STATISTICS IN MEDICINE
Statist. Med. 2008; **27**:6111–6136
Published online 24 September 2008 in Wiley InterScience
(www.interscience.wiley.com) DOI: 10.1002/sim.3441

Meta-analysis of diagnostic test studies using individual patient data and aggregate data

Richard D. Riley[1,*,†], Susanna R. Dodd[1], Jean V. Craig[2], John R. Thompson[3] and Paula R. Williamson[1]

[1]*Centre for Medical Statistics and Health Evaluation, Faculty of Medicine, University of Liverpool, Shelley's Cottage, Brownlow Street, Liverpool L69 3GS, U.K.*
[2]*Evidence-based Child Health Unit, Institute of Child Health, University of Liverpool, RLC NHS Trust, Liverpool, U.K.*
[3]*Centre for Biostatistics and Genetic Epidemiology, Department of Health Sciences, University of Leicester, 2nd Floor, Adrian Building, University Road, Leicester LE1 7RH, U.K.*

SUMMARY

A meta-analysis of diagnostic test studies provides evidence-based results regarding the accuracy of a particular test, and usually involves synthesizing aggregate data (AD) from each study, such as the 2 by 2 tables of diagnostic accuracy. A bivariate random-effects meta-analysis (BRMA) can appropriately synthesize these tables, and leads to clinical results, such as the summary sensitivity and specificity across studies. However, translating such results into practice may be limited by between-study heterogeneity and that they relate to some 'average' patient across studies.

In this paper we describe how the meta-analysis of individual patient data (IPD) from diagnostic studies can lead to clinical results more tailored to the individual patient. We develop IPD models that extend the BRMA framework to include study-level covariates, which help explain the between-study heterogeneity, and also patient-level covariates, which allow one to assess the effect of patient characteristics on test accuracy. We show how the inclusion of patient-level covariates requires a careful separation of within-study and across-study accuracy-covariate effects, as the latter are particularly prone to confounding. Our models are assessed through simulation and extended to allow IPD studies to be combined with AD studies, as IPD are not always available for all studies. Application is made to 23 studies assessing the accuracy of ear thermometers for diagnosing fever in children, with 16 IPD and 7 AD studies. The models reveal that between-study heterogeneity is partly explained by the use of different measurement devices, but there is no evidence that being an infant modifies diagnostic accuracy. Copyright © 2008 John Wiley & Sons, Ltd.

KEY WORDS: meta-analysis; diagnosis; individual patient data (IPD); sensitivity; specificity

*Correspondence to: Richard D. Riley, Centre for Medical Statistics and Health Evaluation, Faculty of Medicine, University of Liverpool, Shelley's Cottage, Brownlow Street, Liverpool L69 3GS, U.K.
†E-mail: richard.riley@liv.ac.uk

Contract/grant sponsor: National Coordinating Centre for Research Capacity Development

Received 21 November 2007
Accepted 17 August 2008

1. INTRODUCTION

Numerous meta-analysis methods exist for combining diagnostic test results from multiple studies [1–8]. Recent research suggests that a bivariate random-effects meta-analysis (BRMA) of sensitivity and specificity is an appropriate method to use [7, 9–11]. This method accounts for the precision of estimates within the studies [9, 12] and any between-study heterogeneity in underlying sensitivity and specificity values; it also models any between-study correlation, which often exists because the threshold value (i.e. the cut-off level defining a 'positive' and 'negative' test result) often differs across studies for explicit and implicit (e.g. due to the use of different measurement devices) reasons. Harbord *et al*. [8] show that the bivariate approach is in some situations identical to the hierarchical summary receiver operating characteristic (HSROC) model [4, 5].

A BRMA of sensitivity and specificity provides informative clinical results; these include a summary sensitivity, a summary specificity, a summary diagnostic odds ratio, and a summary receiver operating curve (SROC) [7]. Such results indicate how well the test performs for some 'average' patient across studies; however, even when the summary sensitivity and specificity are high, it may be hard to translate such results into clinical practice. A key problem for interpretation is the between-study heterogeneity [13, 14], which may exist because of different cut-off levels, different methods of implementing the test (e.g. measurement device), different patient characteristics (e.g. the distribution of age, severity of diseased patients), and also unknown factors [15]. It is thus advisable to explore the reasons for heterogeneity where possible [13, 16], for example using meta-regression [17], and obtain meta-analysis results for relevant subgroups of studies, such as those using the same measuring device.

A further issue is that clinicians often want diagnostic strategies to be tailored for the individual patient, as diagnostic tests may perform better for some patients than others, and so meta-analysis results relating to some 'average' patient may be insufficient. A related problem is that within studies there may be overdispersion, where the sample data indicate more variability than expected if each patient's test response truly followed the same Bernoulli distribution. This may be caused by patient-level covariates affecting the underlying probability of a correct test response [18], and so it is important for meta-analysis to assess if and how patient-level covariates modify diagnostic accuracy; in this manuscript we refer to this as the 'accuracy-covariate effect'. This is akin to the need to estimate treatment–covariate interactions in meta-analysis of therapeutic trials [19]. It is generally recommended that individual patient data (IPD) are needed to estimate such interactions in meta-analysis [19], as it allows the *within*-study relationship between patient-level covariates and diagnostic accuracy to be modelled. This is in contrast to meta-regression of aggregate data (AD), where only the *across-study* relationship between study-level summaries (e.g. mean age, proportion male) and diagnostic accuracy can be modelled. Across-study relationships may differ from the within-study relationships [20] due to ecological bias and confounding [21, 22], and have low power [23].

Despite the potential benefits of IPD for meta-analysis of diagnostic studies, there has been only little consideration of how to synthesize IPD from diagnostic studies [14, 24], and how to assess accuracy-covariate effects. A review of applied diagnostic meta-analysis methods did not identify any that used IPD, leading to a recommendation that IPD methods should be explored in the future [14]. Of course, in practice IPD may not be available from all studies [25], and so guidance is also needed on how to combine IPD studies with AD studies in this situation [26], so to appropriately use all the evidence.

DOI: 10.1002/sim

In this paper we describe models for a BRMA of diagnostic studies when all IPD, or a mixture of IPD and AD are available. We show how to appropriately estimate the accuracy-covariate effects while also exploring between-study heterogeneity, in order to produce clinical results tailored to the individual patient. In Section 2 we introduce a motivating meta-analysis data set of 23 studies assessing the diagnosis of fever using ear thermometers. In Section 3 we describe our IPD models, and show how to separate within-study and across-study relationships. In Section 4 we assess these models via a simulation study, and then apply them to the temperature data set in Section 5. Finally, in Section 6 we discuss our work and suggest potential further research.

2. TEMPERATURE DATA

Craig *et al*. [27] systematically review thermometry studies comparing temperatures taken at the ear and rectum in children. These studies have been considered in a meta-analysis of method comparison studies [28], but a further clinical interest is the accuracy of infrared ear thermometry for diagnosing fever [29]. Twenty-three of the studies provided data regarding sensitivity and specificity, involving a total of 4098 children. Most studies defined fever as being a temperature $\geqslant$38.0°C (Table I), consistent with NHS guidelines for diagnosing fever in children [30]. Rectal temperature was the reference measure, as it is a well-established method of measuring temperature in children, and it guides clinical decisions and fever definition [31]. However, measuring temperature at the ear is clearly less invasive than measuring temperature at the rectum, and so ear measurements would be preferable if its diagnostic accuracy was adequate. Seven of the 23 studies provided AD in the form of 2 by 2 tables summarizing the diagnostic accuracy of ear temperature (Table I). The other 16 studies provided IPD, which contained each patient's 'true' fever diagnosis and their test diagnosis. Eleven of these 16 IPD studies also gave the age of each patient (Table I). The original analysis of this data assessed just the 2 by 2 tables of diagnostic accuracy from all the 23 studies [29]. However, conclusions were limited given considerable between-study heterogeneity in specificity and sensitivity. Such heterogeneity may be caused by, among other factors, different types of ear and rectal measurement devices, and different proportions of infants across studies (Table I). It is of interest to assess how measurement device and infant status modify diagnostic accuracy, so to explain between-study heterogeneity and develop more meaningful conclusions. Models using IPD can help achieve this, as described next.

3. MODELS FOR A BRMA OF SENSITIVITY AND SPECIFICITY USING IPD

We now introduce BRMA models for synthesizing IPD from diagnostic studies, building on IPD models for other study types [26, 32–36]. We note that given IPD for each study, and thus the ability to chose and vary the cut-off level, one could pool ROC curves across studies [2, 37]. This approach is most suitable when meta-analysis aims to assess the cut-off level produces the most accurate test results. In this paper, however, we focus on when there is mainly one particular cut-off of interest, and thus one diagnostic result per patient per study. This situation arises when the cut-off level is well established, as in the temperature data.

Let there be $i=1$ to m studies that perform a diagnostic test on n_{1i} diseased patients and n_{0i} non-diseased patients, whose disease status is truly known. The diagnostic test classifies each patient as either positive or negative, with positive indicating 'diseased' and negative indicating

DOI: 10.1002/sim

Table I. Summary of the 23 temperature studies identified for meta-analysis.

First author*	IPD	r_{11i}	n_{1i}	Sensitivity	r_{00i}	n_{0i}	Specificity	IPD for age?	Proportion of infants in fever group	Proportion of infants in non-fever group	Fever cut-off temperature; rectal/ear (°C)	Rectal thermometer device type	Ear thermometer device type
Bernardo	Y	0	3	0	33	35	0.94	Y	0	0.14	38.0/38.0	Electronic	CoreCheck
Brennan	Y	150	203	0.74	155	167	0.93	N	NA	NA	38.0/38.0	Electronic	FirstTemp
Davis	Y	9	18	0.50	46	48	0.96	Y	0.39	0.50	38.0/38.0	Electronic	FirstTemp
Green	Y	8	9	0.89	12	12	1.00	Y	0	0	38.0/38.0	Electronic	FirstTemp
Hoffman	Y	30	42	0.71	56	58	0.97	Y	0.12	0.43	38.0/38.0	Electronic	FirstTemp
Hoffman	Y	36	62	0.58	32	34	0.94	Y	0.03	0.29	38.0/38.0	Electronic	CoreCheck
Hoffman	Y	41	42	0.98	44	55	0.80	Y	0.29	0.33	38.0/38.0	Electronic	Thermoscan
Hooker	Y	10	15	0.67	24	24	1.00	Y	0.33	0.25	38.0/38.0	Electronic	FirstTemp
Hooker	Y	75	99	0.76	78	81	0.96	Y	0.34	0.69	38.0/38.0	Mercury	Thermoscan
Lanham	Y	53	103	0.51	74	75	0.99	N	NA	NA	38.0/38.0	Electronic	FirstTemp
Loveys	Y	12	30	0.40	44	46	0.96	N	NA	NA	38.0/38.0	Electronic	LightTouch Pedi-Q
Loveys	Y	37	47	0.79	74	93	0.80	N	NA	NA	38.0/38.0	Electronic	LightTouch Pedi-Q
Nypaver	Y	282	425	0.66	445	453	0.98	Y	0.40	0.51	38.0/38.0	Electronic	FirstTemp
Petersen-Smith	Y	9	10	0.90	214	222	0.96	Y	0.60	0.69	38.0/38.0	Mercury	FirstTemp
Rhoads	Y	7	27	0.26	38	38	1.00	Y	0.56	0.55	38.0/38.0	Electronic	FirstTemp
Robinson	Y	1	2	0.50	13	13	1.00	N	NA	NA	38.0/38.0	Probe	CoreCheck
Akinyinka	N	77	105	0.73	259	273	0.95	N	NA	NA	37.5/37.5	Mercury	Thermoscan
Greenes	N	53	109	0.49	193	195	0.99	N	NA	NA	38.0/38.0	Electronic	FirstTemp
Muma	N	48	87	0.55	136	136	1.00	N	NA	NA	38.0/38.0	Electronic	FirstTemp
Selfridge	N	16	18	0.89	75	84	0.89	N	NA	NA	38.1/37.6	Mercury	FirstTemp
Stewart	N	57	59	0.97	20	20	1.00	N	NA	NA	38.0/38.0	Electronic	FirstTemp
Terndrup	N	91	178	0.51	105	125	0.84	N	NA	NA	37.9/37.9	Electronic	FirstTemp
Wilshaw	N	16	16	1.00	60	104	0.58	N	NA	NA	38.0/38.0	Mercury	Ototemp Pedi-Q

IPD = individual patient data.
NA = not available as IPD for age not provided.
Y = yes, N = no.
*Full reference list is available on request.

