Research Article

Statistics in Medicine

Received 6 February 2009, Accepted 21 December 2009 Published online 25 January 2010 in Wiley Interscience

(www.interscience.wiley.com) DOI: 10.1002/sim.3858

Bayesian bivariate meta-analysis of diagnostic test studies using integrated nested Laplace approximations

M. Paul,[a]*† A. Riebler,[a] L. M. Bachmann,[b] H. Rue[c] and L. Held[a]

For bivariate meta-analysis of diagnostic studies, likelihood approaches are very popular. However, they often run into numerical problems with possible non-convergence. In addition, the construction of confidence intervals is controversial. Bayesian methods based on Markov chain Monte Carlo (MCMC) sampling could be used, but are often difficult to implement, and require long running times and diagnostic convergence checks. Recently, a new Bayesian deterministic inference approach for latent Gaussian models using integrated nested Laplace approximations (INLA) has been proposed. With this approach MCMC sampling becomes redundant as the posterior marginal distributions are directly and accurately approximated. By means of a real data set we investigate the influence of the prior information provided and compare the results obtained by INLA, MCMC, and the maximum likelihood procedure `SAS PROC NLMIXED`. Using a simulation study we further extend the comparison of INLA and `SAS PROC NLMIXED` by assessing their performance in terms of bias, mean-squared error, coverage probability, and convergence rate. The results indicate that INLA is more stable and gives generally better coverage probabilities for the pooled estimates and less biased estimates of variance parameters. The user-friendliness of INLA is demonstrated by documented R-code. Copyright © 2010 John Wiley & Sons, Ltd.

Keywords: integrated nested Laplace approximation (INLA); bivariate random effects model; meta-analysis; diagnostic studies; Bayesian methods

1. Introduction

Meta-analyses are used to summarise the results of several separately performed studies. In this paper we consider meta-analyses of diagnostic studies. The majority of these studies report the number of true positives, false positives, true negatives, and false negatives from which pairs of sensitivity and specificity follow immediately. Often sensitivity and specificity are converted into a single measure, e.g. the diagnostic odds ratio [1]. However, using this approach it is not clear how to directly derive summary estimates of sensitivity and specificity. An alternative is to use two univariate analyses. Thereby, a possible correlation between these two measures is ignored [2]. However, sensitivity and specificity tend to be negatively correlated because test thresholds are likely to vary between studies [3–5]. A bivariate random effects model preserves the two-dimensionality of the underlying data [4–9]. There has been much literature on whether the within-study variability of sensitivity and specificity in the bivariate meta-regression model should be modelled using an approximate normal or an exact binomial distribution. An advantage of using a normal likelihood is that the model is less complex, and fitting becomes easier by using a linear mixed model instead of a generalized linear mixed model (GLMM). However, Chu and Cole [8] and Hamza *et al.* [5] showed through simulation studies that the exact binomial likelihood model performed better in terms of bias, mean-squared error (MSE) and coverage. In particular, in the case of only few studies in the meta-analysis or sparse data leading to zero cells in the contingency tables the normal approximation is not adequate. Additionally, the binomial approach does not require a continuity correction.

Mainly likelihood approaches are used for model fitting. One drawback of maximum likelihood (ML) is that the construction of the usual Wald-type confidence intervals is problematic if the required standard errors are underestimated, which is generally the case for small or moderate sample sizes [10]. A profile likelihood approach could help but is not easy to implement for complex models such as the bivariate random effects model. Standard software such as the `PROC NLMIXED` procedure from the SAS

[a]*Biostatistics Unit, Institute of Social and Preventive Medicine, University of Zurich, Switzerland*
[b]*Horten Centre, University of Zurich, Switzerland*
[c]*Department of Mathematical Sciences, Norwegian University of Science and Technology, Trondheim, Norway*
**Correspondence to: M. Paul, Biostatistics Unit, Institute of Social and Preventive Medicine, University of Zurich, Hirschengraben 84, CH-8001 Zurich, Switzerland.*
†*E-mail: michaela.paul@ifspm.uzh.ch*

package [11] does not provide an option to calculate profile likelihood confidence intervals, hence additional programming by the user is necessary [2, 6]. Alternatively, confidence intervals could also be constructed by means of a bootstrap approach. However, this is only feasible if the number of studies in the meta-analysis is sufficiently large [12] and would also require programming by the user. Another problem is that in complex problems numerical maximization algorithms might fail. Bayesian approaches using Markov chain Monte Carlo (MCMC) can be advantageous, especially in applications with few studies [9]. Credible intervals are easily obtained as quantiles of the generated samples. For the univariate scenario Warn *et al.* [1] applied Bayesian methods to random effects meta-analysis of clinical trials. MCMC based inference for bivariate diagnostic meta-analysis using `WinBUGS` is discussed e.g. by Riley *et al.* [13] and Zwinderman and Bossuyt [12].

Recently, a new approximate Bayesian inference approach using integrated nested Laplace approximations (INLA) was proposed by Rue *et al.* [14]. INLA can be a fast, deterministic alternative to MCMC as it directly computes very accurate approximations to the posterior marginal distributions. Convergence diagnostics are not necessary anymore and the results are immediately available as no sampling is needed. Martino and Rue [15] provide freely available software written in *C* which is easy to use via an interface from R [16].

In this paper, we compare `PROC NLMIXED` and `inla` for the analysis of bivariate meta-analyses of diagnostic studies. In Section 2 we first describe the bivariate meta-analysis model based on a binomial likelihood and then discuss the estimation of this model within a frequentist framework using `PROC NLMIXED` and within a Bayesian framework using `inla`. In Section 3 we consider a meta-analysis including 10 studies on the diagnosis of bladder cancer based on the telomerase marker reviewed by Glas *et al.* [17]. For this data set, estimation problems were reported by Riley *et al.* [18] when using the ML procedure `PROC NLMIXED`. We re-analyzed this data set in a fully Bayesian framework with `inla` and also for comparison with MCMC. The corresponding R-code for the analysis with `inla` is documented in the Appendix. In contrast to `PROC NLMIXED`, neither of the Bayesian approaches show an estimation problem and both yield comparable results. We also investigated the influence of the prior distributions in this specific example in a sensitivity analysis. An extensive simulation study is carried out in Section 4 to further investigate the influence of prior distributions on the Bayesian approach and to assess the performance of `inla` and `PROC NLMIXED`. We vary the number of studies per meta-analysis, the overall sensitivity and specificity, the between-study variances, and the correlation between logit sensitivity and logit specificity. The number of participants is sampled for each study separately. Bias, the precision of estimates, MSE, and coverage probabilities are used as comparative measures. The results indicate that the Bayesian approach using INLA is advantageous. Although there are only small differences concerning bias and MSE in sensitivity and specificity, `inla` reaches better coverage for these parameters, produces in part substantially less biased estimates of variance parameters and is also more stable than `PROC NLMIXED`. We end with a discussion in Section 5.

2. The bivariate model

The bivariate model is used in the meta-analysis of diagnostic studies that report pairs of sensitivity and specificity [7]. Preserving the bivariate structure of the data, pairs of sensitivity $\mathrm{Se}=\mathrm{TP}/(\mathrm{TP}+\mathrm{FN})$ and specificity $\mathrm{Sp}=\mathrm{TN}/(\mathrm{TN}+\mathrm{FP})$ are jointly analyzed. Hereby, TP, FN, TN and FP represent the number of true positives, false negatives, true negatives, and false positives, respectively. Any existing correlation between these two measures is taken into account via random effects. As proposed by Chu and Cole [8] and Hamza *et al.* [5] we use the GLMM approach based on a binomial likelihood:

$$\begin{aligned}
&\mathrm{TP}_i|\mathrm{Se}_i \sim \mathrm{Binomial}(\mathrm{TP}_i+\mathrm{FN}_i, \mathrm{Se}_i), \quad \mathrm{logit}(\mathrm{Se}_i)=\mu+\mathbf{U}_i\boldsymbol{\alpha}+\phi_i,\\
&\mathrm{TN}_i|\mathrm{Sp}_i \sim \mathrm{Binomial}(\mathrm{TN}_i+\mathrm{FP}_i, \mathrm{Sp}_i), \quad \mathrm{logit}(\mathrm{Sp}_i)=\nu+\mathbf{V}_i\boldsymbol{\beta}+\psi_i,\\
&\begin{pmatrix}\phi_i\\ \psi_i\end{pmatrix} \sim \mathcal{N}\left[\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}\sigma_\phi^2 & \rho\sigma_\phi\sigma_\psi\\ \rho\sigma_\phi\sigma_\psi & \sigma_\psi^2\end{pmatrix}\right].
\end{aligned} \tag{1}$$

Here $\mathrm{logit}(p)=\log(p/(1-p))$ is the logit function, μ, ν are intercepts for $\mathrm{logit}(\mathrm{Se}_i)$ and $\mathrm{logit}(\mathrm{Sp}_i)$, respectively, and $\mathbf{U}_i, \mathbf{V}_i$ are vectors of possibly available covariates related to Se_i and Sp_i. The index $i=1,\ldots,I$ represents study i in the meta-analysis. Note that the covariance matrix of the random effects ϕ_i and ψ_i in (1) is parameterized in terms of the between-study variances σ_ϕ^2, σ_ψ^2 and the correlation ρ. An equivalent and, especially in Bayesian contexts, frequently used parameterization uses the precisions $\tau_\phi=1/\sigma_\phi^2$ and $\tau_\psi=1/\sigma_\psi^2$.

