methods for adjusting for misclassification in meta-analysis of contingency tables and aggregate data.

3.2 | Adjustment for misclassification in a meta-analysis of contingency tables

Most meta-analyses are based on aggregate data. When the exposures are binary, the aggregate data for the exposure-outcome associations are often presented as counts in contingency tables. Provided that contingency tables for the surrogate-gold standard exposure association are also available, one can adjust for the misclassification in the surrogate exposure-outcome association that is unadjusted for covariates.[21]

3.2.1 | Misclassification assumptions in meta-analysis

As the rate of misclassification may differ across studies, Lian et al. recently developed a model that accounts for clustering and heterogeneity. They relaxed model assumptions such that the frequency of misclassification is not required to be constant across studies but is allowed to vary across studies by applying a random effect.[21] That is, the degree of misclassification is allowed to vary across studies by applying a random effect. The resulting coefficients for the misclassification model and for the exposure-outcome model need not come from the same studies if this can be assumed. This is advantageous, as it implies that studies in which misclassification was not investigated can be included in the analysis.

Although the model of Lian et al. does not assume that misclassification in the measured exposure is common across studies, their model's assumptions nevertheless require that misclassification is independent of any patient-level covariates, given the value of the gold standard measurement of the exposure.[21] In particular, they assume that misclassification depends solely on study-level variables. This is an important distinction, as misclassification that is non-differential given covariates, may be differential when these covariates are not taken into account.[7] Thus, if misclassification rates are different for the levels of the outcome

and patient-level covariates can explain those differences, then these covariates must be taken into account.

3.3 | Adjustment for misclassification in AD-MA

Extending methods that rely on stratified contingency tables to the analysis of covariate-adjusted exposure-outcome associations may be impractical. It would require that studies provide contingency tables that are stratified for the outcome, gold standard measurement of the exposure, surrogate exposure, and every adjustment variable. Clearly, this may be infeasible for a large number of variables. Alternatively, one may opt to adjust for misclassification in a meta-analysis of aggregate data, that is, using exposure-outcome associations (and standard errors) reported in the form of regression coefficients such as (log) risk or odds ratios that have been adjusted for covariates. If all of these reported estimates (including the standard errors) are appropriately adjusted for misclassification in their respective studies, one could analyze these with traditional meta-analysis methods. On the other hand, if the estimation of these covariate adjusted exposure-outcome associations did not include accounting for misclassification, then this would have to occur in the meta-analysis.

If IPD are available for the gold standard and surrogate measurements of the exposure, one might apply a misclassification model to adjust the reported exposure-outcome associations for misclassification, but this would require misclassification to be dependent solely on study level variables.[21] This assumption would clearly be violated in case the misclassification is dependent on participant-level covariates. For instance, in our motivating example, the misclassification of muscle pain was associated with the participant-specific value of joint pain. If the measurement for joint pain is missing for a participant, then the information to estimate the expected value of the missing measurement of muscle pain is missing for that participant. In the case of AD-MA, this implies that the covariate joint pain would be missing for the entire study. Thus, any participant-specific misclassification would not be accounted for. In the next section we describe how the assumption that misclassification in meta-analysis depends on solely on study-level variables can be relaxed if IPD are available.

3.4 | Adjustment for misclassification in a meta-analysis of individual participant data

We extend the methods of Nelson et al.[20] and Lian et al.[21] to incorporate participant-level covariates in a one-stage IPD-MA for potentially misclassified binary exposures. As such, we allow the probability of misclassification to depend on study-level variables and on individual participant-level covariates that are observed without error. Further, modeling of IPD allows us to estimate the adjusted (i.e., multivariable) exposure-outcome associations. For example, suppose that misclassification of muscle pain may occur in the differential diagnosis of dengue.

Let x_{ij} denote the gold standard measurement of the binary exposure (e.g. muscle pain) for participant $i, i = 1, ..., n_j$ in study j, $j = 1, ..., J$. The surrogate exposure is given as x^*_{ij} and represents a possibly misclassified measurement of the exposure. We assume that x^*_{ij} and x_{ij} have been observed for some participants in some studies, and that for some participants in some studies both have been observed. Further, we assume that z_{ij} is a covariate (e.g., joint pain) without ME and that y_{ij} is a binary outcome (e.g., dengue).

Following the approach described by Richardson and Gilks,[27] we specify three submodels to account for misclassification: a measurement submodel, an exposure submodel and an outcome submodel. In the measurement submodel, the surrogate exposure (i.e., the measurement of the exposure that is prone to misclassification) is predicted, conditional on the latent gold standard measurement of the exposure, to determine the extent of misclassification. The measurement submodel models the relation $x^*_{ij} \sim f(x_{ij}, z_{ij})$ (i.e., it imposes parametric assumptions on the distribution of x^*_{ij}). In the exposure submodel, the latent gold standard measurement of the exposure is regressed on covariates that are measured without error, in order to predict the gold standard measurement of the exposure in participants for whom it is missing. Hence, the exposure submodel models the relation $x_{ij} \sim f(z_{ij})$. In the outcome submodel, the outcome is regressed on the latent gold standard measurement of the exposure and on covariates that are measured without error, to determine the exposure-outcome relationship. The outcome submodel models the relation $y_{ij} \sim f(x_{ij}, z_{ij})$. Although our model generalizes to multiple covariates, we restrict our notation to a single covariate for simplicity. We first consider non-differential misclassification models, which assume that y_{ij} is independent of $x*_{ij}$, conditional on x_{ij} and z_{ij}.

3.4.1 | Common effects IPD-MA

We start with describing an IPD-MA misclassification model containing three submodels that assumes common effects across studies. Hence, all data are analyzed as if they were measured in a single study. In this first model, the probability of misclassification only depends on the

value of the gold standard measurement of the exposure. The measurement (sub)model is then given by:

$$\begin{aligned} x_{ij}^{*}|x_{ij} &\sim \text{Bernoulli}\left(p_{ij}^{*}\right), \\ g\left(p_{ij}^{*}\right) &= \lambda_{00}x_{ij} + \phi_{00}\left(1-x_{ij}\right) \end{aligned}, \quad (1)$$

where we choose the following prior distributions for the coefficients: $\lambda_{00} \sim N\left(0, \sigma^2_{\lambda_{00}}\right), \phi_{00} \sim N\left(0, \sigma^2_{\phi_{00}}\right)$ and $g(.)$ is a link function. For instance, one could choose the logit for $g(.)$, such that intercept parameters represent log odds and (exposure) coefficient parameters represent log odds ratios. For coefficients, we use the first subscript to denote the variable it is associated with (0: intercept, 1: z_{ij}, 2: x_{ij}), and the second to denote the level of the effect (*0*: fixed effect, *j*: random effect). This is equivalent to a measurement submodel proposed by Nelson et al.,[20] as λ_{00} and ϕ_{00} are parameters that determine $g(\text{sensitivity})$ and $g(1-\text{specificity})$, respectively. The above parametrization allows us to introduce covariates to the measurement submodel in subsequent steps. We leave the variance parameters unspecified, as fixed values may be supplied for these, though one may also supply prior distributions for the variance parameters.

The exposure submodel aims to estimate the relationship between the gold standard measurement of the exposure and covariate(s). It is simultaneously applied to predict the probability that the exposure is present in participants for whom the gold standard measurement of the exposure status is missing. For participants for whom the gold standard measurement of the exposure status is missing, the expected value given covariates is imputed following this submodel. We note that misclassification models cannot restore the true value of the gold standard, but can account for missing values of the gold standard. This submodel is given by:

$$\begin{aligned} x_{ij}|z_{ij} &\sim \text{Bernoulli}\left(p_{ij}\right), \\ g\left(p_{ij}\right) &= \gamma_{00} + \gamma_{10}z_{ij}, \end{aligned} \quad (2)$$

where we have chosen the priors for the coefficients as $\gamma_{00} \sim N\left(0, \sigma^2_{\gamma_{00}}\right)$ and $\gamma_{10} \sim N\left(0, \sigma^2_{\gamma_{10}}\right)$. Thirdly, of course, we describe the submodel that is designed to assess the (adjusted) exposure-outcome association. This outcome submodel is given by:

$$\begin{aligned} y_{ij}|x_{ij}, z_{ij} &\sim \text{Bernoulli}\left(\pi_{ij}\right), \\ g\left(\pi_{ij}\right) &= \beta_{00} + \beta_{10}z_{ij} + \beta_{20}x_{ij}, \end{aligned} \quad (3)$$