DOI: 10.1002/sim

'non-diseased'. Let y_{1ik} be the test response (1 = positive, 0 = negative) of patient k in study i who truly has the disease, where $k = 1$ to n_{1i}, and let y_{0ij} be the test response (1 = negative, 0 = positive) of patient j in study i who truly does not have the disease, where $j = 1$ to n_{0i}. Thus, y_{1ik} and y_{0ij} are equal to 1 if the test response is correct and 0 otherwise. Summarizing test results over all patients produces AD in the form of r_{11i}, the number of patients in study i with a positive test result who truly have the disease, and r_{00i}, the number of patients in study i with a negative test result who truly do not have the disease. The observed sensitivity in each study is r_{11i}/n_{1i} and the observed specificity is r_{00i}/n_{0i}. Meta-analysis seeks to synthesize these results across studies to make inferences about diagnostic accuracy based on all the evidence.

3.1. Modelling summary sensitivity and specificity

3.1.1. Model specification. Consider first the simple, but potentially unrealistic, scenario where in each study sensitivity is the same for every diseased subject, and similarly specificity is the same for every non-diseased subject. In this situation, a number of authors propose a BRMA to jointly synthesize the sensitivity and specificity across studies [7, 8], with r_{11i} and r_{00i} modelled directly using the binomial distribution [9, 12]. This approach can equivalently be written with subject test responses modelled directly using the Bernoulli distribution:

$$\begin{aligned} y_{1ik} &\sim \text{Bernoulli}(p_{1i}) \\ \text{logit}(p_{1i}) &= \beta_{1i} \\ \beta_{1i} &= \beta_1 + u_{1i} \\ y_{0ij} &\sim \text{Bernoulli}(p_{0i}) \\ \text{logit}(p_{0i}) &= \beta_{0i} \\ \beta_{0i} &= \beta_0 + u_{0i} \\ \begin{pmatrix} u_{1i} \\ u_{0i} \end{pmatrix} &\sim \text{N}\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \boldsymbol{\Omega} \right], \quad \boldsymbol{\Omega} = \begin{pmatrix} \tau_1^2 & \tau_1\tau_0\rho \\ \tau_1\tau_0\rho & \tau_0^2 \end{pmatrix} \end{aligned} \tag{1}$$

Model (1) specifies an underlying logit-sensitivity and logit-specificity in each study, by $\text{logit}(p_{1i})$ and $\text{logit}(p_{0i})$, respectively. Further, u_{1i} and u_{0i} are random effects denoting that $\text{logit}(p_{1i})$ and $\text{logit}(p_{0i})$ are normally distributed about a mean logit-sensitivity of β_1 and a mean logit-specificity of β_0, with between-study variance τ_1^2 and τ_0^2, respectively, and between-study correlation ρ. Note ρ will usually be negative, as the underlying specificity will tend to decrease as the underlying sensitivity increases, and vice-versa, due to explicit and implicit differences in cut-off value across studies [7]. Model (1) can easily be extended to include study-level covariates that affect either the underlying sensitivity or specificity across studies by replacing one or both of β_1 and β_0 with linear predictors in the covariates (see Section 3.2.3, for an example). However, an under-researched issue is how to extend model (1) to account for the potential impact of patient-level covariates on diagnostic accuracy; this motivates Section 3.1.3 onwards.

3.1.2. Model estimation. Model (1), and subsequent models described in this paper, can be fitted in a frequentist framework using maximum likelihood estimation. This will produce, for example, $\hat{\beta}_1$ and $\hat{\beta}_0$ by marginalizing study-specific logit-sensitivity and logit-specificity over the random

DOI: 10.1002/sim

effects [38]. This can be undertaken using, for example, PROC NLMIXED in SAS or the xtmelogit procedure in STATA (code available on request) [9, 39, 40], which fit nonlinear mixed models by maximizing an approximation to the likelihood integrated over the random effects. Different integral approximations are available, with adaptive Gaussian quadrature our method of choice [39]. This requires a number of quadrature points to be specified, with increasing estimation accuracy as the number of points increases, but at the expense of an increased computational time. We generally chose 10 quadrature points for our analyses, as this gave estimates very close to those when using >10 points but in a faster time. Successful convergence of the optimization procedure was assumed when successive iteration estimates differed by $<10^{-7}$, resulting in parameter estimates and their approximate standard errors based on the second derivative matrix of the likelihood function. Estimates of clinical interest from model (1) include the summary estimates of sensitivity and specificity across studies, obtained by $\exp(\hat{\beta}_1)/[1+\exp(\hat{\beta}_1)]$ and $\exp(\hat{\beta}_0)/[1+\exp(\hat{\beta}_0)]$, respectively, and the SROC, with 95 per cent confidence and prediction ellipses around it [7, 41]. The summary diagnostic odds ratio and the likelihood ratio for positive and negative test results can also be estimated [7].

Model (1) and subsequent models can alternatively be fitted in a Bayesian framework, for example using Markov Chain Monte Carlo estimation in WinBUGS [42]. This additionally requires specification of prior distributions for the unknown parameters (i.e. β_1, β_0, and $\mathbf{\Omega}$ in model (1)), and thus allows external information to be incorporated. Vague prior distributions can also be specified, such as N(0, 100 000) for β_1 and β_0; however, specifying a non-informative prior distribution for $\mathbf{\Omega}$ may be difficult [43, 44], as estimation of $\mathbf{\Omega}$ is dependent on the number of studies, which will often be small. Sensitivity analysis of the posterior estimates to the choice of prior distribution for $\mathbf{\Omega}$ is thus advised. Zwinderman and Bossuyt [11] assume $\mathbf{\Omega}^{-1} \sim \text{Wishart}(\mathbf{S}, t)$, a common conjugate prior distribution for precision matrices [44]. The t represents the degrees of freedom, which essentially represents the prior number of studies, and given no prior information, it is usually set to the smallest feasible value, which is 2 here. $\mathbf{S}$ is the scale matrix, which can be considered the prior mean estimate of $\mathbf{\Omega}$. In our Bayesian analyses, we used $\mathbf{\Omega}^{-1} \sim \text{Wish}\left(\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, 2\right)$ and then performed sensitivity analysis to the choice of this prior.

3.1.3. Accounting for overdispersion within studies. Model (1) assumes that in each study p_{1i} is the same for diseased patients and p_{0i} is the same for non-diseased patients, and thus implicitly that patient-level covariates do not affect the probability of a correct test response. If this assumption is incorrect, then model (1) is potentially inappropriate as it does not specifically account for the variation in p_{1i} and p_{0i} across patients within each study [18, 38], and may thus suffer from overdispersion within studies. To appropriately account for such overdispersion, model (1) can be extended to include further random effects, e_{1ik} and e_{0ij}, that allow $\text{logit}(p_{1i})$ and $\text{logit}(p_{0i})$ to vary across subjects within each study, with variances of σ^2_{1i} and σ^2_{0i}, respectively, as follows:

$$\begin{aligned} y_{1ik} &\sim \text{Bernoulli}(p_{1ik}) \\ \text{logit}(p_{1ik}) &= \beta_{1i} + e_{1ik} \\ \beta_{1i} &= \beta_1 + u_{1i} \\ y_{0ij} &\sim \text{Bernoulli}(p_{0ij}) \\ \text{logit}(p_{0ij}) &= \beta_{0i} + e_{0ij} \end{aligned} \tag{2}$$

DOI: 10.1002/sim

$$\beta_{0i} = \beta_0 + u_{0i}$$
$$e_{1ik} \sim \mathrm{N}(0, \sigma^2_{1i})$$
$$e_{0ij} \sim \mathrm{N}(0, \sigma^2_{0i})$$
$$\begin{pmatrix} u_{1i} \\ u_{0i} \end{pmatrix} \sim \mathrm{N}\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \boldsymbol{\Omega} \right], \quad \boldsymbol{\Omega} = \begin{pmatrix} \tau^2_1 & \tau_1\tau_0\rho \\ \tau_1\tau_0\rho & \tau^2_0 \end{pmatrix}$$

Each patient now has their own individual probability of a correct test response, with the errors e_{1ik} or e_{0ij} accounting for their unmeasured patient-level covariates. β_{1i} and β_{0i} thus, respectively, represent the underlying logit-sensitivity and logit-specificity in study i for a patient with unmeasured standardized covariates of zero (i.e. where $e_{1ik}=0$ and $e_{0ij}=0$). Note that β_{1i} and β_{0i} themselves are assumed normally distributed about a mean logit-sensitivity of β_1 and a mean logit-specificity of β_0 across studies, which summarize the performance of the test across all studies, akin to β_1 and β_0 from model (1).

Without multiple test results per subject it may be difficult to estimate σ^2_{1i} and σ^2_{0i} with any accuracy [18], so it may be necessary to simplify the estimation by assuming σ^2_{1i} and σ^2_{0i} are the same across studies. If estimation continues to be difficult, one could impute a range of σ^2_{1i} and σ^2_{0i} values, or in a Bayesian framework a range of different prior distributions for σ^2_{1i} and σ^2_{0i}, and observe how meta-analysis results change. The more pertinent requirement, however, is to alleviate the need for σ^2_{1i} and σ^2_{0i} by extending model (1) to include those patient-level covariates that actually modify the underlying sensitivity and specificity within each study. This issue is now considered further.

