Statistical methods

General principles

Meta-analysis is a two-stage process[1]. In the first stage, a summary statistic is calculated for each study. Unlike controlled trials, in evaluation of diagnostic tests, each study is summarized by a pair of statistics that measures the test's accuracy. The pair is usually either sensitivity and specificity or positive and negative likelihood ratios. In the second stage, the overall test accuracy indexes are calculated as the weighted average of these summary statistics. Meta-analysis should only be performed when studies have recruited clinically similar patients and have used comparable experimental and reference tests. When there is considerable heterogeneity in study results, the reviewer should investigate the reasons for these differences rather than reporting a pooled estimate.

This document describes the different tools implemented by **MetaDiSc**: i) summarizing data from each individual study, ii) investigating the homogeneity of studies graphically and statistically, iii) computing the pooled indexes and iv) exploring heterogeneity.

Summary statistics in individual studies

The results of each individual study should be presented in a 2 x 2 table (Table 1) showing the number of people who have been classified as positive and negative by the experimental test among the groups of participants with and without disease according the reference test.

Table 1: Results in an individual study

	Reference test		
Experimental test	With disease	Without disease	Total
Positive	a	b	P
Negative	c	d	N
Total	D	ND	T

a: number of people with positive test result and disease: True positives (TP).
b: number of people with positive test result and without disease: False positives (FP).
c: number of people with negative test result and disease: False negatives (FN).
d: number of people with negative result and without disease: True negatives (TN).
P: total number of people with positive test result.
N: total number of people with negative test result.
D: total number of people with disease.
ND: total number of people without disease.
T: number of people in the study.

Accuracy can be expressed by *sensitivity* (proportion of positives among people with disease) and *specificity* (proportion of negatives among people without disease).

$$Sen = \frac{a}{D} \qquad Spe = \frac{d}{ND}$$

or by the *likelihood ratios*

$$LR+ = \frac{Sen}{1-Spe} = \frac{\frac{a}{D}}{\frac{b}{ND}} \qquad LR- = \frac{1-Sen}{Spe} = \frac{\frac{c}{D}}{\frac{d}{ND}}$$

The likelihood ratios express how much more frequent the respective result is among subjects with disease than among subjects without disease.

Another measure of the test accuracy, useful in meta-analysis, is the *diagnostic odds ratio* (*DOR*)

$$DOR = \frac{LR+}{LR-} = \frac{a \times d}{b \times c}$$

The *DOR* expresses how much greater the odds of having the disease are for the people with a positive test result than for the people with a negative test result. It is a single measure of diagnostic test performance that combines both likelihood ratios.

Standard errors and confidence intervals

The confidence intervals of sensitivity and specificity are calculated using the *F* distribution method[2] to compute the exact confidence limits for the binomial proportion (*x*/*n*).

$$LL = \left(1 + \frac{n-x+1}{xF_{2x,2(n-x+1),1-\alpha/2}}\right)^{-1} \qquad UL = \left(1 + \frac{n-x}{(x+1)F_{2(x+1),2(n-x),\alpha/2}}\right)^{-1}$$

The distribution of the logarithm of the likelihood ratios are approximately normal and their standard errors are[1] :

$$SE\left(\ln LR+\right) = \sqrt{\frac{1}{a} + \frac{1}{b} - \frac{1}{D} - \frac{1}{ND}}\,; \qquad SE\left(\ln LR-\right) = \sqrt{\frac{1}{c} + \frac{1}{d} - \frac{1}{D} - \frac{1}{ND}}$$

Thus the confidence intervals of the *LR*'s are

$$LR\, \mathrm{e}^{\pm z_{\alpha/2} SE(\ln LR)}$$

The distribution of logarithm of the diagnostic odds ratio is also approximately normal, with standard error given[1] by

$$SE\left(\ln DOR\right) = \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}$$

Thus the confidence interval of the *DOR* is

$$DOR\, \mathrm{e}^{\pm z_{\alpha/2} SE(\ln DOR)}$$

Assessing homogeneity

The degree of variability among study results should first be evaluated graphically by plotting the sensitivity and specificity from each study on a *forest plot.* Some divergence is to be expected by chance, but variation in other factors may increase the observed heterogeneity.