where $\beta_{00} \sim N\left(0, \sigma^2_{\beta_{00}}\right), \beta_{10} \sim N\left(0, \sigma^2_{\beta_{10}}\right), \beta_{20} \sim N\left(0, \sigma^2_{\beta_{20}}\right)$, β_{00} is an intercept, β_{10} is the coefficient for the covariate, and β_{20} is the coefficient (log odds ratio) for the exposure of interest. When an adjusted exposure-outcome association is to be estimated, such as when adjustments for confounding need to be made, the confounder should be included in this submodel, which is represented by z_{ij}. In practice, one may adjust for multiple confounders in this misclassification model, but here we adjust for a single variable for simplicity. Equations (1), (2), and (3) together make up the least complex misclassification model that we consider here and are illustrated in Figure 2. The posterior distribution of this model is given by the product of the likelihoods of the three submodels, and the prior distributions of the three submodels:

$$\begin{aligned} &p(\lambda_{00}, \phi_{00})p(\gamma_{00}, \gamma_{10})p(\beta_{00}, \beta_{10}, \beta_{20}) \prod_j \prod_i \\ &p\left(x_{ij}^{*}|x_{ij}, \lambda_{00}, \phi_{00}\right) \prod_j \prod_i p(x_{ij}|z_{ij}, \gamma_{00}, \gamma_{10}) \prod_j \prod_i \\ &p\left(y_{ij}|x_{ij}, z_{ij}, \beta_{00}, \beta_{10}, \beta_{20}\right) \end{aligned}$$

Although the implementation of aforementioned misclassification models is fairly straightforward in an IPD-MA, their justification becomes problematic when studies differ with respect to case-mix, baseline risk, exposure-outcome associations, or the extent of misclassification. We, therefore, discuss how to adjust the submodels accordingly.

3.4.2 | Accounting for between-study heterogeneity in the distribution of the exposure

A common situation in IPD-MA is the presence of heterogeneity in case-mix distributions.[18] In particular, when the distribution of the gold standard measurement of the exposure variable varies across studies and the exposure submodel does not account for this, then inadequate predictions will be made for the unobserved gold standard measurements. We may model the varying prevalence of the gold standard measurement of the exposure x by applying random intercepts to the exposure submodel, replacing Equation (2) with (note that we omit the random parameters from the conditional notation):

$$\begin{aligned} x_{ij}|z_{ij} &\sim \text{Bernoulli}\left(p_{ij}\right), \\ g\left(p_{ij}\right) &= \gamma_{00} + \gamma_{0j} + \gamma_{10}z_{ij}, \\ \text{with random intercepts}&: \\ \gamma_{0j} &\sim N\left(0, \tau^2_{\gamma_{0j}}\right), \end{aligned} \quad (4)$$

where we choose the following priors for the coefficients: $\gamma_{00} \sim N\left(0, \sigma^2_{\gamma_{00}}\right)$, and $\gamma_{10} \sim N\left(0, \sigma^2_{\gamma_{10}}\right)$. Whereas it is

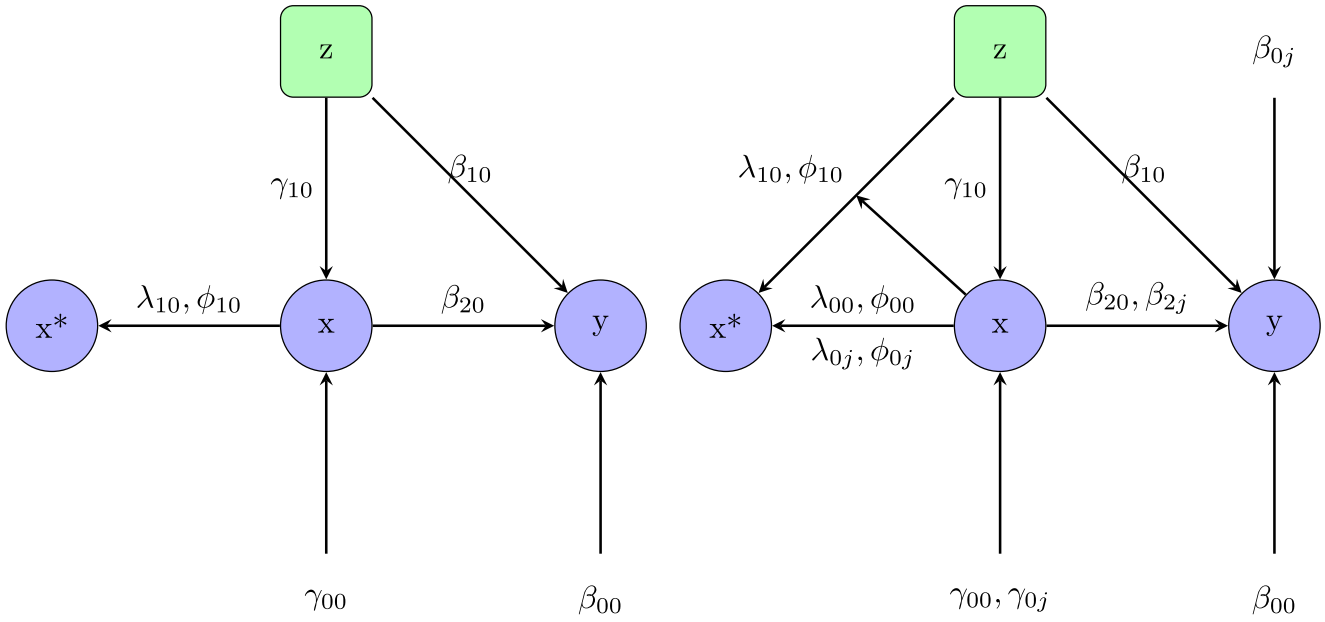

FIGURE 2 Diagrams of model Equations (1), (2), and (3) (left) and (4), (6), and (8) (right). Green squares: fully observed data, blue circles: at least partially observed data, not in boxes: parameters. Variance parameters omitted. Note that in our examples and simulation, x^* and y are fully observed [Colour figure can be viewed at wileyonlinelibrary.com]

common to assume a Normal prior distribution for regression coefficients and the intercept,[12] the choice for a prior distribution for the variance parameters is less straightforward. A prior with too heavy tails will give too much prior weight on high variance, whereas a prior with thin tails will put too much prior weight on a low variance.[28] We here consider a half-Normal (i.e., the positive half) distribution for parameters for heterogeneity between studies, namely $\tau^2_{\gamma_{0j}} \sim \text{half}-N\left(0, \xi_{\gamma_{0j}}\right)$ but would like to highlight that several alternatives have been proposed, such as the half-Cauchy, half-t, and inverse-gamma distributions.[28–30] This submodel may further be expanded by adding random effects for the covariates as well, which we omit here for brevity.

3.4.3 | Accounting for between-study heterogeneity in misclassification

For various reasons, the extent of error in the measurement of the exposure may vary by study in an IPD-MA. This may be modeled by applying random intercepts in the measurement submodel, which can be interpreted as that the log-odds sensitivity and 1−specificity vary by study. The measurement submodel is then given by:

$$\begin{aligned} & x^*_{ij}|x_{ij} \sim \text{Bernoulli}\left(p^*_{ij}\right), \\ & g\left(p^*_{ij}\right) = \left(\lambda_{00}+\lambda_{0j}\right)x_{ij} + \left(\phi_{00}+\phi_{0j}\right)\left(1-x_{ij}\right), \\ & \text{with random intercepts}: \\ & \quad \lambda_{0j} \sim N\left(0, \tau^2_{\lambda_{0j}}\right), \\ & \quad \phi_{0j} \sim N\left(0, \tau^2_{\phi_{0j}}\right), \end{aligned} \tag{5}$$

where we choose the following priors for the coefficients: $\lambda_{00} \sim N\left(0, \sigma^2_{\lambda_{00}}\right), \phi_{00} \sim N\left(0, \sigma^2_{\phi_{00}}\right)$, and the following priors for the heterogeneity parameters: $\tau^2_{\lambda_{0j}} \sim \text{half}-N\left(0, \xi_{\lambda_{0j}}\right)$ and $\tau^2_{\phi_{0j}} \sim \text{half}-N\left(0, \xi_{\phi_{0j}}\right)$.