Mainly ML methods are applied to estimate the parameters of the bivariate model (1). One often used statistical package in this context is `SAS PROC NLMIXED`, see Chu and Cole [8], Arends *et al.* [9], and Hamza *et al.* [5]. Alternatively model (1) can be fitted by a Bayesian approach that incorporates prior beliefs into the analysis. Especially, the uncertainty in estimates of the hyperparameters $\sigma_\phi^2, \sigma_\psi^2$, and ρ is taken into account. Prior distributions have to be specified for the three hyperparameters as well as for $\mu, \nu, \boldsymbol{\alpha}, \boldsymbol{\beta}$. For the latter, typically a locally uniform prior distribution is chosen.

2.1. *Frequentist analysis with* `PROC NLMIXED`

`PROC NLMIXED` fits the GLMM (1) by maximizing an approximation to the likelihood integrated over the random effects as analytical integration is not possible. By default the adaptive Gaussian quadrature is used to approximate the integrated likelihood by a weighted sum of the integrand at predefined abscissas for the random effects [19]. The number of quadrature points can

thereby either be defined by the user or automatically by the SAS software. The more quadrature points are specified the better the approximation becomes but at the cost of increasing the computational time. In the following we used 10 quadrature points as proposed by Hamza *et al.* [5] in the same context.

It is essential to select good starting values in order to facilitate convergence and to shorten the time needed for estimation. As suggested in the user manual of a SAS macro for meta-analysis [20] we used a grid search to obtain good starting values. More specifically, the grid was created from all possible combinations of the values (0.18, 0.73, 0.97) for sensitivity and specificity, the values (0.5, 1) for both between-study variances and the values (−0.5, 0) for the correlation.

In addition, the hyperparameters are not optimized based on the parameterization of the covariance matrix for the random effects given in (1), but based on a more robust reparameterization. The use of Fisher's *z*-transformation for the correlation parameter ρ and the use of $\log(\sigma_\phi)$ and $\log(\sigma_\psi)$ ensure that the original parameters remain within a valid domain which considerably facilitates convergence. Note that `inla` uses analogous parameterizations of the hyperparameters, see Section 2.2.2 for details about Fisher's *z* transformation. As optimization algorithm the dual quasi-Newton algorithm is used. For more information about the estimation procedure, we refer to the extensive SAS manual [11].

2.2. Bayesian analysis with INLA

Up to now, Bayesian hierarchical models often implied the use of MCMC algorithms. MCMC-based model inference can be time-consuming and requires diagnostic checks to ensure good mixing properties and convergence of the simulated samples. These might be some of the reasons why a Bayesian model analysis is not widely used in practice though it might be beneficial [9]. Rue *et al.* [14] recently proposed an alternative deterministic Bayesian inference approach for latent Gaussian models. There is no sampling involved so the above-mentioned issues no longer apply. In the following, we briefly present the main ideas of INLA and then discuss how suitable prior distributions for parameters of the bivariate model could be chosen.

2.2.1. Inference based on deterministic approximations. A latent Gaussian model is a hierarchical model where the response variables are non-Gaussian, but the latent field is Gaussian controlled by a few hyperparameters. In the model presented in (1) the response is given in bivariate form, i.e. $\boldsymbol{y}=(y_1,\ldots,y_I)^\mathsf{T}$ with $y_i=(\mathrm{TP}_i,\mathrm{TN}_i)$ for $i=1,\ldots,I$. Choosing a zero-mean normal distribution with high variance (here we take 0.001^{-1}) for the regression parameters $\mu, \nu, \boldsymbol{\alpha}$, and $\boldsymbol{\beta}$ the latent Gaussian field is $\boldsymbol{x}=(\mu,\nu,\boldsymbol{\alpha}^\mathsf{T},\boldsymbol{\beta}^\mathsf{T},\boldsymbol{\phi}^\mathsf{T},\boldsymbol{\psi}^\mathsf{T})^\mathsf{T}$, where $\boldsymbol{\phi}=(\phi_1,\ldots,\phi_I)^\mathsf{T}$, $\boldsymbol{\psi}=(\psi_1,\ldots,\psi_I)^\mathsf{T}$, and $\dim(\boldsymbol{x})=\dim(\boldsymbol{\alpha})+\dim(\boldsymbol{\beta})+2(1+I)$. There are three hyperparameters $\boldsymbol{\theta}=(\sigma^2_\phi,\sigma^2_\psi,\rho)^\mathsf{T}$.

Assuming that $\{y_i : i=1,\ldots,I\}$ are conditionally independent the posterior distribution is given by

$$\pi(\boldsymbol{x},\boldsymbol{\theta}|\boldsymbol{y}) \propto \pi(\boldsymbol{\theta})\pi(\boldsymbol{x}|\boldsymbol{\theta})\prod_i \pi(y_i|x_i,\boldsymbol{\theta}),$$

where $\pi(\cdot|\cdot)$ denotes the conditional density of its arguments. We are mainly interested in the posterior marginals

$$\pi(x_i|\boldsymbol{y}) = \int_{\boldsymbol{\theta}} \underbrace{\int_{\boldsymbol{x}_{-i}} \pi(\boldsymbol{x},\boldsymbol{\theta}|\boldsymbol{y})d\boldsymbol{x}_{-i}}_{\pi(x_i,\boldsymbol{\theta}|\boldsymbol{y})=\pi(x_i|\boldsymbol{\theta},\boldsymbol{y})\pi(\boldsymbol{\theta}|\boldsymbol{y})} d\boldsymbol{\theta},$$

$$\pi(\theta_j|\boldsymbol{y}) = \int_{\boldsymbol{\theta}_{-j}} \underbrace{\int_{\boldsymbol{x}} \pi(\boldsymbol{x},\boldsymbol{\theta}|\boldsymbol{y})d\boldsymbol{x}}_{\pi(\boldsymbol{\theta}|\boldsymbol{y})} \mathrm{d}\boldsymbol{\theta}_{-j}.$$

Analytical integration of $\pi(\boldsymbol{x},\boldsymbol{\theta}|\boldsymbol{y})$ is usually not possible with the consequence that MCMC sampling is the common tool of choice. To be more specific, MCMC simulates a Markov chain that converges to the posterior marginal distribution. However, INLA can bypass MCMC completely. Instead of sampling it directly approximates

$$\pi(x_i|\boldsymbol{y}) = \int_{\boldsymbol{\theta}} \pi(x_i|\boldsymbol{\theta},\boldsymbol{y})\pi(\boldsymbol{\theta}|\boldsymbol{y})\,\mathrm{d}\boldsymbol{\theta} \tag{2}$$

using mainly Laplace approximations. The INLA scheme proceeds in three steps:

1. Build a Laplace approximation to $\pi(\boldsymbol{\theta}|\boldsymbol{y})$.
2. Build a Laplace approximation (or its simplified version) to $\pi(x_i|\boldsymbol{\theta},\boldsymbol{y})$.
3. To obtain approximations of the marginals of interest x_i the integral in equation (2) can be numerically computed as a finite sum

$$\tilde{\pi}(x_i|\boldsymbol{y}) = \sum_k \tilde{\pi}(x_i|\boldsymbol{\theta}_k,\boldsymbol{y}) \times \tilde{\pi}(\boldsymbol{\theta}_k|\boldsymbol{y}) \times \Delta_k,$$

where $\tilde{\pi}(\cdot|\cdot)$ denotes the approximated conditional density of its arguments. The sum is over values of $\boldsymbol{\theta}$ with area weights Δ_k. Posterior marginals for θ_j can be obtained similarly from $\tilde{\pi}(\boldsymbol{\theta}|\boldsymbol{y})$. For a detailed description of the approximations we refer to Rue *et al.* [14].

In contrast to the approximation error of INLA, the Monte Carlo error of MCMC can be made arbitrarily small by running the Markov chain infinitively long. However, Rue *et al.* [14] found that in typical examples the approximation error made by

INLA is small compared to the Monte Carlo error and negligible in practice. Typical examples include amongst other GLMMs for longitudinal data, time series models, or spatial disease mapping models. For the bivariate model we will compare the results of an example data set obtained by INLA and MCMC in Section 3.

The `inla` program [15] is written in C and built on the open source `GMRFLib` library [21, Appendix B]. The components of the model to be analyzed as well as the options for INLA are specified through an `ini` file which is processed by `inla`. However, there is no need for C programming when applying `inla`. An R-package called `INLA`, which works as an interface, is available. Its usage is similar to the `glm` routine in R, see the Appendix. Both the `inla` program and the R package are available for Unix, Windows, and Mac and can be freely downloaded from http://www.r-inla.org. The website provides documentation and a series of worked out examples. Note that the `inla` program is already included in the R package and needs not be additionally installed. For the bivariate meta-analysis model discussed in this paper a new `inla` model type '2diid' was written and integrated in `inla` by one of the authors (Håvard Rue). For all analyses presented within this paper we used the `inla` version 1.523 built on 11.08.2009 and applied the default approximation settings, including the simplified Laplace approximation.

2.2.2. Choice of prior distributions. As INLA is a fully Bayesian approach, all parameters of equation (1) are treated as random variables and prior distributions need to be assigned. The choice of appropriate prior distributions is a highly discussed topic [1, 9, 22]. When medical expert knowledge exists, the use of informative priors is encouraged [1]. Especially when there is no strong prior information available, the data should dominate the analysis. This desire leads to the use of *reference* or *non-informative* priors which should have virtually no influence on the inference. However, finding suitable reference priors for a binomial-normal hierarchical model is problematic.

For the bivariate normal model with zero-mean the reference prior is $\pi(\rho, \sigma_\phi^2, \sigma_\psi^2) \propto (1-\rho^2)^{-1}\sigma_\phi^{-2}\sigma_\psi^{-2}$ [23, Example 5.22]. This prior is improper and assigns more weight to extreme values of ρ than to values close to 0. It is not suitable at the second level of a hierarchical model as it can lead to an improper posterior distribution.