There is one important extra source of variation in meta-analysis of diagnostic accuracy: the studies included may have used, explicitly or implicitly, different *thresholds* to define positive and negative test results. To explore this source of variation, it is useful to plot sensitivity and specificity on an *ROC* plane. If such a threshold effect exists, the points will show a curvilinear pattern. One can also test for this threshold effect by calculating the Spearman correlation coefficient between sensitivity and specificity. If the threshold effect exists an inverse correlation appears[3]. Combining study results in these cases involves fitting an ROC curve rather than pooling sensitivities and specificities or likelihood ratios.

The homogeneity of the sensitivities and specificities can also be tested applying the likelihood ratio test[4].

(From here on, we will use the notation in Table 1, with subscripts i to designate an individual study and T for overall index).

$$G^2_{Sen} = 2\sum_i \left(a_i \ln \frac{a_i}{\frac{a_T \times D_i}{D_T}} + c_i \ln \frac{c_i}{\frac{c_T \times D_i}{D_T}} \right) \qquad a_T = \sum_i a_i \quad c_T = \sum_i c_i \quad D_T = \sum_i D_i$$

$$G^2_{Spe} = 2\sum_i \left(d_i \ln \frac{d_i}{\frac{d_T \times ND_i}{ND_T}} + b_i \ln \frac{b_i}{\frac{b_T \times ND_i}{ND_T}} \right) \qquad b_T = \sum_i b_i \quad d_T = \sum_i d_i \quad ND_T = \sum_i ND_i$$

In the homogeneity hypothesis, both have asymptotic chi-squared distribution with *k*-1 degrees of freedom (*k* being to the number of the studies).

The homogeneity of likelihood ratios and diagnostic odds ratios can be tested using Cochran's *Q* test based upon inverse variance weights[1], which also has a chi-squared distribution with *k*-1 degrees of freedom.

$$Q = \sum_i w_i \left(\ln \theta_i - \ln \theta_T \right)^2 \qquad w_i = \frac{1}{SE\left(\ln \theta_i \right)^2}$$

where θ is the positive or negative likelihood ratio or the diagnostic odds ratio.

As meta-analyses often include small numbers of studies the power of both tests (G^2 and Q) is low, so they are poor at detecting true heterogeneity among studies as significant. An alternative approach to quantify the effect of heterogeneity is the I^2 index which describes the percentage of total variation across studies that is due to heterogeneity rather than chance[5]. I^2 is calculated

$$I^2 = \frac{\chi^2 - (d.f.)}{\chi^2} \times 100$$

where χ^2 is the G^2 ó Q statistic and *d.f.* its degrees of freedom.

Pooled indexes

Sensitivities, specificities and likelihood ratios should only be pooled in the absence of variability of the diagnostic threshold.

Sensitivity and specificity are pooled by

$$Sen_T = \frac{\sum_i a_i}{\sum_i D_i} \qquad Spe_T = \frac{\sum_i d_i}{\sum_i ND_i}$$

These formulas correspond to weighted averages in which the weight of each study is its sample size.

Likelihood ratios and diagnostic odds ratios can be pooled by the Mantel-Haenszel method (fixed effects model) or by the DerSimonian Laird method (random effects model) to incorporate variation among studies. Both methods compute a weighted average, but the difference lies in the weights used and the "effect size" to be averaged. With the Mantel-Haenszel method, the *DOR*'s or *LR*'s are averaged whereas with the DerSimonian Laird method, the logs of *DOR's* or *LR*'s are averaged[1].

$$\theta_T^{MH} = \frac{\sum_i w_i^{MH} \theta_i}{\sum_i w_i^{MH}} \qquad \ln \theta_T^{DL} = \frac{\sum_i w_i^{DL} \ln \theta_i}{\sum_i w_i^{DL}}$$

The Mantel-Haenszel weights are

$$DOR: \quad w_i^{MH} = \frac{b_i c_i}{T_i} \qquad LR+: \quad w_i^{MH} = \frac{b_i D_i}{T_i} \qquad LR-: \quad w_i^{MH} = \frac{d_i D_i}{T_i}$$

For all statistics, the DerSimonian Laird weights are:

$$w_i^{DL} = \frac{1}{SE\left(\ln \theta_i\right)^2 + \tau^2}$$

where θ_i is the likelihood ratio or diagnostic odds ratio and τ^2 an estimation of between studies variance given by

$$\tau^2 = \begin{cases} \dfrac{Q-(k-1)}{\sum_i w_i - \left(\dfrac{\sum_i w_i^2}{\sum_i w_i}\right)} & \text{if} \quad Q > k-1 \\ 0 & \text{if} \quad Q < k-1 \end{cases}$$

Q stands for the Cochran homogeneity statistic calculated using the Mantel-Haenszel overall estimate and w_i the inverse variance weights.