3.4.4 | Adjusting for participant-specific misclassification

A more complex situation arises when misclassification is related to participant-level covariates. For instance, recall of exposure values may be poorer in the elderly, the answering of questionnaires may be hampered by poor literacy, and measurement instruments might be designed for specific subgroups of participants. Participant-specific misclassification is particularly problematic if the case-mix distributions vary across studies, as estimates of exposure-outcome associations will then be affected differently across

studies. For this reason, the presence of such error can be accounted for by incorporating patient-level covariate effects in the measurement submodel:

$$\begin{aligned} x^*_{ij}|x_{ij},z_{ij} &\sim \text{Bernoulli}\left(p^*_{ij}\right),\\ g\left(p^*_{ij}\right) &= (\lambda_{00}+\lambda_{0j}+\lambda_1 z_{ij})x_{ij}+\left(\phi_{00}+\phi_{0j}+\phi_1 z_{ij}\right)(1-x_{ij}),\\ \text{with random intercepts}&:\\ \lambda_{0j} &\sim N\left(0,\tau^2_{\lambda_{0j}}\right),\\ \phi_{0j} &\sim N\left(0,\tau^2_{\phi_{0j}}\right), \end{aligned} \tag{6}$$

where we choose the following priors for the coefficients: $\lambda_{00}\sim N\left(0,\sigma^2_{\lambda_{00}}\right), \lambda_{10}\sim N\left(0,\sigma^2_{\lambda_{10}}\right), \phi_{00}\sim N\left(0,\sigma^2_{\phi_{00}}\right)$ and $\phi_{10}\sim N\left(0,\sigma^2_{\phi_{10}}\right)$, and the following priors for the heterogeneity parameters: $\tau^2_{\lambda_{0j}}\sim \text{half}-N\left(0,\xi_{\lambda_{0j}}\right)$ and $\tau^2_{\phi_{0j}}\sim \text{half}-N\left(0,\xi_{\phi_{0j}}\right)$.

3.4.5 | Accounting for between-study heterogeneity in outcome frequency

Commonly, in data from an IPD-MA and other clustered data sets the frequency of the outcome varies by study. To account for this effect of clustering within studies, it is generally considered vital that random intercepts for the outcome are applied in an IPD-MA.[18] We can add these to the outcome submodel as follows:

$$\begin{aligned} y_{ij}|x_{ij},z_{ij} &\sim \text{Bernoulli}(\pi_{ij}),\\ g(\pi_{ij}) &= \beta_{00}+\beta_{0j}+\beta_{10}z_{ij}+\beta_{20}x_{ij},\\ \text{with random intercepts}&:\\ \beta_{0j} &\sim N\left(0,\tau^2_{\beta_{0j}}\right), \end{aligned} \tag{7}$$

where we choose the following priors for the coefficients: $\beta_{00}\sim N\left(0,\sigma^2_{\beta_{00}}\right), \beta_{20}\sim N\left(0,\sigma^2_{\beta_{20}}\right), \beta_{10}\sim N\left(0,\sigma^2_{\beta_{10}}\right)$, and $\tau^2_{\beta_{0j}}\sim \text{half}-N\left(0,\xi_{\beta_{0j}}\right)$.

3.4.6 | Accounting for between-study heterogeneity in exposure-outcome associations

Further, the strength of the true exposure-outcome association might also vary by study. To model this, one may adopt a random effects model for the outcome, which does not assume there is a single exposure-outcome association.[31] Instead, it assumes there is a distribution of exposure-outcome associations and it estimates the center and variance of that distribution.

$$\begin{aligned} y_{ij}|x_{ij},z_{ij} &\sim \text{Bernoulli}(\pi_{ij}),\\ g(\pi_{ij}) &= \beta_{00}+\beta_{0j}+\beta_{10}z_{ij}+\beta_{20}x_{ij}+\beta_{2j}x_{ij},\\ \text{with random effects}&:\\ \beta_{0j} &\sim N\left(0,\tau^2_{\beta_{0j}}\right),\\ \beta_{2j} &\sim N\left(0,\tau^2_{\beta_{2j}}\right), \end{aligned} \tag{8}$$

where we choose the following priors for the coefficients: $\beta_{00}\sim N\left(0,\sigma^2_{\beta_{00}}\right), \beta_{10}\sim N\left(0,\sigma^2_{\beta_{10}}\right), \beta_{20}\sim N\left(0,\sigma^2_{\beta_{20}}\right)$, and for the heterogeneity parameters: $\tau^2_{\beta_{0j}}\sim \text{half}-N\left(0,\xi_{\beta_{0j}}\right)$, and $\tau^2_{\beta_{2j}}\sim \text{half}-N\left(0,\xi_{\beta_{2j}}\right)$. In this model β_{20} is the center of the exposure-outcome association distribution and represents the overall association, β_{2j} is the study-specific exposure-outcome association and $\tau^2_{\beta_{2j}}$ is the heterogeneity of the exposure-outcome association across studies. The random effects assumption is commonly adopted in meta-analysis where sources of between-study heterogeneity cannot (fully) be explained using participant-specific information but need to be accounted for. It is also considered a rather safe assumption, as a random effects model will estimate the variance of the exposure-outcome association at near zero when that association does not vary in the sample. Conversely, a common effects model will lead to inadequate estimates when the common effects assumption does not hold. Equations (4), (6), and (8) together are illustrated in Figure 2.

The models considered here are identifiable only if sufficient information is present in the data.[12,32] For instance, to estimate Equations (2) and (4) requires that the gold standard measurement of the exposure x_{ij} is observed for sufficient individuals. Strictly speaking, a single (large) study where the gold standard and surrogate measurements have been observed should be sufficient to estimate the participant-level effects, though more studies would be necessary to estimate the study-level effects. For instance, in our motivating example x_{ij} is available for participants in half of the included studies.

Here we have assumed that the outcome y is available for every participant in every study of the IPD-MA. Though, if unavailable, it could be imputed following Equations (3), (7), or (8). To ensure congeniality this imputation model must at least contain the exposure and covariates of the outcome submodel.[33]

3.4.7 | Accounting for differential misclassification

So far we have assumed the error in the measurement of the exposure is non-differential, that is that conditional on the gold standard measurement of the value of the exposure and on the perfectly measured covariates, the error in the measurement is unrelated to the outcome. In any other case the error is differential. An example of differential error is recall bias in a case–control (or case-referent) study, where individuals may overestimate (or underestimate) their exposure, as a result of a known outcome. A cause for this may be that the recall period for the (self-report of the) exposure differs for controls and cases, which may especially be an issue in a cross-over study[34,35] or in a case–control study.[36]

For example, in a case–control study on breast cancer, Morabia and Flandre observed that defining menopause using the cases' and controls' respective age leads to differential misclassification, as on average menopause occurs later in cases than in controls.[37] Further, Levois and Switzer report on studies on environmental tobacco smoke exposure and lung cancer. They discuss multiple studies in which the probability of misclassification of smoking or magnitude of ME in smoke exposure differed between lung cancer cases and controls.[38] Greenland recommends that in the absence of a reason to assume that misclassification is non-differential, a method that accounts for differential misclassification would be preferred.[39] For example, differential misclassification would be unlikely in a cohort study where data is collected prospectively,[39] though even in this case it is not ensured that the misclassification is non-differential.[40] The methods we described can be extended to allow for differential misclassification, by replacing Equation (6) with:

$$\begin{aligned} x^*_{ij}|x_{ij},y_{ij},z_{ij} &\sim \text{Bernoulli}\left(p^*_{ij}\right),\\ g\left(p^*_{ij}\right) &= \left(\lambda_{00}+\lambda_{0j}+\lambda_{10}z_{ij}+\lambda_{20}y_{ij}\right)x_{ij}\\ &\quad+\left(\phi_{00}+\phi_{0j}+\phi_{10}z_{ij}+\phi_{20}y_{ij}\right)\left(1-x_{ij}\right) \qquad (9)\\ \text{with random intercepts}&:\\ \lambda_{0j} &\sim N\left(0,\tau^2_{\lambda_{0j}}\right),\\ \phi_{0j} &\sim N\left(0,\tau^2_{\phi_{0j}}\right), \end{aligned}$$

where we choose the following priors for the coefficients: $\lambda_{00}\sim N\left(0,\sigma^2_{\lambda_{00}}\right),\lambda_{10}\sim N\left(0,\sigma^2_{\lambda_{10}}\right),\lambda_{20}\sim N\left(0,\sigma^2_{\lambda_{20}}\right),\phi_{00}\sim N\left(0,\sigma^2_{\phi_{00}}\right),\phi_{10}\sim N\left(0,\sigma^2_{\phi_{10}}\right),\phi_{20}\sim N\left(0,\sigma^2_{\phi_{20}}\right)$, and for the heterogeneity parameters we choose: $\tau^2_{\lambda_{0j}}\sim \text{half}-N\left(0,\xi_{\lambda_{0j}}\right)$ and $\tau^2_{\phi_{0j}}\sim \text{half}-N\left(0,\xi_{\phi_{0j}}\right)$. It might seem at first that including the outcome y_{ij} in the measurement submodel leads to a circular model formulation. However, the outcome y_{ij} is used to predict the value of x^*_{ij}, not x_{ij}. As x^*_{ij} is not present in the outcome submodel, the model is not circular. This model is illustrated in Figure 3. It bears much resemblance to (Bayesian) MI. The difference is that in MI no measurement submodel is specified and the surrogate measurement instead appears on the right-hand side of the exposure submodel. That is, in the MI approach the surrogate is treated as just another variable, whereas in our approach it is treated as a surrogate of the gold standard. We provide a description of a stratified differential misclassification model in Supporting Information B in Data S1.

