# Meta-Analysis of Diagnostic Accuracy with mada

**Philipp Doebler**
TU Dortmund University

**Heinz Holling**
WWU Münster

**Bernardo Sousa-Pinto**
University of Porto

**Abstract**

The R-package **mada** is a tool for the meta-analysis of diagnostic accuracy. In contrast to univariate meta-analysis, diagnostic meta-analysis requires bivariate models. An additional challenge is to provide a summary receiver operating characteristic curves that seek to integrate receiver operator characteristic curves of primary studies. The package implements the approach of Reitsma, Glas, Rutjes, Scholten, Bossuyt, and Zwinderman (2005), which in the absence of covariates is equivalent to the HSROC model of Rutter and Gatsonis (2001). More recent models by Doebler, Holling, and Böhning (2012) and Holling, Böhning, and Böhning (2012b) are also available, including meta-regression for the first approach. In addition a range of functions for descriptive statistics and graphics are provided.

*Keywords*: diagnostic meta-analysis, multivariate statistics, summary receiver operating characteristic, R.

## 1. Introduction

While substantial work has been conducted on methods for diagnostic meta-analysis, it has not become a routine procedure yet. One of the reasons for this is certainly the complexity of bivariate approaches, but another reason is that standard software packages for meta-analysis, for example **Comprehensive Meta-Analysis** and **RevMan** (Biostat, Inc. 2006; The Nordic Cochrane Centre 2011), do not include software to fit models appropriate for diagnostic meta-analysis. For the recommended (Leeflang, Deeks, Gatsonis, and Bossuyt 2008) bivariate approach of Rutter and Gatsonis (2001) meta-analysts can use Bayesian approaches (for example in WinBUGS (Lunn, Thomas, Best, and Spiegelhalter 2000) or OpenBUGS (Lunn, Spiegelhalter, Thomas, and Best 2009)), the stata module **metandi** (Harbord and Whiting 2010), or the SAS macro **METADAS** (Takwoingi and Deeks 2011). So currently available software is either relatively complex (WinBUGS/OpenBUGS) or proprietary (stata, SAS).

The open source package **mada** written in R (R Core Team 2012) provides some established and some current approaches to diagnostic meta-analysis, as well as functions to produce descriptive statistics and graphics. It is hopefully complete enough to be the only tool needed for a diagnostic meta-analysis. **mada** has been developed with an R user in mind that has used standard model fitting functions before, and a lot of the output of **mada** will look familiar to such a user. While this paper cannot provide an introduction to R, it is hopefully detailed enough to provide a novice R user with enough hints to perform diagnostic meta-analysis along the lines of it. Free introductions to R are for example available on the homepage of the R project. We assume that the reader is familiar with central concepts of meta-analysis,

like fixed and random effects models (for example Borenstein, Hedges, Higgins, and Rothstein 2009) and ideas behind diagnostic accuracy meta-analysis and (S)ROC curves (starting points could be Sutton, Abrams, Jones, Sheldon, and Song 2000; Walter 2002; Jones and Athanasiou 2005; Leeflang *et al.* 2008).

## 2. Obtaining the package

Once R is installed and an internet connection is available, the package can be installed from CRAN on most systems by typing

```
R> install.packages("mada")
```

Development of **mada** is hosted at `http://r-forge.r-project.org/projects/mada/`; the most current version is available there[1], while only stable versions are available from CRAN. The package can then be loaded:

```
R> library("mada")
```

## 3. Entering data

Primary diagnostic studies observe the result of a *gold standard* procedure which defines the presence or absence of a *condition*, and the result of a *diagnostic test* (typically some kind of low cost procedure, or at least one that is less invasive than the gold standard). Data from such a primary study could be reported in a $2 \times 2$ table, see Table 1.

Table 1: Data from the $i$th study in a $2 \times 2$ table.

| | with condition | without condition |
|---|---|---|
| Test positive | $y_i$ | $z_i$ |
| Test negative | $m_i - y_i$ | $n_i - z_i$ |
| Total | $m_i$ | $n_i$ |

The numbers $y_i$ and $z_i$ are the numbers of true-positives (TP) and false positives (FP), respectively, and $m_i - y_i$ and $n_i - z_i$ are the numbers of false negatives (FN) and true negatives (TN). Often derived measures of diagnostic accuracy are calculated from $2 \times 2$ tables. Using the notation in Table 1, one can calculate

$$p_i = \text{sensitivity of } i\text{th study} = \frac{y_i}{m_i} \tag{1}$$

$$u_i = \text{false positive rate of } i\text{th study} = \frac{z_i}{n_i} \tag{2}$$

$$1 - u_i = \text{specificity of } i\text{th study} = \frac{n_i - z_i}{n_i}. \tag{3}$$

[1]For example by typing `install.packages("mada", repos="http://R-Forge.R-project.org")` at an R prompt.

Basically all functions in the **mada** package need data from 2×2 tables. One can use R to calculate the table given specificities or sensitivities if the sample size in each group is known (sometimes there is insufficient data to reconstruct the 2×2 table). The above formulae for the sensitivity for example implies that

$$y_i = m_i p_i.$$

If a primary study reports a sensitivity of .944 and that there were 142 people with the condition, we can calculate $y$ by

```
R> y <- 142 * .944
R> y

[1] 134.048
```

Since this is not an integer, we need to round it to the nearest integer

```
R> round(y)

[1] 134
```

Note that **mada** is a bit paranoid about the input: it demands that the data and the rounded data are identical to prevent some obvious error. Hence the use of the `round` function should not be omitted.

Let us now assume that the number of TP, FP, FN and TN is known for each primary study. A good way to organise information in R is to use *data frames*, which can hold different variables. In our case each row of the data frame corresponds to one primary study. As an example we enter the data from six studies from a meta-analysis of the AUDIT-C (a short screening test for alcohol problems, Kriston, Hölzel, Weiser, Berner, and Härter 2008) into a data frame

```
R> AuditC6 <- data.frame(TP = c(47, 126, 19, 36, 130, 84),
+                        FN = c(9, 51, 10, 3, 19, 2),
+                        FP = c(101, 272, 12, 78, 211, 68),
+                        TN = c(738, 1543, 192, 276, 959, 89))
R> AuditC6

   TP FN  FP   TN
1  47  9 101  738
2 126 51 272 1543
3  19 10  12  192
4  36  3  78  276
5 130 19 211  959
6  84  2  68   89
```

Note that many central functions in **mada** also accept four vectors of frequencies (TP, FN, FP, TN) as input. Nevertheless, it is convenient to store not only the observed frequencies, but also the study names in the same data frame. The following command shows how to do this for our shortened example:

```
R> AuditC6$names <- c("Study 1", "Study 2", "Study 4",
+                     "Study 4", "Study 5", "Study 6")
```

The full data set with 14 studies is part of **mada**; let's load the data set and have a look at the last six studies:

```
R> data("AuditC")
R> tail(AuditC)
```

```
    TP FN  FP   TN
9   59  5  55  136
10 142 50 571 2788
11 137 24 107  358
12  57  3 103  437
13  34  1  21   56
14 152 51  88  264
```

In the following we will use the `AuditC` data set as a running example.