3.2. Modelling accuracy-covariate effects

Let x_{ik} and x_{ij} be a patient-level covariate, such as age, observed for diseased (k) and non-diseased (j) patients, and consider that x_{ik} and x_{ij} may potentially modify diagnostic accuracy. Given IPD for each study, models (1) or (2) can be extended to assess the effect of this covariate on diagnostic accuracy. However, the specification of within-study and across-study effects needs careful consideration, as in other contexts involving clustering [26, 33, 34, 45, 46]. Patient-level covariates vary both within studies (e.g. the age of patients varies within studies) and across studies (e.g. the mean age of patients varies across studies). Of key interest are the *within-study* effects between diagnostic accuracy and individual covariate values; i.e. the sensitivity-covariate effect, $\gamma_{1\mathrm{W}}$ say, and the specificity-covariate effect, $\gamma_{0\mathrm{W}}$ say (where 'W' emphasizes this is a *within*-study relationship). These effects can explain the within-study variation as specified in model (2) through the random effects, e_{1ik} and e_{0ij}, and thus alleviate the need to estimate σ^2_{1i} and σ^2_{0i}. Also available are the *across-study* effects, but these less meaningfully describe how the mean covariate value in each study (e.g. mean age) is associated with the underlying sensitivity and specificity across studies; i.e. the effect of $\overline{x}_{1i}$ on underlying mean logit-sensitivity across studies, $\gamma_{1\mathrm{A}}$, and the effect of $\overline{x}_{0i}$ on underlying mean logit-specificity across studies, $\gamma_{0\mathrm{A}}$ (where 'A' emphasizes this is an *across-study* relationship). These effects can help explain the between-study variation as specified in model (1) through τ^2_1 and τ^2_0.

Traditional meta-analysis methods have relied on AD, from which only the across-study effects are estimable using, for example, meta-regression [47]. In this situation, the across-study effect estimates obtained are sometimes used to make inferences about the within-study effects, under

DOI: 10.1002/sim

the assumption that $\gamma_{1A}=\gamma_{1W}$ and $\gamma_{0A}=\gamma_{0W}$, so that $\hat{\gamma}_{1A}$ and $\hat{\gamma}_{0A}$ are assumed unbiased estimates of the within-study effects. However, it is well known that ecological bias and confounding can affect this assumption [33, 34]. When IPD are available from diagnostic studies, a major advantage is that one can explicitly model the within-study effects separately to the across-study effects, to avoid or assess the threat of ecological bias, as now described.

3.2.1. Modelling within-study and across-study effects. Let us extend model (1) to include the within-study and across-study effects separately by centering x_{ik} about its mean $\overline{x}_{1i}$, and by centering x_{ij} about its mean $\overline{x}_{0i}$ [26]:

$$\begin{aligned}
y_{1ik} &\sim \text{Bernoulli}(p_{1ik}) \\
\text{logit}(p_{1ik}) &= \beta_{1i}+\gamma_{1W}(x_{ik}-\overline{x}_{1i}) \\
\beta_{1i} &= \alpha_1+\gamma_{1A}\overline{x}_{1i}+u_{1i} \\
y_{0ij} &\sim \text{Bernoulli}(p_{0ij}) \\
\text{logit}(p_{0ij}) &= \beta_{0i}+\gamma_{0W}(x_{ij}-\overline{x}_{0i}) \\
\beta_{0i} &= \alpha_0+\gamma_{0A}\overline{x}_{0i}+u_{0i} \\
\begin{pmatrix} u_1 \\ u_0 \end{pmatrix} &\sim \text{N}\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \boldsymbol{\Omega}\right], \quad \boldsymbol{\Omega}=\begin{pmatrix} \tau_1^2 & \tau_1\tau_0\rho \\ \tau_1\tau_0\rho & \tau_0^2 \end{pmatrix}
\end{aligned} \tag{3}$$

The parameters here are as for model (2), with additionally the within-study effects, γ_{1W} and γ_{0W}, which indicate the change in an individual's logit-sensitivity for a one-unit increase in x_{ik}, and the change in an individual's logit-specificity for a one-unit increase in x_{ij}, respectively. In addition, α_1 and α_0 denote the underlying logit-sensitivity and logit-specificity, respectively, in studies where $x_{ik}=0$ and $x_{ij}=0$ for all subjects. The across-study effects, γ_{1A} and γ_{0A}, denote the change in underlying mean logit-sensitivity and underlying mean logit-specificity for a one-unit increase in $\overline{x}_{1i}$ and $\overline{x}_{0i}$, respectively. The underlying mean logit-sensitivity and underlying mean logit-specificity in a study with covariate means $\overline{x}_{1i}$ and $\overline{x}_{0i}$ is thus $\hat{\alpha}_1+\hat{\gamma}_{1A}\overline{x}_{1i}$ and $\hat{\alpha}_0+\hat{\gamma}_{0A}\overline{x}_{0i}$, respectively. One could set $\overline{x}_{1i}$ and $\overline{x}_{0i}$ to $\overline{x}_1$ and $\overline{x}_0$, respectively, denoting the mean x_{ik} and mean x_{ij} across all patients in all studies, to give a summary logit-sensitivity and summary logit-specificity similar in interpretation to $\hat{\beta}_1$ and $\hat{\beta}_0$ from models (1) and (2).

Modifications and advantages of model (3). The patient-level covariate in model (3) could be continuous or binary, and for the former a linear trend with diagnostic accuracy on the logit-scale is assumed, though nonlinear effects can alternatively be specified. γ_{1W} and γ_{0W} are also assumed fixed across studies, but these could also be made random if desired. Model (3) assumes only one patient-level covariate is important, but extension to two or more patient-level covariates is straightforward, with each covariate centered at the patient-level and its mean added at the study-level. One could also add further random effects within each study, as in model (2), if one believed that unexplained within-study variability still exists across subjects.

Model (3) highlights the theoretical advantages of having IPD. Given AD one can only estimate $\hat{\gamma}_{1A}$ and $\hat{\gamma}_{0A}$, but using IPD one can simultaneously estimate $\hat{\gamma}_{1W}$ and $\hat{\gamma}_{0W}$ alongside $\hat{\gamma}_{1A}$ and $\hat{\gamma}_{0A}$. The benefits of $\hat{\gamma}_{1W}$ and $\hat{\gamma}_{0W}$ over $\hat{\gamma}_{1A}$ and $\hat{\gamma}_{0A}$ are assessed in a simulation study in Section 4. If preferred one can alternatively fit model (3) without the across-study effects, leaving then just

DOI: 10.1002/sim

the within-study effects and summary effects [33]. Indeed, practitioners may not be interested in $\hat{\gamma}_{1A}$ and $\hat{\gamma}_{0A}$, or may consider it unintuitive that model (3) regresses the individual probability of a correct test response against the mean covariate values in a study ($\overline{x}_{1i}$ and $\overline{x}_{0i}$) as well as individual covariate values (x_{ik} and x_{ij}). The inclusion of γ_{1A} and γ_{0A} in model (3) is to explain why the underlying mean logit-sensitivity and underlying mean logit-specificity change across studies. In Section 4.1, we show that if x_{ik} and x_{ij} are important in explaining variations at the patient-level, then theoretically $\overline{x}_{1i}$ and $\overline{x}_{0i}$ should be important in explaining variations at the study-level; thus including $\hat{\gamma}_{1A}$ and $\hat{\gamma}_{0A}$ can reduce between-study variation, making $\hat{\tau}_1^2$ and $\hat{\tau}_0^2$ smaller in model (3) than models (1) or (2), which is desirable [16].

3.2.2. Including additional study-level covariates. Models 1– 3 can also accommodate study-level covariates, such as measurement device, in order to explain the between-study heterogeneity and obtain results for specific subgroups of studies. For example, consider extending model (3) to include a categorical study-level covariate, z_i, which has l categories. We thus include l dummy variables, D_{ic}, where $c=1$ to l, which enable a different diagnostic accuracy for each category:

$$\begin{aligned}
y_{lik} &\sim \text{Bernoulli}(p_{1ik}) \\
\text{logit}(p_{1ik}) &= \beta_{1i}+\gamma_{1W}(x_{ik}-\overline{x}_{1i}) \\
\beta_{1i} &= \left(\sum_{c=1}^{l}\lambda_{1c}D_{ic}\right)+\gamma_{1A}\overline{x}_{1i}+u_{1i} \\
y_{0ij} &\sim \text{Bernoulli}(p_{0ij}) \\
\text{logit}(p_{0ij}) &= \beta_{0i}+\gamma_{0W}(x_{ij}-\overline{x}_{0i}) \\
\beta_{0i} &= \left(\sum_{c=1}^{l}\lambda_{0c}D_{ic}\right)+\gamma_{0A}\overline{x}_{0i}+u_{0i} \\
\begin{pmatrix}u_{1i}\\ u_{0i}\end{pmatrix} &\sim \text{N}\left[\begin{pmatrix}0\\0\end{pmatrix},\boldsymbol{\Omega}\right], \quad \boldsymbol{\Omega}=\begin{pmatrix}\tau_1^2 & \tau_1\tau_0\rho\\ \tau_1\tau_0\rho & \tau_0^2\end{pmatrix}
\end{aligned} \tag{4}$$

Here, $D_{ic}=1$ if z_i is in category c and 0 otherwise, and the other parameters are as specified for model (3), with additionally λ_{1c} and λ_{0c} denoting the mean logit-sensitivity and mean logit-specificity in category c for a study with $\overline{x}_{1i}=0$ and $\overline{x}_{0i}=0$, respectively. If $\hat{\lambda}_{1c}$ and $\hat{\lambda}_{0c}$ are different across categories, then $\hat{\tau}_1^2$ and $\hat{\tau}_0^2$ are likely to be smaller in model (4) than models (1), (2) or (3). Note that γ_{1W} and γ_{0W} are again assumed fixed across categories, and $\hat{\gamma}_{1W}$ and $\hat{\gamma}_{0W}$ will be very close to those from model (3) as their correlation with other study-level estimates is negligible (see Section 4) [26]. Model (4) can be adapted to allow different within-study effects or even a different between-study covariance matrix for each category, but at the expense of increased computational time and potential estimation problems when some categories are rare [5, 10].

3.3. Combining IPD studies and AD studies

IPD may only be available from a proportion of studies in the meta-analysis [25]. The remaining studies may provide AD, which for diagnostic studies are usually in the form of a 2 by 2 diagnostic accuracy table (i.e. n_{1i}, n_{0i}, r_{11i}, and r_{00i}), perhaps with study-level covariates and aggregated patient-level covariates. The question is thus posed about how to combine IPD studies with AD

DOI: 10.1002/sim

studies in this situation. In the AD studies, one can reconstruct IPD from the available 2 by 2 table by creating a row for each subject in the study and delegating them to test responses and disease status that collectively mirror the totals in the table. This enables AD studies to be combined with IPD studies using the framework of models (1) or (2) above, with additional study-level covariates included as necessary, as long as they are available for both IPD and AD studies.