For the variances σ_ϕ^2 and σ_ψ^2 we therefore decide on either a mildly informative inverse gamma distribution with shape 0.25 and rate 0.025 or a less informative inverse gamma distribution with shape and rate equal to 0.001. Note that `inla` parameterizes the covariance matrix in terms of logarithmized precisions and thus we have transformed the returned results accordingly to obtain results for the variances. We also reparameterized ρ using Fisher's z-transformation

$$z = \text{logit}\left(\frac{\rho+1}{2}\right)$$

that takes values over the whole real line and assigned a normal distribution with mean zero and variance r^{-1} to z. Setting r to 0.2 corresponds to a U-shaped prior, $r=0.4$ results in a roughly uniform prior and $r=0.8$ in a bump-shaped prior for ρ, compare Figure 1. Note that $r=0$ corresponds to the reference prior.

Alternatively, the inverse Wishart distribution could be used as prior for the covariance matrix of the random effects. This choice is of particular interest in a MCMC setting as it is conjugate to the multivariate normal model and thus allows Gibbs sampling for the hyperparameters. Within `inla`, this prior could also be chosen, see the `inla` manual [15]. However, note that the inverse Wishart prior is not flexible enough to derive non-informative priors on the elements of the covariance matrix [24].

3. Example: the telomerase data

We re-analyzed 10 studies included in the broad review of Glas *et al.* [17] that use the telomerase marker for the diagnosis of bladder cancer. The data are shown in Table I. For this data set, estimation problems were reported by Riley *et al.* [18] who used `PROC NLMIXED` for the analysis of a generalized bivariate random effects meta-analysis model. The authors tried different starting values all resulting in an estimated between-study correlation of -1, differing parameter estimates and very large standard errors.

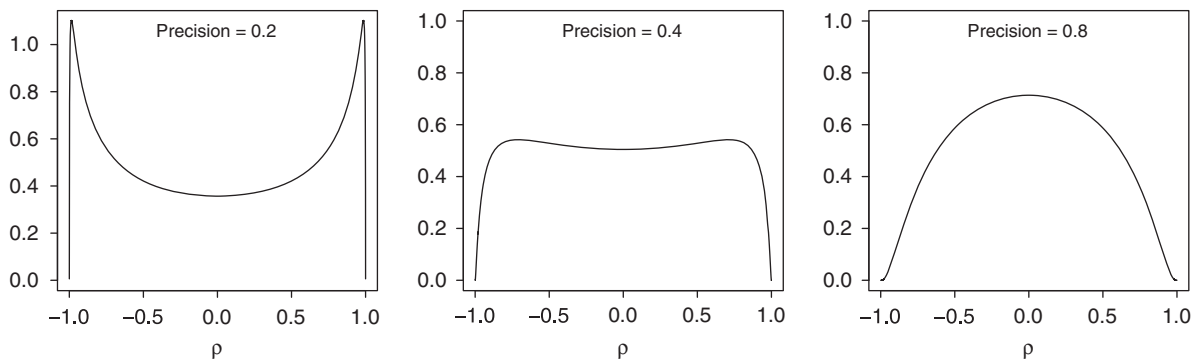

Figure 1. Prior distribution of the correlation parameter ρ derived from a zero-mean Gaussian distribution for z with three different values for the precision.

Table I. The telomerase data taken from the bladder cancer review of Glas *et al.* [17].

Study	P	TP	FN	N	TN	FP
1	33	25	8	26	25	1
2	21	17	4	14	11	3
3	104	88	16	47	31	16
4	26	16	10	83	80	3
5	57	40	17	138	137	1
6	47	38	9	30	24	6
7	42	23	19	12	12	0
8	33	27	6	20	18	2
9	17	14	3	32	29	3
10	44	37	7	29	7	22

Riley *et al.* [18] cited the small number of studies as a cause of their estimation problem, especially the inability to estimate the correlation parameter.

In the following we apply `inla` to the telomerase data. As this data set is sparse the results are potentially sensitive to the choice of prior distribution [22]. Using a sensitivity analysis we investigated the influence of the different prior distributions, presented in Section 2.2.2, on the results of `inla`. Figure 2 shows the posterior marginal distributions of all parameters resulting for the different prior distributions. It is obvious that sensitivity and specificity are very stable for all choices of hyperpriors. In contrast, the hyperparameters are more influenced because they are the parameters the priors are defined for. Thus, there is no separating layer as was the case for sensitivity and specificity. For the correlation parameter ρ there are three pairs of curves caused by the three different priors for ρ. Within each pair the marginals are only slightly influenced by the choice of prior distributions for the variances. Note that for all priors the posterior estimate of the correlation parameter is very high in absolute value. However, the roughly uniform and bump-shaped prior do not provide much density to extreme correlations (see Figure 1) so that posterior estimates of ρ close to -1 are almost impossible. In contrast, the U-shaped prior assigns sufficient probability mass to values close to -1 and 1 thus making even extreme posterior estimates possible. For a data set with lower correlation the influence of the correlation prior is expected to be less. Inspecting the posterior marginals for the variances σ_ϕ^2 and σ_ψ^2 we note that the overall appearance is similar across the priors, but differences occur.

To compare the results obtained by `inla` and MCMC we used the U-shaped prior for ρ and the mildly informative prior for the variances. These priors were also selected after the sensitivity analysis of the extensive simulation study presented in Section 4. The R-code used to fit the model with `inla` can be found in Appendix A. Building and running the model took less than 0.2 seconds. The MCMC sampler is implemented in C and uses the one-block algorithm proposed by Rue and Held [21, Section 4.1] to jointly update the hyperparameters and the latent field. We obtained 10 000 posterior samples using a burn-in period of 5 000 iterations and a thinning of 20. Standard output diagnostics provided within the R-package `coda` [25] were applied to check for convergence. The MCMC approach needed about 10 min. Figure 3 shows the histograms based on the MCMC samples together with the corresponding posterior marginals obtained by `inla` for all parameters. The approximations computed with `inla` are very precise.

In addition we applied `inla` as an empirical Bayes (EB) approach meaning that only the modal configuration is used to integrate over $\pi(\boldsymbol{\theta}|\boldsymbol{y})$ [14]. This approach can also be interpreted from a frequentist perspective where the uncertainty in the hyperparameters is ignored. The results are expected to be similar to those of `PROC NLMIXED`. A comparison of the parameter estimates obtained by MCMC, `inla`, `inla` (EB), and `PROC NLMIXED` is given in Table II. As `PROC NLMIXED` optimizes model (1) in terms of logarithmized standard deviations we present the results for σ_ϕ and σ_ψ in the table. The estimates for μ, ν, and σ_ϕ are very similar over all approaches. Differences only appear for standard deviation σ_ψ and correlation ρ. However, these are the parameters that are the most difficult to estimate because σ_ψ is very high and ρ is close to -1. Owing to the extreme right skewness of the posterior marginals the computation of credible intervals is difficult for these parameters. Especially, the upper limits differ across the different approaches, whereas the point estimates are still similar. Note that `PROC NLMIXED` fails to report reliable estimates for the between-study correlation and estimates ρ to be -1 with confidence interval ranging from -1 to 1, see Riley *et al.* [18]. An estimate $\hat{\rho}=-1$ is *per se* not alarming. However, a confidence interval covering the whole parameter space means that no information could be gained. In contrast, all Bayesian approaches produce comparable results.

4. Simulation study

An extensive simulation study was conducted to investigate in more detail the effect of different prior distributions on the results obtained with `inla` and to compare the performance of `inla` and `PROC NLMIXED`. We varied the number of studies included in the meta-analysis, the mean logit sensitivity and mean logit specificity, the between-study variances, and the correlation between logit sensitivity and logit specificity. No covariate information has been considered for the simulation.

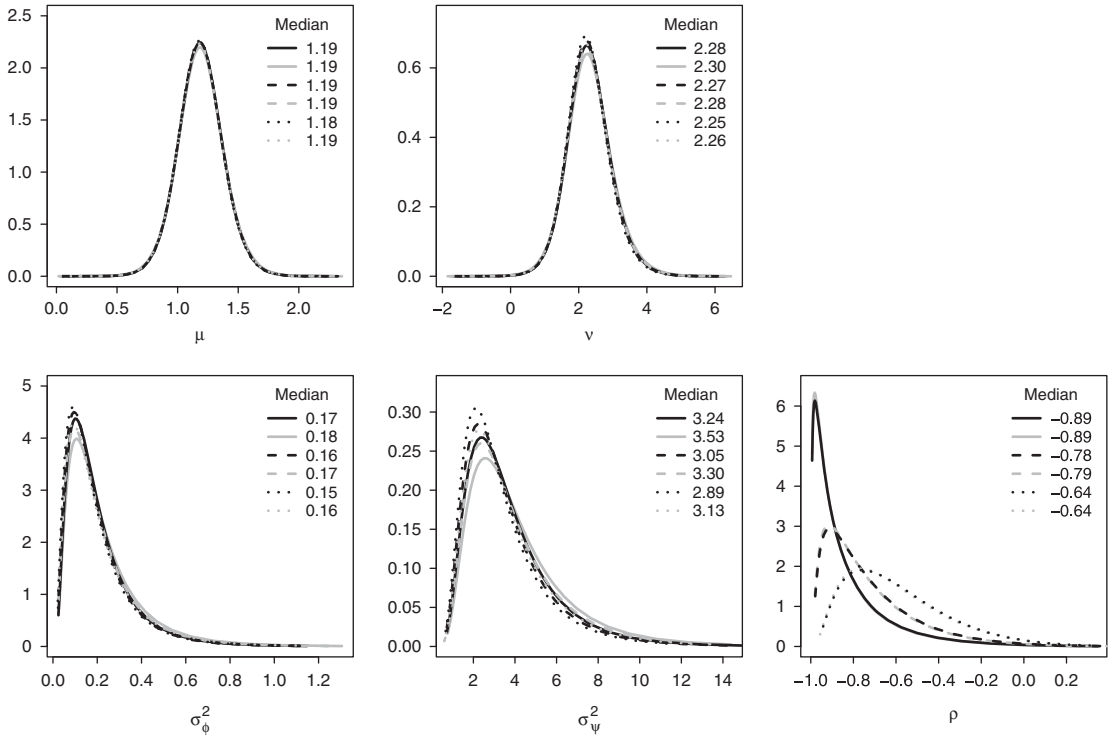

Figure 2. Posterior marginal distributions for the telomerase data set. Black lines correspond to the mildly informative (shape = 0.25, rate = 0.025) and grey lines to the less informative (shape = 0.001, rate = 0.001) inverse gamma prior distributions for the variances. Line types represent the different prior distributions for the correlation parameter: (—) U-shaped, (- -) roughly uniform, (···) bump-shaped.

