Statistical Methods in Medical Research 2008; 1–12

Revisiting youden's index as a useful measure of the misclassification error in meta-analysis of diagnostic studies

Dankmar Böhning Applied Statistics, School of Biological Sciences, University of Reading, Reading, UK and **Walailuck Böhning** and **Heinz Holling** Statistics and Methods, Institute for Psychology IV, University of Münster, Münster, Germany

The paper considers meta-analysis of diagnostic studies that use a continuous score for classification of study participants into healthy or diseased groups. Classification is often done on the basis of a threshold or cut-off value, which might vary between studies. Consequently, conventional meta-analysis methodology focusing solely on separate analysis of sensitivity and specificity might be confounded by a potentially unknown variation of the cut-off value. To cope with this phenomena it is suggested to use, instead, an overall estimate of the misclassification error previously suggested and used as Youden's index and; furthermore, it is argued that this index is less prone to between-study variation of cut-off values. A simple Mantel–Haenszel estimator as a summary measure of the overall misclassification error is suggested, which adjusts for a potential study effect. The measure of the misclassification error based on Youden's index is advantageous in that it easily allows an extension to a likelihood approach, which is then able to cope with unobserved heterogeneity via a nonparametric mixture model. All methods are illustrated at hand of an example on a diagnostic meta-analysis on duplex doppler ultrasound, with angiography as the standard for stroke prevention.

1 Introduction

Meta-analysis has become an important tool in the social, bio-medical and pharmaceutical, life and natural sciences for the analysis, integration and deeper understanding of empirical findings resulting from several studies. Numerous recent books[1–5] and tutorials[6] underline a sustainable interest and development in this field.

However, several niches exist experiencing specific problems, which are not well covered within the existing body of meta-analytic developments. One of these is the meta-analysis of diagnostic studies. Here, interest has developed in the *diagnostic test accuracy* evaluated in a series of diagnostic studies.[7,8] The objective lies in determining the discriminating power of the diagnostic test in separating persons with a specific condition (diseased) from those without this condition (non-diseased). Widely, two measures of diagnostic accurracy are considered:

- the *sensitivity* defined as $Pr(\text{positive}|\text{diseased}) = (1 - \alpha)$ and
- the *specificity* defined as $Pr(\text{negative}|\text{non-diseased}) = (1 - \beta)$.

Address for correspondence: Dankmar Böhning, Applied Statistics, School of Biological Sciences, University of Reading, Harry Pitt Building, Earely Gate, Reading RG6 6FN, UK. E-mail: d.a.w.bohning@reading.ac.uk

10.1177/0962280208081867

The sensitivity measures the capability of the diagnostic test to recognize a diseased person correctly, whereas the specificity measures the capability of diagnosing a healthy person correctly. Consequently, α is the error probability of falsely classifying a diseased person as healthy and β is the error probability of falsely classifying a healthy person as diseased. Ideally, α and β should be small if not zero.

Suppose that a series of k diagnostic studies is available. For study i, let

- $x_{H,i}$ be the frequency of (falsely) positively classified persons out of $n_{H,i}$ healthy ones,
- $x_{D,i}$ be the frequency of (falsely) negatively classified persons out of $n_{D,i}$ diseased ones,

so that natural estimates for α_i and β_i are provided as $\hat{\alpha}_i = x_{D,i}/n_{D,i}$ and $\hat{\beta}_i = x_{H,i}/n_{H,i}$.

1.1 An application study: DDU-meta-analysis

We will illustrate these definitions and concepts at hand of a meta-analysis of 14 studies on the accuracy of duplex doppler ultrasound (DDU) with angiography as the standard for stroke prevention described previously in Hasselblad and Hedges.[9]

Table 1 shows for each study the number $x_{H,i}$ of positively classified persons out of $n_{H,i}$ healthy ones, the number $x_{D,i}$ of negatively classified persons out of $n_{D,i}$ with positive angiography, and the associated estimated error rates. The latter show considerable variability: from 0 to 29%. It will be discussed in the following that a separate analysis of these error rates might be not meaningful, as any meta-analysis of diagnostic accuracy focusing separately on sensitivity and specificity.[10] Sensitivity and specificity might not be directly comparable *between* studies because of the cut-off value problem. This is illustrated in the next section.

