Journal of Statistical Software

MMMMMM YYYY, Volume VV, Issue II. *http://www.jstatsoft.org/*

bamdit: An R Package for Bayesian Meta-Analysis of Diagnostic Test Data

Pablo Emilio Verde

Abstract

In this paper we present the R package **bamdit**, its name stands for "**Ba**yesian **m**eta-analysis of **di**agnostic **t**est data". **bamdit** was developed with the aim of simplifying the use of models in meta-analysis, that up to now have demanded great statistical expertise in Bayesian meta-analysis. The package implements a series of innovative statistical techniques including: the Bayesian Summary Receiver Operating Characteristic curve, the use of prior distributions that avoid boundary estimation problems of variances and correlations of random-effects, analysis of conflict of evidence and robust estimation of model parameters. In addition, the package comes with several published examples of meta-analysis that can be used for illustration or further research in this area.

Keywords: meta-analysis, diagnostic test data, hierarchical models, conflict of evidence, bias modeling, MCMC, JAGS, R.

1. Introduction

One of the most important decisions in the presence of illness is the correct medical diagnosis. Ideally, for a particular diagnostic problem we should have a collection of studies which indicate the best way to proceed. However, this is not the case in clinical and other areas of empirical research. Instead, researchers have to face a heterogeneous and fragmented evidence that has to be analyzed.

Meta-analysis is a branch of statistical techniques that helps researchers to combine evidence from a multiplicity of sources. In particular, meta-analysis of diagnostic test data differs from other types of meta-analysis in several aspects: First, the diagnostic summaries that we aim to combine (e.g. sensitivity and specificity) could be interdependent and a marginal combination by pooling these quantities might be misleading (Irwig, Macaskill, Glasziou, and Fahey 1995). Second, diagnostic studies are usually performed under slightly different diagnostic setups and they can be applied to different patients' populations. Hence, we can expect high heterogeneity between studies' results. In addition, the number of studies included might be small and with different qualities (e.g. they might have different study designs) (Lijmer, Mol, Heisterkamp, Bonsel, Prins, van der Meule, and Bossuyt 1999; Lijmer, Bossuyt, and Heisterkamp 2002; Westwood, Whiting, and Kleijnen 2005). Hence, conducting meta-analysis and combining results from diagnostic studies may become a challenge.

In this paper we present the R package **bamdit** (Verde 2013). The name of the package stands for "**Ba**yesian **m**eta-analysis of **di**agnostic **t**est data". The development of the package started with the following question: "How can we make complex meta-analysis in an automatic fashion?"

The initial release of **bamdit** was the version 1.0 of summer 2011. This version was an experimental package where the aim was to investigate different statistical software architectures to fit complex meta-analysis models. During the last years we have rewritten and updated the package several times with the intention of making the package more user-friendly. The current release corresponds to the version 2.5.0 of summer 2016 which is presented in this paper.

The package may be helpful to practitioners who are not familiar with complex Bayesian modeling and who do not have the skills to implement these models in Bayesian software such as WinBUGS/OpenBUGS (Lunn, Spiegelhalter, Thomas, and Best 2009) or JAGS (Plummer 2003).

For more than a decade, meta-analysis of diagnostic tests has been an active area of research, a gentle introduction is given by Gatsonis and Paliwal (2006) and more recently by Takwoingi, Riley, and Deeks (2015). Statistical methods have fallen into two main approaches: On the one hand we have techniques that focus on making a meta-analysis summary by recovering an underlined Receiver Operating Characteristic (ROC) curve. This is the case of the summary ROC (SROC) curve introduced by Moses, Shapiro, and Littenberg (1993) and the hierarchical ROC (HROC) curve presented in Rutter and Gatsonis (1995, 2001); Macaskill (2004).

On the other hand we have approaches that directly model the diagnostic outcomes as a bivariate meta-analysis (Reitsma, Glas, Rutjes, Scholten, Bossuyt, and Zwinderman 2005; Chu and Guo 2009). The relationships between these two approaches have been investigated by Harbord, Deeks, Egger, Whiting, and Sterne (2007) and Arends, Hamza, Van Houwelingen, Heijenbrik, Hunink, and Stijnen (2008) from the classical perspective and by Novielli, Cooper, Sutton, and Abrams (2010) from the Bayesian perspective.

Recent research in meta-analysis of diagnostic test data has focused on the problem of modeling heterogeneity (Verde 2010b), measuring heterogeneity (Zhou and Dendukuri 2014), assessing publication bias (Buerkner and Doebler 2014), modeling results in the presence of imperfect reference standard (Menten, Boelaert, and Lesaffre 2013), using bivariate and trivariate copulas distributions (Kuss, Hoyer, and Solms 2014; Hoyer and Kuss 2015) and non-parametric approaches (Zapf, Hoyer, Kramer, and Kuss 2015).

Software for meta-analysis has been available for many years. The SROC curve approach is implemented in the free software Meta-Disc (Zamora, Abraira, Muriel, Khan, and Coomarasamy 2006), implementations of the bivariate meta-analysis can be found in commercial software such as the function `metandi` in Stata (Harbord and Whiting 2010) and in the macro `MetaDAS` in SAS (Takwoingi and Deeks 2010).

In R (R Core Team 2013), several packages have been developed for different meta-analytic problems. An extensive list with a comprehensive description of these packages is presented in the CRAN task view "Meta-Analysis" (Dewey 2014). In particular the following R packages have been developed for meta-analysis of diagnostic test data: **mada** (Doebler 2015) implements the bivariate method of Reitsma *et al.* (2005) by using the normal approximation of the observed diagnostic rates in the logit scale, this package also offers bivariate meta-regression functionality. **HSROC** (Schiller and Dendukuri 2015) provides a full Bayesian implementation of the hierarchical summary receiver operating characteristic (HSROC) method of Rutter and Gatsonis (2001). **Metatron** (Huang 2014) includes the implementation of the Reitsma *et al.* (2005) model by fitting a bivariate Generalized Linear Mixed-Effects (GLMM) model, the package includes the case of diagnostic tests with an imperfect reference standard. **metamisc** (Debray 2013) implements the method of Riley, Lambert, Staessen, Wang, Gueyffier, Thijs, and Boutitie (2008) which estimates a common within and between correlation when the within-study correlations are unknown. Approximate Bayesian methods using INLA (Integrated Nested Laplace Approximation) can be found in Paul, Riebler, Bachmann, Rue, and Held (2010). This approch is implemented in the package **meta4diag** (Guo and Riebler 2015). In Section 5 we give more detailed information about these R packages.

Implementation of different Bayesian meta-analysis models for diagnostic test data in BUGS software is discussed in Rutter and Gatsonis (2001), Verde (2008, 2010b) and Novielli *et al.* (2010).

The rest of the paper is organized as follows: In Section 2 we describe the software implementation of **bamdit**. In Section 3 we present methodological details of the Bayesian statistical model. In Section 4 we show how to use **bamdit** in practice. In Section 5 we compare **bamdit** with other R packages for meta-analysis of diagnostic test data. Finally, in Section 6 we give a brief summary of the work and we discuss future developments of the **bamdit** package.

2. Software Characteristics

2.1. Software Implementation

In the implementation of **bamdit** we have considered that the package should be easy to use for practitioners familiar with R, but without a Bayesian statistical background.

We also considered that the package has to be portable between different operating systems. **bamdit** uses JAGS for MCMC (Markov Chain Monte Carlo) computations, therefore the main

system requirement is that JAGS ($\geq$ 3.4.0) is installed on your computer (see http://mcmc-jags.sourceforge.net).

It is important to note that R 3.3.0 introduced a major change in the use of toolchain for Windows. This new toolchain is incompatible with older packages written in C++. As a consequence, if the installed version of JAGS does not match the R installation, then the **rjags** package will spontaneously crash. Therefore, if a user works with R $\geq$ 3.3.0, then JAGS must be installed with the installation program JAGS-4.2.0-Rtools33.exe. For users who continue using R 3.2.4 or an earlier version, the installation program for JAGS is the default installer JAGS-4.2.0.exe.

A single function called `metadiag()` performs the meta-analysis. This function allows to fit bivariate Normal random effects or bivariate scale mixture of Normals. The default link function is the logistic link, but the user can choose between the three classical link functions of binomial data: logistic, complementary log-log or probit.

Internally, this function writes the BUGS script and sends the script to JAGS where MCMC computations are performed and returned to R.

The `metadiag()` function is a generic function implemented in S3 object oriented programming in R. The output of the function is an object of the class `metadiag`, which contains results of the MCMC computations, the data used for analysis and further information from the fitted model. Results from a `metadiag` object can be displayed with its `print`, `summary` and `plot` functions. Further statistical details of the model behind **bamdit** is presented in Section 3.

Convergence of the MCMC computations can be analyzed using the R package **coda** (Plummer, Best, Cowles, and Vines 2006). In addition, we have implemented a series of graphical functions that can be used to summarize results and to compare results between models. We demonstrate this software's functionality in Section 4.

2.2. Some Statistical Advantages of Using bamdit

From the statistical point of view, **bamdit** reduces the risk of having boundary problems in the estimation of the variances and the correlation between random effects of the meta-analysis model. In this regard it can be applied to problems where classical approaches fail (see Section 4 and Section 5).

In addition, **bamdit** is equipped with an automatic analysis of conflict of evidence (Verde 2014) which allows to detect studies with unusual results that have been included in the meta-analysis. The user does not need to exclude these studies, the heavy tailed distributions for random-effects implemented in **bamdit** automatically down-weight conflicting studies, which results in a robust Bayesian technique.

The statistical approach implemented in **bamdit** is fully Bayesian (see Section 3). Therefore, the variability of all parameters in the model are propagated to their posteriors. This contrasts with packages that use classical GLMM such as **metatron**, where standard deviations and correlations are calculated by fixing parameter values at the their estimates.

The likelihood contributions of the models implemented in **bamdit** are exact and normal approximations are not required. Packages such us **mada** and **metamisc** use normal approximations of the likelihood contributions. These approximations can be very misleading in studies with small number of patients or meta-analysis of high technology diagnostic tests

where we expect zero outcomes in true positives or true negatives.

A particular value of **bamdit** is that it calculates the marginal and joint posterior predictive distribution of the sensitivity and specificity. As discussed by Higgins, Thompson, and Spiegelhalter (2009), these predictions are the most important summaries in meta-analysis involving random-effects, which can be used to predict results in a new study.

With respect to predictions, most of the **R** packages for meta-analysis of diagnostic tests provide a graphical summary of the predictive contours at given confidence levels (e.g. 50% and 95%). These contours are calculated by assuming that the predictive distribution of random-effects follows a bivariate normal distribution. The `plot` function of **bamdit** implements this option by using the argument `smooth.par = TRUE`. For `smooth.par = FALSE` a non-parametric smoothing of the bivariate distribution of the predictive sensitivities and specificities is displayed.This last option is very useful when the normal distribution of random-effects is not plausible.

3. Bayesian Meta-Analysis of Diagnostic Test Data

3.1. Data Model for Diagnostic Test Results

We assume that the pieces of evidence that we aim to combine are the results of N diagnostic studies, where results of the ith study ($i = 1, \ldots, N$) are summarized in a 2×2 table as follows:

		Patient status	
		With disease	Without disease
Test outcome	+	tp_i	fp_i
	-	fn_i	tn_i
Sum:		$n_{i,1}$	$n_{i,2}$

where tp_i and fn_i are the number of patients with positive and negative diagnostic results from $n_{i,1}$ patients with disease, and fp_i and tn_i are the positive and negative diagnostic results from $n_{i,2}$ patients without disease.

Assuming that $n_{i,1}$ and $n_{i,2}$ have been fixed by design, we model the tp_i and fp_i outcomes with two independent Binomial distributions:

$$tp_i \sim \texttt{Binomial}(\text{TPR}_i, n_{i,1}) \quad \texttt{and} \quad fp_i \sim \texttt{Binomial}(\text{FPR}_i, n_{i,2}), \tag{1}$$

where TPR_i is the true positive rate or sensitivity of study i and FPR_i it the false positive rate or complementary specificity (1-specificity).

At face value, diagnostic performance of each study is summarized by the empirical true positive rate and true negative rate or specificity,

$$\widehat{\text{TPR}}_i = \frac{tp_i}{n_{i,1}} \quad \text{and} \quad \widehat{\text{TNR}}_i = \frac{tn_i}{n_{i,2}} \tag{2}$$

and the complementary empirical rates of false positive rate and false negative diagnostic results,

$$\widehat{\text{FPR}}_i = \frac{fp_i}{n_{i,2}} \quad \texttt{and} \quad \widehat{\text{FNR}}_i = \frac{fn_i}{n_{i,1}}. \tag{3}$$

The main question in meta-analysis of diagnostic test data is: How can we combine the multiplicity of diagnostic accuracy rates in a single coherent model? In this work we recognize that in order to combine results of different studies we have to explicitly model the variability between studies, which is the topic of the next section.