### 3.1. Zero cells

In the analysis of data in 2×2 tables zero cells often lead to problems or statistical artefacts since certain ratios do not exist. So called *continuity corrections* are added to the observed frequencies; these are small positive numbers. One suggestions in the literature is to use 0.5 as the continuity correction, which is the default value in **mada**. All relevant functions in **mada** allow user specified continuity corrections and the correction can be applied to all studies, or just to those with zero cells.

## 4. Descriptive statistics

Descriptive statistics for a data set include the sensitivity, specificity and false-positive rate of the primary studies and also their positive and negative likelihood ratios ($\mathrm{LR}_+, \mathrm{LR}_-$), and their diagnostic odds ratio (DOR; Glas, Lijmer, Prins, Bonsel, and Bossuyt 2003). These are defined as

$$\mathrm{LR}_+ = \frac{p}{u} = \frac{\text{sensitivity}}{\text{false positive rate}},$$

$$\mathrm{LR}_- = \frac{1-p}{1-u},$$

and

$$\mathrm{DOR} = \frac{\mathrm{LR}_+}{\mathrm{LR}_-} = \frac{\mathrm{TP}\cdot\mathrm{TN}}{\mathrm{FN}\cdot\mathrm{FP}}.$$

All these are easily computed using the `madad` function, together with their confidence intervals. We use the formulae provided by Deeks (2001). `madad` also performs $\chi^2$ tests to assess heterogeneity of sensitivities and specificities, the null hypothesis being in both cases, that all are equal. Finally the correlation of sensitivities and false positive rates is calculated to give a hint whether the cut-off value problem is present. The following output is slightly cropped.

```
R> madad(AuditC)

Descriptive summary of AuditC with 14 primary studies.
Confidence level for all calculations set to 95 %
Using a continuity correction of 0.5 if applicable

Diagnostic accuracies
       sens  2.5% 97.5%  spec  2.5% 97.5%
 [1,] 0.833 0.716 0.908 0.879 0.855 0.899
 [2,] 0.711 0.640 0.772 0.850 0.833 0.866
...
[14,] 0.748 0.684 0.802 0.749 0.702 0.792

Test for equality of sensitivities:
X-squared = 272.3603, df = 13, p-value = <2e-16
Test for equality of specificities:
X-squared = 2204.8, df = 13, p-value = <2e-16

Diagnostic OR and likelihood ratios
           DOR   2.5%     97.5%  posLR  2.5%  97.5% negLR  2.5% 97.5%
 [1,]   36.379 17.587    75.251  6.897 5.556  8.561 0.190 0.106 0.339
...
[14,]    8.850  5.949    13.165  2.982 2.448  3.632 0.337 0.264 0.430

Correlation of sensitivities and false positive rates:
   rho  2.5 % 97.5 %
 0.677  0.228  0.888
```

The `madad` function has a range of options with respect to computational details; for example one can compute 80% confidence intervals:

```
R> madad(AuditC, level = 0.80)
```

Also note that all the output of `madad` is available for further computations if one assigns the output of `madad` to an object. For example the false positive rates with their confidence intervals can be extracted using the `$` construct (output cropped):

```
R> AuditC.d <- madad(AuditC)
R> AuditC.d$fpr

$fpr
 [1] 0.12083333 0.15005507 0.06097561 0.22112676 0.18061486 0.43354430
 [7] 0.20988806 0.52006770 0.28906250 0.17008929 0.23068670 0.19131238
[13] 0.27564103 0.25070822

$fpr.ci
            2.5%     97.5%
```

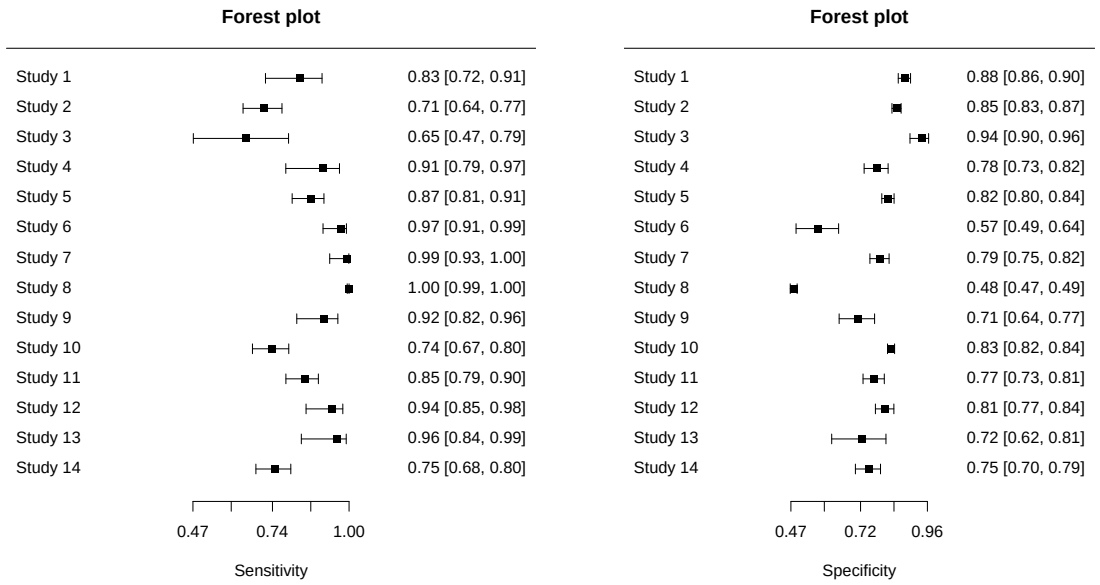


Figure 1: Paired forest plot for AUDIT-C data.

```
 [1,] 0.10050071 0.1446182
...
[14,] 0.20834216 0.2984416
```

### 4.1. Descriptive graphics

For the AUDIT-C data, the $\chi^2$ tests already suggested heterogeneity of sensitivities and specificities. The corresponding *forest plots* confirm this:

```
R> forest(madad(AuditC), type = "sens")
R> forest(madad(AuditC), type = "spec")
```

These plots are shown in Figure 1.

