UHASSELT
KNOWLEDGE IN ACTION

www.uhasselt.be
Universiteit Hasselt
Campus Hasselt:
Martelarenlaan 42 | 3500 Hasselt
Campus Diepenbeek:
Agoralaan Gebouw D | 3590 Diepenbeek

UHASSELT
KNOWLEDGE IN ACTION

Maastricht University

Faculty of Sciences
School for Information Technology
Master of Statistics

Master's thesis

Meta-Research on Statistical Methods of Combining Diagnostic Studies

Philippe Ferdinand Tadger Viloria

Thesis presented in fulfillment of the requirements for the degree of Master of Statistics, specialization Biostatistics

SUPERVISOR :

Prof. dr. Emmanuel LESAFFRE

SUPERVISOR :

Prof. Pablo VERDE

2019
2020

Faculty of Sciences

School for Information Technology

Master of Statistics

Master's thesis

Meta-Research on Statistical Methods of Combining Diagnostic Studies

Philippe Ferdinand Tadger Viloria

Thesis presented in fulfillment of the requirements for the degree of Master of Statistics, specialization Biostatistics

SUPERVISOR :

Prof. dr. Emmanuel LESAFFRE

SUPERVISOR :

Prof. Pablo VERDE

Contents

List of Tables

Abstract

Introduction: The current study has the following aims: systematically investigating the extent to which results of recently published meta-analyses of diagnostic test accuracy could be biased when the authors have applied classical hierarchical models that implicitly assumed normality. Another aim is to compare results with a ready to use Bayesian hierarchical model. Finally, the study aims to assess the impact of its internal validity when classical hierarchical models are used for meta-analyses and results are compared with an Bayesian hierarchical approach.

Methods: The risk for normality assumption was scored in each study trough a new proposed risk score tool. Difference between methods was explored checking convergence status, profile likelihoods, non-positive-definite random-effect covariance matrix and Bland-Altman analysis between estimations. A visual inspection and classification of the profile-likelihoods models was done. The risk score of each study is used for the prediction of the convergence status, profile likelihoods shape,and the non-positive-definite random-effect covariance matrix. In addition, we performed an exploratory analysis of disagreement between different models of meta-analysis. Results are presented using Bland-Altman and principal component analysis.

Results: We found a relationship between the binary risk score of normality, the profile-likelihood findings, and the disagreement between models' estimates.

Conclusion: In classical hierarchical models, studies with a high risk for normality assumption may have: non-convergence issues, profile likelihood irregularities, and non-positive-definite random-effect covariance matrix. In Bayesian hierarchical models, the same studies don't present any difficulties in the fitting or estimation process.

Keywords: Meta-analyses methodology, mixed effect, hierarchical model, Bayes, normality assumptions.

1 Chapter 1: Introduction on meta-analyses of diagnostic test accuracy

A meta-analysis (MA) of diagnostic test accuracy (DTA) studies consists of a synthesis of the quantitative evidence through pooled estimations of the primary studies. The primary studies compare an index test with a standard gold test, resulting in 2×2 contingency tables. Such a diagnostic test produces dichotomized results against true disease status, from which measures like sensitivity, specificity, positive or negative likelihood ratio, receiving operating curve (ROC) can be calculated.

False diagnosis has a significant impact on the subject and in the health system. A false negative error can be life-threatening, because the patients fail to obtain prompt treatment, but also a false positive test may result in physical, emotional, and/or financial burdens (Christensen et al. 2010). In particular, the "MA of DTA" is a critical study to help the decision-making process on whether to implement a test or not.

We may expect that sensitivity and specificity of diagnostic tests based on binarizing a continuous scale are inversely correlated. This correlation could be affected by the diagnostic test setup. Often sensitivity and specificity are analyzed separately but, given their correlation, it is more efficient and insightful to analyze these measures jointly.

Nowadays in MA of DTA studies it is prevalent/common to find features such as rare diseases in target conditions, high accuracy tests, few studies, small studies, tests with different thresholds, or a combination of all those elements. Such factors have been reported previously as factors that can compromise normality assumptions in MA of DTA.

This master-thesis has the following aims: 1) to systematically investigate the extent to which results of recently published meta-analyses could have been biased when the authors applied classical hierarchical models that implicitly assumed normality; 2) to compare results with a ready to use Bayesian hierarchical model; 3) to assess the impact of studies' internal validity when classical hierarchical models are used for meta-analyses and results are compared with an Bayesian hierarchical approach; 4) to propose a risk score for the normality assumptions in meta-analyses of diagnostic test accuracies; 5) to evaluate the predictive skills of the proposed risk score; 6) to assess the agreement between the average estimates in random effect (RE) models under normal vs non-normal assumptions.

The chapters are organized as follows: Chapter 2 reviews the statistical methods used in a MA of DTA. Chapter 3 provides a revision of the normality assumptions in a MA of DTA, and a new "risk tool" for normality assumption in MA of DTA is presented. Chapter 4 is a review of the statistical software for MA of DTA is given. Chapter 5 describes the meta-research experiment, and Chapter 6 shows the results.

2 Chapter 2: Statistical methods of meta-analyses of diagnostic test accuracy

2.1 Introduction

The most common outcomes in a MA of DTA are sensitivity and specificity, which are defined as true positive rate (TPR) and true negative rate (TNR), respectively. The concept can be grasped easily through a 2x2 contingency table (see Table 1), and the following definitions:

$$Sensitivy := \widehat{TPR_i} = \frac{tp_i}{n_{i,1}} = \frac{tp_i}{tp_i + fn_i}, \quad Specificity := \widehat{TNR_i} = \frac{tn_i}{n_{i,2}} = \frac{tn_i}{tn_i + fp_i}$$

Table 1: Contingency table of patient status (columns) vs test outcomes (rows)

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

Besides, sensitivity and specificity, the accuracy of a test can also be measured with diagnostic odd-ratio (DOR), positive or negative likelihood ratio, and the Youden index. The positive likelihood ratio is defined as: $LR+ = \frac{sensitivity}{1-specificity}$, or also as: $LR+ = \frac{P(T+|D+)}{P(T+|D-)}$. Similarly negative likelihood ratio is defined as:$LR- = \frac{1-sensitivity}{specificity}$ or also as:$LR- = \frac{P(T-|D+)}{P(T-|D-)}$. Finally, the most popular univariate index are: Youden index as: $J = sensitivity + specificity - 1$, and the DOR which is defined as $DOR = \frac{TP/FP}{FN/TN}$. All these measures are transformation from sensitivity and specificity, because of that, we will focus only on sensitivity and specificity as the main outcomes of our models.

Usually, the descriptive analysis in MA of DTA is done through a forest plot which presents the measure of effect (e.g. sensitivity) for each DTA studies incorporating confidence intervals represented by horizontal lines. In addition, a common practice is to present a receiver operating characteristic plot (ROC plot), which shows a scatter-plot of the pairs of sensitivity and (1- specificity). Usually, the ROC plot is done sub-grouping at various threshold settings or clinical subgroups categories.

Several models have been proposed to summarize the diagnostic measures (Harbord et al. 2007; Houwelingen, Zwinderman, and Stijnen 1993; Littenberg and Moses, n.d.; Reitsma et al. 2005; Sheu and Suzuki 2001; Verde 2010). All of these models reflect two specific features of this kind of data(Harbord et al. 2007): a negative correlation between sensitivity and specificity and substantial heterogeneity between-studies can be expected because DTA studies are differently designed. Despite the different inferential methods in the classical or Bayesian approaches, both models need to reflect the same data features: heterogeneity between and within studies but also a correlation between the outcomes.

In particular, meta-analyses on diagnosis studies bring specific challenges like high dependency between diagnostic summaries (intrinsically related by the diagnostic threshold used), also the usual sources of heterogeneity (study designs and population characteristics). This high dependency on diagnostic measures theoretically forces us to use multivariate models and reject univariate models due to this intrinsic correlation.

In general, meta-analyses can be conducted with either a fixed-effects model or a random-effect model. A usual approach is to decide according to heterogeneity indexes. However, this decision can be utterly misleading because it has been shown that the heterogeneity indexes are not adequate enough(Jackson 2006).

Since the beginning of the development of MA of DTA the dominant model was the Moses-Littenberg model (Littenberg and Moses, n.d.). Nowadays it's criticized because it assumes a simple linear regression to model the DTA outcomes through a fixed-effect model. Hierarchical or RE models have been proposed to overcome the limitations of the fixed-effects models. So even when it is the case that a fixed effect phenomenon is present the hierarchical model can be easily simplified to a fixed-effect model when it shows a between variability (τ^2) equal to zero (Jackson 2006).

The convergence of the classic binomial-normal or normal-normal RE for a MA of DTA is not always possible. Riley et al. (2007) recommends using a univariate model when convergence on the bivariate model is not met. Of course, we can have better options: Bayesian models bring an excellent opportunity to handle not only the issue of convergence but also to bring more flexibility into the modeling process of the random-effects, with non-Gaussian distributions (Verde 2018, 2010).

Because the main objective of this study is to evaluate normality assumptions, the models considered are only mixed-effects models (frequentist and also Bayesian) in the two general modalities: normal-normal, and more "exact" models (binomial-normal, or binomial-Mixture of normal). Although summary ROC is a recommended model when there is a threshold effect, this model is equivalent to fitting a bivariate model with an exchangeable covariance structure (Takwoingi et al. 2017), so it will not be taken into account in this thesis either.

2.2 Hierarchical random-effects model

In the case of general meta-analysis the mixed models have been proposed initially by Houwelingen, Zwinderman, and Stijnen (1993) , and in the specific context of MA of DTA by Reitsma et al. (2005). In the frequentist framework, the most popular options to model the accuracy proportions, sensitivity, and specificity, are to use a normal-normal or a binomial-normal (exact distribution) distribution. The use of normal approximation has been during decades the standard practices, and have been changed during the last decade. When the normal distribution is used to model sensitivity and specificity, a continuity correction needs to be made, which produces the first source of bias, because it forces a non-normal phenomenon to behave like a Gaussian experiment. Also, it needs to be used in a proper transformation: probit, logit, or arcsine; the standard practice is to use logit; in this work we used the logit link function. Mainly because the observed data can behave more frequently like a log-normal, than normal in most of the cases.

2.2.1 The bivariate normal-normal model

The bivariate normal-normal (Reitsma et al. 2005) model are defined as follows:

$$\begin{pmatrix} logit(\widehat{Se_i}) \\ logit(\widehat{Sp_i}) \end{pmatrix} \equiv \begin{pmatrix} logit(\widehat{TPR_i}) \\ logit(\widehat{TNR_i}) \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_{A_i} \\ \mu_{B_i} \end{pmatrix}, \Sigma_i \right) \quad i = 1, .., N \text{ studies,}$$

$$\text{with} \quad \begin{pmatrix} \mu_{A_i} \\ \mu_{B_i} \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_A \\ \mu_B \end{pmatrix}, \Psi \right), \quad \text{and} \quad \Psi = \begin{pmatrix} \sigma^2_{A_i} & \sigma_{AB_i} \\ \sigma_{AB_i} & \sigma^2_{B_i} \end{pmatrix}$$

Where $\mu = (\mu_A, \mu_B)^T$ and the covariance matrix Ψ are both estimated from the data by maximum likelihood estimation (MLE). The estimators μ and Ψ are the fixed effect outcomes, and the variance-covariance matrix respectively, the principal estimators of this model.

Also, in the Reitsma model the covariance matrix (Σ_i) is assumed known, calculated from the observed data with the delta method:

$$\Sigma_i = \begin{pmatrix} S^2_{i,1} & 0 \\ 0 & S^2_{i,2} \end{pmatrix}, \text{ where } \quad S^2_{i,1} = \frac{1}{n_{i,1}\widehat{TPR_i}(1-\widehat{TPR_i})} \quad \text{and} \quad S^2_{i,2} = \frac{1}{n_{i,2}(\widehat{TNR_i}(1-\widehat{TNR_i}))}$$

2.2.2 An univariate normal-normal model

A simplified version of the bivariate model will be considered only to produce a reduced version of the MLE in such a case. We can express the $logit(sensitivity_i)$ as:

$$logit(sen_i) = X_i\beta + Z_i\mu_{Ai} + \epsilon_i, \quad i = 1, ...N\mu_{Ai} \sim N(\mu_A, \Sigma), \epsilon_i \sim N(0, \sigma^2 I)$$

β is the p-dimensional vector of fixed effects, μ_{Ai} is the q-dimensional vector of random effects, X_i (of size $n_i p$) and Z_i (of size $n_i q$) are known fixed-effects and random-effects regression matrices,

To facilitate the expression for MLE, a more convenient form to express the variance-covariance matrix as a relative precision factor of Δ, which is the matrix that satisfies:

$$\frac{\Psi^{-1}}{1/\sigma^2} = \Delta^T\Delta$$

If Ψ is positive-definite, then the matrix Δ exist, but not in a unique way. This is the main reason why a non-positive definite random matrix can produce uncertainty about the uniqueness of the estimates obtained. The Cholesky factor of $\sigma^2\Psi^{-1}$ can only have one Δ related. A similar model can be used for $logit(specificity_i)$

2.2.3 The bivariate binomial-normal model

Since the bivariate linear mixed model for MA of DTA was proposed by Reitsma (Reitsma et al. 2005), Chu and Cole (2006) suggest instead (as a letter to the editor) to use a bivariate binomial-normal (BN) generalized mixed model avoiding the normal-normal approximation. The bivariate BN does not require the implementation of continuity correction when the number of events[1] is zero in a study (Chu and Cole 2006). The BN bivariate model (Chu and Cole 2006) can be expressed as:

$$TP_i \sim Bin(TPR_i, n_{i,1}) \quad \text{and} \quad TN_i \sim Bin(TNR_i, n_{i,2})$$

$$\mu_i \equiv \begin{pmatrix} logit(TPR_i) \\ logit(TNR_i) \end{pmatrix} \equiv \begin{pmatrix} \mu_{Ai} \\ \mu_{Bi} \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_A \\ \mu_B \end{pmatrix}, \Psi\right) \quad \text{with} \quad \Psi = \begin{pmatrix} \sigma^2_A & \sigma_{AB} \\ \sigma_{AB} & \sigma^2_B \end{pmatrix}$$

Both previous hierarchical models are random intercept models, where a bivariate normal distribution models the heterogeneity between studies. Different "outcomes" are used in each model, in the BN model a bivariate binomial distribution models the TP and TN of each study; in the NN model, the $logit(Sensitivity)$ and $logit(specificity)$ are modeled by a bivariate normal distribution. As usual, a fixed effects model can be seen as a specific case of the mixed models when the covariance matrix is assumed to be equal to zero in the covariance elements.

[1] true positives, true negatives, false positives, or false negatives

2.2.4 Estimation procedure for classical models

In the classical estimation of the parameters' model, we will focus on the maximum likelihood estimates (MLE) or the restricted maximum likelihood estimates (REML). MLE was used when REML is not available (like in `PROC NLMIXED`). Both methods are the most used methods for estimation in RE models through maximization mathematical methods. In the case of NN MLE, the marginalized distribution of the data is calculated as follows::

$$\begin{pmatrix} logit(\widehat{Se_i}) \\ logit(\widehat{Sp_i}) \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_{A_i} \\ \mu_{B_i} \end{pmatrix}, \Sigma_i + \Psi\right) \quad i = 1,..,N \text{ studies, } \Psi = \begin{pmatrix} \sigma^2_{A_i} & \sigma_{AB_i} \\ \sigma_{AB_i} & \sigma^2_{B_i} \end{pmatrix}$$

So the likelihood function ($\mathcal{L}$) can be written as (Pinheiro, Bates, and Pinheiro 1995):

$$\mathcal{L}(\mu_A, \beta, \Psi \mid y) = p(y \mid \mu_A, \beta, \Psi),$$

where p is the probability density, and y the complete n-dimensional outcome vector, where each element is $logit(sen_i)$; $\mathcal{L}$ can be maximized on μ_A, and Ψ.

$$\mathcal{L}(\beta, \theta, \sigma^2 \mid y) = \prod_{i=1}^{N} p(y \mid \beta, \theta, \sigma^2) = \prod_{i=1}^{N} \int p(y_i \mid \mu_{Ai}, \beta, \sigma^2) p(\mu_{Ai} \mid \theta, \sigma^2) d\mu_{Ai}$$

We use θ to represent an unconstrained set of parameters that determine Δ. It has been shown (Pinheiro, Bates, and Pinheiro 1995) that with additional arithmetic simplification the likelihood can be expressed as:

$$\mathcal{L}(\beta, \theta, \sigma^2 \mid y) = \frac{1}{(2\pi\sigma^2)^{N/2}} exp\Big(- \sum_{i=1}^{N} \mid\mid \tilde{y}_i - \tilde{X}_i \beta_i - \tilde{Z}_i \hat{\mu}_{Ai} \mid\mid \Big) \prod_{i=1}^{N} \frac{abs \mid \Delta \mid}{\sqrt{\mid \tilde{Z}_i^T \tilde{Z}_i \mid}} \qquad (1)$$

The previous expression can be used in an optimization routine to calculate MLE for the parameters of the model. Still, the optimization process can be simplified through profiling the likelihood. On the other hand, for the BN model, this method combines numerical integration to calculate y_i and optimization to estimate μ_A, μ_B, and Ψ.

MLE yields to an estimation of both the fixed effects and the variance components by maximizing the likelihood with respect to each element of μ and with respect to each of the variance component Ψ(Corbeil and Searle 1976). The REML is a transformation of the MLE which partitions the likelihood under normality into two terms, a likelihood that involves the mean parameter (μ_A, μ_B) and a residual likelihood that includes only the variance parameter Ψ. So the first likelihood can be maximized to estimate the mean parameter and the residual likelihood can be maximized to estimate the variance parameter. MLE and REML "both have the same merits of being based on the likelihood principle, which leads to useful properties such as consistency, asymptotic normality, and efficiency"(Verbeke and Molenberghs 2000). The REML has the advantage to produce unbiased estimates of variance and covariance parameters. The only disadvantage of the REML is that the solutions to variance components are not closed-form, so numerical computations are more challenging than MLE.

2.2.5 The Bayesian hierarchical model

The Bayesian hierarchical model for MA of DTA was originally presented in Verde (2008) and generalized to have non-normal random effects in Verde (2010). The Bayesian hierarchical model assume that the studies that we aim to combine are the results of N diagnostic studies shown in Table 1, similarly to classics model tp_i and fp_i outcomes can be modelled with two independent binomial distributions: $tp_i \sim Bin(TPR_i, n_{i,1})$ and $fp_i \sim Bin(FPR_i, n_{i,2})$, keeping the same previously defined notation

$$Sensitivy := \widehat{TPR_i}, \quad Specificity := \widehat{TNR_i}, \quad n_{i,1} = tp_i + fn_i \quad \text{and} \quad n_{i,2} := tn_i + fp_i.$$

Similarly to classical model, the Bayesian model can be expressed as follows:

$$\mu_i \equiv \begin{pmatrix} logit(TPR_i) \\ logit(TNR_i) \end{pmatrix} \equiv \begin{pmatrix} \mu_{Ai} \\ \mu_{Bi} \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_A \\ \mu_B \end{pmatrix}, \Psi \right) \quad \text{with} \quad \Psi = \begin{pmatrix} \sigma_A^2 & \sigma_{AB} \\ \sigma_{AB} & \sigma_B^2 \end{pmatrix}$$

The formulation of the previous model is completed by specifying the priors for the hyperparameters $\mu_A, \mu_B, \sigma_A, \sigma_B, \rho$. Assuming that parameters are independent and using the following set of priors:

$$\mu_A \sim Logistic(m_1, v_1), \quad \mu_B \sim Logistic(m_2, v_2) \quad \sigma_A \sim U(0, u_1), \quad \text{and} \quad \sigma_B \sim U(0, u_2)$$

The correlation parameter ρ is transformed by using the Fisher transformation as follows:

$$\rho = \frac{\sigma_{AB}}{\sqrt{\sigma_A \sigma_B}}, \quad z = logit(\frac{\rho + 1}{2}) \quad \text{where the prior for z is:} \quad z \sim N(m_r, v_r)$$

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 (Verde 2018).

In the priors, 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 chosen, following the recommendation made by Verde (2018):

$$m_1 = m_2 = m_r = 0, v_1 = v_2 = 1 \quad \text{and} \quad v_r = \sqrt{1.7}$$

taking $v_r = \sqrt{1.7}$ gives an uniform distribution for ρ between -0.9 and 0.9 which is clinically logical. Specifically this limitation for ρ avoid problems with estimation on the space frontier. The previous scenarios is for a weakly informative prior setup, but other approaches can be implemented when more information of previous studies (like previous MA of DTA are available).

Finally for bamdit (version 3.3.2) implementation (Verde 2018) we give the following prior to the degrees of freedom ν parameter: $U = 1/\nu$ and a uniform distribution for U: $U \sim Uniform(a, b)$ with $a = 1/df.upper$ and $b = 1/df.lower$. The default values in bandit df.lower = 3 and df.upper = 30 was used for the modelling process, which allowed us to explore random effect distributions that go from a t distribution with 3 degrees of freedom to a normal distribution(Verde 2018).

The binomial-normal mixed Bayesian model provides a similar structure like the frequentist binomial-normal (random intercept with two levels) but extends the model allowing us to incorporate the uncertainty in estimates of the model through priors and hyper-priors.

We need to realize that no asymptotic assumptions were made in the previous Bayesian analysis, so the results are valid and independent of the number of included studies. This is particularly useful in the case of meta-analyses of a few studies

2.2.6 Profile likelihood

MLE and REML extension may be one of the most used methods in the health sciences(Cole, Chu, and Greenland 2014). The profile likelihood (PL) function is a marginal probability, wherein the iterative optimization process uses the nuisance parameters as a fixed value, in order to be able to focus on the parameter of interest through a grid exploration in an interval. The PL exploration in random-effect models is an alternative method to analyze the evidence of heterogeneity in the studies included in the reviews. The same PL exploration has been previously used by Abrams, Myles, and Spiegelhalter (2004) and Curcio and Verde (2011). Each PL plot summarizes the support for each DTA from different values of between-studies standard deviation.

In general the PL function is given by

$$L_1(\tau; X) = \sup_{\theta \in \Theta : \theta_1 = r} p_\theta(X)$$

(Bijma, Jonker, and Vaart 2017) For a fixed value of θ_1 the PL $L_1(\theta_1; X)$ is equal to the maximum of the "usual" likelihood $p_\theta(X)$ over the remaining parameters $\theta_2, \theta_3, \ldots, \theta_n$. In particular we are interest in the PL of the covariance matrix parameters, which can be expressed as:

$$\begin{aligned} PL_{\sigma_A} &:= \sup_{\sigma_A \in \Theta : \sigma_{A,1} = r} p_{\sigma_A}(X) \\ PL_{\sigma_B} &:= \sup_{\sigma_B \in \Theta : \sigma_{B,1} = r} p_{\sigma_B}(X) \\ PL\sigma_{AB} &:= \sup_{\sigma_{AB} \in \Theta : \sigma_{AB,1} = r} p_{\sigma_{AB}}(X) \end{aligned}$$

The parameters $\beta(\theta)$ and $\sigma^2(\theta)$ can be determined from standard linear regression theory (Pinheiro, Bates, and Pinheiro 1995) and substituted in the equation 1, given us the PL for θ as:

$$L(\theta) = L(\hat{\beta}, \theta, \hat{\sigma}^2(\theta)) = \frac{exp(-N/2)}{[2\pi\hat{\sigma}^2(\theta)]^{N/2}} \prod_{i=1}^{N} \frac{abs \mid \Delta \mid}{\sqrt{\mid \tilde{Z}_i^T \tilde{Z}_i \mid}}$$

Which shows how the likelihood function can be used to calculate the PL of the models, and it has been applied in chapter 3 and 4 to evaluate the normality assumptions in the MA of DTA. The PL can provide two valuable information, firstly how the model itself marginally explores the probability of each parameter present in the model. This can give us an idea of how robust (or not) the process of inference in observed data according to the presented model. Also, another useful point is that the PL function provides a point of comparison between classical (PL function) and Bayesian models (posterior probability). Both densities provide "comparable" information and explore the probability function of each estimate marginally according to the model. So it will be interesting to see how these densities can be similar in some studies, or contrarily how these probabilities can be different in location or shape.

2.2.7 Heterogeneity test

Generally, the between-study heterogeneity of MA of DTA studies is larger than the therapeutic/interventional studies (Lee et al. 2015; Verde 2010). According to Zhou and Dendukuri (2014), the source of heterogeneity in MA of DTA is due to 1) the non-randomized design of most diagnostic studies and 2) the natural variation in sensitivity and specificity across positivity thresholds. The random effect model tries to capture as best as possible this potential heterogeneity. However, a more meaningful way to explain heterogeneity will be adding a meta-regression or subgroup analysis. A brief review of the concept of heterogeneity indexes can help us to understand the connection with the estimation of the covariance matrix.

2.2.7.1 Q statistics Also denominated Cochran's χ^2 statistic assess the differences between the observed effect sizes existing due to within-study error; which is equivalent to testing whether all studies share a common effect size, i.e. the heterogeneity is 0 (Spineli and Pandis 2020). A p-value is frequently quoted as an indication of the extent of between-study variability. Q is expressed as follows

$$Q = \sum_{i=1}^{k} w_i(Y_i - M)^2 = \sum_{i=1}^{k} \left(\frac{Y_i - M}{\nu_i}\right) = \sum_{i=1}^{k} w_i Y_i^2 - \left(\left(\sum_{i=1}^{k} w_i Y_i\right)^2 \Big/ \sum_{i=1}^{k} w_i\right)$$

where Y_i is the effect size in study i, w_i, is the weight of the study defined as the inverse of its variance $(1/\nu)$, and M is the summary effect size of the MA of DTA of k studies under the fixed-effect model. It can be used for continuous and binary outcomes (in MA of DTA by $logit(TPR_i)$ or TP_i).

Furthermore, the standard tests for the presence of heterogeneity have low power(Jackson 2006), such affirmation comes not only from simulations studies but also from the direct closed-analytic calculation of the power of such test. Choosing between fixed and random effects models is generally discouraged with such heterogeneity tests (Jackson, White, and Riley 2012).

In MA, the between-study variance quantifies the differences in the studies' results that cannot be explained by within-study variation alone(Jackson 2006). An alternative definition for heterogeneity can be "the dispersion of the true effect sizes across the studies included in a meta-analyses" (Spineli and Pandis 2020). Under the random-effect model, only two sources of variance exist: within-study and between-study variability. Such between-study variability is the variation in the true effect sizes in the included studies, also called heterogeneity (Spineli and Pandis 2020). When heterogeneity between-studies is assumed to be zero, the random effect is transformed into a fixed effect, where there is only within-variability' study.

2.2.7.2 The between variability τ^2 The between-study variability or τ^2 parameter can be estimated with different methods; one way can be using maximum likelihood or the method suggested by DerSimonian and Laird. The method of DerSimonian and Laird for heterogeneity calculation is a non-iterative method, done through the methods of moments. The DesSimonian and Laird method to estimate τ^2 is as follows: $T^2 = \frac{Q-df}{C}$, where $C = \sum_{i=1}^{k} w_i - \left(\sum_{i=1}^{k} w_i^2 \Big/ \sum_{i=1}^{k} w_i\right)$, where df is degree of freedom and C is a factor that incorporates the metric of the effect. The τ^2 index presents an absolute measure of the heterogeneity.

2.2.7.3 The Higgins I^2 index This index tries to tackle the heterogeneity issue, presenting a relative index:

$$I^2 = \left(\frac{Q - df}{Q}\right) = \frac{\hat{\tau}^2}{\hat{\tau}^2 + \hat{\sigma}^2}$$

Where $\hat{\tau}^2$ are the between-variability estimation and $\hat{\sigma}^2$ is the within-variability estimation. A rule of thumb for the cut-off point for I^2 is the following when it can be classified as following: low (0-50%), moderate (50-75%) or high (>75%). Some work to support of the impact of this measure has been done in simulation studies (Diaz 2015; Kontopantelis and Reeves 2012), where changes in the I^2 true values produce different results in the coverage probability and/or bias on the estimation.

2.2.7.4 A bivariate index of heterogeneity Most of the literature about heterogeneity in MA is related to univariate outcomes, but have been extended to the general context of multivariate MA (Jackson, White, and Riley 2012), and also in the context of MA of DTA by Reitsma et al. (2005). The previous index was presented because they are frequently used in the reviewed studies, and for historical reasons. However, in the context of MA of DTA, a more proper index can be a bivariate I^2 index proposed by Zhou and Dendukuri (2014):

$$I^2_{E(Biv)} = \frac{\mid \hat{T} \mid^{1/2}}{\mid \widehat{E(\Sigma)} \mid^{1/2} + \mid \hat{T} \mid^{1/2}} \quad \text{where} \quad \Sigma = \begin{pmatrix} \sigma^2_{iA} & 0 \\ 0 & \sigma^2_{iB} \end{pmatrix},$$

is the within-study variance matrix; $\mid E(\Sigma) \mid$ is the determinant of the expected within-study variance and $\mid T \mid$ is the determinant of the between-study variance-covariance matrix(Zhou and Dendukuri 2014). In the same study (Zhou and Dendukuri 2014), simulations were done comparing the coverage of the univariate and bivariate indexes. The conclusion was that in most of the cases the univariate version underestimated the heterogeneity (from the true values of the simulation).

All the presented measures are dependent on the estimates in the variance-covariance matrix. So, any bias in the covariance matrix will affect directly the estimated I^2 index and can affect the decision process in MA of DTA authors between the selection of models. Using a univariate Higgins I^2 alone to choose between fixed and random model is never recommended. The use of bivariate I^2 can be the starting point of the explorations of the possible sources of such heterogeneity.

Notably, we will explore the heterogeneity of the studies in the context and impact for the risk of normality assumptions in MA, using different tools like `risk inference tools` and PL or Posterior distributions of the covariance matrix elements.

3 Chapter 3: Evaluation of normality assumptions in meta-analyses of diagnostic test accuracy

3.1 Introduction

Meta-analytic studies have mainly focused on examining when the fixed-effect assumption or random effect can be implemented. However, they rarely check the normality assumption for the normal approximation of the likelihood of the parameter of interest and the normal distribution of the random effects.

In this chapter, a revision on normality assumptions in MA will be conducted, but mostly in MA of DTA. We are going to extend the risk tool proposed by Jackson and White (2018) for univariate MA, with a new risk tool to been able to evaluate MA of DTA. Such a tool will help to predict if the statistical inference in MA of DTA has been compromised by the normality assumptions.

A closer method to our actual work are studies based on an empirical assessment of previously published MA of DTA, to posteriorly check the agreement or correlation between estimations. These kinds of studies are more related, with more "real case" scenarios, and can be used as a source of information for simulation studies. We found four studies that follow this methodology (Dahabreh, Trikalinos, and Lau, n.d.; Harbord et al. 2008; Karahalios et al. 2017; Spineli 2019). The specific approach of comparisons of the methods estimations were made through scatter plots and histograms (Dahabreh, Trikalinos, and Lau, n.d.) of 308 MA of DTA (coming from 157 systematic reviews studies); simple descriptive comparison between estimations of 8 MA of DTA studies (Harbord et al. 2008); Bland-Altman comparisons(Karahalios et al. 2017) between the estimations related to network MA (cumulative ranking curves) in 456 network MA methods; Bland-Altman method (Spineli 2019) on 31 network-MA. None of this study uses an assessment of the risk for normality assumptions, only comparisons through scatter plots or Bland-Altman analysis.

Another approach to compare models in MA of DTA is through simulation, which provides a larger sample size, and the true value of the parameters that the study is trying to estimate. Simulations studies can measure the bias between the true value and model estimates. We focus our attention on simulations done in the MA of DTA context like Takwoingi et al. (2017), Riley et al. (2007), and Hamza, van Houwelingen, and Stijnen (2008). The performance of the NN and BN (univariate or bivariate models) was checked in a simulation (Takwoingi et al. 2017) observing that convergence is affected by a few studies or sparse data. This study concludes that univariate methods were recommended for sensitivity and specificity if the bivariate model failed to converge; they only used classical methods, not Bayesian.

Riley et al. (2007) in their simulation concludes that the normal-normal maximum likelihood estimator is sensibly (non-convergence and unstable pooled estimates) when elements of the between-study covariance matrix are truncated on the boundary of its parameter space, which is a common situation when few studies or sparse data are present. The authors suggest that when non-convergence issues appear "then the best option may be a generalized univariate meta-analysis for sensitivity and specificity separately" (Riley et al. 2007). Also, they didn't use Bayesian methods. Finally, the standard NN model was compared to the BN model in a simulation study (Hamza, van Houwelingen, and Stijnen 2008), concluding that the exact likelihood performs better (coverage probability and PL) than the approximate approach and gives unbiased estimates. Also, the authors conclude that the standard NN method provides "huge bias with very poor coverage probability in many cases" (like smaller within-study sample sizes, larger between-studies variance) and larger values of the overall sensitivity.

A theoretical revision made by Jackson and White (2018) about the hidden normality assumptions in meta-analyses is a clear invitation to think about the influence of the normality assumptions in MA of

Table 2: Original risk of compromised statistical inference tool

Assumption	Most serious implication for μ if false	Especially dubious when
Y_i unbiased for μ_i	Biased pooled estimate	Sparse non-continuous data
Variances s_i^2 known	Inaccurate variance for $\hat{\mu}$	Small studies, sparse or skew data
$Y_i \mid \mu_i$ normal	Inaccurate likelihood-based inference	Small studies, sparse or skew data
μ_i normal	Inaccurate likelihood-based inference	Outlying studies are present
$\hat{\mu}$ unbiased for μ	Biased pooled estimate	Y_i biased for μ_i
Variance of $\hat{\mu}$ known	Inaccurate confidence interval	Few studies present; imprecise s_i^2
$\hat{\mu}$ normal	Inaccurate confidence interval	Few studies present
μ_{new} normal	Inaccurate prediction interval	Outlying studies are present

DTA. The Jackson and White (2018)' article is a cornerstone in our study because also produced the first "risk tool" (Table 2) to evaluate if the statistical inference is compromised by normality assumptions in univariate interventional MA. The article has been quoted 31 times, where 13 of these articles are responses to the article. However, as far as we know, the "risk of Inference tool" proposed by Jackson and White (2018) has not been used to evaluate the risk of the normality assumptions in any previous MA or MA of DTA study . Moreover, this risk of evaluation tool has not been extended previously to MA of DTA.

3.2 A new "risk of compromised statistical inference tool" for meta-analyses of diagnostic test accuracy

In this chapter, we present a tool to assess the risk of statistical inference in MA of DTA. Our tool can be seen as an extension for bivariate models based on the univariate tool proposed by Jackson and White (2018).

The proposed "risk tool" evaluates when the statistical inference for a specific MA of DTA has been compromised. The risk tool is presented in Table 3 and it has 17 items. The first eight items correspond mostly with features related to the statistical inference of logit(sensitivity). The items from 9-16 correspond symmetrically to the logit(specificity). The last item are related to the covariance between both outcomes. Each item can have three levels of risk: low, moderate, and high; each level of risk can be tabulated with the colours green, yellow, and red, respectively. Additional to the categorical score, a numeric and binary version for the total score is proposed.

The MA of DTA risk tool shown in table 3 presents the assumptions usually made by conventional methods for MA for sensitivity (TPR, items from 1 to 8), specificity (FPR, items from 9 to 16) and their covariance (item 17). In Table 3, the items from 1-3 (9-10) are related to within-study assumptions, and items from 4-8 (11-16) are associated with the between-study assumptions; also the item 17 (covariance) is related with between-study assumptions.

The risk-tool will be explained and the assumptions evaluated by each item for the sensitivity part of the risk tool: from 1 to 8, the items related to specificity will be in parenthesis. Items 1-3 (9-10) are related when a model assumes that the variability of the within-study is normal. The widespread use of within-study approximations is perhaps one of the biggest concerns of the current standard practice in MA (Jackson and White 2018). The items 4 (11) are related to the between-study distributional assumption or also called the random-effects distributions. The items from 5 to 8 (13 to 16) are associated with the use of normal distribution when making inferences (Jackson and White 2018). Finally, item 17 is related to the covariance (between logit(sen) and logit (spec)).

1. The items 1 (and 8) evaluates when the assumption that $logit(\widehat{TPR_i})$ is an unbiased estimator

for $logit(TPR_i)$ is risky. This situations occurs when *"sparse non-continuous data"* or presence of zero in the sensitivity (/specificity) contingency table. The continuity correction that needs to be implemented here is introducing a bias between $logit(\widehat{TPR_i})$ and $logit(TPR_i)$.

2. The item 2 (and 9) evaluates if the assumption that the variances $s^2_{logit(\widehat{TPR_i})}$ are known, is a realistic assumption (the use of the s_i^2 as if they are the σ_i^2). This it is dubious when *"small studies, sparse or skew data"*, which is associated with cells equal to zero in table 2x2 for the sub-population of non-healthy (/healthy), but also when studies are small, and finally when the prevalence between the highest prevalent and least prevalent ratio is large. Assuming the simplest possible situation that the Y_i are sample means of normally distributed observations then a t-distribution (instead of a normal distribution) is required to make inferences for μ_i in situations where the population variance is unknown (Jackson and White 2018). The items 1 and 2 are the assumptions for the first two moments.

3. The item 3 (and 11) is dubious when *"small studies, sparse or skew data"* are present. Item 3 evaluated the assumption that $logit(\widehat{TPR_i}) \mid logit(TPR_i)$ is normal; this is the assumption where the shape of the normal distribution is assumed, not just the first two moments (Jackson and White 2018).

4. The item 4 (and 12) checks the assumption that $logit(TPR_i)$ is normal, it is dubious when *outlying studies* are present under a model where the random-effects follow a normal distribution; it is dubious when a high frequency of influential DTA in each MA is present.

5. The item 5 (and 13) evaluates the assumption that the pooled estimate is unbiased, where it's a potential cause for concern for fixed and mixed models ($logit(\widehat{TPR})$ is an unbiased estimator for $logit(TPR)$, in the sensitivity case); this assumption will raise concerns in the presence of small studies and also when there exists a correlation between the outcome and their variance(Jackson and White 2018).

6. The item 6 (and 14) checks that assumption that the variance of the pooled estimate is known ($logit(\widehat{TPR})$ is known for the sensitivity case), it is especially dubious when *few studies are present.* "In the second stage of analysis when pooling the Y_i, we approximate σ_i^2 with s_i^2 and, in random-effects meta-analyses, τ^2 with $\hat{\tau}^2$"(Jackson and White 2018), but these standard errors are not truly known, and the accuracy of the statistical approximations will now "require reasonably large studies so that the s_i^2 are precisely estimated in both common-effect and random-effects meta-analyses" (Jackson and White 2018).

7. The item 7 (and 15) checks the assumption that $logit(\widehat{TPR})$ is normal, and it's especially dubious when *few studies are present.* This item checks when the shape of the normal distribution is assumed for the pooled estimate, not only for the first two moments (previous two items).

8. The item 8 (and 16) checks the assumption that $logit(TPR_{new})$ is normal, also it's especially dubious when *outlying studies are present.* This item checks for a specific type of statistical inference: the prediction interval for the true effect in a new study.

9. Item 17 checks the assumption that the covariance ($s_{TPR-FPR}$) between the two outcomes (logit(sen) and logit(spec)) is known. This is especially dubious when small studies, sparse data, and different prevalence between the studies are observed.