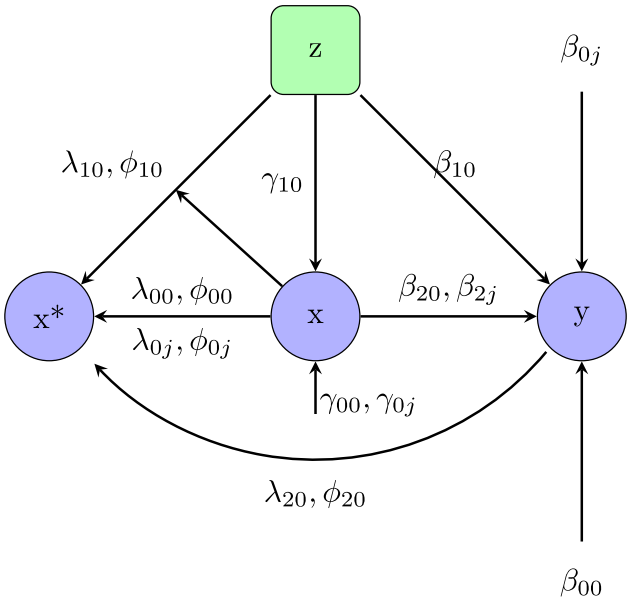

FIGURE 3 Diagrams of model Equations (4), (9), and (8). Green squares: fully observed data, blue circles: at least partially observed data, not in boxes: parameters. Variance parameters omitted. Note that in our examples and simulation, x^* and y are fully observed. [Colour figure can be viewed at wileyonlinelibrary.com]

We have implemented our methodology in the misclass R package, which is available on Github (github.com/VMTdeJong/misclass).

4 | MOTIVATING EXAMPLE: APPLICATION OF METHODS TO DENGUE IPD-MA

To illustrate the impact of misclassification on observed exposure-outcome associations in an IPD-MA, we apply

several modeling strategies to estimate the association between muscle pain (x_{ij}) and dengue (y_{ij}) in patients suspected of dengue, where muscle pain is possibly misclassified (x^*_{ij}). Hereto, we generated three scenarios for a dengue IPD-MA using real data on dengue as described in Section 2. In all scenarios we allowed the true prevalence of muscle pain and the true probability of misclassification to vary across studies. In the first scenario we defined the heterogeneity parameters such that all studies have the same (true) prevalence of dengue conditional on the exposure and the covariate (z_{ij}) and the same (true) exposure-outcome association of muscle pain, conditional on the covariate. In the second scenario we allowed for heterogeneity in the true prevalence of dengue conditional on the exposure and the covariate but not in the true exposure-outcome association, conditional on the covariate. In the third scenario we allowed for the presence of heterogeneity in both the true prevalence of dengue conditional on the exposure and covariate as well as the true exposure-outcome association of muscle pain, conditional on the covariate.

We aim to highlight the ability of the methodology we have presented here to restore this association and its uncertainty, while simultaneously accounting for the clustering of participants within studies and allowing for heterogeneity in the muscle pain-dengue association.

4.1 | Methods

We apply 11 Bayesian binary logistic modeling strategies to estimate the muscle pain-dengue association and its heterogeneity across studies. First, we model the full data with a mixed effects model as if the gold standard measurement was observed for all participants in all studies. In reality, this would not be possible as the gold standard would not be observed for some participants, but here it serves as a reference for comparison with the models that are restricted to the observed data. Second, we apply a mixed effects model on the subset of the data for which the gold standard measurement of the exposure was observed, that is, we apply a so-called complete case analysis. Third, we apply a naive mixed effects modeling strategy, in which we take the surrogate measurement as a proxy for any participant for whom the gold standard measurement is not observed. Finally, we apply the eight models described in Section 3.4. These models range from not accounting for heterogeneity and accounting for the simplest form of misclassification to accounting for heterogeneity in all submodels and for a differing extent and nature of misclassification. Although many more combinations of the submodels exist, for brevity we chose to apply them in the order as outlined, which results in eight full models for accounting for misclassification. We note that some alternative specifications would not be sensible, as the exposure submodel needs to contain at least the variables that are included in the outcome submodel. We use prior distributions as described above, except that we applied inverse-gamma distributions for the parameters for heterogeneity across studies. We chose a value of 0.1 for each of the σ parameters for the prior precision of the Normal distributions of the coefficients. We chose a value of 0.001 for the shape and rate parameters of the inverse gamma distributions for the variances of the random effects and random intercepts.

We estimated all the models with a Gibbs sampler with two independent chains. After 1000 adaptation and 1000 warm-up samples, 25,000 samples for the estimation of the parameters were performed in each chain. To reduce autocorrelation, we thinned the samples by a factor 5. The presented estimates are based on the remaining 2×5000 samples. The code and data for our motivating example is available on Github (github.com/VMTdeJong/Misclassification-Dengue).

4.2 | Results

In each of the scenarios (see Section 2), all models yielded positive estimates with 95% credibility intervals that excluded zero, which in each case may lead to the conclusion that muscle pain is positively associated with dengue. However, we observed considerable differences between the point estimates and estimated 95% credibility intervals of the different models, especially for the common muscle pain-dengue association.

4.2.1 | Scenario 1: homogeneous conditional baseline prevalence and exposure-outcome associations across studies

In the first scenario, the estimated association (log-odds ratio) between muscle pain and dengue in the full data was 0.82 (95% CI: 0.67: 0.98, Table 1). The complete case analysis (0.64, 95% Credibility Interval: 0.41: 0.87) and especially the naive analysis (0.47, 95% CI: 0.34: 0.60) underestimated this association. The misclassification methods were able to restore the muscle pain-dengue association to various degrees. The model entitled Adjusting for participant-specific misclassification (comprising Equations (6), (4), and (3)), which was the correctly specified model, estimated the log odds ratio for the association at 0.72 (95% CI: 0.54: 0.90). Surprisingly, the underspecified misclassification models estimated the association with similar or even less error. The overspecified (i.e., models with excess parameters) misclassification errors estimated the

TABLE 1 Multivariable log odds ratio and heterogeneity estimates (95% Credibility Interval) for the presence of muscle pain for diagnosing dengue in Scenario 1

Model	β_{20} (95% CI)	$\tau_{\beta_{2j}}$ (95% CI)
Full data (reference)	0.82 (0.67: 0.98)	0.05 (0.02: 0.14)
Complete cases	0.64 (0.41: 0.87)	0.06 (0.02: 0.23)
Naive	0.47 (0.34: 0.60)	0.06 (0.02: 0.16)
Misclassification models		
Common effects	0.74 (0.55: 0.93)	
Accounting for between-study heterogeneity in the distribution of the exposure	0.72 (0.53: 0.91)	
Accounting for between-study heterogeneity in misclassification	0.75 (0.56: 0.93)	
Adjusting for participant-specific misclassification	0.72 (0.54: 0.90)	
Accounting for between-study heterogeneity in outcome frequency	0.71 (0.54: 0.90)	
Accounting for between-study heterogeneity in exposure-outcome associations	0.71 (0.53: 0.91)	0.05 (0.02: 0.16)
Accounting for differential misclassification	0.70 (0.52: 0.90)	0.05 (0.02: 0.16)
Accounting for stratified differential misclassification	0.66 (0.45: 0.88)	0.05 (0.02: 0.15)

Note: The center of the distribution was estimated by the median of the posterior distribution. Empty cells for $\tau_{\beta_{2j}}$ (95% CI) indicate it is assumed to equal zero in the respective model. The names for the misclassification models refer to the respective sections in the main text.