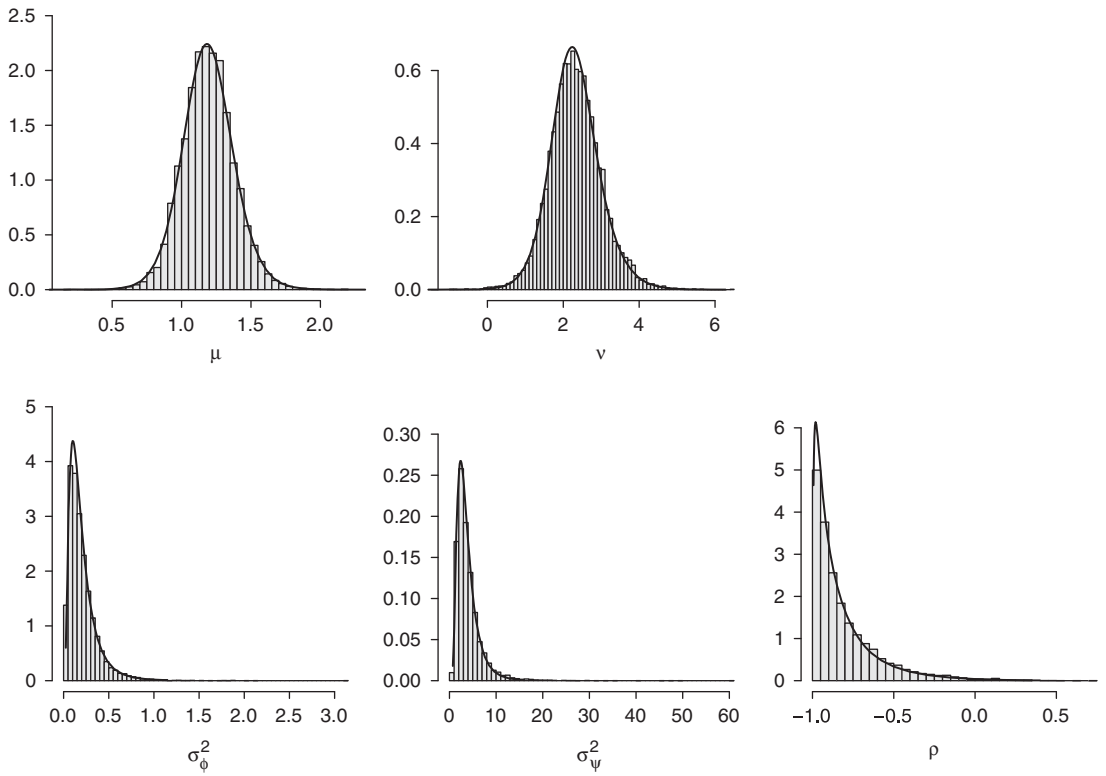

Figure 3. Approximated posterior marginals obtained by `inla` and corresponding histograms of 10 000 samples obtained from a MCMC run with 5 000 burn-in iterations and a thinning of 20. As prior distributions the U-shaped prior for ρ and the mildly informative inverse gamma prior for the variances were used.

Table II. Summary of parameter estimates using MCMC, `inla`, `inla` using an empirical Bayes (EB) approach and `PROC NLMIXED`. As prior distributions the U-shaped prior for ρ and the mildly informative inverse gamma prior for the variances were used.

	μ	ν	σ_ϕ	σ_ψ	ρ
MCMC	$_{0.84}1.19_{1.56}$	$_{1.03}2.30_{3.78}$	$_{0.18}0.40_{0.82}$	$_{1.10}1.81_{3.43}$	$_{-0.99}-0.88_{-0.18}$
`inla`	$_{0.82}1.19_{1.57}$	$_{1.07}2.28_{3.72}$	$_{0.21}0.41_{0.79}$	$_{1.08}1.80_{3.07}$	$_{-0.99}-0.89_{-0.26}$
`inla` (EB)	$_{0.84}1.19_{1.54}$	$_{1.14}2.32_{3.49}$	$_{0.21}0.42_{0.81}$	$_{1.03}1.74_{2.94}$	$_{-0.98}-0.86_{-0.19}$
`PROC NLMIXED`	$_{0.84}1.19_{1.55}$	$_{1.10}2.34_{3.59}$	$_{0.21}0.43_{0.85}$	$_{1.04}1.82_{3.18}$	$_{-1.00}-1.00_{+1.00}$

Note: The triple notation of ${}_L P_U$ denotes the point estimate P with 95 per cent confidence limits (L,U) for `PROC NLMIXED`, or the posterior median P with 2.5 per cent quantile L and 97.5 per cent quantile U for MCMC and `inla`.

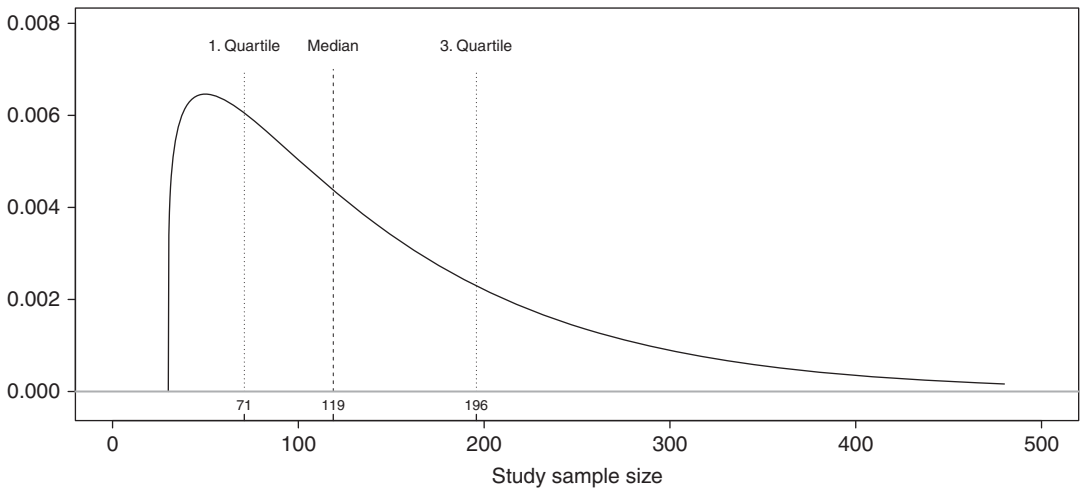

Figure 4. Shifted gamma distribution with shape = 1.2, rate = 0.01, and lag = 30 to generate the study sample size. (Median and quartiles are indicated.)

4.1. Data simulation

To generate the data for one meta-analysis we first specified the parameters μ, ν, σ_ϕ^2, σ_ψ^2, ρ, and the number of studies I included in the meta-analysis. The simulation of the contingency table for each study $i = 1,\ldots,I$ is as follows:

1. Sample the study size $n_i = P_i + N_i$ composed of diseased and non-diseased study participants from a shifted gamma distribution and round off to the nearest integer.
2. Based on n_i determine P_i from a binomial distribution (see below).
3. Generate Se_i and Sp_i from model (1) and likewise TP_i and TN_i.
4. Calculate FN_i and FP_i.

Figure 4 shows the distribution of the study sample size n_i used in the simulation study. Median and interquartile range are comparable to the corresponding values found by Bachmann *et al.* [26] in a literature survey of studies on diagnostic accuracy. Based on the study sample size the number of diseased persons P_i is drawn from a binomial distribution with probability 0.43 that corresponds to the median prevalence found by Bachmann *et al.* [26]. To ensure a minimum of 10 persons in each group P_i and N_i were appropriately adapted if needed.

Dinnes *et al.* [27] noted a median number of 22 studies per meta-analysis in a review including 189 meta-analyses of diagnostic test or screening studies. In the simulation study we thus selected $I = 25$ studies per meta-analysis. Additionally, we looked at more difficult situations with only $I = 10$ studies. In total we considered 72 different scenarios each consisting of 1 000 meta-analyses. The specifications of all scenarios are given in Table III.

4.2. Sensitivity analysis

Using a sensitivity analysis we investigated the influence of the different prior distributions presented in Section 2.2.2 on the results of `inla`. We used all scenarios with $\rho = -0.95, -0.4, 0.0$, resulting in 36 out of the 72 scenarios presented in Table III.

The estimated results were compared using bias (mean difference between the posterior median and the true value), standard deviation of the posterior medians (SD), MSE (sum of squared bias and squared SD), and coverage probability (frequency in which the true value is within the 95 per cent credible interval).

Table III. The different scenarios used in the simulation study. Each subset of six scenarios correponds to the correlation values $\rho = -0.95, -0.8, -0.6, -0.4, -0.2, 0.0$.