Table 1 Studies of duplex doppler ultrasound using angiography as the standard

Study i	$x_{H,i}$	$n_{H,i}$	$\hat{\alpha}_i$	$x_{D,i}$	$n_{D,i}$	$\hat{\beta}_i$	Accuracy
1	2	85	0.02	4	30	0.13	high
2	2	7	0.29	1	12	0.08	low
3	8	42	0.19	3	71	0.04	low
4	0	111	0.00	12	86	0.14	high
5	13	112	0.12	20	104	0.19	low
6	7	48	0.15	3	43	0.07	low
7	9	118	0.08	1	17	0.06	high
8	15	221	0.07	20	116	0.17	low
9	2	59	0.03	2	13	0.15	high
10	5	62	0.08	5	96	0.05	high
11	3	45	0.07	9	55	0.16	low
12	2	95	0.02	1	16	0.06	high
13	16	137	0.12	10	68	0.15	low
14	1	75	0.01	4	30	0.13	high

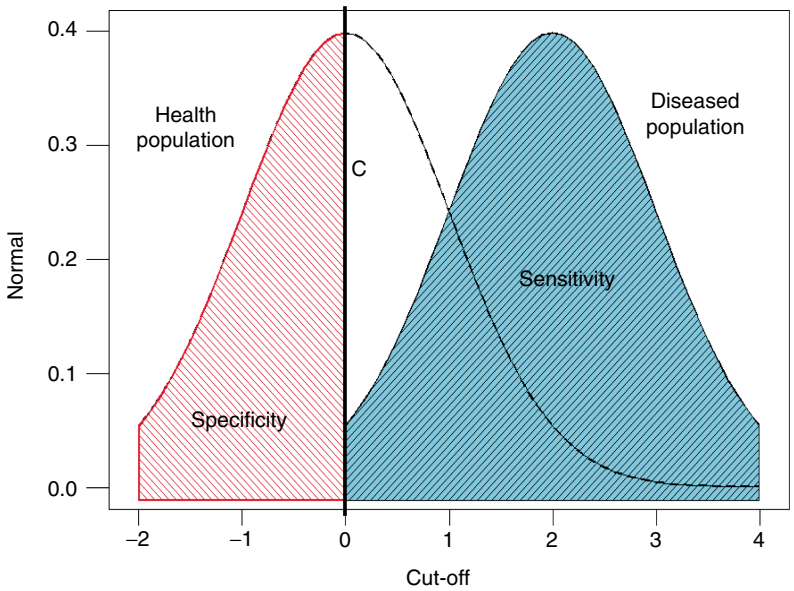

Figure 1 Diagnostic situation illustrated with two normal distributions with variance of 1: one has mean 0 (healthy population), the other has mean 2 (diseased population).

2 The problem

We suppose that the diagnostic test is providing a continuous outcome. For example, a psychological test is used (potentially among other procedures) to determine a medical condition such as the presence of dementia in an elderly person. Often these diagnostic tests deliver a score and a cut-off value is used to decide the presence or absence of the condition. This is illustrated in Figure 1, where two normal distributions with equal variance of one and different means are used. A cut-off value c, here $c = 0$, determines sensitivity as $1 - \alpha = 1 - \Phi((c - \mu^D)/\sigma^D)$ and specificity as $1 - \beta = \Phi((c - \mu^H)/\sigma^H)$, assuming that values above c indicate positivity of the test and Φ is the cumulative distribution function of the standard normal. As has been underlined in the *Guidelines for Meta-analyses Evaluating Diagnostic Tests*,[8] if c is shifted to the right, the sensitivity decreases, whereas the specificity increases, and *vice versa* if c is shifted to the left. This fact implies that quite different sensitivities and specificities might have been observed in the different studies though the underlying diagnostic test might have entirely identical discriminating power. Unless it is verifiable that a common cut-off value has been used a separate meta-analysis of sensitivity and specificity might lead to biased findings such as artificial heterogeneity or spurious variance inflation.

Recently, two alternative approaches were reviewed:[11] the bivariate random-effects meta-analysis[12,13] and the summary receiver operating characteristic model.[14] It has been suggested, earlier, to use techniques based on the *Receiver Operating Characteristic* (ROC)[10] to cope with different cut-off values. As can be seen from Figure 1–if c ranges from $-\infty$ to $+\infty$, a continuum of $(1 - \alpha, \beta)$-pairs arises and, if plotted in a $(1 - \alpha) \times \beta$-diagram, the ROC-curve occurs. The advantage of a ROC-analysis lies in the fact that

the cut-off value problem is incorporated. Two ROC-curves can be directly compared, and if one is uniformly above the other, the discriminatory power can be uniformly evaluated. Consequently, the use of the ROC-analysis has been recommended.[1,15] However, disadvantages of the technique lie in the fact that pairs arising from different studies might not lie on any smooth curve (as the theory might suggest) and smoothing and fitting techniques of parametric[16,17] or semi- and non-parametric kind[18] might need to be involved.[19,20] Also, different diagnostic tests might have incomparable associated ROCs in the sense that for some range of cut-off values the first ROC-curve is above the second, whereas for some other range the second ROC-curve is above the first. In addition, there seems to be a desire from the practical side to have a summary measure available for each diagnostic test to be compared. Consequently, measures like the *Area Under the Curve* (AUC) have been suggested and discussed.[20]