TABLE 2 Multivariable log odds ratio and heterogeneity estimates (95% Credibility Interval) for the presence of muscle pain for diagnosing dengue in Scenario 2

Model	β_{20} (95% CI)	$\tau_{\beta_{2j}}$ (95% CI)
Full data (reference)	0.76 (0.61: 0.92)	0.07 (0.02: 0.22)
Complete cases	0.66 (0.42: 0.89)	0.08 (0.02: 0.33)
Naive	0.56 (0.42: 0.70)	0.06 (0.02: 0.19)
Misclassification models		
Common effects	0.75 (0.58: 0.92)	
Accounting for between-study heterogeneity in the distribution of the exposure	0.69 (0.53: 0.86)	
Accounting for between-study heterogeneity in misclassification	0.76 (0.59: 0.93)	
Adjusting for participant-specific misclassification	0.73 (0.57: 0.90)	
Accounting for between-study heterogeneity in outcome frequency	0.74 (0.58: 0.91)	
Accounting for between-study heterogeneity in exposure-outcome associations	0.75 (0.57: 0.94)	0.09 (0.02: 0.27)
Accounting for differential misclassification	0.72 (0.54: 0.91)	0.08 (0.02: 0.25)
Accounting for stratified differential misclassification	0.67 (0.48: 0.88)	0.07 (0.02: 0.23)

Note: The center of the distribution was estimated by the median of the posterior distribution. Empty cells for $\tau_{\beta_{2j}}$ (95% CI) indicate it is assumed to equal zero in the respective model. The names for the misclassification models refer to the respective sections in the main text.

association with a larger error, though the errors were still smaller than the naive and complete case analyses.

All misclassification models estimated the between-study heterogeneity of the muscle pain-dengue association well, as the estimates were very similar to the reference estimate of 0.05 (95% CI: 0.02: 0.14) in the full data. Note that all models overestimated the between-study heterogeneity of the muscle pain-dengue association, as the true values were equal to 0 in this scenario. This may be a result of the influence of the prior distributions for the heterogeneity parameter $\tau_{\beta_{2j}}$ or due to some heterogeneity existing in the sample. The 95% CI of $\tau_{\beta_{2j}}$ in the complete case analysis was wider (0.02: 0.23) than the 95% CI for the other models. This is unsurprising as it uses only a subset of the available data.

4.2.2 | Scenario 2: heterogeneous baseline prevalence across studies

In this second scenario, the estimated association (log-odds ratio) between muscle pain and dengue in the full

TABLE 3 Multivariable log odds ratio and heterogeneity estimates (95% Credibility Interval) for the presence of muscle pain for diagnosing dengue in Scenario 3

Model	β_{20} (95% CI)	$\tau_{\beta_{2j}}$ (95% CI)
Full data (reference)	0.87 (0.60: 1.14)	0.32 (0.18: 0.61)
Complete cases	1.02 (0.67: 1.38)	0.23 (0.05: 0.73)
Naive	0.60 (0.31: 0.89)	0.37 (0.21: 0.69)
Misclassification models		
Common effects	0.81 (0.63: 0.99)	
Accounting for between-study heterogeneity in the distribution of the exposure	0.58 (0.41: 0.75)	
Accounting for between-study heterogeneity in misclassification	1.09 (0.92: 1.28)	
Adjusting for participant-specific misclassification	1.04 (0.88: 1.22)	
Accounting for between-study heterogeneity in outcome frequency	0.97 (0.73: 1.20)	
Accounting for between-study heterogeneity in exposure-outcome associations	0.79 (0.48: 1.11)	0.35 (0.19: 0.67)
Accounting for differential misclassification	0.80 (0.48: 1.10)	0.35 (0.18: 0.68)
Accounting for stratified differential misclassification	0.82 (0.48: 1.14)	0.34 (0.17: 0.67)

Note: The center of the distribution was estimated by the median of the posterior distribution. Empty cells for $\tau_{\beta_{2j}}$ (95% CI) indicate it is assumed to equal zero in the respective model. The names for the misclassification models refer to the respective sections in the main text.

data was 0.76 (95% CI: 0.61: 0.92, Table 2). Again, the complete case analysis (0.66, 95% CI: 0.42: 0.89) and naive analysis (0.56, 95% CI: 0.42: 0.70) underestimated this association. The misclassification models all estimated the common muscle pain-dengue association with less error than the naive and complete case analysis. The model entitled "Accounting for between-study heterogeneity in outcome frequency" (comprising Equations (6), (4), and (7)), which was the correctly specified model, estimated the association at 0.74 (95% CI: 0.58: 0.91), which was nearly identical to the estimates by the analysis on the full data. In addition, all misclassification models had narrower 95% Credibility Intervals than the complete case analysis.

All considered models estimated the (lack of) between-study heterogeneity in the muscle pain-dengue association adequately. In the analysis on the full data this heterogeneity was estimated at 0.07 (95% CI: 0.02: 0.22). Again, the 95% CI for the complete case analysis was the widest (95% CI: 0.02: 0.33).

4.2.3 | Scenario 3: heterogeneous baseline prevalence and exposure effects across studies)

In this final scenario, the analysis on the full data yielded a muscle pain-dengue association of 0.87 (95% CI: 0.60: 1:14), whereas the complete case analysis estimated it at 1.02 (95% CI: 0.67: 1.38, Table 3) This neatly illustrates that the error in the muscle pain-dengue association estimated by complete case analysis is caused by an increased variance rather than bias, as the estimate by the complete case analysis is now increased with respect to the analysis on the full data, whereas in the other scenarios it was underestimated. As expected, the naive analysis underestimated the association yet again, at 0.60 (95% CI: 0.31: 0.89).

Three of the misclassification models' point estimates were further away from the point estimate by the full data than the complete case analysis' point estimate, which highlights that applying a misclassification model is not guaranteed to reduce the error in the point estimate. Yet, these were all underspecified models that did not account for the various forms of heterogeneity. The correctly specified model entitled "Accounting for between-study heterogeneity in exposure-outcome associations" (comprising Equations (6), (4), and (8)) estimated the muscle pain-dengue association at 0.79 (95% CI: 0.48: 1.11), which was close to the estimate on the full data. The overspecified models yielded similar estimates.

Except for the complete case analysis, all models that estimated the between-study heterogeneity for the muscle pain-dengue association yielded adequate estimates for this variance, as compared to the reference. The complete case analysis underestimated the amount of between-study heterogeneity, whereas the underspecified misclassification models (wrongly) assumed it to be equal to 0.

4.3 | Summary

Overall, the results of this motivating example on the association between muscle pain and dengue highlight

the impact of misclassification on an exposure-outcome association. The misclassification models estimated the exposure-outcome association with less error (where the full data are taken as reference) than both the complete-case and naive approaches, with the exception for some models that were underspecified in Scenario 3. This suggests that even in these scenarios for relatively small IPD-MAs, the more complex (possibly overspecified) models seem more suitable than the simpler (possibly underspecified) models.

In general, the models provided adequate estimates of the heterogeneity of the muscle pain-dengue association. The exception was the complete case analysis, which yielded different point estimates due the fact that these estimates were based on different data and which yielded wider credibility intervals due to the fact that these interval estimates were based on less data. In conclusion, the misclassification methods that accounted for heterogeneity in the various submodels gave the best available estimates of the muscle pain-dengue association and its heterogeneity.

5 | SIMULATION STUDY

We performed a simulation study to assess the impact of misclassification on estimated associations (β_{20}) and the heterogeneity of this association ($\tau^2_{\beta_{2j}}$) in an IPD-MA and to assess the validity of our methodology. We highlight the ability of misclassification models to provide unbiased estimators of these associations while propagating the uncertainty induced by misclassification and the various forms of heterogeneity, to facilitate valid inference. We provide more details and a more extensive simulation study in Supporting information C in Data S1.

The data were simulated similar to that in Scenario 3 of the motivating example on the diagnosis of dengue: there was heterogeneity in the distribution of the exposure of interest (muscle pain), in the true prevalence of the exposure conditional on the exposure and covariate (joint pain), and in the true exposure-outcome association conditional on the covariate.

We applied three models. First, we applied complete case analysis, that is only on the participants for whom the gold standard exposure was observed. Second, we applied a naive model in which the surrogate measurement of the muscle pain was used for participants for whom the gold standard measurement was not available. Third, we applied the misclassification model we describe in section "Accounting for between-study heterogeneity in exposure-outcome associations", which is given by Equations (6), (4), and (8).

We estimated all the models with a Gibbs sampler with two independent chains using JAGS 4.3.0. We performed 1000 replications of the simulation in R 3.5.2.[41] The code for our simulation study is available on Github (github.com/VMTdeJong/Misclassification-IPDMA).

5.1 | Simulation results of the summary estimate of the exposure-outcome relation

As expected, the estimator of complete case analysis was unbiased (Figure 4, left). The estimator of the naive method that used x^* where x was not available was biased, and the misclassification model was nearly unbiased. Due to the reduced sample size for the complete case analysis, the variance of the estimates increased, which increased the RMSE (Figure 4, right). As a result, this method had the largest RMSE. The misclassification method had the lowest RMSE. The proportion of 95% credibility intervals that covered the true effect size was very high for the complete case analysis (Figure 4, middle). The estimates for the variance were frequently overestimated (not shown), perhaps as a result of the influence from the prior for the variance. The coverage rate for the misclassification model was also too high, though it was closer to nominal. The naive method had nominal coverage. In conclusion, the misclassification model provided the best available estimates of the exposure-outcome association.