Scenario	$\text{logit}^{-1}(\mu)$	$\text{logit}^{-1}(\nu)$	σ^2_ϕ	σ^2_ψ	I
1–6	0.8	0.7	1	1	10
7–12	0.9	0.4	1	1	10
13–18	0.95	0.3	1	1	10
19–24	0.8	0.7	1	1	25
25–30	0.9	0.4	1	1	25
31–36	0.95	0.3	1	1	25
37–42	0.8	0.7	0.5	1	10
43–48	0.9	0.4	0.5	1	10
49–54	0.95	0.3	0.5	1	10
55–60	0.8	0.7	0.5	1	25
61–66	0.9	0.4	0.5	1	25
67–72	0.95	0.3	0.5	1	25

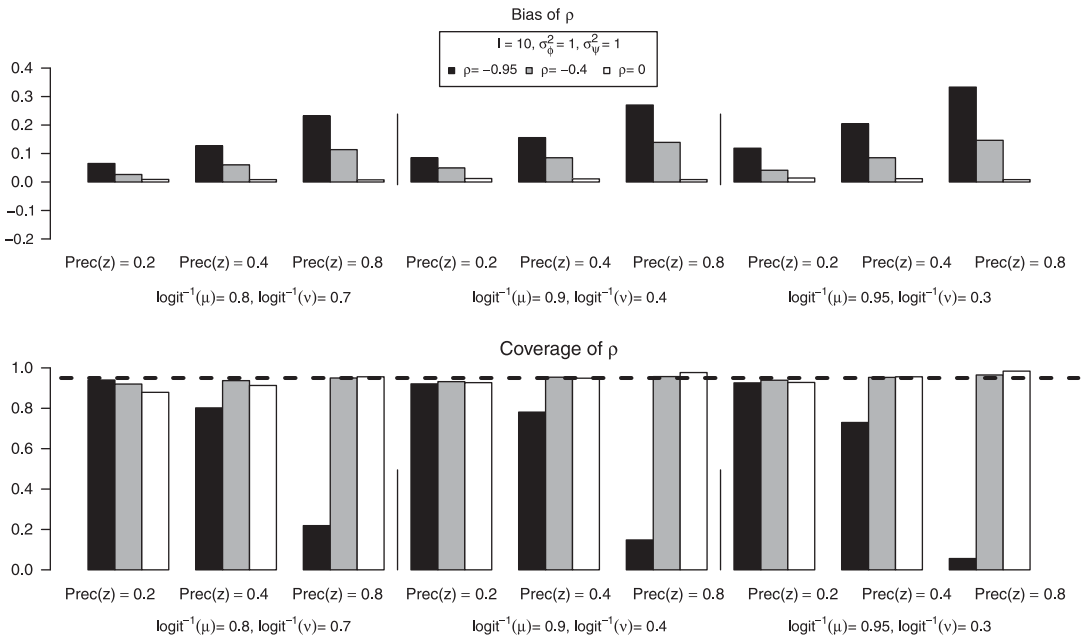

Figure 5. Bias and coverage of correlation parameter ρ for different values of ρ, $\text{logit}^{-1}(\mu)$, $\text{logit}^{-1}(\nu)$ in the data simulation. The dashed line in the coverage plot corresponds to the nominal level of 95 per cent. Analyses were performed with `inla` using the mildly informative prior distribution for σ^2_ϕ and σ^2_ψ.

The results across the different priors are very similar for both sensitivity and specificity. We only found differences concerning the correlation parameter ρ and the between-study variances which we will illustrate with selected scenarios. The results for the remaining scenarios are comparable and available from the authors on request. The U-shaped prior distribution for the correlation yielded nearly unbiased estimates of ρ. Inspecting the coverage we observed that the U-shaped prior distribution obtained good coverage for correlation parameter ρ, whereas the bulk-shaped prior distribution yielded very low coverage for scenarios simulated with extreme correlation values, see Figure 5. As already noted in Section 3 for the telomerase data, this can be explained by the fact that the U-shaped prior distribution also gives prior weight to extreme values of ρ. We also observed this behaviour in the SD and the MSE. For intermediate values of ρ, the SD and MSE using the U-shaped prior were a little higher, whereas for extreme values of ρ these measures were a little lower in comparison to the bulk-shaped prior.

For the variance parameter σ^2_ϕ the SD and the MSE are smaller using the mildly informative inverse gamma prior in comparison to the less informative prior distribution. Figure 6 shows the MSE of σ^2_ϕ for both prior distributions in nine selected scenarios. Taking into account that `inla` returned an error in 10 cases out of 36 000 analyses because of numerical problems using the less informative prior for σ^2_ϕ and σ^2_ψ and twice using the mildly informative prior, we chose the mildly informative prior for the comparison with `PROC NLMIXED`. Regarding the correlation parameter ρ we picked the U-shaped prior distribution.

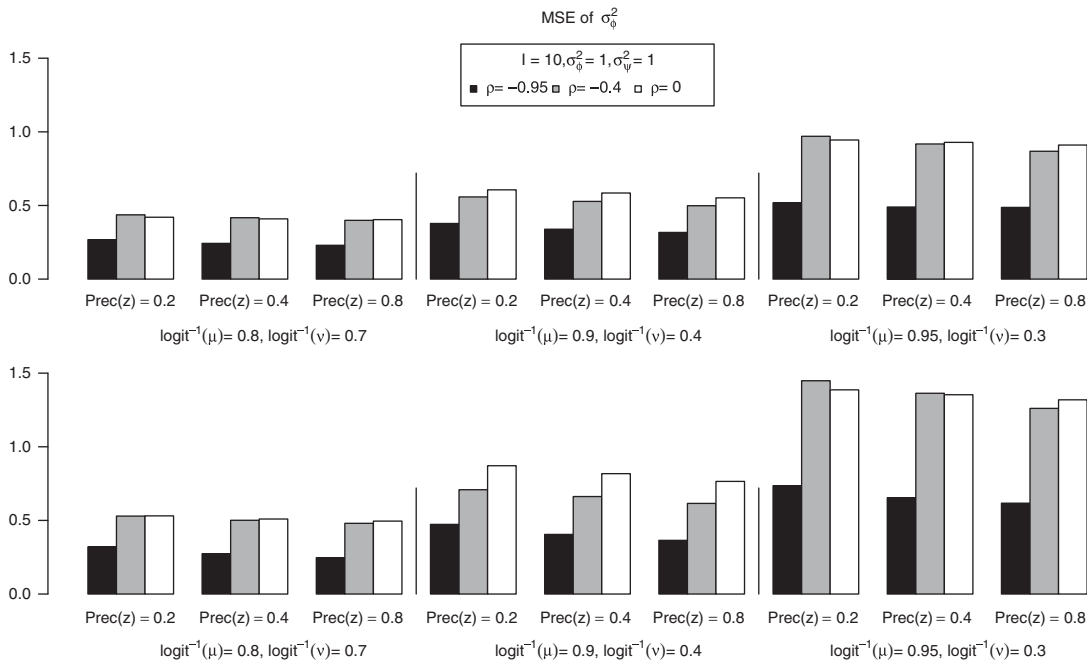

Figure 6. MSE for variance parameter σ^2_ϕ analyzed with `inla` using the mildly informative prior for σ^2_ϕ and σ^2_ψ (top) and the less informative prior for σ^2_ϕ and σ^2_ψ (bottom).

4.3. Comparison of `PROC NLMIXED` *and* `inla`

To compare the performance of `inla` and `PROC NLMIXED` we again looked at the measures used in Section 4.2 for the sensitivity analysis. We also considered the average run-time and the number of failures per scenario, i.e. how often did the method not produce reliable results due to numerical problems such as non-convergence of the optimization method. The 1 000 data sets generated for each scenario were analyzed with both methods. We defined `inla` to have failed if the programme returned an error. Similarly, a failure was defined for `PROC NLMIXED` if either an error or a warning concerning convergence or the positive definiteness of the returned Hessian matrix was issued. The warnings imply that at least one parameter estimate or standard error could not be computed. Throughout the simulation study we used the settings for `PROC NLMIXED` and `inla` described in Sections 2.1 and 2.2.

All analyses were run on a laptop with Intel(R) Core(TM) 2 Duo T7200 processor 2.00 GHz. `PROC NLMIXED` took on average 42.2 min (minimum 23.1, maximum 74.7) for one scenario with 1 000 meta-analyses, whereas `inla` took on average 3.8 min (2.2–6.0). Both methods needed more time to analyze scenarios with $I=25$ studies per meta-analysis compared with scenarios with $I=10$ studies, respectively. Concerning the non-convergence rate `inla` proved to be very stable and only failed twice, once in scenario 8 and once in scenario 13. In contrast, convergence of the optimization algorithm was problematic for `PROC NLMIXED` particularly in scenarios with large correlation. An increasing number of studies per meta-analysis yielded less failures. Overall, the non-convergence rate of `PROC NLMIXED` based on the above definition of a failure was 0.6 per cent (409 out of 72 000 analyses). However, although `PROC NLMIXED` rarely returned a direct non-convergence error there were many data sets (11 037 out of 72 000) for which the correlation was estimated to be -1 with very large standard errors. As a result confidence intervals cover the whole domain of ρ. This problem predominantly occurred in scenarios with high correlation and a low number of studies and is also discussed in Riley *et al.* [18] and Verde [28].

Tables IV and V display the results obtained for the sensitivity and specificity, for a selection of scenarios. The results are only shown for scenarios 37–42, 49–60, and 67–72 with pairs of (0.7, 0.8) and (0.95, 0.3) for sensitivity and specificity, respectively, and variances $\sigma^2_\phi=0.5$, $\sigma^2_\psi=1$ that include the most difficult scenarios with only 10 studies, sensitivity close to 1 and high between-study correlation. The conclusions for the remaining scenarios are similar and the results for the whole set of scenarios are available from the authors on request. Note that bias, MSE, and coverage obtained with `PROC NLMIXED` are only based on the #*d* data sets where no convergence errors or warnings were issued, whereas the results for `inla` are based on all 1 000 data sets.