3 Revisiting Youden's index

Here, we suggest to use as a measure the simple sum of sensitivity and specificity: $(1-\alpha)+(1-\beta)$, or, equivalently, the sum of the misclassification errors α and β. Note that $(1-\alpha)+(1-\beta)=1+J$, where J is Youden's index.[21] The motivation for this suggestion is as follows. It is known that the best cut-off value (in the sense of maximizing the sum of sensitivity and specificity) is found as one of the points of intersection between the two normal curves describing diseased and healthy population. If the normal distributions have the same variance, this point is simply the arithmetic mean of the two normal distributions, otherwise it is some weighted average of the two means. Therefore, it is seems likely that some 'near to optimal' cut-off value has been chosen in the individual diagnostic studies. If we now look at changes in sensitivity, specificity and their sum as is visualized in Figure 2, we see that the sum remains fairly constant, whereas the individual measures undergo considerable changes.

It can, therefore, be expected that variance inflation in specificity and sensitivity due to cut-off value variation is diminished for the sum of the two, in other words, when using Youden's index.

Youden's index has appreciated theoretical interest over many years. Biggerstaff[22] pointed out that it is in a certain sense the best available summary measure. Hilden and Glasziou[23] give a good geometric characterization of Youden's index as *area under the curve*. They also provide a clinical interpretation of Youden's index as maximum proportional reduction in expected regret. However, clinicians appreciate Youden's index in terms of simplicity and clarity (for an example in diagnostic studies of asthma see Pekkanen and Pearce).[24]

3.1 Diagnostics of cut-off value problem

In practice, the cut-off value itself is often not reported and some other device is required as indicator for a cut-off value problem. To detect a cut-off value problem it was suggest to plot specificity against sensitivity, since variation in the cut-off value would lead to higher values of sensitivity corresponding to lower values of specificity and *vice versa*. To diagnose if taking the sum of specificity and sensitivity has diminished the cut-off value problem, it is suggest to plot specificity and sensitivity and their

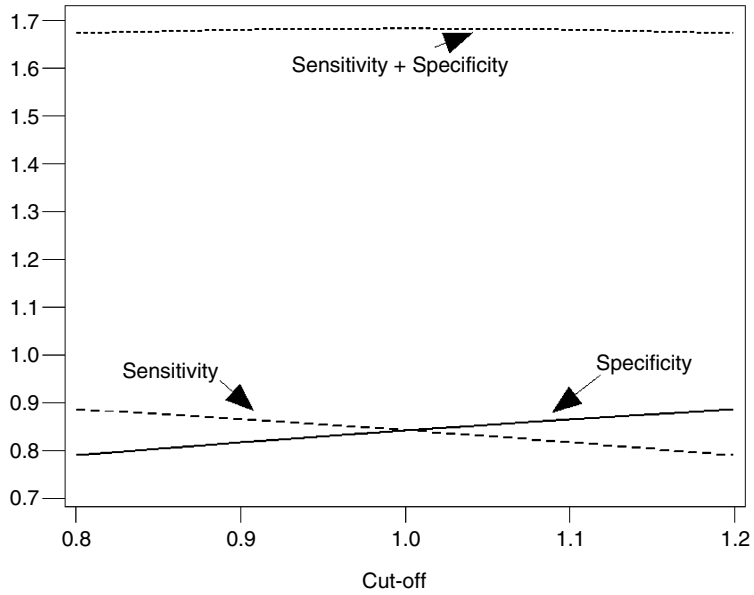

Figure 2 Sensitivity, Specificity and their sum as a function of the cut-off value.

sum against the difference of the two, since when the cut-off values varies from $-\infty$ to $+\infty$, the difference between specificity and sensitivity will vary from -1 to $+1$. In other words, it is suggested to use the change in difference between specificity and sensitivity as a surrogate measure for change in the cut-off value.

3.1.1 DDU-meta-analysis (continued) These techniques at hand of the meta-analysis of 14 studies on the accuracy of duplex doppler ultrasound (DDU) with angiography as the standard for stroke prevention will be demonstrated, as described above. Figure 3 shows a negative trend, at least for higher values of sensitivity. If the sum of specificity and sensitivity is plotted against their difference, the envisaged effect of cut-off value variation is diminished, at least for a wider range of the study data of the DDU-meta-analysis.

