Meta-Analysis of Binary Data Based upon Dichotomized Criteria

Heinz Holling[1], Dankmar Böhning[2], and Walailuck Böhning[1]

[1]University of Münster, Germany, [2]University of Reading, UK

Abstract. This paper considers meta-analysis of binary data that use a dichotomized continuous score. Classification into two categories, e.g., qualified or not qualified, is often based upon a threshold or cut-off value. This threshold might vary between studies since intentionally different values are used. However, conventional meta-analysis methodology analyzing sensitivity and specificity separately might then be confounded by a potentially unknown variation of the cut-off value.
In order to cope with varying thresholds, an overall estimate of the misclassification error is suggested instead, which is equivalent to the well-known Youden index. It is argued that this index is less prone to between-study variation of cut-off values. To adjust for potential study effects a Mantel-Haenszel estimator of the overall misclassification error is suggested. Arguments are illustrated using, as an example, the diagnosis of alcoholism using the Alcohol Use Disorders Identification Test (AUDIT).

Keywords: Mantel-Haenszel approach, Youden index, meta-analysis

Introduction

In everyday life the outcomes of many decisions are binary. Medical decisions may result in healthy or diseased, aptitude tests in qualified or not qualified, and applications for a job in accepted or not accepted. Often, such decisions are based on continuous criteria such as scores of medical or psychological tests that are dichotomized by using a cut-off value. Sensitivity and specificity are appropriate measures to determine validity of such diagnoses. In general, *sensitivity* is defined as P(decision is positive | condition is given) = $(1 - \alpha)$ and *specificity* as P(decision is negative | condition is not given) = $(1 - \beta)$.

Since both concepts are very often used to validate clinical decisions, sensitivity can be specifically defined as P (positive|diseased) and specificity as P (negative|nondiseased), respectively. Then sensitivity refers to the capability of the diagnostic test to recognize a diseased person correctly, whereas specificity measures the capability of diagnosing a healthy person correctly with the diagnostic test. Accordingly, α is the error probability of falsely identifying a diseased person as healthy and β is the error probability of falsely classifying a healthy person as diseased. Ideally, both α and β should be as small as possible. Note that the sum $\alpha + \beta$ is equivalent to the *Youden index* (Youden, 1950), which is defined as $Y = (1 - \alpha) + (1 - \beta) - 1$. The subtraction of one is done in order to ease the interpretation of the measure. Y varies between –1 (maximum error) and +1 (minimum error).

Let diagnostic study i be part of a series of k studies, x_i^H the frequency of (falsely) positively classified persons out of n_i^H healthy ones and x_i^D be the frequency of (falsely) negatively classified persons out of n_i^D diseased ones. Then, natural estimates for α_i and β_i are provided as $\hat{\alpha}_i = x_i^D/n_i^D$ and $\hat{\beta}_i = x_i^H/n_i^H$.

Meta-analysis is a common method to integrate results from different studies, (e.g., Cooper & Hedges, 1994; Hunter & Schmidt, 2004; Schulze, Holling, & Böhning, 2003). Various statistics such as standardized mean differences, correlation coefficients, or probabilities can be pooled in a meta-analysis. In addition to these more common statistics, meta-analytic methods for summarizing coefficients representing relationships of binary variables, like the odds ratio, phi coefficient, and so forth, have also been developed (see Haddock, Rindskopf, & Shadish, 1998; Sánchez-Meca, Marín-Martínez, & Chacón-Moscoso, 2003).

However, pooling of sensitivity and specificity of studies may be invalid if different cut-off values were used across studies to dichotomize the underlying continuous variables (e.g., Lijmer, Bossuyt, & Heisterkamp, 2002). Many researchers are aware of this problem and, as a consequence, they include only studies with an identical cut-off score in their meta-analyses. But omitting studies with different cut-off values, which are optimal for just theses studies, might lead to underestimating the validity.