5.2 | Simulation results of the heterogeneity of the exposure-outcome relation

All methods' estimators for the heterogeneity of the exposure-outcome relation were biased (τ_{β_x}), though the magnitude varied (Figure 5, left). The estimator of the complete case analysis was the most biased, and that of the naive method was similar to that of the misclassification model.

The complete case analysis the highest RMSE (Figure 5, right). The naive IPD method and the misclassification model had the lowest RMSE. For all of the methods, the 95% Credibility Intervals had below nominal coverage for estimating the heterogeneity of the exposure-outcome relation (Figure 5, middle).

6 | DISCUSSION

As ME or misclassification may cause bias in estimated exposure-outcome associations, standard errors and between-study heterogeneity in IPD-MA, it is essential to

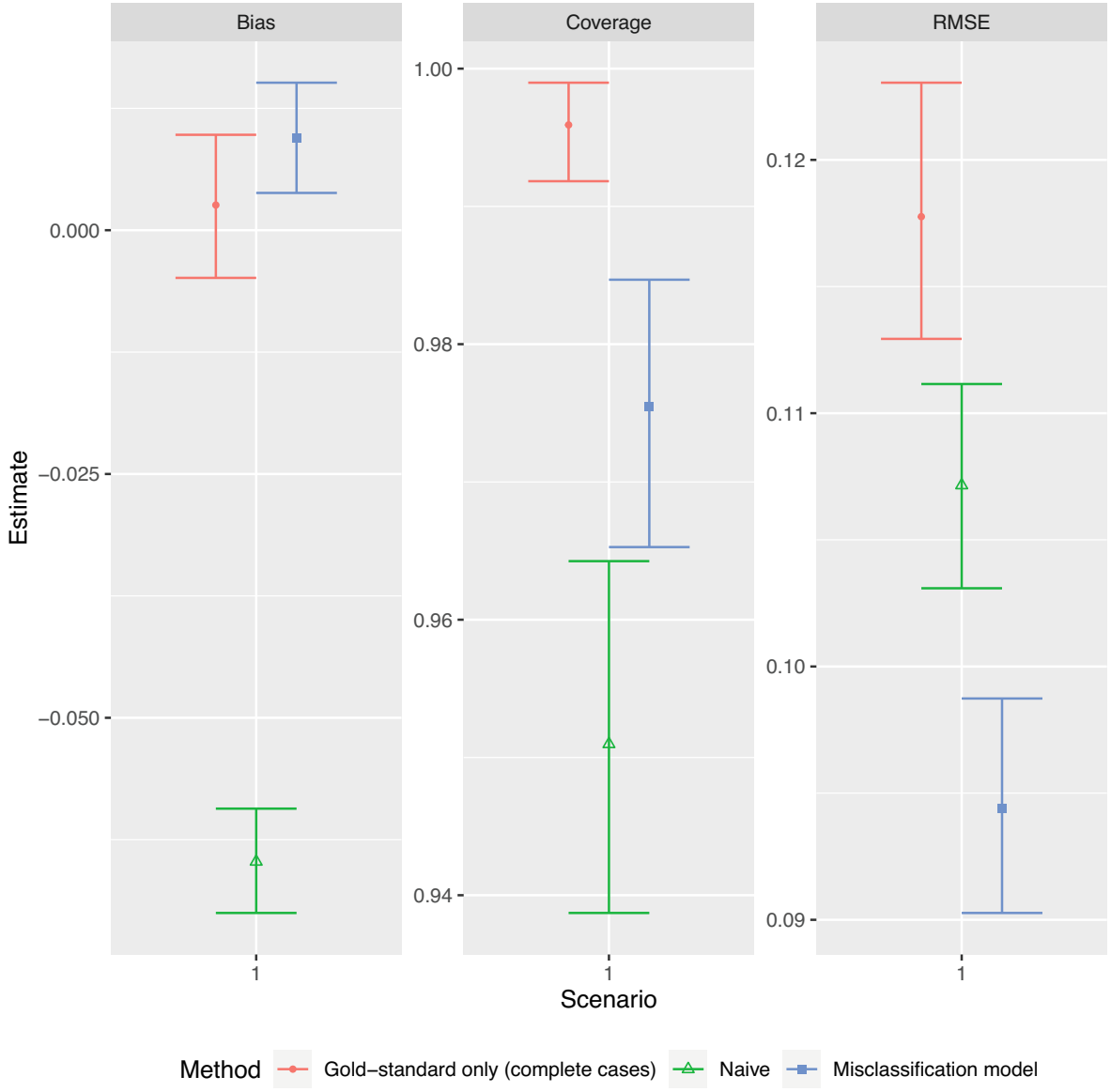

FIGURE 4 Bias, Coverage of the 95% credibility interval and root mean square error (RMSE) for the summary estimate of the exposure-outcome relation. AD, aggregate data; IPD, individual participant data. Gold indicates the model only used studies for which the gold standard x was available. The naive models used x^* for each observation for which x was not available [Colour figure can be viewed at wileyonlinelibrary.com]

account for this. We have unified methods for misclassification in meta-analysis in a one-stage Bayesian meta-analysis framework. Our methodology allows for the incorporation of covariates on the individual participant level to facilitate valid inference regarding therapeutic and etiologic effects, and added diagnostic and prognostic value. This modeling of the individual participant outcome, exposure and covariate values occurs via three sub-models: one for modeling the measurements, one for modeling the (gold standard) exposure, and one for modeling the outcome of interest. By doing so, both individual-level and study-level effects are accounted for in each part of the analysis. This, in turn, may restore the association between the exposure and the outcome.

In our motivating example data sets, the association between muscle pain and dengue was estimated with reduced error by applying the proposed misclassification models with individual participant covariate effects. These models account for the potential between-study heterogeneity in the prevalence of dengue and yielded adequate estimates of between-study heterogeneity of the muscle pain-dengue association.

In our simulations, we considered multiple scenarios where baseline outcome prevalence conditional on