3.2 Estimation of $\alpha + \beta$ under homogeneity

With this motivation in mind, it is assumed that there are study-specific error rates, α_i and β_i, but there is a homogeneous total of both rates over all studies: $\alpha_i + \beta_i = \lambda$ for all $i = 1, ..., k$. One is interested in developing a *good* estimator for λ, where 'good' needs to be specified. A simple, pooled estimator of the form $\sum_i x_{D,i}/\sum_i n_{D,i} + \sum_i x_{H,i}/\sum_i n_{H,i}$ might be confounded by study effects and should, therefore, be avoided. We, therefore, prefer an estimator of the form

$$\frac{\sum_{i=1}^{k} w_i \left(x_{D,i}/n_{D,i} + x_{H,i}/n_{H,i}\right)}{\sum_{i=1}^{k} w_i} \tag{1}$$

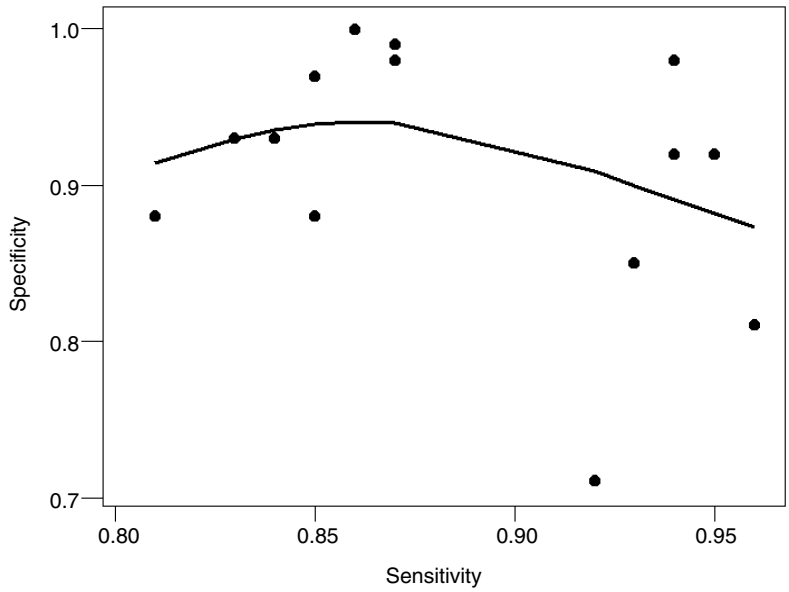

Figure 3 Sensitivity versus Specificity for DDU-meta analysis (line corresponds to LOWESS-smoother).

with $w_i \geq 0$ for all $i = 1, \ldots, k$. Following Böhning and Sarol[25] for the risk difference one write

$$\frac{x_{D,i}}{n_{D,i}} + \frac{x_{H,i}}{n_{H,i}} = \frac{(n_{D,i}x_{H,i} + n_{H,i}x_{D,i})/n_{+i}}{n_{H,i}n_{D,i}/n_{+i}}$$

for all i, with $n_{+i} = n_{D,i} + n_{H,i}$ to arrive at the Mantel–Haenszel-type estimator by taking sums before ratios:

$$\hat{\lambda}_{MH} = \frac{\sum_{i=1}^{k} \left(n_{D,i}x_{H,i} + n_{H,i}x_{D,i}\right)/n_{+i}}{\sum_{i=1}^{k} \left(n_{H,i}n_{D,i}\right)/n_{+i}}. \tag{2}$$

Note that equation (2) is a weighted estimator of the form (1) with $w_i = (n_{D,i}n_{H,i})/n_{+i}$, and, since the weights are non-random, it is *unbiased*. In addition, its variance is readily available:

$$\mathrm{Var}(\hat{\lambda}_{MH}) = \frac{\sum_{i=1}^{k} \left(n_{D,i}^2 n_{H,i}\beta_i(1-\beta_i) + n_{H,i}^2 n_{D,i}\alpha_i(1-\alpha_i)\right)/n_{+i}^2}{\left(\sum_{i=1}^{k} n_{H,i}n_{D,i}/n_{+i}\right)^2} \tag{3}$$

and, estimating $\beta_i(1-\beta_i)$ by $x_{H,i}/n_{H,i}$ and $\alpha_i(1-\alpha_i)$ by $x_{D,i}/n_{D,i}$ (assuming that α_i