Issues with Using Different Cut-Off Scores

If different cut-off scores are used, sensitivity and specificity are not directly comparable. This is illustrated in Figure

DOI 10.1027/0044-3409.215.2.122

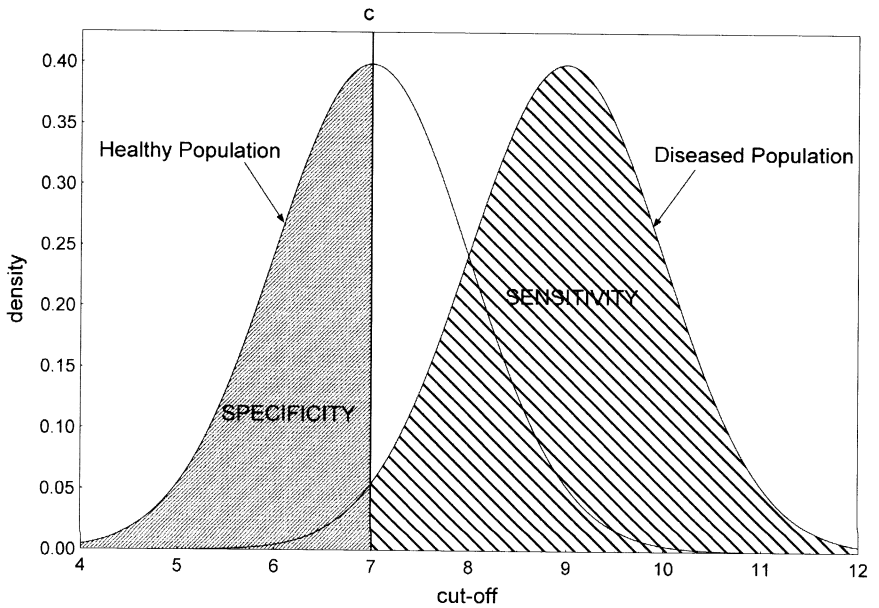

Figure 1. Diagnostic situation with two normal score distributions: The left distribution shows the healthy population (mean seven, variance of one) and the right one shows the diseased population (mean of nine and variance of one).

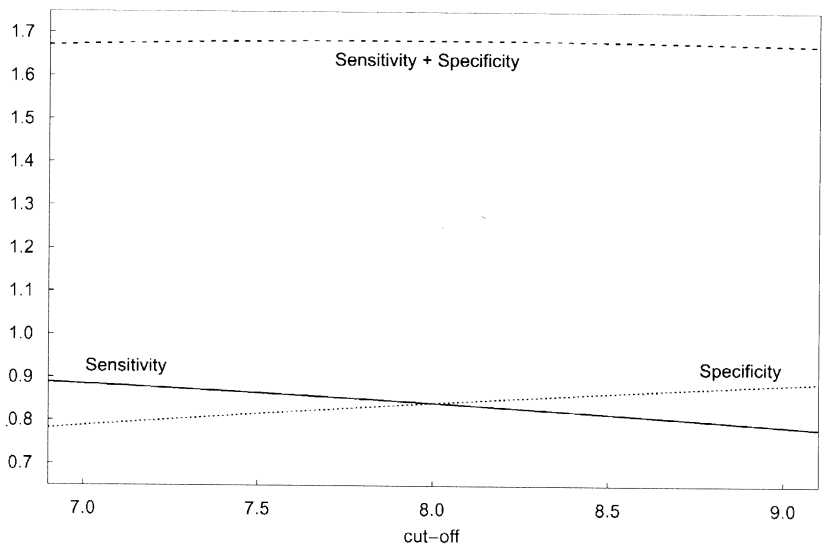

Figure 2. Sensitivity, specificity, and their sum as a function of the cut-off value.

1 where two normal distributions are shown with equal variance of 1 and different means of 7 and 9, respectively.

Assuming that values above the threshold c indicate positivity of the test, a cut-off value c (in Figure 1 $c = 7$) leads to sensitivity given by $1 - \alpha = 1 - \Phi\,((c - \mu^D)/\sigma^D)$ and specificity as given by $1 - \beta = \Phi\,((c - \mu^H)/\sigma^H)$, where Φ is the cumulative distribution function of the corresponding normal distribution. A shift of c to the right leads to a decrease of the sensitivity, whereas the specificity would increase. A shift of c to the left would result in opposite consequences. Thus, quite different sensitivities and specificities might be used in different studies due to the fact that a different cut-off score is used to make a decision. Furthermore, the underlying diagnostic test might have entirely identical discriminating power. To illustrate, consider Figure 1. Here, the discriminating power is shown for a diagnostic test with cut-off value $c = 7$. The discriminating power is identical if c changes to value of $c = 9$, since the sum of the areas under the curve associated with sensitivity and specificity is identical to sum of the corresponding areas shown in Figure 1. Unless it is verifiable that a common true cut-off value is used in different studies, a separate meta-analysis of sensitivity and specificity might lead to biased findings such as artificial heterogeneity of effects or spurious variance inflation.