Apart from these univariate graphics **mada** provides a variety of plots to study the data on ROC space. Note that for exploratory purposes it is often useful to employ color and other features ofR's plotting system. Two high level plots are provided by **mada**: `crosshair` to produce crosshair plots (Phillips, Stewart, and Sutton 2010), and `ROCellipse`. The following is an example of a call of `crosshair` that produces (arbitrarily) colored crosshairs and makes the crosshairs wider with increased sample size; also only a portion of ROC space is plotted.

```
R> rs <- rowSums(AuditC)
R> weights <- 4 * rs / max(rs)
R> crosshair(AuditC, xlim = c(0,0.6), ylim = c(0.4,1),
+            col = 1:14, lwd = weights)
```

Figure 2 displays this plot and the next descriptive plot: `ROCellipse` plots confidence regions which describe the uncertainty of the pair of sensitivity and false positive rate. These regions

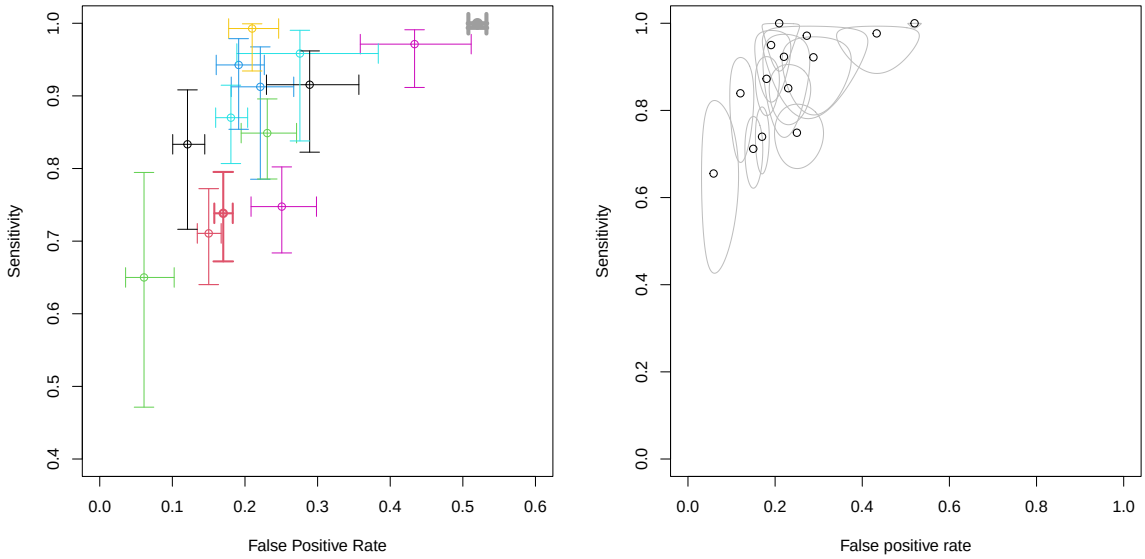


Figure 2: A "weighted" crosshair plot with (arbitrary) coloring and a plot with confidence regions for primary study estimates.

are ellipses on logit ROC space, and by back-transforming them to regular ROC space the (sometimes oddly shaped) regions are produced. By default this function will also plot the point estimates. The following example is a bit contrived, but showcases the flexibility of `ROCellipse`: here the plotting of the point estimates is suppressed manipulating the `pch` argument, but then points are added in the next step.

```
R> ROCellipse(AuditC, pch = "")
R> points(fpr(AuditC), sens(AuditC))
```

## 5. Univariate approaches

Before the advent of the bivariate approaches by Rutter and Gatsonis (2001) and Reitsma *et al.* (2005), some univariate approaches to the meta-analysis of diagnostic accuracy were more popular. Bivariate approaches cannot be recommended if the sample size is too small. The bivariate model of Reitsma *et al.* (2005) for example has 5 parameters, which would clearly be too much for a handful of studies. Hence **mada** provides some univariate methods. Since pooling sensitivities or specificities can be misleading (Gatsonis and Paliwal 2006), options for the univariate meta-analysis of these are not provided. **mada** does provide approaches for the DOR (Glas *et al.* 2003), the positive and negative likelihood ratios, and $\theta$, the accuracy parameter of the proportional hazards model for diagnostic meta-analysis (Holling *et al.* 2012b). In this vignette we explain the details on the DOR methodology and the methods for $\theta$.

### 5.1. Diagnostic odds ratio

In analogy to the meta-analysis of the odds ratio (OR) methods for the meta-analysis of the DOR can be developed (Glas *et al.* 2003). For the *fixed effects* case a Mantel-Haenszel (MH; see for example Deeks 2001) is provided by **mada**. The underlying fixed effects model has the form

$$\mathrm{DOR}_i = \mu + \epsilon_i,$$

where $\mu$ is true underlying DOR and the $\epsilon_i$ are independent errors with mean 0 and study specific variance. The MH estimator is a weighted average of DORs observed in the primary studies and is robust to the presence of zero cells. It takes the form

$$\hat{\mu} = \sum_i \frac{\omega_i^{MH} \mathrm{DOR}_i}{\sum_i \omega_i^{MH}},$$

where $\omega_i^{MH} = \frac{z_i(m_i - y_i)}{m_i + n_i}$ are the Mantel-Haenszel weights.

One obtains an estimator for a *random effects* model following the approach of DerSimonian and Laird (DSL; DerSimonian and Laird 1986). Here the underlying model is in terms of the log DORs. One assumes

$$\log \mathrm{DOR}_i = \mu + \epsilon_i + \delta_i,$$

where $\mu$ is the mean of the log DORs, $\epsilon_i$ and $\delta_i$ are independent with mean 0; the variance $\sigma_i^2$ of $\epsilon_i$ is estimated as

$$\hat{\sigma}_i^2 = \frac{1}{y_i} + \frac{1}{m_i - y_i} + \frac{1}{z_i} + \frac{1}{n_i - z_i},$$

and the variance $\tau^2$ of $\delta_i$ is to be estimated. The DSL estimator then is a weighted estimator, too:

$$\hat{\mu} = \sum_i \frac{\omega_i^{DSL} \mathrm{DOR}_i}{\sum_i \omega_i^{DSL}},$$

where

$$\omega_i^{DSL} = \frac{1}{\hat{\sigma}_i^2 + \tau^2}.$$

The variance $\tau^2$ is estimated by the Cochran $Q$ statistic trick.

The function `madauni` handles the meta-analysis of the DOR (and the negative and positive likelihood ratios). One can use `madauni` in the following fashion:

```
R> (fit.DOR.DSL <- madauni(AuditC))

Call:
madauni(x = AuditC)

   DOR  tau^2
26.337  0.311

R> (fit.DOR.MH <- madauni(AuditC, method = "MH"))
```

```
Call:
madauni(x = AuditC, method = "MH")

     DOR
17.93335
```

Note that the brackets around `fit.DOR.DSL <- madauni(AuditC)` are a compact way to print the fit. The `print` method for `madauni` objects is not very informative, only the point estimate is returned along with (in the random effects case) an estimate of the $\tau^2$, the variance of the random effects. Note that estimation in the random effects case is performed on log-DOR scale, so that $\tau^2$ of the above DSL fit is substantial. To obtain more information the `summary` method can be used:

```
R> summary(fit.DOR.DSL)

Call:
madauni(x = AuditC)

Estimates:
      DSL estimate  2.5 % 97.5 %
DOR         26.337 17.971 38.596
lnDOR        3.271  2.889  3.653
tau^2        0.311  0.000  3.787
tau          0.557  0.000  1.946

Cochran's Q: 19.683 (13 df, p = 0.103)
Higgins' I^2: 33.955%
```

In addition to the confidence intervals, Cochran's $Q$ statistic (Cochran 1954) can be seen and Higgins $I^2$ (Higgins, Thompson, Deeks, and Altman 2003). Producing a forest plot of the (log-)DOR values together with the summary estimate is straightforward using the `forest` method for the `madauni` class:

```
R> forest(fit.DOR.DSL)
```

The resulting plot is shown in Figure 3.