3.2. Random-Effects Model

We model between studies' variability with the following random components:

$$D_i = g(\text{TPR}_i) - g(\text{FPR}_i) \quad \text{and} \quad S_i = g(\text{TPR}_i) + g(\text{FPR}_i), \tag{4}$$

where $g(\cdot)$ corresponds to a link function which maps the diagnostic rates to the real scale $(-\infty, \infty)$. The canonical link function used in this work is the logistic link $g(p) = \log(p/(1-p))$, but other links are also possible (e.g. the complementary log-log link function $g(p) = \log(-\log(1-p))$.

The random component D_i represents the study effect associated with the diagnostic discriminatory power. For example, the logistic link function of D_i corresponds to the diagnostic odds ratio in the logarithmic scale:

$$D_i = \log\left(\frac{\text{TPR}_i}{1-\text{TPR}_i}\right) - \log\left(\frac{\text{FPR}_i}{1-\text{FPR}_i}\right). \tag{5}$$

Meta-analysis based on odds ratios is a common practice for therapeutic outcomes and for diagnostic studies one could also follow this approach. However, diagnostic results are sensitive to diagnostic settings (e.g. the use of different thresholds) and to populations where the diagnostic procedure under investigation is applied. These issues are associated with the *external validity* of diagnostic results.

Following the footsteps of Moses *et al.* (1993), Verde (2010a) introduced the random effect S_i. This random effect quantifies variability produced by patients' characteristics, study design and diagnostic setup, that may produce a correlation between the observed $\widehat{\text{TPR}}$s and $\widehat{\text{FPR}}$s. In short, we called S_i **the threshold effect** of study i and it represents an adjustment of external validity in the meta-analysis.

Conditionally to a study weight w_i, the study effects D_i and S_i are modeled as exchangeable between studies and they follow a *scale-mixture of bivariate Normal* distributions with mean and variance:

$$E\left[\begin{pmatrix} D_i \\ S_i \end{pmatrix} \middle| \, w_i\right] = \begin{pmatrix} \mu_D \\ \mu_S \end{pmatrix}, \quad \texttt{and} \quad var\left[\begin{pmatrix} D_i \\ S_i \end{pmatrix} \middle| \, w_i\right] = \frac{1}{w_i}\begin{pmatrix} \sigma_D^2 & \rho\sigma_D\sigma_S \\ \rho\sigma_D\sigma_S & \sigma_S^2 \end{pmatrix} = \Sigma_i, \tag{6}$$

and scale mixing density

$$w_i \sim p(w_i). \tag{7}$$

The inclusion of the random weights w_i into the model was proposed by Verde (2010a), where $p(w_i)$ allows for a great flexibility to model the marginal distribution of D_i and S_i. Two important cases are: $w_i \sim \chi^2(\nu)$, which corresponds to a marginal bivariate t-distribution

with known degrees of freedom ν, and $p(w_i = 1) = 1$ which corresponds to a bivariate Normal distribution.

In the case of the bivariate t-distribution when the degrees of freedom parameter is fixed to a constant, by integrating w_i from the conditional distribution of $(D_i, S_i|w_i)$ we have a marginal variance of

$$var\left[\begin{pmatrix} D_i \\ S_i \end{pmatrix}\right] = \frac{\nu}{\nu - 2}\begin{pmatrix} \sigma_D^2 & \rho\sigma_D\sigma_S \\ \rho\sigma_D\sigma_S & \sigma_S^2 \end{pmatrix}. \tag{8}$$

In this case, we have to restrict $\nu > 2$ in order to have finite marginal variance in the random effects. However, in the scale mixture of normal distribution we can fix $\nu = 1$ and have a bivariate Cauchy distribution with infinite marginal variances for a particular study. This is equivalent to excluding a study from the meta-analysis.

Another generalization of the random-effects distribution happens when we put a prior on the degrees of freedom parameter ν (see, Section 3.4). This corresponds to an adaptive robust distribution of the random-effects.

The use of the scale mixture of normal distributions as a statistical robust technique has been used in Bayesian statistics for a long time. There is a substantial literature in this area and a good starting point is the recent review by O'Hagan and Pericchi (2012).

3.3. The Directed Acyclic Graph of The Model

Figure 1 displays the Directed Acyclic Graph (DAG) of the model presented in this section. In the usual DAG notation, elliptical nodes represent random variables (parameters and data), rectangular nodes represent fixed parameters, single arrows correspond to stochastic dependencies between nodes and double arrows correspond to deterministic relationships. Model parameters with priors are depicted with dashed ellipses. Repeated structures of the graph are represented by the central plate, where each 2×2 table is modeled as the result of diagnostic parameters (TPR_i and FPR_i) which are the result of random study effects (D_i and S_i). The model of interest is framed with a rectangle containing the hyperparameters of the model $(\mu_D, \mu_S, \sigma_D, \sigma_S, \rho, \nu)$.

The DAG of Figure 1 links the statistical model to the MCMC computations implemented in JAGS. Using an automated theorem proof algorithm, JAGS factorized the joint posterior distribution in a set of conditional distributions which are used for Gibbs sampling. In addition the DAG representation helps to understand how to extend the model of interest. For example, the *pooled Sensitivity* and the *pooled Specificity* are the result of functional parameters of the hyperparameters (see Section 3.7).

3.4. Priors for Hyperparameters

The formulation of the model for aggregate data is completed by specifying the priors for the hyperparameters $\mu_D, \mu_S, \sigma_D, \sigma_S$ and ρ. We assume that parameters are independent and we use the following set of priors:

$$\mu_D \sim \texttt{Logistic}(m_1, v_1), \quad \mu_S \sim \texttt{Logistic}(m_2, v_2) \tag{9}$$

and

$$\sigma_D \sim \texttt{Uniform}(0, u_1), \quad \sigma_S \sim \texttt{Uniform}(0, u_2). \tag{10}$$

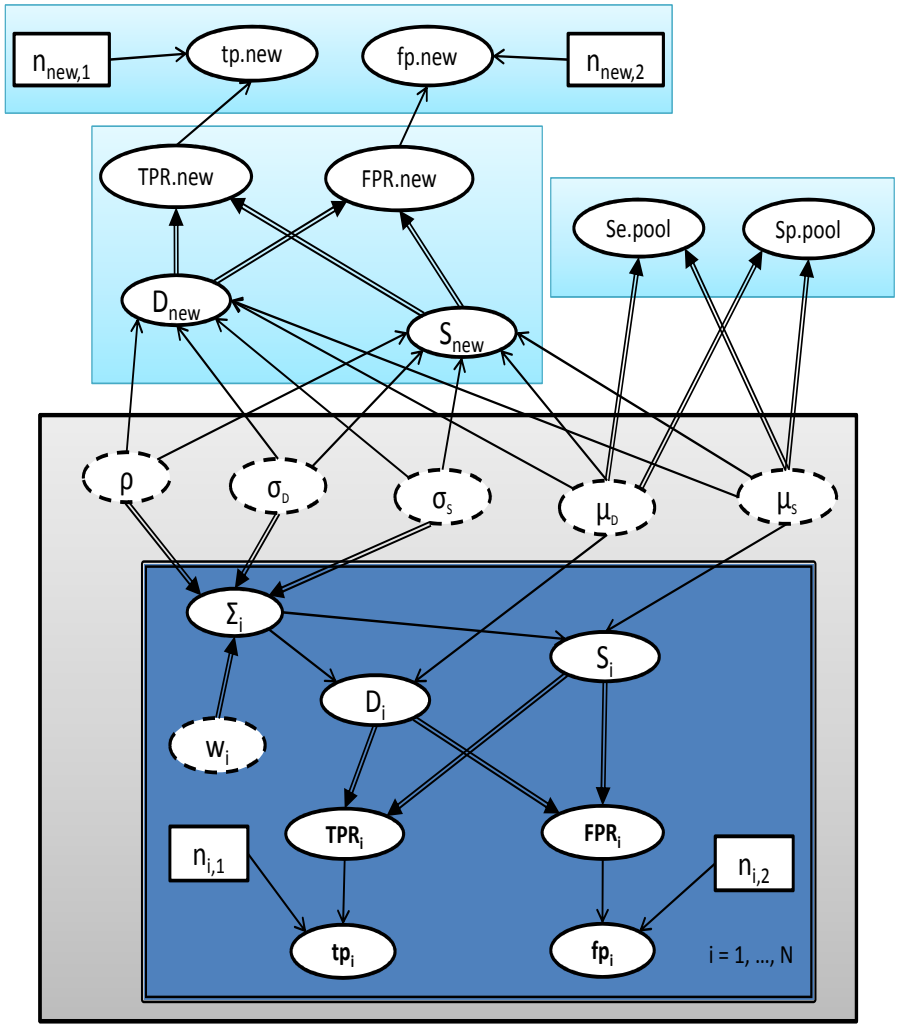

Figure 1: DAG for the model which combines diagnostic accuracy results. Elliptical nodes represent random variables (parameters and data), rectangular nodes represent fixed parameters, single arrows correspond to stochastic dependencies between nodes and double arrows correspond to deterministic relationships. Model parameters with priors are depicted with dashed ellipses. Repeated structures of the graph are represented by the central plate. The model of interest is framed with a rectangle containing the hyperparameters of the model $(\mu_D, \mu_S, \sigma_D, \sigma_S, \rho, \nu)$.

The correlation parameter ρ is transformed by using the Fisher transformation,

$$z = \texttt{logit}\left(\frac{\rho + 1}{2}\right) \tag{11}$$

and a Normal prior is used for z:

$$z \sim \texttt{Normal}(m_r, v_r). \tag{12}$$

Modeling priors in this way guarantees that in each MCMC iteration the variance-covariance matrix of the random effects θ_1 and θ_2 is positive definite. The values of the constants $m_1, v_1, m_2, v_2, u_1, u_2, m_r$ and v_r have to be given. They can be used to include valid prior information which might be empirically available or they could be the result of expert elicitation. If such information is not available, we recommend setting these parameters to values that represent weakly informative priors. In this work, we use $m_1 = m_2 = m_r = 0$, $v_1 = v_2 = 1$ and $v_r = \sqrt{1.7}$ as weakly informative prior setup.

These values are fairly conservative, in the sense that they induce prior uniform distributions for TPR_i and FPR_i. They give locally uniform distributions for μ_1 and μ_2; uniforms for σ_1 and σ_2; and a symmetric distribution for ρ centered at 0. In our experience, the most difficult parameter to estimate in this model is ρ. Therefore, we recommend to make *a prior to posterior sensitivity analysis* by giving different values for m_r and v_r in order to understand their influence in the analysis. Taking $v_r = \sqrt{1.7}$ gives approximately a uniform distribution for ρ between -0.9 and 0.9, and less than 1.5% probability that ρ is less than -0.95 or greater than 0.95. This setup protects the computations from being trapped into impossible values of ρ.

Finally, in the current implementation of **bamdit** we give the following prior to the degrees of freedom ν parameter:

$$U = 1/\nu \tag{13}$$

and a uniform distribution for U:

$$U \sim \texttt{Uniform}(a, b) \tag{14}$$

with $a = 1/df.upper$ and $b = 1/df.lower$. The default values in **bamdit** are $df.lower = 3$ and $df.upper = 30$, this setup allows to explore random-effects distributions that goes from a t-distribution with 3 degrees of freedom to a Normal distribution. This prior is designed to favor long-tailed distributions and to explore conflict of evidence in meta-analysis. In addition, we provide the option to give a fixed value of ν, where the default is $\nu = 4$, or to disable the scale mixture random effects and to use a bivariate Normal distribution in the meta-analysis.

3.5. Interpretation of The Studies' Weights as Measures of Conflict of Evidence

An important aspect of w_i is its interpretation as **estimated bias correction**. *A priori* all studies included in the review have a mean of $E(w_i) = 1$. We can expect that studies which are unusually heterogeneous will have posteriors substantially greater than 1. In **bamdit** we report the posterior $pr(w_i > 1|data)$ to indicate studies' heterogeneity. In our working experience, if this posterior probability is greater than 0.7 a study could be atypical.

In addition, if the model is not corrected by the influence of unusual study results, then the meta-analysis may produce biased results. The use of scale mixtures of random-effects automatically down-weights the influence of outliers in the meta-analysis and produces a robust estimation of the fixed-effects.

Unusual studies' results could be produced by factors that may affect the quality of the study, such as errors in recording diagnostic results, confounding factors, loss to follow-up, etc. For that reason, the studies' weights w_i can be interpreted as an adjustment of studies' **internal validity bias**.

3.6. Splitting the Studies' Weights

In Verde (2014) I conjectured that one way to perform conflict of evidence in a multi-parameter meta-analysis model was to extend the random effects distribution by using a scale mixture of normal distributions per random effect. I have called this technique *"splitting the studies' weights"* and it is implemented in the **bamdit** function `metadiag()` by using the argument `split.w = TRUE`.

The study's weight w_i is now "split" into two components weights $w_{i,1}$ and $w_{i,2}$, these weights measure individual conflict for the components D_i and S_i respectively. For example, if the sources of conflict are studies with unusual specificity the posteriors of $w_{i,2}$ will be further away from a prior mean $E(w_{i,2}) = 1$, while the corresponding posteriors of $w_{i,1}$ will be concentrated around the prior mean.