An Alternative Measure

Instead of separately analyzing sensitivity and specificity, we suggest using the sum of sensitivity and specificity:

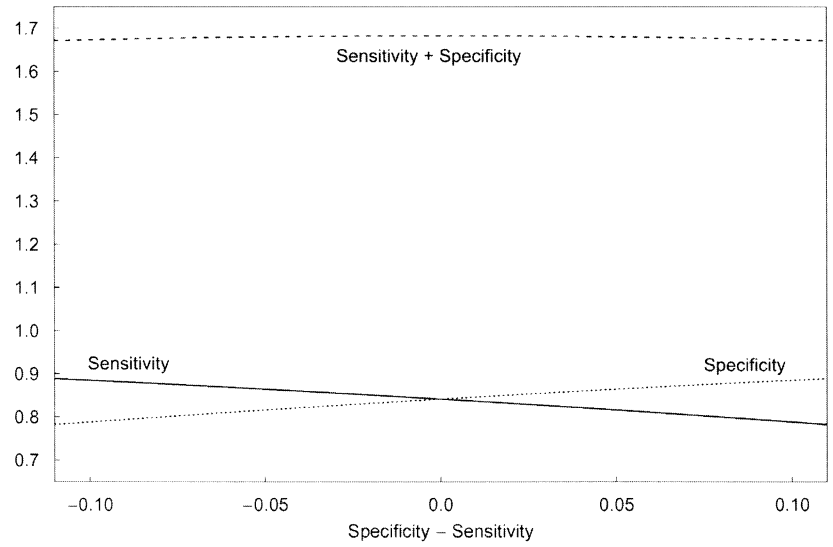

Figure 3. Sensitivity, specificity, and their sum versus their difference.

$(1 - \alpha) + (1 - \beta)$, or, equivalently, the sum of the misclassification errors α and β. This suggestion is motivated as follows. It is well known that the ideal cut-off value (in the sense of maximizing the sum of sensitivity and specificity) is the point of intersection between the two normal curves describing the diseased and healthy subpopulations (Hasselblad & Hedges, 1995). When both normal distributions have the same variance this point is simply the mean of the two normal distributions, otherwise it is some weighted average of the two means. As a consequence, it is likely that some near optimal cut-off value is chosen in individual diagnostic studies. Looking 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 show considerable changes. It can, therefore, be expected that the inflation of the variance in specificity and sensitivity caused by cut-off value variation is attenuated for the sum of the two.

In practice the cut-off value itself is sometimes not reported. In such situations some other indicator for the cut-off value variation can be determined. To detect the cut-off value variation we suggest plotting specificity against sensitivity values reported in the k studies since variation in the cut-off value will lead to higher values of sensitivity corresponding to lower values of specificity, and vice versa.

In order to decide if taking the sum of specificity and sensitivity has diminished the cut-off value problem, we recommend plotting specificity and sensitivity and their sum against their difference. The reason is that 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, we suggest using the change in difference between specificity and sensitivity as a surrogate measure for change in the cut-off value. This is demonstrated in Figure 3 for values in the vicinity of the optimal cut-off value.

An Example: Using the New Measure in Meta-Analysis of Studies on the Alcohol Use Disorder Identification Test (AUDIT)

In 1982, the World Health Organization asked an international group of researchers to develop a simple screening instrument to identify people with hazardous and harmful alcohol use, as well as possible dependence (Babor, de la Fuente, Saunders, & Grant, 1989). The resulting instrument, the Alcohol Use Disorder Identification Test (AUDIT), was carefully developed and evaluated over a period of 20 years. Special attention was paid to gender appropriateness and cross-national generalizability. The AUDIT has also been translated into several languages.