We observed that the bias for sensitivity (Table IV) as well as specificity (Table V) is rather small. If more studies are available for a meta-analysis, then estimates are more precise, the MSE is lower, and coverage higher. Both bias and MSE depend on the choice of μ and ν. In general, estimates are less biased and have smaller MSE for values closer to 1 or 0 than for values close to 0.5. There is almost no difference between both methods concerning bias and MSE. Note that the precision of the estimates and therefore also the MSE are hardly influenced by the value of the correlation ρ. Over all 72 scenarios the coverage probabilities for sensitivity obtained by `inla` are higher and closer to the nominal value of 95 per cent than those obtained by `PROC NLMIXED`

Table IV. Simulation results for the sensitivity $\text{logit}^{-1}(\mu)$. Shown are bias, MSE, and coverage for scenarios with variances $\sigma_\phi^2 = 0.5, \sigma_\psi^2 = 1$. The results for `PROC NLMIXED` are based on #*d* and for `inla` on 1000 data sets.

True values					Bias		MSE × 10		Coverage (per cent)	
I	$\text{logit}^{-1}(\mu)$	$\text{logit}^{-1}(\nu)$	ρ	#*d*	NLMIXED	inla	NLMIXED	inla	NLMIXED	inla
10	0.80	0.70	−0.95	974	−0.001	−0.001	0.017	0.017	91.5	92.8
			−0.80	992	−0.005	−0.005	0.018	0.018	89.5	92.1
			−0.60	996	−0.004	−0.005	0.019	0.018	89.9	93.0
			−0.40	997	−0.005	−0.005	0.017	0.017	90.5	93.6
			−0.20	1000	−0.003	−0.004	0.017	0.017	91.1	94.6
			0.00	997	−0.002	−0.002	0.016	0.016	90.6	92.7
	0.95	0.30	−0.95	958	−0.001	−0.001	0.002	0.002	92.6	93.7
			−0.80	981	−0.001	−0.001	0.002	0.002	92.2	93.6
			−0.60	986	−0.001	−0.001	0.003	0.003	92.3	93.4
			−0.40	988	−0.001	−0.001	0.002	0.002	91.4	93.5
			−0.20	992	−0.002	−0.002	0.002	0.002	92.1	93.5
			0.00	990	−0.002	−0.002	0.003	0.003	89.1	91.8
25	0.80	0.70	−0.95	995	−0.002	−0.002	0.007	0.007	93.5	94.1
			−0.80	999	−0.001	−0.001	0.007	0.006	93.6	94.4
			−0.60	1000	−0.002	−0.002	0.007	0.007	93.1	94.5
			−0.40	1000	−0.001	−0.001	0.007	0.007	93.3	94.3
			−0.20	1000	−0.000	−0.001	0.007	0.007	92.7	94.5
			0.00	1000	−0.002	−0.002	0.007	0.007	93.3	94.5
	0.95	0.30	−0.95	993	−0.000	−0.000	0.001	0.001	93.4	93.7
			−0.80	1000	−0.001	−0.001	0.001	0.001	94.1	94.1
			−0.60	998	−0.001	−0.002	0.001	0.001	94.1	94.7
			−0.40	1000	−0.001	−0.001	0.001	0.001	94.1	94.5
			−0.20	1000	−0.001	−0.001	0.001	0.001	92.7	93.4
			0.00	1000	−0.001	−0.001	0.001	0.001	93.3	93.8

(see Table IV). This statement concerning coverage probabilities generally also applies for specificity (Table V). The better coverage for the estimates of sensitivity and specificity by `inla` is a consequence of the fact that fully Bayesian approaches take into account that variance parameters are estimated. Accordingly, this uncertainty is considered when constructing credibility intervals. In contrast, second-order ML methods generally underestimate the variability of the estimated main regression coefficients, here μ and ν, when sample sizes are small to moderate [10]. Thus, confidence intervals obtained by `PROC NLMIXED` are too narrow.

The results for the between-study correlation parameter ρ are shown in Table VI. The absolute value of the correlation tends to be underestimated by `inla` and the bias increases with increasing $|\rho|$. On the other hand estimates are more precise for a large correlation and, accordingly, the MSE decreases with increasing $|\rho|$. Note that `PROC NLMIXED` returned estimates $\hat{\rho} = -1$ with confidence interval $(-1, 1)$ in at least 5 out of 1 000 analyses per scenario and up to 68 per cent in scenarios with extreme correlation. As one consequence, coverage probabilities are too high and converge towards 100 per cent. In general, we can observe that `PROC NLMIXED` tends to overestimate the nominal coverage value of 95 per cent, whereas `inla` often provides too low coverage probabilities.

The results for the standard deviations σ_ϕ and σ_ψ are shown in Tables BI and BII in Appendix B. Both methods underestimate the variability though the bias by `PROC NLMIXED` is, especially for σ_ψ (Table BII), larger than the bias by `inla`, see also [10]. Coverage probabilities for σ_ϕ obtained by `PROC NLMIXED` are higher than the nominal value of 95 per cent. In particular, in scenarios with few studies ($I = 10$) or if there is little between-study variation concerning sensitivity ($\sigma_\phi^2 = 0.5$ and $\text{logit}^{-1}(\mu) = 0.95$) coverage converges towards 100 per cent (Table BI). This is due to the very wide confidence intervals. For σ_ψ, coverage tends to be too low (Table BII). For `inla`, coverage probabilities for both σ_ϕ and σ_ψ tend to be too low.

5. Discussion

In this paper we considered bivariate meta-analysis of diagnostic test accuracy studies. This approach is closely related to the hierarchical sROC model [29, 30], as shown by Harbord *et al.* [4]. Because of its intuitive model formulation the bivariate approach is very common. We have discussed the use of a new deterministic Bayesian approach using integrated nested Laplace approximations. There are good reasons for preferring Bayesian methods to non-Bayesian approaches. One reason is that the

Table V. Simulation results for the specificity $\text{logit}^{-1}(\nu)$. Shown are bias, MSE, and coverage for scenarios with variances $\sigma^2_\phi = 0.5$, $\sigma^2_\psi = 1$. The results for PROC NLMIXED are based on #*d* and for inla on 1000 data sets.

True values					Bias		MSE × 10		Coverage (per cent)	
I	$\text{logit}^{-1}(\mu)$	$\text{logit}^{-1}(\nu)$	ρ	#*d*	NLMIXED	inla	NLMIXED	inla	NLMIXED	inla
10	0.80	0.70	−0.95	974	−0.006	−0.006	0.051	0.051	91.0	92.3
			−0.80	992	−0.002	−0.002	0.045	0.045	91.1	93.1
			−0.60	996	−0.000	−0.001	0.047	0.046	90.8	93.6
			−0.40	997	−0.003	−0.004	0.052	0.052	88.8	91.7
			−0.20	1000	−0.007	−0.007	0.050	0.050	89.3	93.1
			0.00	997	−0.005	−0.006	0.051	0.051	89.8	93.4
	0.95	0.30	−0.95	958	0.007	0.008	0.050	0.050	90.0	92.4
			−0.80	981	0.006	0.006	0.050	0.049	89.9	92.6
			−0.60	986	0.004	0.005	0.049	0.049	91.0	93.9
			−0.40	988	0.000	0.001	0.049	0.049	90.8	93.1
			−0.20	992	0.001	0.002	0.050	0.049	91.6	93.7
			0.00	990	−0.000	0.000	0.045	0.045	90.6	94.2
25	0.80	0.70	−0.95	995	0.000	0.000	0.020	0.020	92.6	92.8
			−0.80	999	−0.003	−0.004	0.021	0.021	92.5	93.1
			−0.60	1000	−0.000	−0.000	0.018	0.018	94.3	95.5
			−0.40	1000	−0.001	−0.002	0.020	0.020	91.9	94.0
			−0.20	1000	−0.002	−0.002	0.018	0.018	94.6	95.2
			0.00	1000	−0.001	−0.002	0.020	0.020	92.9	94.1
	0.95	0.30	−0.95	993	0.001	0.001	0.020	0.020	93.9	93.9
			−0.80	1000	0.001	0.001	0.019	0.019	94.2	94.6
			−0.60	998	0.003	0.003	0.018	0.018	94.1	95.5
			−0.40	1000	−0.001	−0.000	0.017	0.017	95.7	95.9
			−0.20	1000	0.001	0.001	0.019	0.019	93.8	94.9
			0.00	1000	0.003	0.003	0.020	0.020	92.7	93.7

uncertainty in all parameter estimates is taken into account which is especially important in the case of sparse data [1, 22]. Problems known from MCMC that may have led to the rare use of Bayesian methods no longer apply when using INLA as no Monte Carlo inference is involved.

Starting from a motivating example we tested the influence of different prior distributions and illustrated the very good agreement of the results obtained by INLA and MCMC. An extensive simulation study was conducted to compare the performance of inla and PROC NLMIXED for meta-analyses comprising 10 and 25 studies. Both methods exhibit quite similar performance with respect to bias and MSE concerning sensitivity and specificity. However, as expected, inla showed better coverage for these parameters because PROC NLMIXED does not adequately account for uncertainty of the variance components. Concerning the hyperparameters, PROC NLMIXED was often not able to produce reliable estimates for the correlation parameter and by trend underestimated variance parameters. In contrast, inla practically always produced reliable estimates for the correlation and generally less downwardly biased variance estimates. Further, work could be spent on studying bivariate meta-analyses containing less than 10 studies. However, the correlation cannot be accurately estimated anymore with too few studies. Moreover, the likelihood is presumed to be extremely flat, so that even more failures or unreliable estimates are expected for the frequentist approach. For the Bayesian approach the choice of prior distributions is crucial and an even wider sensitivity analysis becomes indispensable [22].