## 5.2. Proportional hazards model approach

The proportional hazards model approach (PHM; see Holling *et al.* 2012b) builds on the assumption of a simple form of the ROC curves. The so called *Lehmann model* (Le 2006) is assumed. Let $p_i$ and $u_i$ denote the $i$th study's sensitivity and false positive rate respectively. The relationship of $p_i$ and $u_i$ is then assumed to be

$$p_i = u_i^{\theta_i},$$

where $\theta_i > 0$ is a diagnostic accuracy parameter. The smaller $\theta$, the larger the area under the ROC curve and thus the more accurate the diagnostic test. For the meta-analysis of $\theta$

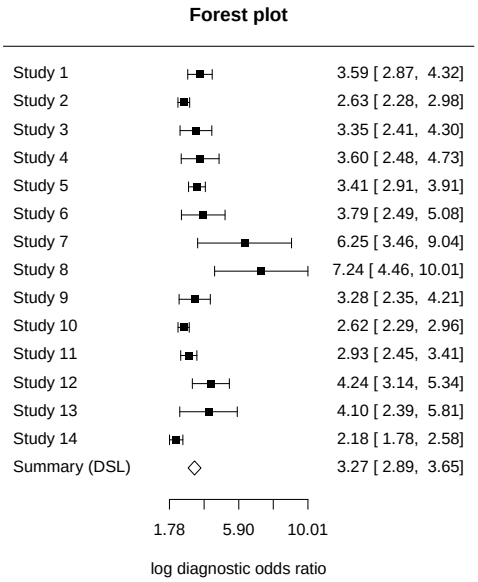


Figure 3: Forest plot for a univariate random effects meta-analysis of the AUDIT-C data using the diagnostic odds ratio.

the APMLE estimator is implemented in **mada** for the case of homogeneity (i.e., fixed effects) and heterogeneity (i.e., random effects). Again the standard output of the `phm` function is rather sparse:

```
R> (fit.phm.homo <- phm(AuditC, hetero = FALSE))

Call:
phm.default(data = AuditC, hetero = FALSE)

Coefficients:
      theta
0.004586893

R> (fit.phm.het <- phm(AuditC))

Call:
phm.default(data = AuditC)

Coefficients:
      theta     taus_sq
0.084631351 0.003706143
```

The `summary` method is more informative:

```
R> summary(fit.phm.homo)
```

```
Call:
phm.default(data = AuditC, hetero = FALSE)

         Estimate       2.5 %     97.5 %
theta 0.004586893 0.003508507 0.00566528

Log-likelihood: -61.499 on 1 degrees of freedom
AIC:  125
BIC:  125.6

        Chi-square goodness of fit test (Adjusted Profile Maximum
        Likelihood under homogeneity)

data:  x
Chi-square = 222.47, df = 1, p-value < 2.2e-16


   AUC  2.5 % 97.5 %   pAUC  2.5 % 97.5 %
 0.995  0.997  0.994  0.994  0.995  0.992
```

The $\chi^2$ test goodness of fit test rejects the assumption of homogeneity, but the fit of the model for heterogeneity is better:

```
R> summary(fit.phm.het)

Call:
phm.default(data = AuditC)

           Estimate        2.5 %      97.5 %
theta   0.084631351  0.047449859 0.121812844
taus_sq 0.003706143 -0.001277798 0.008690085

Log-likelihood: 31.121 on 2 degrees of freedom
AIC:  -58.2
BIC:  -57

        Chi-square goodness of fit test (Adjusted Profile Maximum
        Likelihood under heterogeneity)

data:  x
Chi-square = 13.726, df = 2, p-value = 0.3185


   AUC  2.5 % 97.5 %   pAUC  2.5 % 97.5 %
 0.922  0.955  0.891  0.891  0.937  0.848
```

The estimation of $\theta$ results in an SROC curve; plotting this curve together with confidence bands obtained from the confidence interval of $\theta$ in the summary is done with the `plot`

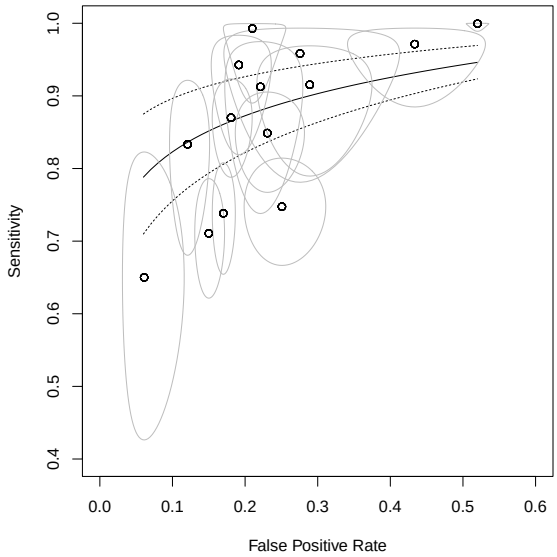


Figure 4: Summary plot for the analysis of the AUDIT-C data with the PHM model.

method. We also add the original data on ROC space with confidence regions and only plot a portion of ROC space.

```
R> plot(fit.phm.het, xlim = c(0,0.6), ylim = c(0.4,1))
R> ROCellipse(AuditC, add = TRUE)
```

The resulting plot is shown in Figure 4.

Note that the SROC curve is not extrapolated beyond the range of the original data. The area under the SROC curve, the AUC, is also part of the summary above. For the PHM it is calculated by

$$\text{AUC} = \frac{1}{\theta + 1},$$

and by the same relation a confidence interval for the AUC can be computed from the confidence interval for $\theta$. The **mada** package also offers the `AUC` function to calculate the AUC of other SROC curves which uses the trapezoidal rule.

## 6. A bivariate approach

Typically the sensitivity and specificity of a diagnostic test depend on each other through a cut-off value: as the cut-off is varied to, say, increase the sensitivity, the specificity often decreases. So in a meta-analytic setting one will often observe (negatively) correlated sensitivities and specificities. This observation can (equivalently) also be stated as a (positive) correlation of sensitivities and false positive rates. Since these two quantities are interrelated, bivariate approaches to the meta-analysis of diagnostic accuracy have been quite successful (Rutter and Gatsonis 2001; Van Houwelingen, Arends, and Stijnen 2002; Reitsma *et al.*

2005; Harbord, Deeks, Egger, Whiting, and Sterne 2007; Arends, Hamza, Van Houwelingen, Heijenbrok-Kal, Hunink, and Stijnen 2008).

One typically assumes a binomial model conditional on a primary studies true sensitivity and false positive rates, and a bivariate normal model for the logit-transformed pairs of sensitivities and false positive rates. There are two ways to cast the final model: as a non-linear mixed model or as linear mixed model (see for example Arends *et al.* 2008). The latter approach is implemented in **mada**'s `reitsma` function, so we give some more details. We note that more generally the following can be seen as a multivariate meta-regression and so the the package `mvmeta` (Gasparrini, Armstrong, and Kenward 2012) serves as a basis for our implementation.