It is worth mentioning that the mixture of normal distributions introduced by using the splitting argument changes the distribution of the random-effects. For example, if this splitting option is used with a t-distribution with $\nu = 4$, then we look at outliers in two orthogonal directions: the direction of D and S. If the splitting option is not used then the multivariate t-distribution looks at outliers in any direction of the space (D, S).

We report the posteriors $pr(w_{i,1} > 1|data)$ and $pr(w_{i,1} > 1|data)$ to indicate the direction of the studies' heterogeneity. We illustrate how to use this technique in the examples of Section 4.

Conditionally to the study weights $w_{i,1}$ and $w_{i,2}$, the study effects D_i and S_i are modeled as exchangeable between studies. As a common scale mixing density, we use a χ^2 distribution:

$$w_{1,1}, \ldots, w_{N,1}, w_{1,2} \ldots, w_{N,2} \sim \chi^2(\nu), \tag{15}$$

conditionally to the degrees of freedom ν.

3.7. Pooled and Predictive Summaries

In meta-analysis of diagnostic data we are interested in summarizing the overall accuracy of the test in terms of the *pooled Sensitivity* and the *pooled Specificity*.

These quantities are calculated as functions of μ_D and μ_S as following:

$$\text{Sensitivity}^{pooled} = g^{-1}[(\mu_D + \mu_S)/2], \quad \text{Specificity}^{pooled} = 1 - g^{-1}[(\mu_D - \mu_S)/2]. \tag{16}$$

In Figure 1 these quantities are represented as functions of logical nodes, statistical inference is based on sampling from their marginal posterior distributions:

$$p(\text{Sensitivity}^{pooled}|\texttt{Data}) \quad p(\text{Specificity}^{pooled}|\texttt{Data}). \tag{17}$$

Another important summary is the predicted pair of rates (FPR, TPR) for a study that has not been included in the meta-analysis. Statistical inference of these quantities is based on sampling from the bivariate predictive posterior

$$p(\text{TPR}^{new}, \text{FPR}^{new}|\texttt{Data}). \tag{18}$$

In Figure 1 we display how this posterior is built by defining a stochastic node (D^{new}, S^{new}) which is used to calculate $\text{TPR}^{new}, \text{FPR}^{new}$ in each MCMC iteration.

The predictive posterior (18) can be used graphically in order to report the predictive surface at a given credibility level (e.g. 95%). We call this summary the Bayesian Predictive Surface (BPS). Clearly, in this model framework we can calculate the marginal predictive posteriors $p(\text{TPR}^{new}|\texttt{Data})$ and $p(\text{FPR}^{new}|\texttt{Data})$.

The predictive posterior (18) can be used to generate predictive data. This process is described at the top of Figure 1. A total number of patients is fixed in each group n_1^{new} and n_2^{new} and the predictive number of true positive and false positive results is generated by using two independent Binomial distributions with predictive rates $\text{TPR}^{new}, \text{FPR}^{new}$. These predictive data can be used to assess what is expected in a new diagnostic study with n_1^{new} and n_2^{new} patients per group.

Data prediction can be extended to generate N studies with the same number of $n_{i,1}$ and $n_{i,2}$ as the original ones $(i = 1, ..., n)$. The resulting predictive data can be compared with the observed data to assess model misfit.

3.8. Conditional Summaries and the Bayesian SROC Curve (BSROC) and the Bayesian Area Under the Curve (BAUC)

The most common statistical technique used by practitioners to summarize meta-analysis of diagnostic data is the Summary Receiving Operating Characteristic (SROC) curve introduced by Moses *et al.* (1993). The model presented in Section 3 allows to build the Bayesian version of the SROC curve introduced by Verde (2008).

An alternative representation of the marginal model presented in Section 3.2 is the model based on the conditional distribution of $(D_i|S_i = x)$ and the marginal distribution of S_i. The conditional mean of $(D_i|S_i = x)$ is given by:

$$E(D_i|S_i = x) = \text{A} + \text{B}\, x \tag{19}$$

where the functional parameters A and B are

$$\text{A} = \mu_D, \quad \text{and} \quad \text{B} = \rho\frac{\sigma_D}{\sigma_S}. \tag{20}$$

We define the *Bayesian SROC Curve* (BSROC) by transforming back results from (S, D) to (FPR, TPR) with

$$\text{BSROC}(\text{FPR}) = \text{g}^{-1}\left[\frac{A}{(1-B)} + \frac{B+1}{(1-B)}\,\text{g}(\text{FPR})\right]. \tag{21}$$

The BSROC curve is obtained by calculating TPR in a grid of values of FPR which gives a posterior conditionally on each value of FPR. Therefore, it is straightforward to give credibility intervals for the BSROC for each value of FPR.

One important aspect of the BSROC is that it incorporates the variability of the model's parameters, which influences the width of its credibility intervals. In addition, given that FPR is modeled as a random variable, the curve is corrected by measurement error bias in FPR.

Finally, we can define a *Bayesian Area Under the SROC Curve* (BAUC) by numerically integrating the BSROC for a range of values of the FPR:

$$\text{BAUC} = \int_{fpr_0}^{fpr_1} \text{BSROC}(x)\, dx. \tag{22}$$

We recommend to use the limits fpr_0 and fpr_1 within the observed values of $\widehat{\text{FPR}}$s.

We have implemented these conditional summaries in the function `bsroc()`, the function plots the study results with the fitted SROC curve, its credibility intervals and the posterior distribution of the BAUC. We illustrate this functionality in Section 4.

3.9. Further Parametrization of Random-Effects

In the bivariate Normal distribution case, the random effects distribution is similar to the bivariate model introduced by Reitsma *et al.* (2005) where the authors modeled random effects on the logistic transformed sensitivities (se_i) and specificities (sp_i). From equation (4) we have:

$$g(se_i) = (D_i + S_i)/2 \quad g(sp_i) = 1 - (S_i - D_i)/2. \tag{23}$$

Taking $g(\cdot) = \texttt{logit}(\cdot)$ we have the same random effects distribution as in Reitsma *et al.* (2005). However, the likelihood contributions of each study in the Reitsma model are assumed to be approximately normal, while in our model the likelihood contributions are exactly binomial. Moreover, given that our model is a full Bayesian hierarchical model with priors on the hyperparameters (see Section 3.4) the resulting estimation could be different (see Section 5).

The model implemented in **bamdit** generalizes the Reitsma model in the following aspects: by allowing the random-effects to be non-normal; by relaxing the normality assumption of the likelihood contributions; and by introducing priors on hyper-parameters, which reduces the risk of having numerical problems in the estimation of the random-effects distributions (e.g. variances equal to zero or correlations equal to one).

The argument `re.model` in the function `metadiag()` allows to choose between two parametrization of the random-effects: taking `re.model = "SeSp"` parametrizes the model in terms of $(g(se_i), g(sp_i))$ while taking `re.model = "DS"` parametrizes the model as presented in Section 3.2 where the default value is `re.model = "DS"`. The hyperpriors are automatically adapted according to the choice of parametrization.

4. Application of bamdit in Practice

4.1. Example: Diagnostic of bladder cancer

Glas, Lijmer, Prins, Bonsel, and Bossuyt (2003) performed a systematic review to investigate diagnostic procedures for tumor markers used for diagnosing bladder cancer. One of these markers was telomerase, a ribonucleoprotein enzyme, which was evaluated in 10 studies. Riley, Abrams, Sutton, and Thompson (2007) used this example to present issues regarding boundary problems in the estimation of the correlation between random effects. Paul *et al.* (2010) illustrate the use of INLA computations in this example as well.

Looking at The Data

The data of this meta-analysis can be found in the `glas` data frame in **bamdit**. We can have a quick view of the different subgroups of markers by using the function `plotdata()`, here we present some of its functionality:

```
R> library(bamdit)
R> data(glas)
R> head(glas)
```

```
  tp n1 fp  n2   Author cutoff(U/ml) marker
1  1  2 15  52 Kirollos         <NA>    BTA
2 17 60  9  70 Johnston         <NA>    BTA
3  8 28  7  34   Murphy         <NA>    BTA
4 19 47  8  30  Landman         <NA>    BTA
5 33 41 27 304     Leyh         <NA>    BTA
6  8 12 12  35    Chong         <NA>    BTA
```

```
R> plotdata(glas,                  # Data frame
+           group = glas$marker, # groupping variable
+           max.size = 5)        # scale of circles
```

We extract the subset of studies which have reported results by using the telomerase marker:

```
R> glas.t <- glas[glas$marker == "Telomerase", 1:4]
```

and we plot this subgroup by

```
R> plotdata(glas.t)
```

Fitting Bayesian Meta-Analysis Models

A single function called `metadiag()` is used to fit different type of Bayesian meta-analysis models. Below we illustrate some of the arguments of this function. For example, to fit a model, with bivariate Normal distribution with logistic link function, and random effects on D_i and S_i type:

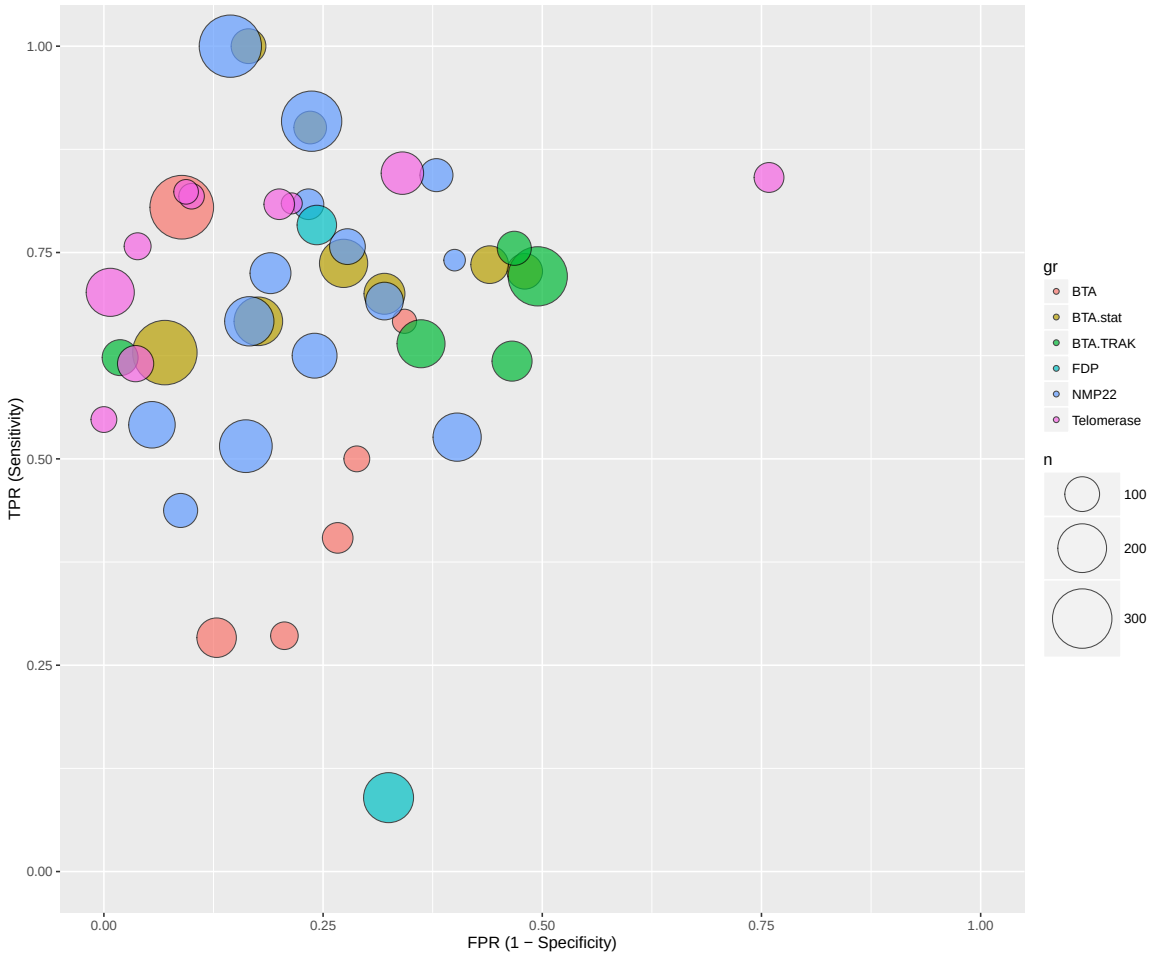

Figure 2: Display of the meta-analysis results of the data frame glas: each circle identifies the true positive rate vs. the false positive rate of each study. Different colours are used for different markers and different sizes for sample sizes.