Standard errors and confidence intervals of pooled indexes

The confidence intervals of overall sensitivity and specificity are also calculated using the *F* distribution method[2] to compute the exact confidence limits for the binomial proportion. However, **Meta-DiSc** optionally computes them using overdispersion correction. In this case, it uses the normal approximation to binomial, i.e.

$$SE\left(Sen_T\right) = \sqrt{\frac{Sen_T\left(1-Sen_T\right)}{\sum_i D_i}} \qquad SE\left(Spe_T\right) = \sqrt{\frac{Spe_T\left(1-Spe_T\right)}{\sum_i ND_i}}$$

the confidence intervals[6] corrected by overdispersion are:

$$Sen_T \pm z_{\alpha/2}\varphi_{Sen}SE(Sen_T) \qquad Spe_T \pm z_{\alpha/2}\varphi_{Spe}SE(Spe_T)$$

with the correction factors

$$\varphi_{Sen} = \sqrt{\frac{\chi^2_{Sen}}{k-1}} \quad \text{with} \quad \chi^2_{Sen} = \sum_i \left(\frac{\left(a_i - \dfrac{a_T \times D_i}{D_T}\right)^2}{\dfrac{a_T \times D_i}{D_T}} + \frac{\left(c_i - \dfrac{c_T \times D_i}{D_T}\right)^2}{\dfrac{c_T \times D_i}{D_T}} \right)$$

$$\varphi_{Spe} = \sqrt{\frac{\chi^2_{Spe}}{k-1}} \quad \text{with} \quad \chi^2_{Spe} = \sum_i \left(\frac{\left(d_i - \dfrac{d_T \times ND_i}{ND_T}\right)^2}{\dfrac{d_T \times ND_i}{ND_T}} + \frac{\left(b_i - \dfrac{b_T \times ND_i}{ND_T}\right)^2}{\dfrac{b_T \times ND_i}{ND_T}} \right)$$

The distribution of the logarithm of the Mantel-Haenszel overall likelihood ratios and overall *DOR* are approximately normal with standard error given [1] by:

$$SE\left(\ln LR+\right) = \sqrt{\frac{P}{U \times V}} \qquad SE\left(\ln LR-\right) = \sqrt{\frac{P}{U' \times V'}}$$

$$SE\left(\ln DOR\right)=\sqrt{\frac{1}{2}\left(\frac{E}{R^2}+\frac{F+G}{R\times S}+\frac{H}{S^2}\right)}$$

where

$$P=\sum\frac{\left(D_i\times ND_i\left(a_i+b_i\right)-a_ib_iT_i\right)}{T_i^2} \qquad U=\sum\frac{a_iND_i}{T_i} \qquad V=\sum\frac{c_iD_i}{T_i}$$

$$U'=\sum\frac{b_iND_i}{T_i} \qquad V'=\sum\frac{d_iD_i}{T_i}$$

$$R=\sum\frac{a_id_i}{T_i} \qquad S=\sum\frac{b_ic_i}{T_i} \qquad E=\sum\frac{\left(a_i+d_i\right)a_id_i}{T_i^2}$$

$$F=\sum\frac{\left(a_i+d_i\right)b_ic_i}{T_i^2} \qquad G=\sum\frac{\left(b_i+c_i\right)a_id_i}{T_i^2} \qquad H=\sum\frac{\left(b_i+c_i\right)b_ic_i}{T_i^2}$$

Thus the confidence intervals are:

$$LR\ \mathrm{e}^{\pm z_{\alpha/2}SE(\ln LR)} \qquad DOR\ \mathrm{e}^{\pm z_{\alpha/2}SE(\ln DOR)}$$

The distribution of the logarithm of the DerSimonian-Laird overall likelihood ratios and overall *DOR* are also approximately normal with standard error given[1] by:

$$SE\left(\ln\theta_T^{DL}\right)=\frac{1}{\sqrt{\sum w_i^{DL}}}$$

Thus the confidence intervals are:

$$\theta\ \mathrm{e}^{\pm z_{\alpha/2}SE(\ln\theta)}$$

ROC curves

If there is any evidence of diagnostic threshold variation among studies, the best summary of study results will be an ROC curve rather than a single point. The shape of the ROC curve depends on the underlying distribution of test results in patients with and without the disease[7]. There are two methods of fitting the ROC curve.