In terms of estimating the within-study effects, it may be difficult to incorporate AD studies within models (3) or (4) as patient-level covariates cannot usually be reconstructed within the AD studies. In such situations, it is most appropriate to allow only IPD studies to estimate $\gamma_{1\mathrm{W}}$ and $\gamma_{0\mathrm{W}}$, but both IPD and AD studies to estimate all other parameters. To achieve this, a framework is required that simultaneously fits a model for IPD studies with a model for AD studies, with the two models linked by common parameters [26, 48, 49]. For example, given one patient-level covariate and one categorical study-level covariate, model (4) could be fitted to the IPD studies; simultaneously, one could also fit model (2) to the reconstructed IPD from the AD studies while also including the study-level covariate and aggregated patient-level covariate. This analysis can be specified as:

$$
\begin{aligned}
\textbf{IPD studies}:\ y_{1ik} &\sim \text{Bernoulli}(p_{1ik}) \\
\text{logit}(p_{1ik}) &= \beta_{1i} + \gamma_{1\mathrm{W}}(x_{ik} - \overline{x}_{1i}) \\
y_{0ij} &\sim \text{Bernoulli}(p_{0ij}) \\
\text{logit}(p_{0ij}) &= \beta_{0i} + \gamma_{0\mathrm{W}}(x_{ij} - \overline{x}_{0i}) \\
\textbf{AD studies}:\ y_{1ik} &\sim \text{Bernoulli}(p_{1ik}) \\
\text{logit}(p_{1ik}) &= \beta_{1i} + e_{1ik} \\
y_{0ij} &\sim \text{Bernoulli}(p_{0ij}) \\
\text{logit}(p_{0ij}) &= \beta_{0i} + e_{0ij} \\
e_{1ik} &\sim \mathrm{N}(0, \sigma^2_{1i}) \\
e_{0ij} &\sim \mathrm{N}(0, \sigma^2_{0i}) \\
\textbf{All studies}:\ \beta_{0i} &= \left(\sum_{c=1}^{l} \lambda_{0c} D_{ic}\right) + \gamma_{1\mathrm{A}}\overline{x}_{1i} + u_{0i} \\
\beta_{1i} &= \left(\sum_{c=1}^{l} \lambda_{1c} D_{ic}\right) + \gamma_{0\mathrm{A}}\overline{x}_{0i} + u_{1i} \\
\begin{pmatrix} u_{1i} \\ u_{0i} \end{pmatrix} &\sim \mathrm{N}\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \boldsymbol{\Omega}\right], \quad \boldsymbol{\Omega} = \begin{pmatrix} \tau_1^2 & \tau_1\tau_0\rho \\ \tau_1\tau_0\rho & \tau_0^2 \end{pmatrix}
\end{aligned}
\tag{5}
$$

The approach enables *all* studies to estimate the summary sensitivity and summary specificity for each category of the study-level covariate, and also the across-study effects and the between-study covariance matrix; however, just IPD studies will estimate the within-study accuracy-covariate effects. Thus, both IPD studies and AD studies are contributing valuable information. The approach works because $\hat{\gamma}_{1\mathrm{W}}$ and $\hat{\gamma}_{0\mathrm{W}}$ have little correlation with other parameter estimates (see Section 4.1) [26]. As discussed for model (3), model (5) can alternatively be fitted without the across-study

DOI: 10.1002/sim

effects if preferred, and this is one approach to take if AD studies do not provide $\overline{x}_{1i}$ and $\overline{x}_{0i}$. Note that e_{1ik} and e_{0ij} are needed within the AD model to account for the unavailable patient-level covariates [18]. Extensions of the IPD model may also be needed, as described in Section 3.2.1, for example, to allow additional random effects that account for remaining unexplained within-study overdispersion, as shown in Section 3.1.3.

4. A SIMULATION STUDY

We now describe two simulation studies to assess model (3) further. Previous simulation studies of BRMA focus on the estimation of the summary sensitivity and specificity, and the between-study covariance matrix [9, 10]. Our simulation results for these parameters are consistent with these reports, and so we focus here on the accuracy-covariate effects.

4.1. Simulation 1: one patient-level covariate

4.1.1. Simulation description. Each simulation generated IPD for 1000 meta-analysis data sets, each containing $i=1$ to m studies. A binary response (i.e. y_{0ij} for non-diseased and y_{1ik} for diseased patients) and a continuous patient-level covariate were generated for each patient in each of the m studies using an 8-step process (Figure 1). In steps 1, 3, and 6 parameter values were required to be chosen. For all simulations our choices for $\beta_1, \beta_0, \gamma_{1W}, \gamma_{0W}, \tau_1^2, \tau_0^2$, and ρ were fixed, relating to values akin to those in the temperature data analysis (Section 5). We chose $\beta_1=2$ and $\beta_0=0.4$, relating to a summary sensitivity of 88 per cent and a summary specificity of 60 per cent across studies. We chose $\tau_1^2=\tau_0^2=1$, indicating large between-study heterogeneity in underlying diagnostic accuracy, and a between-study correlation of $\rho=-0.5$. The within-study accuracy-covariate effects were set as $\gamma_{1W}=0.05$ and $\gamma_{0W}=-0.05$. Thus, as the covariate increases, sensitivity improves but specificity decreases. For example, an increase in the patient-level covariate of 10 increases an initial logit-sensitivity of 2 (sensitivity 88 per cent) to 2.5 (sensitivity 92 per cent), and decreases an initial logit-specificity of 0.4 (specificity 60 per cent) to -0.1 (specificity 47 per cent).

In our simulations we did vary V_1 ($=25$ or 100), V_2 ($=25$ or 100), m ($=10$ or 20), and the number of patients across studies ($=$'small' or 'large'). We simulated and assessed 1000 meta-analysis data sets for different combinations of V_1, V_2, m, and the number of patients (Table II). V_1 is the variance of the mean covariate value across studies, whereas V_2 is the variance of the covariate value within studies. The ratio of these values is known to influence the power of interaction estimates in meta-analysis [50, 51]. The number of studies chosen reflects that often observed in practice, and we defined a 'large' number of patients to be about 3000 patients across studies in total, and a 'small' number to be about 300 patients in total. We let the number of patients within each study vary while ensuring the total patients was as specified. For example, given 10 studies and a 'large' number of patients, within the 10 studies were 1000, 750, 500, 250, 200, 150, 100, 76, 50, and 20 patients (a combined total of 3096), reflecting the temperature studies (Table I).

4.1.2. Simulation results. To each of the 1000 meta-analysis data sets in each simulated setting, we applied model (3) and the results are given in Table II. In all settings, $\hat{\gamma}_{1W}$ and $\hat{\gamma}_{0W}$ are approximately unbiased estimates of the accuracy-covariate effects, as their mean bias is close to zero. The mean bias of $\hat{\gamma}_{1A}$ and $\hat{\gamma}_{0A}$ is also close to zero in all $m=20$ settings, and also in the

DOI: 10.1002/sim

Step 1: Values were chosen for m, β_1, β_0, γ_{1W}, γ_{0W}, τ_1^2, τ_0^2, and ρ. The number of patients in each study ($n_{1i} + n_{0i}$) was also decided; for simplicity the number of diseased and non-diseased patients was set equal in each study.

Step 2: Underlying mean logit-sensitivity and logit-specificity values, β_{1i} and β_{0i}, were simulated for each of the m studies using the bivariate framework:

$$\beta_{1i} = \beta_1 + u_{1i}$$
$$\beta_{0i} = \beta_0 + u_{0i}$$
$$\begin{pmatrix} u_1 \\ u_0 \end{pmatrix} \sim N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \boldsymbol{\Omega}\right], \quad \boldsymbol{\Omega} = \begin{pmatrix} \tau_1^2 & \tau_1\tau_0\rho \\ \tau_1\tau_0\rho & \tau_0^2 \end{pmatrix}$$

Step 3: An underlying mean patient-level covariate value ($\overline{X}_i$) was simulated for each study from a normal distribution with mean 0 and variance V_1, with V_1 chosen prior to the simulation. For simplicity, $\overline{X}_i$ and V_1 were assumed the same in diseased and non-diseased patients. A mean of 0 was chosen for simplicity and the covariate assumed to be some continuous measure that could take negative or positive values.

Step 4: For each patient in the study, their patient-level covariate (x_{ik} for diseased patients and x_{ij} for non-diseased patients) was simulated from a normal distribution with mean $\overline{X}_i$ and variance V_2, with $\overline{X}_i$ available from Step 3 and V_2 prespecified.

Step 5: For diseased patients, their individual probability (p_{1ik}) of having a positive test response was specified as logit(p_{1ik}) = $\beta_{1i} + \gamma_{1W} x_{ik}$. For non-diseased patients, their individual probability (p_{0ij}) of having a negative test response was specified as logit(p_{0ij}) = $\beta_{0i} + \gamma_{0W} x_{ij}$. Thus the probability of a correct test response depended on the underlying sensitivity and specificity, and importantly also the patient-level covariate.

Step 6: In each study, the binary response, y_{1ik}, for each diseased patient was then sampled from a Bernoulli(p_{1ik}) distribution, with p_{1ik} as defined in Step 5; similarly, for each non-diseased patient the test response, y_{0ij}, was sampled from a Bernoulli(p_{0ij}) distribution, with p_{0ij} as defined in Step 5. A simulated value of 1 for y_{1ik} or y_{0ij} denoted a correct test result, and 0 indicated an incorrect test result.

Step 7: Steps 2 to 6 were repeated 1000 times, to obtain 1000 meta-analysis datasets each containing m studies providing patient data of y_{1ik} and x_{ik} for diseased patients and y_{0ij} and x_{ij} for non-diseased patients.

Step 8: Model (3) was applied to each of the 1000 simulated meta-analysis datasets. The parameter estimates from across all 1000 simulations were assessed by calculating their mean bias, their mean-square error (MSE), and their mean standard error.

Figure 1. Description of the simulation procedure.

$m = 10$ setting, where V_1 is large relative to V_2. However, as V_2 becomes relatively large, there appears to be a very small bias in $\hat{\gamma}_{1A}$ and $\hat{\gamma}_{0A}$ as their means are attenuated toward zero, as noted elsewhere [18, 52]; for example, given $m = 10$, a 'large' number of patients, $V_1 = 25$ and $V_2 = 100$, the mean $\hat{\gamma}_{1A}$ is 0.044 and the mean $\hat{\gamma}_{0A}$ is -0.044, compared with the true effects of 0.05 and -0.05, respectively. In all settings there is a little agreement between the within-study and across-study effect estimates from the same meta-analysis; their correlation is between -0.06 and 0.06 in any meta-analysis, with a mean of zero.

DOI: 10.1002/sim

Table II. Summary of the accuracy-covariate effect estimates from model (3) in the simulations of Section 4.1 (no across-study confounding); comparison of $\hat{\gamma}_{1W}$ and $\hat{\gamma}_{0W}$ with $\hat{\gamma}_{1A}$ and $\hat{\gamma}_{0A}$, and $\hat{\gamma}_{1WA}$ and $\hat{\gamma}_{0WA}$.