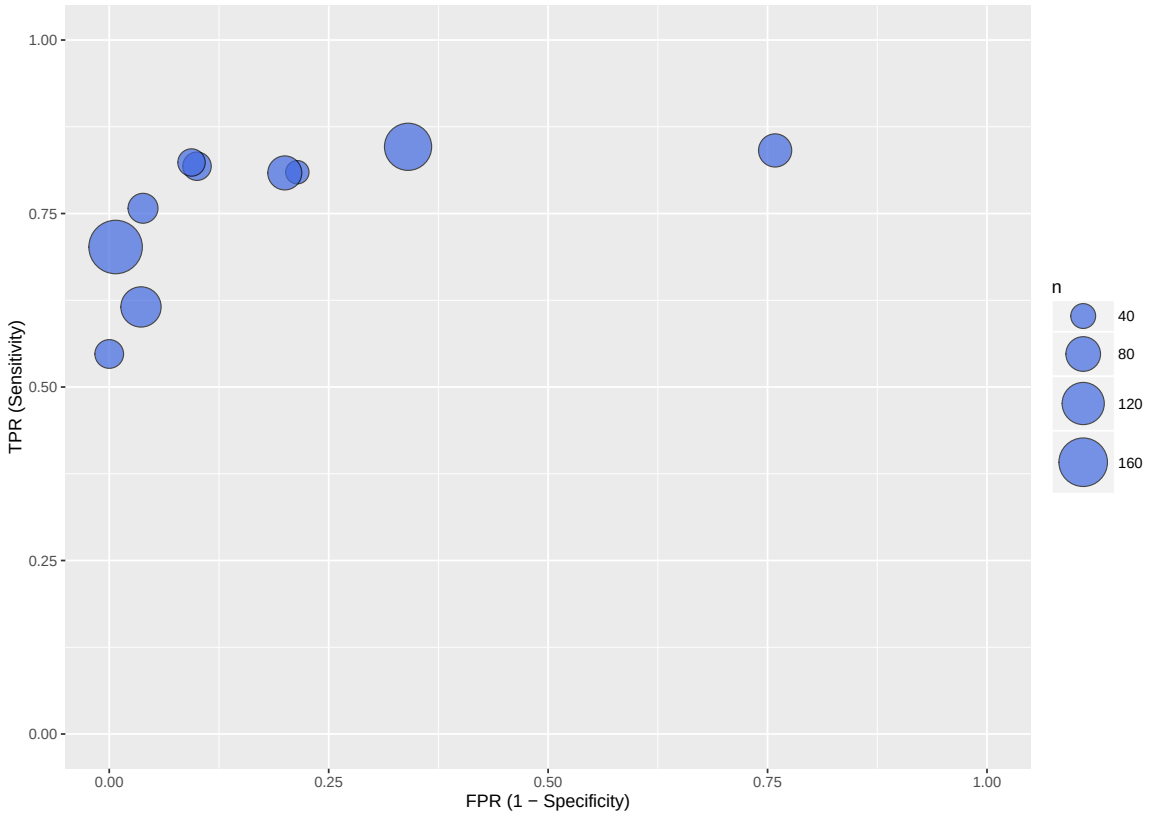

Figure 3: Display of the meta-analysis results of studies with the telemerase marker in the data frame glas.

```
R> glas.m1 <- metadiag(glas.t,                 # Data frame
+                      re = "normal",         # Random effects distribution
+                      re.model = "DS",       # Random effects on D and S
+                      link = "logit",        # Link function
+                      sd.Fisher.rho   = 1.7, # Prior standard deviation of correlation
+                      nr.burnin = 1000,      # Iterations for burnin
+                      nr.iterations = 10000, # Total iterations
+                      nr.chains = 4,         # Number of chains
+                      r2jags = TRUE)         # Use r2jags as interface to jags

module glm loaded

Compiling model graph
   Resolving undeclared variables
   Allocating nodes
Graph information:
   Observed stochastic nodes: 20
   Unobserved stochastic nodes: 28
   Total graph size: 314

Initializing model
```

To see the results of this computations just print the object by typing:

```
R> summary(glas.m1, digits = 3)

Inference for Bugs model at "5", fit using jags,
 4 chains, each with 10000 iterations (first 1000 discarded)
 n.sims = 36000 iterations saved
           mean     sd   2.5%    25%    50%    75%  97.5% Rhat n.eff
deviance 80.287  5.503 71.372 76.327 79.687 83.580 92.791    1  7700
fp.new   11.973 13.124  0.000  2.000  7.000 18.000 47.000    1 36000
mu.D      3.109  0.526  2.018  2.784  3.123  3.449  4.128    1 13000
mu.S     -0.573  0.727 -1.970 -1.047 -0.588 -0.119  0.917    1  6300
rho      -0.862  0.144 -0.991 -0.956 -0.907 -0.821 -0.465    1  3700
se.new    0.758  0.126  0.444  0.695  0.778  0.845  0.944    1 18000
se.pool   0.778  0.042  0.691  0.753  0.779  0.805  0.854    1 10000
sigma.D   1.531  0.534  0.769  1.161  1.439  1.792  2.836    1  1900
sigma.S   2.443  0.746  1.406  1.934  2.313  2.807  4.214    1  4700
sp.new    0.760  0.258  0.079  0.643  0.865  0.958  0.997    1 36000
sp.pool   0.849  0.077  0.656  0.814  0.864  0.902  0.951    1  3400
tp.new   37.876  6.967 21.000 34.000 39.000 43.000 48.000    1 19000

For each parameter, n.eff is a crude measure of effective sample size,
and Rhat is the potential scale reduction factor (at convergence, Rhat=1).
```

```
DIC info (using the rule, pD = var(deviance)/2)
pD = 15.1 and DIC = 95.4
DIC is an estimate of expected predictive error (lower deviance is better).
```

We can see that hyper-parameters, like the component of variances (σ_D and σ_S) and the correlation between random effects (ρ) are estimated without boundary problems.

If we need to directly calculate the correlation between the pooled sensitivity and the pooled specificity, then we can attach the object `glas.m1` by using the package **R2jags** and directly calculate the correlation:

```
R> library(R2jags)
R> attach.jags(glas.m1)
R> cor(se.pool, sp.pool)

           [,1]
[1,] -0.431
```

Displaying meta-analysis summaries

It is very useful to display the the Bayesian Predictive Surface by contours at different credibility levels and compare these curves with the observed data. The function `plot` displays parametric or non-parametric predictive contours:

```
R> plot(glas.m1,                    # Fitted model
+      level = c(0.5, 0.75, 0.95), # Credibility levels
+      parametric.smooth = TRUE)   # Parametric curve
```

The function `plotsesp()` is a user friendly function in **bamdit** which displays the posterior distribution of the pooled sensitivity and specificity and their predictive posteriors. We can display these posteriors as follows:

```
R> plotsesp(glas.m1)
```

Figure 5 shows the output, clearly the low number of studies influence the ability to predict the result of a future study.

The BSROC curve and its area under the curve are useful summaries of a meta-analysis, we can easily display these summaries by using the function `bsroc()` as follows:

```
R>  bsroc(glas.m1,                        # Fitted model
+        level = c(0.025, 0.5, 0.975),  # Credibility levels
+        plot.post.bauc = TRUE,         # Include the posterior of the AUC
+        fpr.x = seq(0.01, 0.75, 0.01), # Grid of values for FPR
+        lower.auc = 0,                 # Lower limit for the BAUC
+        upper.auc = 0.99)              # Upper limit for the BAUC
```

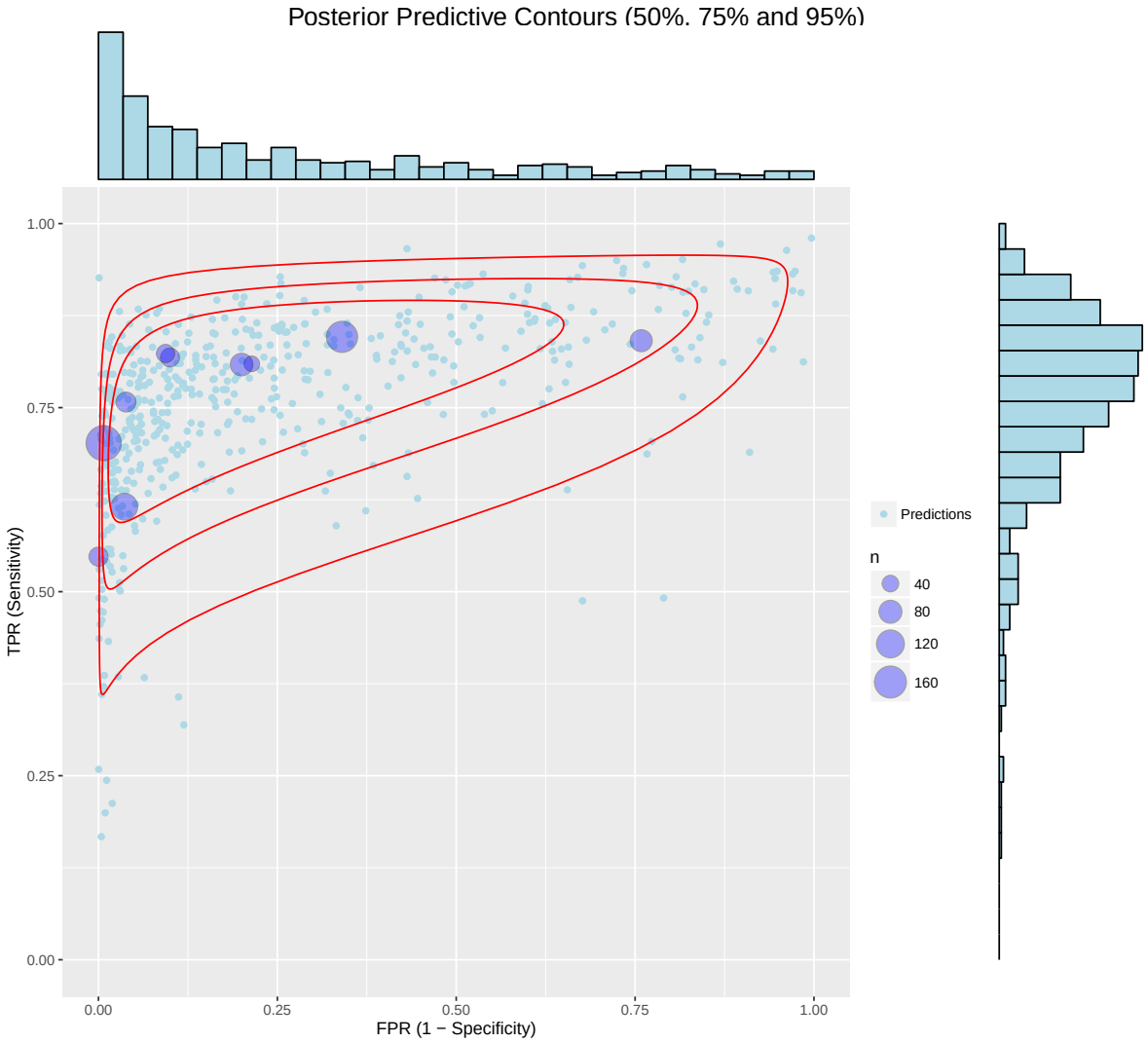

Figure 4: Results of the meta-analysis: Bayesian Predictive Surface by contours at different credibility levels.

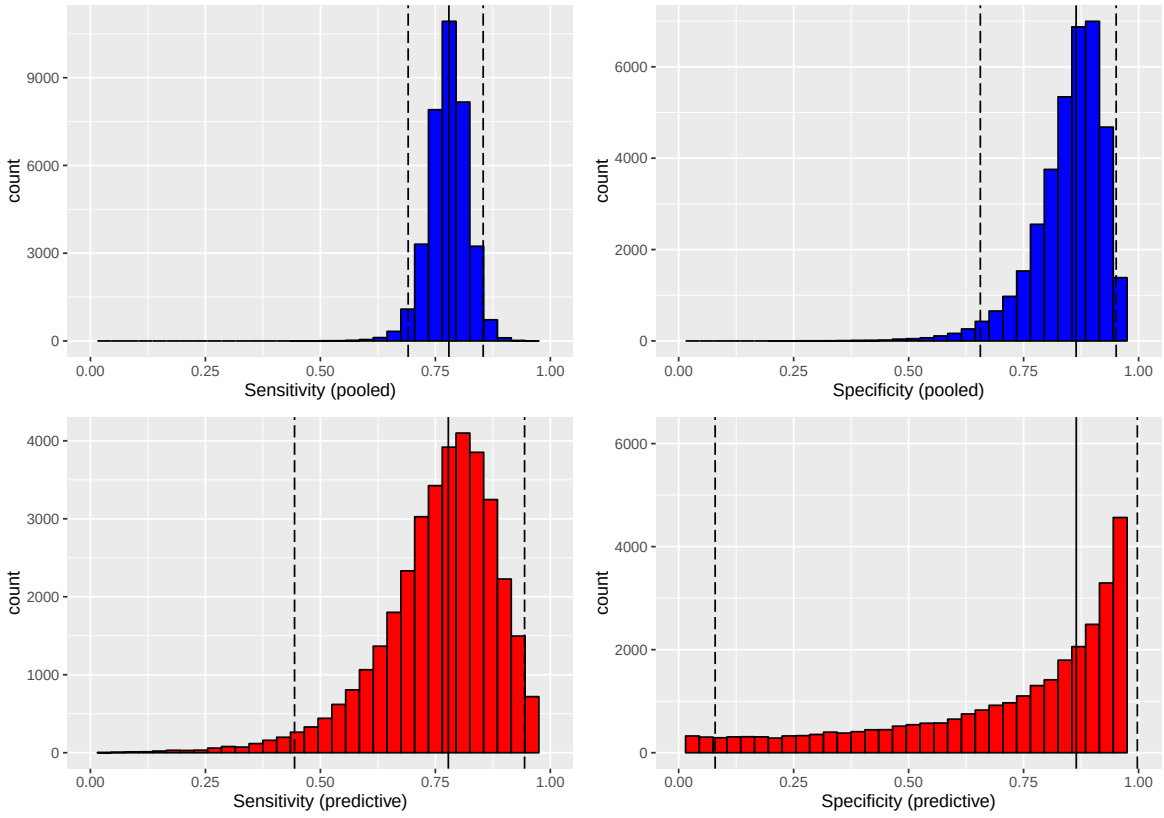

Figure 5: Results of the meta-analysis: Posterior distributions for the pooled sensitivity and specificity and their predictive posteriors.