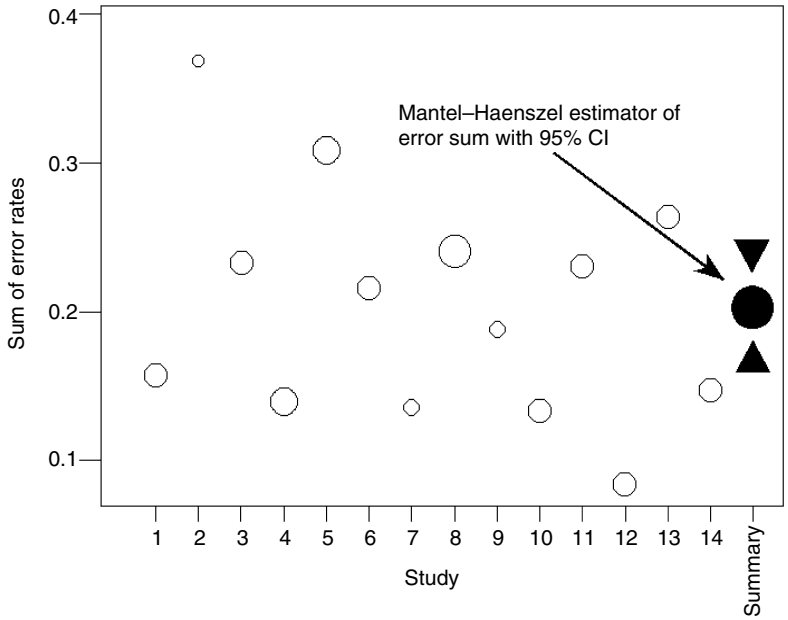

Figure 4 Sum of error rates $\hat{\alpha}_i + \hat{\beta}_i$ for studies $i = 1, ..., 14$ and Mantel–Haenszel estimator of summary error rate with 95% CI.

and β_i are small), the variance estimator

$$\widehat{\mathrm{Var}}(\hat{\lambda}_{MH}) = \frac{\sum_{i=1}^{k} \left(n_{D,i}^2 x_{H,i} + n_{H,i}^2 x_{D,i} \right) / n_{+i}^2}{\left(\sum_{i=1}^{k} n_{H,i} n_{D,i} / n_{+i} \right)^2} \tag{4}$$

can be obtained. Note, as a 'good' property that not only equation (2) is less affected by the occurrence of zeros than the optimal, inverse-variance weighted estimator – as one would expect from a Mantel–Haenszel-type estimator, but also the variance-estimate (4) has this property.

3.2.1 The Mantel–Haenszel estimator for the DDU-meta-analysis Let one consider equation (2) for the data from the DDU-meta-analysis. One finds here that $\hat{\lambda}_{MH} = 0.2101$ with estimated variance $\widehat{\mathrm{Var}}(\hat{\lambda}_{MH}) = 0.00027$ and associated 95%-confidence interval 0.1778–0.2424. This is illustrated in Figure 4, which also illustrates the gain in efficiency achieved by the Mantel–Haenszel-estimator, if homogeneity holds.

4 A likelihood approach coping with heterogeneity

Let $X_i = x_{D,i} + x_{H,i}$ denote the sum of misclassifications in study i. Then, $E(X_i) = n_{D,i}\alpha_i + n_{H,i}\beta_i$ holds, and, if the studies are balanced ($n_{D,i} = n_{H,i} = n_i$), then

$$E(X_i) = n_{D,i}\alpha_i + n_{H,i}\beta_i = n_i(\alpha_i + \beta_i) = n_i\lambda_i,$$

where $\lambda_i = \alpha_i + \beta_i$ is the sum of the two misclassification rates in study i. Furthermore, it appears justified to consider a Poisson assumption for X_i *conditional upon study* i, so that

$$P(X_i = x_i| \text{ study } i) = e^{-\lambda_i n_i}(\lambda_i n_i)^{x_i}/x_i!. \tag{5}$$

Furthermore, let λ arise from a random variable Λ_i with probability density function $q(\lambda)$, so that the joint distribution of X_i and Λ_i is provided as

$$P(X_i = x_i, \Lambda_i = \lambda) = P(X_i = x_i|\Lambda_i = \lambda) \quad q(\lambda) = \frac{e^{-\lambda n_i}(\lambda n_i)^{x_i}}{x_i!} q(\lambda)$$

with associated marginal distribution over Λ_i as

$$P(X_i = x_i) = \int_\lambda \frac{e^{-\lambda n_i}(\lambda n_i)^{x_i}}{x_i!} q(\lambda) \mathrm{d}(\lambda). \tag{6}$$

One considers the *nonparametric likelihood* for any distribution Q with associated probability density function $q(\lambda)$

$$L(Q) = \prod_{i=1}^{k} \int_\lambda \frac{e^{-\lambda n_i}(\lambda n_i)^{x_i}}{x_i!} q(\lambda) \mathrm{d}(\lambda), \tag{7}$$

which needs to be maximized over all probability distributions Q. For the approach in general, see Lindsay[26] or Böhning.[27] Technically, one considers the likelihood for discrete mixtures

$$L(Q) = \prod_{i=1}^{k} \sum_{j=1}^{m} \frac{e^{-\lambda_j n_i}(\lambda_j n_i)^{x_i}}{x_i!} q_j, \tag{8}$$

which needs to be maximized in $\lambda_1, ..., \lambda_m$, $q_1, ..., q_m$, and m, where the restriction on discrete distributions is no limitation of generality. The general theory delivers here a unique nonparametric maximum likelihood estimator with a certain number of discrete mass points m. The EM algorithm[28] might be used with a specific number of mass points m. This may be combined with a forward strategy, starting with the smallest possible number of mass points $m = 1$, the case of homogeneity, and then working upwards by increasing m one by one.[29] Note that in the case of homogeneity, the maximum likelihood estimator and the proposed Mantel–Haenszel estimator (2) coincide.