				Model (3), estimating within-study and across-study effects separately									
				Within-study effects						Across-study effects			
	Simulation setting			Sensitivity			Specificity			Sensitivity		Specificity	
No. of studies, m	No. of patients across studies	V_1	V_2	Mean of $\hat{\gamma}_{1W}$	Mean s.e. of $\hat{\gamma}_{1W}$	MSE of $\hat{\gamma}_{1W}$	Mean of $\hat{\gamma}_{0W}$	Mean s.e. of $\hat{\gamma}_{0W}$	MSE of $\hat{\gamma}_{0W}$	Mean of $\hat{\gamma}_{1A}$	Mean s.e. of $\hat{\gamma}_{1A}$	MSE of $\hat{\gamma}_{1A}$	Mean of $\hat{\gamma}_{0A}$	Mean s.e. of $\hat{\gamma}_{0A}$	MSE of $\hat{\gamma}_{0A}$
10	Large	100	25	0.050	0.015	0.0003	−0.050	0.012	0.0002	0.048	0.035	0.0019	−0.049	0.034	0.0015
10	Large	25	25	0.049	0.015	0.0003	−0.050	0.012	0.0001	0.045	0.069	0.0077	−0.047	0.066	0.0061
10	Large	25	100	0.050	0.008	0.00006	−0.050	0.006	0.00004	0.044	0.065	0.0068	−0.044	0.061	0.0052
20	Large	100	25	0.050	0.015	0.0003	−0.050	0.012	0.0001	0.049	0.025	0.0008	−0.050	0.024	0.0006
20	Large	25	25	0.050	0.015	0.0002	−0.050	0.012	0.0001	0.049	0.050	0.0033	−0.049	0.047	0.0025
20	Large	25	100	0.050	0.008	0.00006	−0.050	0.006	0.00004	0.047	0.047	0.0029	−0.047	0.043	0.0022
10	Small	100	25	0.050	0.051	0.0028	−0.051	0.039	0.0015	0.047	0.045	0.0031	−0.049	0.039	0.0020
10	Small	25	25	0.050	0.051	0.0028	−0.052	0.038	0.0016	0.043	0.087	0.0116	−0.047	0.075	0.0079
10	Small	25	100	0.051	0.026	0.0008	−0.051	0.020	0.0004	0.042	0.078	0.0088	−0.045	0.066	0.0058
20	Small	100	25	0.050	0.053	0.0028	−0.054	0.040	0.0017	0.051	0.036	0.0016	−0.052	0.030	0.0011
20	Small	25	25	0.049	0.053	0.0029	−0.052	0.040	0.0012	0.050	0.068	0.0056	−0.051	0.057	0.0037
20	Small	25	100	0.051	0.027	0.0007	−0.052	0.021	0.0005	0.049	0.057	0.0039	−0.049	0.048	0.0026

DOI: 10.1002/sim

Table II. *Continued.*

Simulation setting				Model (3) mixing within-study and across-study effect estimates					
				Amalgamated effects					
				Sensitivity			Specificity		
No. of studies, m	No. of patients across studies	V_1	V_2	Mean of $\hat{\gamma}_{1WA}$	Mean s.e. of $\hat{\gamma}_{1WA}$	MSE of $\hat{\gamma}_{1WA}$	Mean of $\hat{\gamma}_{0WA}$	Mean s.e. of $\hat{\gamma}_{0WA}$	MSE of $\hat{\gamma}_{0WA}$
10	Large	100	25	0.050	0.014	0.0002	−0.050	0.011	0.0001
10	Large	25	25	0.049	0.015	0.0002	−0.050	0.011	0.0001
10	Large	25	100	0.050	0.008	0.00006	−0.050	0.006	0.00004
20	Large	100	25	0.050	0.014	0.0002	−0.050	0.011	0.0001
20	Large	25	25	0.050	0.014	0.0002	−0.050	0.011	0.0001
20	Large	25	100	0.050	0.008	0.00006	−0.050	0.006	0.00003
10	Small	100	25	0.049	0.034	0.0014	−0.051	0.027	0.0009
10	Small	25	25	0.048	0.044	0.0021	−0.052	0.034	0.0013
10	Small	25	100	0.050	0.025	0.0007	−0.051	0.019	0.0004
20	Small	100	25	0.050	0.030	0.0010	−0.053	0.025	0.0007
20	Small	25	25	0.049	0.042	0.0018	−0.052	0.033	0.0011
20	Small	25	100	0.050	0.025	0.0006	−0.052	0.019	0.0004

s.e. = standard error; MSE = mean-square error; V_1 is the variance of the mean covariate value across studies; V_2 is the variance of the patient covariate values within a study; the true within-study accuracy-covariate effect is 0.05 for sensitivity and −0.05 for specificity.

'Large' relates to a total of 3096 patients across studies; 'small' relates to a total of 320 patients across studies.

DOI: 10.1002/sim

The standard error and mean-square error (MSE) are generally much smaller for $\hat{\gamma}_{1W}$ and $\hat{\gamma}_{0W}$ than for $\hat{\gamma}_{1A}$ and $\hat{\gamma}_{0A}$. For example, given $m=10$, a 'large' number of patients, $V_1=25$, and $V_2=25$, the mean standard error and MSE of $\hat{\gamma}_{1W}$ are 0.015 and 0.0003, respectively, compared with 0.069 and 0.077 for $\hat{\gamma}_{1A}$. This highlights that $\hat{\gamma}_{1W}$ and $\hat{\gamma}_{0W}$ generally have the greater power and thus the reason why IPD is usually beneficial. However, the power of $\hat{\gamma}_{1A}$ and $\hat{\gamma}_{0A}$ increases as the number of studies increase, and where V_1 is large relative to V_2, the standard error and MSE of $\hat{\gamma}_{1A}$ and $\hat{\gamma}_{0A}$ are comparable, and in some situations slightly superior to those for $\hat{\gamma}_{1W}$ and $\hat{\gamma}_{0W}$. The standard error and MSE of $\hat{\gamma}_{1W}$ and $\hat{\gamma}_{0W}$ increase as the number of patients decrease and as V_2 decreases.

4.2. Simulation 2: Confounding across studies

One can extend model (3) to adjust for factors that may confound the accuracy-covariate effects of interest. Unfortunately, confounding factors are often unknown or unavailable in the IPD. Unknown patient-level factors may confound observed within-study effects, and similarly the mean of these factors may confound observed across-study effects. Unknown study-level factors may further confound across-study effects. For example, consider the effect of age on diagnostic accuracy when studies use different measuring devices. Within studies patients use the same device, so the within-study accuracy-age effects are not confounded by device. Yet, if the mean age in each study is somehow related to the device used, and the choice of device is itself related to diagnostic accuracy, then the across-study effect of mean age on underlying diagnostic accuracy will be confounded by device. To adjust for this a study-level covariate for device is necessary in the model, but such study-level confounders are often unknown.

4.2.1. Simulation description. To assess effect estimates from model (3) given across-study confounding, we simulated data as in Section 4.1 but made patient test responses also dependent on their study's measurement device (A or B). Studies with a mean patient-level covariate above zero were designated as device A, and studies with a mean covariate of zero or below were designated as device B. For device A, the mean logit-sensitivity and logit-specificity were increased by 1, giving an improved summary sensitivity of 95 per cent and a summary specificity of 80 per cent For each simulated setting, we generated 1000 IPD meta-analysis data sets and to each we fitted model (3), including the within-study and across-study effects for the patient-level covariate, but without a study-level covariate for device. There was thus 'unknown' confounding across studies due to the measurement device.

4.2.2. Simulation results. Results for some of the $m=10$ and $m=20$ simulations are shown in Table III. In each setting, $\hat{\gamma}_{1W}$ and $\hat{\gamma}_{0W}$ are still approximately unbiased estimates of the accuracy-covariate effects, as their mean bias is close to zero. However, due to the confounding, $\hat{\gamma}_{1A}$ and $\hat{\gamma}_{0A}$ are clearly biased in every setting, with means above the true accuracy-covariate effect values of 0.05 and -0.05, even when V_1 is large relative to V_2. For example, given $m=10$, $V_1=100$, and $V_2=25$, and a 'large' number of patients, the mean $\hat{\gamma}_{1A}$ is 0.094 and the mean $\hat{\gamma}_{0A}$ is -0.005. This suggests that an increase in the mean patient-level covariate of 10 increases an initial underlying logit-sensitivity of 2 (sensitivity 88 per cent) to 2.94 (sensitivity 95 per cent), rather than to the correct value of 2.5 (sensitivity 92 per cent), and decreases an initial underlying logit-specificity of 0.4 (specificity 60 per cent) to 0.35 (specificity 59 per cent), rather than to the correct value of -0.1 (specificity 47 per cent). Thus, there is clear ecological bias here, and it is even larger in settings where V_1 is relatively small.

DOI: 10.1002/sim

Table III. Summary of the accuracy-covariate effect estimates from model (3) in the simulations of Section 4.2 (across-study confounding); comparison of $\hat{\gamma}_{1W}$ and $\hat{\gamma}_{0W}$ with $\hat{\gamma}_{1A}$ and $\hat{\gamma}_{0A}$, and $\hat{\gamma}_{1WA}$ and $\hat{\gamma}_{0WA}$.

				Model (3), estimating within-study and across-study effects separately									
				Within-study effects					Across-study effects				
	Simulation setting			Sensitivity			Specificity		Sensitivity			Specificity	
No. of studies, m	No. of patients across studies	V_1	V_2	Mean of $\hat{\gamma}_{1W}$	Mean s.e. of $\hat{\gamma}_{1W}$	MSE of $\hat{\gamma}_{1W}$	Mean of $\hat{\gamma}_{0W}$	Mean s.e. of $\hat{\gamma}_{0W}$	MSE of $\hat{\gamma}_{0W}$	Mean of $\hat{\gamma}_{1A}$	Mean s.e. of $\hat{\gamma}_{1A}$	MSE of $\hat{\gamma}_{1A}$	Mean of $\hat{\gamma}_{0A}$	Mean s.e. of $\hat{\gamma}_{0A}$	MSE of $\hat{\gamma}_{0A}$
10	Large	100	25	0.050	0.018	0.0003	−0.050	0.012	0.0002	0.094	0.040	0.004	−0.005	0.035	0.004
10	Large	25	25	0.049	0.018	0.0004	−0.050	0.012	0.0002	0.133	0.077	0.016	0.039	0.070	0.015
10	Large	25	100	0.050	0.009	0.00008	−0.050	0.006	0.0004	0.127	0.074	0.014	0.035	0.066	0.013
20	Small	100	25	0.053	0.060	0.004	−0.054	0.041	0.002	0.093	0.044	0.004	−0.010	0.031	0.003
20	Small	25	25	0.049	0.061	0.004	−0.049	0.042	0.002	0.128	0.081	0.014	0.025	0.059	0.010
20	Small	25	100	0.051	0.031	0.0009	−0.050	0.022	0.0005	0.105	0.067	0.009	0.005	0.050	0.006

DOI: 10.1002/sim

Table III. *Continued.*

Simulation setting				Model (3) mixing within-study and across-study effect estimates					
				Amalgamated effects					
				Sensitivity			Specificity		
No. of studies, m	No. of patients across studies	V_1	V_2	Mean of $\hat{\gamma}_{1WA}$	Mean s.e. of $\hat{\gamma}_{1WA}$	MSE of $\hat{\gamma}_{1WA}$	Mean of $\hat{\gamma}_{0WA}$	Mean s.e. of $\hat{\gamma}_{0WA}$	MSE of $\hat{\gamma}_{0WA}$
10	Large	100	25	0.060	0.016	0.0004	−0.044	0.012	0.0002
10	Large	25	25	0.055	0.017	0.0004	−0.047	0.012	0.0002
10	Large	25	100	0.052	0.009	0.00009	−0.049	0.006	0.00004
20	Small	100	25	0.080	0.035	0.002	−0.026	0.025	0.001
20	Small	25	25	0.081	0.048	0.003	−0.025	0.035	0.002
20	Small	25	100	0.062	0.029	0.001	−0.042	0.020	0.0005

s.e. = standard error; MSE = mean-square error; V_1 is the variance of the mean covariate value across studies; V_2 is the variance of the patient covariate values within a study; the true within-study accuracy-covariate effect is 0.05 for sensitivity and −0.05 for specificity.
'Large' relates to a total of 3096 patients across studies; 'small' relates to a total of 320 patients across studies.