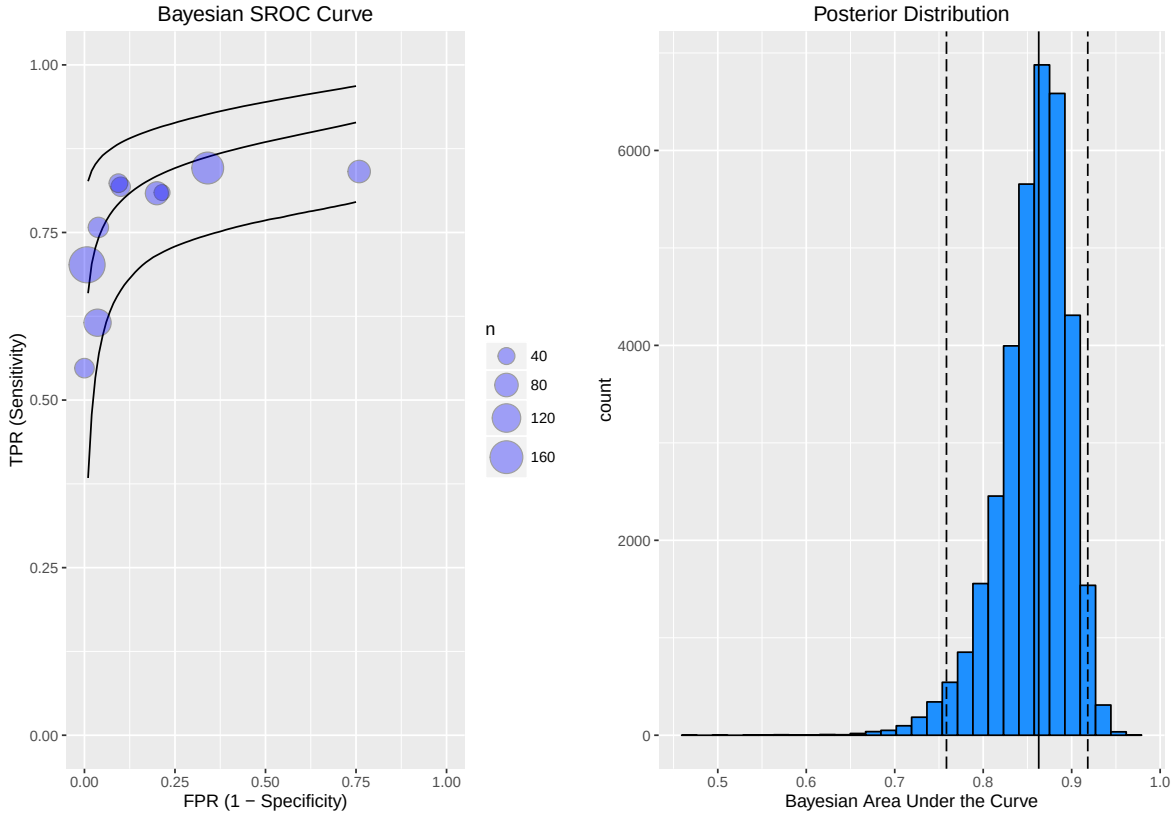

Figure 6: Conditional summaries: Left panel shows the BSROC curve, the central line corresponds to the posterior median and the upper and lower curves correspond to the quantiles of the 2.5 and 97.5 precent respectively. The right panel displays the posterior distribution of the area under the BSROC curve.

```
Summary results for the Bayesian Area Under the Curve (BAUC)
------------------------------------------------------------
 2.5%   25%   50%   75% 97.5%
0.758 0.836 0.863 0.885 0.918
------------------------------------------------------------
```

Interestingly, the BAUC results and the BSROC , which is displayed in Figure 6, show promising diagnostic ability of this marker.

Hyper-Parameters Posteriors

If we are interested in visualizing the posterior distributions of all hyper-parameters simultaneously, we can use one of the alternative matrix plot function in R. For example, we can use the `ggpairs()` function from the package **GGally** as follows:

```
R> library(ggplot2)
```

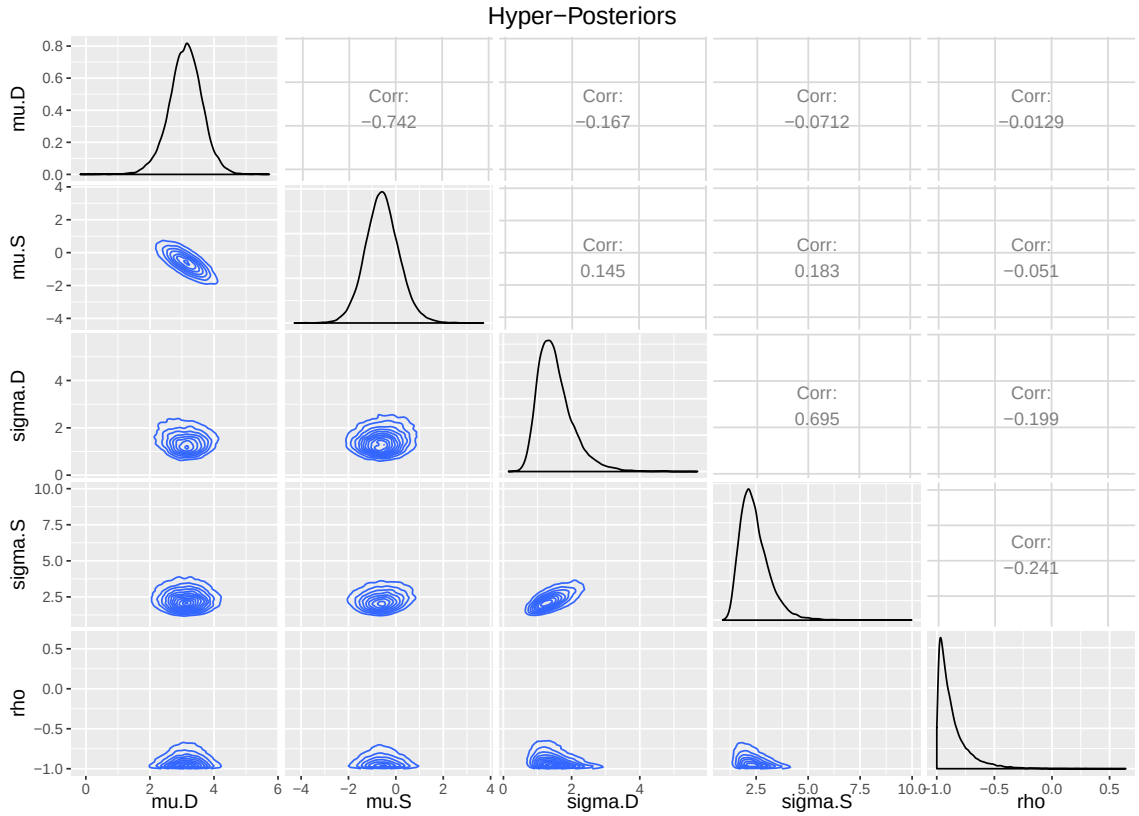

Figure 7: Posterior distributions for the hyperparameters of the model.

```
R> library(GGally)
R> library(R2jags)

R> attach.jags(glas.m1)
R> hyper.post <- data.frame(mu.D, mu.S, sigma.D, sigma.S, rho)
R>
R> ggpairs(hyper.post,                    # Data frame with MCMC realizations
+ title = "Hyper-Posteriors",            # title of the graph
+ lower = list(continuous = "density")) # contour plots
```

In Figure 7 we can also see in the lower diagonal panels the correlation structure of this multivariate posterior. The main diagonal of this matrix plot contains the posterior densities of each parameter. One interesting aspect is the posterior of the correlation coefficient ρ, which clearly shows a negative correlation in the random-effects.

Conflict of Evidence Analysis by Using Scale Mixtures Random-Effects

We can fit a model with scale mixtures as random effects to investigate if there are conflict of evidence between the studies included in the systematic review. The following code gives an example:

```
R> glas.m2 <- metadiag(glas.t, # Data frame
+               re = "sm",     # Scale mixture of normals
+             link = "logit",  # Link function
+    sd.Fisher.rho = 1.7,      # Prior standard deviation of correlation
+       df.estimate = TRUE,    # Degrees of freedom estimated from the data
+          split.w = TRUE,     # Different weights for each component
+        nr.burnin = 1000,     # Iterations for burnin
+    nr.iterations = 20000,    # Total iterations
+        nr.chains = 1,        # Number of chains
+           r2jags = TRUE)     # Use r2jags as interface to jags
```

The results are printed as usual:

```
R> glas.m2
```

```
Inference for Bugs model at "6", fit using jags,
 1 chains, each with 20000 iterations (first 1000 discarded)
 n.sims = 19000 iterations saved
         mean   sd 2.5%  25%  50%  75% 97.5%
df        7.6  4.2  3.1  4.2  6.2 10.0  18.0
fp.new   14.4 13.5  0.0  4.0 10.0 22.0  47.0
mu.D      2.6  0.5  1.4  2.3  2.6  2.9   3.5
mu.S     -0.1  0.6 -1.3 -0.5 -0.1  0.4   1.3
...
p.w2[5]   0.7  0.4  0.0  0.0  1.0  1.0   1.0
p.w2[6]   0.5  0.5  0.0  0.0  0.0  1.0   1.0
p.w2[7]   0.7  0.4  0.0  0.0  1.0  1.0   1.0
...
p.w2[10]  0.7  0.5  0.0  0.0  1.0  1.0   1.0
...
rho      -0.9  0.1 -1.0 -1.0 -0.9 -0.8  -0.5
se.new    0.8  0.1  0.4  0.7  0.8  0.8   0.9
se.pool   0.8  0.0  0.7  0.8  0.8  0.8   0.9
sigma.D   1.5  0.7  0.7  1.1  1.4  1.8   3.2
sigma.S   2.2  0.8  1.1  1.7  2.1  2.6   4.2
sp.new    0.7  0.3  0.1  0.6  0.8  0.9   1.0
sp.pool   0.8  0.1  0.5  0.7  0.8  0.8   0.9
...
w2[5]     2.0  2.7  0.5  1.0  1.4  2.2   7.4
w2[6]     1.3  1.3  0.4  0.7  1.0  1.4   3.8
w2[7]     2.2  3.4  0.5  1.0  1.4  2.3   8.8
...
pD = 16.1 and DIC = 96.6
```

Although this model shows similar results as the model with bivariate normal random effects, there is about 5% of reduction of the standard deviations of the pool summaries and we have the additional information coming from the posterior weights. The posterior probability that