4.1 The mixture approach for the DDU-meta-analysis

To apply the mixture approach on X_i, for this example, one needs to cope with the situation that most of the studies involved in the meta-analysis are unbalanced so that one needs to use an approximation such as $\overline{n}_i = (n_{D,i} + n_{H,i})/2$. A comparison between the estimated Poisson variances on the basis of $n_{D,i}$ and $n_{H,i}$ with those on the basis of $\overline{n}_i$ shows that in most studies the approximation appears to be reasonably good.

Table 2 Heterogeneity estimated via the nonparametric mixture approach; m are the number of components in the mixing distribution and $\hat{\lambda}_j$ and $\hat{q}_j$ are the estimates of the parameters in the mixing distribution for $j = 1, ..., m$; for $m = 2$ the NPMLE is reached

m	$\hat{\lambda}_j$	$\hat{q}_j$	log-likelihood
1	0.1824	1	−44.1221
2	0.1181	0.5053	−40.1592
	0.2325	0.4947	

Table 2 shows the results for an analysis using mixture models of the form (8). For the homogeneity case $m = 1$, the maximum likelihood estimator is simply $\hat{\lambda} = \sum_i x_i / \sum_i \overline{n}_i = 0.1824$. For $m = 2$, NPMLE is already found to indicate that increasing the number of components will not lead to any increase in the likelihood. NPMLE gives almost equal mass to a cluster of studies with a small total misclassification error and a cluster with a larger size of misclassification error. The last column in Table 1 shows the classification based on the maximum posterior probability (MAP). There appears to be a good separation between studies with high and low diagnostic accuracy.

5 Discussion

5.1 Alternative measures of misclassification error

Frequently, a different measure for the discriminatory power is suggested: the sum of log-odds of sensitivity and log-odds of specificity. The suggestion goes back to Hasselblad and Hedges[9] and has found its entry in the literature including guidelines[8,30,31] under the term *diagnostic odds ratio* (DOR). Glas[32] *et al.* recommend the DOR as a single measure of diagnostic test performance, in particular for application in meta-analysis. It is also used more frequently in practical meta-analysis.[33] It also plays a core part in estimating the *summary receiver operating characteristic* (SROC).[1,34] The motivation behind this measure is that it relates the odds that the test is positive in the diseased population to the odds that the test is positive in the healthy population. It was noted by Edwards[35] as well as Hasselblad and Hedges[9] amongst others that this measure is remarkably constant in the vicinity of the optimal cut-off value. However, care must be taken if it is used as an advice for selecting an optimal cut-off value, as Figure 5 shows. Here, at the optimal cut-off value the lowest value of the log-odds ratio is found.

This is in contrast to the sum of log-sensitivity and log-specificity, where the optimal cut-off value occurs at the largest value of the sum of sensitivity and specificity. Both approaches become equivalent when specificity and sensitivity are high (such as when both populations are well seperated).

5.2 The approach in the light of alternatives

Note that in the balanced case $n_{D,i} = n_{H,i} = n_i$ for all $i = 1, \ldots, k$ and for the situation of homogeneity the Mantel–Haenszel estimator (2) agrees with the maximum

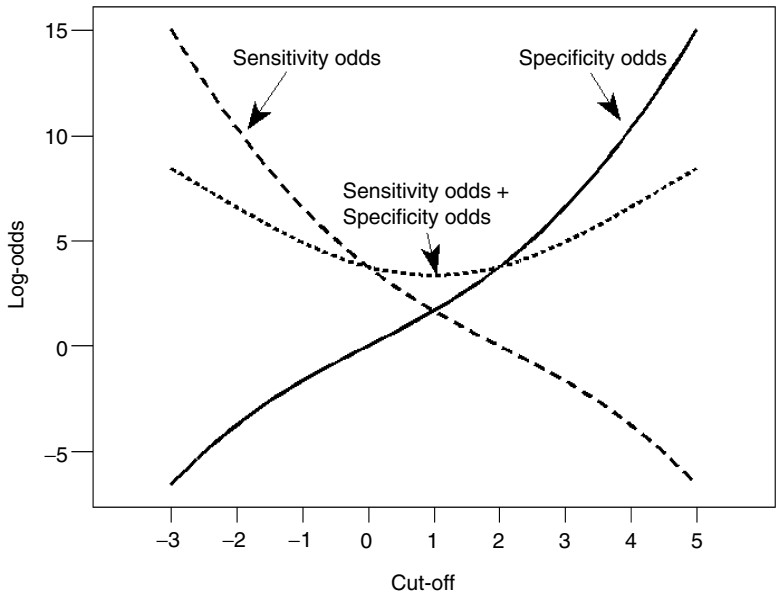

Figure 5 Sensitivity odds, Specificity odds and their sum (on log-scale) as a function of the cut-off value for two normal distributions with mean 0 (healthy) and mean 2 (diseased).