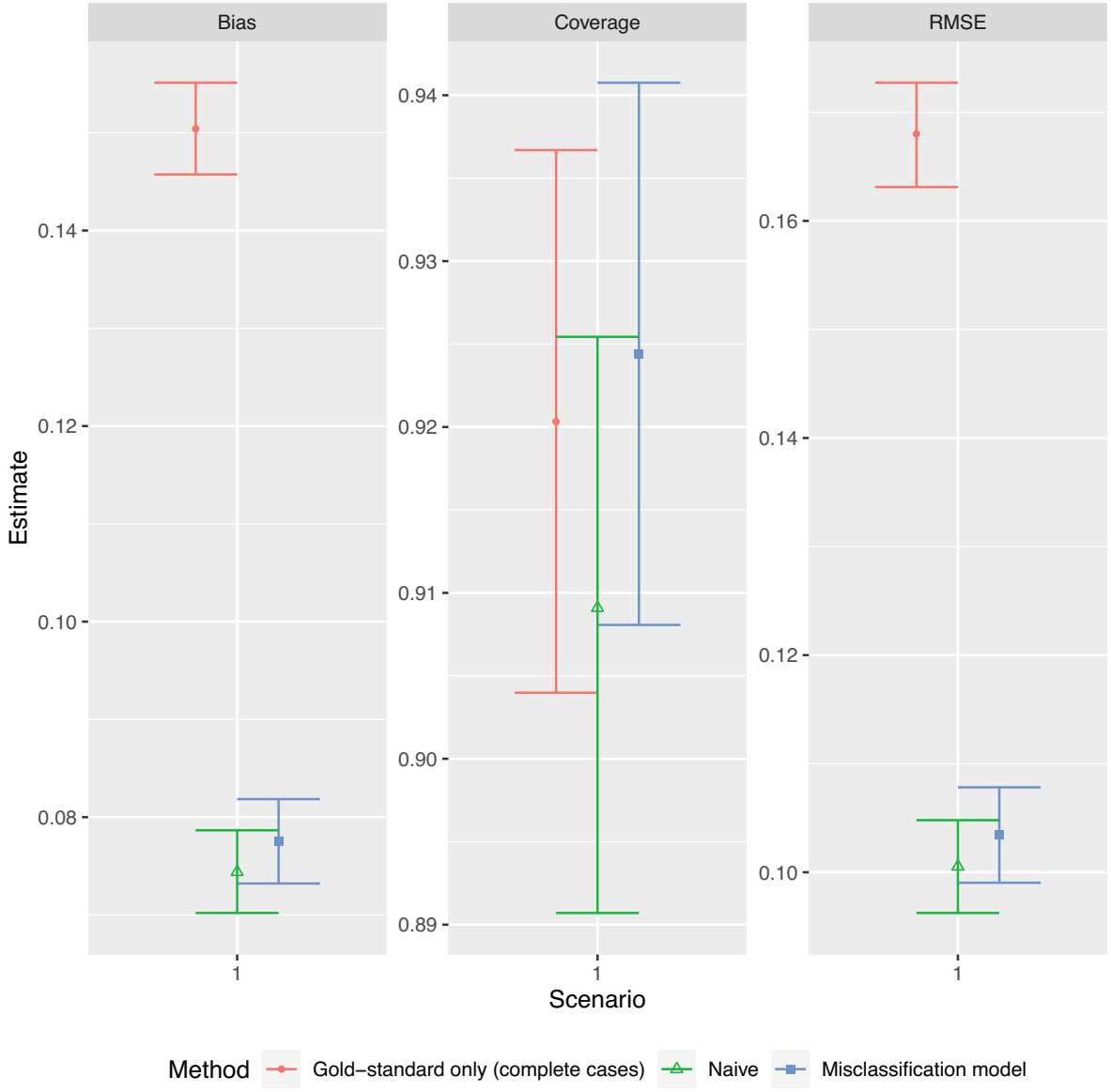

FIGURE 5 Bias, Coverage of the 95% Credibility Interval and Root Mean Square Error (RMSE) for estimating the heterogeneity of the exposure outcome relation across studies. AD, aggregate data; IPD, individual participant data. Gold indicates the model only used studies for which the gold standard x was available. The naive models used x^* for each observation for which x was not available. [Colour figure can be viewed at wileyonlinelibrary.com]

covariates and exposure effects were heterogeneous across studies, as well as the exposure-outcome association and the degree of misclassification, and compared the performance of several models. We found that analyses that only used data from studies in which the gold standard was measured, that is, complete case analyses, produced unbiased summary estimators of the exposure-outcome relation, but did so with considerably increased error unless in at least 7 out of 10 studies the gold standard was measured. Hence, the feasibility of restricting the analysis to patients with complete data for the (gold standard) exposure will depend on the remaining sample size. If this number is low, the variance of the resulting estimates will be large. In the extreme case, gold standard measurements are entirely unavailable for participants for whom the outcome is available, making this method impossible. In addition, the validity of a complete case analysis may become challenging when patients (or studies) for which only surrogate exposure are available differ with respect to covariates that are not part of the outcome model. As expected, our simulations also showed that naively using a possibly misclassified

surrogate measurement for the exposure when the gold standard is not available, introduced bias in the estimators for the exposure-outcome association.

This bias could be avoided or mitigated by using our proposed methodology for misclassification models. As the misclassification models use all available data, the resulting standard error is smaller than that of models that use only the observations for which only the gold standard is available. In our simulation, this resulted in estimates that were consistently better than those of the naive and gold standard methods, or that were equally good. However, we did observe bias for estimators of the heterogeneity of the exposure-outcome relation (using all approaches), which is no surprise as estimating this parameter is notoriously difficult in meta-analyses of few studies.[29,42] We found that applying a misclassification model on the individual participant data were particularly beneficial when the misclassification was strongly related to covariate values and when the number of studies where the gold standard exposure was measured was low. We found that three studies that have measured the gold standard measurement is sufficient (Scenario 2) to greatly reduce the bias resulting from misclassification, but the bias could not be prevented entirely in this case.

In general, over-specification should not induce bias in the estimators, provided that the sample contains enough information to estimate all parameters. Nor should it affect the coverage as the models appropriately account for the uncertainty. However, we stress that if we had applied an underspecified misclassification model, we would expect to have observed (some) bias in the estimators for the exposure-outcome association, as well as less favorable statistical properties in terms of RMSE and coverage. After all, although the misclassification was non-differential given covariates, once those covariates are removed from the model the misclassification may become differential.[6]

Contrary to our expectations, the fully Bayesian method that naively used the possibly misclassified x^* when the gold standard x was not available, was very well able to estimate the heterogeneity in the heterogeneity of the exposure-outcome relation. However, this may have been a result of two biases canceling each other out, as all methods estimators for the heterogeneity of the exposure-outcome relation were positively biased due to a limited sample size per study, influence from the priors and a relatively small amount of studies, and the naive methods produced estimates for the summary estimates of the exposure outcome-relation that were biased toward zero. When all estimates in a meta-analysis are drawn toward the null and the SE is kept constant, the heterogeneity estimate is guaranteed to be drawn toward zero.

Another surprise is that the estimation of heterogeneity of the exposure-outcome relation by the naive and misclassification methods was hardly affected by the number of studies for which the gold standard was observed; 3 of 10 (the lowest in our simulations) was sufficient. This is a sharp contrast with the methods that relied on the gold standard, which needed at least seven studies, or perhaps nine.

6.1 | Limitations and future directions

Although we recommend the implementation of misclassification models, an alternative strategy is to implement models that require fewer assumptions and do not depend on Bayesian MCMC sampling methods. Two such methods, RC and Multiple Imputation for Measurement Error correction (MIME), do not specify measurement submodels and require fewer distributional assumptions, and are therefore described as functional methods[12] or reclassification methods.[43] In contrast, in structural methods such as ours, a model is specified, which when analyzed with Bayesian methods allows for the appropriate propagation of uncertainty. However, this requires assumptions on the distribution of the gold standard measurement of the exposure and its surrogate measurement.[12] However, we focused on the scenario where the exposure is a binary variable that is potentially misclassified, which is common in epidemiology. This binary variable is assumed to follow a Bernoulli distribution, so specification of an exposure submodel does not add a major assumption[6] aside from congeniality, which is also required for RC and MIME. Although both of these methods have been applied to account for misclassification in single studies, neither has yet been adapted to the heterogeneous setting that is IPD-MA. This would require the specification of multiple heterogeneity parameters. We suggest that further research may focus on integrating these into the IPD-MA framework.

In case the exposure is a continuous variable which has been transformed into a binary variable at a specific cut-off point, alternative assumptions are needed for modeling the distribution of the exposure and its ME (see e.g., Reference 12). Our method could be further extended in case multiple surrogate exposure measurements are available for some or each participant, by specifying a measurement submodel for each surrogate measurement.

In the simulation study, we applied only one misclassification model as this simulation was intended as a *proof of concept*, not to assess the relative performance of all the described models in a variety of scenarios. All of the methods discussed here require covariates that predict the value of the gold standard measurement of the

exposure to be fruitful. If the available covariates are not predictive of the missing gold standard exposure or the surrogate exposure, only noise would be added by including individual participant covariate effects in the exposure and measurement submodel, respectively.

Due to the influence of misclassification on exposure-outcome associations and the presence of between-study heterogeneity, and the increase in parameters that are required to account for this, a larger amount of data are necessary than in an IPD-MA where misclassification is absent. This should be especially the case for the more complex misclassification models. In our simulation study, however, 3000 individual participants spread over 10 studies was sufficient for the estimators of the exposure-outcome relation to be unbiased and this bias was very small for 1000 participants. In a typical IPD-MA, where the sample size is often much larger, there should be enough information to estimate the more complex misclassification models.

6.2 | Conclusion

In an IPD-MA, the gold standard measurement of an exposure may be entirely unavailable for all participants in some studies, or unavailable for some participants in all studies, leaving the researcher with only surrogate measurements for these participants. If ignored, this induces bias in the estimators for exposure-outcome associations and other parameters of interest, which must be accounted for. Our Bayesian methodology can be applied to participant level data to reduce the error in the estimate of the exposure-outcome association compared with analyses restricted to participants for whom the gold standard measurement is observed, while appropriately propagating uncertainty for all parameters. This may provide unbiased estimators of the exposure-outcome association and coverage of the true effect by the 95% CI, provided that the model is specified correctly.

AUTHOR CONTRIBUTIONS

VJ, HC, LM, PG, TD conceived and designed the study, TJ acquired the data, VJ performed the analyses and drafted the first version of the article, all revised the article and agree to publication of the manuscript.

ACKNOWLEDGMENTS

This project has received funding from the European Union's Horizon 2020 research, the Canadian Institutes of Health Research, Institute of Genetics (CIHR-IG) grant agreement No 01886-000 and innovation programme under ReCoDID grant agreement No 825746. We thank the IDAMS consortium for providing aggregate data on the diagnosis of dengue. We would like to thank the Editor in Chief, the Associate Editor, and the reviewers for their helpful comments that have substantially improved the manuscript.