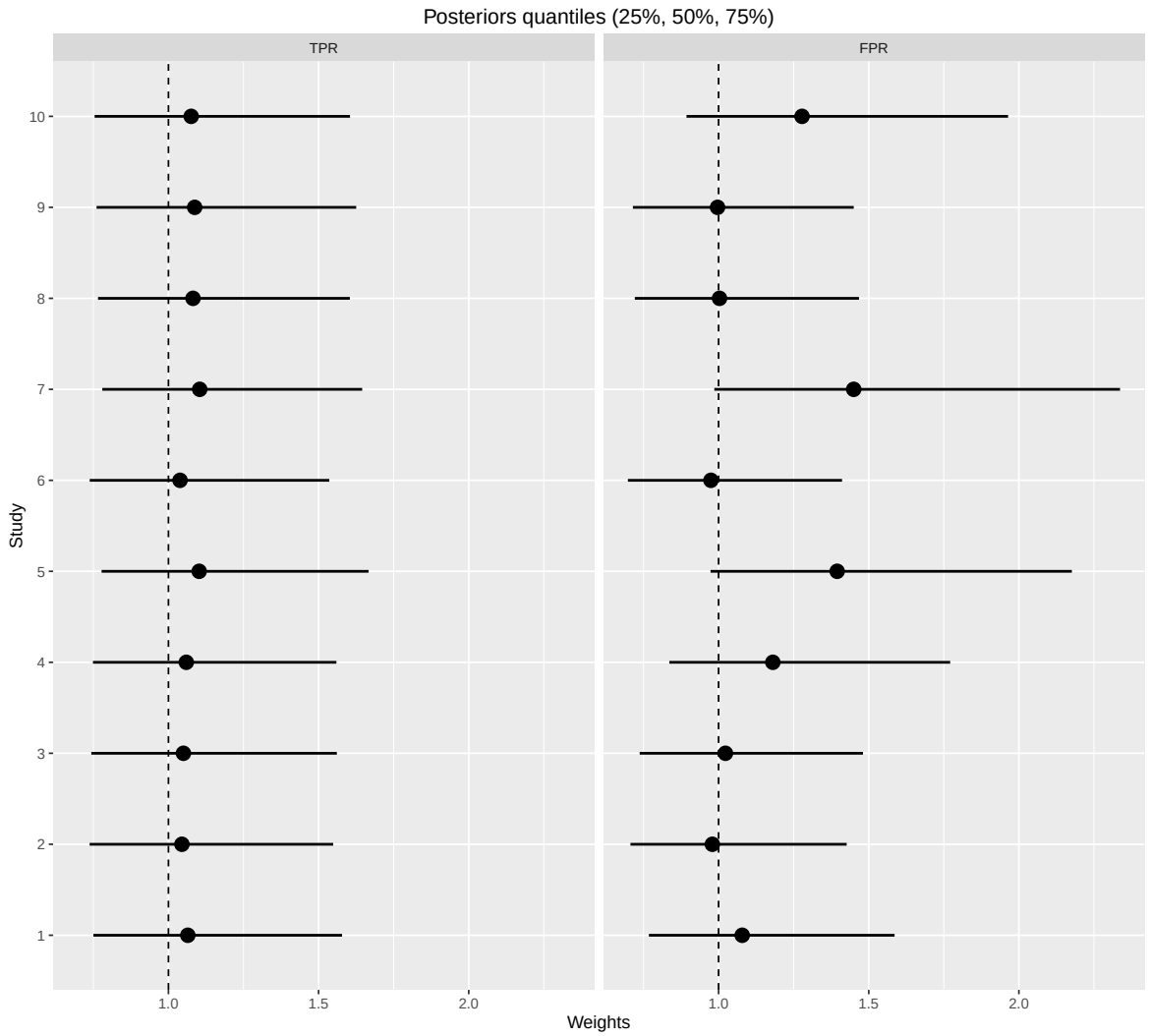

Figure 8: Posterior distributions of the component weights: It is expected that the posterior is centered at 1. Studies 5 and 7 showed a moderate deviation and Study 10 a clear deviation.

w_2 is greater than one is over 0.7 for observations 5, 7 and 10. Than indicates that those studies may contain unusual results. The function `plotw` plots a the posteriors of the weights:

```
R>   plotw(m = glas.m2)
```

Figure 8 summarize the results of the component weights w_1 and w_2. If the bivariate normal random effects is correct, then we expect that the posteriors are centered at 1. Studies 5 and 7 showed a moderate deviation and Study 10 a clear deviation. We can print the original data to explain these results:

```
R> glas.t[c(5, 7, 10), ]

   tp n1 fp  n2
38 40 57  1 138
```

```
40 23 42  0  12
43 37 44 22  29
```

and calculate the empirical rates

```
R> dat.hat <- data.frame(tpr = glas.t[ , 1]/glas.t[ , 2],
+                        fpr = glas.t[ , 3]/glas.t[, 4],
+                        n = glas.t[ , 2] + glas.t[ , 4])
R> dat.hat[c(5, 7, 10), ]
```

```
     tpr     fpr   n
5  0.702 0.00725 195
7  0.548 0.00000  54
10 0.841 0.75862  73
```

Studies 5 and 7 have a very low false positive rate, maybe too low to be true! Study 10 has over 75% false positive rate, which is extreme for these data. We can use the function `plotcompare()` to display the differences between two models with respect to the predictive posterior contours:

```
R> plotcompare(m1 = glas.m1,                              # Model 1 object
+              m2 = glas.m2,                              # Model 2 object
+              m1.name = "Binomial + Normal",             # Label for Model 1
+              m2.name = "Binomial + Scale mixtures", # Label for Model 2
+              level = 0.95)
```

Figure 9 shows that the model with the scale mixture random effects extends the predictive contours in the lower direction of sensitivity and in the upper direction of the false positive rate.

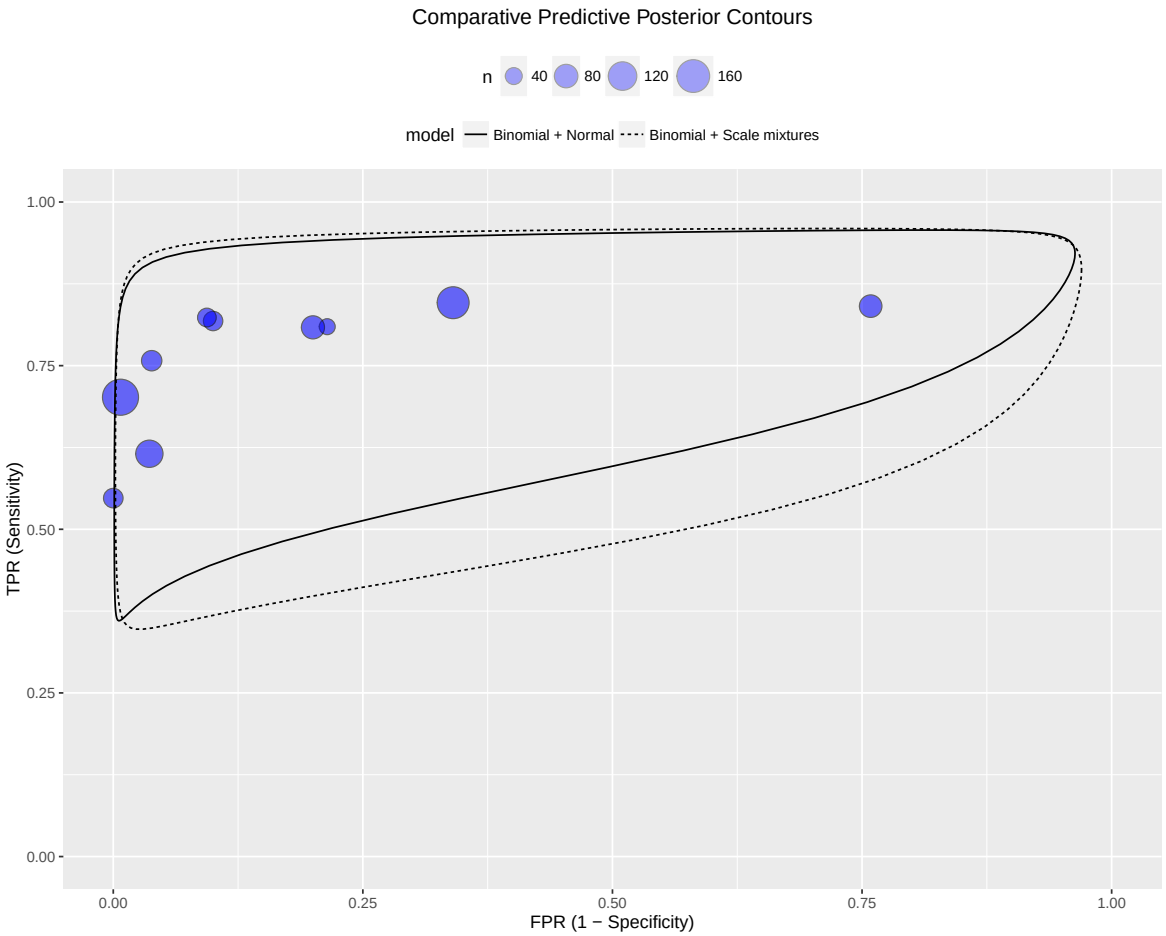

Figure 9: Comparative results of the Bayesian Predictive Surface at the 95 percent credibility level. The Normal random effects model corresponds to the solid line and the scale mixtures of random effects to the dotted line.

Computer Tomography (CT) Scans in The Diagnosis of Appendicitis

This example refers to a meta-analysis 51 studies investigating the accuracy performance of Computer Tomography (CT) scans in the diagnosis of appendicitis Verde (2008).

One characteristic of this meta-analysis is the combination of disparate data. From the 51 studies 22 were retrospective and 29 were prospective. Verde (2008) analyzed this characteristic and found that retrospective studies had substantial more heterogeneity than prospective ones, which led to the structural dispersion model of Verde (2010a). Recently, Zhou and Dendukuri (2014) used this data to illustrate measurement heterogeneity in a bivariate random effects meta-analysis.

Looking at the data

The data of this meta-analysis can be found in the `ct` data frame in **bamdit**. In addition to the test performance results, this data frame contains information about study characteristics, patient characteristics, study design, and diagnostic setup.

```
R> data(ct)
R> gr <- with(ct, factor(design,
+                        labels = c("Retrospective study", "Prospective study")))
R>
R> plotdata(ct,                 # Data frame
+           group = gr,         # Grouping variable
+           y.lo = 0.75,        # Lower limit of y-axis
+           x.up = 0.75,        # Upper limit of x-axis
+           alpha.p = 0.5,      # Transparency of the balls
+           max.size = 5)       # Scale the circles
```

Analyzing conflict of evidence of studies with different design

We analyze these data to show how to compare the posterior weights for different groups of studies. In the following example we compare these posteriors by using the function `plotw`. We give to the argument `group` the factor variable which indicates if a study has prospective or retrospective design.

```
R> ct.m <- metadiag(ct,
+                   re = "sm",      # Scale mixture of normals
+                 link = "logit",   # Link function
+       df.estimate = TRUE,         # Degrees of freedom estimated from the data
+           split.w = TRUE,         # Different weights for each component
+         nr.burnin = 1000,         # Iterations for burnin
+     nr.iterations = 10000,        # Total iterations
+         nr.chains = 4,            # Number of chains
+            r2jags = TRUE)         # Use r2jags as interface to jags

R> plotw(m = ct.m,                 # The fitted model
+       group = gr                  # The groupping factor
+       )
```

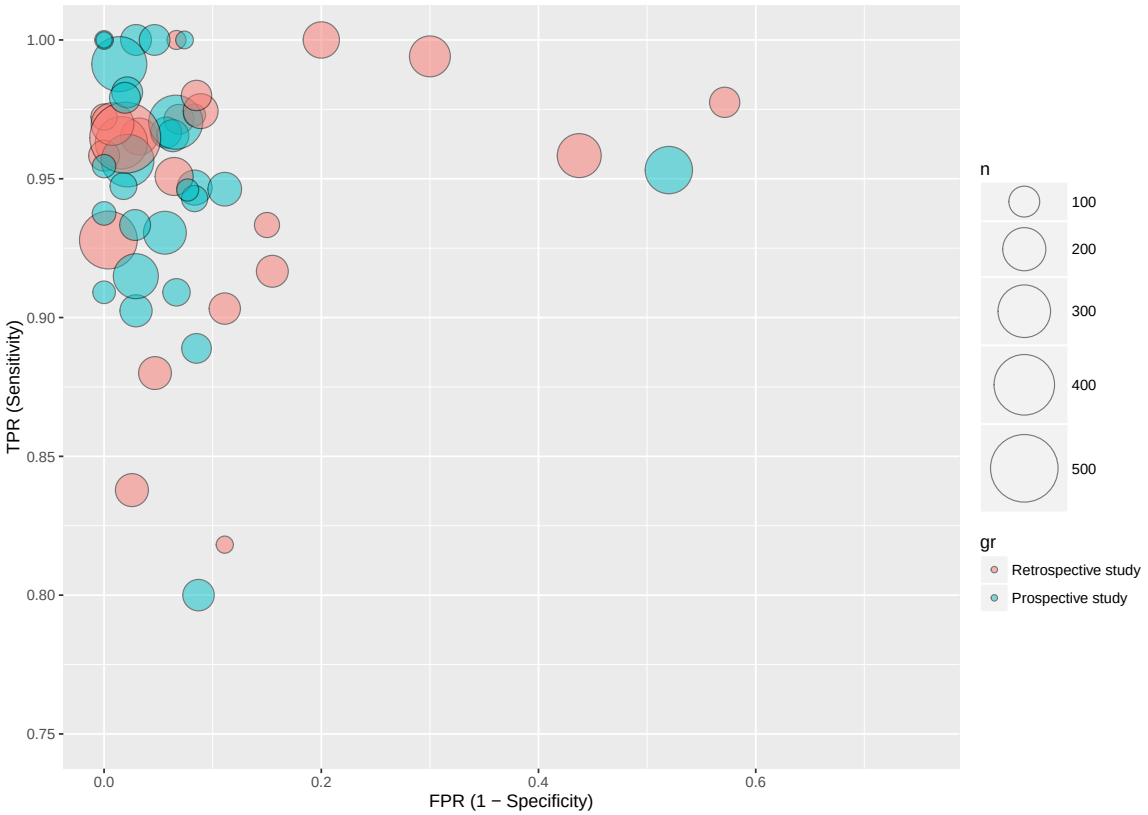

Figure 10: Display of the meta-analysis results of the data frame ct: Each circle identifies the true positive rate vs. the false positive rate of each study. Different colours are used for different study designs and different sizes for sample sizes.