An additional observation is that the presented tool measure risk of inference is very useful for statisticians but not of much utility to a clinician, that are usually guided by threshold values (Mavridis 2018). This was an additional reason to include a detailed descriptive step in the Methodology design, not only to support the decision of risk score but also to provide intuitive threshold values to score each item.

Table 3: Adapted 'risk of compromised statistical inference tool' for MA of DTA. Where μ is logit(TPR) or logit(FPR)

Item	Assumption	if assumption is false > inference on μ is compromised	Especially dubious when
1	$logit(\widehat{TPR_i})$ unbiased for $logit(TPR_i)$	Biased pooled estimate	Sparse non-continuous data,
2	Variances $s^2_{logit(\widehat{TPR_i})}$ known	Inaccurate variance for $logit(\widehat{TPR})$	Small studies, sparse or skew data
3	$logit(\widehat{TPR_i}) \mid logit(TPR_i)$ normal	Inaccurate likelihood-based inference	Small studies, sparse or skew data
4	$logit(TPR_i)$ normal	Inaccurate likelihood-based inference	Outlying studies are present,
5	$logit(\widehat{TPR})$ unbiased for $logit(TPR)$	Biased pooled estimate	$logit(\widehat{TPR_i})$ biased for $logit(TPR_i)$,
6	Variance of $logit(\widehat{TPR})$ known	Inaccurate confidence interval	Few studies present; imprecise $s^2_{\widehat{TPR_i}}$,
7	$logit(\widehat{TPR})$ normal	Inaccurate confidence interval	Few studies present,
8	$logit(TPR_{new})$ normal	Inaccurate confidence interval	Outlying studies are present,
9	$logit(\widehat{FPR_i})$ unbiased for $logit(FPR_i)$	Biased pooled estimate	Sparse non-continuous data,
10	Variances $s^2_{logit(\widehat{FPR_i})}$ known	Inaccurate variance for $logit(\widehat{FPR})$	Small studies, sparse or skew data
11	$logit(\widehat{FPR_i}) \mid logit(FPR_i)$ normal	Inaccurate likelihood-based inference	Small studies, sparse or skew data
12	$logit(FPR_i)$ normal	Inaccurate likelihood-based inference	Outlying studies are present,
13	$logit(\widehat{FPR})$ unbiased for $logit(FPR)$	Biased pooled estimate	$logit(\widehat{FPR_i})$ biased for $logit(FPR_i)$,
14	Variance of $logit(\widehat{FPR})$ known	Inaccurate confidence interval	Few studies present; imprecise $s^2_{\widehat{FPR_i}}$,
15	$logit(\widehat{FPR})$ normal	Inaccurate confidence interval	Few studies present,
16	$logit(FPR_{new})$ normal	Inaccurate prediction interval	Outlying studies are present,
17	Covariance $s_{TPR-FPR}$ known	Inaccurate covariance for $logit(\widehat{TPR})$ and $logit(\widehat{FPR})$	Small studies, sparse data, different prevalence

4 Chapter 4: Statistical software for meta-analyses of diagnostic test accuracy

4.1 Introduction

After reviewing the concepts behind the MA of DTA, we can think that the journey stops here but it is not finished. The next important step is to find a software that can successfully estimate the parameters of our models with the related assumptions. Most of the time, this selection is not freely done, because it mainly depends on the skills of the operators and the budget. Usually, when an author has a deep understanding of the models and assumptions, the budget is not a limitation, and the coding skills can be achieved because the motivation is clear. If the researcher understands the possible models that can be fitted, choosing software is not a challenging quest. As follows, we reflect on the usual software option and potential implications in the model selection and assumptions.

For general users without a statistical background it can be confusing why some software is not useful to provide estimation in MA of DTA, and even dangerous to do descriptive explorations. In general, it is crucial to realize the model provided by each software/package in MA of DTA. A common feature in the MA of DTA studies reports, observed in this thesis, it was a complete lack of description of the statistical model. For example "bi-variate random effect" model is the most commonly used term in the methodological chapter of the selected studies, but never goes deeper if the bivariate model is finally a NN or BN model. In the explored studies, the original studies estimates don't have a high agreement (see Figure 29) with any of the proposed models and software (NN/BN or BBN).

4.2 Statistical software review

In general, all the usual software (`Revman`, `Meta-Disc`, or more complex software like `R`, `Stata`, and `SAS`) can provide simple pooling (univariate) fixed. But if we want to produce a summary estimate with a bivariate random effect model, only software like `R`, `Stata` or `SAS` can be used.

4.2.1 MetaDisc

`Metadisc 1.4` is standalone and free software created in 2004 (Zamora et al. 2006), broadly used [2], but noways their developers [3] discourage using it for inference purposes and only recommend using it as a descriptive approach. MetaDisc can conduct non-iterative estimation methods (DerSimonian and Laird method) and iterative methods (MLE and REML) but only *in a univariate fashion.* The use of Meta-Disc is prevalent in MA of DTA, specifically for plotting forest plots for the sensitivity and specificity. The plots and estimations provided by this software include *univariate pooled estimates.* Still, most of the MA of DTA studies that we revised for the actual meta-research were not aware of the univariate nature of the estimation (used in the article and not reported as such).

4.2.2 Revman

`Revman` is a standalone and free software provided by Cochrane, which cannot be used to perform hierarchical random-effects models; only exploratory analyses can be undertaken. Revman only conducts no iterative estimation methods, specifically only the DerSimonian and Laird method is available (Veroniki

[2] It can be challenging to quantify how many times this software has been used. Still, the article that explains the methodology behind the software (Zamora et al., 2006), it has been cited in 1599 according to Google Scholar until the current date, July 29, 2020

[3] ftp://ftp.hrc.es/pub/programas/metadisc/Metadisc_update.htm

et al. 2016). Cochrane recommends “the definitive analyses need to be undertaken in commercial software packages and sophisticated statistical programming environments such as SAS, Stata, S-Plus, R, MLwiN or WinBUGS/OpenBUGS” (Leeflang et al. 2013). According to Hoaglin (2018), Revman, and Meta-Disc are prevalent and problematic software, because both continue to offer to the clinician the DerSimonian-Laird method as the default method for random-effects MA, “without warning users of its shortcomings or of calls to abandon it”.

4.2.3 R packages

Rstudio as all the R’ packages are free and available to use, but require that the user know how to code in R. The number of packages in R for MA and MA of DTA can be overwhelming. So we’ll do a brief description of the packages for MA of DTA. A CRAN Task View has been made for meta-analyses (Dewey 2020) in general. Specifically, a subchapter for MA of DTA can be found which mentions the following software: mada, Metatron, metamisc, bamdit, meta4diag, CopulaREMADA, diagmeta, CopulaDTA, and NMADiagT. Particularly mada and Metatron allow to fit Reitsma models. bamdit, NMADiagT and meta4diag aloud to do MA of DTA Baysian models. CopulaREMADA and CopulaDTA offer extensions through copula based mixed model distribution in MA of DTA. metamisc offer to fit Riley MA of DTA models; and diagmeta aloud to do MA of DTA with multiple cutoff points. Other packages can be used to fit MA of DTA, like glmer, but require more knowledge of the fitted model by the user.

4.2.4 bamdit

bamdit (version 3.3.2) provides “functions for Bayesian meta-analyses of diagnostic test data which are based on a scale mixtures bivariate random-effects model” (using JAGS to implement the MCMC method)(Verde 2010). Also, graphical methods are provided. This package was developed to simplify “the use of meta-analyses models that up to now have demanded great statistical expertise in Bayesian meta-analyses”(Verde 2010). Specifically, in bamdit we use the metadiag function, which performs a Bayesian meta-analysis of diagnostic test data by fitting a bivariate random-effects model (Verde 2010). “The number of true positives and false positives is modeled with two conditional binomial distributions and the random-effects are based on a bivariate scale mixture of normal”(Verde 2018). Computations are done by calling JAGS to perform Markov Chain Monte Carlo sampling and returning an object of the class mcmc.list.

4.3 SAS details for fitting a meta-analysis of diagnostic test accuracy model

SAS is one of the most flexible options to fit mixed models. As follow, we’ll describe the details to fit the MA of DTA model that will use our meta-research experiment.

One of the most classical Reitsma model code was done with PROC MIXED included in the original Reitsma et al. (2005) model article. The steps in the code consist of data set transformation to produce Continuity correction in each cell. In this model, the degree of freedom is considered arbitrarily high (df=1000) to be able to produce a normal distribution (SAS provide a t-distribution by default). To facilitate the interpretation of the estimate no-intercept statement model is used (noint). The repeated statement is used because each study have outcomes logit(sen) and logit(spe) (each outcome has been previously duplicated in two different lines), this is the multivariate aspect of the model. A random statement is being used to produce the modeling of the response variable logit (logit(sen) or logit (spe)) through a random intercept model. A dummy (dis and non_dis) variable represents the modeling of the disease (sensitivity) or non-disease (specificity).

In all the three classic procedures: `PROC MIXED`, `PROC GLIMMIX`, and `PROC NLMIXED`, the selected type of covariance structure is unstructured (no restrictions in shape). In `PROC MIXED` to guarantees a positive definite covariance matrix an unstructured covariance structure is used, specifically, a Cholesky root (Stroup et al. 2018)(a type of unstructured structure), that constrain all diagonal values to be positive (`Type=CHOL` statement). If the model does not converge using `CHOL`, the option `UN` (Menke 2010) option can be used; but both options are unstructured covariance.

To be able to make a fair comparison between procedures, it will be essential to select the most similar method of optimization, whenever it is possible. For example, in `PROC MIXED` and `PROC GLIMMIX` we can perform both MLE and REML estimation, but `PROC NLMIXED` only allows MLE. In `PROC NLMIXED` there is no direct analogue to the REML method (Wolfinger, n.d.)[4]. For `PROC NLMIXED`, one of the most recommended integration approximation methods is the adaptive Gaussian quadrature, which allows us to control the number of quadratures. In `PROC MIXED` the selected method will be a Restricted Maximum Likelihood or `REML` using a ridge-stabilized Newton Raphson algorithm, with the default number of iterations (50). In `PROC MIXED` we included the residual variance as part of the Newton-Raphson iterations, which occurs when the HOLD= or EQCONS option is used in the PARMS statement, to use the within-variance calculated from data as known. In MA of DTA, the Gaussian modeling requires the assumption that "the within variance is known" (which is done through HOLD or EQCONS statement); this assumption is not required in the non-linear cases, because it is directly estimated. Also `PROC GLIMMIX` can be used as an iterative method, maximum likelihood estimation by adaptive Gaussian quadrature using the statement METHOD=QUAD, sharing the quadrature method with `PROC NLMIXED`.

We follow recommendations by Diaz (2015) simulation-study about using `PROC NLMIXED` with an optimization algorithm that uses double derivatives like TRUREG, NEWRAP or NRRIDG. Because this method provide a more accurate estimation than the default option (`QUANEW` which performs a quasi-Newton optimization) in `PROC NLMIXED` . Additionally, we will set 20 points of quadrature, which can provide a good value of precision in Random effects models (Lesaffre and Spiessens 2001).

The estimations of sensitivity, specificity, and uncertainty measures were done in each Bayesian model (binomial-normal and binomial-mixture of normal) using the bivariate mixed-effects regression model, fitted with `bamdit::metadiag`. The common model features will be done with 10.000 iterations, 1000 adaptions, 1000 Burn-in, no thin (1), (original default options of `bamdit`), but the priors for the variance of the logit will be selected from 0 to 5 (`sigma.D.upper = 5`, `sigma.S.upper = 5`); The possible values selected for the priors of the variance give equal probabilities within a range between 0 and 5, which includes the case of "non-heterogeneity between studies' results" to "impossible to combine results" (Spiegelhalter, Abrams, and Myles, n.d.; Verde 2019). We used three chains to check for convergence. Convergence will be reviewed by inspecting trace-plots for covariance matrix element and pooled estimations. `bamdit` computations are done by calling `JAGS` (Just Another Gibbs Sampler) to perform MCMC (Markov Chain Monte Carlo).

In both approaches, we will be performing a random-effects meta-analysis, even when the authors originally performed a fixed-effects analysis. The sub-grouping analysis will be done in the same way as the original 21 MA of DTA studies. A subgroup-analysis will be avoided if there is a general estimate. Subgroup estimations will be conducted if in the original MA of DTA a global estimation was avoided, to be able to compare when it is possible but also to respect the decision of the authors. Some minor exception was found when no estimation was provided, like in all subgroup in MA of DTA 1, and partially in some subgroups of MA of DTA 12, where some estimations were not presented (only four out nine subgroups was omitted). Still, we decided to include these subgroups just in case no original estimations were provided to compare with our estimations, to fulfill the rest of the model comparisons.

[4] The reason for this is related with common sense, 'PROC NLMIXED' is a procedure for non-linear models, and if a REML method could hypothetically be used in 'PROC NLMIXED' would involve a high dimensional integral over all of the fixed-effects parameters, and this integral is typically not available in the closed-form [@wolfingerFittingNonlinearMixed]

4.3.1 Profile likelihood

The profile-likelihood can be calculated in a close solution, without the need to obtain it from the likelihood grid evaluation of the software, but we consider it as not the best option. To be able to evaluate the model fitted in each software, it is necessary to assess the PL function in such a context. Specifically, exploring the grid of probabilities of the specific PL function in each software, instead of the theoretical close PL solution.

The approach to producing the PL in SAS was mainly made using code provided by Millar (2011) and SAS manual procedures for `MIXED`, `PROC NLMIXED` and `PROC GLIMMIX` (SAS 2015).

So, it is important to stress that for MA of DTA the use of more robust software like `SAS Stat`, `Stata`, `WinBugs/Jags`, or `R`, is mandatory clearly knowing the assumptions behind the model fitting. Meta-Disc and Revman are the most used softwares, these only provide "descriptive analysis", fixed effect estimation, or univariate analysis, with no possibilities to run hierarchical models.

Also, SAS functions provide support to evaluate the likelihood function in the fitted models in `PROC MIXED`, `PROC GLIMMIX`, and `PROC NLMIXED`, which allows us to produce a PL. The only limitation to provide PL in `PROC MIXED` and `PROC GLIMMIX` is that it doesn't allow the PL of the central outcomes as logit(sen) or logit(spec) fixed effects, but can be done with `PROC NLMIXED`.

5 Chapter 5: Design of the meta-research experiment

5.1 Introduction

In this chapter, we present the design and data collection of a meta-research experiment to evaluate the risk of statistical inference in MA of DTA through RIT score.

5.2 Data selection

A meta-research of recent meta-analyses of medical diagnostic test accuracy was performed. For each meta-analysis included in the review, the diagnostic data results, study population's characteristics, study quality evaluation, details on the statistical procedures, and statistical software that have been applied were extracted and summarized.

A Medline search was done with the phrase: ("Sensitivity and Specificity"[Mesh]) AND "Diagnosis"[Mesh] AND "accuracy"[Title/Abstract], to search the last 21 DTA MA, and to analyze if in each study the sensitivity and specificity estimation can be compromised (or not) according to (Jackson and White 2018). Inclusion criteria on the MA: presents contingency tables for each study (Author, year, TP, FN, TN, FP) which need to use the same reference standard test. The exclusion criteria are; multiple disease categories as target conditions, systematic review without meta-analyses, articles that only describe the protocol for a future MA of DTA, MA of DTA of individual patients, MA of DTA with language different from English, Spanish or French. We manually extracted the data for this thesis from the MA of DTA study that met the inclusion criteria for further analysis.

5.3 General analysis steps

The selected studies were evaluated according to qualitative criteria and quantitative exploration to see how normality assumptions can compromise in general the statistical analysis. A qualitative appraise was applied in 21 MA of DTA (and their 55 subgroups) according to the proposed RIT score (see Table 3). These qualitative explorations allowed us to present possible reasons or scenarios that explain why the results of published meta-analyses may be compromised under normality assumptions. The selected MA of DTA studies were fitted through different models, evaluating convergence and profile likelihood characteristics. Concordance analysis of the estimate and uncertainty measure was done according to Bland and Altman (1986). A final analysis was conducted for connections between the' findings. The methodological steps are summarized in Figure 1.

5.4 Detailed analysis plan

5.4.1 Qualitative analysis

The qualitative evaluation of the 55 studies has consisted of a descriptive exploration followed by the application of our "risk of compromised statistical inference tool" (RIT) in each study. Each item was scored as high (3 or red), medium (2 or yellow), or low (1 or green) risk, given a total risk of 17*3=51 maximum score. To recognize the supposed impact of normality assumptions from a qualitative perspective, the RIT score was applied based on descriptive measures' studies, to guide the decision in each risk for RIT (previous Table 3 and Figure 4); these rule of thumb decisions are presented as follows:

1. Item 1 and 9: The assumption of these items does not hold when we are dealing with "sparse non-continuous data". A MA of DTA was classified as "high risk" if more than 40% of the studies

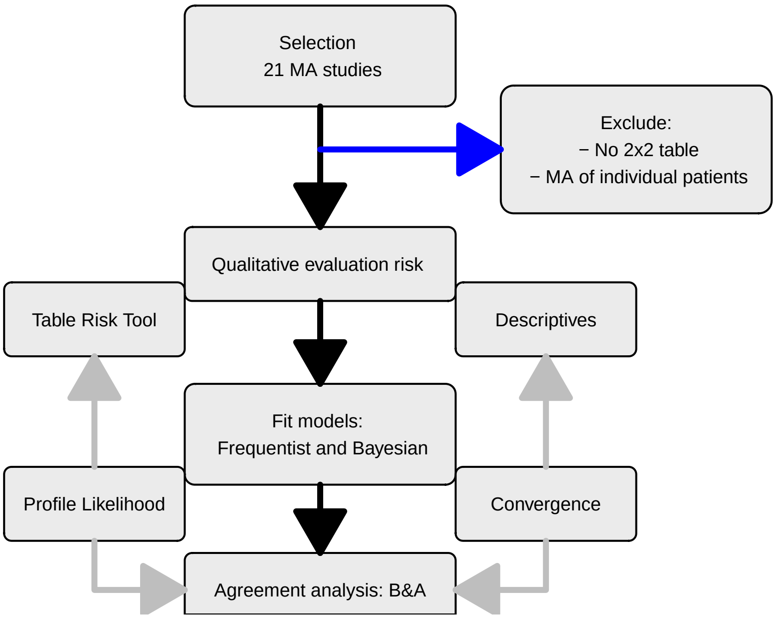


Figure 1: Flow diagram of steps that were followed for the meta-research on statistical methods of the selected MA of DTA studies

contained zeros in the 2x2 table. A MA of DTA was classified as "medium risk" if 20-40% of the studies contained zeros. Otherwise, a MA of DTA was classified as "Low risk". The name of the variable is `rare cases` in Table 5.1.

2. Item 2, 3, 10, 11, and 17: The assumption of these items does not hold when we see "small studies, sparse or skew data". This item was assumed not only when studies provided cells (from a 2x2 table) equal to zero, but also when studies were small. A DTA study was considered small when having less than 30 patients, a rate of small studies greater than 0.3 were considered as medium risk and 0.6 and higher risk. `Rate small st` and `rare cases` are the variables of interest in Table 5.1.

3. Item 4, 8, 12, and 16. The assumption of these items does not hold when we are facing "outlying studies". A MA of DTA was scored as "high risk" if we had 0.60 or more of the studies classified as influential. A MA of DTA was at "medium risk" if we had between 0.30-0.60 of the studies classified as influential; or else, a MA of DTA was classified as "low risk". The name of the variable is `Rate of influentials` in Table 5.1 .

4. Item 5 and 13. The assumption that tp_i (or tn_i) is unbiased for TPR_i(or TPN_i) did not hold when small studies were present, and also when a correlation between Logit(sensitivity) and their variance was present. A MA of DTA was at "High risk" when the correlation was greater +0.7 or lesser than -0.7, but also jointly has the presence of small studies (item 2 of this list). If only one of the previous elements were present it was considered as "medium risk" and "Low risk" if no elements were present. Finally to be scored as "high risk" both elements needed to be present. `Rate small`, `Corr Se`, and `Corr Sp` are the variables of interest in Table 5.1.

5. Item 6 and 14. The assumption of these items does not hold when "few studies are present" or when we have an imprecise σ_{SE}. An imprecise σ_{SE} doesn't occur when there are reasonably extensive studies (Jackson and White 203318); a cutpoint of at least 30 patients in each study were considered, as previously stated with the same rate. The used variable is Rate small st: a rate of small studies greater than 2/3 were regarded as "high risk" between 1/3 and 2/3 "medium risk", a rate of small studies below 1/3 were viewed as "low risk". The name of the variable is `N Dta` in Table 5.1.

6. Item 7 and 15. The assumption of these items is not safe when "few studies are present". A MA of DTA with less than 10 DTA studies was considered as "high risk". A MA of DTA with less than 15 was regarded as a "medium risk", and finally a MA of DTA with greater than 15 DTA studies was considered as "low risk". The name of the variable is `N Dta` in Table 5.1.

5.4.2 Quantitative analysis:

The quantitative modelling was conducted following the frequentist and Bayesian bivariate random effect models previously described. The variations on the normality assumption were compared through a normal-normal (all normality assumption are made) with a binomial-normal (frequentist and Bayesian, only normality assumption are made in the higher level). Finally, the previous models were compared with a Bayesian binomial-mixture of normal. Four random-effects models were fitted in each MA of DTA study: frequentist normal-normal (NN), frequentist binomial_normal (BN), Bayesian binomial-normal (BBN), and Bayesian binomial-Mixture of normal (BBM). Each model was checked for convergence and was explored through the profile likelihood (frequentist) or Posterior density function (Bayesian) (See Figure 1). Also, each fitted model was compared through their estimations: logit(sensitivity), logit(specificity), variance(logit(sensitivity)), variance(logit(specificity)), covariance, and correlation. The comparison was made through the Bland & Altman method of agreement.

The estimations of sensitivity, specificity, and uncertainty measures was done in each frequentist model (normal-normal and binomial-normal) using a bivariate mixed-effects regression model, fitted with `PROC MIXED`, `PROC NLMIXED` and `PROC GLIMMIX` with SAS/STAT version 9.4. Such models were done following the previously published codes by Reitsma et al. (2005), Riley et al. (2007) and Menke (2010), respectively.

6 Chapter 6: Results of the meta-research experiment

In this chapter, we present the results of our meta-research experiment, organizing the results as follows:

1. Summary of the MA of DTA (1.1 Data characteristics, 1.2 Methods, and software used)
2. Results of risk of statistical inference (2.1 Qualitative results: RIT, 2.2 Quantitative results)
3. Agreement analysis
4. Relationship between the presented findings

From the original 21 MA of DTA, we get 55 estimations due to the subgroup estimations. In Table 6 (appendix) we can see the subgroups nomenclature present in each MA. Each one of the subgroup-estimations was considered as an independent study for the modelling estimations (Tables 8, 9, 10, 11, and 12, from appendix), using the ID notation (`Id sub studies` column) presented in the Table 6.The same ID number is used for the estimation Tables (Tables 8, 9, 10, 11, and 12, from appendix) and also as identification in all the plots presented as results.

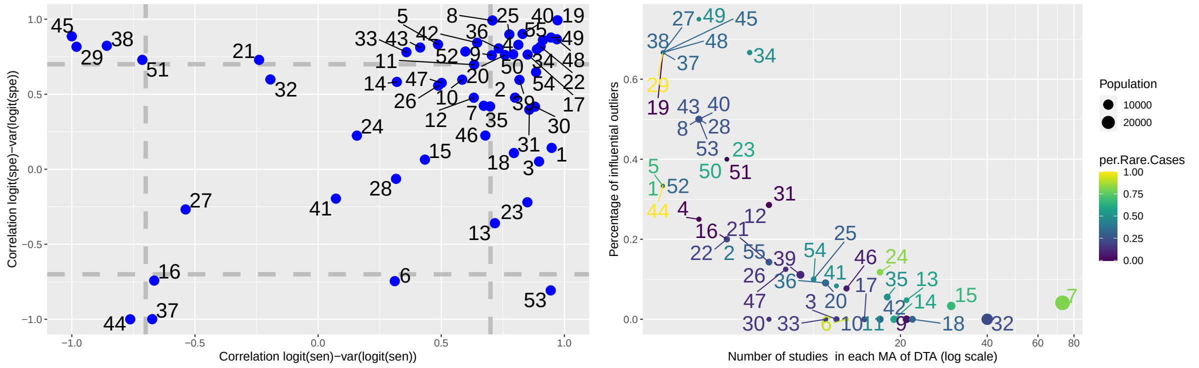
6.1 Data characteristics: descriptive features

With the purpose to provide support to the qualitative RIT scoring, we provide descriptive measures, that can help to discriminate the statistical risk related to each study, based on the descriptive numerical findings. Such descriptive measures are based on the cutpoint presented in the previous section.

The descriptive values are presented in Table 5.1, coloured with the potential risk of each measure. In Table 5.1, the numbers printed in orange represent a medium risk, and the numbers written in red denote high risk. An explanation of the used variables in Table 5.1 is described as follows. *N DTA* represents the number of DTA studies in each subgroup selected (or all the studies in case no subgroups, see Table 6). The variable *Ratio prev* represents the ratio between the lower prevalent DTA study divided by the prevalence of the highest, a small *Ratio prev* shows more risk for normality assumptions (RIT). The columns *rate.RC.Se* and *rate.RC.Sp* are the number of "rare cases" or cells equal to zero for sensitivity or specificity related cells (FN=0 or TP=0 for Sen, and TN=0 or FP= for Spec). The *Rate Infl* variable represents the rate of studies where sensitivity/specificity outcomes show an influence trend in influential plots (done with `metafor:rma`, plots are not presented for space reasons only summarized as *Rate Infl*). The *Rate small st* variable represents DTA studies with less than 30 patients divided by all studies. Finally, the *Corr Se* and *Corr Sp* variables are the Pearson correlation index between the logit of the accuracy measure and its variance, an index higher than 0.7 (or lower than -0.7) is considered to provide risk. In Table 5.1, the numbers printed in orange represent a medium risk, and the numbers written in red denote high risk. A brief exploration of the descriptive measures are presented in the Figures 2 and 3, and will be explained as follows.

In the studies, we observed that a high correlation (an index greater than 0.7 or lesser than -0.7) (Figure 2a) between accuracy (logit) measure and the variability is a common situation, and only a few amounts of studies present low correlation. In the correlation plot (Figure 2b), presents at the same time three risk factors: number of studies, rate of influential and rate of rare cases. We can see that only taking these three measures, a low amount of studies show a low trend risk. Low-risk MA of DTA studies is recognized when present in more than 15 DTA studies. Also, we can see that when the number of studies is lower than 10, the probabilities to have influential outliers increases.

An additional descriptive plot is the treemap plot (see Figure 3), which allowed us to visualize prevalence and DTA studies size. The treemap is a hierarchically organized plot where each MA of DTA contains their DTA studies coloured with the prevalence in each DTA. Three plots are presented to allow that

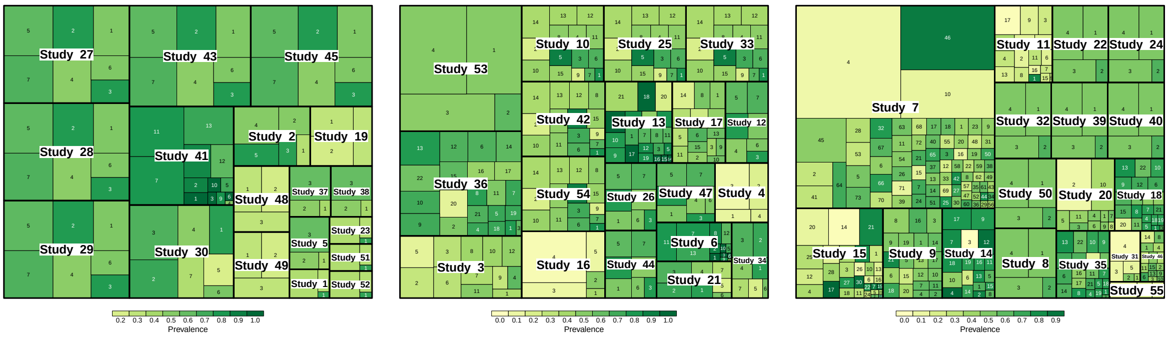


(a) Pearson correlation indexes in each MA of DTA between the acurracy and their variance

(b) Scatterplot between number studies and percentage of rare cases (DTA studies with 0 events in one or more cells)

Figure 2: Descriptives explorations about risk of compromised statistical inference tool

similar studies (in population size) are compared in the same treemap, to avoid scalability issues between larger and shorter studies. These plots bring us a valuable insight into the collected sample of studies: it's more likely to have more heterogeneity in the prevalence and in the size rates when studies are "larger" (like studies: 7, 15, 14, 11) which also doesn't have any sub-grouping. Also, we can see a more homogeneous pattern (prevalence and size) when the MA of DTA has less studies and patients. We think that this trend is not casual, and reflect the true nature and purpose of sub-grouping in the original MA: to bring a homogeneous comparison.



(a) MA of DTA Studies 29, 1, 52, 5, 23, 37, 38, 48, 27, 51, 45, 19, 2, 49, 28, 43, 41 and 30

(b) MA of DTA studies 21, 6, 4, 26, 12, 25, 34, 44, 33, 47, 17, 54, 13, 16, 42, 10, 53, 36 and 3

(c) MA of DTA Studies 40, 31, 46, 22, 24, 55 ,35, 18, 20, 8, 11, 14, 9,39, 15, 32, 7 and 50

Figure 3: Population proportion size in each MA, colored with the prevalence of each DTA, divided in three group for scalability reasons

6.2 Risk score for the meta-analyses of diagnostic test accuracy studies

We have seen the amount of descriptive that each study can provide (Table 5.1). It can be a challenge to summarize these indexes and conclude the potential impact on the inference process under the normality assumption with the collected raw data. Now, a measure that collects and weighs these characteristics, relative to inference risk assumptions, seems valuable. One benefit of the RIT is to allow us to summarize the descriptive measures presented in the context of risk of normality assumptions.

In this subchapter, the RIT score previously defined was calculated for the MA of DTA studies, using the previous descriptive measures. The already presented cutpoint for such descriptive measures is in relationship to each item of the RIT for MA of DTA. The results of the RIT score in the selected studies are presented in Figure 4 using the red colour for high risk, yellow for medium risk, and green for low risk. An alternative summary of the RIT Table was done transforming the three categories into numbers: three (3) as red, two (2) as yellow and one (1) as green.

A connection between the binary score and the previous descriptive variables are presented in Figure 5, which represent the "closeness" between different descriptive measures but coloured with a binary RIT score of the MA of DTA studies. The principal component analysis (PCA) Figure suggests an interesting connection between variables like rate of influential studies and rate prevalence [5], rate of small studies and rate of "rare cases" or cells equal to zero for sensitivity or specificity [6], and finally the number of studies and the total population in a MA of DTA study. Furthermore, the "cloud" of low-risk studies (blue triangles) has a higher eigenvalue for the population and number of DTA studies, and the "cloud" of high-risk studies (red circles) has more rate of influential studies, rate of prevalence and rate of small studies. This highlights that these variables produce a more significant impact in the scoring process than correlation, which somehow is more prevalent in "low-risk studies" according to the binary RIT score. Finally, the MA of DTA studies that have a higher population/number of DTA are the ones that have a lower amount of rate of influentials and low rate between studies prevalence.

Different PCA with a different cutpoint was done to decide the optimal measure in our 55 subgroup samples. In Figure 5 the PCA is coloured with the chosen cutpoints: a RIT score 33 or bigger is considered as high, and 32 or lesser as low. The cutpoint of 33 for RIT produces a binary risk index. The binary RIT was used in the following analysis to predict how the models would perform in each MA of DTA. The model performance in each MA of DTA consists of convergence of each model in the observed data, presence of non-positive definite (G or H matrix) and PL irregularities.

A triplot or ternary diagram (Figure 7) is presented which shows the percentage of high/medium/low score items in each MA of DTA subgroup. This plot can provide us with a simple fact that a high score can be achieved not only by a high percentage of red scored items (like studies 29, 43, 40, 50 and 28), but also by studies with a high percentage of yellow scored items (like 27, 37, 49, 23), and finally, a low score is possible only with a high percentage of green score item.

[5] which make sense because if the prevalence between studies is quite different, it can potentially convert into influential in the MA of DTA analysis

[6] which also make sense because this small studies can have more chance to present zero cells

Table 5.1: Descriptives measures summary of the 55 studies (21 MA)

Study	MA	N Dta	Pop	Pop prom	Rate preval	rate.RC.Se	rate.RC.Sp	Rate Infl	Rate small st	Corr Se	Corr Sp
1	1	3	62	21	0.48	0.33	0.33	0.33	0.67	0.95	0.14
2	2	5	245	49	0.44	0.20	0.20	0.20	0.20	0.80	0.48
3	3	12	1432	119	0.24	0.17	0.00	0.00	0.00	0.90	0.05
4	4	4	598	150	0.59	0.00	0.00	0.25	0.00	0.81	0.83
5	5	3	113	38	0.55	0.33	0.33	0.33	0.67	0.49	0.83
6	6	13	553	43	0.50	0.08	0.92	0.00	0.62	0.31	-0.75
7	7	73	27091	371	0.10	0.03	0.79	0.04	0.00	0.67	0.42
8	8	4	2892	723	0.81	0.00	0.25	0.50	0.00	0.71	0.99
9	9	21	3579	170	0.28	0.00	0.05	0.00	0.00	0.71	0.76
10	10	15	1181	79	0.31	0.07	0.27	0.00	0.13	0.59	0.60
11	11	17	2982	175	0.08	0.18	0.35	0.00	0.06	0.63	0.70
12	12	7	658	94	0.11	0.00	0.14	0.29	0.00	0.63	0.48
13	13	21	1034	49	0.28	0.19	0.38	0.05	0.38	0.72	-0.36
14	14	19	3173	167	0.03	0.05	0.58	0.00	0.00	0.32	0.58
15	15	30	5250	175	0.05	0.50	0.37	0.03	0.20	0.43	0.06
16	16	5	1054	211	0.17	0.00	0.00	0.20	0.00	-0.67	-0.74
17	17	15	820	55	0.32	0.20	0.00	0.00	0.07	0.85	0.76
18	18	22	2411	110	0.22	0.14	0.23	0.00	0.05	0.80	0.11
19	19	3	198	66	0.62	0.00	0.00	0.67	0.00	0.97	0.99
20	20	11	2844	259	0.30	0.27	0.00	0.09	0.00	0.76	0.76
21	21	7	522	75	0.40	0.00	0.14	0.14	0.14	-0.24	0.73
22	8	5	1662	332	0.51	0.00	0.20	0.20	0.00	0.90	0.82
23	1	5	147	29	0.48	0.20	0.40	0.40	0.60	0.85	-0.22
24	8	17	1988	117	0.26	0.00	0.82	0.12	0.06	0.16	0.22
25	10	10	670	67	0.36	0.10	0.20	0.10	0.00	0.78	0.90
26	12	7	625	89	0.40	0.00	0.14	0.14	0.00	0.49	0.56
27	12	3	196	65	0.62	0.00	0.33	0.67	0.33	-0.54	-0.27
28	12	4	320	80	0.60	0.00	0.25	0.50	0.25	0.32	-0.06
29	12	3	61	20	0.62	0.33	1.00	0.67	0.67	-0.98	0.82
30	21	7	507	72	0.40	0.00	0.14	0.00	0.14	0.88	0.42
31	16	7	1488	213	0.34	0.00	0.00	0.29	0.00	0.86	0.40
32	8	40	13561	339	0.05	0.07	0.17	0.00	0.00	-0.19	0.60
33	10	11	742	67	0.18	0.00	0.18	0.00	0.09	0.36	0.78
34	21	6	726	121	0.24	0.33	0.50	0.67	0.00	0.89	0.80
35	18	18	2326	129	0.13	0.11	0.44	0.06	0.06	0.70	0.42
36	18	11	1329	121	0.17	0.18	0.18	0.09	0.00	0.73	0.81
37	5	3	153	51	0.46	0.00	0.33	0.67	0.33	-0.67	-1.00
38	5	3	153	51	0.46	0.00	0.33	0.67	0.33	-0.86	0.82
39	8	9	4172	464	0.24	0.11	0.11	0.11	0.00	0.82	0.60
40	8	4	1476	369	0.39	0.00	0.25	0.50	0.00	0.83	0.90
41	6	12	398	33	0.00	0.25	0.33	0.08	0.67	0.07	-0.20
42	10	17	1159	68	0.17	0.06	0.29	0.00	0.12	0.65	0.84
43	12	4	321	80	0.39	0.00	0.25	0.50	0.25	0.41	0.81
44	12	3	736	245	0.44	0.33	1.00	0.33	0.67	-0.76	-1.00
45	12	3	197	66	0.44	0.00	0.33	0.67	0.33	-1.00	0.89
46	17	13	1507	116	0.19	0.08	0.00	0.08	0.08	0.68	0.22
47	12	8	756	94	0.20	0.00	0.12	0.12	0.00	0.50	0.58
48	19	3	190	63	0.39	0.33	0.00	0.67	0.00	0.97	0.87
49	19	4	308	77	0.31	0.50	0.00	0.75	0.00	0.95	0.88
50	8	5	505	101	0.26	0.00	0.60	0.40	0.00	0.79	0.77
51	1	5	196	39	0.26	0.00	0.00	0.40	0.40	-0.71	0.73
52	1	3	105	35	0.39	0.33	0.00	0.33	0.33	0.60	0.78
53	8	4	1282	320	0.31	0.25	0.25	0.50	0.00	0.94	-0.81
54	10	10	1005	100	0.22	0.10	0.40	0.10	0.00	0.89	0.65
55	4	7	2256	322	0.36	0.29	0.00	0.14	0.00	0.91	0.86

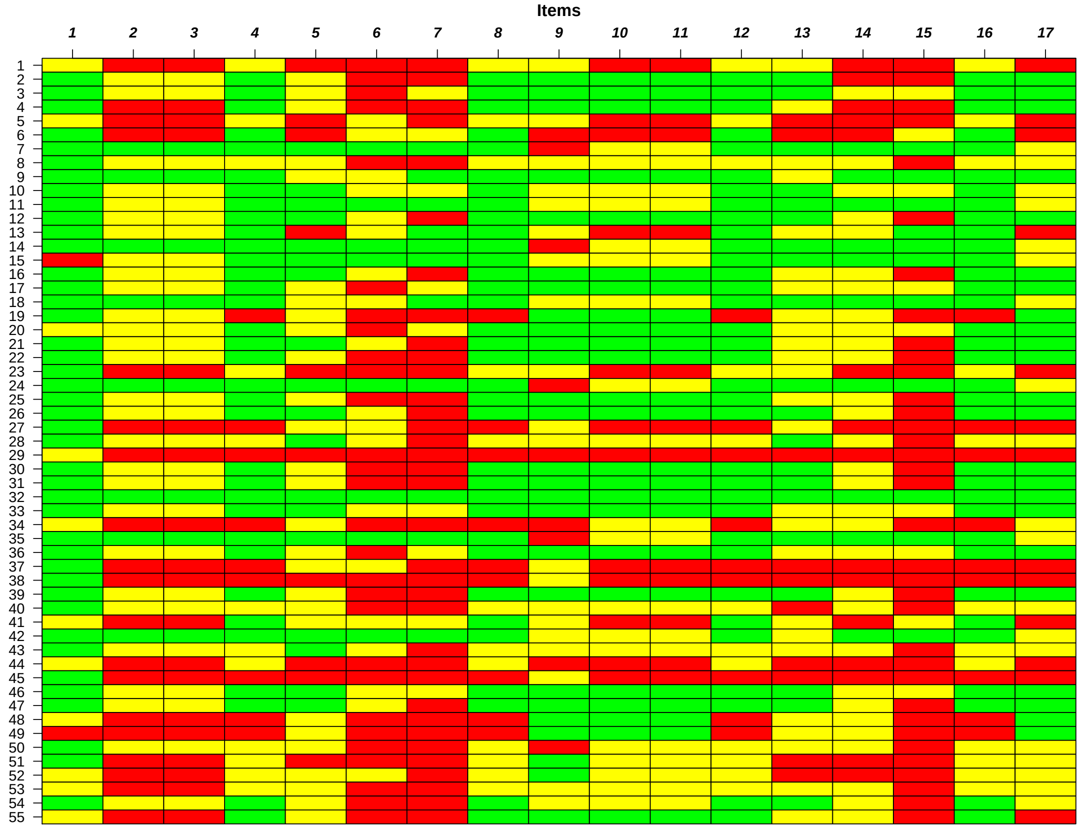


Figure 4: Risk of compromised statistical inference tool applied into 55 MA studies: colored and transposed version. Risks items from 1 to 17 are kept in the same order as the original table

According to Figure 4 24 studies have a RIT score higher than 33 (43% can be considered as risky to make normality assumption); and the rest 31 studies that can be considered as a low risk to make normality assumption, according to such cutpoint.