Diagnostic tests where the *DOR* is constant regardless of the diagnostic threshold have symmetrical curves around the "*Sen*=*Spe*" line. In these situations, it is possible to combine *DOR's* by the Mantel-Haenszel or the DerSimonian Laird methods to estimate the overall *DOR* and hence to determine the best-fitting ROC curve[8]. The equation of curve is given by

$$Sen = \frac{1}{1 + \frac{1}{DOR_T \times \left(\frac{1 - Spe}{Spe}\right)}}$$

When the *DOR* changes with diagnostic threshold, the ROC curve is asymmetrical. To study *DOR* variation in according to threshold, and thereby fit symmetrical or asymmetrical curves, the Moses- Shapiro-Littenberg method[8] is used.

The method consists of studying this relationship by fitting the straight line

$$D = a + bS$$

where *D* is the log of *DOR* and *S* a measure of threshold given by

$$S = \ln\left(\frac{Sen}{1 - Sen} \times \frac{1 - Spe}{Spe}\right)$$

Estimates of parameters a and b and their standard errors and covariance are obtained by ordinary or weighted least squares method using the NAG C library[9]. The weights can be simply the sample size or the inverse of variance of the log of the *DOR*. Optionally, random effects between studies can be taken into account using one of three different iterative methods (Restricted maximum likelihood, Maximun likelihood and Empirical Bayes)[10].

Testing the hypothesis of whether or not diagnostic performance (measured by *DOR*) varies with threshold is equivalent to testing whether parameter b = 0. If b=0, there is no variation, the method yields a symmetrical ROC curve, and e^a is an estimate of the overall *DOR*. However if $b \neq 0$, variation exists and the ROC curve is asymmetrical, given by equation

$$Sen = \frac{1}{1 + \frac{1}{e^{\frac{a}{1-b}} \times \left(\frac{1 - Spe}{Spe}\right)^{\frac{1+b}{1-b}}}}$$

A useful statistic in pooling studies by means of the ROC curve is the area under the curve (AUC) which summarizes the diagnostic performance as a single number[11]: a perfect test will have an AUC close to 1 and poor tests have AUCs close to 0.5. The AUC is computed by numeric integration of the curve equation by the trapezoidal method.

Another useful statistic is the *Q** index, defined by the point where sensitivity and specificity are equal, which is the point closest to the ideal top-left corner of the ROC space. It is calculated[12] by:

$$Q^* = \frac{\sqrt{DOR_T}}{1 + \sqrt{DOR_T}}$$

Standard errors of AUC and Q* and confidence intervals of ROC curve

The standard error of the area under the symmetrical ROC curve is given[12] by

$$SE(AUC_{sym}) = \frac{DOR_T}{(DOR_T - 1)^3}\left[(DOR_T + 1)\ln DOR_T - 2(DOR_T - 1)\right]SE(\ln DOR_T)$$

but if the curve is asymmetrical its standard error is given by

$$SE(AUC_{asy}) = \sqrt{A^2 \operatorname{var}(a) + B^2 \operatorname{var}(b) + 2AB\operatorname{cov}(a,b)}$$

where *A* and B are:

$$A = \left(\frac{1}{1-b}\right)\exp\left(\frac{a}{1-b}\right)\int_0^1 \frac{\left(\frac{x}{1-x}\right)^p}{\left[1+\left(\frac{x}{1-x}\right)^p \exp\left(\frac{a}{1-b}\right)\right]^2}dx$$

$$B = \left(\frac{1}{1-b}\right)^2 \exp\left(\frac{a}{1-b}\right)\int_0^1 \frac{\left(\frac{x}{1-x}\right)^p \left[a + 2\ln\left(\frac{x}{1-x}\right)\right]}{\left[1+\left(\frac{x}{1-x}\right)^p \exp\left(\frac{a}{1-b}\right)\right]^2}dx$$

with $p = \frac{1+b}{1-b}$.