likelihood estimator for the 'singular' mixture $m = 1$. Consequently, one might consider the mixture model as a generalization of the Mantel–Haenszel approach. The measure *sum of the error rates* was chosen since it is less prone to the cut-off value problem. Consequently, any residual heterogeneity is less likely spurious in the sense that it has been created by a variation of the cut-off value. In the case of the meta-analysis of the DDU-studies, heterogeneity consisting of two components with high and low diagnostic accuracy could be identified and it seems unlikely that this heterogeneity can be explained entirely on the basis of a cut-off value variation.

It has been mentioned that ROC (or SROC in meta-anaylsis) is used frequently. To achieve a summary measure for a given ROC, the area-under-the-curve (AUC) is used. Greiner[20] mentions the result of Hilden and Glasziou[23] that if there is only one point in the ROC-space and the ROC-curve is estimated by connecting the three existing points, then the estimated AUC corresponds to the average of estimated sensitivity and specificity. Pepe[36] points out the connection of Youden's index to the Kolomogorov-Smirnov statistic.

In a recent contribution, Le[37] recovers and summarizes several favourable properties of Youden's index. He points out that a point $(u, v) = (\beta, 1 - \alpha)$ on the ROC curve is connected to J as

$$J = v - u = 1 - \alpha - \beta$$

so that a natural modelling of the ROC curve by $v = R(u)$, where $R(.)$ denotes such a model for the ROC curve, might be used to find an optimal (minimal) cut-off value.

Finally, one notes that the proposed measure allows an interesting way for constructing a confidence interval. Let X again denote the sum of the errors for one study. Assuming $E(X) = \mathrm{Var}(X) = n\lambda$, it is straightforward to construct a confidence interval on the basis of the pivotal statistic $(X - n\lambda)/(\sqrt{n\lambda})$:

$$\mu_{L,R} = X + z^2/2 \pm z\sqrt{X + z^2/4}$$

provides a $(1-\alpha)100\%$ confidence interval for $n\lambda$, where $z = \Phi^{-1}(1-\alpha)$. This is in contrast to the risk difference, for example, where this construction is not possible since the variance is not a function of the risk difference only. Here, the variance is only the function of the sum of the error rates and allows the above pivotal construction of a confidence interval.

Acknowledgments

This work was supported by the *German Research Foundation*.

References

1 Sutton AJ, Abrams KR, Jones DR, Sheldon TA, Song F. *Methods for Meta-Analysis in Medical Research*. Wiley, 2000.
2 Schulze R, Holling H, Böhning D. eds. *Meta-Analysis. New Developments and Applications in Medical and Social Sciences*. Hogrefe & Huber, 2003.
3 Stangl DK, Berry, DA. eds. *Meta-Analysis in Medicine and Health Policy*. Marcel Dekker, 2000.
4 Whitehead A. *Meta-Analysis of Controlled Clinical Trials*. Wiley, 2002.
5 Böhning D, Malzahn U, Schlattmann P, Dammann U-P, Mehnert W, Holling H, Schulze R. The application of statistical methods of meta-analysis for heterogeneity modelling in Medicine and Pharmacy, Psychology, Quality Control and Assurance. In: W Jäger, H-J Krebs eds. *Mathematics, Key Technology for the Future, Joint Projects between Universities and Industry*. Springer, 2003: 533–553.
6 Normand S-LT. Tutorial in Biostatistics: Meta-Analysis: Formulating, Evaluating, Combing, and Reporting. *Statistics in Medicine* 1999; **18**: 321–9.
7 Deeks JJ. Systematic Reviews of Evaluation of Diagnostic and Screening Tests. *British Medical Journal* 2001; **323**: 157–62.
8 Irwig L, Tosteson AN, Gatsonis C, Lau J, Colditz G, Chalmers TC, Mosteller F. Guidelines for Meta-analyses Evaluating Diagnostic Tests. *Annals of Internal Medicine* 1994; **120**: 667–76.
9 Hasselblad V, Hedges LV. Meta-Analysis of Screening and Diagnostic Tests. *Psychological Bulletin* 1995; **117**: 167–78.
10 Midgette AS, Stukel TA, Littenberg B. A Meta-analytic Method for Summarizing Diagnostic Test Performances: Receiver-operating-characteristc-summary point estimates. *Medical Decision Making* 1993; **13**: 253–7.
11 Harbord RM, Deeks JJ, Egger M, Whiting P, Sterne JAC. A Unification of Models for Meta-Analysis of Diagnostic Accuracy Studies. *Biostatistics* 2006; **1**: 1–21.
12 Van Houwelingen HC, Zwinderman KH, Stijnen T. A bivariate approach to meta-analysis. *Statistics in Medicine* 1993; **12**: 2273–84.
13 Reitsma JB, Glas AS, Rutjes AWS, Scholten RJPM, Bossuyt PM, Zwinderman AH. Bivariate analysis of sensitivity and specificity produces informative measures in diagnostic reviews. *Journal of Clinical Epidemiology* 2005; **58**: 982–90.