Let $p_i$ and $u_i$ denote the $i$th study's true sensitivity and false positive rate respectively, and let $\hat{p}_i$ and $\hat{u}_i$ denote their estimates from the observed frequencies. Then, since a binomial model is assumed conditional on the true $p_i$, the variance of $\text{logit}(\hat{p}_i)$ can be approximated[2] by

$$\frac{1}{m_i \hat{p}_i (1 - \hat{p}_i)},$$

and the variance of $\text{logit}(\hat{u}_i)$ is then

$$\frac{1}{n_i \hat{u}_i (1 - \hat{u}_i)}.$$

So on the within study level one assumes, conditional on $p_i$ and $u_i$, that the observed variation is described by these variances and a normal model; let $D_i$ denote a diagonal $2 \times 2$ matrix with the two variances on the diagonal. On the study level, one assumes that a global mean

$$\mu = (\mu_1, \mu_2)^\mathsf{T}$$

and covariance matrix

$$\Sigma = \begin{pmatrix} \sigma_1^2 & \sigma \\ \sigma & \sigma_2^2 \end{pmatrix}$$

describe the heterogeneity of the pairs $(\text{logit}(p_i), \text{logit}(u_i))$. So the model for the $i$th study is then

$$(\text{logit}(\hat{p}_i), \text{logit}(\hat{u}_i))^\mathsf{T} \sim \text{N}(\mu, \Sigma + D_i).$$

Fitting this model in **mada** is similar to the other model fitting functions:

```
R> (fit.reitsma <- reitsma(AuditC))

Call:  reitsma.default(data = AuditC)

Fixed-effects coefficients:
              tsens     tfpr
(Intercept)  2.0997  -1.2637

14 studies, 2 fixed and 3 random-effects parameters
  logLik       AIC       BIC
 31.5640  -53.1279  -46.4669
```

[2] This uses the delta method.

The `print` method for `reitsma` objects has a scarce output. More information is offered by the `summary` method:

```
R> summary(fit.reitsma)

Call:  reitsma.default(data = AuditC)

Bivariate diagnostic random-effects meta-analysis
Estimation method: REML

Fixed-effects coefficients
                  Estimate Std. Error      z Pr(>|z|) 95%ci.lb
tsens.(Intercept)    2.100      0.338  6.215    0.000    1.438
tfpr.(Intercept)    -1.264      0.174 -7.249    0.000   -1.605
sensitivity          0.891          -      -        -    0.808
false pos. rate      0.220          -      -        -    0.167
                  95%ci.ub
tsens.(Intercept)    2.762 ***
tfpr.(Intercept)    -0.922 ***
sensitivity          0.941
false pos. rate      0.285
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Variance components: between-studies Std. Dev and correlation matrix
      Std. Dev tsens  tfpr
tsens    1.175 1.000     .
tfpr     0.638 0.854 1.000

 logLik     AIC     BIC
 31.564 -53.128 -46.467

AUC:  0.887
Partial AUC (restricted to observed FPRs and normalized):  0.861

I2 estimates
Zhou and Dendukuri approach:  40.4 %
Holling sample size unadjusted approaches:  35.6 - 79.3 %
Holling sample size adjusted approaches:  0.2 - 2.4 %
```

Note the sensitivity and false positive rate returned in this summary are just the back-transformed $\mu_1$ and $\mu_2$. In addition, please note that the methodology for estimating the I2 in the context of diagnostic test accuracy meta-analyses is not yet fully established. The obtained estimates are based on the approach described by Zhou and Dendukuri (2014) for bivariate meta-analysis and based on the approaches described by Holling, Bohning, Masoudi, Bohning, and Sangnawakij (2019). For the latter, we compute both sample size-unadjusted and adjusted estimated according to all provided formulae for computing within-study variability. One can

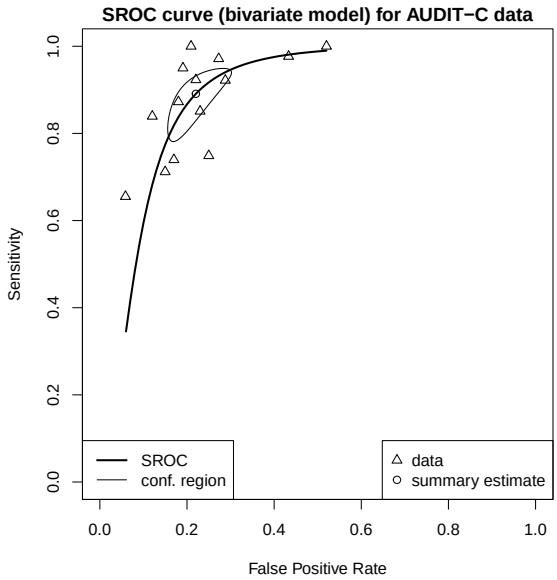


Figure 5: SROC curve for the Reitsma *et al.* (2005) model.

then proceed to plot the SROC curve of this model. By default the point estimate of the pair of sensitivity and false positive rate is also plotted together with a confidence region. In the following example the SROC curve is plotted a bit thicker using the `sroclwd` argument, a caption is added to the plot and also the data and a legend. By default the SROC curve is not extrapolated beyond the range of the original data:

```
R> plot(fit.reitsma, sroclwd = 2,
+       main = "SROC curve (bivariate model) for AUDIT-C data")
R> points(fpr(AuditC), sens(AuditC), pch = 2)
R> legend("bottomright", c("data", "summary estimate"), pch = c(2,1))
R> legend("bottomleft", c("SROC", "conf. region"), lwd = c(2,1))
```

The output is shown in Figure 5.

## 6.1. Comparing SROC curves

We show how to compare SROC curves. Patrick, Cheadle, Thompson, Diehr, Koepsell, and Kinne (1994) conducted a meta-analysis to (among other things) investigate the efficacy of self administered and interviewer administered questionnaires to detect nicotine use. The data sets `SAQ` and `IAQ` are the respective subsets of this data. First one fits bivariate models to the data sets:

```
R> data("IAQ")
R> data("SAQ")
R> # both datasets contain more than one 2x2-table per study
R> # reduce (somewhat arbitrarily) to one row per study by
R> # using the first coded table only:
```