The better time performance that we observed for inla refers to the analysis of 1 000 meta-analyses. When analysing only one meta-analysis the time argument is not essential. Note that because sensitivity and specificity are expected to be negatively correlated our simulation study was based on negative correlations. Results for the corresponding positive correlations are similar.

In the simulation study, inla had virtually no problems in obtaining sensible estimates, whereas PROC NLMIXED often had estimation problems in the presence of extreme correlation between logit sensitivity and logit specificity. Keeping in mind that the bivariate model has been propagated for use in diagnostic meta-analyses especially because of its ability to incorporate correlation, the increasing failure rate of PROC NLMIXED for increasing correlations seems to be counterproductive. We tried different initial values and changed the specifications of the optimization routine to obtain convergence for the example application. However, the results for the correlation parameter were never reliable. For the simulation study we used a grid of plausible initial values to find a good starting configuration.

As in diagnostic test studies the sensitivity and specificity are usually negatively correlated, a joint analysis may borrow strength from the related outcome data by incorporating their correlation. Reduced standard errors are to be expected compared with

Table VI. Simulation results for the correlation parameter ρ. Shown are bias, MSE, and coverage for scenarios with variances $\sigma^2_\phi = 0.5, \sigma^2_\psi = 1$. The results for `PROC NLMIXED` are based on #*d* and for `inla` on 1000 data sets.

True values					Bias		MSE × 10		Coverage (per cent)	
I	$\text{logit}^{-1}(\mu)$	$\text{logit}^{-1}(\nu)$	ρ	#*d*	NLMIXED	inla	NLMIXED	inla	NLMIXED	inla
10	0.80	0.70	−0.95	974	−0.001	0.088	0.103	0.193	97.7	93.1
			−0.80	992	−0.009	0.073	0.578	0.490	97.5	96.5
			−0.60	996	−0.024	0.040	1.129	0.862	97.3	94.7
			−0.40	997	−0.012	0.026	1.682	1.233	97.6	92.3
			−0.20	1000	0.026	0.044	2.058	1.496	96.1	90.6
			0.00	997	−0.010	−0.005	2.197	1.593	95.7	90.5
	0.95	0.30	−0.95	958	0.030	0.182	0.594	0.729	96.6	90.0
			−0.80	981	−0.006	0.137	1.337	0.934	98.8	96.4
			−0.60	986	−0.011	0.096	2.346	1.319	99.2	97.2
			−0.40	988	−0.010	0.067	3.429	1.730	98.7	95.1
			−0.20	992	−0.006	0.030	3.678	1.806	99.2	95.7
			0.00	990	0.008	0.003	4.236	2.014	99.5	94.0
25	0.80	0.70	−0.95	995	−0.007	0.029	0.030	0.036	97.8	94.3
			−0.80	999	−0.006	0.022	0.145	0.134	97.5	95.3
			−0.60	1000	0.003	0.021	0.344	0.320	94.8	93.4
			−0.40	1000	−0.004	0.007	0.485	0.452	95.3	93.9
			−0.20	1000	−0.000	0.004	0.605	0.573	93.4	91.3
			0.00	1000	0.011	0.010	0.636	0.608	94.3	92.7
	0.95	0.30	−0.95	993	0.005	0.068	0.078	0.115	93.8	89.9
			−0.80	1000	−0.028	0.031	0.295	0.244	97.5	96.2
			−0.60	998	−0.025	0.016	0.596	0.479	97.1	95.2
			−0.40	1000	−0.025	0.000	1.029	0.763	96.8	93.9
			−0.20	1000	−0.015	−0.001	1.062	0.875	96.3	92.7
			0.00	1000	−0.023	−0.021	1.194	0.934	96.6	93.6

univariate approaches in the presence of correlation. Somewhat surprisingly, the simulation study showed that the actual value of the correlation between logit sensitivity and logit specificity had only a small effect on the precision of $\hat{\mu}$ and $\hat{\nu}$ and also did not influence the bias. It would be interesting to examine in detail how the Bayesian bivariate model performs compared to the simpler model without correlation, i.e. two separate univariate analyses of sensitivity and specificity. Riley *et al.* [18, 31] showed within a slightly different frequentist setting the benefits and limitations of a bivariate approach.

We demonstrated the user-friendliness of `inla` in an application to a published meta-analysis. The incorporation of covariates into the analysis is also straightforward. In the INLA-manual the Scheidler *et al.* [32] data set is included with indicators for different imaging types, i.e. binary covariates. As is also possible with `PROC NLMIXED`, continuous covariates can be included in the same manner. Using `inla` it is additionally possible to estimate covariate effects not only parametrically, say as linear functions, but also non-parametrically as smooth functions.

It is worthy to mention that Chu *et al.* [33] recently extended the bivariate model to jointly analyze sensitivity, specificity, and the disease prevalence. The authors used `PROC NLMIXED` for the analysis, but noted that the use of a Bayesian approach might be helpful to get more accurate standard errors. The integration of the trivariate model in `inla` seems straightforward.

Appendix A: R-interface

The interface to `inla` from R can be loaded with

```
> library(INLA)
```

The telomerase meta-analysis for the diagnosis of bladder cancer is presented in the broad review of Glas *et al.* [17] and shown in Table I. For the analysis with `inla`, we need to transform the data in a way such that each study is represented by two consecutive rows of the data set. The first row contains the number of diseased persons P_i, true positives TP_i, and an index. Accordingly, the second row contains the number of non-diseased persons N_i, true negatives TN_i, and an index. The index runs

from 1,...,2*I*. The first rows of the data-frame are as follows:

```
> head(telomerase)
    N   Y id tp tn
1  33  25  1  1  0
2  26  25  2  0  1
3  21  17  3  1  0
4  14  11  4  0  1
5 104  88  5  1  0
6  47  31  6  0  1
```

Using a formula object the model is specified as

```
> formula <- Y ~ f(id, model = "2diid", param = c(0.25, 0.025, 0.25, 0.025, 0, 0.2)) +
+               tp + tn - 1
```

The function `f(id, model = "2diid",...)` is used to define the bivariate random effects $(\phi_1,\psi_1),(\phi_2,\psi_2),\ldots,(\phi_I,\psi_I)$ by setting `model = "2diid"`. The argument `param` contains the two parameters of the prior distributions for each of the three hyperparameters in the order τ_ϕ, τ_ψ, and ρ. Note that `inla` works with and returns precisions and not variances. As the intercepts μ and ν of model formulation (1), denoted by `tp` and `tn` in the telomerase data-frame, are assumed to have a fixed effect they are simply added to the formula object. Note that the automatically added intercept has to be explicitly removed with -1. Once the formula is defined, we call the main function `inla()` where the likelihood family of the model and some additional parameters are specified.

```
> model <- inla(formula, family = "binomial", Ntrials = N, data = telomerase,
+          quantiles = c(0.025, 0.5, 0.975))
```

A summary of the fitted model with estimates $\widehat{\texttt{tp}}$, $\widehat{\texttt{tn}}$, and hyperparameters $\hat{\tau}_\phi$, $\hat{\tau}_\psi$, $\hat{\rho}$ can be obtained with:

```
> summary(model)
```

Note that `inla` displays a warning that the approximations might not be very accurate. This warning is automatically generated whenever the equivalent number of replicates is smaller than two and does not imply unreliable results. Note that the telomerase data set includes 10 studies and we have more than 5 effective parameters.

The meta-analysis conducted by Scheidler *et al.* [32] and analyzed with a bivariate approach e.g. by Reitsma *et al.* [7], Chu and Cole [8], and Harbord *et al.* [4], is included in the `INLA` R-package and documented in detail in the `inla` manual [15]. In this example three different diagnostic imaging types are compared for detecting lymph node metastases in patients with cervical cancer.

Appendix B: Simulation results

Simulation results for the standard deviations σ_ϕ and σ_ψ are found in Tables BI and BII, respectively.

Table BI. Simulation results for the standard deviation σ_ϕ. Shown are bias, MSE, and coverage for scenarios with variances $\sigma^2_\phi = 0.5, \sigma^2_\psi = 1$. The results for `PROC NLMIXED` are based on #*d* and for `inla` on 1000 data sets.

True values					Bias		MSE × 10		Coverage (per cent)	
I	$\text{logit}^{-1}(\mu)$	$\text{logit}^{-1}(\nu)$	ρ	#*d*	NLMIXED	inla	NLMIXED	inla	NLMIXED	inla
10	0.80	0.70	−0.95	974	−0.057	−0.064	0.496	0.469	96.0	93.1
			−0.80	992	−0.080	−0.072	0.516	0.480	95.8	92.2
			−0.60	996	−0.067	−0.050	0.532	0.516	97.1	93.9
			−0.40	997	−0.080	−0.055	0.557	0.540	97.1	92.5
			−0.20	1000	−0.070	−0.039	0.581	0.578	97.5	93.1
			0.00	997	−0.087	−0.057	0.600	0.596	97.5	91.7
	0.95	0.30	−0.95	958	−0.046	−0.064	0.828	0.762	99.4	97.6
			−0.80	981	−0.073	−0.082	0.955	0.866	98.4	95.9
			−0.60	986	−0.082	−0.073	0.999	0.905	99.2	96.1
			−0.40	988	−0.100	−0.089	1.098	0.989	99.0	95.6
			−0.20	992	−0.090	−0.069	1.174	1.102	98.7	93.5
			0.00	990	−0.132	−0.112	1.125	0.976	99.2	95.8

Table BI. *Continued*.