DOI: 10.1002/sim

4.3. The importance of separating within-study and across-study effects

Some articles (e.g. [48]) propose models akin to model (3), but assume $\gamma_{1W}=\gamma_{1A}$ ($=\gamma_{1WA}$, say) and $\gamma_{0W}=\gamma_{0A}$ ($=\gamma_{0WA}$, say). This makes the strong assumption that across-study effects are unbiased estimates of within-study effects (i.e. no ecological bias exists) and estimation gives $\hat{\gamma}_{1WA}$, an amalgam of $\hat{\gamma}_{1W}$ and $\hat{\gamma}_{1A}$, and $\hat{\gamma}_{0WA}$, an amalgam of $\hat{\gamma}_{0W}$ and $\hat{\gamma}_{0A}$. To assess $\hat{\gamma}_{1WA}$ and $\hat{\gamma}_{0WA}$, we fitted the amalgamated model to each simulated data set from Simulations 1 and 2, and the results are summarized in Tables II and III. When $\hat{\gamma}_{1A}$ and $\hat{\gamma}_{0A}$ are truly unbiased, as in simulation 1, $\hat{\gamma}_{1WA}$ and $\hat{\gamma}_{0WA}$ have a mean standard error and MSE smaller than $\hat{\gamma}_{1W}$ and $\hat{\gamma}_{0W}$ (Table II). The reduction is only slight when the number of patients is 'large', but is greater when the number is 'small'. For example, given $m=10$, a 'small' number of patients, $V_1=100$, and $V_2=25$, the standard error and MSE of $\hat{\gamma}_{1WA}$ are 0.034 and 0.0014, respectively, compared with 0.051 and 0.0028 for $\hat{\gamma}_{1W}$.

However, where $\hat{\gamma}_{1A}$ and $\hat{\gamma}_{0A}$ are biased by across-study confounding (simulation (2)), the amalgamated model gives $\hat{\gamma}_{1WA}$ and $\hat{\gamma}_{0WA}$ that are also biased in those settings, where $\hat{\gamma}_{1A}$ and $\hat{\gamma}_{0A}$ are influential, i.e. when $m=20$ or V_1 is large relative to V_2 (Table III). For example, given $m=20$, $V_1=100$, $V_2=25$, and a 'small' number of patients, the mean $\hat{\gamma}_{1WA}$ is 0.080 and the mean $\hat{\gamma}_{0WA}$ is -0.026, around 50 per cent above the true effect values. Though the standard error of $\hat{\gamma}_{1WA}$ and $\hat{\gamma}_{0WA}$ is often smaller than $\hat{\gamma}_{1W}$ and $\hat{\gamma}_{0W}$, this only arises by utilizing the biased $\hat{\gamma}_{1A}$ and $\hat{\gamma}_{0A}$. It is thus clear that inferences about how patient-level covariates modify the diagnostic accuracy are best based solely on $\hat{\gamma}_{1W}$ and $\hat{\gamma}_{0W}$ if across-study confounding is at all a concern.

5. APPLICATION TO THE TEMPERATURE DATA SET

We now apply some of the models described in Section 3 to the temperature data. We begin by assessing the overall effects, and then study-level and patient-level covariates.

5.1. Summary sensitivity and specificity results

Model (1) was applied to all 23 studies in a frequentist framework with IPD reconstructed in the seven AD studies as described in Section 3.3. This analysis is equivalent to reducing IPD studies to AD and fitting a BRMA to the AD from all 23 studies [9]. The results give a summary sensitivity of 0.71 (95 per cent CI: 0.60 to 0.82) and a summary specificity of 0.96 (95 per cent CI: 0.93 to 0.98) across studies; the SROC is shown in Figure 2. The low summary sensitivity suggests that ear temperature is not an accurate tool for diagnosing children with fever. However, the between-study heterogeneity limits clinical interpretation, as $\hat{\tau}_1^2=1.23$ and $\hat{\tau}_0^2=1.47$. Estimation in a Bayesian framework, using the prior distributions specified in Section 3.1.2, gives posterior results very similar to the frequentist estimates and robust to changes in the 'vague' prior distributions. Comparing the meta-analysis results from all the 23 studies to those from just the 16 IPD studies reveals similar summary estimates and confidence intervals, but smaller between-study variance estimates in the IPD-only analysis (Table IV). We initially thought that three of the AD studies must increase heterogeneity by using cut-off levels other than 38.0°C (Table I); yet excluding these three studies gives even larger between-study variance estimates (Table IV), so other differences between AD and IPD studies must be causing this, such as measurement device.

There were eight different pairs of ear and rectal devices across studies (Table I), and it is important to assess if they cause some of the between-study heterogeneity. Only one study considered the Ototemp Pedi-Q ear device and only one the indwelling Probe rectal device. In both

DOI: 10.1002/sim

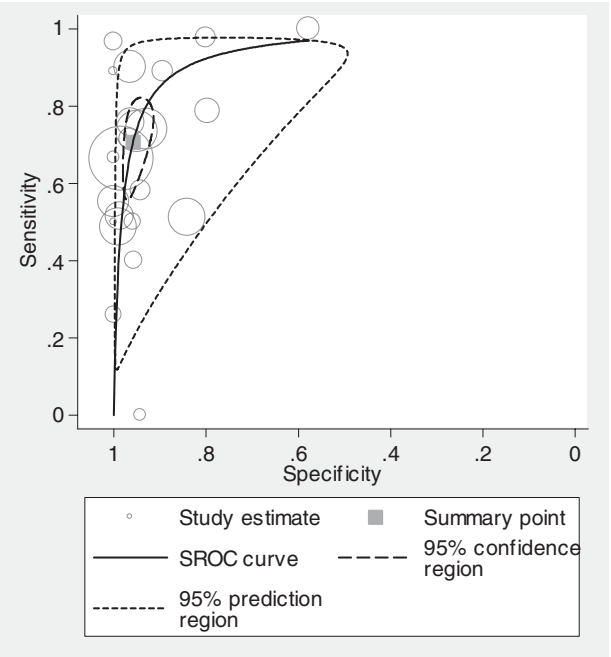

Figure 2. Summary Receiver Operating Curve (SROC) obtained from the BRMA of model (1) for the 23 temperature studies. The size of each circle is proportional to the number of patients in the study.

these there was a zero cell in the 2 by 2 table, so to ensure their inclusion in the following analysis a continuity correction of 0.5 was added to each cell in these studies. Model (1) was extended to include a categorical study-level covariate denoting device pair, and was fitted to all 23 studies in a frequentist framework. This gave $\hat{\tau}_1^2 = 0.49$ and $\hat{\tau}_0^2 = 0.54$ (Table IV), a significant reduction of 60 and 63 per cent, respectively, from model (1) ($\Delta - 2\text{LL} = 27.3$, $p = 0.02$ on 14 degrees of freedom). Summary sensitivity and specificity estimates vary considerably across device pairs (Table IV), but their wide 95 per cent confidence intervals make it impossible to define the most appropriate pair. Further research clearly must establish a consensus for which devices should be used in practice.

5.2. The potential impact of within-study overdispersion

To assess overdispersion within-studies, model (2) was fitted in a frequentist framework, but convergence was difficult, even when σ_{1i}^2 and σ_{0i}^2 were assumed common across studies (see Section 3.1.3). Estimation was possible within a Bayesian framework, though the posterior estimates were strongly influenced by the prior distributions for σ_{1i}^2 and σ_{0i}^2 (Table IV). A range of prior distributions for σ_{1i}^2 and σ_{0i}^2 was thus used, and it was observed that the larger σ_{1i}^2 and σ_{0i}^2, the greater meta-analysis results from model (2) differed to those from model (1). For example, consider a Uniform(0, 1) prior for σ_{1i} and σ_{0i}, which at the upper extreme infers that in a study with an underlying logit-sensitivity of 0.88 (sensitivity = 70 per cent) an individual's logit-sensitivity can approximately vary between -1.12 and 2.88 (sensitivity between 25 and 95 per cent). Focusing on the sensitivity results, this leads to a posterior median estimate of $\hat{\tau}_1^2 = 1.40$, compared with

DOI: 10.1002/sim

Table IV. Application of models (1) and (2) to the temperature studies.

Model	Included studies	Estimation method	Ear temperature device	Rectal temperature device	Mean logit-sensitivity $\hat{\mu}_1$ (s.e.)	Mean logit-specificity $\hat{\mu}_0$ (s.e.)	Summary sensitivity [95 per cent interval]	Summary specificity [95 per cent interval]	Between-study variance for logit-sensitivity τ_1^2	Between-study variance for logit-specificity τ_0^2	Between-study correlation ρ
Model (1)	All	Frequentist	—	—	0.88 (0.26)	3.14 (0.31)	0.71 [0.60, 0.81]	0.96 [0.93, 0.98]	1.23	1.47	−0.63
Model (1)	20 studies with cut-off level of 38.0°C	Frequentist	—	—	0.88 (0.28)	3.34 (0.35)	0.71 [0.57, 0.84]	0.97 [0.94, 0.99]	1.32	1.64	−0.77
Model (1)	16 IPD studies	Frequentist	—	—	0.67 (0.24)	3.26 (0.31)	0.66 [0.56, 0.77]	0.96 [0.94, 0.99]	0.71	0.86	−0.80
Model (2)	All	Bayesian using a Uniform(0, 0.1) prior for σ_{1i} and σ_{0i}	—	—	0.89 (0.27)	3.31 (0.35)	0.71 [0.60, 0.81]	0.96 [0.93.0.98]	1.19	1.74	−0.53
Model (2)	All	Bayesian using a Uniform(0, 1) prior for σ_{1i} and σ_{0i}	—	—	0.94 (0.27)	3.50 (0.36)	0.71 [0.59, 0.81]	0.97 [0.94, 0.99]	1.40	1.99	−0.57
Model (1) but extended to include device pair as a categorical study-level covariate	All	Frequentist	Thermoscan	Electronic	3.84 (1.24)	1.43 (0.81)	0.98 [0.93, 1.00]	0.81 [0.56, 1.00]	0.49	0.54	−0.22
			Thermoscan	Mercury	1.09 (0.52)	3.11 (0.60)	0.75 [0.56, 0.94]	0.96 [0.91, 1.00]			
			Core-Check	Electronic	−0.23 (0.67)	2.91 (0.74)	0.44 [0.12. 0.77]	0.95 [0.88, 1.00]			
			Core-Check	Probe	0.04 (1.39)	3.47 (1.64)	0.51 [0, 1.00]	0.97 [0.88, 1.00]			
			FirstTemp	Electronic	0.54 (0.23)	3.68 (0.33)	0.63 [0.53, 0.74]	0.98 [0.96, 0.99]			
			FirstTemp	Mercury	2.23 (0.81)	2.75 (0.58)	0.90 [0.76, 1.00]	0.94 [0.88, 1.00]			
			LightTouch Pedi-Q	Electronic	0.47 (0.55)	2.11 (0.61)	0.62 [0.38, 0.87]	0.89 [0.78, 1.00]			
			Ototemp Pedi-Q	Mercury	3.68 (1.63)	0.29 (0.76)	0.98 [0.90, 1.00]	0.57 [0.21, 0.94]			

IPD = individual patient data; s.e. = standard error; The Bayesian posterior estimates are taken from a 20 000 sample, following a 20 000 burn-in period.