When the range of ROC curve is constrained to some limits (upper-left quadrant or user defined limits) the standard error of AUC is computed using the formula of asymmetrical curve, substituting accordingly the integration limits. Of note, the standard error of AUC for constrained curves can be only calculated when DOR_T is computed from Moses's model.

The standard error of *Q** is

$$SE(Q^*) = \frac{\sqrt{DOR_T}}{2\left(1+\sqrt{DOR_T}\right)^2}SE(\ln DOR_T)$$

Confidence interval of symmetrical ROC curve is calculated introducing the upper and lower limits of confidence interval of overall DOR in the equation of curve. For asymmetrical ROC curves obtained by Moses-Shapiro-Littenberg model, Mitchell[13] suggests that a confidence interval for curve is the back-transformed on the confidence band for the linear regression. The back-transformed is given by

$$Sen = \frac{1}{1+e^{-\frac{D+S}{2}}} \qquad Spe = \frac{1}{1+e^{-\frac{D-S}{2}}}$$

Meta-regression

To explore sources of heterogeneity in the studies, the Moses-Shapiro-Littenberg method can be extended by adding covariates[14] to the model. The antilogarithm transformations of the resulting estimated parameters can be interpreted as a relative *DOR* (*RDOR*) of the corresponding covariable. They indicate the change in diagnostic performance of the test under study per unit increase in the covariate.

Note about correction of cells with zero

If any study has a table with a 0 value in any cell, some statistics implemented by **MetaDisc** cannot be calculated. A solution to this problem, suggested by Cox[15], is to add 0.5 to **all** cells in the table. **MetaDisc** allows users to select from the following options: I) eliminating meta-analysis of all studies with 0 in any cell, ii) adding 0.5 to all cells in the studies where any cells were 0, iii) adding 0.5 to all cells in all studies.

In any case, this correction does not apply to calculations of sensitivity and specificity except for the SROC graph, where points correspond to sensitivities and specificities are calculated with the selected correction.

References

1. Deeks JJ. Systematic reviews of evaluations of diagnostic and screening tests. In Egger M, Smith GD, Altman DG (eds). Systematic Reviews in Health Care. Meta-analysis in context. London: BMJ Books; 2001:248-282.
2. Leemis LM, Trivedi KS. A Comparison of Approximate Interval Estimators for the Bernoulli Parameter. Am Stat 1996; 50:63-68.
3. Devillé WL, Buntinx F, Bouter LM, Montori VM, de Vet HC, van der Windt DA, Bezemer PD. Conducting systematic reviews of diagnostic studies: didactic guidelines. BMC Med Res Methodol 2002; 2:9.
4. Agresti A. Analysis of ordinal categorical data. New York: John Wileys & Sons; 1984.
5. Higgins JPT, Thompson SG, Deeks JJ, Altman DG. Measuring inconsistency in meta-analyses. BMJ 2003; 327:557-560.
6. McCullagh P, Nelder JA. Generalized Linear Models. Boca Raton: Chapman & Hall; 1989.
7. Glas AS, Lijmer JG, Prins MH, Bonsel GJ, Bossuyt PM. The diagnostic odds ratio: a single indicator of test performance. J Clin Epidemiol 2003; 56:1129-1135.
8. Moses LE, Shapiro D, Littenberg B. Combining independent studies of a diagnostic test into a summary ROC curve: data-analytic approaches and some additional considerations. Stat Med 1993; 12:1293-1316.
9. The Numerical Algorithms Group. NAG C Library. Oxford: http://www.nag.co.uk; 2004.
10. Thompson SG, Sharp SJ. Explaining heterogeneity in meta-analysis: a comparison of methods. Stat Med 1999; 18:2693-2708.
11. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 1982; 143:29-36.
12. Walter SD. Properties of the summary receiver operating characteristic (SROC) curve for diagnostic test data. Stat Med 2002; 21:1237-1256.
13. Mitchell MD. Validation of the summary ROC for diagnostic test meta-analysis: A Monte Carlo simulation. Acad Radiol 2003; 10:25-31.
14. Lijmer JG, Bossuyt PM, Heisterkamp SH. Exploring sources of heterogeneity in systematic reviews of diagnostic tests. Stat Med 2002; 21:1525-1537.
15. Cox DR. The analysis of binary data. London: Methuen; 1970.