6.3 Methods and software used in the original published meta-analyses of diagnostic test accuracy

In this chapter, a description of the methods and software used for the original MA of DTA was done. In general, we can say that only seven (7) studies provide general pooled estimates: (Bellini et al. 2019; Bin et al. 2019; Farahani and Baloch 2019; He et al. 2019; Lee et al. 2019; Li et al., n.d.; Tsou et al. 2019).The rest of the studies (15) made subgroup analyses or ROC curve analysis. All the studies provide only frequentist analysis.

A brief review of the statistical decisions of each study can provide us with key information to understand the potential associated risk of bias. From the 21 MA of DTA selected studies, general features of each study were collected (see Table 7 in the appendix). From the selected studies 11/21 (52,38%) used `Revman` and/or `Meta-Disc`. The only study that uses exclusively descriptive analysis is Barnsley and Barnsley (2019), which doesn't produce any pool or line estimation. All the rest, blend "descriptive/exploration" software like Revman or Meta-Disc with software that could potentially perform a hierarchical analysis like `Stata` or `R`. That several MA of DTA studies base their decision on a univariate heterogeneity index is the most obvious red flag in this Table. These studies are: Bellini et al. (2019), Faias, Pereira, Luís, Chaves, et al. (2019), Faias, Pereira, Luís, Cravo, et al. (2019), He et al. (2019), Li et al. (n.d.), Li et al. (2019), Wei, Zhao, and Wang (n.d.), Yoon et al. (2019), and Zheng et al. (2019). Model selection based

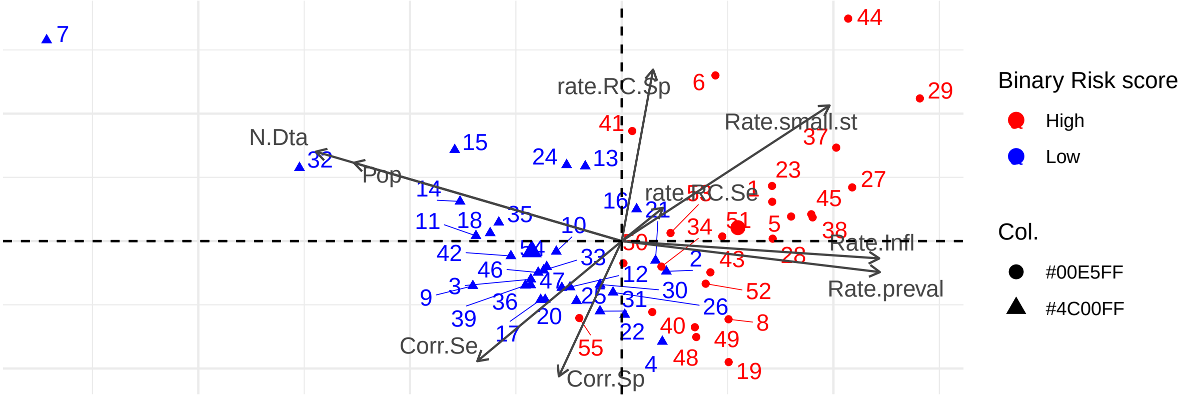


Figure 5: Principal components analysis biplot of the descriptives measures in each MA with eigen-vectors axis included

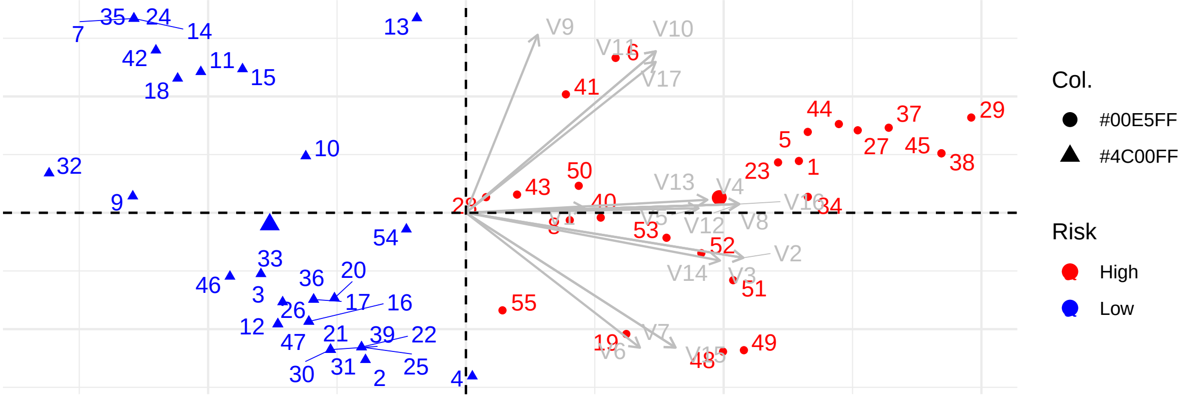


Figure 6: Representation of the score for the 55 MA studies with Biplot of RIT Score. Red dots represent a higher binary risk assign by RIT, and blue ones a lower RIT risk. The eigenvectors are the items of the RIT score.

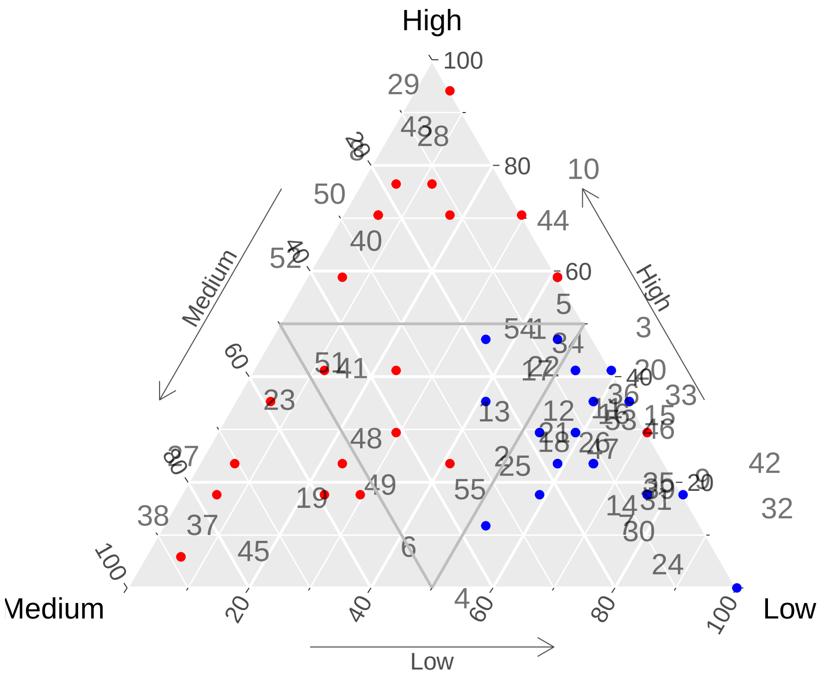


Figure 7: Triplot or Ternary diagram of Rit Score. Red dots represent a higher binary risk assign by RIT, and blue ones a lower RIT risk

on a univariate heterogeneity index is a potential source of bias. This flaw can explain the difference between the original estimated values of each study, and the random effect model estimation provided in the following sections.

Another methodological flaw observed in the reviewed MA of DTA comes from the use of Meta-Disc software forest-plots. Such forest plots produce univariate analysis pooled estimates. However, authors of the reviewed MA of DTA didn't explicitly report it, which we assume is a common practice. The studies reported (Bellini et al. 2019, @binValueThreeDimensionalUltrasound2019a, @faiasGeneticTestingVs2019, @faiasKRASCystFluid2019, @zhuDiagnosticValueVarious2020) a "bivariate" model, but a forest-plot was presented with a univariate estimation.

6.4 Quantitative analysis: model performance

We consider for the model performance in each MA of DTA, three features: convergence of each model in the observed data, presence of non-positive definitive (G or H matrix) and PL irregularities.

6.4.1 Convergence of the models

In general, all the main convergence characteristics of each model can be seen in Figure 8B: convergence, and positive-definite of the G (random effect variance-covariance matrix) and H (Hessian) matrix . In SAS, a positive H matrix confirms that the estimate is a maximum. In SAS, G matrix is designated for the variance-covariance matrix for subject-specific effects. When G matrix is non-positive-definite, it means that one or more variance component in the random statement is estimated to be zero

[7], due to that in a hierarchical model PROC MIXED or PROC GLIMMIX apply a logical lower boundary to all variance components making them greater than 0.

The convergence status and the positive-definite of the G/H matrix are information provided automatically by SAS procedure and collected for all the fitted models. In Figure 8B, the significance of the acronyms are NN.Status/GBN.Status= convergence status in NN model/`GLIMMIX` BN model, NN.pdG/GBN.pdG= Positive-definite in the G diagonal matrix for the NN model/`GLIMMIX` BN model, NN.pdH = Positive-definite in H matrix for the NN model. The interpretation for the colour in the Figure 8B and C are grey in the case to have a positive-definite matrix results and red for non-positive definite status.

The ideal convergence model scenario is not only to have an "ok" general status converge but also to present a positive-definite G and H matrix. This feature confirms not only convergence of the model, but the stability of the covariance estimates (G matrix), also that the estimates are a maximum (H matrix). The failure in convergence in the NN model, could not be improved using manual starting values. In particular, BN classic model in GLIMMIX procedure, the SAS default starting values were enough to achieve convergence but not positive-definite for all the cases (see Table 8). With the NLMIXED procedure, the fitting process requires more work because with the default starting values, several models didn't converge. So, the estimates from GLIMMIX were used in NLMIXED as starting values in 53 studies, except the study 29 and 44. Such studies required modification of the starting values which allowed us to achieve convergence. The reason for the failure of the GLIMMIX starting values in NLMIXED with the model 29 and 44, was that the central estimation in GLIMMIX presented overestimation, despite to have a "convergence" status. Once again, here we would like to stress that a convergence status does not mean that we are in a safe zone, it is required to check that G/H matrices are positive-definite. When a message of a non-positive-definite matrix is received, it means that the quality of the model convergence is questionable. Possible reasons behind this can be: 1) the estimate is not a maximum (H matrix) 2) the specified model cannot be adequately estimated with your data (G matrix). Convergence' Bayesian models were achieved without the need to work on the default starting values in bandit in all the cases.

For the process of exploring the PL functions in the BN classical models, three scenarios were chosen: a GLIMMIX BN model with default starting values, a BN NLMIXED model with default starting values, a BN NLMIXED model with the GLIMMIX starting values or manual when needed.The most surprising fact was the repercussion of the starting values on the PL in NLMIXED, which in general improves the observed discontinuities in PL when better starting values are provided.

In Figure 8B we can see that despite that all BN models converge in all the cases (GLIMMIX and NLMIXED) not all of them present a positive-definite G matrix. For NN model several studies did not converge (1, 5, 8, 15, 16, 27, 28, 29, 34, 37, 38, 44, 45, 48, and 52), we even had convergence in NN model without having positive-definite (like studies: 2,3, 6, 12, 19, 20, 21, 23, 30, 41, 49). We can realize that convergence and positive-definite are not common in our selected sample under the proposed models: in the NN model 58% (30/51) are convergent and positive-definite, and in the GBN model 58% (30/51) are convergent and positive-definite; and positive-definite convergent models in both models only a 49% (25/51) of the studies.

6.4.2 Profile likelihood and posterior density functions of the models

In the Figure 8C a summary of all PL functions irregularities can be seen for the 55 models: numerical discontinuities, and non-informative patterns. The PL irregularities were visually evaluated

[7] Probably because not enough variation in the response to attribute any variation to the random effect, producing an estimation equal to zero. It's a kind of mismatch between the model and the observed data that impact the variance estimation producing in most of the cases an underestimation of the true value.

by the author (in the appendix chapter from Figure 37 to 91). In Figure 8C, the significance of the acronyms are NN.disnum/GBN.disnum= Numerical discontinuities for NN model/GLIMMIX BN model, NN.NI/GBN.NI= PL NN model/PL GLIMMIX BN model is not informative.

A key aspect in the current analysis is the recognition of PL functions that may compromise the estimation of each study. In general, a PL that does not show a clear maximum upon simple inspection, such as flat curves, multiple peaks, or maximums on the parameter boundary present this type of problem. Such non-informative probabilities do not allow the finding of a reliable estimate and generally coincide with non-convergence issues or large heterogeneity in the studies. These situations have previously been described as a risky scenario for estimation by Gelman et al. (2013), Curcio and Verde (2011) and Riley et al. (2007).
A PL is assessed as non-informative when the peak' PL-function is observed on the frontier of the space parameter (close to 0 in the case of variance, or correlation close to +1/-1). When the PL function is non-informative in most of the cases it produces an estimation for the parameter equal to zero (Curcio and Verde 2011) when other values also can be possible, producing an underestimation of the variance in most of the cases; this situation usually produces a positive-definite G matrix. Furthermore, the PL function is also considered non-informative when the PL function is completely flat. A PL with numerical discontinuities is observed in the iterations process through a grid when the PL in some points produces a non- convergence , that can be seen as a discontinuity in the PL function plotting like Figure 37, 48, 57, 59, 65.

The PL function plot can be used to check how plausible and reliable can be the provided estimation by the software, under the normality assumptions in NN and BN models, and contrasting with the Bayesian posterior distribution of the same parameter, which relaxes these assumptions in the whole hierarchical structure. The PL function needs to provide enough and non-conflicting information to support a numerical estimation by the software procedure. The ideal scenario for a PL function is the one that presents a unique peak in the space parameter. Another non-ideal scenario occurs when the function maximum is observed in the space frontier' parameter [8]. If the normality assumptions are not risky in the observed data of each study, we expect to see similar distribution between the four models. Specifically, a normal assumption can be safely made when the variance of the logit(sensitivity) or logit(specificity) follows a normal shape or at least a gamma distribution shape. The same occurs for the PL in the covariance between logit(sensitivity)-logit(specificity), where we expected to see a beta shape distribution. In general, we can see (in the appendix chapter from Figure 37 to 91) that these optimistic scenarios rarely occur in our selected sample studies.

A PL and posterior function exploration of the different fitted models are presented case by case in the 55 studies (in the appendix chapter from Figure 37 to 91). In each PL panel study, additional information is summarized like descriptive measures and RIT scores; the blue vertical lines represent the estimation points for each model. In PL NN Mixed when the model is not convergent no line is printed , because PROC MIXED do not provide an estimation. However, when a BN NLMIXED does not converge, PROC NLMIXED still can give an estimation in this case lines estimation are red. We have blank areas in the panels because in PROC MIXED and PROC GLIMMIX there's no standard procedure to find a PL for the central estimations. Also, the PL NLMIXED with manual starting values was limited to a random selection between all the MA, labeled studies 1-21. Because it's a "time-demanding task", it was avoided from study 22 to 55, which are additional subgroups of the same original 21 MA. Although we only did two versions of NLMIXED from study 1 to 21 (not for 22 until 55), it is possible to have a clear understanding of the potential impact of better starting values to improve the PL function in BN NLMIXED models. This is one of the significant insights in this exploration: the starting values usually

[8] The frontier' paramater is the value that such parameters can not assume; for example for variances zero is the frontier, and for correlation such values are -1 or 1.

Table 4: RIT prediction for model performance features in classical models, with different optimal cutoff points , and their sensitivity and specificity

Features	Optimal.cut.off	Sensitivity	Specificity
NN.Status	39	0.73	0.90
NN.pdG	36	0.65	0.86
NN.pdH	25	1.00	0.20
NN.disnum	NA	NA	NA
NN.NI	28	0.70	0.56
GBN.Status	NA	NA	NA
GBN.pdG	38	0.56	0.93
GBN.disnum	39	0.53	0.86
GBN.NI	24	0.85	0.67

change the PL function, which also leads to better chances to have more reliable estimation. [9].

6.4.3 Prediction skills of the binary risk score

This subchapter presents an exploration of the predictive abilities of the binary RIT (with a cutoff point of 33) for non-convergence, non-positive definite matrix, and PL irregularities in the classical models. One additional goal of this chapter is to explore other possibilities for a cutoff point, according to the previous model's outcomes.

One of the possible benefits of the proposed RIT score in MA of DTA is to predict situations in which normality assumptions in a classical bivariate model can produce non-convergence or non-positive-definite in the G or H matrix or PL irregularities. Outcomes with a constant result (see Figure 8) will not be possible to predict; like GBN.Status (convergence of Glimmix BN model) and NN.Disnum (numerical discontinuities for the NN model). An increment of the sample size may produce more variability in such measures, but according to this reduced sample size of studies such variables doesn't provide any change.

In Figure 9 the optimal cutpoints were determined for different model performance features. An "optimal" cutpoint for the risk score is the one that maximizes both accuracy measures. Using the cutpoints of Table 9, an average of the most frequents optimal cutpoint (36, 38, 39, and 39) was done to propose an alternative cutpoint of 38. Table 4, and 5 presents the accuracy measures for both cutpoints. The cutpoints RIT=33 and RIT=38 present a similar Youden index (sensitivity +specificity-1) (see Table 5) for each predicted feature, which means that they have similar general accuracy. The fact that both cutpoints share similar precision measures ensures that the prediction skills for the most conservative cutpoint (33) share some close similarities with one of the "best" cutpoint (38).

Similar reasoning can be done with the cutpoint equal to 24 for the model performance features: `GBN.NI` and `NN.pdH`, which noticeably increases the accuracy skills (for `GBN.NI`). Such cutpoint findings are only post-hoc explorations that can not be validated in our study, because the explorations are done after looking at the data. Future studies with bigger sample sizes could explore these additional cutpoints.

Additionally, we can stress that the plots for optimal cutoff point for RIT score (see Figure 9) show more a plateau curve shape, rather than a precise peak shape. The only exception could be `GBN.NI` which clearly present a peak when the score is equal to 24. The previous reasons can be used to reassure the original cutpoint of 33 (which is more conservative than 38), and only suggest additional cutpoint that could be explored by future research (like cutpoint of 24 or cutpoint of 38).

[9] If the use of starting values is limited only when convergence issues arise, neglecting important messages of non-positive-definite G or H matrix, potentially can jeopardize the trust in the variance-covariance estimations when G matrix is non-positive-definite, and the certainty of having a maximum when H matrix is non-positive-definite.

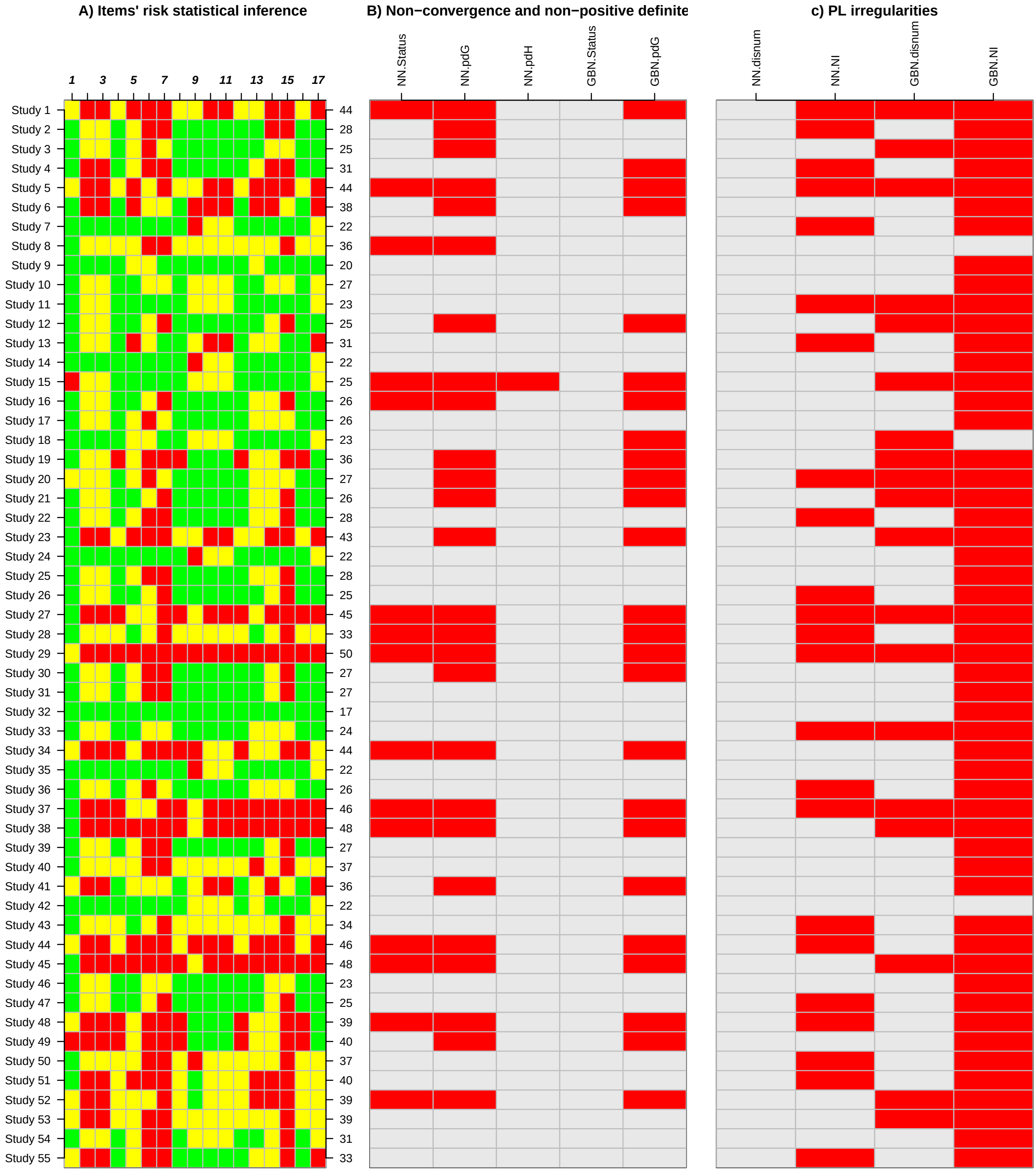


Figure 8: Summarize of the A) RIT-risk inference tool, B) non-convergence and non-positive definite matrix C) profile likelihood visual irregularities under visual inspection. Colour legend: in A) red is high risk, yellow is medium risk and green is low risk; in B) red means non-convergence or non-positive definite matrix, in C) red means numerical discontinuities, or PL non-informative

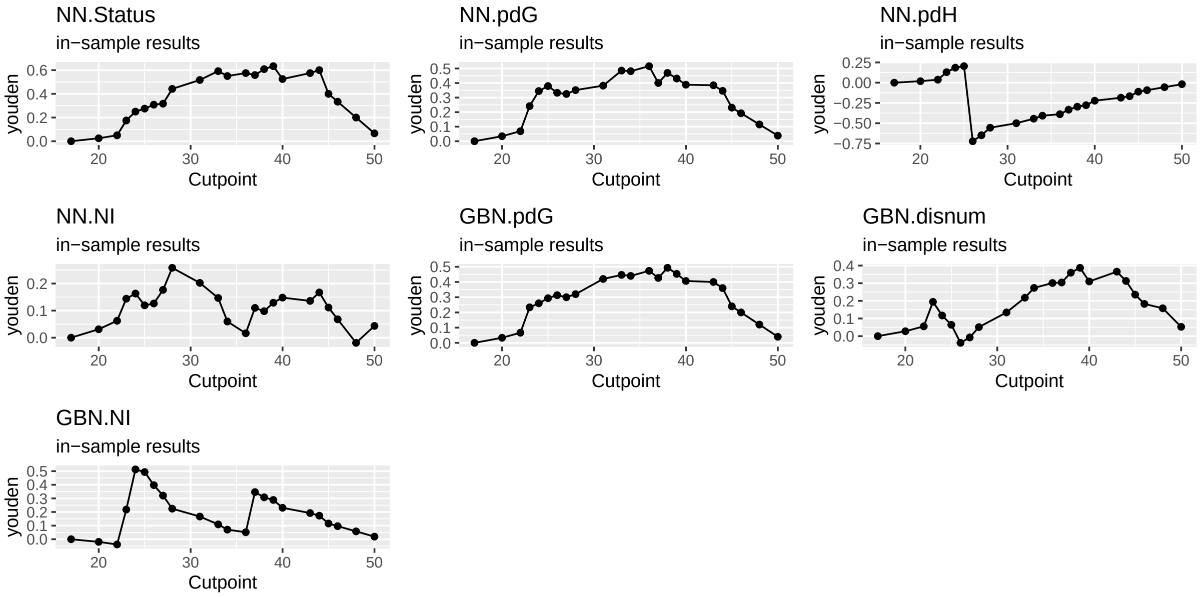


Figure 9: Optimal cutoff point for risk score to predict convergence features in classical models

Table 5: Prediction sensitivity and specificity for convergence features with risk score in classical models with a cutt-off=33 and cut-foo=38,

	Cutoff=33			Cutoff=38		
Features	Sensitivity1	Specificity1	Youden1	Sensitivity2	Specificity2	Youden2
NN.Status	0.72	0.87	0.59	0.88	0.73	0.61
NN.pdG	0.79	0.69	0.48	0.93	0.54	0.47
NN.pdH	0.56	0.00	-0.44	0.70	0.00	-0.30
NN.NI	0.62	0.52	0.14	0.75	0.35	0.10
GBN.pdG	0.77	0.68	0.45	0.93	0.56	0.49
GBN.disnum	0.64	0.58	0.22	0.83	0.53	0.36
GBN.NI	0.67	0.44	0.11	1.00	0.31	0.31

It's important to remind that the cutoff point equal to 33 provided a better separation in the PCA exploration of the RIT score (no information was used about the convergence or the PL irregularities at that analysis stage), and could be considered as a validated cutoff point for medium risk. The actual exploration for a new cutoff point can be considered for a high-risk level. Introducing a new cut off point post-adhoc changes the validity of the last cutoff point, and only can be considered as an exploratory approach. An additional study can bring validation of the sensitivity and specificity values for the second cutoff point, and confirm the value for the first one proposed already.

6.4.4 Agreement between measures

The evaluation in the proposed MA of DTA methods was done through Bland-Altman analysis(B&A plot) (Bland and Altman 1986), which is a graphical method to quantify the agreement between two quantitative measurements by constructing an interval of agreement. B&A plots help to evaluate a potential bias between the mean differences between two methods and to estimate an agreement interval through a confidence interval (Bland and Altman 1986). The use of the correlation coefficient (Pearson) can be misleading because such an index measures the relationship between two variables, not the agreement between the measures. Correlation studies the relationship between one variable and another, not the differences, and it is not recommended as a method for assessing the comparability between methods (Giavarina 2015).For example, a change in the scale provides the same correlation but not the same agreement; also, a perfect agreement only is shown in the case of a correlation coefficient equal to one. The Bland-Altman method only defines the intervals of agreements but does not conclude if the limits are acceptable or not. Such a decision needs to be done according to a clinical necessity (Giavarina 2015).

The B&A plot is a scatter plot between the difference between the two paired measurements (y-axis) and the average of these measures (X-axis). The original authors recommended that 95 of the data points should lie within $\pm 2SD$ of the mean difference (Bland and Altman 1986). In the case of a good agreement, all the dots are around the zero line with a dispersion degree not greater than intervals of agreements proposed; also it is essential to check that the degree of dispersion is uniform.

The differences between the estimation of equivalent models theoretically should be equal to zero. But this is almost, because each estimation method could implicitly have some degree of error. "However, if the variability of the differences were only linked to analytically imprecision of each of the two methods, the average of these differences should be zero" (Giavarina 2015).

6.4.4.1 Global agreement

In the Figures 16, 17, 18, 19, 20, and 21 several global B&A agreement analysis between all the model estimates in each MA of DTA are considered jointly (risky and non-risky MA of DTA for normal assumptions), to evaluate the effect as whole of the model assumption, independent of RIT score in each study.

When all the estimation comparisons are considered together, according to the B&A statistics it is difficult to find a statistical significance with 95% normal confidence interval (CI). In the agreement between Freq BN and the Bayes BN estimation of the variance of logit(sensibility), can be shown a slight underestimation pattern, where the CI does not include zero. This can be interpreted as the frequentist BN model can produce an underestimation of the variance, compared with the same Bayesian model (using non-informative priors).

A more clear trend is achieved when each MA of DTA is evaluated case by case in a detailed B&A analysis (next section). Also, the estimation of the CI using a normal interval assumption can be a bold statement, so we decide to consider a non-parametric bootstrap CI with a significance level of 90%. A 90% CI allows us a more sensitive agreement analysis, producing a larger amount of dis-agreement between studies than a 95% CI.

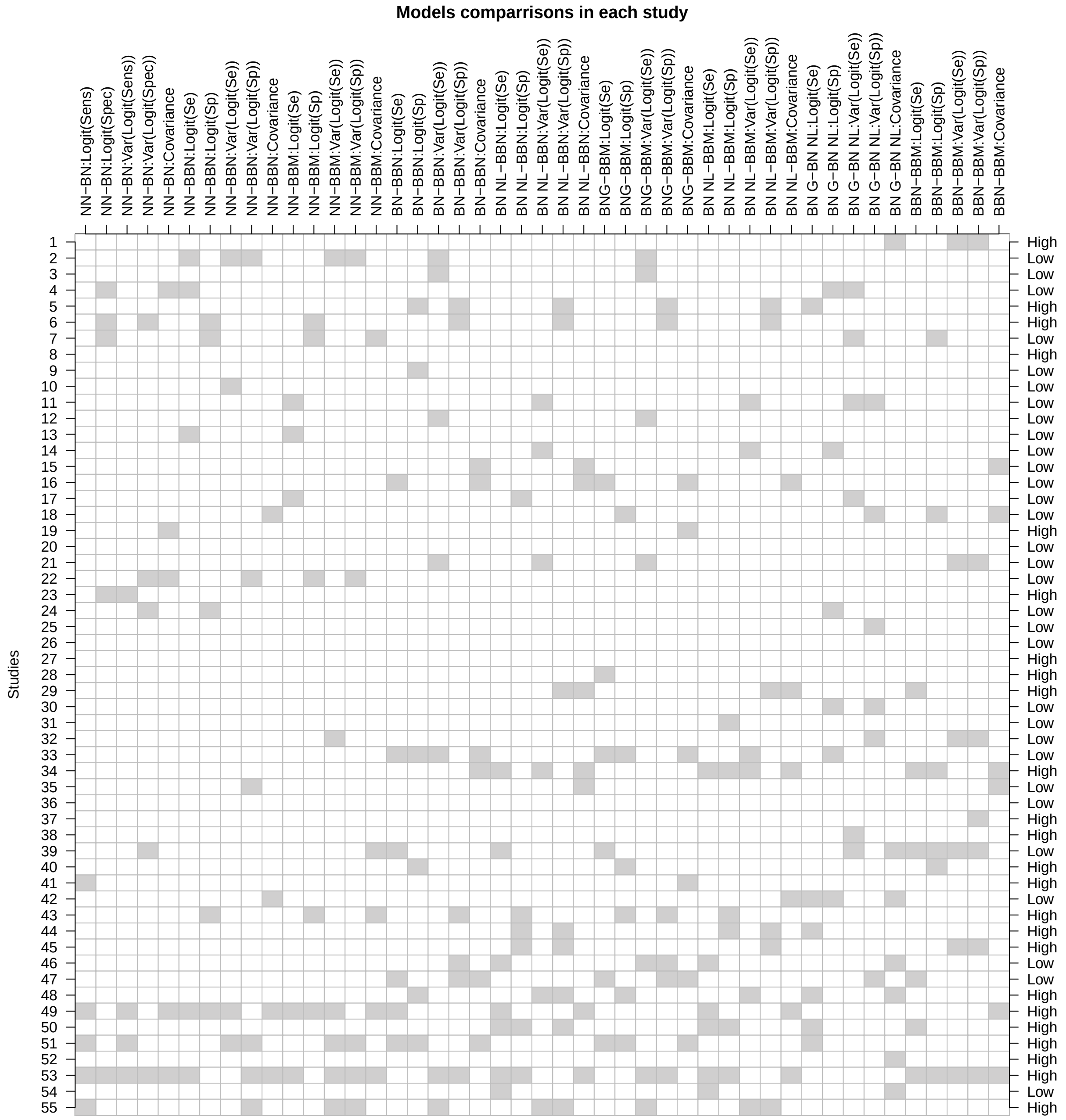


Figure 10: Disagreement according to Bland-Altman analysis between models in each study, RIT score with a cutoff point=33

6.4.4.2 Case by case agreement The agreement between two methods, according to the B&A method, occurs when the confidence interval contains the differences between each pair of measures. When two methods provide a complete agreement we can see a homogeneous cloud of dots inside the CI of the B&A plot. Special cases of agreement can be seen when the pattern is not homogeneous, when the trend is linear it can be interpreted as two methods that present "a proportional constant error, overlapped with the same proportional variability" (Giavarina 2015). In both cases, if the CI contains the dots, we can sustain that an agreement is accomplished between methods.

The confidence interval was one with a non-parametric bootstrap, to increase the accuracy in the confidence interval without the risk to assume normality in the difference between the estimation between models. The confidence level for the interval was done with a 90%; all the plots are provided coloured by the binary RIT risk (cutoff point=33) as usual.

The reasons for excluding a study in this analysis were: non-convergence in the Reistma model, no reported estimations on the original studies, and extreme outliers estimations. In the NN model fitted through `PROC MIXED`, studies without convergence (1, 5, 8, 15, 16, 27, 28, 29, 34, 37, 38, 44, 45, 48 and 52) were naturally excluded, because the software does not provide an estimation. `GLIMMIX` always provides convergence (FCONG or GCONV) and `NLMIXED` also converges in all the cases (studies 29 and 44 require manual starting values). As we previously stated studies without original estimates: 1, 23, 27, 29, 44, 45, 51, and 52 were excluded from original estimates comparisons. The excluded studies common in"non-convergence models NN" and "no original estimates" are: 1, 5, 8, 15, 16, 23, 27, 28, 29, 34, 37, 38, 44, 45, 48, 51 and 52. Specifically, between the fitted models in GLIMMIX, we found two studies estimations that require an exclusion "by hand" in the plotting process, not for convergence issues. Still, because of their extreme numeric estimation, that does not allow us to understand the pattern of the rest of the studies. Still, these estimations were considered as a disagreement in the following analysis (Table 10). Comparisons are made only in one specific direction; for example, the comparison BN (as reference) vs NN was not done because it was previously done with NN (as reference) vs BN.

A fair comparison between correlations estimations always presents an additional challenge. If the correlation estimation is calculated indirectly with the covariance and variance estimations, (or if the estimation is provided directly by the software), some estimates were equal to infinite when the variance is zero; so, these are not-defined correlations values, which are excluded from agreement analysis between correlations estimation. A Fisher transformation was applied to the correlations estimations, which allowed us to convert a skewed distribution of the sample correlation into a more normal shape distribution.

All the disagreements are summarized in the Table 10, where a grey $cell_{i,j}$ that a disagreement outside the non-parametric bootstrap CI of 90% was noticed in $study_i$ (row_i of the Table) under the model comparison of the Table $column_j$. In Figure 10, the right axis presents again the binary risk score. The original B&A plots can be seen (Figure 18 to 30) in the appendix.

No RIT score prediction for the disagreements is shown, despite it being done, because it was not significant in all the cases. We argue that a possible reason for a failure to achieve prediction skills in the RIT score for the disagreement, could be related to a lower sample size of the studies. It can also be because it cannot be predicted with the RIT score. Only a bigger study with a bigger sample size can answer this.

As follows, we explore possible relations between the number of disagreements (between all the models) in each study and: a) the RIT risk binary score b) the number of PL irregularities in model NN and c) the number of PL irregularities in model BN. These explorations can be seen in Figures 11, 12, and 13. In the Figure 11, the studies with higher disagreement: 53, 49, 55, 51, 43, 6, 5, 50, 29, 44, 48, 45 all have a RIT risk score higher than 33; but studies 33, 16, 46, and 47 are an exception which provides a low-risk score can also have a high number of disagreements.

A shallow trend also is observed that studies with PL irregularities could have more dis-agreement according to the PCA 11,but no formal conclusions can be extracted from this exploration of the data. But also “outliers” MA of DTA studies can be observed with a low RIT score, but also with a high amount of disagreements like 16, 33, 46, and 47. A similar trend can be observed when we consider the same PCA coloured by the number of PL irregularities, where studies with more disagreement present also more PL irregularities. The Figure 11 shows an asymmetrical distribution of the studies, observing a low amount of studies with higher amount of disagreement, which can reflect the disagreement rarely occurs in all the estimates, where a partial disagreement is the most common scenario. As an exploratory analysis, it is evident that we cannot appreciate a strong trend between agreement estimation between models, and RIT scores or PL irregularities of the observed studies. An improvement of this approach can be made through a simulation study that can achieve “certainty” in this type of exploration increasing the sample size of considered studies.