DOI: 10.1002/sim

$\hat{\tau}_1^2 = 1.23$ from model (1), and a posterior mean logit-sensitivity estimate of 0.94 (S.E. 0.27), compared with 0.88 (S.E. 0.26) from model (1). However, on the original sensitivity scale, the difference is negligible, with both models giving a summary sensitivity of 71 per cent across all studies with a very similar 95 per cent interval (Table 4). Thus, where interest lies only in the summary estimates of sensitivity and specificity across studies, model (1) appears to be a suitable approximation to model (2) even when large overdispersion exists within studies. Whether this is true in other data sets is an issue for further research.

5.3. Assessment of the accuracy-infant effect

It is of clinical interest whether diagnostic accuracy of ear temperature is different for infants (<1 year of age) and non-infants, as alternatives to rectal temperature devices are perhaps more important for non-infants, due to their increased emotional awareness. Eleven IPD studies provided age, which allow an assessment of the accuracy-infant effect both within studies and across studies. The other 12 studies did not provide the proportion of infants in each disease group, so could not even contribute to the across-study effects and are thus excluded in the following analyses. Model (5) was fitted in a frequentist framework to the 11 IPD studies, with a study-level covariate for device pair and with age as a binary patient-level covariate, such that x_{ik} and x_{ij} were one for infants and zero for non-infants, and $\overline{x}_{1i}$ and $\overline{x}_{0i}$ were the proportion of infants in the diseased and non-diseased groups. The within-study effect estimates do not provide significant evidence that either the sensitivity-infant effect ($\hat{\gamma}_{1W} = 0.10$, s.e. $(\hat{\gamma}_{1W}) = 0.18$, $p = 0.57$) or the specificity-infant effect ($\hat{\gamma}_{0W} = 0.12$, s.e. $(\hat{\gamma}_{0W}) = 0.36$, $p = 0.74$) are important. The across-study effect estimates are very different to the within-study effects; $\hat{\gamma}_{1A}$ is equal to -3.81 (s.e. $(\hat{\gamma}_{1A}) = 1.32$, $p = 0.004$) and $\hat{\gamma}_{0A}$ is equal to -1.36 (s.e.$(\hat{\gamma}_{0A}) = 2.64$, $p = 0.60$). These estimates are very large and the former provides significant evidence that underlying sensitivity decreases in studies with a large proportion of infants.

These results emphasize the benefit of IPD and the importance of separating within-study and across-study effects. Firstly, the IPD enables $\hat{\gamma}_{1W}$ and $\hat{\gamma}_{0W}$ to be estimated, and these have standard errors about 86 per cent smaller than $\hat{\gamma}_{1A}$ and $\hat{\gamma}_{0A}$; this is due to V_1 being smaller than V_2 for the infant covariate, as across studies the proportion of infants varies between 0 and 0.69, whereas within studies the infant status usually varies from 0 to 1. In such situations, the across-study effects have low power (Section 4.1) and so it is important to obtain IPD. Secondly, it is a concern that across-study confounding may be affecting $\hat{\gamma}_{1A}$ and $\hat{\gamma}_{0A}$, as they are in the opposite direction to $\hat{\gamma}_{1W}$ and $\hat{\gamma}_{0W}$, and much larger. For example, $\hat{\gamma}_{1W}$ infers that logit-sensitivity increases by 0.10 for infants compared with non-infants, such that when non-infants have a summary sensitivity of 70 per cent then infants have a summary sensitivity of 72 per cent. However, $\hat{\gamma}_{1A}$ infers that studies with only infants have an underlying logit-sensitivity decreased by 3.81 compared with studies with only non-infants, such that if non-infant studies have an underlying sensitivity of 70 per cent then infant studies have an underlying sensitivity of just 5 per cent. Thus, had IPD not been available here, such that $\hat{\gamma}_{1W}$ and $\hat{\gamma}_{0W}$ were unobtainable, using $\hat{\gamma}_{1A}$ and $\hat{\gamma}_{0A}$ to make inferences about the effect of being an infant on diagnostic accuracy would have led to very different conclusions.

6. DISCUSSION

Dinnes *et al*. [14] state that 'the use of IPD meta-analysis in diagnostic test accuracy reviews should be explored to allow heterogeneity to be considered in more detail'. In this paper, we have

DOI: 10.1002/sim

described models for a BRMA of diagnostic studies for when IPD, or a mixture of IPD and AD are available. We have shown how IPD can be used to examine study-level covariates, to help explain the between-study heterogeneity, and patient-level covariates, to assess the effect of patient characteristics on test accuracy. This facilitates diagnostic accuracy results tailored to the individual patient, thus allowing meta-analysis to inform individual diagnostic strategies. Our analysis of the temperature data illustrates this, and we identified that accuracy of fever diagnosis depends on the measurement devices used, but there is no evidence it depends on a child's infant status. Our work focused on how covariates affect sensitivity, specificity, and between-study heterogeneity and correlation. Other informative results may also be assessed in a BRMA, like the diagnostic odds ratio, the positive and negative likelihood ratio tests, and the SROC [7].

A key component of our work is how an IPD meta-analysis can estimate the effect of patient-level covariates on diagnostic accuracy. If patient-level covariates truly modify diagnostic accuracy, the BRMA of model (1) is misspecified as it assumes a fixed logit-sensitivity and logit-specificity across patients within each study. One must rather account for within-study overdispersion by placing random effects that allow diagnostic accuracy to vary across patients (model (2)). An even better approach is to examine those patient-level covariates that may influence diagnostic accuracy. In this regard, model (3) importantly shows how to separate within-study and across-study effects. The former is more clinically meaningful, as they relate patient characteristics to individual test accuracy. They are only observable from studies providing IPD including patient-level covariates or, in the special case of a categorical patient-level covariate, studies reporting a separate 2 by 2 diagnostic accuracy table for each category. In other situations only the across-study effects may be observable, that is the relationship across studies between aggregated patient-level covariates (e.g. mean age) and diagnostic accuracy. Within-study effects usually have more power than across-study effects, and are less prone to confounding (Section 4), but without them a decision is needed whether the observed across-study effects can be assumed to reflect the true within-study relationships.

Section 4.1 showed that in ideal conditions the across-study effects are close to unbiased estimates of the true within-study effect, with a slight attenuation toward zero when the number of studies is small or there is only a small variation across studies in the covariate mean. They are most powerful when the number of studies is large and also when the variation in covariate means across studies is large; in such scenarios they may even out perform the within-study effect estimates from IPD (Section 4.1). This is consistent with previous work [23, 50], and a Q-statistic is available to help detect such situations [50]. However, it is known that across-study effects may not reflect within-study effects [20], as they are subject to ecological bias and confounding [21]. This was exemplified in Section 4.2, where simulated across-study confounding caused across-study effect estimates to be biased, and there may be other reasons why ecological bias may occur [22]. Such bias was also evident in the temperature analysis when the infant covariate was assessed. For this reason, practitioners are recommended to be cautious when interpreting across-study effects, and generally within-study effects should be preferred [19]. We acknowledge, however, that sometimes there is little or no IPD, or perhaps only small variation of covariate values within IPD studies, such that the across-study effects may then be the major source of information available. In such situations, before interpreting the across-study effects, due thought must be given to whether ecological bias may exist and whether all potential study-level confounders have been adjusted for.

Another important aspect of our work is how to combine IPD and AD studies (Section 3.3), which is important as IPD are not always available [25]. Where only study-level parameters are

DOI: 10.1002/sim

modelled, the AD and IPD studies can be combined reasonably easily by converting the AD to IPD so to use models (1) or (2). Where patient-level covariates are to be assessed, more complex modelling is required, as usually only IPD studies provide the necessary patient-level covariate. Model (6) builds on previous work [26, 33, 34, 48, 51] and allows all studies to estimate the summary sensitivity and summary specificity across studies, the impact of study-level covariates, and also the across-study effects; however, only IPD studies estimate the within-study effects. Some practitioners may argue that IPD is more reliable than AD, as one can assess the quality of data and ensure within-study analyses are done correctly, such that including AD alongside IPD may lead to bias. With this in mind, one may wish to assess the sensitivity of meta-analysis results to the inclusion of AD studies, and explore any differences between IPD and AD studies.

Further research. Models (3–5) utilize $\overline{x}_{1i}$ and $\overline{x}_{0i}$, the covariate means of the study sample, but ideally they should be replaced with $\overline{X}_{1i}$ and $\overline{X}_{0i}$, the true covariate means of the study population. In this context, $\overline{x}_{1i}$ and $\overline{x}_{0i}$ are actually estimates, which more precisely reflect the underlying population means when the study size is large. Models (3–5) do not account for this measurement error, and further research should consider the impact this has and suitable model extensions. It is conceivable that such measurement error may have only a little impact, as small studies already have less contribution to the likelihood than large studies; however, this issue needs to be examined in detail. Note that the need to separate within-study and across-study effects remains a potentially pertinent issue regardless of the measurement error concern of $\overline{x}_{1i}$ and $\overline{x}_{0i}$; Begg and Parides [46] state that concerns about measurement error should not take precedence over accounting for individual-level and cluster-level effects.

Our models assume that random effects are normally distributed, which is a common meta-analysis assumption, but this may not always be appropriate. Proc NLMIXED in SAS and xtmelogit in STATA currently only allow a normal distribution for the random effects, but Lee and Thompson [53] show that alternative distributions can be specified in a Bayesian framework. Some of our models were more easily fitted within a Bayesian approach, but one must be cautious about the impact of 'vague' prior distributions as they may be undesirably influential, especially when the number of studies is small [43, 44, 54]. Other extensions to our work may include the use of IPD to evaluate multiple diagnostic tests [2, 5], and the development of IPD models to pool ROC curves from each study while assessing study-level and patient-level covariates [37]. In both these situations, one would need to account for the correlation between multiple responses from the same patient. It would also be interesting to translate our BRMA models to the alternative but related HSROC framework [4, 5]. Harbord *et al.* [8] describe when the bivariate model including study-level covariates is equivalent to the HSROC approach, and further consideration of this issue in relation to patient-level covariates is needed.