True values					Bias		MSE × 10		Coverage (per cent)	
I	$\text{logit}^{-1}(\mu)$	$\text{logit}^{-1}(\nu)$	ρ	#d	NLMIXED	inla	NLMIXED	inla	NLMIXED	inla
25	0.80	0.70	−0.95	995	−0.021	−0.022	0.175	0.167	95.3	93.4
			−0.80	999	−0.036	−0.034	0.199	0.196	94.2	92.7
			−0.60	1000	−0.025	−0.018	0.192	0.194	95.2	94.5
			−0.40	1000	−0.028	−0.018	0.182	0.188	95.6	94.7
			−0.20	1000	−0.032	−0.021	0.198	0.205	95.3	94.7
			0.00	1000	−0.034	−0.023	0.189	0.195	95.6	94.9
	0.95	0.30	−0.95	993	−0.021	−0.028	0.287	0.293	97.6	95.3
			−0.80	1000	−0.046	−0.052	0.368	0.371	97.6	94.2
			−0.60	998	−0.044	−0.051	0.392	0.402	96.9	93.0
			−0.40	1000	−0.040	−0.047	0.403	0.416	98.2	93.4
			−0.20	1000	−0.054	−0.062	0.387	0.414	98.5	93.3
			0.00	1000	−0.050	−0.057	0.418	0.438	98.6	93.0

Table BII. Simulation results for the standard deviation σ_ψ. Shown are bias, MSE, and coverage for scenarios with variances $\sigma^2_\phi = 0.5, \sigma^2_\psi = 1$. The results for PROC NLMIXED are based on #d and for inla on 1000 data sets.

True values					Bias		MSE × 10		Coverage (per cent)	
I	$\text{logit}^{-1}(\mu)$	$\text{logit}^{-1}(\nu)$	ρ	#d	NLMIXED	inla	NLMIXED	inla	NLMIXED	inla
10	0.80	0.70	−0.95	974	−0.085	−0.085	0.739	0.708	91.7	91.1
			−0.80	992	−0.083	−0.062	0.719	0.692	91.7	92.1
			−0.60	996	−0.074	−0.037	0.773	0.781	91.6	92.1
			−0.40	997	−0.079	−0.030	0.716	0.742	92.2	93.2
			−0.20	1000	−0.074	−0.021	0.771	0.821	91.9	92.1
			0.00	997	−0.074	−0.019	0.764	0.819	91.9	92.2
	0.95	0.30	−0.95	958	−0.069	−0.060	0.740	0.732	92.6	92.4
			−0.80	981	−0.083	−0.063	0.710	0.698	92.5	92.1
			−0.60	986	−0.087	−0.054	0.748	0.759	91.2	90.8
			−0.40	988	−0.083	−0.044	0.746	0.770	91.6	91.8
			−0.20	992	−0.086	−0.040	0.792	0.821	91.3	91.7
			0.00	990	−0.075	−0.028	0.715	0.759	92.4	93.1
25	0.80	0.70	−0.95	995	−0.028	−0.028	0.286	0.277	92.3	90.5
			−0.80	999	−0.040	−0.030	0.251	0.243	94.9	94.5
			−0.60	1000	−0.032	−0.014	0.290	0.289	92.2	92.9
			−0.40	1000	−0.034	−0.011	0.285	0.288	92.1	92.7
			−0.20	1000	−0.028	−0.004	0.269	0.277	94.2	93.9
			0.00	1000	−0.036	−0.012	0.277	0.281	93.2	94.2
	0.95	0.30	−0.95	993	−0.032	−0.032	0.270	0.265	92.8	92.3
			−0.80	1000	−0.033	−0.026	0.300	0.295	91.8	91.6
			−0.60	998	−0.034	−0.019	0.299	0.298	92.0	92.2
			−0.40	1000	−0.022	−0.004	0.284	0.290	94.4	94.3
			−0.20	1000	−0.015	0.006	0.293	0.304	93.3	93.1
			0.00	1000	−0.038	−0.017	0.295	0.298	92.4	93.1

Acknowledgements

We are grateful to Sara Martino for integrating the model type '2diid' used to perform bivariate meta-analysis into the R-package INLA. We thank the two anonymous reviewers for valuable comments and suggestions on an earlier version of the paper.

References

1. Warn DE, Thompson SG, Spiegelhalter DJ. Bayesian random effects meta-analysis of trials with binary outcomes: methods for the absolute risk difference and relative risk scales. *Statistics in Medicine* 2002; **21**(11):1601–1623.
2. 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**(1):41–51.
3. Irwig L, Macaskill P, Glasziou P, Fahey M. Meta-analytic methods for diagnostic test accuracy. *Journal of Clinical Epidemiology* 1995; **48**(1):119–130.
4. Harbord RM, Deeks JJ, Egger M, Whiting P, Sterne JAC. A unification of models for meta-analysis of diagnostic accuracy studies. *Biostatistics* 2007; **8**(2):239–251.
5. Hamza TH, Reitsma JB, Stijnen T. Meta-analysis of diagnostic studies: a comparison of random intercept, normal-normal, and binomial-normal bivariate summary ROC approaches. *Medical Decision Making* 2008; **28**(5):639–649.
6. Van Houwelingen HC, Arends LR, Stijnen T. Advanced methods in meta-analysis: multivariate approach and meta-regression. *Statistics in Medicine* 2002; **21**(4):589–624.
7. Reitsma JB, Glas AS, Rutjes AWS, Scholten RJPM, Bossuyt PM, Zwinderman AH. Bivariate analysis of sensitivity and specificity produces informative summary measures in diagnostic reviews. *Journal of Clinical Epidemiology* 2005; **58**(10):982–990.
8. 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**(12):1331–1333.
9. Arends LR, Hamza TH, van Houwelingen JC, Heijenbrok-Kal MH, Hunink MGM, Stijnen T. Bivariate random effects meta-analysis of ROC curves. *Medical Decision Making* 2008; **28**(5):621–638.
10. Morris CN. Discussion of Lee and Nelder (1996). *Journal of the Royal Statistical Society, Series B* 1996; **58**(4):669.
11. SAS Institute Inc. The NLMIXED procedure. *SAS/STAT 9.2 User's Guide*. SAS Institute Inc.: Cary, NC., 2008.
12. Zwinderman AH, Bossuyt PM. We should not pool diagnostic likelihood ratios in systematic reviews. *Statistics in Medicine* 2008; **27**(5):687–697.
13. Riley RD, Dodd SR, Craig JV, Thompson JR, Williamson PR. Meta-analysis of diagnostic test studies using individual patient data and aggregate data. *Statistics in Medicine* 2008; **27**(29):6111–6136.
14. Rue H, Martino S, Chopin N. Approximate Bayesian inference for latent Gaussian models using integrated nested Laplace approximations (with Discussion). *Journal of the Royal Statistical Society, Series B* 2009; **71**(2):319–392.
15. Martino S, Rue H. Implementing Approximate Bayesian Inference using Integrated Nested Laplace Approximation: a Manual for the `inla` Program, Department of Mathematical Sciences, NTNU, Norway 2009.
16. R Development Core Team. *R*: *A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria 2009; Available from: http://www.R-project.org, ISBN: 3-900051-07-0.
17. Glas AS, Roos D, Deutekom M, Zwinderman AH, Bossuyt PMM, Kurth KH. Tumor markers in the diagnosis of primary bladder cancer. A systematic review. *Journal of Urology* 2003; **169**(6):1975–1982.
18. 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.
19. Pinheiro JC, Bates DM. Approximations to the loglikelihood function in the nonlinear mixed effects model. *Journal of Computational and Graphical Statistics* 1995; **4**:12–35.
20. METADAS, A SAS macro for meta-analysis of diagnostic accuracy studies. *User Guide Version 1.0 Beta*, December 2008.
21. Rue H, Held L. *Gaussian Markov Random Fields*: *Theory and Applications*. Monographs on Statistics and Applied Probability. vol. 104. Chapman & Hall, CRC Press: London, 2005.
22. 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 using WinBUGS. *Statistics in Medicine* 2005; **24**(15):2401–2428.
23. Bernardo JM, Smith AFM. *Bayesian Theory*. Wiley: Chichester, 2000.
24. Dominici F, Zanobetti A, Zeger SL, Schwartz J, Samet JM. Hierarchical bivariate time series models: a combined analysis of the effects of particulate matter on morbidity and mortality. *Biostatistics* 2004; **5**(3):341–360.
25. Plummer M, Best N, Cowles K, Vines K. *coda*: *Output Analysis and Diagnostics for MCMC* 2008, R package version 0.13-2, 2008.
26. Bachmann LM, Puhan MA, ter Riet G, Bossuyt PM. Sample sizes of studies on diagnostic accuracy: literature survey. *British Medical Journal* 2006; **332**(7550):1127–1129.
27. 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**(12):1–113.
28. Verde PE. Meta-analysis of diagnostic test data: Modern statistical approaches. *Ph.D. Thesis*, University of Düsseldorf, 2008. Available from: http://docserv.uni-duesseldorf.de/servlets/DocumentServlet?id=8494.
29. Rutter CM, Gatsonis CA. A hierarchical regression approach to meta-analysis of diagnostic test accuracy evaluations. *Statistics in Medicine* 2001; **20**(19):2865–2884.
30. 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**(9):925–932.
31. Riley RD, Abrams KR, Lambert PC, Sutton AJ, Thompson JR. An evaluation of bivariate random-effects meta-analysis for the joint synthesis of two correlated outcomes. *Statistics in Medicine* 2007; **26**(1):78–97.
32. Scheidler J, Hricak H, Yu KK, Subak L, Segal MR. Radiological evaluation of lymph node metastases in patients with cervical cancer: a meta-analysis. *Journal of the American Medical Association* 1997; **278**(13):1096–1101.
33. Chu H, Nie L, Cole SR, Poole C. Meta-analysis of diagnostic accuracy studies accounting for disease prevalence: alternative parameterizations and model selection. *Statistics in Medicine* 2009; **28**(18):2384–2399.