14 Rutter CM, Gatsonis CA. A hierarchical regression approach to meta-analysis of diagnostic test accuracy evaluations. *Statistics in Medicine* 2001; **20**: 2865–84.
15 Irwig L, Macaskill P, Glasziou P, Fahey M. Meta-analytic Methods for Diagnostic Test Accuracy. *Journal of Clinical Epidemiology* 1995; **48**: 119–30.
16 McCullagh P. Regression Models for Ordinal Data. *Journal of the Royal Statistical Society, Series B* 1980; **42**: 109–42.
17 Tosteson A, Begg C. A General Regression Methodology for ROC Curve Estimation. *Medical Decision Making* 1998; **8**: 204–15.
18 Zou K, Hall W, Shapiro D. Smooth Non-parametric ROC curves for continuous diagnostic tests. *Statistics in Medicine* 1997; **16**: 2143–56.
19 Pepe MS. Receiver Operating Characteristic Methodology. *Journal of the American Statistical Association* 2000; **95**: 308–11.
20 Greiner M. *Serodiagnostische Tests*. Springer, 2003.
21 Youden D. Index for Rating Diagnostic Tests. *Cancer* 1950; **3**: 32–5.
22 Biggerstaff BJ. Comparing diagnostic tests: a simple graphic using likelihood ratios. *Statistics in Medicine* 2000; **19**: 649–63.
23 Hilden J, Glasziou P. Regret graphs, diagnostic uncertainty and Youden's index. *Statistics in Medicine* 1996; **15**: 969–86.
24 Pekkanen J, Pearce N. Defining asthma in epidemiological studies. *European Respiratory Journal* 1999; **14**: 951–57.
25 Böhning D, Sarol J. Estimating risk difference in multicenter studies under baseline heterogeneity. *Biometrics* 2000; **56**: 304–08.
26 Lindsay BG. *Mixture Models: Theory, Geometry, and Applications*. Institute of Statistical Mathematics, 1995.
27 Böhning D. *Computer-Assisted Analysis of Mixtures and Applications. Meta-Analysis, Disease Mapping and Others*. Chapman & Hall/CRC, Boca Raton, 2000.
28 Dempster AP, Laird NM, Rubin DB. Maximum Likelihood Estimation from Incomplete Data Via the EM Algorithm (with Discussion). *Journal of the Royal Statistical Society B*: 1977; **39**: 1–38.
29 Böhning D. The EM Algorithm with Gradient Function Update for Discrete Mixtures with Known (Fixed) Number of Components. *Statistics and Computing* 2003; **13**: 257–65.
30 Deville WL, Buntinx F, Bouter LM, Montori VM, de Vet HCW, van der Windt DAWM, Bezemer PD. Conducting Systematic Reviews of Diagnostic Studies: Didactic Guidelines. *BMC Medical Research Methodology* 2002; **2**: 9.
31 NRC Committee on Applied and Theoretical Statistics. *Combining Information: Statistical Issues and Opportunities for Research*, National Academy Press, 1992.
32 Glas AS, Lijmer JG, Prins MH, Bonsel GJ, Bossuyt PMM. The diagnostic odds ratio: a single indicator of test performance. *Journal of Clinical Epidemiology* 2003; **56**: 1129–35.
33 Cruciani M, Marcati P, Malena M, Bosco O, Serpelloni G, Mengoli C. Meta-analysis of Diagnostic Procedures for *Pneumocystis carninii* penumonia in HIV-1-infected Patients. *European Respiratory Journal* 2002; **20**: 982–89.
34 Dukic V, Gatsonis C. Meta-analysis of Diagnostic Test Accuracy Assessment Studies with Varying Number of Thresholds. *Biometrics* 2003; **59**: 936–46.
35 Edwards JH. Some Taxonomic Implications of a Curious Feature of the Bivariate Normal Surface. *British Journal of Prevention and Social Medicine* 1966; **20**: 42.
36 Pepe MS. *The Statistical Evaluation of Medical Tests for Classification and Prediction*. Oxford University Press, 2003.
37 Le CT. A solution for the most basic optimization problem associated with an ROC curve. *Statistical Methods in Medical Research* 2006; **15**: 571–84.