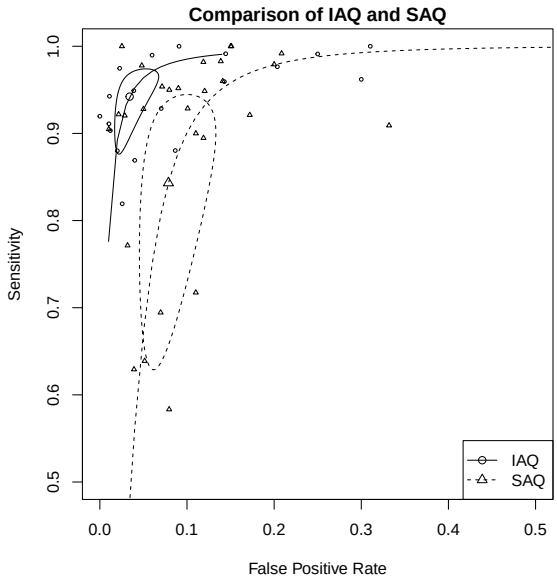


Figure 6: Comparison of interviewer and self-adminstered smoking questionaires with SROC curves.

```
R> IAQ1 <- subset(IAQ, IAQ$result_id == 1)
R> SAQ1 <- subset(SAQ, SAQ$result_id == 1)
R> fit.IAQ <- reitsma(IAQ1)
R> fit.SAQ <- reitsma(SAQ1)
```

Then one plots the SROC curves of these fits, beginning with the fit of the IAQ and adding the SAQ curve. Note that the `lty` arguments is used so that the curves can be distinguished.

```
R> plot(fit.IAQ, xlim = c(0,.5), ylim = c(.5,1),
+       main = "Comparison of IAQ and SAQ")
R> lines(sroc(fit.SAQ), lty = 2)
R> ROCellipse(fit.SAQ, lty = 2, pch = 2, add = TRUE)
R> points(fpr(IAQ1), sens(IAQ1), cex = .5)
R> points(fpr(SAQ1), sens(SAQ1), pch = 2, cex = 0.5)
R> legend("bottomright", c("IAQ", "SAQ"), pch = 1:2, lty = 1:2)
```

Figure 6 contains the resulting plot. The summary estimates are well separated, though the confidence regions slightly overlap. It would nevertheless be safe to conclude that IAQ is a more reliable way to measure smoking than SAQ.

## 6.2. Bivariate meta-regression

We demonstrate diagnostic meta-regression also using the data of Patrick *et al.* (1994). We use the complete data set, which is loaded by

```
R> data("smoking")
```

```
R> # again reduce to one result per study:
R> smoking1 <- subset(smoking, smoking$result_id == 1)
```

The `data.frame` contains the same variables as the `SAQ` and `IAQ` subsets, but the type is coded by the variable `type`:

```
R> summary(smoking1$type)

IAQ SAQ
 10  16
```

We use the factor `type` as a covariate in diagnostic meta-regression:

```
R> fit.smoking.type <- reitsma(smoking1,
+                              formula = cbind(tsens, tfpr) ~ type)
```

Note that the left hand side of the `formula` object always has to be of the form `cbind(tsens, tfpr)`, where `tsens` and `tfpr` are for *transformed* sensitivity and false positive rate respectively. We generate detailed output by:

```
R> summary(fit.smoking.type)

Call:  reitsma.default(data = smoking1, formula = cbind(tsens, tfpr) ~
    type)

Bivariate diagnostic random-effects meta-analysis
Estimation method: REML

Fixed-effects coefficients
                  Estimate Std. Error       z Pr(>|z|) 95%ci.lb
tsens.(Intercept)    2.813      0.491   5.735    0.000    1.852
tsens.typeSAQ       -1.166      0.634  -1.838    0.066   -2.409
tfpr.(Intercept)    -3.337      0.311 -10.733    0.000   -3.946
tfpr.typeSAQ         0.882      0.389   2.269    0.023    0.120
                  95%ci.ub
tsens.(Intercept)    3.775 ***
tsens.typeSAQ        0.077   .
tfpr.(Intercept)    -2.727 ***
tfpr.typeSAQ         1.645   *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Variance components: between-studies Std. Dev and correlation matrix
      Std. Dev tsens  tfpr
tsens    1.508 1.000     .
tfpr     0.875 0.551 1.000
```

```
  logLik      AIC      BIC
  70.721 -127.441 -113.783


I2 estimates
Zhou and Dendukuri approach:  66 %
Holling sample size unadjusted approaches:  83.7 - 96.8 %
Holling sample size adjusted approaches:  5.3 - 12.7 %
```

This output can be interpreted as follows: The $z$ value for the regression coefficient for the false-positive rates is significant, indicating that the interviewer administered questionnaires offer a better false-positive rate (the coefficient for the difference in false-positive rate for SAQ is positive, so the false positive rates are higher for the SAQ and, hence, lower for the IAQ). Interestingly the point estimate for the sensitivities does not indicate any effect.

Note that once meta-regression is used, one cannot reasonably plot SROC curves, since fixed values for the covariates would have to be supplied to do so. Also (global) AUC values do not make sense.

We can also compare the fit of two bivariate meta-regressions with a likelihood-ratio test. For this, we have to refit the models with the maximum likelihood method, as the likelihood-ratio test relies on asymptotic theory that is only valid if this estimation method is employed.

```
R> fit.smoking.ml.type <- reitsma(smoking1,
+                             formula = cbind(tsens, tfpr) ~ type,
+                             method = "ml")
R> fit.smoking.ml.intercept <- reitsma(smoking1,
+                                      formula = cbind(tsens, tfpr) ~ 1,
+                                      method = "ml")
R> anova(fit.smoking.ml.type, fit.smoking.ml.intercept)

Likelihood-ratio test
Model 1: cbind(tsens, tfpr) ~ type
Model 2: cbind(tsens, tfpr) ~ 1

 ChiSquared Df Pr(>ChiSquared)
      13.25  2         0.00133 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The meta-regression confirms that type explains some of the heterogeneity between the primary studies.

### 6.3. Transformations beyond the logit

All bivariate approaches explained so far use the conventional logit transformation. The `reitsma` function offers the parametric $t_\alpha$ family (Doebler *et al.* 2012) of transformations as alternatives. The family is defined by

$$t_\alpha(x) := \alpha \log(x) - (2 - \alpha) \log(1 - x), \quad x \in (0, 1), \alpha \in [0, 2].$$

For $\alpha = 1$, the logit is obtained. In many cases the fit of a bivariate meta-regression can be improved upon by choosing adequate values for $\alpha$. The rational behind this is, that especially sensitivities tend to cluster around values like .95 and the symmetric logit transformation does not necessarily lead to normally distributed transformed proportions. As an example we study the `smoking` data again using maximum-likelihood estimation:

```
R> fit.smoking1 <- reitsma(smoking1, method = "ml")
R> fit.smoking2 <- reitsma(smoking1,
+                          alphasens = 0, alphafpr = 2,
+                          method = "ml")
R> AIC(fit.smoking1)

[1] -120.0473

R> AIC(fit.smoking2)

[1] -120.1002
```

The almost identical AIC values indicates, that the fit of the models is comparable. For purpose of inference, we likelihood-ratio tests are recommended, which are discussed for this type of transformation by Doebler *et al.* (2012).

### 6.4. Estimating Likelihood Ratios and Diagnostic Odds Ratio

Based on the bivariate model for diagnostic test accuracy, it is possible to obtain pooled likelihood ratios and diagnostic odds ratio. The `SummaryPts` function applies a sampling based approach (as proposed by Zwinderman and Bossuyt (2008)) to estimate the aforementioned properties. As an example, applying the `SummaryPts` function to the `AuditC` data, we would get:

```
R> summary_pts_audit <- SummaryPts(reitsma(AuditC))
R> summary(summary_pts_audit)

            Mean Median    2.5%  97.5%
posLR      4.060   4.03  3.2600  4.990
negLR      0.145   0.14  0.0811  0.234
invnegLR   7.440   7.14  4.2800 12.300
DOR       29.700  28.90 18.4000 45.300
```

### 6.5. Estimating Pooled Predictive Values

Predictive values are particularly useful in the clinical practice. The negative predictive value indicates the probability of a patient with a negative test not having a certain disease/condition, while the positive predictive value indicates the probability of a patient with a positive test having that certain disease/condition. The predictive values are dependent not only on the sensitivity and specificity of the diagnostic test, but also on the prevalence of

the disease/condition in the setting being studied. The `predv_r` and the `predv_d` functions project probability distributions of predictive values based i. on pooled sensitivities and specificities (obtained using the bivariate approach) and ii. on prevalence ranges or distributions. An application of this approach has been used by Sousa-Pinto, Tarrio, Blumenthal, Azevedo, Delgado, and Fonseca (2021) As an example, we will use the `AuditC` data. Let us consider that the prevalence of alcohol problems (the condition being assessed in the AuditC example) ranges between 5% and 15%. We can use the `predv_r` function to obtain distributions for the negative and positive predictive values of the screening test for each prevalence value within that range:

```
R> pred_audit1 <- predv_r(AuditC, prop_min=0.05, prop_max=0.15)
R> summary(pred_audit1)

Estimates of predictive values
Minimum prevalence:[1] 0.05

Maximum prevalence:[1] 0.15

NPV
   prevalence  mean    sd  p2.5    p5   p10   p25   p50   p75   p90
1        0.05 0.992 0.002 0.988 0.989 0.990 0.991 0.993 0.994 0.995
2        0.06 0.991 0.002 0.985 0.986 0.988 0.989 0.991 0.993 0.994
3        0.07 0.989 0.003 0.983 0.984 0.985 0.988 0.990 0.991 0.993
4        0.08 0.988 0.003 0.980 0.982 0.983 0.986 0.988 0.990 0.992
5        0.09 0.986 0.004 0.977 0.979 0.981 0.984 0.986 0.989 0.990
6        0.10 0.984 0.004 0.975 0.976 0.979 0.982 0.985 0.987 0.989
7        0.11 0.982 0.005 0.972 0.974 0.976 0.980 0.983 0.986 0.988
8        0.12 0.981 0.005 0.969 0.971 0.974 0.978 0.981 0.984 0.987
9        0.13 0.979 0.006 0.966 0.969 0.971 0.975 0.979 0.983 0.986
10       0.14 0.977 0.006 0.963 0.966 0.969 0.973 0.978 0.981 0.984
11       0.15 0.975 0.007 0.960 0.963 0.966 0.971 0.976 0.980 0.983
     p95 p97.5
1  0.995 0.996
2  0.994 0.995
3  0.993 0.994
4  0.992 0.993
5  0.991 0.992
6  0.990 0.991
7  0.989 0.990
8  0.988 0.989
9  0.987 0.988
10 0.986 0.987
11 0.985 0.986

PPV
   prevalence  mean    sd  p2.5    p5   p10   p25   p50   p75   p90
1        0.05 0.176 0.016 0.146 0.151 0.156 0.165 0.175 0.186 0.196
```

```
2         0.06 0.205 0.018 0.172 0.177 0.183 0.193 0.205 0.217 0.228
3         0.07 0.233 0.019 0.197 0.202 0.209 0.220 0.233 0.246 0.259
4         0.08 0.260 0.021 0.221 0.227 0.234 0.246 0.260 0.274 0.287
5         0.09 0.286 0.022 0.244 0.250 0.258 0.270 0.285 0.300 0.314
6         0.10 0.310 0.023 0.266 0.273 0.280 0.294 0.309 0.325 0.340
7         0.11 0.333 0.024 0.287 0.294 0.302 0.316 0.332 0.349 0.364
8         0.12 0.355 0.025 0.308 0.315 0.324 0.338 0.355 0.372 0.387
9         0.13 0.376 0.025 0.328 0.335 0.344 0.359 0.376 0.393 0.409
10        0.14 0.397 0.026 0.347 0.355 0.364 0.379 0.396 0.414 0.430
11        0.15 0.416 0.026 0.366 0.373 0.382 0.398 0.416 0.434 0.450
     p95 p97.5
1  0.202 0.208
2  0.235 0.242
3  0.266 0.273
4  0.296 0.303
5  0.323 0.331
6  0.349 0.357
7  0.373 0.381
8  0.397 0.405
9  0.419 0.427
10 0.440 0.448
11 0.460 0.469
```

We observe that the, for the defined prevalence range, the mean estimate for the negative predictive value ranges between 98% and 99%, while the mean estimate for the positive predictive value ranges between 18% and 42%.

Now let us consider that we do not want to project predictive values based on a prevalence range, but rather based on a prevalence probability distribution. We know that the prevalence of alcohol problems is given by a distribution, with a mean value of 10% and a standard-deviation of 5. We can use the `predv_d` function to obtain probability distributions for the negative and positive predictive values:

```
R> pred_audit2 <- predv_d(AuditC, prop_m=0.10, prop_sd=0.05)
R> summary(pred_audit2)