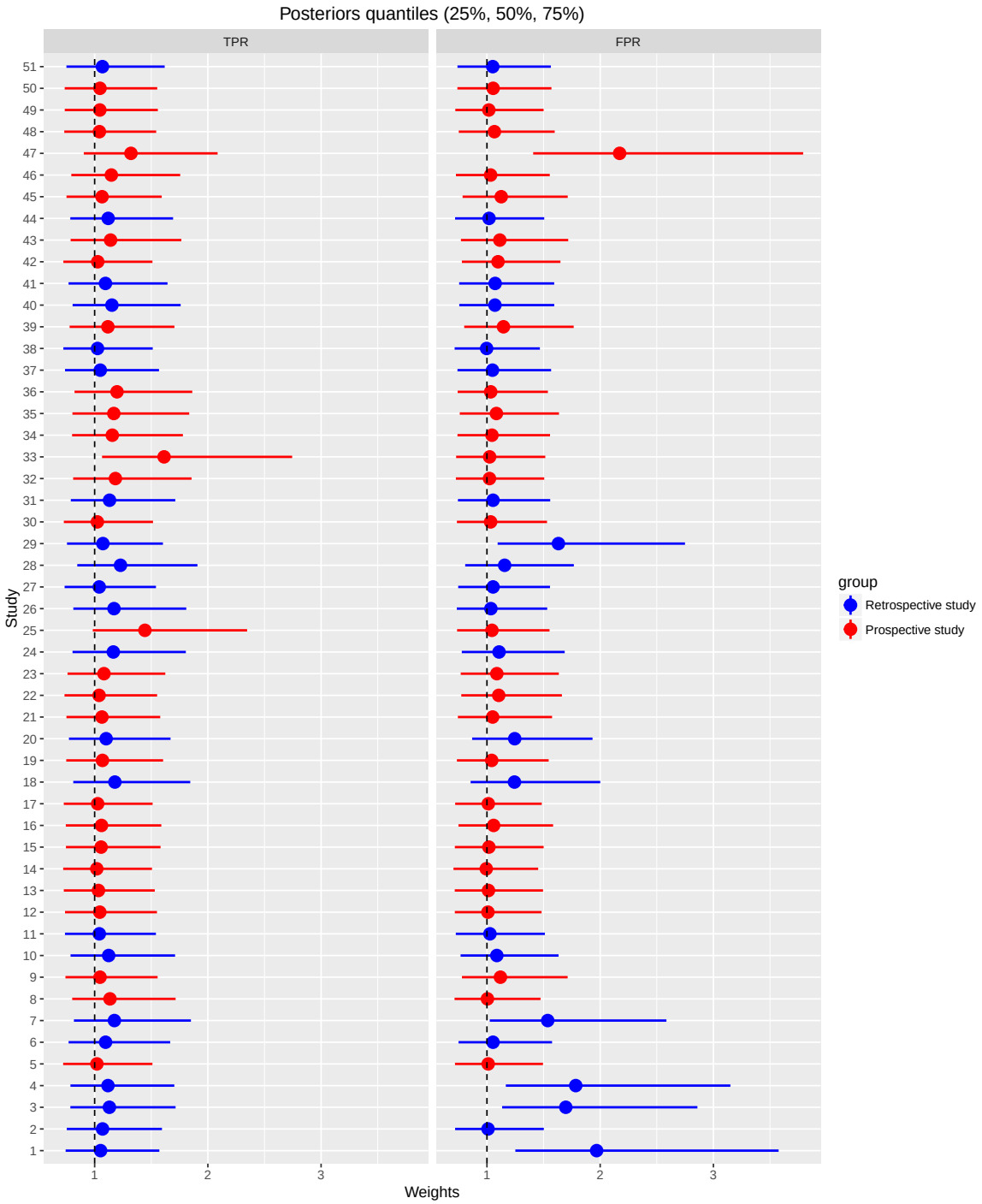

Figure 11: Posterior distributions of the component weights: It is expect that the posterior is centered at 1. Studies with retrospective design tend to present deviations in FPR.

Figure 11 displays the posteriors of each components' weights. The right panel shows that prospective studies number 25 and 33 deviate with respect to the prior mean of 1, while on the left panel we see that a prospective study (number 47) and five retrospective studies have substantial variability.

The function `plotcompare()` can be used to compare the predictive differences between retrospective and prospective studies:

```
R> m1.ct <- metadiag(ct[ct$design==1, 1:4]) # Restrospective studies
R> m2.ct <- metadiag(ct[ct$design==2, 1:4]) # Prospective studies
R> plotcompare(m1.ct, m2.ct,
+             m1.name = "Retrospective design",
+             m2.name = "Prospective design",
+               group = gr,
+            limits.x = c(0, 0.75), limits.y = c(0.65, 1))
```

Finally, Figure 12 presents the 95% predictive posterior contours for studies with retrospective and prospective design, we can clearly see the effects of study design in the meta-analysis. In synthesis, retrospective studies are less specific and more uncertain than prospective ones.

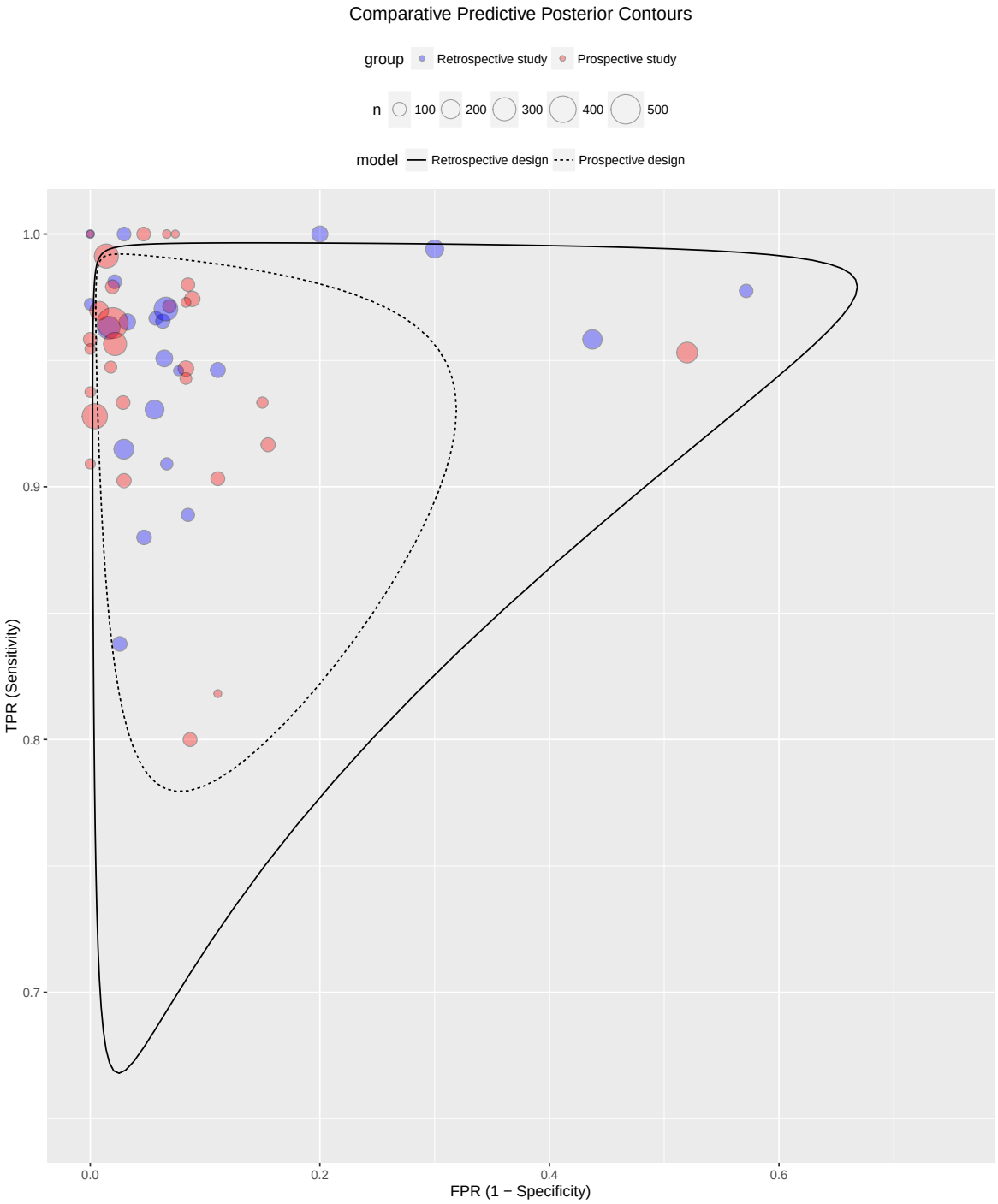

Figure 12: Predictive posteriors contours at 95 credibility level: Two models with Normal random effects are fitted to studies with retrospective (blue points) and prospective (red points) design.

5. Comparison with other R packages

The aim of this section is to present a brief comparison of results between different R packages that can be used for meta-analysis of diagnostic tests. For this aim we use the example of Glas *et al.* (2003) presented in Section 4.

Different packages implement different parametrization of random-effects distributions, they use different estimation techniques, different numerical procedures and different inferential approaches. Therefore, in order to make results comparable we harmonize results in the following way: We use the logistic link function for sensitivity and specificity, the bivariate normal distribution is used for random-effects, and the parametrization of random-effects is based on sensitivity and specificity in the logistic scale. We apply the default settings for all functions and we present results with three significant decimal digits.

For example **metatron** uses sensitivity and false positive rate, therefore the estimated correlation was multiplied by -1 to obtain the correlation between sensitivity and specificity in the logistic scale. The same was used for **mada** and **metamisc**. The R script from Kuss *et al.* (2014) presents results in the probability scale, so we use the delta-method to back-transform the variances in the logistic scale. The package **HSROC** implements the calculations in the probit scale, therefore we re-scale results and we use the formulas (4.16)-(4.20) from Harbord *et al.* (2007) for means, variances and correlation.

Table 1 summarizes the results of our analysis. Of course the results are not conclusive and it is not the intention to make a systematic comparison between packages, but we can give the following remarks:

- *Estimation of the pooled means:* The parameters μ_{SE} and μ_{SP} represent the pooled sensitivity and specificity in the logistic scale respectively. All packages estimate μ_{SE} similarly, with respect to μ_{SP} results are similar too, but **metatron** tends to overestimate this parameter. In the probability scale we can see that the pooled sensitivity and specificity are approximately similar accross the packages, with the exception again of **metatron**.

- *Estimation of the standard deviations:* The parameters σ_{SE} and σ_{SP} are the standard deviations of the random-effects of the sensitivity and specificity in the logistic scale. Looking across the results we clearly see that **metatron** gives implausible values close to zero for both parameters. As we saw in Section 4 there is an outlier in the direction of the specificity, which makes the estimates of σ_{SP} more variable across packages, **bamdit** gives the larger value of 2.207 and the **HSROC** the smallest value 1.422.

- *Estimation of the correlations:* As mentioned in Section 3.4 the correlation parameter ρ is the most difficult parameter to estimate. The package **mada** gives the impossible value -1. The same happens with **metatron** with a value close to 0. The other packages managed to estimate ρ between -0.857 to -0.615.

- *Bayesian approaches:* There are similarities between results of **meta4diag** and **bamdit**. This is not by chance, both packages implement the same model for random-effects. The differences come from the priors used for the variance covariance matrix, **meta4diag** uses a Wishart distribution while **bamdit** uses a conditional model to implement the priors. **HSROC** deliver similar results for the pooled estimates of sensitivity and specificity with posterior intervals similar to **meta4diag** and **bamdit**.

package	μ_{SE}	μ_{SP}	σ_{SE}	σ_{SP}	ρ	*Sensitivity*	*Specificity*
`mada`	1.137	1.962	0.434	1.540	-1.000	0.757 [0.686;0.816]	0.877 [0.717;0.952]
`metatron`	1.192	2.343	0.005	0.005	-0.001	0.767 -	0.912 -
`metamisc`	1.115	1.974	0.429	1.580	-0.752	0.753 [0.681; 0.813]	0.878 [0.712; 0.954]
`KussHoyer`	1.231	1.883	0.308	1.522	-0.857	0.774 [0.721; 0.820]	0.868 [0.708; 0.947]
`meta4diag`	1.178	2.187	0.407	1.770	-0.819	0.764 [0.699; 0.820]	0.897 [0.744; 0.968]
`HSROC`	1.180	2.050	0.566	1.422	-0.615	0.765 [0.643; 0.871]	0.886 [0.752; 0.982]
`Predictions`						0.825 [0.382; 1.000]	0.817 [0.388; 1.000]
`bamdit`	1.232	2.006	0.513	2.207	-0.731	0.772 [0.696; 0.841]	0.862 [0.631; 0.969]
`Predictions`						0.758 [0.510; 0.920]	0.766 [0.052; 0.999]

Table 1: Estimation of model parameters using different packages in R. Results of the mean parameters μ_{SE}, μ_{SP}, the scale parameters σ_{SE}, σ_{SP} and the correlation ρ are presented in the logistic scale. The models are parametrized in terms of sensitivity and specificity.