CONFLICT OF INTEREST

We have no conflicts of interest to declare.

DATA AVAILABILITY STATEMENT

The code and data for our motivating example is available on Github (github.com/VMTdeJong/Misclassification/Dengue).

DISCLAIMER

The views expressed in this paper are the personal views of the authors and may not be understood or quoted as being made on behalf of or reflecting the position of the regulatory agency/agencies or organizations with which the authors are employed/affiliated.

ORCID

Valentijn M. T. de Jong https://orcid.org/0000-0001-9921-3468
Lauren Maxwell https://orcid.org/0000-0002-0777-2092
Thomas P. A. Debray https://orcid.org/0000-0002-1790-2719

REFERENCES

1. Stewart LA, Tierney JF. To IPD or not to IPD? Advantages and disadvantages of systematic reviews using individual patient data. *Eval Health Prof*. 2002;25(1):76-97.
2. Sjoding MW, Cooke CR, Iwashyna TJ, Hofer TP. Acute respiratory distress syndrome measurement error. Potential effect on clinical study results. *Ann Am Thorac Soc*. 2016;13(7):1123-1128.
3. Simmons N, Donnell D, Ss O, et al. Assessment of contamination and misclassification biases in a randomized controlled trial of a social network peer education intervention to reduce HIV risk behaviors among drug users and risk partners in Philadelphia, PA and Chiang Mai, Thailand. *AIDS Behav*. 2015;19(10):1818-1827.
4. Choudhry NK. Randomized, controlled trials in health insurance systems. *N Engl J Med*. 2017;377(10):957-964. doi:10.1056/NEJMra1510058
5. Keys A, Kihlberg JK. Effect of misclassification on estimated relative prevalence of a characteristic: part I. two populations infallibly distinguished. Part II. Errors in two variables. *Am J Public Health*. 1963;53(10):1656-1665.
6. Carroll R, Ruppert D, Stefanski LA, Crainiceanu CM. *Measurement Error in Nonlinear Models. A Modern Perspective*. Chapmn & Hall/CRC; 2006.
7. Gustafson P, Greenland S. Misclassification. In: Wolfgang A, Iris P, eds. *Handbook of Epidemiology*. Springer New York; 2014:639-658.
8. van Smeden M, Lash TL, Groenwold RHH. *Five Myths about Measurement Error in Epidemiologic Research*. ResearchGate; 2019.

9. 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.
10. Bross I. Misclassification in 2 x 2 tables. *Biometrics*. 1954;10(4):478-486.
11. Copeland KT, Checkoway H, McMichael AJ, Holbrook RH. Bias due to misclassification in the estimation of relative risk. *Am J Epidemiol*. 1977;105(5):488-495.
12. Gustafson P. *Measurement Error and Misclassification in Statistics and Epidemiology: Impacts and Bayesian Adjustments*. CRC Press; 2003.
13. Dosemeci M, Wacholder S, Lubin JH. Does nondifferential misclassification of exposure always bias a true effect toward the null value? *Am J Epidemiol*. 1990;132(4):746-748.
14. Birkett NJ. Effect of nondifferential misclassification on estimates of odds ratios with multiple levels of exposure. *Am J Epidemiol*. 1992;136(3):356-362.
15. Weinberg CA, Umbach DM, Greenland S. When will nondifferential misclassification of an exposure preserve the direction of a trend? *Am J Epidemiol*. 1994;140(6):565-571.
16. Debray TPA, de Jong VMT, Moons KGM, Riley RD. Evidence synthesis in prognosis research. *Diagn Progn Res*. 2019;3(1):13.
17. Campbell H, de Jong VMT, Maxwell L, Jaenisch T, Debray TPA, Gustafson P. Measurement error in meta-analysis (MEMA)—a Bayesian framework for continuous outcome data subject to non-differential measurement error. *Res Synth Methods*. 2021;12(6):796-815.
18. Abo-Zaid G, Guo B, Deeks JJ, et al. Individual participant data meta-analyses should not ignore clustering. *J Clin Epidemiol*. 2013;66(8):865-873.e4.
19. Falley BN, Stamey JD, Beaujean AA. Bayesian estimation of logistic regression with misclassified covariates and response. *J Appl Stat*. 2018;45(10):1756-1769.
20. Nelson T, Song JJ, Chin YM, Stamey JD. Bayesian correction for misclassification in multilevel count data models. *Comput Math Methods Med*. 2018;2018:1-6.
21. Lian Q, Hodges JS, MacLehose R, Chu H. A Bayesian approach for correcting exposure misclassification in meta-analysis. *Stat Med*. 2019;38(1):115-130.
22. Jaenisch T, Tam DTH, Kieu NTT, et al. Clinical evaluation of dengue and identification of risk factors for severe disease: protocol for a multicentre study in 8 countries. *BMC Infect Dis*. 2016;16(1):120.
23. Anders KL, Nguyet NM, Van Vinh CN, et al. Epidemiological factors associated with dengue shock syndrome and mortality in hospitalized dengue patients in Ho Chi Minh City, Vietnam. *Am J Trop Med Hyg*. 2011;84(1):127-134.
24. Sudeep AB, Parashar D. Chikungunya: an overview. *J Biosci*. 2008;33(4):443-449.
25. Keogh RH, White IR. A toolkit for measurement error correction, with a focus on nutritional epidemiology. *Stat Med*. 2014;33(12):2137-2155.
26. Cole SR, Chu H, Greenland S. Multiple-imputation for measurement-error correction. *Int J Epidemiol*. 2006;35(4):1074-1081.
27. Richardson S, Gilks WR. A Bayesian approach to measurement error problems in epidemiology using conditional independence models. *Am J Epidemiol*. 1993;138(6):430-442.
28. Williams DR, Rast P, Bürkner PC. Bayesian Meta-Analysis with Weakly Informative Prior Distributions. PsyArXiv 2018.
29. Gelman A. Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). *Bayesian Anal*. 2006;1(3):515-534.
30. Polson NG, Scott JG. On the half-Cauchy prior for a global scale parameter. *Bayesian Anal*. 2012;7(4):887-902.
31. DerSimonian R, Laird N. Meta-analysis in clinical trials. *Control Clin Trials*. 1986;7(3):177-188.
32. Richardson S, Gilks WR. Conditional independence models for epidemiological studies with covariate measurement error. *Stat Med*. 1993;12(18):1703-1722.
33. Meng XL. Multiple-imputation inferences with uncongenial sources of input. *Stat Sci*. 1994;9:538-558.
34. Maclure M. The case-crossover design: a method for studying transient effects on the risk of acute events. *Am J Epidemiol*. 1991;133(2):144-153.
35. Möller J, Hessén-Söderman AC, Hallqvist J. Differential misclassification of exposure in case-crossover studies. *Epidemiology*. 2004;15(5):589-596.
36. Garcia-Closas M, Thompson WD, Robins JM. Differential misclassification and the assessment of gene-environment interactions in case-control studies. *Am J Epidemiol*. 1998;147(5):426-433.
37. Morabia A, Flandre P. Misclassification bias related to definition of menopausal status in case-control studies of breast cancer. *Int J Epidemiol*. 1992;21(2):222-228.
38. Levois M, Switzer P. Differential exposure misclassification in case-control studies of environmental tobacco smoke and lung cancer. *J Clin Epidemiol*. 1998;51(1):37-54.
39. Greenland S. Variance estimation for epidemiologic effect estimates under misclassification. *Stat Med*. 1988;7(7):745-757. doi: 10.1002/sim.4780070704
40. Wacholder S, Dosemeci M, Lubin JH. Blind assignment of exposure does not always prevent differential misclassification. *Am J Epidemiol*. 1991;134(4):433-437.
41. R Core Team. *R: A Language and Environment for Statistical*. R Foundation for Statistical Computing; 2020.
42. Langan D, Higgins JPT, Jackson D, et al. A comparison of heterogeneity variance estimators in simulated random-effects meta-analyses. *Res Synth Methods*. 2018;10(1):83-98.
43. Spiegelman D, Rosner B, Logan R. Estimation and inference for logistic regression with covariate misclassification and measurement error in main study/validation study designs. *J Am Stat Assoc*. 2000;95(449):51-61.

SUPPORTING INFORMATION

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

How to cite this article: de Jong VMT, Campbell H, Maxwell L, Jaenisch T, Gustafson P, Debray TPA. Adjusting for misclassification of an exposure in an individual participant data meta-analysis. *Res Syn Meth*. 2023;14(2):193-210. doi:10.1002/jrsm.1606