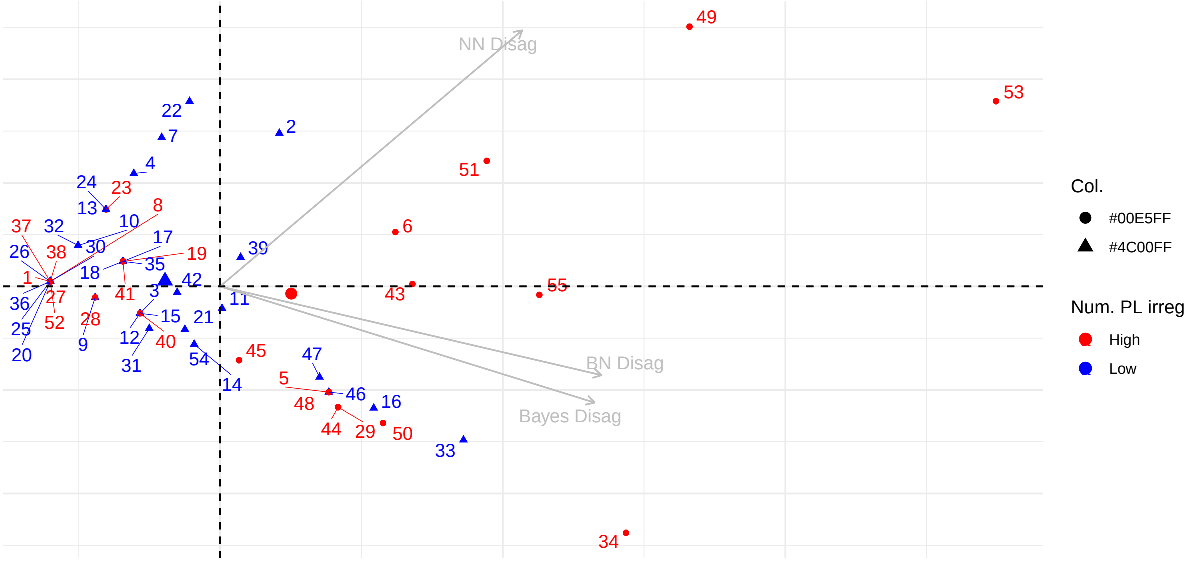


Figure 11: PCA explorations for the disagremments between models related with binary RIT risk score of 33

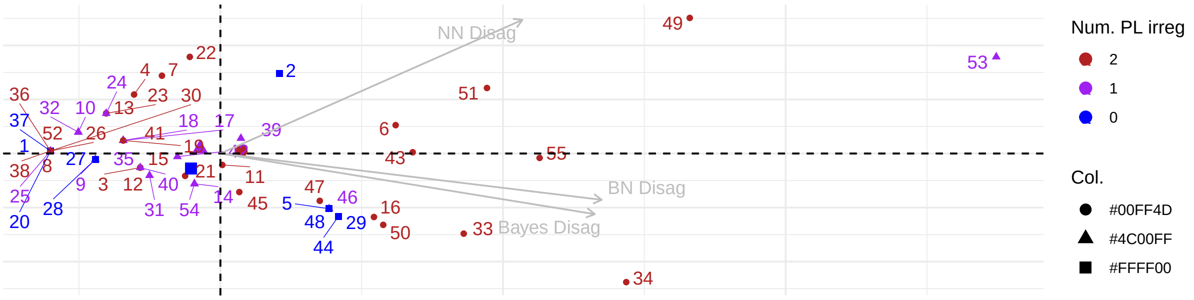


Figure 12: PCA explorations for the disagremments between models related with PL findinds on each model colored with a score for the findings on PL of the NN model

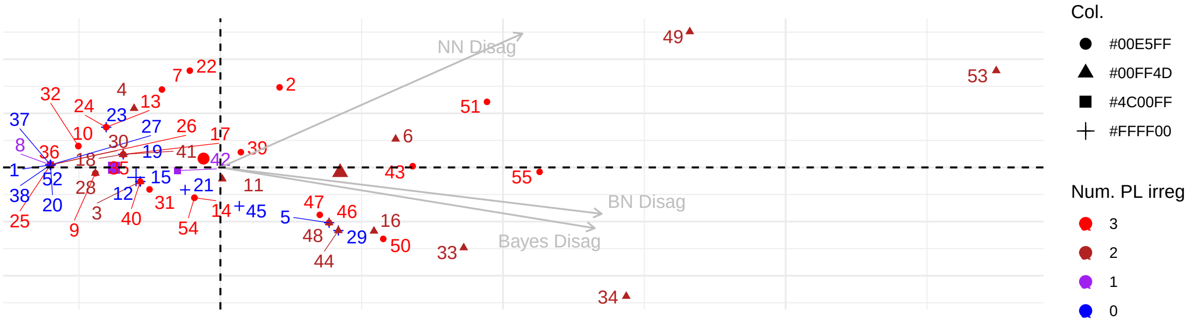


Figure 13: PCA explorations for the disagreements between models related with PL findings on each model, colored with a score for the findings on PL of the BN model

7 Discussion

In this work, we have introduced a new method to evaluate the risk of the normality assumption in MA of DTA studies. We also explored the potential risk for compromising the statistical inference when such assumptions are made. We have performed a meta-research experiment to investigate the characteristics of recently published MA of DTA.

These MA of DTA and their subgroup estimates have been used to test our risk score. We have found that 43% of the recently published MA have a higher risk (RIT score greater than 33) of delivering potentially misleading results. We validate these findings checking the predictive skills for the RIT binary score for the model performance features (the non-convergence status, non-positive definite and PL irregularities). The model performance features in the observed studies show evidence of substantial instability in the proposed classical model results. These findings are in agreement with studies that had demonstrated that convergence could be an issue under non-favorable circumstances (a small number of observations, small number of studies, small studies, heterogeneity) (Riley et al. 2007; Takwoingi et al. 2017; Hamza, van Houwelingen, and Stijnen 2008 ; Röver and Friede 2018).

The 42% of the MA of DTA presented non-convergence and non-positive-definite G matrix. Only 49% of the studies had convergence and not-positive-definite G matrix in both models (NN Mixed and BN Glimmix).

A cutpoint of 33 in the RIT risk score in MA of DTA can predict the appearance of a non-positive definite in the NN model having a sensitivity=0.75, and specificity=0.742. Finally, the prediction for the non-positive-definite for the GLIMMIX BN model presented a sensitivity=0.708, and specificity=0.742. PL irregularities produce a very low accuracies measures (for NN sen=0.62 and spe=0.52, and for BN sen=0.67, spe=0.44).This feature seems to be the most difficult features to predict in both models. The PL in a model is not a “fixed reality”, depends not only on the observed data and the proposed model which it’s obvious, but also depend on the starting values in the fitted models. Therefore, the starting values in the proposed classical models in MA of DTA are essential to improve the trust in the estimates when the normality assumption is at risky. In a post-hoc exploration, two additional cutpoint was presented 24 and 38, but no validated accuracies measures can be presented.

No clear trends were found in relation to the dis-agreement between statistical methods when a high RIT score or when PL irregularities were presented. Previous simulations studies have shown that in the case of high heterogeneity; few or small studies, the BN model showed better coverage probability, and/or less bias than a NN model (Riley et al. 2007; Takwoingi et al. 2017; Hamza, van Houwelingen, and Stijnen 2008) . These are not necessarily conflicting evidence with our findings, because our study is not a simulation, so we don’t have true values for the estimations to be compared to the models estimation.

8 Conclusions

According to the proposed RIT risk score with a cutoff point of 33, 43% (24/55) of the studies can be considered with a high risk to make a normality assumption, that can compromise the statistical inference under such assumptions.

In the NN and BN models, only 58% (30/51) of the studies had convergence and not-positive-definite G matrix. But only 49% (25/51) of the studies had convergence and not-positive-definite G matrix in both models (NN Mixed and BN Glimmix).

In MA of DTA a cutpoint of 33 in the RIT risk score can have moderate accuracy to predict non-positive-definite in the NN and BN model. RIT score predict the presence of non-positive definite in the NN model with a sensitivity=0.79 and specificity= 0.69. Also the RIT score predicts the non-convergence in the NN model have a sensitivity=0.72, and specificity=0.87. Moreover, it was possible to predict a non-positive-definite for the GLIMMIX BN model with a sensitivity=0.77, and specificity=0.68.

Lower accuracy measures are present when the risk is used to predict numerical discontinuities in BN Glmmix model (sen=0.64, and spec= 0.58), and the rest of the convergence features. We also explore additional cutpoint of 24 and 38, post-hoc, so no validation can be proven in such cutpoint additional studies could explore such cutpoints.

It's important to highlight that in all the cases where the convergence was not achieved, or non-positive definite was present in the observed studies under classical models, these studies were possible to fit under Bayesian methods, without any difficulties. We don't recommend the use of univariate models when convergence is not achieved in bivariate classical models; it's evident that the Bayesian methods are a more robust solution in this case.

The value of the RIT predictive ability is related to the potential use of the RIT score for researchers planning to conduct a MA of DTA. A researcher without the knowledge to run a full diagnosis to choose the most appropriate model can be benefited with the application of the RIT score before selecting the model. The researcher can know in advance the risk in the normality assumptions even before fitting the model when the RIT score is used accordingly.

9 Recommendations

The essential advantage of our scoring is that we can assess the risk of statistical inference being compromised in a single MA of DTA. This advantage is precious to researchers who usually don't know which statistical method is the most appropriate. In the event that a model needs to be chosen prior to the model fitting, conducting a risk assessment with RIT score can help to choose the proper model. If the RIT score greater than 33 we recommend to use the hierarchical Bayesian models.

In MA of DTA applying the hierarchical Bayesian models is the wisest choice, because they are robust against data sparsity, outliers, and in general studies features that produce a risk to compromise the statistical inference when normality assumptions are made.

It is possible in MA of DTA to conduct a "sensitivity analysis" between classical and Bayesian estimates, to check the agreement of the estimates under the observed data. Historically in MA, it's a common practice to compare between fixed and random models estimates, so we encourage this comparison with more relevant models, like the Bayesian methods.

For future research, it's important to replicate this study with greater sample size, with simulations or empirical studies or both, to reassure the validity of the cutpoint proposed. Also, it's essential to continue to explore the evidence behind the proposed "rule of thumb" (based on previous simulation studies) because it will facilitate the connection between the descriptive measures and the RIT score punctuation. A reliability study (intra or extra-rater reliability) in the use of the RIT score can also help to support the implementation of the RIT score in MA of DTA.

10 Appendix

10.1 General tables about the studies

10.2 General tables and estimations tables

As follow are presented general tables about the studies (subgroup in table 6, statistical features of the studies in table 7), estimation tables (Frequentist in table 8, 9, and 10; and Bayesian table 11 and 12).

Table 6: Descriptions of presence or absense of subgroups in each MA, an the sub-groups for each study

n	Study	PMID	Groups	MA	Id sub studies
1	Barnsley & Barnsley, 2019	30976835	18F-FDG-PET,Bone Scintigraphy,Radionucleide artrogram,SPEC/CT artrogram	1	1, 23, 51, 52
2	Bellini et al., 2019	31496631	No subgoup	2	2
3	Bin et al., 2019	30807546	No subgoup	3	3
4	Chen et al., 2020	31654803	less 12 h after,24 h after	4	4, 55
5	Faias, Pereira, Luis, Chaves, et al., 2019b	31341368	Low risk + High risk,High risk + low risk vs benign,High risk + Non High risk	5	5, 37, 38
6	Faias, Pereira, Luis, Cravo, et al., 2019	31206466	Malignant +Mucinous,Malignant	6	6, 41
7	Farahani and Baloch, 2019	31301215	No subgoup	7	7
8	Gurung et al., 2019	31158516	FLA-ABS,Agglutinattion,Conventional PCR,ELISA,Lateral Flow,Lepromin Skin Reaction,qPCR,T Cell inmunological response assays	8	8, 22, 24, 32, 39, 40, 50, 53
9	He et al., 2019	31574828	No subgoup	9	9
10	Issa et al., 2017	28130609	EUS,CT,ERCP,MRCP,US	10	10, 25, 33, 42, 54
11	Lee et al., 2019	31714947	No subgoup	11	11
12	J. Li et al., 2019	31250172	MRI1,CT 1,CT 4,CT2,CT3,MRI2,MRI3,MRI4,Overall CT	12	12, 26, 27, 28, 29, 43, 44, 45, 47
13	Li et al., 2020	31939891	No subgoup	13	13
14	Shen et al., 2019	31437232	No subgoup	14	14
15	Tsou et al., 2019	31182360	No subgoup	15	15
16	Wang et al., 2019	31335700	Cervical Length,Elastography	16	16, 31
17	Wei et al., 2020	31939890	Tumor Staging,Node Staging	17	17, 46
18	Xu et al., 2019	31702649	FNA-Tg,FNAC,FNAC+FNA-Tg	18	18, 35, 36
19	Yoon et al., 2019	31470804	CRP,PCT,Presepsin	19	19, 48, 49
20	Zheng et al., 2019	31441886	No subgoup	20	20
21	Zhu et al., 2020	32011436	Overall CTCs,ctDNA,overall exosomes	21	21, 30, 34

Table 7: General analysis features of the 55 MA DTA sub-groups studies

MA	Study	PMID	Software	Model	Pooled stimated	ROC	Heterogeneity	Threshold effect	Extra analysis
1	Barnsley & Barnsley, 2019	30976835	Probvably Revman, no available aditional software	SROC	No, only SROC. Pick visually the higher curve	SROC	Not mention		Subgroup
2	Bellini et al., 2019	31496631	Comprehensive MA 2.2.064, Excel 365, and MetaDiSc 1.4	Bivariate RE/FE, heterogeneity based decission	Yes	SROC	χ^2, I^2	Not mention	No
3	Bin et al., 2019	30807546	Metadisc, Stata 14.0	Bivariate MA: FE (MantelHaenszel method) & RE (DerSimionan and Laird method) and SHROC	Yes	sHROC	χ^2, I^2	Spearman correlation Sen-Spec	Metaregression, Sub-group analysis
4	Chen et al., 2020	31654803	Stata mendi	Bivariate RE, Hierarchichal RE.	Yes	HSROC	I^2		Subgroup
5	Faias, Pereira, Luis, Chaves, et al., 2019	31341368	Comprehensive MA 2.0 and MetaDiSc 1.4	RE (DerSimonian-Laird)/FE (Mantel-Haenszel method), heterogeneity based decission	Yes	SROC	Explored in ROC, χ^2, Cochran-Q, I^2	Not mention	Subgroup
6	Faias, Pereira, Luis, Cravo, et al., 2019	31206466	Meta-Disc 1.4, SPSS Statistics 23 (Deeks' test)	RE (DerSimonian-Laird), heterogeneity based decission	Yes	SROC	Explored in ROC, χ^2, I^2		Subgroup
7	Farahani & Baloch, 2019	31301215	Not mention	Univariate and bivariate random-effect model		SROC, HSROC	Forest plot, Cochran-Q, I^2	Spearman correlation Sen-Spec	Metaregression
8	Gurung et al., 2019	31158516	R 3.4.2 mada, Revman 5.3	Bivariate RE		SROC	ROC plots, metaregression	Different threshold was present, summary estimates of sensitivity and specificity are summary estimates belonging to the 'average' threshold used	Metaregression
9	He et al., 201	31574828	Stata 12.0, Revman 5.3	Bivariate RE/FE, heterogeneity based decission	Yes	SROC	I^2, χ^2 and a bivariate box-plot	Spearman rank correlation analysis	Metaregression, Sub-group analysis
10	Issa et al., 2017	28130609	Revman 5.3, SAS 9.3	RE/FE, compared with a paired z-test	Yes	SROC	I^2		Subgroup
11	Lee et al., 2019	31714947	R 3.4.2 meta mada	RE		SROC	I^2	Spearman correlation Sen-Spec	Metaregression, Sub-group analysis
12	J. Li et al., 2019	31250172	Revman 5.3, Stata 15.1	RE/FE, heterogeneity based decission		SROC	χ^2, I^2		Subgroup
13	Li et al., 2020	31939891	Stata 10.0	RE/FE, heterogeneity based decission	Yes, in case of no threshold effect	SROC (when threshold effect observed)	χ^2, I^2. High heterogeneity $p-value < .05$, $I^2 \geqslant .5$	Correlation (Spearman's) between the logit (Sn) & logit(1 - Sp). When observed plot SROC	Subgroup
14	Shen et al., 2019	31437232	Stata 14.0 midas	bivariate random-effects models		SROC	I^2		Metaregression, Sub-group analysis
15	Tsou et al., 2019	31182360	Stata 12.0, R	bivariate random-effects models	Yes	HSROC	I^2		Metaregression, Sub-group analysis
16	Wang2019	31335700	Stata 14	SROC		SROC	I^2	If heterogeneity then Spearman correlation coefficient	No metaregression (limited number studies)
17	Wei et al., 2020	31939890	Stata 10.0	RE/FE, heterogeneity based decission		SROC (when threshold effect observed)	χ^2, I^2	If heterogeneity then Spearman correlation coefficient > threslhold effect observed	No
18	Xu et al., 2019	31702649	Stata 14.0	bivariate mixed-effects model. Mosesâ€“Littenberg SROC	Yes	SROC	Cochran-Q, I^2		Subgroup, Metaregression
19	Yoon et al., 2019	31470804	R 3.5.2	Bivariate RE/FE and HSROC model, heterogeneity and spearmen based decission (respectvely)		HSROC	χ^2	Different cut off. Spearman correlation coefficient then threslhold effect observed	Subgroup
20	Zheng et al., 2019	31441886	Stata 12, SPSSS 17.0, Metadisc 1.4	RE/FE, heterogeneity based decission	Yes	SROC	Cochran-Q, I^2	Spearman correlation Sen-Spec	Metaregression, Sub-group analysis
21	Zhu et al., 2020	32011436	Stata 14.2, Meta-Disc 1.4	RE	Yes	SROC	I^2	Spearman correlation Sen-Spec. If $I^2 > .5$ a non-threshold effect would exist.	Subgroup

Table 8: Frequentist normal-normal (mixed) model

Study ID	Meta analysis	logit(Sens)	logit(Spec)	Var(logit(Sens))	Var(logit(Spec))	SD(logit(Sens))	SD(logit(Spec))	Covariance	Correlation
Study 1	1	NA	NA	NA	NA	NA	NA	NA	NA
Study 2	2	-0.09895	1.87637	0.9827127	0.35487	0.99132	0.5957	-0.74968	-1.00000
Study 3	3	2.30120	1.88536	0.0000000	0.19351	0.00000	0.4399	0.03043	0.00000
Study 4	4	1.37119	0.57389	1.3038799	0.34959	1.14188	0.5913	0.64111	0.94959
Study 5	5	NA	NA	NA	NA	NA	NA	NA	NA
Study 6	6	-0.30479	2.01946	0.2210948	0.25062	0.47021	0.5006	0.39423	1.00000
Study 7	7	0.88383	4.43853	1.0352880	1.11609	1.01749	1.0565	-0.04722	-0.04393
Study 8	8	NA	NA	NA	NA	NA	NA	NA	NA
Study 9	9	0.99040	1.67943	0.3089410	0.56115	0.55582	0.7491	-0.29551	-0.70973
Study 10	10	1.43189	1.94892	1.1718241	0.90357	1.08251	0.9506	-0.83057	-0.80717
Study 11	11	2.21648	2.91461	0.1239627	1.71024	0.35208	1.3078	0.18175	0.39472
Study 12	12	1.82151	2.51444	0.0049754	0.38425	0.07054	0.6199	-0.17909	-1.00000
Study 13	13	2.10887	1.23615	0.4445871	0.97343	0.66677	0.9866	-0.25269	-0.38411
Study 14	14	1.38135	3.47554	0.1309960	0.26085	0.36193	0.5107	0.03019	0.16334
Study 15	15	NA	NA	NA	NA	NA	NA	NA	NA
Study 16	16	NA	NA	NA	NA	NA	NA	NA	NA
Study 17	17	1.09813	0.60801	0.7586781	1.13607	0.87102	1.0659	-0.28996	-0.31233
Study 18	18	2.45420	1.78126	0.7859012	1.15382	0.88651	1.0742	-0.28700	-0.30139
Study 19	19	0.05111	1.43843	0.8852712	1.20179	0.94089	1.0963	-1.04924	-1.00000
Study 20	20	2.58451	2.53921	0.0000000	1.13922	0.00000	1.0673	0.33773	0.00000
Study 21	21	0.98592	1.63929	0.0000000	0.34010	0.00000	0.5832	-0.09161	0.00000
Study 22	8	0.94752	2.25541	0.6093005	3.80142	0.78058	1.9497	-1.01678	-0.66809
Study 23	1	1.47625	1.34679	0.0006538	0.00000	0.02557	0.0000	-0.22854	0.00000
Study 24	8	1.11989	2.89332	0.4939542	0.10382	0.70282	0.3222	0.04860	0.21460
Study 25	10	1.01685	1.93373	0.2648039	1.16930	0.51459	1.0813	-0.21016	-0.37768
Study 26	12	1.05104	2.02824	0.4061422	0.30500	0.63729	0.5523	-0.16743	-0.47572
Study 27	12	NA	NA	NA	NA	NA	NA	NA	NA
Study 28	12	NA	NA	NA	NA	NA	NA	NA	NA
Study 29	12	NA	NA	NA	NA	NA	NA	NA	NA
Study 30	21	0.89027	2.22913	1.7144725	0.00000	1.30938	0.0000	0.03549	0.00000
Study 31	16	1.55337	1.49220	1.1649911	1.79939	1.07935	1.3414	-0.12540	-0.08661
Study 32	8	0.57793	2.36351	1.3120555	1.66494	1.14545	1.2903	-0.35631	-0.24108
Study 33	10	1.46985	2.44476	0.1710566	0.87787	0.41359	0.9369	-0.27970	-0.72177
Study 34	21	NA	NA	NA	NA	NA	NA	NA	NA
Study 35	18	1.36696	2.42438	0.5028210	3.00472	0.70910	1.7334	-0.59248	-0.48202
Study 36	18	3.15736	2.21774	0.4027007	0.78745	0.63459	0.8874	0.09491	0.16855
Study 37	5	NA	NA	NA	NA	NA	NA	NA	NA
Study 38	5	NA	NA	NA	NA	NA	NA	NA	NA
Study 39	8	0.75352	1.89659	0.3009241	2.93219	0.54857	1.7124	-0.76688	-0.81640
Study 40	8	-0.75688	-0.05688	1.0626781	3.01298	1.03086	1.7358	-0.32420	-0.18118
Study 41	6	-0.17942	0.47230	0.3659786	0.00000	0.60496	0.0000	0.07374	0.00000
Study 42	10	1.08157	2.58063	0.4222193	1.37586	0.64978	1.1730	0.45730	0.59999
Study 43	12	1.25108	2.48155	0.2958170	2.42673	0.54389	1.5578	-0.18083	-0.21343
Study 44	12	NA	NA	NA	NA	NA	NA	NA	NA
Study 45	12	NA	NA	NA	NA	NA	NA	NA	NA
Study 46	17	1.08053	1.19142	0.5455933	1.52848	0.73864	1.2363	-0.42573	-0.46620
Study 47	12	0.78280	2.23527	1.0082616	0.54872	1.00412	0.7408	-0.59237	-0.79640
Study 48	19	NA	NA	NA	NA	NA	NA	NA	NA
Study 49	19	2.65236	0.91756	0.9844540	2.34992	0.99220	1.5329	1.79922	1.00000
Study 50	8	1.31495	2.20825	0.7720253	2.62943	0.87865	1.6216	0.34924	0.24512
Study 51	1	-0.36252	1.52221	2.4528422	0.04612	1.56616	0.2147	0.09466	0.28146
Study 52	1	NA	NA	NA	NA	NA	NA	NA	NA
Study 53	8	1.11368	-1.94983	2.6367844	7.40359	1.62382	2.7210	-4.20439	-0.95158
Study 54	10	0.45976	2.88906	0.1847429	2.69304	0.42982	1.6410	0.39013	0.55311
Study 55	4	1.50231	1.99276	1.4949757	0.45421	1.22269	0.6740	-0.23032	-0.27950

Table 9: Frequentist binomial-normal (Glimmix) model

Study ID	Meta analysis	logit(Sens)	logit(Spec)	Var(logit(Sens))	Var(logit(Spec))	SD(logit(Sens))	SD(logit(Spec))	Covariance	Correlation
study 1	1	1.15569	1.0311	0.81616	2.635e-01	0.9034	0.5133	-2.012e-01	-4.339e-01
study 2	2	0.14993	2.0631	0.38760	1.571e-01	0.6226	0.3964	-9.866e-02	-3.998e-01
study 3	3	2.50140	1.9394	0.03107	2.642e-02	0.1763	0.1625	-3.565e-03	-1.244e-01
study 4	4	1.50438	0.5568	0.50474	8.367e-02	0.7105	0.2893	1.544e-01	7.515e-01
study 5	5	2.39342	2.4243	0.33887	4.822e+00	0.5821	2.1958	-4.585e-01	-3.587e-01
study 6	6	-0.27185	5.3885	0.04889	5.645e+00	0.2211	2.3759	1.555e-01	2.961e-01
study 7	7	0.99763	7.6505	0.02102	6.614e-01	0.1450	0.8133	-1.208e-02	-1.025e-01
study 8	8	-0.67565	1.8273	0.02415	1.012e+00	0.1554	1.0059	8.334e-02	5.332e-01
study 9	9	1.01505	1.8435	0.01888	4.466e-02	0.1374	0.2113	-1.244e-02	-4.285e-01
study 10	10	1.54338	2.2828	0.10482	1.325e-01	0.3238	0.3640	-5.829e-02	-4.946e-01
study 11	11	2.52672	3.6035	0.04977	3.018e-01	0.2231	0.5494	2.021e-03	1.649e-02
study 12	12	1.87102	2.6214	0.02454	1.258e-01	0.1567	0.3546	-1.114e-02	-2.005e-01
study 13	13	2.52524	1.7878	0.07136	2.011e-01	0.2671	0.4484	-1.370e-02	-1.144e-01
study 14	14	1.44820	4.8758	0.01399	4.048e-01	0.1183	0.6362	-1.184e-02	-1.574e-01
study 15	15	3.67146	3.7944	0.04529	2.171e-01	0.2128	0.4659	5.151e-02	5.195e-01
study 16	16	-0.57366	1.7743	0.16302	3.412e-01	0.4038	0.5841	-1.937e-01	-8.215e-01
study 17	17	1.45283	0.7200	0.14226	1.202e-01	0.3772	0.3467	-1.904e-02	-1.456e-01
study 18	18	2.76027	2.0427	0.08498	1.877e-01	0.2915	0.4333	-4.684e-02	-3.708e-01
study 19	19	0.07713	1.4798	0.30279	3.560e-01	0.5503	0.5967	-2.516e-01	-7.663e-01
study 20	20	2.77341	2.8011	0.03388	1.591e-01	0.1841	0.3989	2.769e-02	3.771e-01
study 21	21	1.01794	1.8682	0.01762	1.375e-01	0.1327	0.3709	-4.733e-03	-9.615e-02
study 22	8	0.95562	2.6155	0.10864	1.069e+00	0.3296	1.0338	-1.825e-01	-5.356e-01
study 23	1	1.62526	1.2117	0.09984	7.646e-02	0.3160	0.2765	-1.040e-05	-1.190e-04
study 24	8	1.17171	4.9533	0.03634	1.438e+00	0.1906	1.1991	-2.555e-02	-1.118e-01
study 25	10	1.13102	2.3286	0.06077	2.665e-01	0.2465	0.5162	-2.841e-02	-2.232e-01
study 26	12	1.07551	2.1762	0.06621	1.003e-01	0.2573	0.3167	-2.281e-02	-2.799e-01
study 27	12	0.22693	2.2287	0.03268	1.583e-01	0.1808	0.3978	-2.050e-08	-2.851e-07
study 28	12	-0.15678	2.6217	0.10103	2.762e-01	0.3179	0.5256	-9.650e-02	-5.777e-01
study 29	12	-0.05548	102.4907	1.69682	7.220e+06	1.3026	2687.0223	-1.515e+03	-4.329e-01
study 30	21	0.97974	2.3762	0.28405	5.358e-02	0.5330	0.2315	-1.256e-02	-1.018e-01
study 31	16	1.68940	1.5314	0.23096	2.518e-01	0.4806	0.5018	-1.831e-02	-7.593e-02
study 32	8	0.56703	2.6225	0.05150	6.913e-02	0.2269	0.2629	-1.500e-02	-2.513e-01
study 33	10	1.53815	2.7859	0.03745	1.936e-01	0.1935	0.4400	-1.971e-02	-2.315e-01
study 34	21	3.20320	3.9254	1.02353	2.329e+00	1.0117	1.5262	1.101e+00	7.129e-01
study 35	18	1.47016	3.3028	0.04446	6.336e-01	0.2109	0.7960	-6.532e-02	-3.891e-01
study 36	18	3.42395	2.6125	0.10702	1.804e-01	0.3271	0.4248	2.686e-02	1.933e-01
study 37	5	1.11044	0.9694	0.04748	1.254e-01	0.2179	0.3541	0.000e+00	0.000e+00
study 38	5	0.28769	1.9459	0.08333	8.791e-02	0.2887	0.2965	0.000e+00	0.000e+00
study 39	8	0.96062	2.0489	0.08374	3.687e-01	0.2894	0.6072	-5.262e-02	-2.994e-01
study 40	8	-0.76353	0.3334	0.20947	1.298e+00	0.4577	1.1394	-9.196e-04	-1.764e-03
study 41	6	-0.20779	0.5269	0.09238	1.987e-02	0.3039	0.1410	1.201e-02	2.802e-01
study 42	10	1.25869	3.1376	0.06068	2.300e-01	0.2463	0.4796	4.649e-02	3.935e-01
study 43	12	1.27205	2.9978	0.09645	1.118e+00	0.3106	1.0571	-4.652e-03	-1.417e-02
study 44	12	-0.12757	26.0039	8.78963	0.000e+00	2.9647	0.0000	0.000e+00	0.000e+00
study 45	12	1.62076	2.6962	0.07188	1.806e+00	0.2681	1.3440	1.258e-01	3.492e-01
study 46	17	1.21209	1.2288	0.09336	1.280e-01	0.3055	0.3578	-4.923e-02	-4.503e-01
study 47	12	0.80290	2.3865	0.12847	1.253e-01	0.3584	0.3539	-6.799e-02	-5.360e-01
study 48	19	1.89656	1.1848	0.70558	4.373e-02	0.8400	0.2091	3.207e-02	1.826e-01
study 49	19	3.50721	0.9560	1.51310	5.460e-01	1.2301	0.7389	6.428e-01	7.072e-01
study 50	8	1.35677	3.9068	0.15587	3.416e+00	0.3948	1.8483	4.190e-02	5.742e-02
study 51	1	-0.42375	1.7683	0.57271	1.132e-01	0.7568	0.3364	-3.616e-02	-1.420e-01
study 52	1	1.70799	1.4832	0.50752	2.942e-01	0.7124	0.5424	1.366e-01	3.536e-01
study 53	8	1.59568	-3.0788	1.45715	5.405e+00	1.2071	2.3248	-2.695e+00	-9.603e-01
study 54	10	0.70015	3.7538	0.09120	7.212e-01	0.3020	0.8492	4.023e-02	1.568e-01
study 55	4	1.96724	2.0204	0.54270	6.701e-02	0.7367	0.2589	-5.587e-02	-2.930e-01

Table 10: Frequentist binomial-normal (Nlmixed) model

Study ID	Meta analysis	logit(Sens)	logit(Spec)	Var(logit(Sens))	Var(logit(Spec))	SD(logit(Sens))	SD(logit(Spec))	Covariance	Correlation
Study 1	1	0.94118	0.79995	0.743609	1.753e-01	0.86233	0.4186601	-3.610e-01	1.0000
Study 2	2	-0.15023	1.94595	0.513053	7.177e-02	0.71628	0.2678982	-9.950e-02	-1.0000
Study 3	3	2.48018	1.91781	0.061213	6.090e-02	0.24741	0.2467806	6.106e-02	NA
Study 4	4	1.48854	0.08029	0.231002	1.864e-01	0.48063	0.4316940	2.437e-02	NA
Study 5	5	2.75414	2.13302	0.143077	4.836e+00	0.37826	2.1990216	-5.077e-01	NA
Study 6	6	-0.66383	5.31776	0.143441	5.653e+00	0.37874	2.3776133	3.375e-01	NA
Study 7	7	0.77542	7.67959	0.571069	3.917e-01	0.75569	0.6258620	5.528e-04	NA
Study 8	8	-0.75135	1.76218	0.001999	9.948e-01	0.04471	0.9973742	4.460e-02	-1.0000
Study 9	9	0.95985	1.79076	0.071063	8.096e-02	0.26658	0.2845294	7.585e-02	NA
Study 10	10	1.19902	2.15254	0.431272	4.469e-02	0.65671	0.2113958	8.376e-02	-1.0000
Study 11	11	2.29137	3.58096	0.760432	4.162e-02	0.87203	0.2040167	-1.339e-01	NA
Study 12	12	1.67943	2.50491	0.100116	2.110e-02	0.31641	0.1452432	-4.596e-02	1.0000
Study 13	13	2.40544	1.57846	0.025081	3.914e-01	0.15837	0.6256110	9.908e-02	-1.0000
Study 14	14	1.35273	4.96727	0.469887	9.282e-02	0.68548	0.3046653	-7.430e-02	-1.0000
Study 15	15	3.63741	3.75986	0.102516	2.768e-01	0.32018	0.5261087	1.684e-01	NA
Study 16	16	-0.57223	1.77475	0.240090	4.112e-01	0.48999	0.6412405	-4.508e-02	0.6995
Study 17	17	1.27819	0.28552	0.021412	3.504e-01	0.14633	0.5919744	2.133e-02	1.0000
Study 18	18	2.64104	1.78164	0.015429	5.288e-01	0.12422	0.7272097	8.257e-02	0.6443
Study 19	19	-0.05998	1.36323	0.279094	3.284e-01	0.52829	0.5730250	-3.027e-01	1.0000
Study 20	20	2.63819	2.71833	0.223723	8.726e-02	0.47299	0.2954006	1.397e-01	-1.0000
Study 21	21	0.72720	1.70583	0.101838	1.155e-02	0.31912	0.1074511	-3.429e-02	1.0000
Study 22	8	0.56525	2.57397	0.174820	8.292e-01	0.41812	0.9106058	-2.376e-02	-1.0000
Study 23	1	1.44410	0.98663	0.044342	7.382e-02	0.21058	0.2717060	-5.721e-02	-1.0000
Study 24	8	0.86741	5.04937	0.362552	1.266e+00	0.60212	1.1250179	1.534e-01	-1.0000
Study 25	10	0.86751	2.27443	0.294077	1.240e-02	0.54229	0.1113520	-6.038e-02	NA
Study 26	12	0.77945	2.01507	0.168804	6.407e-02	0.41086	0.2531258	2.909e-02	-1.0000
Study 27	12	-0.04895	2.08373	0.075622	5.776e-02	0.27499	0.2403277	-6.609e-02	1.0000
Study 28	12	-0.34326	2.55925	0.281515	8.542e-02	0.53058	0.2922695	-1.551e-01	NA
Study 29	12	-0.05548	1.00000	1.696820	1.511e-08	1.30262	0.0001229	-4.128e-06	NA
Study 30	21	0.68079	2.23388	0.514944	1.377e-02	0.71760	0.1173312	7.314e-02	-1.0000
Study 31	16	1.45278	0.88174	0.136182	7.459e-01	0.36903	0.8636314	1.638e-01	NA
Study 32	8	0.44462	2.53903	0.196347	6.238e-02	0.44311	0.2497612	1.107e-01	NA
Study 33	10	1.29621	2.70700	0.313446	7.179e-04	0.55986	0.0267935	-8.941e-03	0.6879
Study 34	21	3.04135	3.44825	0.817178	2.969e+00	0.90398	1.7231618	9.672e-01	NA
Study 35	18	1.20286	3.31858	0.253660	2.145e-01	0.50365	0.4631090	1.321e-01	NA
Study 36	18	3.35675	2.49347	0.032197	3.379e-01	0.17943	0.5813301	1.043e-01	-0.9998
Study 37	5	0.88934	0.73916	0.021915	1.034e-01	0.14804	0.3215469	-4.760e-02	1.0000
Study 38	5	0.12316	1.82283	0.067686	4.713e-02	0.26017	0.2170974	-5.648e-02	NA
Study 39	8	0.48077	2.02855	0.753024	1.134e-01	0.86777	0.3366787	1.504e-03	NA
Study 40	8	-0.80137	0.30287	0.054166	1.195e+00	0.23274	1.0931403	-2.544e-01	NA
Study 41	6	-0.21260	0.51368	0.054400	4.201e-02	0.23324	0.2049721	-4.906e-03	NA
Study 42	10	0.95835	3.08935	0.619010	7.969e-02	0.78677	0.2823025	4.247e-02	NA
Study 43	12	0.73909	3.19180	0.122591	9.937e-01	0.35013	0.9968634	2.459e-01	-1.0000
Study 44	12	-0.21453	1.08573	10.015898	1.534e-01	3.16479	0.3916131	-5.002e-06	0.8371
Study 45	12	1.50688	2.64582	0.049641	1.740e+00	0.22280	1.3191130	6.664e-02	NA
Study 46	17	0.66169	0.83685	0.319584	1.365e-01	0.56532	0.3694577	1.699e-01	1.0000
Study 47	12	0.50383	2.27859	0.347509	5.171e-03	0.58950	0.0719127	-4.031e-02	NA
Study 48	19	1.79719	0.95291	0.539566	2.867e-02	0.73455	0.1693169	-1.063e-01	1.0000
Study 49	19	3.49779	0.97924	1.384970	9.690e-01	1.17685	0.9843703	5.263e-01	NA
Study 50	8	0.66227	3.68759	0.553592	3.451e+00	0.74404	1.8578079	4.737e-01	-1.0000
Study 51	1	-0.63389	1.63368	0.522579	9.059e-02	0.72290	0.3009851	-9.873e-02	-1.0000
Study 52	1	1.85917	1.12753	0.628872	1.641e-01	0.79301	0.4050879	2.855e-01	NA
Study 53	8	1.67635	-3.02828	1.454199	5.369e+00	1.20590	2.3170308	-2.755e+00	NA
Study 54	10	0.19143	3.83817	0.511833	4.791e-01	0.71542	0.6921930	4.297e-01	NA
Study 55	4	1.75885	1.88087	0.621256	1.537e-02	0.78820	0.1239705	-4.321e-02	NA

Table 11: Bayesian binomial-normal model (bamdit)