- *Predictions:* Only two packages, **HSROC** and **bamdit**, presented the posterior predictive intervals of sensitivity and specificity. The predictive interval for sensitivity reported by **bamdit** is close to the observed data, which have a range of 0.54 to 0.84. The predictive interval for specificity reported by **HSROC** excludes one observed specificity equal to 0.24, indicating that the model over-predict specificity in this example.

We can highlight, that in this example, the packages which implement a classical approach based on GLMM, or its approximation, have problems with the estimation of variances and correlations of random-effects. The exception is the package **metamisc**, which gives results similar to **meta4diag** and **bamdit**.

A casual practitioner may only look at the pooled sensitivity and specificity and find no difference between packages. However, the correlation structure gives important information about the heterogeneity of the studies included in the meta-analysis and the prediction of future studies. Therefore, packages with problems in the estimation of this part of the model will predict impossible results.

6. Conclusions

When developing **bamdit**, our aim was to simplify the application of a meta-analysis model which was accessible to practitioners but which up to now had required a large amount of statistical expertise. The package implements a series of innovative statistical techniques to avoid boundary estimation of parameters, conflict of evidence and robust estimation of model parameters.

The first example in Section 4 shows that the MCMC algorithm implemented in **bamdit** outperforms a classical bivariate random-effects approach based on REML estimation, which can be unreliable when the meta-analysis contains a small number of studies with a large heterogeneity (Riley *et al.* 2007). Moreover, the flexible random-effects distribution used in **bamdit** helps to better understand the studies' results by pointing out unusual results.

The conflict of evidence assessment is the deconstructionist side of meta-analysis, where each piece of evidence is put aside from the full model and compared to the rest of the evidence. One possibility for this type of analysis is to embed a meta-analysis model in a more general model where the non-conflict situation is a particular case. Both examples in Section 4 demonstrated that we could apply a double scale mixture of bivariate normal distributions and we made conflict diagnostics by direct interpretation of the scale weights.

One important topic currently not implemented in **bamdit** is the meta-regression and the indirect comparison of several diagnostic procedures. These topics are linked to the problematic of ecological bias and are topics of current research. However, we plan to update **bamdit** to include this functionality soon.

Actually, there is no other statistical software such as R that has implemented such a universe of possibilities to make meta-analyses of diagnostic tests. Unfortunately, R is not the most popular software in meta-analysis of diagnostic test data. Recently, we reviewed 68 published meta-analyses of diagnostic test data in medical journals between October 2015 and February 2016. We found that 29 papers chose Stata for fitting the bivariate meta-analysis of Reitsma *et al.* (2005), 9 used Meta-Disc alone, and 19 papers combined the use of Meta-Disc with Stata. Among the remaining 11 papers, 3 used SAS and 5 R. We found that practitioners publish statistical results in medical journals with convergence errors, variance equal to zero, integrated AUC in a range that is not empirically plausible, and so on. There is an imperative need to improve statistical results in this area and we hope that **bamdit** can help with this.

Acknowledgments

I am grateful to the following talented and motivated students from the Institute of Bioinformatics at the University of Düsseldorf: Arnold Sykosch, who helped me to design the first version of the **bamdit** package, and Marc Daxer, who implemented new plot functions. My acknowledgment to Martyn Plummer for developing JAGS. This work was supported by the German Research Foundation project DFG VE 986/1-1.

References

Arends L, Hamza T, Van Houwelingen J, Heijenbrik K, Hunink M, Stijnen T (2008). "Bivariate random effects meta-analysis of ROC curves." *Medical Decision Making*, **28**(5), 621–638.

Buerkner P, Doebler P (2014). "Testing for publication bias in diagnostic meta-analysis: a simulation study." *Statistics in Medicine*, **33**, 3061–3077.

Chu H, Guo H (2009). "Letter to the editor." *Biostatistics*, **10**(1), 201–203.

Debray T (2013). *metamisc: Diagnostic and prognostic meta analysis (metamisc)*. R package version 0.1.1, URL https://CRAN.R-project.org/package=metamisc.

Dewey M (2014). "CRAN Task View: Meta-Analysis." Version 2014-07-25, URL http://CRAN.R-project.org/view=MetaAnalysis.

Doebler P (2015). *mada: Meta-Analysis of Diagnostic Accuracy*. R package version 0.5.7, URL https://CRAN.R-project.org/package=mada.

Gatsonis C, Paliwal P (2006). "Meta-analysis of diagnostic and screening test accuracy evaluations: methodologic primer." *AJR: American Journal of Roentgenology*, **187**, 271–281.

Glas A, Lijmer J, Prins M, Bonsel G, Bossuyt P (2003). "The diagnostic odds ratio: a single indicator of test performance." *Journal of Clinical Epidemiology*, **56**(11), 1129–1135.

Guo J, Riebler A (2015). *meta4diag: Meta-Analysis for Diagnostic Test Studies*. R package version 1.0.20, URL https://CRAN.R-project.org/package=meta4diag.

Harbord R, Deeks J, Egger M, Whiting P, Sterne J (2007). "A unification of models for meta-analysis of diagnostic accuracy studies." *Biostatistics*, **1**, 1–21.

Harbord R, Whiting P (2010). "metandi: Meta-Analysis of Diagnostic Accuracy Using Hierarchical Logistic Regression." *Stata Journal*, **9**, 211–229.

Higgins J, Thompson S, Spiegelhalter D (2009). "A re-evaluation of random-effects meta-analysis." *J.R. Statist. Soc. A*, **172**, 127–159.

Hoyer A, Kuss O (2015). "Meta-analysis of diagnostic tests accounting for disease prevalence: a new model using trivariate copulas." *Statistics in Medicine*, **34**(11), 1912–1924. ISSN 1097-0258. doi:10.1002/sim.6463.

Huang H (2014). *Metatron: Meta-analysis for Classification Data and Correction to Imperfect Reference*. R package version 0.1-1, URL https://CRAN.R-project.org/package=Metatron.

Irwig L, Macaskill P, Glasziou P, Fahey M (1995). "Meta-analytic methods for diagnostic test accuracy." *Journal of Clinical Epidemiology*, **48**, 119–130.

Kuss O, Hoyer A, Solms A (2014). "Meta-analysis for diagnostic accuracy studies: a new statistical model using beta-binomial distributions and bivariate copulas." *Statistics in medicine*, **33**(1), 17–30.

Lijmer J, Bossuyt P, Heisterkamp S (2002). "Exploring sources of heterogeneity in systematic reviews of diagnostic tests." *Statistics in Medicine*, **21**, 1525–1537.

Lijmer J, Mol B, Heisterkamp S, Bonsel G, Prins M, van der Meule J, Bossuyt P (1999). "Empirical evidence of design-related bias in studies of diagnostic tests." *The Journal of the American Medical Association*, **282**, 1061–1066.

Lunn D, Spiegelhalter D, Thomas A, Best N (2009). "The BUGS project: Evolution, critique and future directions." *Statistics in Medicine*, **28 (25)**, 3049–3067.

Macaskill P (2004). "Empirical Bayes estimates generated in a hierarchical summary ROC analysis agreed closely with those of a full Bayesian analysis." *Journal of Clinical Epidemiology*, **57**(9), 925–932.

Menten J, Boelaert M, Lesaffre E (2013). "Bayesian meta-analysis of diagnostic tests allowing for imperfect reference standards." *Statistics in Medicine*, **32**, 5398–5413.

Moses L, Shapiro D, Littenberg B (1993). "Combining independent studies of a diagnostic test into a summary ROC curve: data-analytic approaches and some additional considerations." *Statistics in Medicine*, **12**, 1293–1316.

Novielli N, Cooper NJ, Sutton AJ, Abrams K (2010). "Bayesian model selection for meta-analysis of diagnostic test accuracy data: application to Ddimer for deep vein thrombosis." *Research Synthesis Methods*, **1**, 226–238.

O'Hagan A, Pericchi L (2012). "Bayesian heavy-tailed models and conflict resolution: A review." *Braz. J. Probab. Stat.*, **26**(4), 372–401. doi:10.1214/11-BJPS164.

Paul M, Riebler A, Bachmann L, Rue H, Held L (2010). "Bayesian bivariate meta-analysis of diagnostic test studies using integrated nested Laplace approximations." *Statistics in Medicine*, **29**(12), 1325–1339.

Plummer M (2003). "JAGS : A Program for Analysis of Bayesian Graphical Models Using Gibbs Sampling." *Proceedings of the 3rd international workshop on distributed statistical computing*, **124**, 125. URL http://www.ci.tuwien.ac.at/Conferences/DSC-2003/Drafts/Plummer.pdf.

Plummer M, Best N, Cowles K, Vines K (2006). "CODA: Convergence Diagnosis and Output Analysis for MCMC." *R News*, **6**(1), 7–11. URL http://CRAN.R-project.org/doc/Rnews/.

R Core Team (2013). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Reitsma J, Glas A, Rutjes A, Scholten R, Bossuyt P, Zwinderman A (2005). "Bivariate analysis of sensitivity and specificity produces informative summary measures in diagnostic reviews." *Journal of Clinical Epidemiology*, **58**, 982–990.

Riley R, Abrams K, Sutton Lambert P, Thompson J (2007). "Bivariate random-effects meta-analysis and the estimation of between-study correlation." *BMC Medical Research Methodology*, **7**, 3.

Riley RD, Lambert PC, Staessen JA, Wang J, Gueyffier F, Thijs L, Boutitie F (2008). "Meta-analysis of continuous outcomes combining individual patient data and aggregate data." *Statistics in Medicine*, **27**(11), 1870–1893. ISSN 1097-0258. doi:10.1002/sim.3165. URL http://dx.doi.org/10.1002/sim.3165.

Rutter C, Gatsonis C (1995). "Regression methods for meta-analysis of diagnostic test data." *Academic Radiology*, **2**, 48–56.

Rutter C, Gatsonis C (2001). "A hierarchical regression approach to meta-analysis of diagnostic test accuracy evaluations." *Statistics in Medicine*, **20**, 2865–2884.

Schiller I, Dendukuri N (2015). *HSROC: Joint meta-analysis of diagnostic test sensitivity and specificity with or without a gold standard reference test*. R package version 2.1.8.

Takwoingi Y, Deeks J (2010). "MetaDAS: a SAS macro for meta-analysis of diagnostic accuracy studies. User Guide Version 1.3. 2010 July." *Available from: http://srdta.cochrane.org/*.

Takwoingi Y, Riley RD, Deeks JJ (2015). "Meta-analysis of Diagnostic Accuracy Studies in Mental Health." *Evid Based Mental Health*, **18**(4), 103–109.

Verde PE (2008). "Meta-analysis of diagnostic test data: modern statistical approaches." *Deutsche Nationalbibliothek*.

Verde PE (2010a). "An introduction of Bayesian data analysis with R and BUGS: a simple worked example." *Estadistica*, **62**, 21–44.

Verde PE (2010b). "Meta-analysis of diagnostic test data: a bivariate Bayesian modeling approach." *Statistics in Medicine*, **29**(30), 3088–3102. doi:10.1002/sim.4055. URL http://doi.wiley.com/10.1002/sim.4055.

Verde PE (2013). *bamdit: Bayesian meta-analysis of diagnostic test data*. R package version 1.1, URL http://CRAN.R-project.org/package=bamdit.

Verde PE (2014). "A comment mentioning possible application in meta-analysis of Dirichlet t-distributions." *Bayesian Analysis*, **9**(3), 589–590.

Westwood M, Whiting P, Kleijnen J (2005). "How does study quality affect the results of a diagnostic meta-analysis?" *BMC Medical Research Methodology*, **5**, 1471–2288.

Zamora J, Abraira V, Muriel A, Khan K, Coomarasamy A (2006). "Meta-disc: a software for meta-analysis of test accuracy data." *BMC Medical Research Methodology*, **6**(31), 1–12.

Zapf A, Hoyer A, Kramer K, Kuss O (2015). "Nonparametric meta-analysis for diagnostic accuracy studies." *Statistics in Medicine*, **34**(29), 3831–3841. ISSN 1097-0258.

Zhou Y, Dendukuri N (2014). "Statistics for quantifying heterogeneity in univariate and bivariate meta-analyses of binary data: the case of meta-analyses of diagnostic accuracy." *Statistics in Medicine*, **33**, 2701–2717.

Affiliation:

Pablo Emilio Verde
Coordination Center for Clinical Trials
University of Duesseldorf
Moorenstr. 5
40225, Duesseldorf
Germany
E-mail: pabloemilio.verde@hhu.de

Journal of Statistical Software http://www.jstatsoft.org/
published by the American Statistical Association http://www.amstat.org/
Volume VV, Issue II *Submitted:* yyyy-mm-dd
MMMMMM YYYY *Accepted:* yyyy-mm-dd