ACKNOWLEDGEMENTS

While undertaking this work, Richard Riley was funded as a Research Scientist in Evidence Synthesis by the Department of Health's National Coordinating Centre for Research Capacity Development. We would like to thank Roger Harbord for helpful discussions and making available his STATA program for fitting a bivariate meta-analysis of diagnostic studies. We also thank the Editor, the Associate Editor and two reviewers, whose comments have greatly improved the paper. We are also extremely grateful to those researchers who provided IPD for the meta-analysis of ear temperature studies.

DOI: 10.1002/sim

REFERENCES

1. Deeks JJ. Systematic reviews in health care: systematic reviews of evaluations of diagnostic and screening tests. *British Medical Journal* 2001; **323**:157–162.
2. Irwig L, Macaskill P, Glasziou P, Fahey M. Meta-analytic methods for diagnostic test accuracy. *Journal of Clinical Epidemiology* 1995; **48**:119–130; discussion 131–112.
3. Moses LE, Shapiro D, Littenberg B. Combining independent studies of a diagnostic test into a summary ROC curve: data-analytic approaches and some additional considerations. *Statistics in Medicine* 1993; **12**: 1293–1316.
4. Rutter CM, Gatsonis CA. A hierarchical regression approach to meta-analysis of diagnostic test accuracy evaluations. *Statistics in Medicine* 2001; **20**:2865–2884.
5. Macaskill P. Empirical Bayes estimates generated in a hierarchical summary ROC analysis agreed closely with those of a full Bayesian analysis. *Journal of Clinical Epidemiology* 2004; **57**:925–932.
6. Hellmich M, Abrams KR, Sutton AJ. Bayesian approaches to meta-analysis of ROC curves. *Medical Decision Making* 1999; **19**:252–264.
7. Reitsma JB, Glas AS, Rutjes AW, Scholten RJ, Bossuyt PM, Zwinderman AH. Bivariate analysis of sensitivity and specificity produces informative summary measures in diagnostic reviews. *Journal of Clinical Epidemiology* 2005; **58**:982–990.
8. Harbord RM, Deeks JJ, Egger M, Whiting P, Sterne JA. A unification of models for meta-analysis of diagnostic accuracy studies. *Biostatistics* 2007; **8**:239–251.
9. Chu H, Cole SR. Bivariate meta-analysis of sensitivity and specificity with sparse data: a generalized linear mixed model approach. *Journal of Clinical Epidemiology* 2006; **59**:1331–1332.
10. Riley RD, Abrams KR, Sutton AJ, Lambert PC, Thompson JR. Bivariate random-effects meta-analysis and the estimation of between-study correlation. *BMC Medical Research Methodology* 2007; **7**:3.
11. Zwinderman AH, Bossuyt PM. We should not pool diagnostic likelihood ratios in systematic reviews. *Statistics in Medicine* 2008; **27**:687–697.
12. Hamza TH, Van Houwelingen HC, Stijnen T. The binomial distribution of meta-analysis was preferred to model within-study variability. *Journal of Clinical Epidemiology* 2008; **61**:41–51.
13. Lijmer JG, Bossuyt PM, Heisterkamp SH. Exploring sources of heterogeneity in systematic reviews of diagnostic tests. *Statistics in Medicine* 2002; **21**:1525–1537.
14. Dinnes J, Deeks J, Kirby J, Roderick P. A methodological review of how heterogeneity has been examined in systematic reviews of diagnostic test accuracy. *Health Technology Assessment* 2005; **9**:1–113, iii.
15. Whiting P, Rutjes AW, Reitsma JB, Glas AS, Bossuyt PM, Kleijnen J. Sources of variation and bias in studies of diagnostic accuracy: a systematic review. *Annals of Internal Medicine* 2004; **140**:189–202.
16. Thompson SG. Why sources of heterogeneity in meta-analysis should be investigated. *British Medical Journal* 1994; **309**:1351–1355.
17. Berkey CS, Hoaglin DC, Mosteller F, Colditz GA. A random-effects regression model for meta-analysis. *Statistics in Medicine* 1995; **14**:395–411.
18. Neuhaus JM, Kalbfleisch JD, Hauck WW. A comparison of cluster-specific and population-averaged approaches for analyzing correlated binary data. *International Statistical Review* 1991; **59**:25–35.
19. Thompson SG, Higgins JP. Treating individuals 4: can meta-analysis help target interventions at individuals most likely to benefit? *Lancet* 2005; **365**:341–346.
20. Lau J, Ioannidis JP, Schmid CH. Summing up evidence: one answer is not always enough. *The Lancet* 1998; **351**: 123–127.
21. Berlin JA, Santanna J, Schmid CH, Szczech LA, Feldman HI. Individual patient-versus group-level data meta-regressions for the investigation of treatment effect modifiers: ecological bias rears its ugly head. *Statistics in Medicine* 2002; **21**:371–387.
22. Greenland S, Morgenstern H. Ecological bias, confounding, and effect modification. *International Journal of Epidemiology* 1989; **18**:269–274.
23. Lambert PC, Sutton AJ, Abrams KR, Jones DR. A comparison of summary patient-level covariates in meta-regression with individual patient data meta-analysis. *Journal of Clinical Epidemiology* 2002; **55**:86–94.
24. Khan KS, Bachmann LM, ter Riet G. Systematic reviews with individual patient data meta-analysis to evaluate diagnostic tests. *European Journal of Obstetrics and Gynecology and Reproductive Biology* 2003; **108**: 121–125.

DOI: 10.1002/sim

25. Riley RD, Simmonds MC, Look MP. Evidence synthesis combining individual patient data and aggregate data: a systematic review identified current practice and possible methods. *Journal of Clinical Epidemiology* 2007; **60**:431–439.
26. Riley RD, Lambert PC, Staessen JA, Wang J, Gueyffier F, Thijs L, Boutitie F. Meta-analysis of continuous outcomes combining individual patient data and aggregate data. *Statistics in Medicine* 2008; **27**:1870–1893.
27. Craig JV, Lancaster GA, Taylor S, Williamson PR, Smyth RL. Infrared ear thermometry compared with rectal thermometry in children: a systematic review. *The Lancet* 2002; **360**:603–609.
28. Williamson PR, Lancaster GA, Craig JV, Smyth RL. Meta-analysis of method comparison studies. *Statistics in Medicine* 2002; **21**:2013–2025.
29. Dodd SR, Lancaster GA, Craig JV, Smyth RL, Williamson PR. In a systematic review, infrared ear thermometry for fever diagnosis in children finds poor sensitivity. *Journal of Clinical Epidemiology* 2006; **59**:354–357.
30. http://www.nhsdirect.nhs.uk/articles/article.aspx?articleId=1633.
31. Committee of Cincinnati Children's Hospital Medical Center: Evidence based clinical practice guideline for fever of uncertain source in children 2 to 36 months of age. *Cincinatti Children's Hospital Medical Center*. Available from: http//wwwguidelinesgov, 2003.
32. Turner RM, Omar RZ, Yang M, Goldstein H, Thompson SG. A multilevel model framework for meta-analysis of clinical trials with binary outcomes. *Statistics in Medicine* 2000; **19**:3417–3432.
33. Schmid CH, Stark PC, Berlin JA, Landais P, Lau J. Meta-regression detected associations between heterogeneous treatment effects and study-level, but not patient-level, factors. *Journal of Clinical Epidemiology* 2004; **57**: 683–697.
34. Simmonds MC. Statistical methodology of individual patient data. *Ph.D. Thesis*, University of Cambridge, 2005.
35. Tudur-Smith C, Williamson PR, Marson AG. Investigating heterogeneity in an individual patient data meta-analysis of time to event outcomes. *Statistics in Medicine* 2005; **24**:1307–1319.
36. Higgins JP, Whitehead A, Turner RM, Omar RZ, Thompson SG. Meta-analysis of continuous outcome data from individual patients. *Statistics in Medicine* 2001; **20**:2219–2241.
37. Kester AD, Buntinx F. Meta-analysis of ROC curves. *Medical Decision Making* 2000; **20**:430–439.
38. Zeger SL, Liang KY, Albert PS. Models for longitudinal data: a generalized estimating equation approach. *Biometrics* 1988; **44**:1049–1060.
39. Pinheiro JC, Bates DM. Approximations to the log-likelihood function in the nonlinear mixed-effects model. *Journal of Computational and Graphical Statistics* 1995; **4**:12–35.
40. StataCorp. *Statistical Software*: *Release 10.0.* Stata Corporation, College Station, TX, 2007.
41. Van Houwelingen HC, Arends LR, Stijnen T. Advanced methods in meta-analysis: multivariate approach and meta-regression. *Statistics in Medicine* 2002; **21**:589–624.
42. Lunn DJ, Thomas A, Best N, Spiegelhalter D. WinBUG—a Bayesian modelling framework: concepts, structure, and extensibility. *Statistics and Computing* 2000; **10**:325–337.
43. Lambert PC, Sutton AJ, Burton PR, Abrams KR, Jones DR. How vague is vague? A simulation study of the impact of the use of vague prior distributions in MCMC. *Statistics in Medicine* 2005; **24**:2401–2428.
44. Browne WJ, Draper D. Implementation and performance issues in the Bayesian and likelihood fitting of multilevel models. *Computational Statistics* 2000; **15**:391–420.
45. Neuhaus JM, Kalbfleisch JD. Between- and within-cluster covariate effects in the analysis of clustered data. *Biometrics* 1998; **54**:638–645.
46. Begg MD, Parides MK. Separation of individual-level and cluster-level covariate effects in regression analysis of correlated data. *Statistics in Medicine* 2003; **22**:2591–2602.
47. Berkey CS, Hoaglin DC, Antczak-Bouckoms A, Mosteller F, Colditz GA. Meta-analysis of multiple outcomes by regression with random effects. *Statistics in Medicine* 1998; **17**:2537–2550.
48. Sutton AJ, Kendrick D, Coupland CA. Meta-analysis of individual- and aggregate-level data. *Statistics in Medicine* 2008; **27**:651–669.
49. Jackson C, Best N, Richardson S. Hierarchical related regression for combining aggregate and individual data in studies of socio-economic disease risk factors. *Journal of the Royal Statistical Society*, *Series A* 2008; **171**:159–178.
50. Simmonds MC, Higgins JP. Covariate heterogeneity in meta-analysis: criteria for deciding between meta-regression and individual patient data. *Statistics in Medicine* 2007; **26**:2982–2999.
51. Jackson C, Best N, Richardson S. Improving ecological inference using individual-level data. *Statistics in Medicine* 2006; **25**:2136–2159.

DOI: 10.1002/sim

52. Carroll RJ, Stefanski LA. Measurement error, instrumental variables and corrections for attenuation with applications to meta-analyses. *Statistics in Medicine* 1994; **13**:1265–1282.
53. Lee KJ, Thompson SG. Flexible parametric models for random-effects distributions. *Statistics in Medicine* 2008; **27**:418–434.
54. Gelman A. Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis* 2006; **1**:515–534.

DOI: 10.1002/sim