Study ID	Meta analysis	logit(Sens)	logit(Spec)	SD(logit(Sens))	SD(logit(Spec))	Covariance	Correlation
study 1	1	0.81194	0.8962	2.0649	1.5807	-0.5739291	-0.175835
study 2	2	0.28406	1.9054	2.4220	1.3505	-0.6056690	-0.185176
study 3	3	2.55662	1.9173	0.3969	0.5278	-0.0326169	-0.155712
study 4	4	0.96000	0.4735	1.7548	1.0881	0.8568490	0.448724
study 5	5	1.92709	0.9984	1.4335	3.5064	-0.4167665	-0.082913
study 6	6	-0.26744	4.1732	0.7592	2.3108	0.3557108	0.202756
study 7	7	1.00235	7.4184	1.2097	2.8665	-0.8007165	-0.230907
study 8	8	-0.68910	1.1459	0.6554	2.9720	0.5821037	0.298838
study 9	9	1.01650	1.8325	0.6372	0.9868	-0.2837360	-0.451235
study 10	10	1.54137	2.2637	1.3172	1.4265	-1.0009027	-0.532665
study 11	11	2.49135	3.3708	0.6052	2.2896	0.0601805	0.043428
study 12	12	1.84124	2.4977	0.3593	0.8757	-0.0431892	-0.137255
study 13	13	2.49720	1.6598	1.0546	1.7141	-0.2633660	-0.145691
study 14	14	1.45881	4.6781	0.4844	1.7465	-0.2289519	-0.270652
study 15	15	3.66515	3.5571	0.9524	2.3558	1.6750311	0.746538
study 16	16	-0.39078	1.3226	1.3658	1.9780	-1.6489123	-0.610357
study 17	17	1.39218	0.6872	1.6154	1.4971	-0.3436867	-0.142105
study 18	18	2.74711	1.9475	1.3229	2.1427	-1.2425276	-0.438343
study 19	19	0.15588	1.0375	1.7962	2.0932	-1.4228860	-0.378444
study 20	20	2.79013	2.5966	0.6716	1.5332	0.4050590	0.393385
study 21	21	1.00855	1.7758	0.1992	1.2261	-0.0197238	-0.080761
study 22	8	0.96538	1.7203	1.1841	3.1879	-1.2553125	-0.332551
study 23	1	1.68961	1.0664	0.9303	1.5568	-0.1608150	-0.111036
study 24	8	1.16831	4.6311	0.8278	1.8832	-0.3497655	-0.224357
study 25	10	1.15232	2.1142	0.8029	1.9534	-0.3816942	-0.243369
study 26	12	1.03605	2.0897	0.8980	0.8720	-0.1880809	-0.240201
study 27	12	0.21157	1.7432	0.6997	1.5929	-0.0485653	-0.043576
study 28	12	-0.04236	2.2201	1.1396	1.4779	-0.5805206	-0.344678
study 29	12	-0.03051	0.5504	2.7106	3.1900	-3.0209002	-0.349360
study 30	21	0.82676	2.3273	1.9399	0.4654	-0.0636494	-0.070504
study 31	16	1.46716	1.2631	1.7375	1.9305	-0.1202720	-0.035855
study 32	8	0.56179	2.5895	1.4943	1.6310	-0.6611249	-0.271269
study 33	10	1.51713	2.6221	0.4698	1.5326	-0.1484258	-0.206156
study 34	21	1.69899	2.1290	2.8948	3.0612	6.7659788	0.763514
study 35	18	1.51841	2.8354	0.9323	3.3249	-1.3412839	-0.432713
study 36	18	3.33076	2.3967	1.0027	1.5816	0.4165958	0.262692
study 37	5	0.81313	0.6828	1.2232	1.4071	0.1046300	0.060793
study 38	5	0.09635	1.7379	1.7537	1.4317	-0.2448683	-0.097528
study 39	8	0.97146	1.6885	1.2477	2.5439	-0.3860083	-0.121621
study 40	8	-0.57966	0.2609	1.7917	3.1201	0.0295033	0.005278
study 41	6	-0.22602	0.6035	0.8922	0.7270	-0.0008218	-0.001267
study 42	10	1.18135	2.9346	0.9064	2.0287	0.9166214	0.498507
study 43	12	1.07921	1.7707	1.1629	2.9812	0.2321075	0.066949
study 44	12	0.05409	2.3203	3.8632	2.0854	0.3038309	0.037713
study 45	12	1.41011	1.3013	0.7886	3.1343	0.4696374	0.190001
study 46	17	1.17103	1.1742	1.1467	1.4591	-0.7757035	-0.463618
study 47	12	0.78748	2.3060	1.2889	1.0027	-0.6749339	-0.522249
study 48	19	1.19756	1.0257	2.5506	0.8651	0.2310496	0.104715
study 49	19	2.22470	0.2908	2.9022	2.1689	3.6874582	0.585811
study 50	8	1.10040	2.1103	1.4902	3.3360	0.5390948	0.108441
study 51	1	-0.28123	1.5987	2.3958	1.0303	-0.2071755	-0.083932
study 52	1	1.21808	1.0584	1.6545	1.8891	0.6350852	0.203189
study 53	8	0.66906	-0.9873	2.8895	3.4130	-6.7886917	-0.688375
study 54	10	0.63986	3.0542	1.1304	2.9242	0.5225060	0.158068
study 55	4	1.56250	1.9551	2.6362	1.0267	-0.5840479	-0.215796

Table 12: Bayesian binomial-mixture of normal model (bamdit)

Study ID	Meta analysis	logit(Sens)	logit(Spec)	SD(logit(Sens))	SD(logit(Spec))	Covariance	Correlation
study 1	1	0.79115	0.91150	2.0399	1.5661	-0.513473	-0.160720
study 2	2	0.16814	1.89651	2.2925	1.1881	-0.483763	-0.177609
study 3	3	2.53218	1.91884	0.3499	0.4521	-0.017541	-0.110888
study 4	4	1.00040	0.47399	1.6450	1.0373	0.747354	0.437989
study 5	5	1.97045	0.96414	1.3039	3.3703	-0.395703	-0.090042
study 6	6	-0.25597	4.18676	0.6460	2.2805	0.294051	0.199586
study 7	7	0.84818	7.02959	0.9790	2.0787	-0.418540	-0.205665
study 8	8	-0.72332	0.94597	0.5906	2.6545	0.359641	0.229394
study 9	9	1.03814	1.74063	0.5185	0.8855	-0.224699	-0.489442
study 10	10	1.55315	2.21220	1.0865	1.2565	-0.658539	-0.482378
study 11	11	2.50630	3.27885	0.5518	2.0905	0.065522	0.056799
study 12	12	1.83941	2.51988	0.2808	0.7776	-0.024679	-0.113040
study 13	13	2.46834	1.56594	0.9346	1.4636	-0.169261	-0.123735
study 14	14	1.43444	4.61588	0.4136	1.5361	-0.161920	-0.254875
study 15	15	3.67564	3.52979	0.6888	1.6325	0.773067	0.687507
study 16	16	-0.41817	1.35486	1.3069	1.9234	-1.516513	-0.603301
study 17	17	1.37335	0.55700	1.4335	1.2972	-0.207844	-0.111778
study 18	18	2.58822	2.16025	0.9886	1.2520	-0.305860	-0.247123
study 19	19	0.19108	0.99326	1.6910	1.9423	-1.152186	-0.350815
study 20	20	2.75247	2.44133	0.5665	1.2019	0.244425	0.359003
study 21	21	1.01849	1.73616	0.1769	1.1255	-0.018323	-0.092048
study 22	8	0.89430	1.76730	1.1292	2.9555	-0.948777	-0.284294
study 23	1	1.68131	1.02517	0.7653	1.5859	-0.128953	-0.106245
study 24	8	1.17359	4.47454	0.7158	1.6066	-0.258506	-0.224796
study 25	10	1.10803	2.13075	0.7199	1.7916	-0.266391	-0.206553
study 26	12	1.02245	2.05702	0.8167	0.8842	-0.191262	-0.264858
study 27	12	0.20975	1.78812	0.6019	1.5067	-0.045605	-0.050290
study 28	12	-0.01106	2.18751	1.0646	1.3964	-0.481262	-0.323734
study 29	12	0.06816	0.50136	2.5299	3.1996	-2.812396	-0.347441
study 30	21	0.78468	2.31222	1.7750	0.4557	-0.075328	-0.093124
study 31	16	1.38895	1.34764	1.6230	1.6659	0.010321	0.003817
study 32	8	0.53370	2.65652	1.1298	1.2754	-0.233023	-0.161723
study 33	10	1.51797	2.57428	0.3950	1.3506	-0.089138	-0.167071
study 34	21	1.44607	1.84468	2.5895	2.9287	5.676953	0.748555
study 35	18	1.41554	2.89693	0.7283	2.8751	-0.629202	-0.300470
study 36	18	3.31893	2.35372	0.8610	1.4037	0.282721	0.233935
study 37	5	0.84718	0.72267	1.1838	1.3723	0.070130	0.043172
study 38	5	0.12607	1.70524	1.6570	1.3746	-0.236400	-0.103795
study 39	8	0.73834	1.93707	0.9038	1.9503	-0.181753	-0.103106
study 40	8	-0.58233	0.03352	1.6470	2.9812	-0.187097	-0.038105
study 41	6	-0.24180	0.61210	0.7894	0.7184	-0.005377	-0.009483
study 42	10	1.14916	2.86319	0.8105	1.8332	0.772801	0.520114
study 43	12	1.11358	1.64582	1.0500	2.9034	0.193529	0.063485
study 44	12	-0.07096	2.29114	3.6033	2.2214	0.110114	0.013757
study 45	12	1.39442	1.21540	0.7668	3.0119	0.436512	0.189001
study 46	17	1.17714	1.13045	0.9184	1.2152	-0.462351	-0.414260
study 47	12	0.84902	2.24150	1.1170	0.9544	-0.530283	-0.497434
study 48	19	1.14240	1.04401	2.5016	0.8188	0.216418	0.105652
study 49	19	2.07290	0.23586	2.8209	2.0138	3.145515	0.553715
study 50	8	1.17486	2.06348	1.3427	3.2213	0.474767	0.109768
study 51	1	-0.30143	1.58263	2.2674	0.9086	-0.195747	-0.095017
study 52	1	1.22220	1.05398	1.5739	1.8412	0.587937	0.202894
study 53	8	0.45134	-0.74970	2.4489	3.1459	-4.890807	-0.634830
study 54	10	0.62101	3.10022	0.9630	2.6657	0.365017	0.142186
study 55	4	1.40046	1.90537	2.5451	0.8288	-0.482344	-0.228674

10.3 Agreement estimation analysis

As follow general agreement (figure 18, 19, 20 and 21) and detailed agreement analysis (figure 22, 23, 24, 25, 26, 27, 28, 29, and 30) according to Bland-Altman analysis; additionally correlation matrix plots (figure 31, 32, 33, 34, 35, and 36) and general histograms are presented.

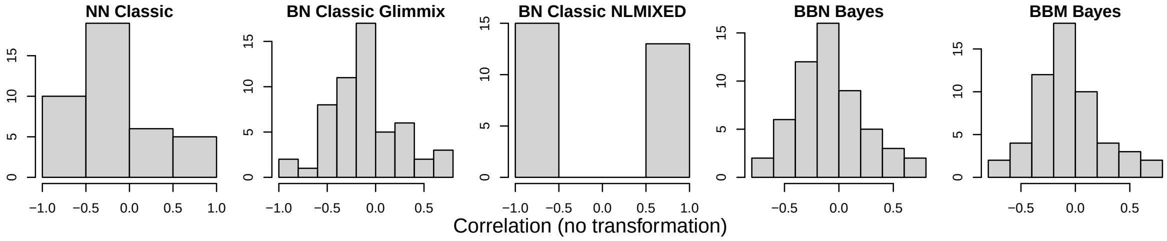


Figure 14: Histograms of correlation values in the three frequentist models

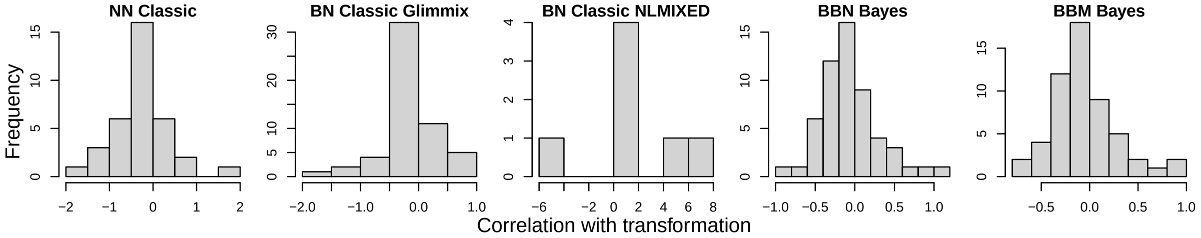


Figure 15: Histograms of correlation values in the three frequentist models

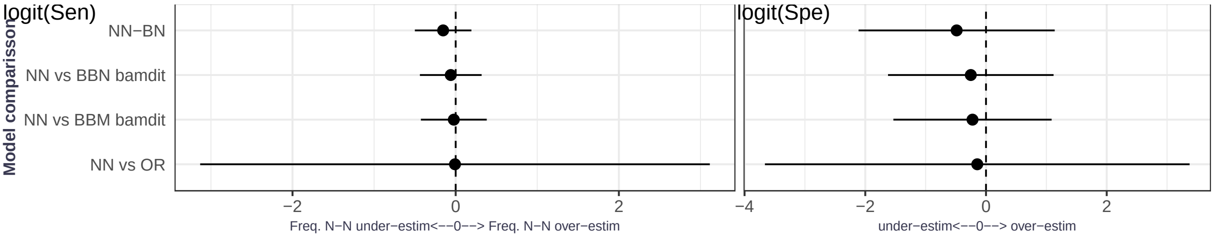


Figure 16: B-A global agreement of frequentist normal-normal model as a reference: logit(sen) and logit (spe)

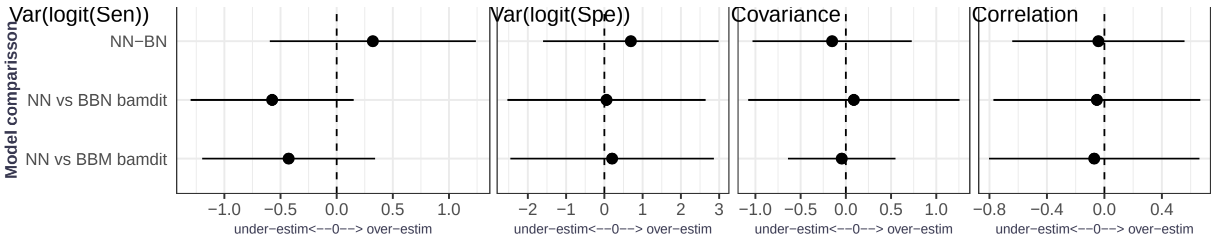


Figure 17: B-A global agreement of frequentist normal-normal model as a reference: covariance parameters

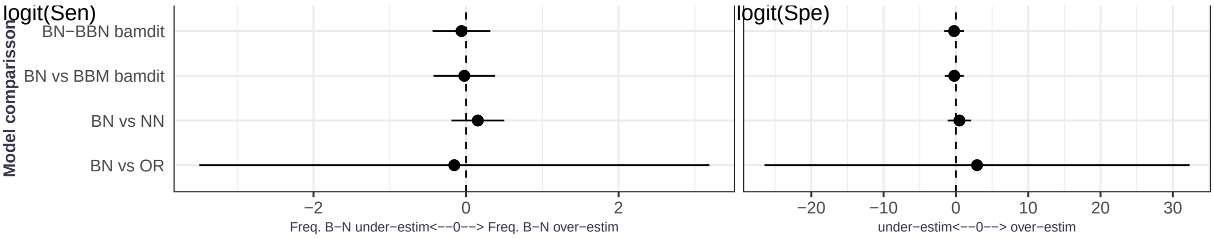


Figure 18: B-A global agreement of frequentist binomial-normal model as a reference: logit(sen) and logit (spe)

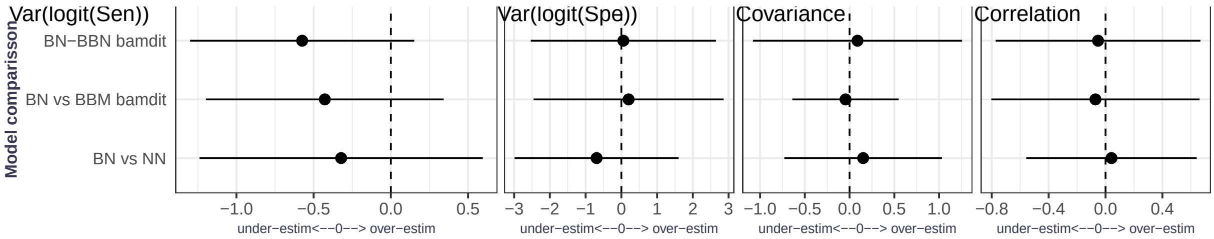


Figure 19: B-A global agreement of frequentist binomial-normal model as a reference: covariance parameters

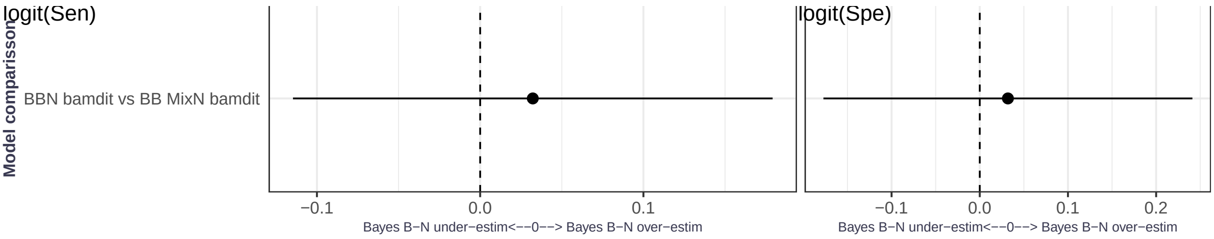


Figure 20: B-A global agreement of Bayesian binomial-normal model as a reference: covariance parameters

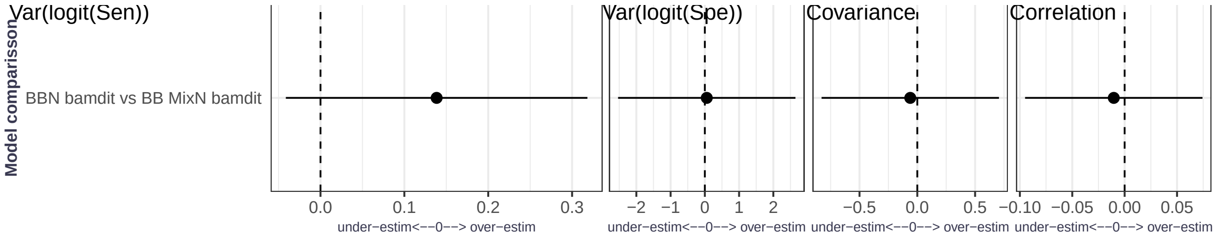


Figure 21: B-A global agreement of Bayesian binomial-normal model as a reference: logit(sen) and logit (spe)

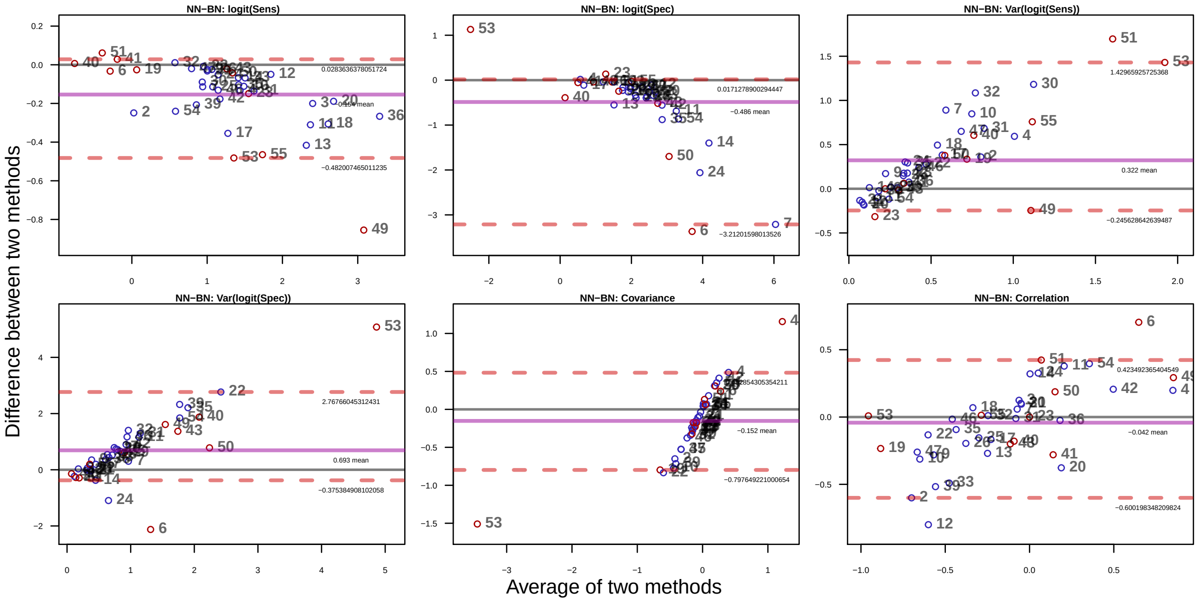


Figure 22: Detailed Bland-Altman agreement between NN (reference model) vs BN: six parameters

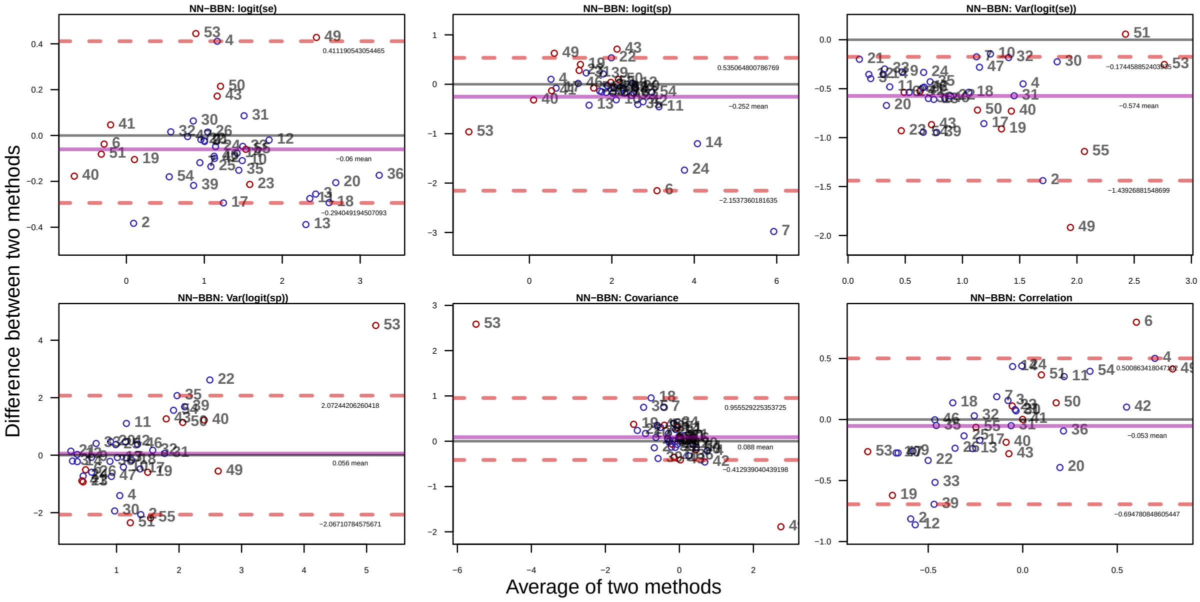


Figure 23: Detailed Bland-Altman agreement between NN (reference model) vs BBN bamdit: six parameters

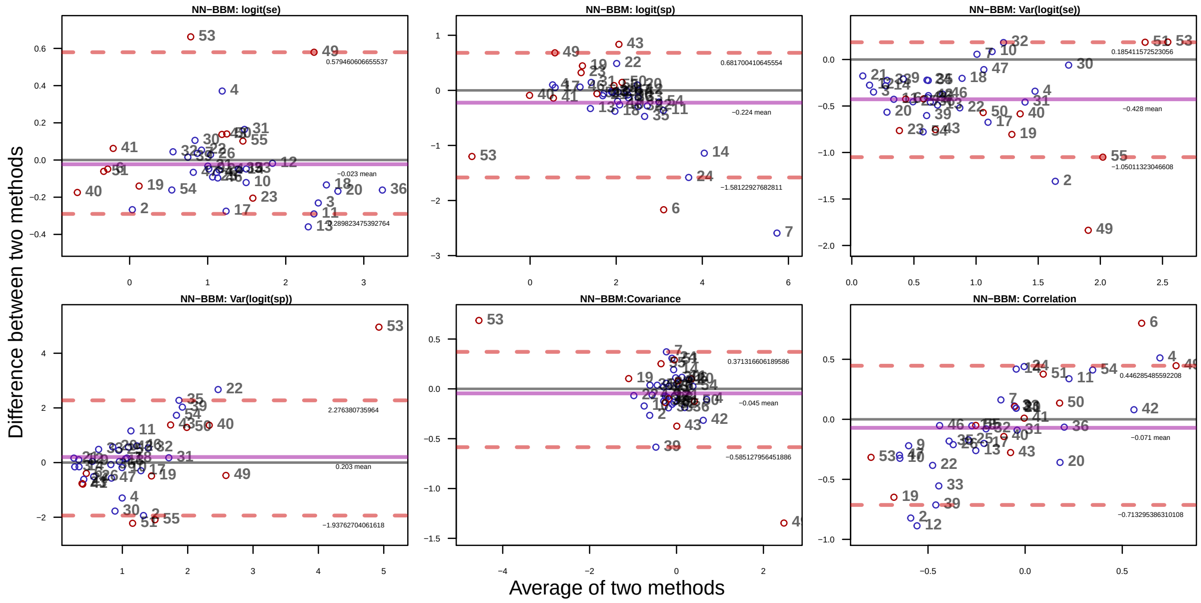


Figure 24: Detailed Bland-Altman agreement between NN (reference model) vs BBN bamdit: six parameters

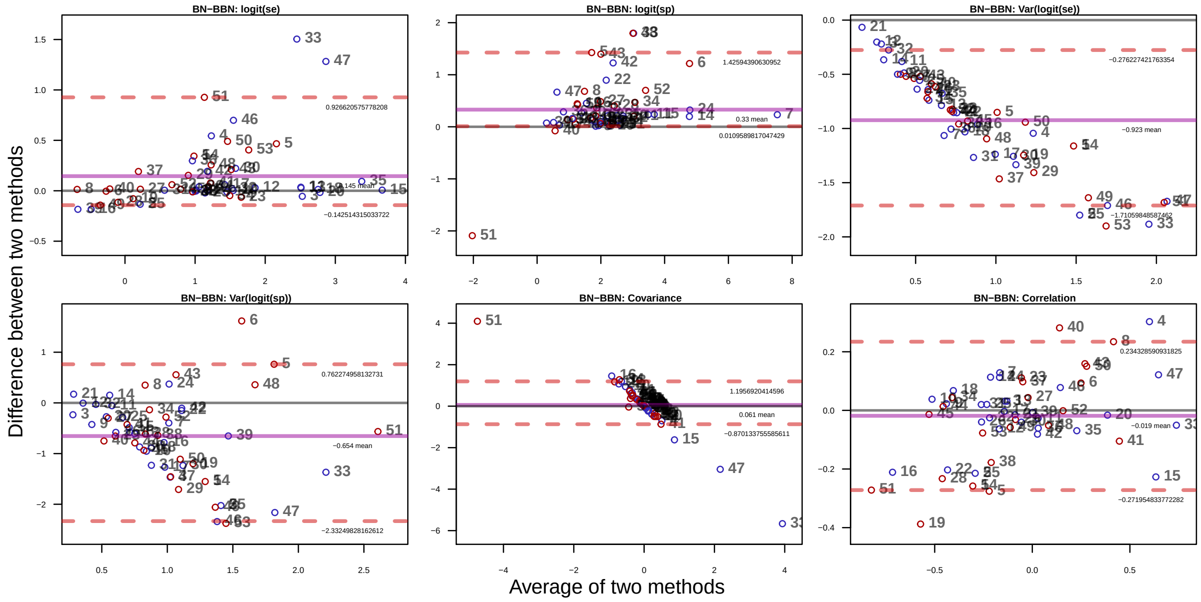


Figure 25: Detailed Bland-Altman agreement between BN Glimmix (reference model) vs BBN bamdit: six parameters. Studies 29 and 44 (extreme numeric outliers) was exclude from the plot, but not from the analysis

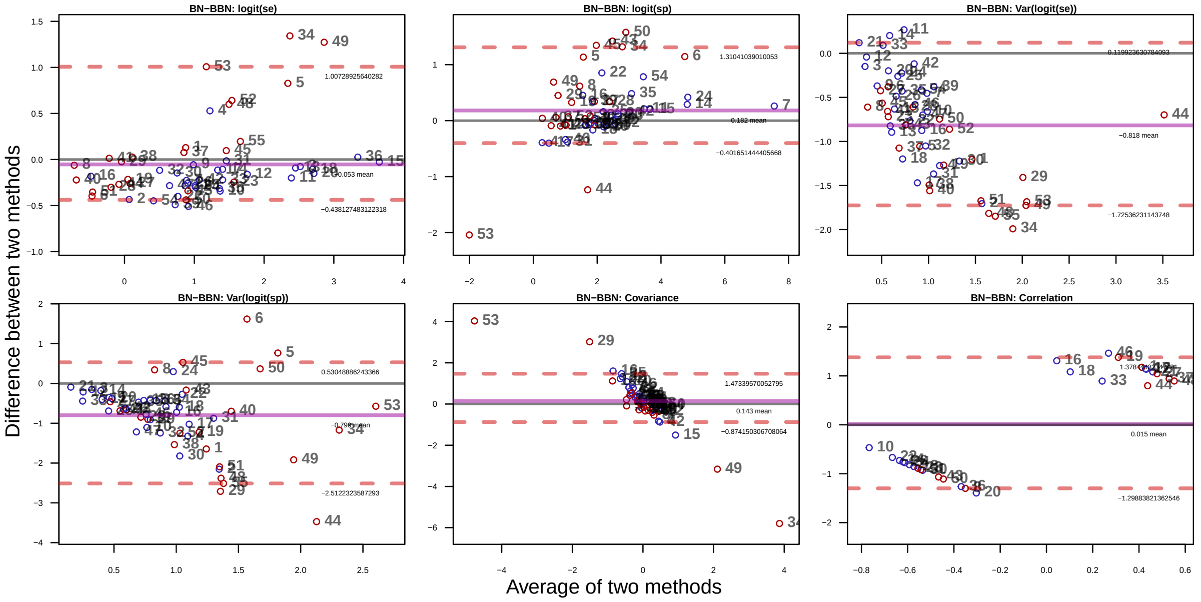


Figure 26: Detailed Bland-Altman agreement between BN Nlmixed (reference model) vs BBN bamdit: six parameters

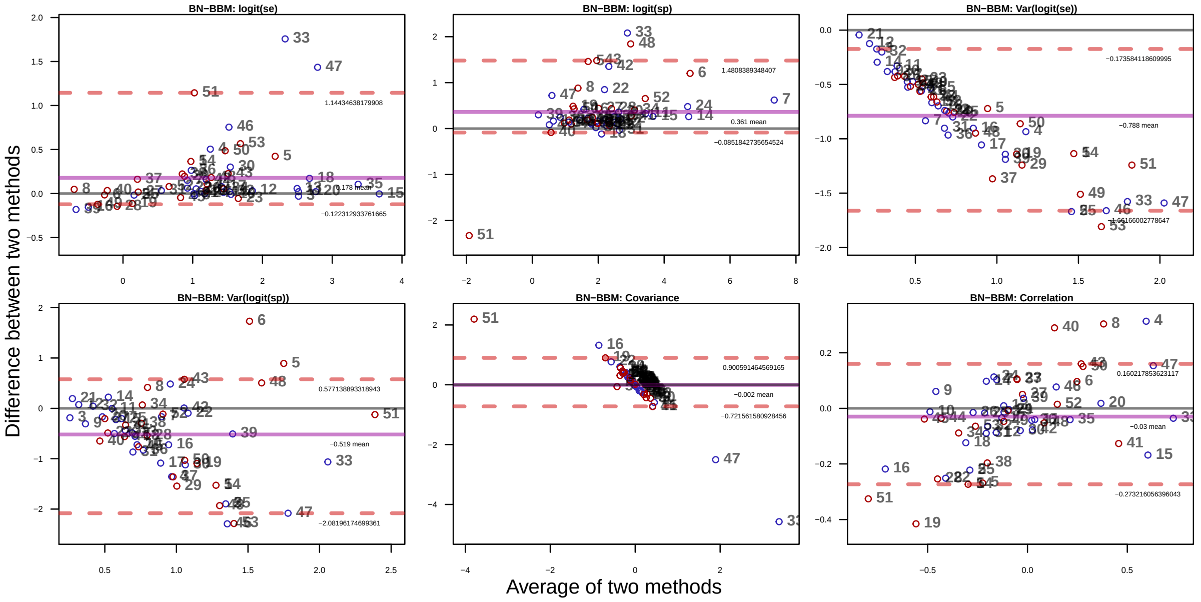


Figure 27: Detailed Bland-Altman agreement between BN Glimmix (reference model) vs BBM bamdit: six parameters. Studies 29 and 44 (extreme numeric outliers) was exclude from the plot, but not from the analysis

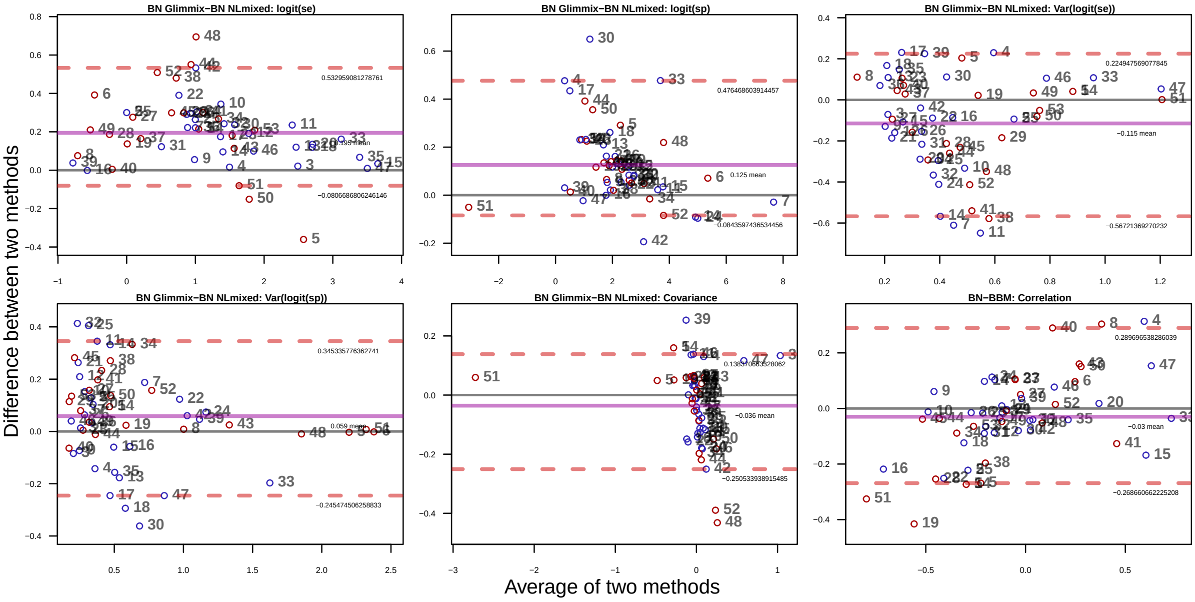


Figure 28: Detailed Bland-Altman agreement between BN Glimmix (reference model) vs BN NLMIXED: six parameters. Studies 29 and 44 (extreme numeric outliers) was exclude from the plot, but not from the analysis

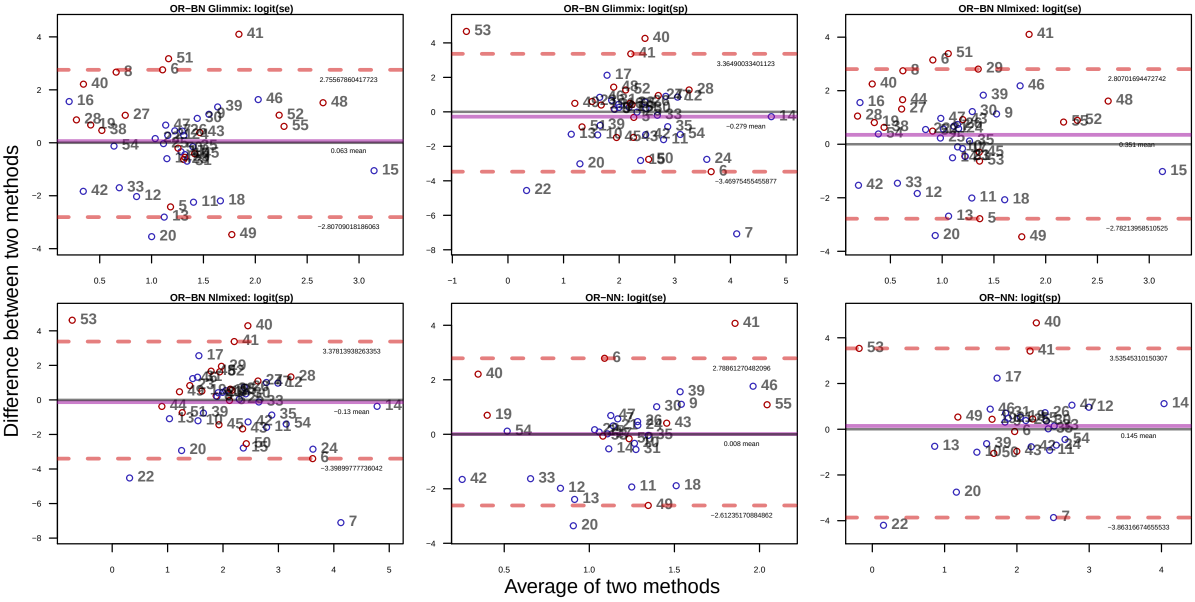


Figure 29: Detailed Bland-Altman agreement between Original estimates vs NN Mixed, BN Glimmix or Nlmixed (reference model) vs: central parameters. Studies 29 and 44 (extreme numeric outliers) was exclude from the plot, but not from the analysis

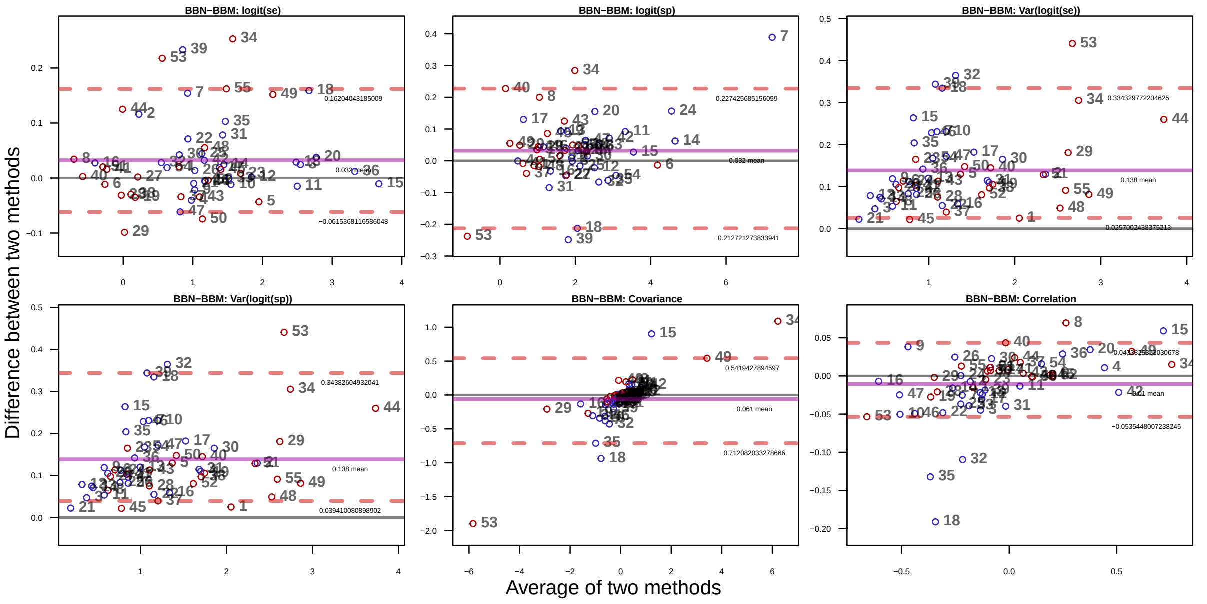


Figure 30: Detailed Bland-Altman agreement between BBN Bandit (reference model) vs BBM: six parameters