     Mean    SD  p2.5    p5  p10   p25   p50   p75   p90   p95 p97.5
NPV 0.984 0.010 0.957 0.964 0.97 0.979 0.986 0.991 0.994 0.996 0.997
PPV 0.297 0.115 0.094 0.118 0.15 0.212 0.291 0.375 0.452 0.497 0.534
```

We obtain a distribution of negative predictive values defined by a mean of 98% and a standard-deviation of 1% (95%CrI=96-100%), and a distribution of positive predictive values defined by a mean of 30% and a standard-deviation of 12% (95%CrI=10-53%).

By default, the Zwindermann & Bossuyt approach is used to generate samples based of sensitivities and false positive rates Zwinderman and Bossuyt (2008), based on which distributions of predictive values are obtained. For faster results, such an approach may not be used (`zb=FALSE`). Expected differences are small, especially if the number of participants of primary studies is sufficiently high.

If prevalence inputs (minimum and maximum, or mean and standard-deviation) are not provided, the `predv_r` and the `predv_d` functions will estimate those inputs based on data from primary studies. However, this may not be advisable, as studies focusing on diagnostic test accuracy are typically not specifically designed for prevalence assessment (case-control studies are particularly troublesome in this context).

## 7. Further development

In the future **mada** will support the mixture approach of Holling, Böhning, and Böhning (2012a) and Bayesian approaches.

## Acknowledgements

This work was funded by the DFG project HO 1286/7-2.

## References

Arends L, Hamza T, Van Houwelingen J, Heijenbrok-Kal M, Hunink M, Stijnen T (2008). "Bivariate Random Effects Meta-Analysis of ROC Curves." *Medical Decision Making*, **28**, 621–638.

Biostat, Inc (2006). "Comprehensive Meta-Analysis (**CMA**), Version 2." Computer program.

Borenstein M, Hedges L, Higgins J, Rothstein H (2009). *Introduction to Meta-Analysis*. John Wiley & Sons.

Cochran W (1954). "The Combination of Estimates from Different Experiments." *Biometrics*, **10**, 101–129.

Deeks J (2001). "Systematic Reviews of Evaluations of Diagnostic and Screening Tests." *British Medical Journal*, **323**, 157–162.

DerSimonian R, Laird N (1986). "Meta-Analysis in Clinical Trials." *Controlled Clinical Trials*, **7**, 177–188.

Doebler P, Holling H, Böhning D (2012). "A Mixed Model Approach to Meta-Analysis of Diagnostic Studies With Binary Test Outcome." *Psychological Methods*.

Gasparrini A, Armstrong B, Kenward MG (2012). "Multivariate Meta-Analysis for Non-Linear and other Multi-Parameter Associations." *Statistics in Medicine*, **Epub ahead of print**(doi: 10.1002/sim.5471).

Gatsonis C, Paliwal P (2006). "Meta-Analysis of Diagnostic and Screening Test Accuracy Evaluations: Methodologic Primer." *American Journal of Roentgenology*, **187**, 271–281.

Glas A, Lijmer J, Prins M, Bonsel G, Bossuyt P (2003). "The Diagnostic Odds Ratio: A Single Indicator of Test Performance." *Journal of Clinical Epidemiology*, **56**, 1129–1135.

Harbord R, Deeks J, Egger M, Whiting P, Sterne J (2007). "A Unification of Models for Meta-Analysis of Diagnostic Accuracy Studies." *Biostatistics*, **8**, 239–251.

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

Higgins J, Thompson S, Deeks J, Altman D (2003). "Measuring Inconsistency in Meta-Analyses." *British Medical Journal*, **327**, 557–560.

Holling H, Böhning W, Böhning D (2012a). "Likelihood-Based Clustering of Meta-Analytic SROC Curves." *Psychometrika*, **77**, 106–126.

Holling H, Böhning W, Böhning D (2012b). "Meta-Analysis of Diagnostic Studies Based upon SROC-Curves: A Mixed Model Approach Using the Lehmann Family." *Statistical Modelling*, **12**, 347–375.

Holling H, Bohning W, Masoudi E, Bohning D, Sangnawakij P (2019). "Evaluation of a new version of I2 with emphasis on diagnostic problems." *Communications in Statistics - Simulation and Computation*, (doi: 10.1080/03610918.2018.1489553).

Jones C, Athanasiou T (2005). "Summary Receiver Operating Characteristic Curve Analysis Techniques in the Evaluation of Diagnostic Tests." *The Annals of Thoracic Surgery*, **79**, 16–20.

Kriston L, Hölzel L, Weiser A, Berner M, Härter M (2008). "Meta-Analysis: Are 3 Questions Enough to Detect Unhealthy Alcohol Use?" *Annals of Internal Medicine*, **149**, 879–888.

Le C (2006). "A Solution for the Most Basic Optimization Problem Associated with an ROC Curve." *Statistical Methods in Medical Research*, **15**, 571–584.

Leeflang M, Deeks J, Gatsonis C, Bossuyt P (2008). "Systematic Reviews of Diagnostic Test Accuracy." *Annals of Internal Medicine*, **149**, 889–897.

Lunn D, Spiegelhalter D, Thomas A, Best N (2009). "The BUGS Project: Evolution, Critique and Future Directions." *Statistics in Medicine*, **28**(25), 3049–3067.

Lunn D, Thomas A, Best N, Spiegelhalter D (2000). "WinBUGS – a Bayesian Modelling Framework: Concepts, Structure, and Extensibility." *Statistics and computing*, **10**, 325–337.

Patrick D, Cheadle A, Thompson D, Diehr P, Koepsell T, Kinne S (1994). "The Validity of Self-reported Smoking: A Review and Meta-Analysis." *American Journal of Public Health*, **84**, 1086–1093.

Phillips B, Stewart L, Sutton A (2010). "Cross Hairs Plots for Diagnostic Meta-Analysis." *Research Synthesis Methods*, **1**, 308–315.

R Core Team (2012). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/.

Reitsma J, Glas A, Rutjes A, Scholten R, Bossuyt P, Zwinderman A (2005). "Bivariate Analysis of Sensitivity and Specificity Produces Informative Summary Measures in Diagnostic Reviews." *Journal of Clinical Epidemiology*, **58**, 982–990.

Rutter C, Gatsonis C (2001). "A Hierarchical Regression Approach to Meta-Analysis of Diagnostic Test Accuracy evaluations." *Statistics in Medicine*, **20**, 2865–2884.

Sousa-Pinto B, Tarrio I, Blumenthal K, Azevedo L, Delgado L, Fonseca J (2021). "Accuracy of penicillin allergy diagnostic tests: A systematic review and meta-analysis." *Journal of Allergy and Clinical Immunology*, **147**(doi: 10.1016/j.jaci.2020.04.058), 296–308.

Sutton A, Abrams K, Jones D, Sheldon T, Song F (2000). *Methods for Meta-Analysis in Medical Research*. John Wiley & Sons.

Takwoingi Y, Deeks J (2011). "**METADAS**: An SAS Macro for Meta-Analysis of Diagnostic Accuracy Studies, Version 1.3." Computer program.

The Nordic Cochrane Centre (2011). "Review Manager (**RevMan**), Version 5.1." Computer program.

Van Houwelingen H, Arends L, Stijnen T (2002). "Advanced Methods in Meta-Analysis: Multivariate Approach and Meta-Regression." *Statistics in Medicine*, **21**, 589–624.

Walter S (2002). "Properties of the Summary Receiver Operating Characteristic (SROC) Curve for Diagnostic Test Data." *Statistics in Medicine*, **21**, 1237–1256.

Zhou Y, Dendukuri N (2014). "Statistics for quantifying heterogeneity in univariate and bivariate meta-analyses of binary data: The case of meta-analyses of diagnostic accuracy." *Statistics in Medicine*, **33**(doi: 10.1002/sim.6115), 2701–2717.

Zwinderman A, Bossuyt P (2008). "We should not pool diagnostic likelihood ratios in systematic reviews." *Statistics in Medicine*, **27**(doi: 10.1002/sim.2992), 687–697.

## Affiliation:

Philipp Doebler
Fachbereich Psychologie und Sportwissenschaft
Westfälische Wilhelms-Universität Münster
D-48149 Münster, Germany
E-mail: doebler@uni-muenster.de
URL: http://wwwpsy.uni-muenster.de/Psychologie.inst4/AEHolling/personen/P_Doebler.html

Heinz Holling
Fachbereich Psychologie und Sportwissenschaft
Westfälische Wilhelms-Universität Münster
D-48149 Münster, Germany
E-mail: holling@uni-muenster.de
URL: http://wwwpsy.uni-muenster.de/Psychologie.inst4/AEHolling/personen/holling.html