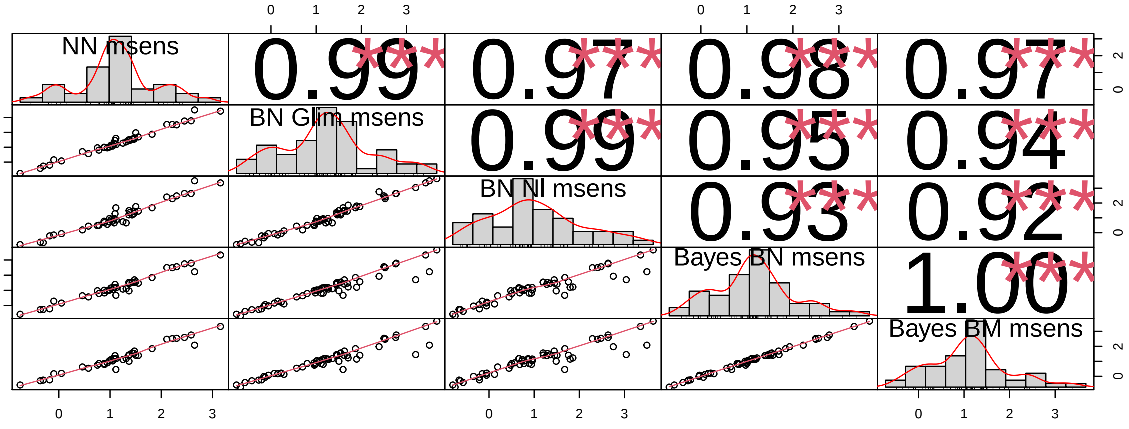


Figure 31: Correlation between models estimates: logit(sens)

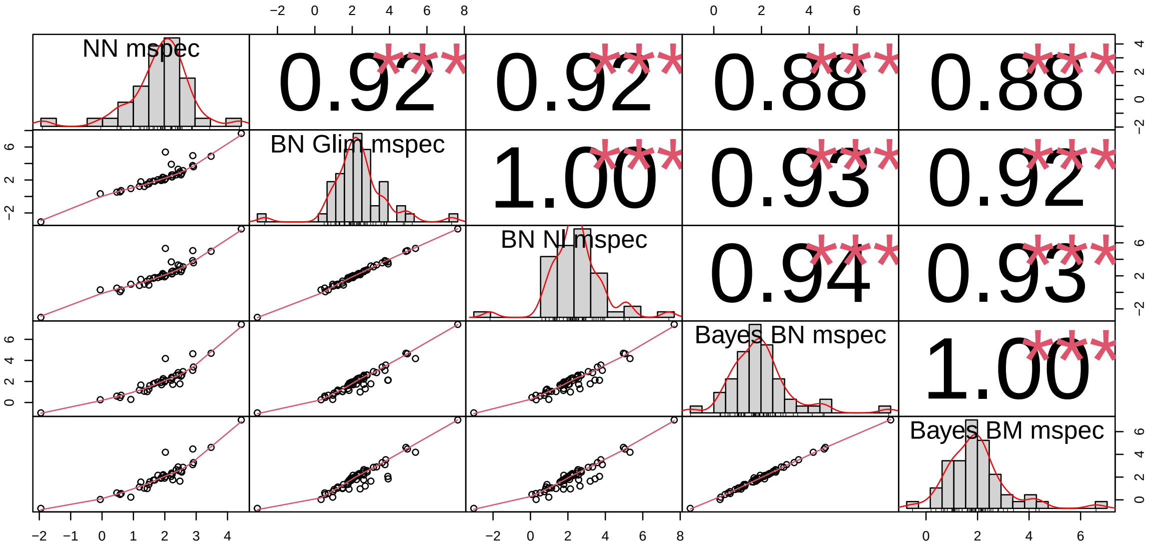


Figure 32: Correlation between models estimates: logit(spec). Studies 29 and 44 (extreme numeric outliers) was exclude from the plot

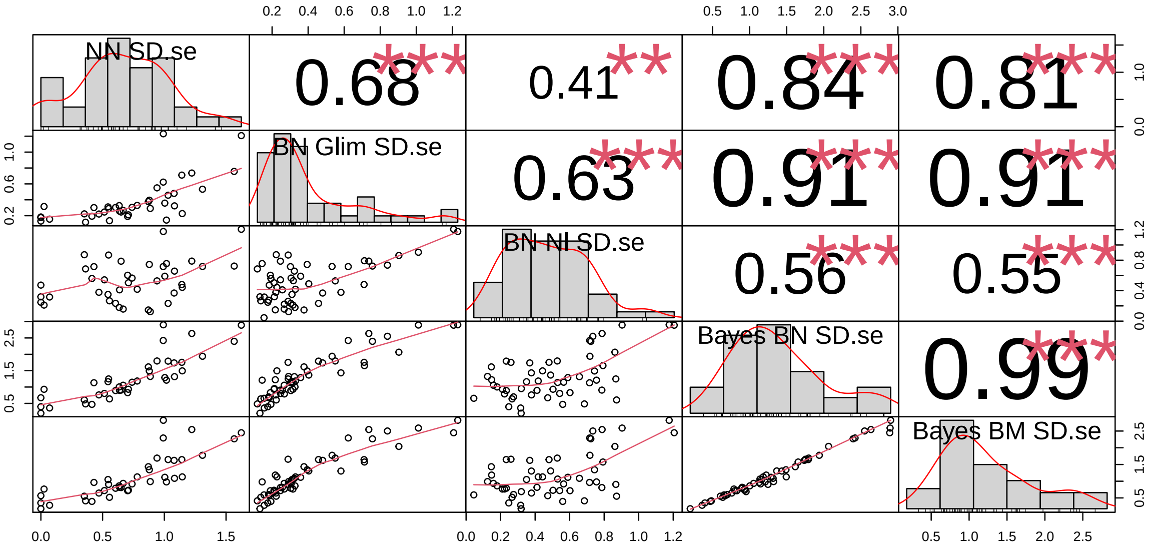


Figure 33: Correlation between models estimates: SD of logit(sens). Studies 29 and 44 (extreme numeric outliers) was exclude from the plot

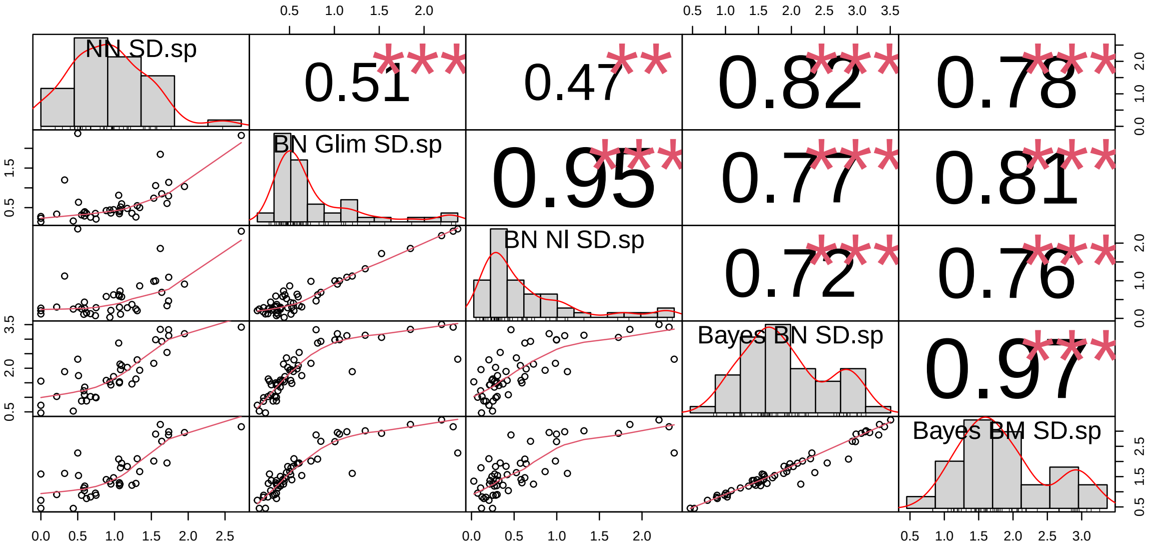


Figure 34: Correlation between models estimates: SD of logit(spec). Studies 29 and 44 (extreme numeric outliers) was exclude from the plot

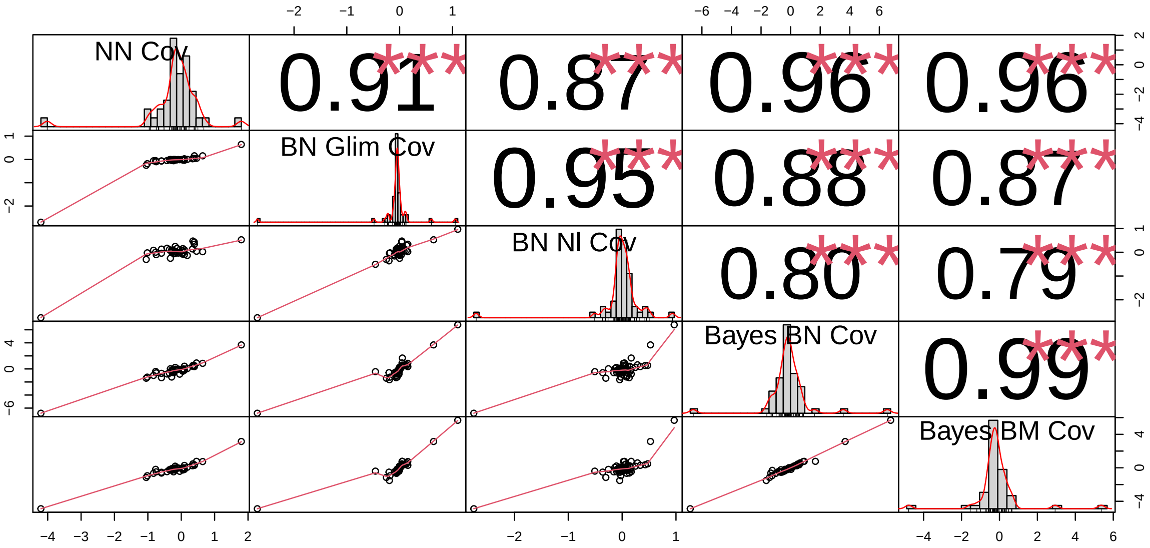


Figure 35: Correlation between models estimates: Covariance. Studies 29 and 44 (extreme numeric outliers) was exclude from the plot

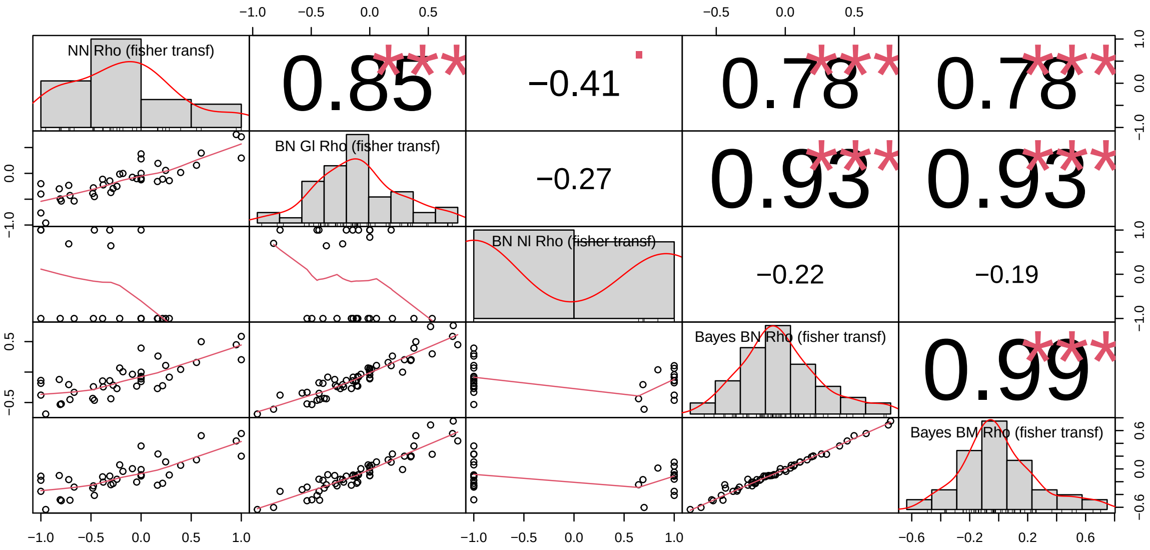


Figure 36: Correlation between models estimates: Rho (fisher transformation)

10.4 Codes

As following the core essential codes are provided.

10.4.1 Classical models

10.4.1.1 Reitsma model

```
data meta ;
  set LAG_data;
  if tp eq 0 or fp eq 0 or fn eq 0 or tn eq 0 then do;
  tp=tp+0.5; fp=fp+0.5; fn=fn+0.5; tn=tn+ 0.5; end;
  sens=tp/(tp +  fn); spec=tn /(tn +  fp);
  log_sens=log(sens/(1-sens)); var_log_sens=1/(sens*(1-sens)*(tp + fn));
  log_spec=log(spec/(1-spec)); var_log_spec=1/(spec*(1-spec)*(tn+  fp));
run;

data meta; rename group=modality studies=study_id;set meta;  run;

data bi_meta;
  set meta;
  dis = 1; non_dis= 0;
  logit = log_sens;
  var_logit = var_log_sens;
  rec + 1;output;
  dis = 0; non_dis= 1;
  logit = log_spec;
  var_logit = var_log_spec;
  rec + 1;
  output;
run;

data c.cov;
  if _n_ eq 1 then do; est  =   0; output; est =0; output; est  =   0; output; end;
  set bi_meta;
  est =    var_logit; output;
  keep est;
run;

data _null_;SET c.cov; call symput(’nCov_par’, _n_);run;
```

10.4.1.1.1 Continuity correction

```
proc mixed data=bi_meta method=REML cl ;
  class study_id  ;
  model logit = dis  non_dis / noint cl df=1000, 1000, 1000, 1000, 1000, 1000;
  random dis non_dis / subject=study_id type=un V VCorr G GCORR;
  repeated / group=rec;
  parms / parmsdata=c.cov hold=4 to &nCov_par;
  ods output
              V=one_reitsma_covmat
  SolutionF=one_reitsma_estimates
  CovParms=one_reitsma_CovParms
  G=one_NN_GMixed
  GCorr=one_NN_GCorr;
run;
```

10.4.1.1.2 Reitsma model

10.4.1.1.3 Profile likelihood Reitsma model The -2log(likelihood) values are extracted separated for each covariance matrix element, making fix the rest of the elements, and then aloud to increase the element under consideration. A grid is provided to PROC MIXED for the variance of Logit(sens) (0.01 to 25.00 by 0.01) equally for specificity, but for covariance the evaluated interval was (-2.00 to 2.00 by 0.01). The parameters are manipulated in R, 1) To backtransform output=-2log(likelihood), to likelihood: as follow $exp^{-output/2}$ 2) After the backtransformation this vector is normalized diving by the maximimun element of all the elements in the vector, allowing to plot side by sile all the PL in the same frame. The parms used in the model correspond to the hugher and lower levels of the Hierarchical model, the first three are Var (logit(sen))), Covariance and Var (logit(spec))), and the rest correspond to the within variance of each study, calculated with a previous data procedure (Continuity correction section, provide the matrix dataframe: c.cov) from the observed data (DTA studies of the MA).

```
proc mixed data = bi_meta NOPROFILE;
  class study_id;
  model logit = dis  non_dis / noint cl df = 1000, 1000, 1000, 1000, 1000, 1000;
  random dis non_dis / subject = study_id type = un V VCorr;
  repeated / group = rec;
  # first cov parameter
  parms  (0.01 to 25.00 by 0.01)(-0.2203)(0.1021)
  (1.06666666666667) (2.66666666666667) (0.253968253968254)
  (0.243478260869565) (2.18181818181818) (0.404040404040404)
  / eqcons = 2 to & nCov_par;
  ods output
  V = one_reitsma_covmat
  SolutionF = one_reitsma_estimates
  CovParms = one_reitsma_CovParms
  ParmSearch = pl_nopro.one_parmsearch1;
run;

proc mixed data = bi_meta NOPROFILE;
  class study_id;
  model logit = dis  non_dis / noint cl df = 1000, 1000, 1000, 1000, 1000, 1000;
  random dis non_dis / subject = study_id type = un V VCorr;
  repeated / group = rec;
  # second cov parameter
  parms  (1.3103)(-2.00 to 2.00 by 0.01)(0.1021)
  (1.06666666666667) (2.66666666666667) (0.253968253968254)
  (0.243478260869565) (2.18181818181818) (0.404040404040404)
  / eqcons = 1, 3 to & nCov_par;
  ods output
  V = one_reitsma_covmat
  SolutionF = one_reitsma_estimates
  CovParms = one_reitsma_CovParms
  ParmSearch = pl_nopro.one_parmsearch2;
run;

proc mixed data = bi_meta NOPROFILE;
  class study_id;
  model logit = dis  non_dis / noint cl df = 1000, 1000, 1000, 1000, 1000, 1000;
  random dis non_dis / subject = study_id type = un V VCorr;
  repeated / group = rec;
  # third cov parameter
  parms  (1.3103)(-0.2203) (0.01 to 25.00 by 0.01)
  (1.06666666666667) (2.66666666666667) (0.253968253968254)
  (0.243478260869565) (2.18181818181818) (0.404040404040404)
  / eqcons = 1, 2, 4 to & nCov_par;
  ods output
  V = one_reitsma_covmat
  SolutionF = one_reitsma_estimates
  CovParms = one_reitsma_CovParms
  ParmSearch = pl_nopro.one_parmsearch3;
run;
```

10.4.1.2 Binomial-Normal model

```
PROC GLIMMIX data=LAG_bidata method=quad;
  title 'Bivariate generalized linear random-effects model';
  class study status;
  model true/total = status / noint s cl corrb covb ddfm=bw;
  random status / subject=study S type=chol G;
  /* if the model does not converge then replace 'chol' by 'un' */
    estimate 'logit_sens' status 1 0 / cl ilink;
    estimate 'logit_spec' status 0 1 / cl ilink; estimate 'LOR' status 1 1 / cl exp;
  ods output
  parameterestimates = one_BN_p1
  estimates=one_BN_p2
  CovB=one_BN_GMixed
  CorrB=one_BN_GCorr
  CovParms=one_BN_CovaParms
  FitStatistics=c;
   covtest  / parms;
run;
```

10.4.1.2.1 Frequentist binomial-normal model

10.4.1.2.2 Profile likelihood frequentist binomial-normal model

This code was construct by the author of this project based on the code done by Millar (2011) (page 213)[10]

```
data tdata;do i = 1 to 1001;output;end;run;
data tdata;set tdata; if i=1 then do;      covp1=0;   end;
else covp1 + 0.025; drop i;run;
data tdata2;do i = 1 to 1001;output;end;run;
data tdata2;set tdata2; if i=1 then do;      covp2=-2;   end;
else covp2 + 0.004; drop i;run;
data tdata3;do i = 1 to 1001;output;end;run;
data tdata3;set tdata3; if i=1 then do;      covp3=0;   end;
else covp3 + 0.025; drop i;run;

PROC GLIMMIX data=LAG_bidata method=quad;
    title 'Bivariate generalized linear random-effects model';
    class study status;
    model true/total = status / noint s cl corrb covb DDFM=bw;
    random status / subject=study S type=chol G;
    covtest tdata=tdata / parms;
    ods output  covtests=ct;
run;
PROC GLIMMIX data=LAG_bidata method=quad;
    title 'Bivariate generalized linear random-effects model';
    class study status;
    model true/total = status / noint s cl corrb covb   DDFM=bw;
    random status / subject=study S type=chol G;
    covtest tdata=tdata2 / parms;
    ods output  covtests=ct2;   *cambiar aqui;
run;
PROC GLIMMIX data=LAG_bidata method=quad;
    title 'Bivariate generalized linear random-effects model';
    class study status;
    model true/total = status / noint s cl corrb covb  DDFM=bw;
    random status / subject=study S type=chol G;
    covtest tdata=tdata3 / parms;
    ods output  covtests=ct3;
run;
data ct; set ct;drop CovP2 CovP3;run;
data ct2; set ct2;drop CovP1 CovP3;run;
data ct3; set ct3;drop CovP1 CovP2;run;
```

[10] PROC MIXED DATA=estrone NOPROFILE; MODEL y= / SOLUTION; RANDOM INT / SUBJECT=person; PARMS (0.0 TO 0.2 BY 0.0001) (0.0015 TO 0.0060 BY 0.00005); RUN;

```
library(metafor)
plot(influence(rma(measure="PLO",data=DataMA,xi=DataMA$TP,ni=DataMA$FN+DataMA$TP)))
```

10.4.1.2.3 For influential plots

10.4.2 Bayesian code

```
library(bamdit)
## Normal random effect
data <- Barnsley
set.seed(2020)
metadiag(data,
    re = "normal",
    re.model = "SeSp",
    two.by.two = TRUE,
    link = "logit",
    sd.Fisher.rho = 1.7,
    nr.burnin = 1000,
    nr.thin = 1,
    nr.iterations = 10000,
    nr.chains = 3,
    sigma.D.upper = 5,
    sigma.S.upper = 5,
    r2jags = TRUE)

## Mixture Normal random effect
metadiag(data,
    re = "sm",
    re.model = "SeSp",
    two.by.two = TRUE,
    link = "logit",
    sd.Fisher.rho = 1.7,
    nr.burnin = 1000,
    nr.thin = 1,
    nr.iterations = 10000,
    nr.chains = 3,
    sigma.D.upper = 5,
    sigma.S.upper = 5,
    r2jags = TRUE)
```

10.5 Sug-groups from triplot graph

10.6 Profile-likelihoods plots

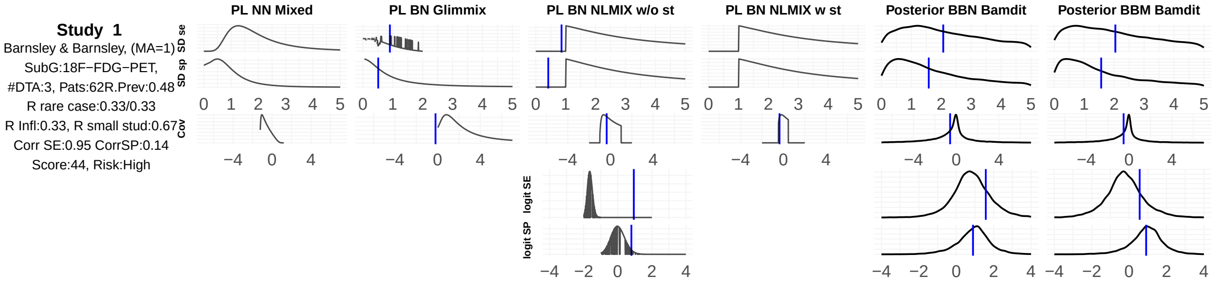


Figure 37: Study 1 Frequentist PL and Bayesian Posterior comparisson between SD(logit(se)), SD(logit(sp)), Covariance, logit(se) and logit(sp)

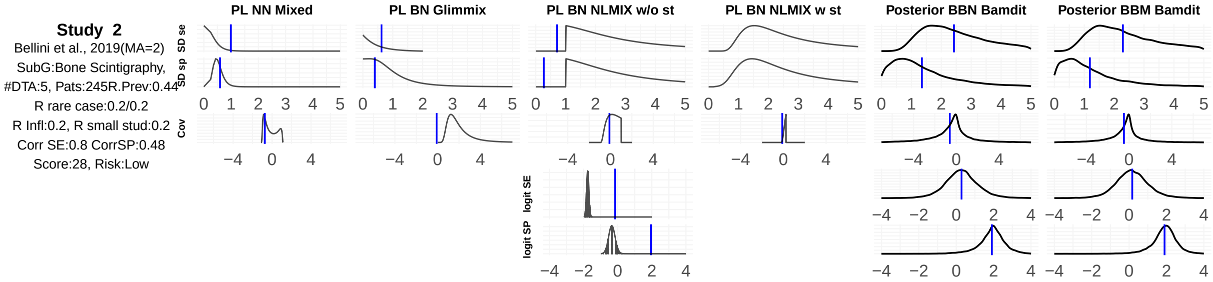


Figure 38: Study 2 Frequentist PL and Bayesian Posterior comparisson between SD(logit(se)), SD(logit(sp)), Covariance, logit(se) and logit(sp)

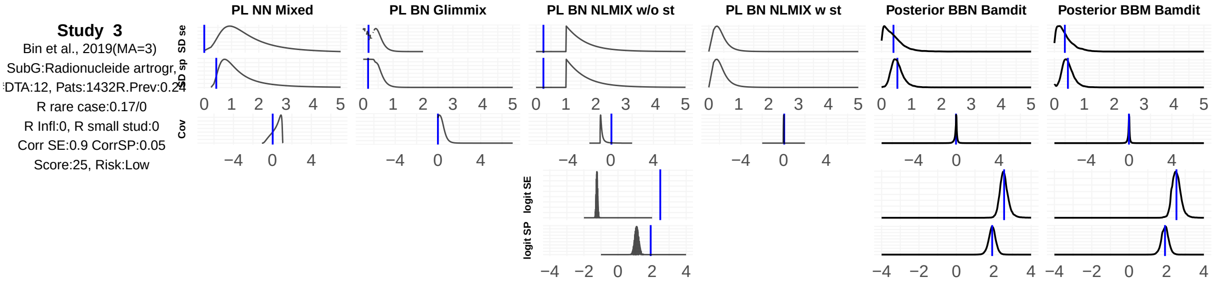


Figure 39: Study 3 Frequentist PL and Bayesian Posterior comparisson between SD(logit(se)), SD(logit(sp)), Covariance, logit(se) and logit(sp)

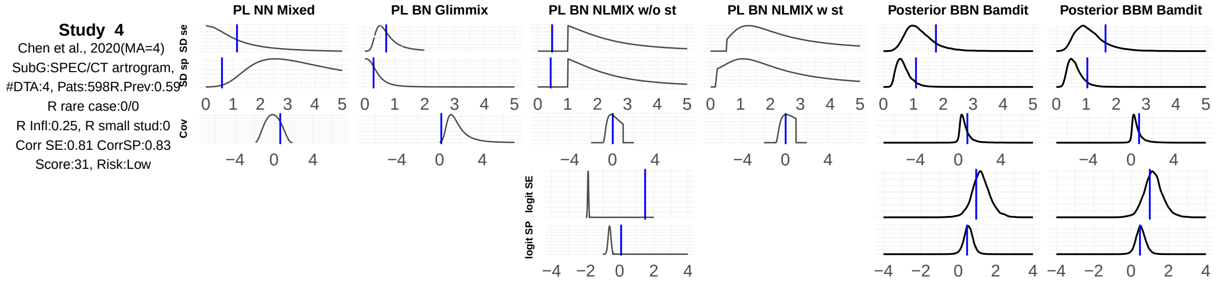


Figure 40: Study 4 Frequentist PL and Bayesian Posterior comparisson between SD(logit(se)), SD(logit(sp)), Covariance, logit(se) and logit(sp)

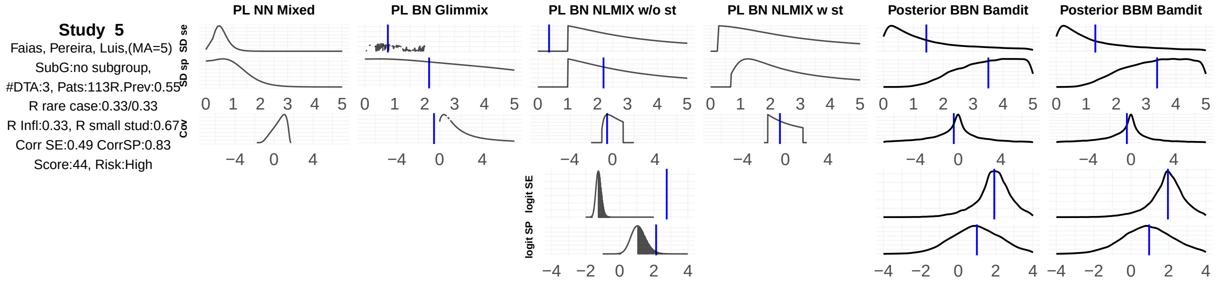


Figure 41: Study 5 Frequentist PL and Bayesian Posterior comparisson between SD(logit(se)), SD(logit(sp)), Covariance, logit(se) and logit(sp)

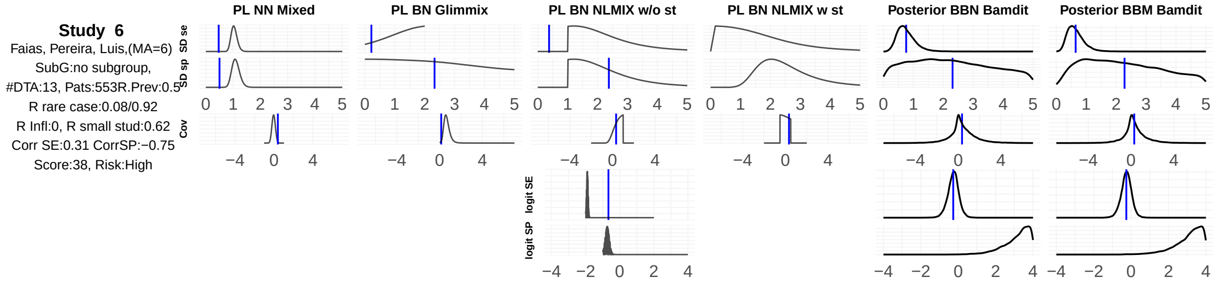


Figure 42: Study 6 Frequentist PL and Bayesian Posterior comparisson between SD(logit(se)), SD(logit(sp)), Covariance, logit(se) and logit(sp)

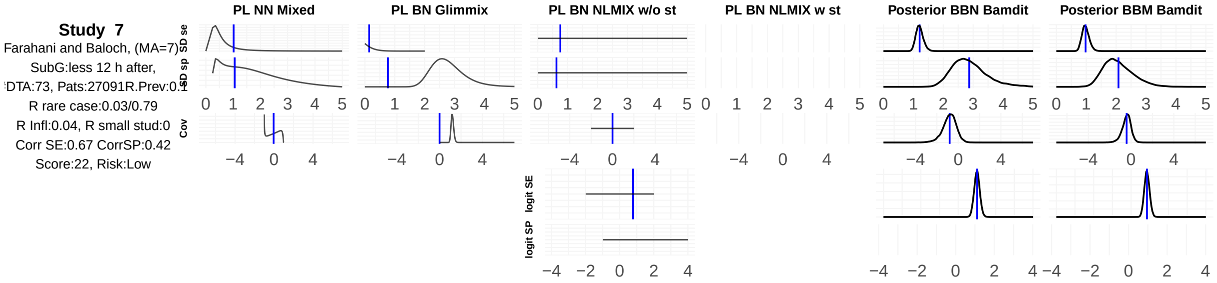


Figure 43: Study 7 Frequentist PL and Bayesian Posterior comparisson between SD(logit(se)), SD(logit(sp)), Covariance, logit(se) and logit(sp)

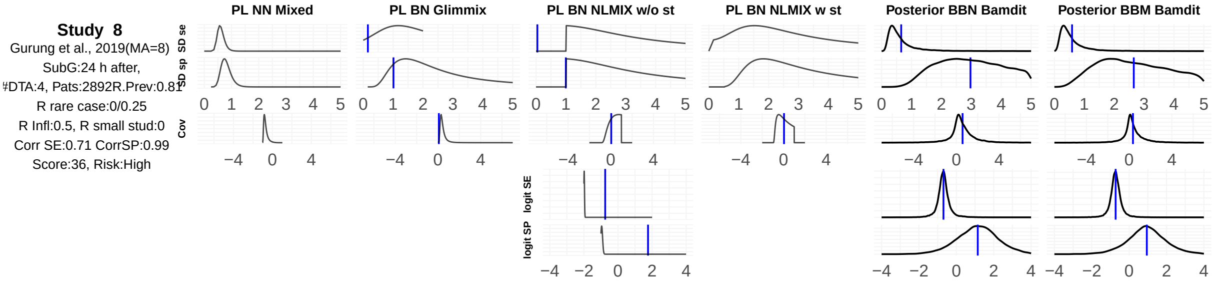


Figure 44: Study 8 Frequentist PL and Bayesian Posterior comparisson between SD(logit(se)), SD(logit(sp)), Covariance, logit(se) and logit(sp)

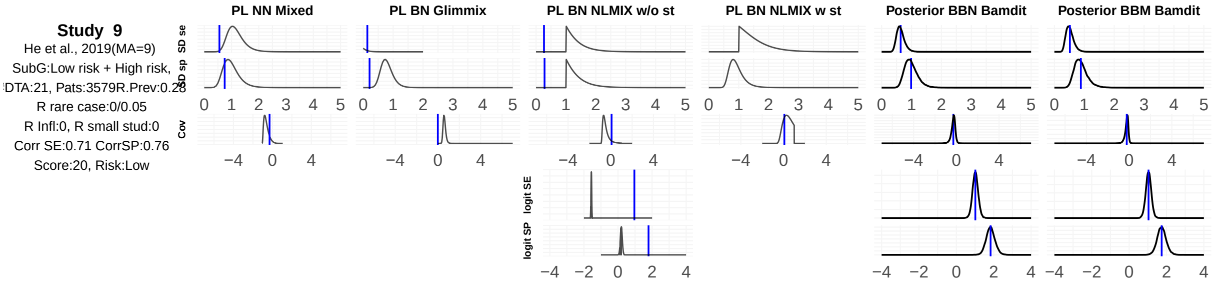


Figure 45: Study 9 Frequentist PL and Bayesian Posterior comparisson between SD(logit(se)), SD(logit(sp)), Covariance, logit(se) and logit(sp)

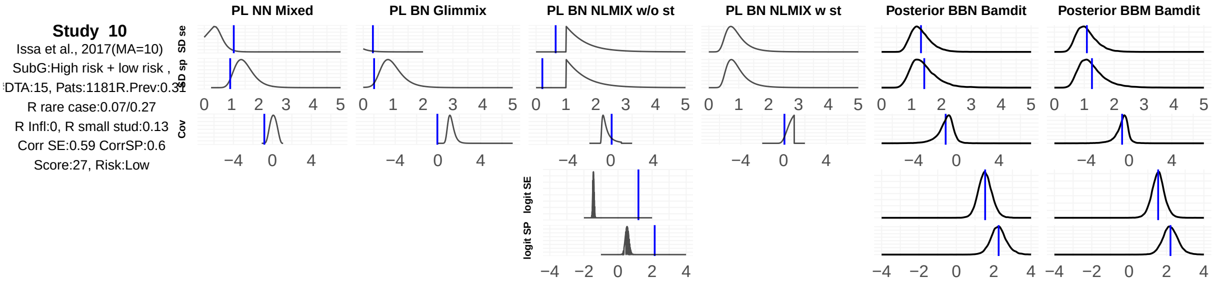


Figure 46: Study 10 Frequentist PL and Bayesian Posterior comparisson between SD(logit(se)), SD(logit(sp)) and Covariance

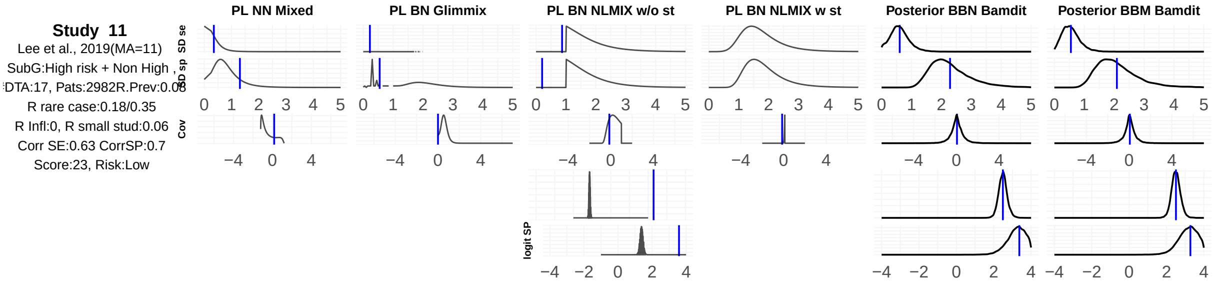


Figure 47: Study 11 Frequentist PL and Bayesian Posterior comparisson between SD(logit(se)), SD(logit(sp)) and Covariance

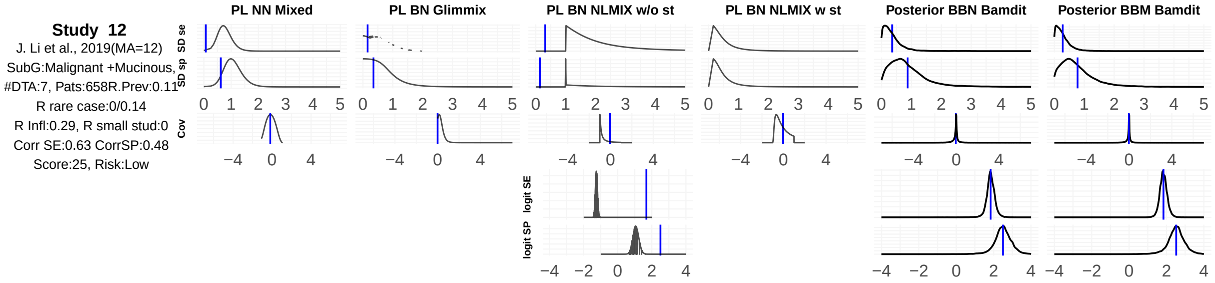


Figure 48: Study 12 Frequentist PL and Bayesian Posterior comparisson between SD(logit(se)), SD(logit(sp)), Covariance, logit(se) and logit(sp)

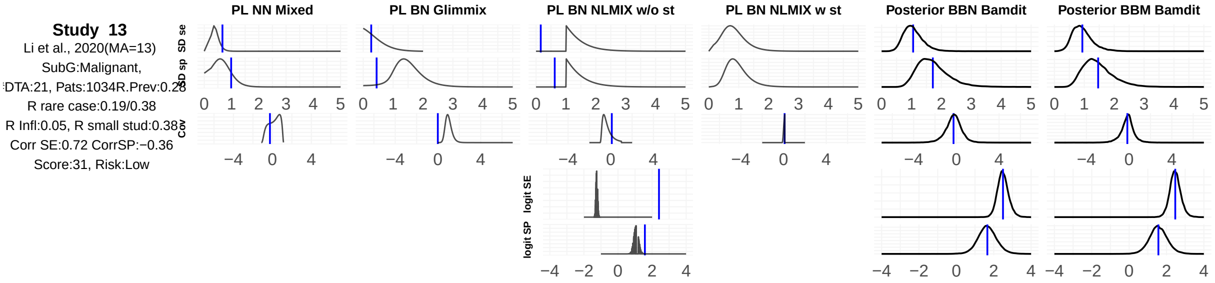


Figure 49: Study 13 Frequentist PL and Bayesian Posterior comparisson between SD(logit(se)), SD(logit(sp)), Covariance, logit(se) and logit(sp)

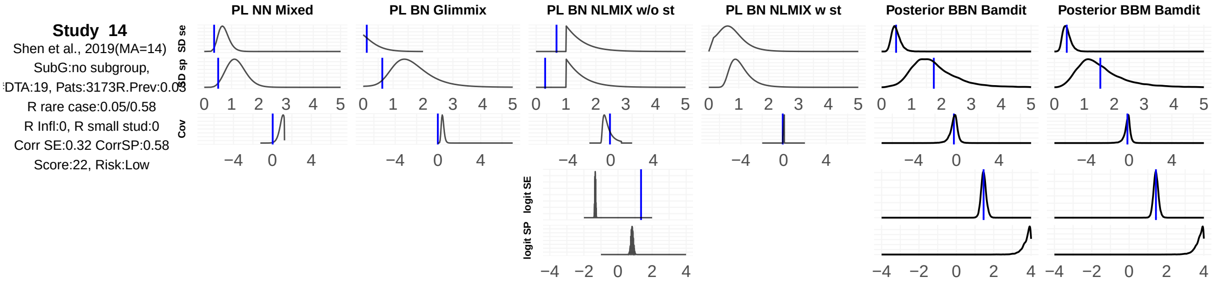


Figure 50: Study 14 Frequentist PL and Bayesian Posterior comparisson between SD(logit(se)), SD(logit(sp)), Covariance, logit(se) and logit(sp)

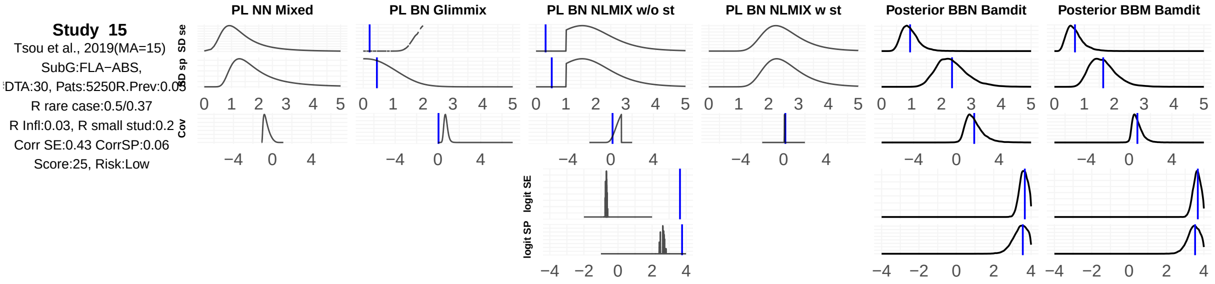


Figure 51: Study 15 Frequentist PL and Bayesian Posterior comparisson between SD(logit(se)), SD(logit(sp)), Covariance, logit(se) and logit(sp)

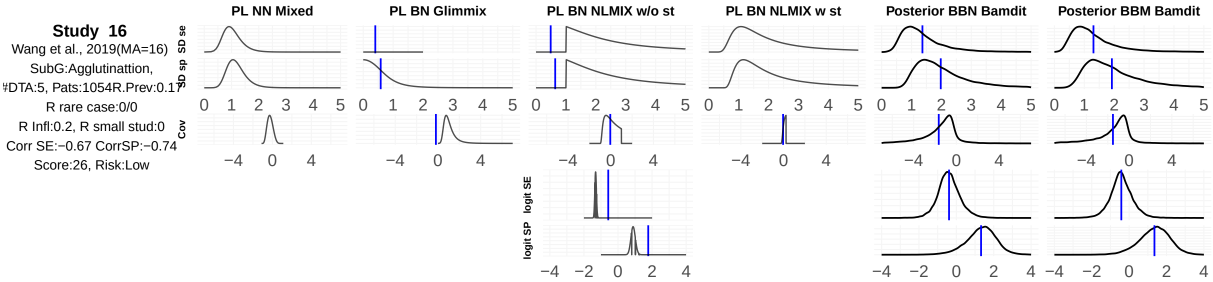


Figure 52: Study 16 Frequentist PL and Bayesian Posterior comparisson between SD(logit(se)), SD(logit(sp)), Covariance, logit(se) and logit(sp)

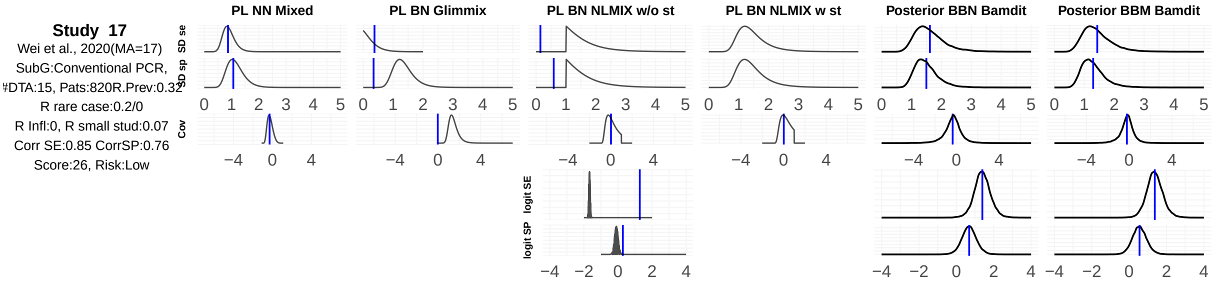


Figure 53: Study 17 Frequentist PL and Bayesian Posterior comparisson between SD(logit(se)), SD(logit(sp)), Covariance, logit(se) and logit(sp)

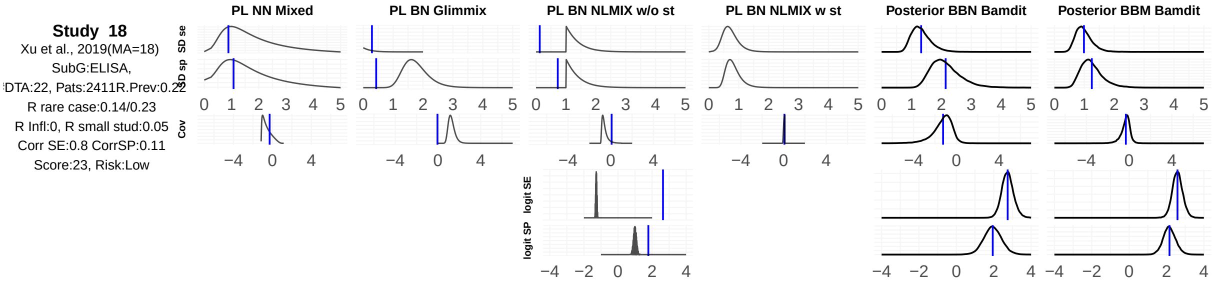


Figure 54: Study 18 Frequentist PL and Bayesian Posterior comparisson between SD(logit(se)), SD(logit(sp)), Covariance, logit(se) and logit(sp)

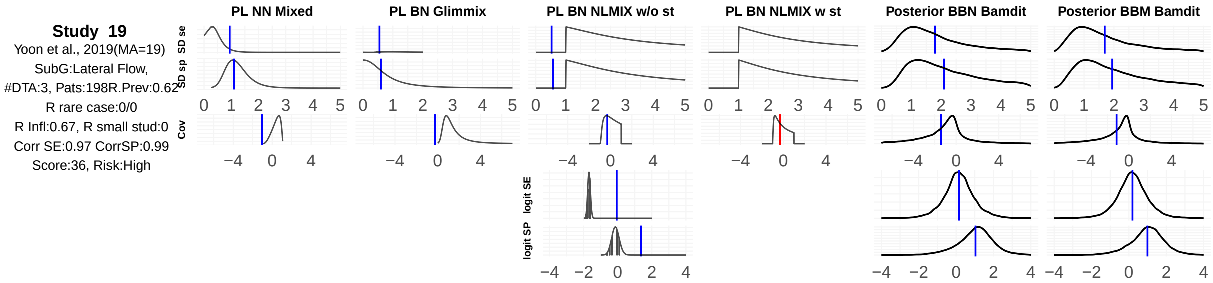


Figure 55: Study 19 Frequentist PL and Bayesian Posterior comparisson between SD(logit(se)), SD(logit(sp)), Covariance, logit(se) and logit(sp)

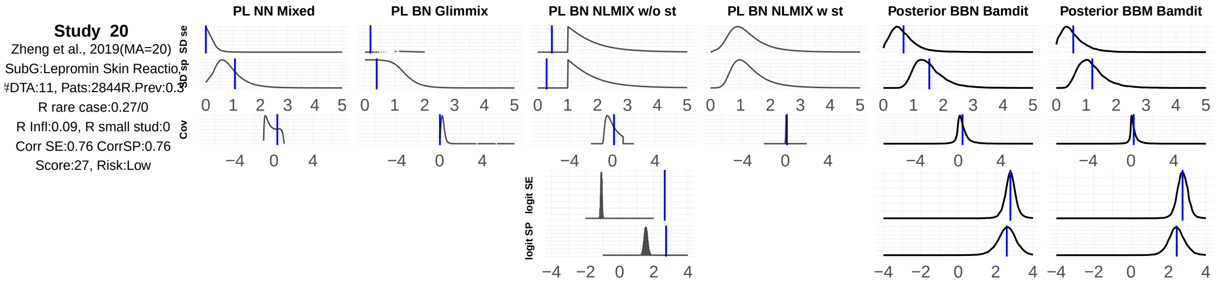


Figure 56: Study 20 Frequentist PL and Bayesian Posterior comparisson between SD(logit(se)), SD(logit(sp)), Covariance, logit(se) and logit(sp)

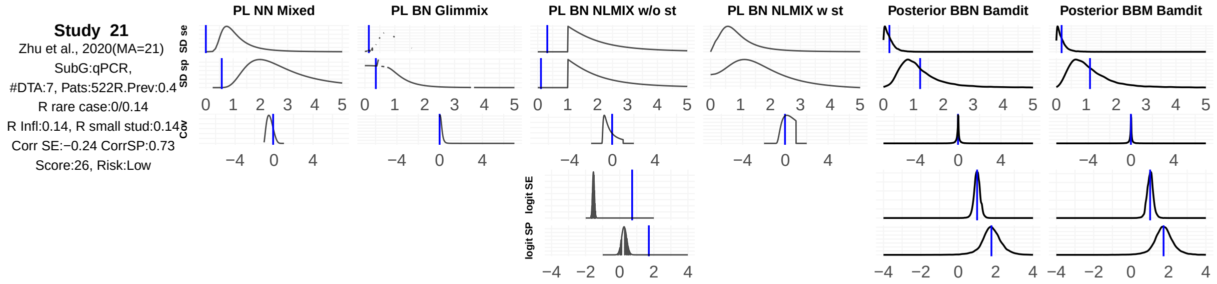


Figure 57: Study 21 Frequentist PL and Bayesian Posterior comparisson between SD(logit(se)), SD(logit(sp)), Covariance, logit(se) and logit(sp)

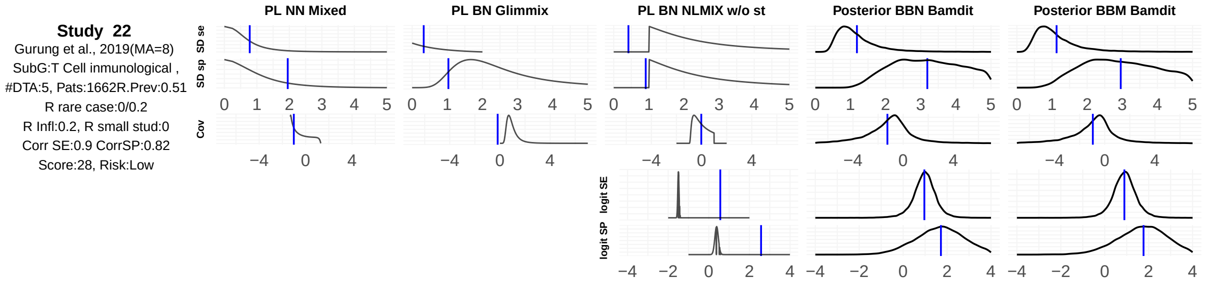


Figure 58: Study 22 Frequentist PL and Bayesian Posterior comparisson between SD(logit(se)), SD(logit(sp)), Covariance, logit(se) and logit(sp)

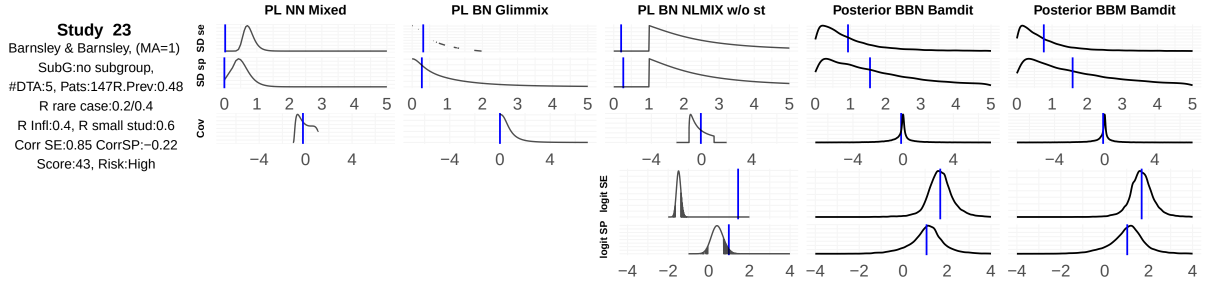


Figure 59: Study 23 Frequentist PL and Bayesian Posterior comparisson between SD(logit(se)), SD(logit(sp)), Covariance, logit(se) and logit(sp)

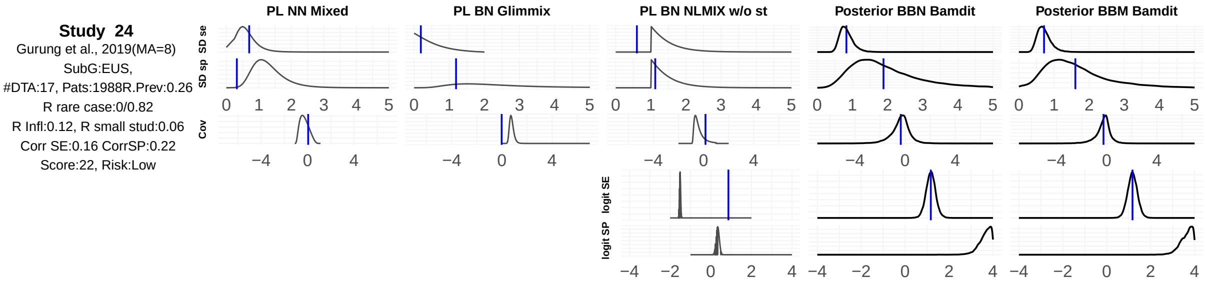


Figure 60: Study 24 Frequentist PL and Bayesian Posterior comparisson between SD(logit(se)), SD(logit(sp)), Covariance, logit(se) and logit(sp)

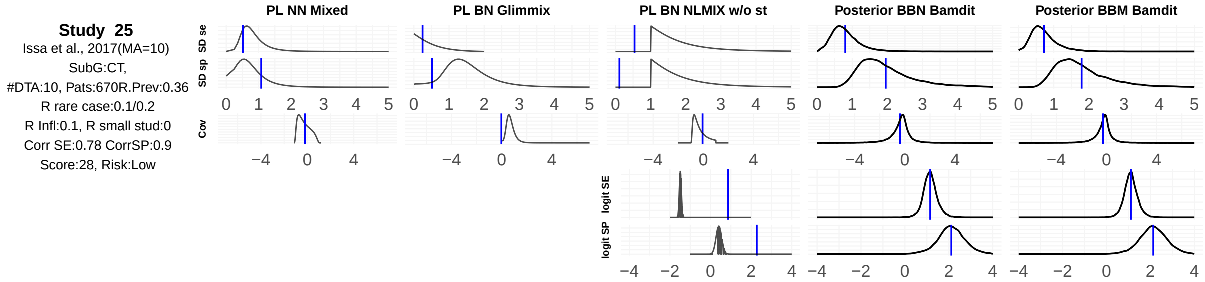


Figure 61: Study 25 Frequentist PL and Bayesian Posterior comparisson between SD(logit(se)), SD(logit(sp)), Covariance, logit(se) and logit(sp)

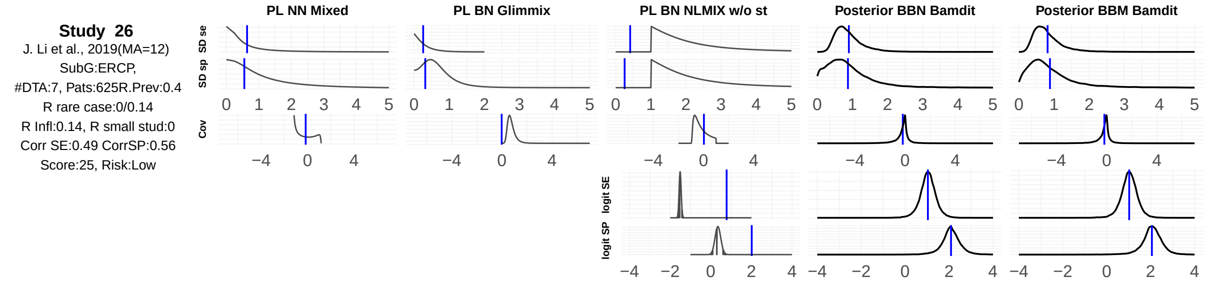


Figure 62: Study 26 Frequentist PL and Bayesian Posterior comparisson between SD(logit(se)), SD(logit(sp)), Covariance, logit(se) and logit(sp)

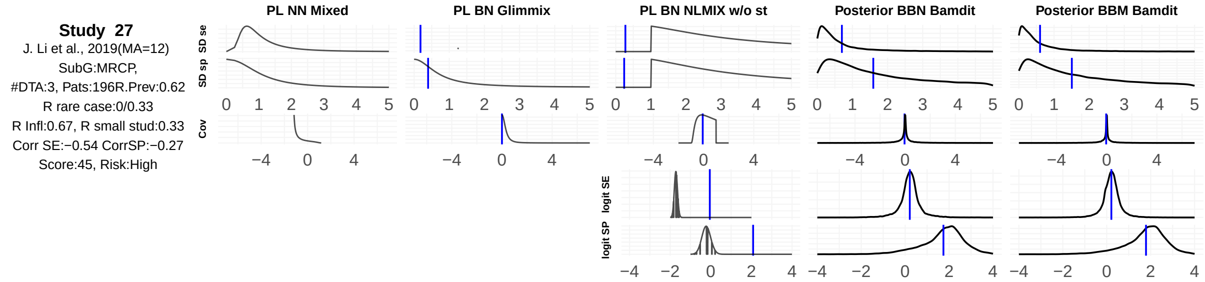


Figure 63: Study 27 Frequentist PL and Bayesian Posterior comparisson between SD(logit(se)), SD(logit(sp)), Covariance, logit(se) and logit(sp)

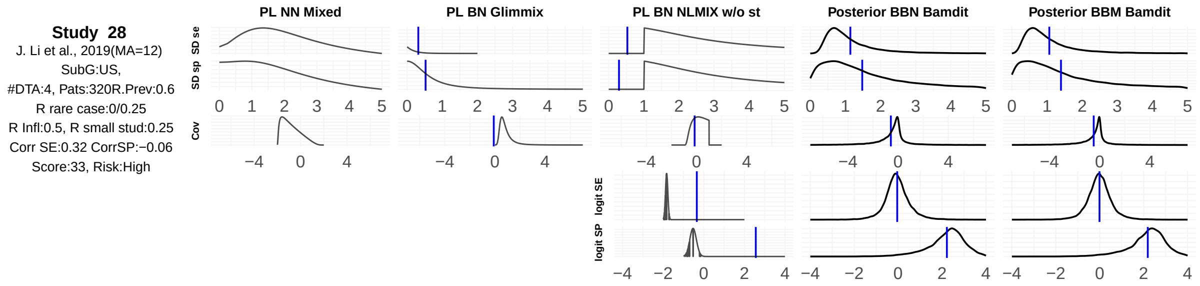


Figure 64: Study 28 Frequentist PL and Bayesian Posterior comparisson between SD(logit(se)), SD(logit(sp)), Covariance, logit(se) and logit(sp)

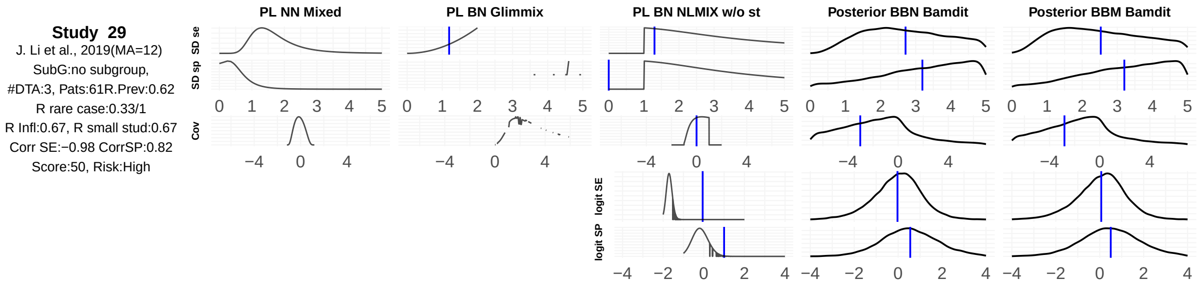


Figure 65: Study 29 Frequentist PL and Bayesian Posterior comparisson between SD(logit(se)), SD(logit(sp)), Covariance, logit(se) and logit(sp)

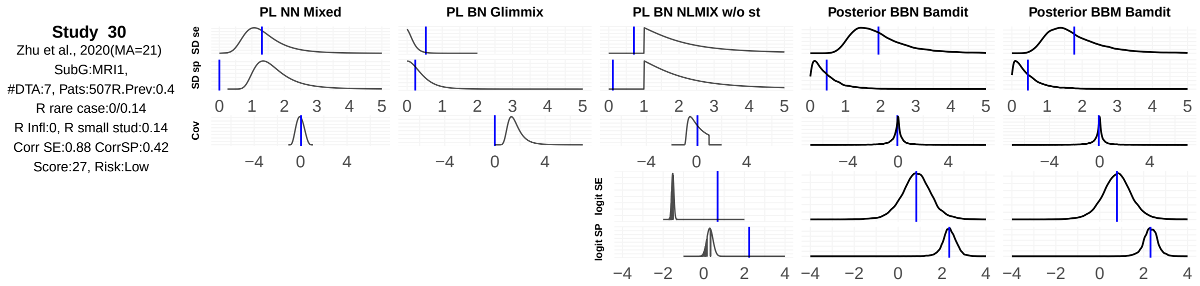


Figure 66: Study 30 Frequentist PL and Bayesian Posterior comparisson between SD(logit(se)), SD(logit(sp)), Covariance, logit(se) and logit(sp)

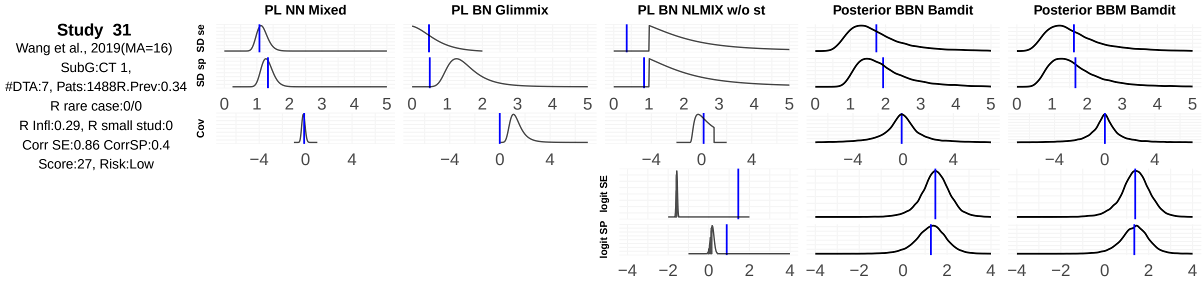


Figure 67: Study 31 Frequentist PL and Bayesian Posterior comparisson between SD(logit(se)), SD(logit(sp)), Covariance, logit(se) and logit(sp)

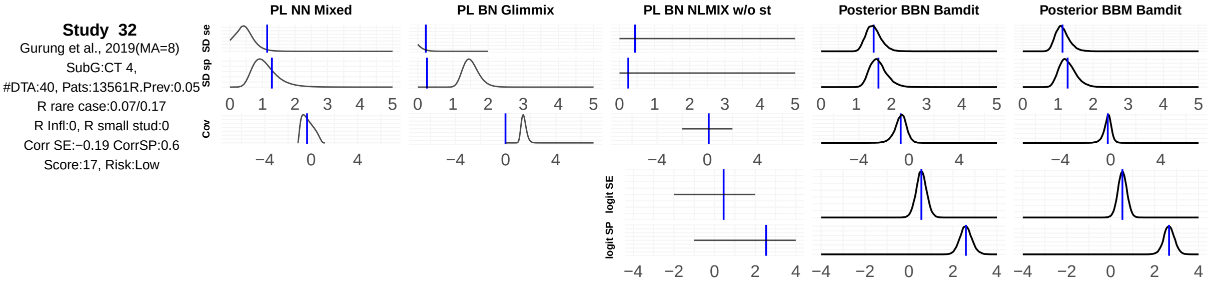


Figure 68: Study 32 Frequentist PL and Bayesian Posterior comparisson between SD(logit(se)), SD(logit(sp)), Covariance, logit(se) and logit(sp)

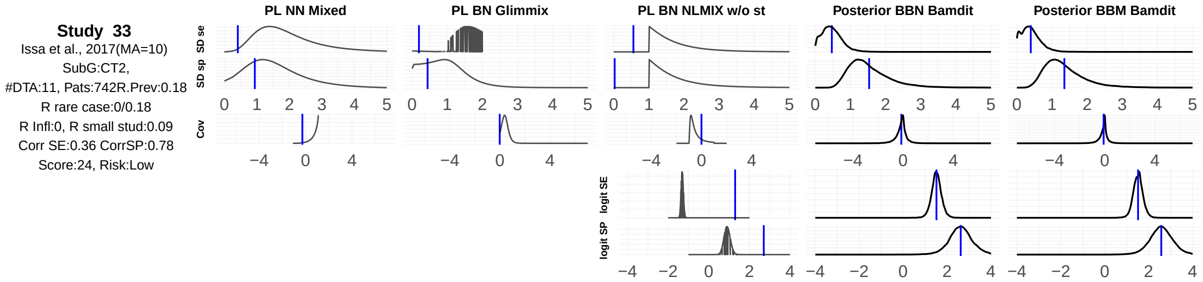


Figure 69: Study 33 Frequentist PL and Bayesian Posterior comparisson between SD(logit(se)), SD(logit(sp)), Covariance, logit(se) and logit(sp)

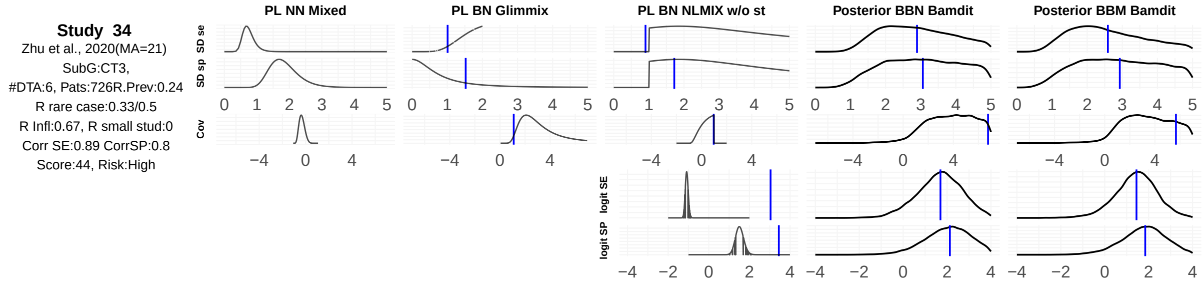


Figure 70: Study 34 Frequentist PL and Bayesian Posterior comparisson between SD(logit(se)), SD(logit(sp)), Covariance, logit(se) and logit(sp)

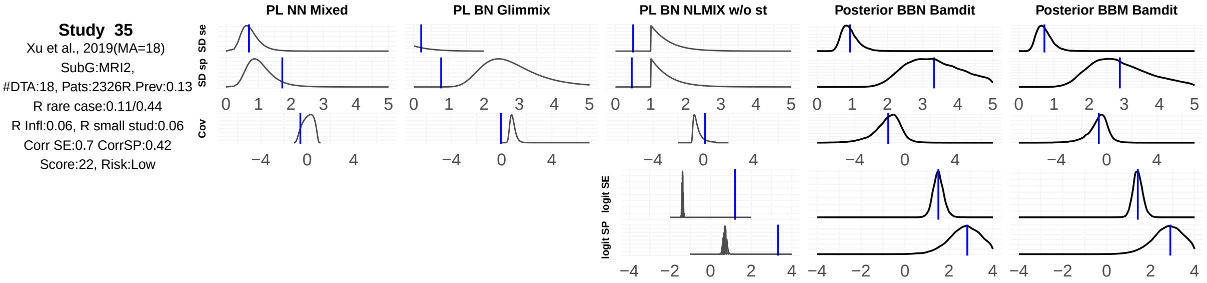


Figure 71: Study 35 Frequentist PL and Bayesian Posterior comparisson between SD(logit(se)), SD(logit(sp)), Covariance, logit(se) and logit(sp)

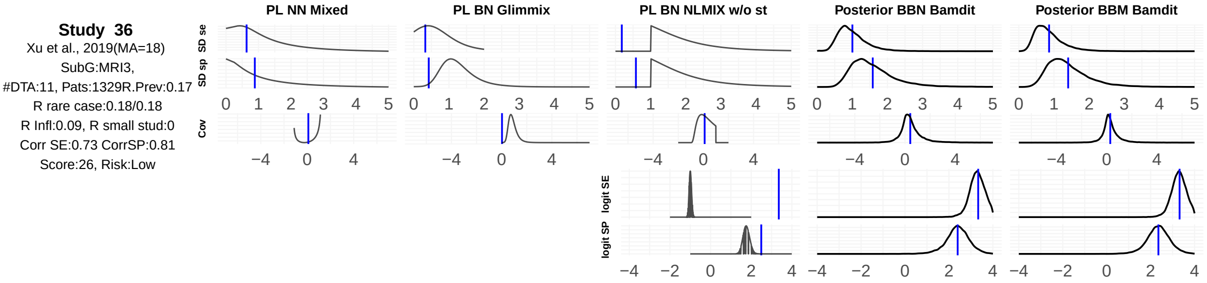


Figure 72: Study 36 Frequentist PL and Bayesian Posterior comparisson between SD(logit(se)), SD(logit(sp)), Covariance, logit(se) and logit(sp)

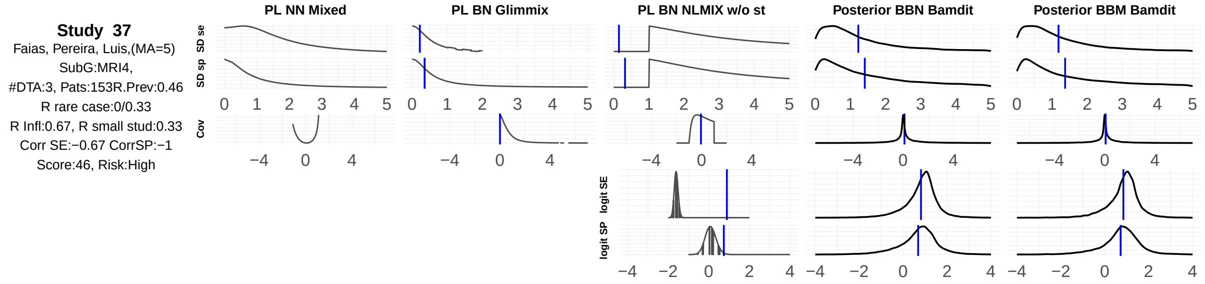


Figure 73: Study 37 Frequentist PL and Bayesian Posterior comparisson between SD(logit(se)), SD(logit(sp)), Covariance, logit(se) and logit(sp)

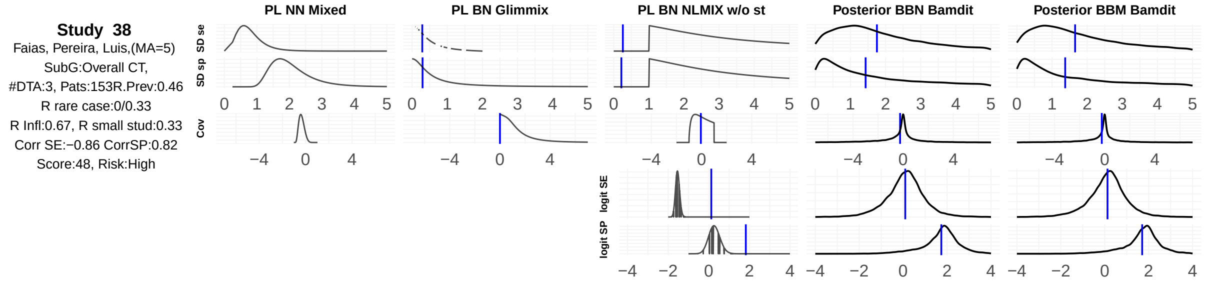


Figure 74: Study 38 Frequentist PL and Bayesian Posterior comparisson between SD(logit(se)), SD(logit(sp)), Covariance, logit(se) and logit(sp)

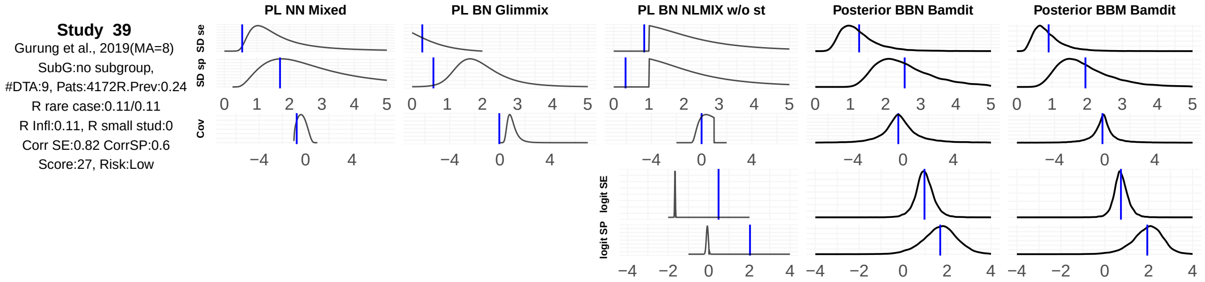


Figure 75: Study 39 Frequentist PL and Bayesian Posterior comparisson between SD(logit(se)), SD(logit(sp)), Covariance, logit(se) and logit(sp)

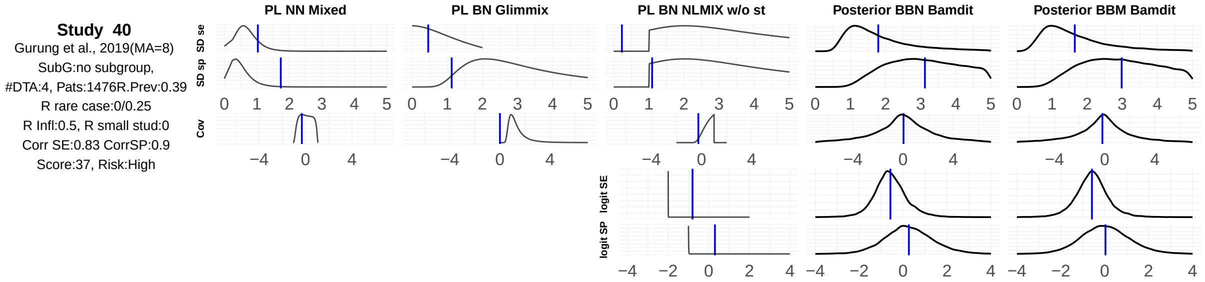


Figure 76: Study 40 Frequentist PL and Bayesian Posterior comparisson between SD(logit(se)), SD(logit(sp)), Covariance, logit(se) and logit(sp)

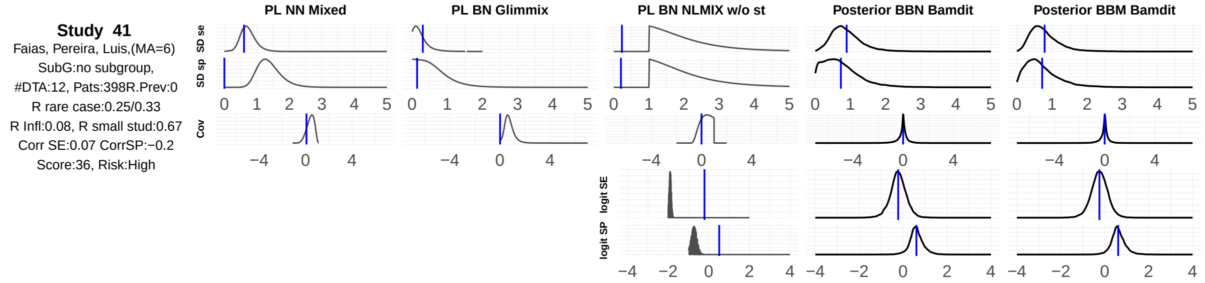


Figure 77: Study 41 Frequentist PL and Bayesian Posterior comparisson between SD(logit(se)), SD(logit(sp)), Covariance, logit(se) and logit(sp)

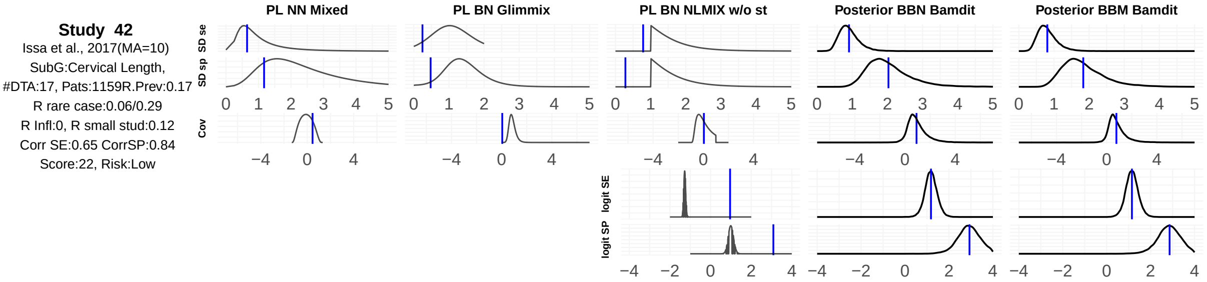


Figure 78: Study 42 Frequentist PL and Bayesian Posterior comparisson between SD(logit(se)), SD(logit(sp)), Covariance, logit(se) and logit(sp)

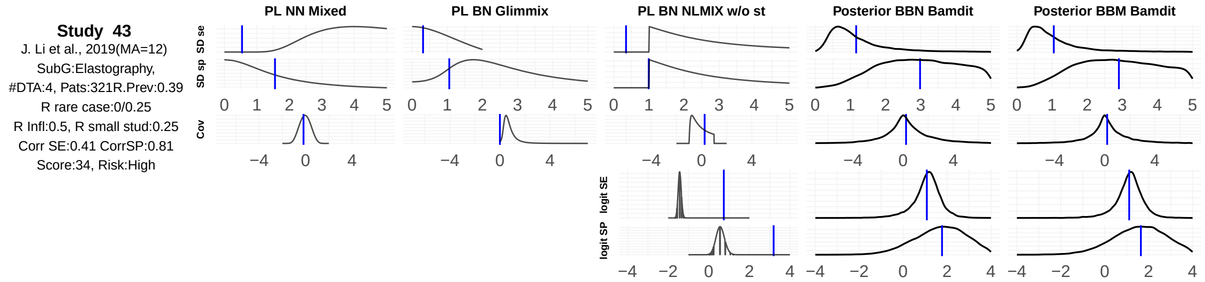


Figure 79: Study 43 Frequentist PL and Bayesian Posterior comparisson between SD(logit(se)), SD(logit(sp)), Covariance, logit(se) and logit(sp)

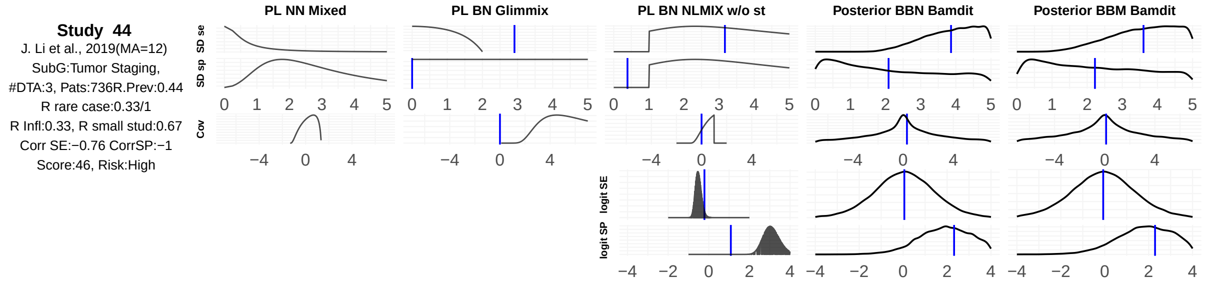


Figure 80: Study 44 Frequentist PL and Bayesian Posterior comparisson between SD(logit(se)), SD(logit(sp)), Covariance, logit(se) and logit(sp)

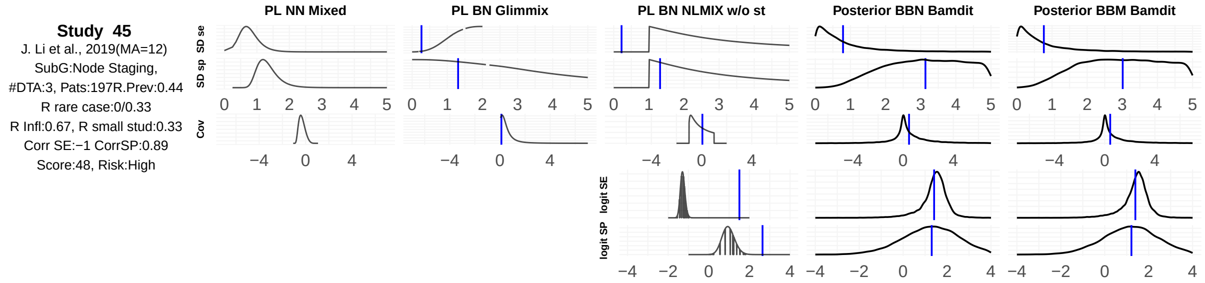


Figure 81: Study 45 Frequentist PL and Bayesian Posterior comparisson between SD(logit(se)), SD(logit(sp)), Covariance, logit(se) and logit(sp)

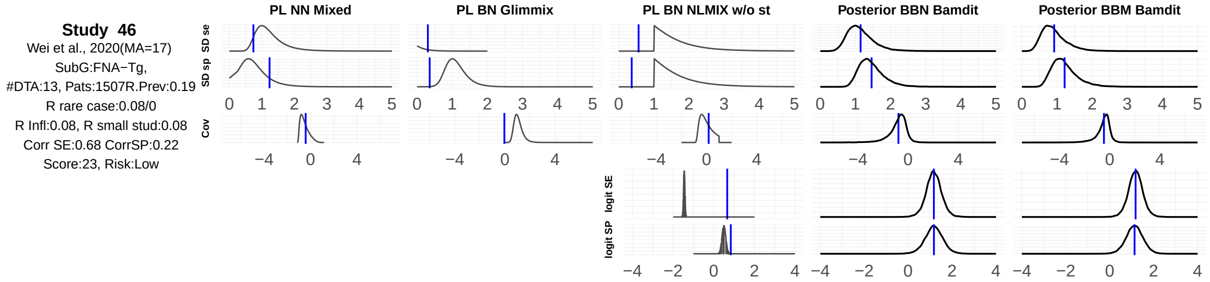


Figure 82: Study 46 Frequentist PL and Bayesian Posterior comparisson between SD(logit(se)), SD(logit(sp)), Covariance, logit(se) and logit(sp)

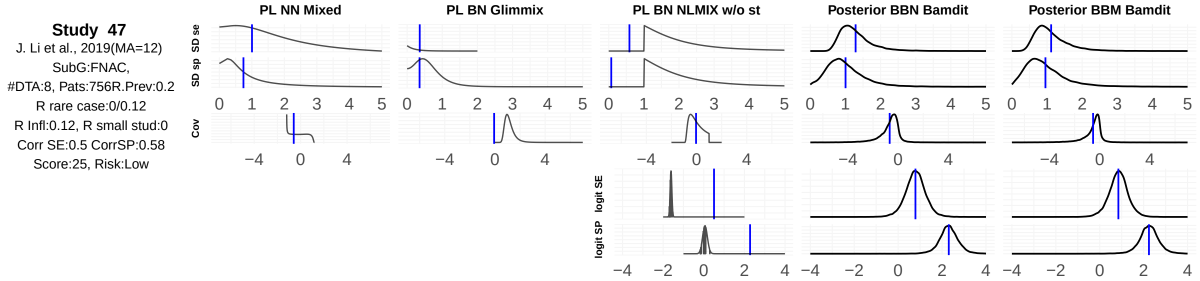


Figure 83: Study 47 Frequentist PL and Bayesian Posterior comparisson between SD(logit(se)), SD(logit(sp)), Covariance, logit(se) and logit(sp)

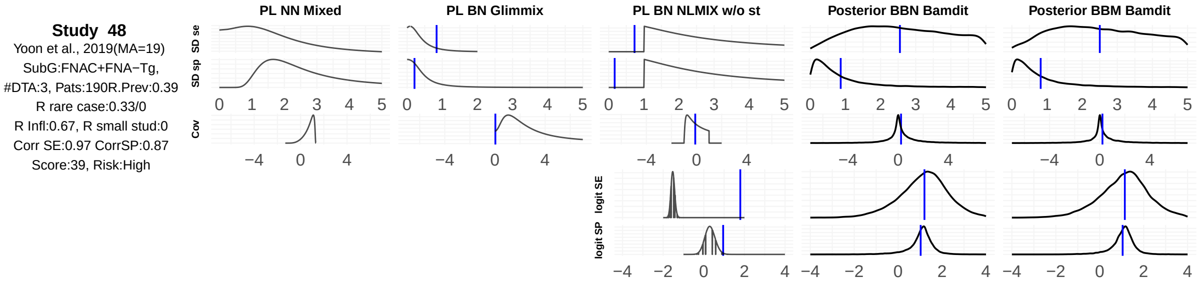


Figure 84: Study 48 Frequentist PL and Bayesian Posterior comparisson between SD(logit(se)), SD(logit(sp)), Covariance, logit(se) and logit(sp)

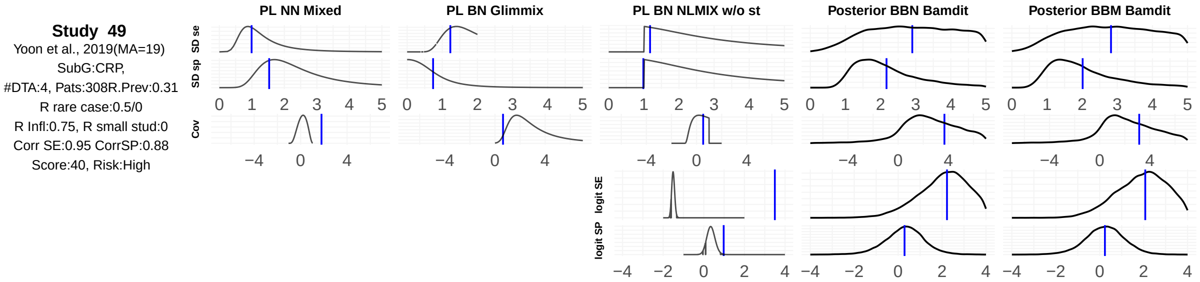


Figure 85: Study 49 Frequentist PL and Bayesian Posterior comparisson between SD(logit(se)), SD(logit(sp)), Covariance, logit(se) and logit(sp)

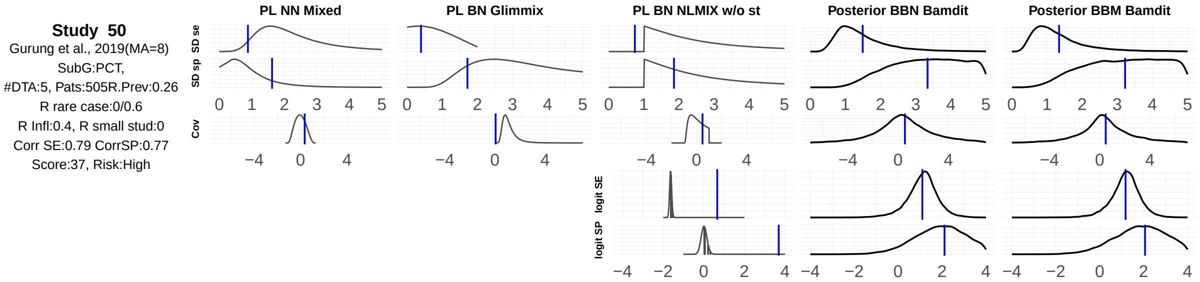


Figure 86: Study 50 Frequentist PL and Bayesian Posterior comparisson between SD(logit(se)), SD(logit(sp)), Covariance, logit(se) and logit(sp)

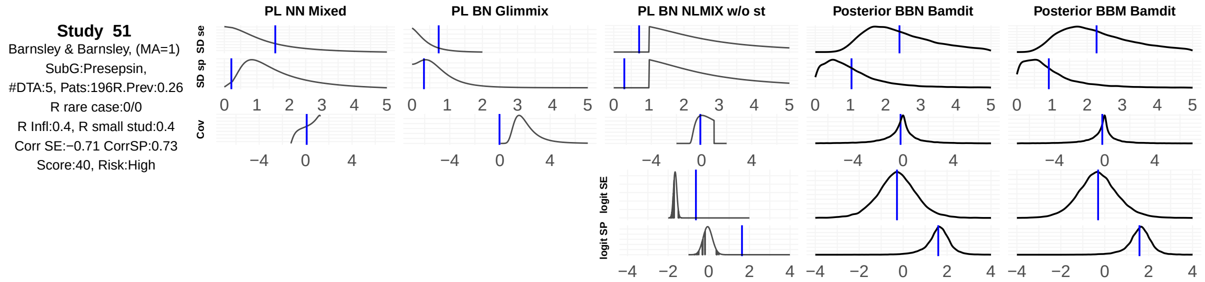


Figure 87: Study 51 Frequentist PL and Bayesian Posterior comparisson between SD(logit(se)), SD(logit(sp)), Covariance, logit(se) and logit(sp)

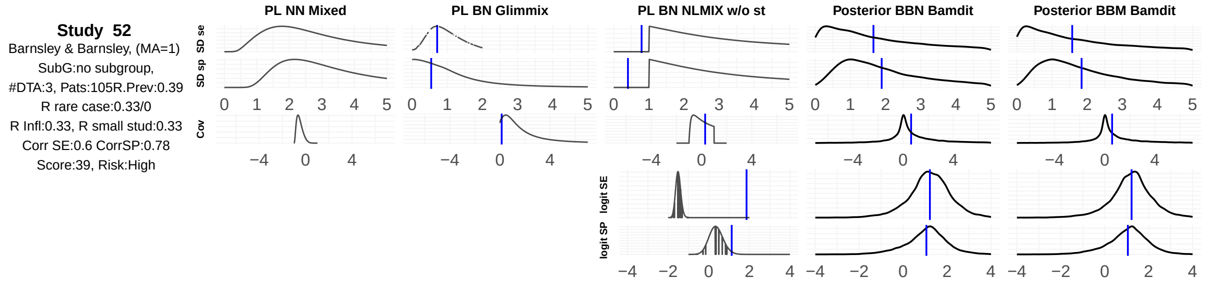


Figure 88: Study 52 Frequentist PL and Bayesian Posterior comparisson between SD(logit(se)), SD(logit(sp)), Covariance, logit(se) and logit(sp)

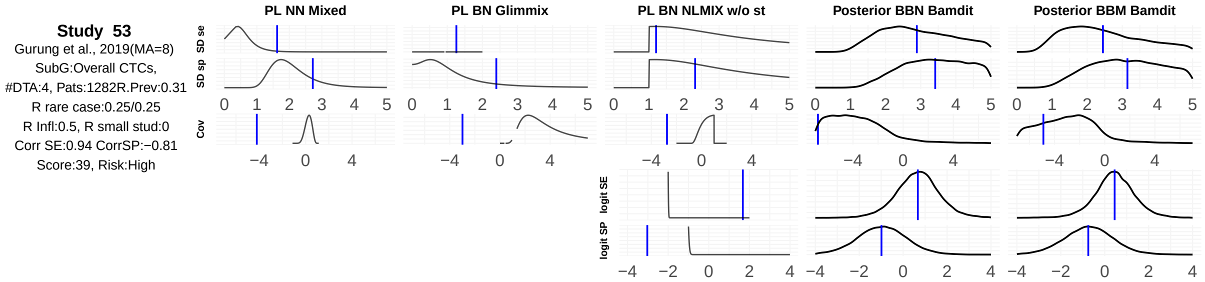


Figure 89: Study 53 Frequentist PL and Bayesian Posterior comparisson between SD(logit(se)), SD(logit(sp)), Covariance, logit(se) and logit(sp)

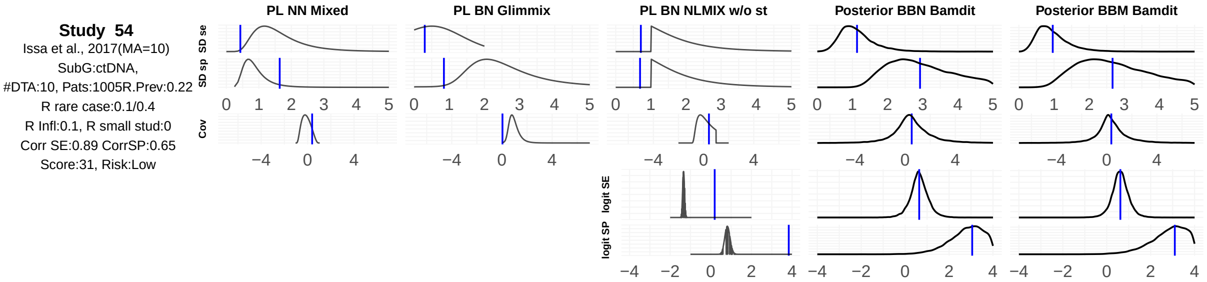


Figure 90: Study 54 Frequentist PL and Bayesian Posterior comparisson between SD(logit(se)), SD(logit(sp)), Covariance, logit(se) and logit(sp)

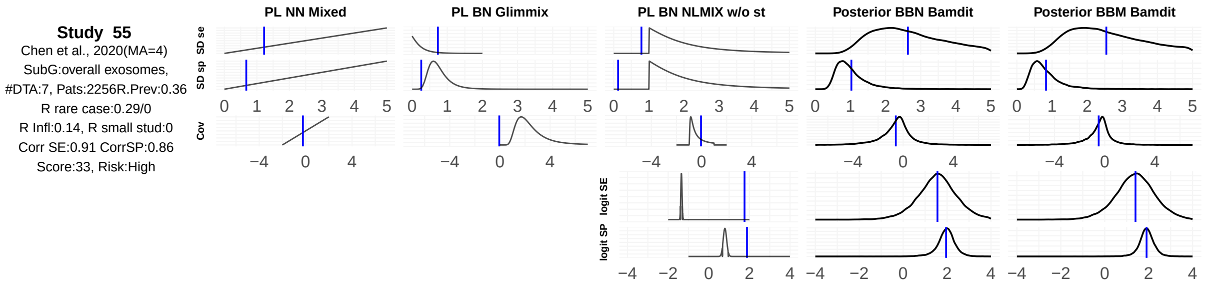


Figure 91: Study 55 Frequentist PL and Bayesian Posterior comparisson between SD(logit(se)), SD(logit(sp)), Covariance, logit(se) and logit(sp)

11 References

R (Version 3.6.3; R Core Team 2020) and the R-packages *DescTools* [Version 0.99.31], *mada* [Version 0.5.9][11], *ggtern* (Version 3.3.0; Hamilton and Ferry 2018)[12], *boot* (Version 1.3.25; Davison and Hinkley 1997)[13], *coda* [Version 0.19.3][14], *bamdit* [Version 3.3.2; Verde P. 2019], *cutpointr* (Version 1.0.32; Thiele 2020)[15] *ggplot2* (Version 3.3.2; Wickham 2016), *treemap* (Version 2.4.2; Tennekes 2017)[16], *ztable* (Version 0.2.1; Moon 2020)[17], *factoextra* (Version 1.0.7; Kassambara and Mundt 2020), *FactoMineR* (Version 2.3; Lê, Josse, and Husson 2008)[18], *plotrix* (Version 3.7.8; J 2006)[19], *metafor* [version 2.4-0] [20]. [21]

Abrams, Keith R., Jonathan P. Myles, and D. J. Spiegelhalter. 2004. *Bayesian Approaches to Clinical Trials and Health-Care Evaluation.* Chichester ; Hoboken, NJ: John Wiley & Sons.

Barnsley, Lara, and Les Barnsley. 2019. "Detection of Aseptic Loosening in Total Knee Replacements: A Systematic Review and Meta-Analysis." *Skeletal Radiology* 48 (10): 1565–72. https://doi.org/10.1007/s00256-019-03215-y.

Bellini, Davide, Flaminia Rivosecchi, Nicola Panvini, Marco Rengo, Damiano Caruso, Iacopo Carbone, Riccardo Ferrari, Pasquale Paolantonio, and Andrea Laghi. 2019. "Layered Enhancement at Magnetic Resonance Enterography in Inflammatory Bowel Disease: A Meta-Analysis." *World Journal of Gastroenterology* 25 (31): 4555–66. https://doi.org/10.3748/wjg.v25.i31.4555.

Bijma, Fetsje, Marianne Jonker, and Aad van der Vaart. 2017. *An Introduction to Mathematical Statistics.* Translated by Reinie Erné. Amsterdam University Press.

Bin, Lian, Yang Huihui, Yang Weiping, Wei Changyuan, Qin Qinghong, and Meng Weiyu. 2019. "Value of Three-Dimensional Ultrasound in Differentiating Malignant from Benign Breast Tumors." *Ultrasound Quarterly* 35 (1): 68–73. https://doi.org/10.1097/RUQ.0000000000000433.

Bland, J. Martin, and DouglasG Altman. 1986. "Statistical Methods for Assessing Agreement Between Two Methods of Clinical Measurement." *The Lancet* 327 (8476): 307–10. https://doi.org/10.1016/S0140-6736(86)90837-8.

Christensen, Ronald, Wesley Johnson, Adam Branscum, and Timothy E. Hanson. 2010. *Bayesian Ideas and Data Analysis: An Introduction for Scientists and Statisticians.* 1 edition. Boca Raton, FL: CRC Press.

Chu, Haitao, and Stephen R. Cole. 2006. "Bivariate Meta-Analysis of Sensitivity and Specificity with Sparse Data: A Generalized Linear Mixed Model Approach." *Journal of Clinical Epidemiology* 59 (12): 1331–2. https://doi.org/10.1016/j.jclinepi.2006.06.011.

Cole, Stephen R., Haitao Chu, and Sander Greenland. 2014. "Maximum Likelihood, Profile Likelihood, and Penalized Likelihood: A Primer." *American Journal of Epidemiology* 179 (2): 252–60. https://doi.org/10.1093/aje/kwt245.

Corbeil, R. R., and S. R. Searle. 1976. "Restricted Maximum Likelihood (REML) Estimation of Variance Components in the Mixed Model." *Technometrics* 18 (1): 31–38. https://doi.org/10.2307/1267913.

[11] Only used for exploratory reasons not for results

[12] For triplot diagram

[13] For CI bootstrap in Bland-Altman analysis

[14] Only for MCMC convergency check purpose

[15] To determine and evaluate optimal cutpoints

[16] For the treemap plots in the descriptive analysis

[17] For the descriptive table colored with logical conditions

[18] For all the PCA plots used

[19] For color2D.matplotnumeric function that provide 2D matrix or data frame as colored rectangles

[20] For the influentials plots, code in the appendix

[21] These are a list of the core R packages used during the current Thesis work

Curcio, Daniel, and Pablo E. Verde. 2011. "Comment on: Efficacy and Safety of Tigecycline: A Systematic Review and Meta-Analysis." *Journal of Antimicrobial Chemotherapy* 66 (12): 2893–5. https://doi.org/10.1093/jac/dkr368.

Dahabreh, Issa J, Thomas A Trikalinos, and Joseph Lau. n.d. "An Empirical Assessment of Bivariate Methods for Meta-Analysis of Test Accuracy," 96.

Davison, A. C., and D. V. Hinkley. 1997. *Bootstrap Methods and Their Applications.* Cambridge: Cambridge University Press. http://statwww.epfl.ch/davison/BMA/.

Dewey, Michael. 2020. "CRAN Task View: Meta-Analysis," July. https://CRAN.R-project.org/view=MetaAnalysis.

Diaz, Mireya. 2015. "Performance Measures of the Bivariate Random Effects Model for Meta-Analyses of Diagnostic Accuracy." *Computational Statistics & Data Analysis* 83 (March): 82–90. https://doi.org/10.1016/j.csda.2014.09.021.

Faias, Sandra, Luisa Pereira, Ângelo Luís, Paula Chaves, and Marília Cravo. 2019. "Genetic Testing Vs Microforceps Biopsy in Pancreatic Cysts: Systematic Review and Meta-Analysis." *World Journal of Gastroenterology* 25 (26): 3450–67. https://doi.org/10.3748/wjg.v25.i26.3450.

Faias, Sandra, Luisa Pereira, Ângelo Luís, Marília Cravo, António Dias Pereira, and Joana Torres. 2019. "KRAS in Cyst Fluid Obtained by Endoscopic Ultrasound-Fine-Needle Aspiration in Pancreatic Cystic Lesions: A Systematic Review and Meta-Analysis." *Pancreas* 48 (6): 749–58. https://doi.org/10.1097/MPA.0000000000001325.

Farahani, Sahar J., and Zubair Baloch. 2019. "Are We Ready to Develop a Tiered Scheme for the Effusion Cytology? A Comprehensive Review and Analysis of the Literature." *Diagnostic Cytopathology* 47 (11): 1145–59. https://doi.org/10.1002/dc.24278.

Gelman, Andrew, John B. Carlin, Hal S. Stern, David B. Dunson, Aki Vehtari, and Donald B. Rubin. 2013. *Bayesian Data Analysis.* 3 edition. Boca Raton: Chapman and Hall/CRC.

Giavarina, Davide. 2015. "Understanding Bland Altman Analysis." *Biochemia Medica* 25 (2): 141–51. https://doi.org/10.11613/BM.2015.015.

Hamilton, Nicholas E., and Michael Ferry. 2018. "ggtern: Ternary Diagrams Using ggplot2." *Journal of Statistical Software, Code Snippets* 87 (3): 1–17. https://doi.org/10.18637/jss.v087.c03.

Hamza, Taye H., Hans C. van Houwelingen, and Theo Stijnen. 2008. "The Binomial Distribution of Meta-Analysis Was Preferred to Model Within-Study Variability." *Journal of Clinical Epidemiology* 61 (1): 41–51. https://doi.org/10.1016/j.jclinepi.2007.03.016.

Harbord, Roger M., Jonathan J. Deeks, Matthias Egger, Penny Whiting, and Jonathan A. C. Sterne. 2007. "A Unification of Models for Meta-Analysis of Diagnostic Accuracy Studies." *Biostatistics (Oxford, England)* 8 (2): 239–51. https://doi.org/10.1093/biostatistics/kxl004.

Harbord, Roger M., Penny Whiting, Jonathan A. C. Sterne, Matthias Egger, Jonathan J. Deeks, Aijing Shang, and Lucas M. Bachmann. 2008. "An Empirical Comparison of Methods for Meta-Analysis of Diagnostic Accuracy Showed Hierarchical Models Are Necessary." *Journal of Clinical Epidemiology* 61 (11): 1095–1103. https://doi.org/10.1016/j.jclinepi.2007.09.013.

He, Yong-Peng, Li-Xian Li, Jia-Xi Tang, Lin Yi, Yi Zhao, Hai-Wei Zhang, Zhi-Juan Wu, et al. 2019. "HE4 as a Biomarker for Diagnosis of Lung Cancer: A Meta-Analysis." *Medicine* 98 (39): e17198. https://doi.org/10.1097/MD.0000000000017198.

Hoaglin, David C. 2018. "Contribution to the Discussion of 'When Should Meta-Analysis Avoid Making Hidden Normality Assumptions?'." *Biometrical Journal* 60 (6): 1083–4. https://doi.org/10.1002/bimj.

201800188.

Houwelingen, Hans C. Van, Koos H. Zwinderman, and Theo Stijnen. 1993. "A Bivariate Approach to Meta-Analysis." *Statistics in Medicine* 12 (24): 2273–84. https://doi.org/10.1002/sim.4780122405.

J, Lemon. 2006. "Plotrix: A Package in the Red Light District of R." *R-News* 6 (4): 8–12.

Jackson, Dan. 2006. "The Power of the Standard Test for the Presence of Heterogeneity in Meta-Analysis." *Statistics in Medicine* 25 (15): 2688–99. https://doi.org/10.1002/sim.2481.

Jackson, Dan, and Ian R. White. 2018. "When Should Meta-Analysis Avoid Making Hidden Normality Assumptions?" *Biometrical Journal* 60 (6): 1040–58. https://doi.org/10.1002/bimj.201800071.

Jackson, Dan, Ian R White, and Richard D Riley. 2012. "Quantifying the Impact of Between-Study Heterogeneity in Multivariate Meta-Analyses." *Statistics in Medicine* 31 (29): 3805–20. https://doi.org/10.1002/sim.5453.

Karahalios, Amalia (Emily), Georgia Salanti, Simon L. Turner, G. Peter Herbison, Ian R. White, Areti Angeliki Veroniki, Adriani Nikolakopoulou, and Joanne E. Mckenzie. 2017. "An Investigation of the Impact of Using Different Methods for Network Meta-Analysis: A Protocol for an Empirical Evaluation." *Systematic Reviews* 6 (June). https://doi.org/10.1186/s13643-017-0511-x.

Kassambara, Alboukadel, and Fabian Mundt. 2020. *Factoextra: Extract and Visualize the Results of Multivariate Data Analyses.* https://CRAN.R-project.org/package=factoextra.

Kontopantelis, Evangelos, and David Reeves. 2012. "Performance of Statistical Methods for Meta-Analysis When True Study Effects Are Non-Normally Distributed: A Comparison Between DerSimonianLaird and Restricted Maximum Likelihood." *Statistical Methods in Medical Research* 21 (6): 657–59. https://doi.org/10.1177/0962280211413451.

Lee, Juneyoung, Kyung Won Kim, Sang Hyun Choi, Jimi Huh, and Seong Ho Park. 2015. "Systematic Review and Meta-Analysis of Studies Evaluating Diagnostic Test Accuracy: A Practical Review for Clinical Researchers-Part II. Statistical Methods of Meta-Analysis." *Korean Journal of Radiology* 16 (6): 1188–96. https://doi.org/10.3348/kjr.2015.16.6.1188.

Lee, Woojoo, Eun Shin, Bo-Hyung Kim, and Hyunchul Kim. 2019. "Diagnostic Accuracy of SOX11 Immunohistochemistry in Mantle Cell Lymphoma: A Meta-Analysis." *PLoS ONE* 14 (11). https://doi.org/10.1371/journal.pone.0225096.

Leeflang, Mariska MG, Jonathan J Deeks, Yemisi Takwoingi, and Petra Macaskill. 2013. "Cochrane Diagnostic Test Accuracy Reviews." *Systematic Reviews* 2 (October): 82. https://doi.org/10.1186/2046-4053-2-82.

Lesaffre, Emmanuel, and Bart Spiessens. 2001. "On the Effect of the Number of Quadrature Points in a Logistic Random Effects Model: An Example." *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 50 (3): 325–35. https://doi.org/10.1111/1467-9876.00237.

Lê, Sébastien, Julie Josse, and François Husson. 2008. "FactoMineR: A Package for Multivariate Analysis." *Journal of Statistical Software* 25 (1): 1–18. https://doi.org/10.18637/jss.v025.i01.

Li, Huai-Feng, Hai-Jia Mao, Li Zhao, Dan-Ling Guo, Bo Chen, and Jian-Feng Yang. n.d. "The Diagnostic Accuracy of PET(CT) in Patients with Neuroblastoma: A Meta-Analysis and Systematic Review." *Journal of Computer Assisted Tomography* 44 (1): 111–17. https://doi.org/10.1097/RCT.0000000000000973.

Li, Jiangfa, Jiming Wang, Liping Lei, Guandou Yuan, and Songqing He. 2019. "The Diagnostic Performance of Gadoxetic Acid Disodium-Enhanced Magnetic Resonance Imaging and Contrast-Enhanced Multi-Detector Computed Tomography in Detecting Hepatocellular Carcinoma: A Meta-Analysis of

Eight Prospective Studies." *European Radiology* 29 (12): 6519–28. https://doi.org/10.1007/s00330-019-06294-6.

Littenberg, B., and L. E. Moses. n.d. "Estimating Diagnostic Accuracy from Multiple Conflicting Reports: A New Meta-Analytic Method." *Medical Decision Making: An International Journal of the Society for Medical Decision Making* 13 (4): 313–21. https://doi.org/10.1177/0272989X9301300408.

Mavridis, Dimitris. 2018. "Contribution to the Discussion of 'When Should Meta-Analysis Avoid Making Hidden Normality Assumptions?'." *Biometrical Journal* 60 (6): 1081–2. https://doi.org/10.1002/bimj.201800187.

Menke, J. 2010. "Bivariate Random-Effects Meta-Analysis of Sensitivity and Specificity with SAS PROC GLIMMIX." *Methods of Information in Medicine* 49 (1): 54–64. https://doi.org/10.3414/ME09-01-0001.

Millar, Russell B. 2011. *Maximum Likelihood Estimation and Inference: With Examples in R, SAS and ADMB*. 1 edition. Chichester, West Sussex: Wiley.

Moon, Keon-Woong. 2020. *Ztable: Zebra-Striped Tables in Latex and Html Formats*. http://github.com/cardiomoon/ztable.

Pinheiro, José C., Douglas M. Bates, and Jose C. Pinheiro. 1995. "Approximations to the Log-Likelihood Function in the Nonlinear Mixed-Effects Model." *Journal of Computational and Graphical Statistics* 4 (1): 12. https://doi.org/10.2307/1390625.

R Core Team. 2020. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Reitsma, Johannes B., Afina S. Glas, Anne W. S. Rutjes, Rob J. P. M. Scholten, Patrick M. Bossuyt, and Aeilko H. Zwinderman. 2005. "Bivariate Analysis of Sensitivity and Specificity Produces Informative Summary Measures in Diagnostic Reviews." *Journal of Clinical Epidemiology* 58 (10): 982–90. https://doi.org/10.1016/j.jclinepi.2005.02.022.

Riley, Richard D, Keith R Abrams, Alexander J Sutton, Paul C Lambert, and John R Thompson. 2007. "Bivariate Random-Effects Meta-Analysis and the Estimation of Between-Study Correlation." *BMC Medical Research Methodology* 7 (1): 3. https://doi.org/10.1186/1471-2288-7-3.

Röver, Christian, and Tim Friede. 2018. "Contribution to the Discussion of 'When Should Meta-Analysis Avoid Making Hidden Normality Assumptions?' A Bayesian Perspective." *Biometrical Journal* 60 (6): 1068–70. https://doi.org/10.1002/bimj.201800179.

SAS. 2015. *SAS/STAT 14.1. User's Guide*. SAS Institute Inc Cary, NC.

Sheu, Ching-Fan, and Sawako Suzuki. 2001. "Meta-Analysis Using Linear Mixed Models." *Behavior Research Methods, Instruments, & Computers* 33 (2): 102–7. https://doi.org/10.3758/BF03195354.

Spiegelhalter, David J, Keith R Abrams, and Jonathan P Myles. n.d. "Bayesian Approaches to Clinical Trials and Health-Care Evaluation," 408.

Spineli, Loukia M. 2019. "An Empirical Comparison of Bayesian Modelling Strategies for Missing Binary Outcome Data in Network Meta-Analysis." *BMC Medical Research Methodology* 19 (April). https://doi.org/10.1186/s12874-019-0731-y.

Spineli, Loukia M., and Nikolaos Pandis. 2020. "Statistical Heterogeneity: Notion and Estimation in Meta-Analysis." *American Journal of Orthodontics and Dentofacial Orthopedics* 157 (6): 856–859.e2. https://doi.org/10.1016/j.ajodo.2020.03.009.

Stroup, Walter W., George A. Milliken, Elizabeth A. Claassen, and Russell D. Wolfinger. 2018. *SAS for Mixed Models: Introduction and Basic Applications*. SAS Institute. http://books.google.com?id=j_eCDwAAQBAJ.

Takwoingi, Yemisi, Boliang Guo, Richard D Riley, and Jonathan J Deeks. 2017. "Performance of Methods for Meta-Analysis of Diagnostic Test Accuracy with Few Studies or Sparse Data." *Statistical Methods in Medical Research* 26 (4): 1896–1911. https://doi.org/10.1177/0962280215592269.

Tennekes, Martijn. 2017. *Treemap: Treemap Visualization*. https://CRAN.R-project.org/package=treemap.

Thiele, Christian. 2020. *Cutpointr: Determine and Evaluate Optimal Cutpoints in Binary Classification Tasks*. https://CRAN.R-project.org/package=cutpointr.

Tsou, Po-Yang, Yu-Hsun Wang, Yu-Kun Ma, Julia K. Deanehan, Jason Gillon, Eric H. Chou, Tzu-Chun Hsu, Yuan-Chun Huang, Judy Lin, and Chien-Chang Lee. 2019. "Accuracy of Point-of-Care Ultrasound and Radiology-Performed Ultrasound for Intussusception: A Systematic Review and Meta-Analysis." *The American Journal of Emergency Medicine* 37 (9): 1760–9. https://doi.org/10.1016/j.ajem.2019.06.006.

Verbeke, Geert, and Geert Molenberghs. 2000. *Linear Mixed Models for Longitudinal Data*. Springer Series in Statistics. New York: Springer-Verlag. https://doi.org/10.1007/978-1-4419-0300-6.

Verde, Pablo. 2018. "Bamdit : An R Package for Bayesian Meta-Analysis of Diagnostic Test Data." *Journal of Statistical Software* 86 (September). https://doi.org/10.18637/jss.v086.i10.

Verde, Pablo E. 2008. "Meta-Analysis of Diagnostic Test Data: Modern Statistical Approaches."

———. 2019. "The Hierarchical Metaregression Approach and Learning from Clinical Evidence." *Biometrical Journal* 61 (3): 535–57. https://doi.org/10.1002/bimj.201700266.

———. 2010. "Meta-Analysis of Diagnostic Test Data: A Bivariate Bayesian Modeling Approach." *Statistics in Medicine* 29 (30): 3088–3102. https://doi.org/10.1002/sim.4055.

Veroniki, Areti Angeliki, Dan Jackson, Wolfgang Viechtbauer, Ralf Bender, Jack Bowden, Guido Knapp, Oliver Kuss, Julian PT Higgins, Dean Langan, and Georgia Salanti. 2016. "Methods to Estimate the Between-Study Variance and Its Uncertainty in Meta-Analysis." *Research Synthesis Methods* 7 (1): 55–79. https://doi.org/10.1002/jrsm.1164.

Wei, Ming-Zhu, Zhen-Hua Zhao, and Jian-Yun Wang. n.d. "The Diagnostic Accuracy of Magnetic Resonance Imaging in Restaging of Rectal Cancer After Preoperative Chemoradiotherapy: A Meta-Analysis and Systematic Review." *Journal of Computer Assisted Tomography* 44 (1): 102–10. https://doi.org/10.1097/RCT.0000000000000964.

Wickham, Hadley. 2016. *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org.

Wolfinger, Russell D. n.d. "Fitting Nonlinear Mixed Models with the New NLMIXED Procedure," 10.

Yoon, Seo Hee, Eun Hwa Kim, Ha Yan Kim, and Jong Gyun Ahn. 2019. "Presepsin as a Diagnostic Marker of Sepsis in Children and Adolescents: A Systemic Review and Meta-Analysis." *BMC Infectious Diseases* 19 (1): 760. https://doi.org/10.1186/s12879-019-4397-1.

Zamora, Javier, Victor Abraira, Alfonso Muriel, Khalid Khan, and Arri Coomarasamy. 2006. "Meta-DiSc: A Software for Meta-Analysis of Test Accuracy Data." *BMC Medical Research Methodology* 6 (July): 31. https://doi.org/10.1186/1471-2288-6-31.

Zheng, Qiang, Weibiao Kang, Changyu Chen, Xinxin Shi, Yang Yang, and Changjun Yu. 2019. "Diagnosis Accuracy of Raman Spectroscopy in Colorectal Cancer." *Medicine* 98 (34). https://doi.org/10.1097/MD.0000000000016940.

Zhou, Yan, and Nandini Dendukuri. 2014. "Statistics for Quantifying Heterogeneity in Univariate and Bivariate Meta-Analyses of Binary Data: The Case of Meta-Analyses of Diagnostic Accuracy: Y. ZHOU AND N. DENDUKURI." *Statistics in Medicine* 33 (16): 2701–17. https://doi.org/10.1002/sim.6115.

Zhu, Yuzhou, Hao Zhang, Nan Chen, Jianqi Hao, Hongyu Jin, and Xuelei Ma. 2020. "Diagnostic Value of Various Liquid Biopsy Methods for Pancreatic Cancer: A Systematic Review and Meta-Analysis." *Medicine* 99 (3): e18581. https://doi.org/10.1097/MD.0000000000018581.