The AUDIT is consistent with ICD-10 definitions of alcohol dependence and harmful alcohol use and focuses on recent alcohol consumption. It is designed as a screening test specifically for use in primary care settings. The AUDIT consists of 10 questions about recent alcohol use, alcohol dependence symptoms, and alcohol-related problems (see Table 1).

Each of the questions has a set of five response options, where the associated scores range from 0 to 4. All the scored responses are added up to arrive at the total AUDIT score. Sensitivities and specificities can be computed for several criteria (e.g., average daily alcohol consumption, recurrent intoxication, presence of at least one dependence symptom, or diagnosis of alcohol abuse or dependence). Also, various cut-off points in total scores can be considered to identify the value with optimal sensitivity and specificity to distinguish hazardous and harmful alcohol use. As a result of this variation of the cut-off for finding an optimal value, total scores of 8 or higher are recommended as indi-

Table 1. Items of the Alcohol Use Disorders Identification Test

Items	Responses 0	1	2	3	4
1. How often do you have a drink containing alcohol?	Never	Monthly or less	2–4 times a month	2–3 times a week	4 or more times a week
2. How many drinks containing alcohol do you have on a typical day when you are drinking?	1 or 2	3 or 4	5 or 6	7 to 9	10 or more
3. How often do you have six or more drinks on one occasion?	Never	Less than monthly	Monthly	Weekly	Daily or almost daily
4. How often during the last year have you found that you were not able to stop drinking once you had started?	Never	Less than monthly	Monthly	Weekly	Daily or almost daily
5. How often during the last year have you failed to do what was normally expected of you because of drinking?	Never	Less than monthly	Monthly	Weekly	Daily or almost daily
6. How often during the last year have you needed a first drink in the morning to get yourself going after a heavy drinking session?	Never	Less than monthly	Monthly	Weekly	Daily or almost daily
7. How often during the last year have you had a feeling of guilt or remorse after drinking?	Never	Less than monthly	Monthly	Weekly	Daily or almost daily
8. How often during the last year have you been unable to remember what happened the night before because of your drinking?	Never	Less than monthly	Monthly	Weekly	Daily or almost daily
9. Have you or someone else been injured because of your drinking?	No		Yes, but not in the last year		Yes, during the last year
10. Has a relative, friend, doctor or other health care worker been concerned about your drinking or suggested you cut down?	No		Yes, but not in the last year		Yes, during the last year

cators of hazardous and harmful alcohol use, as well as possible alcohol dependence.

Several studies have reported reliability indices for the AUDIT total scores (for an overview, see Shields & Caruso, 2003). The results indicate a reliability of about .80.

Berner, Kriston, Bentele, and Härter (2007) recently published a very carefully conducted meta-analysis of previous validity studies on the AUDIT. They identified 429 potentially relevant papers published between 1995 and 2004 by electronic database searches. According to the authors, a set of 19 studies turned out to be appropriate with respect to their inclusion and exclusion criteria. The authors performed two meta-analyses: (1) for 17 studies using a fixed cut-off of 8 as recommended in the AUDIT test manual ("general cut-off") and (2) for 16 studies (see Table 2) using the cut-off value that has been determined as optimal for the examined study population ("specific cut-off").

The studies were classified into seven subgroups according to their setting (see Table 3): general primary care, general medical inpatients, emergency department/trauma center, adolescents/college students, elderly patients, patients with mental disorder, and other highly selected samples.

Berner et al. (2007) pooled sensitivities and specificities by computing weighted averages in homogeneous study results only. Statistical heterogeneity was assessed using likelihood ratio tests (Higgins, Thompson, Deeks, & Altmann, 2003). Furthermore, the authors computed positive likelihood ratios $LR^+ = Sen/(1 - Spec)$, negative likelihood ratios $LR^- = (1 - Sens)/Spe$ as well as diagnostic odds ratios $DOR = LR^+/LR^-$. Positive (negative) likelihood ratios measure the frequency of a positive (negative) diagnosis among subjects with a disease compared to those without a disease. Diagnostic odds ratios combine both likelihood ratios to a single measure of diagnostic test performance and indicate how much greater the odds of having a disease are

Table 2. Studies of the Alcohol Use Disorder Identification Test for Detecting At-Risk Drinking taken from the meta-analysis by Berner, Kriston, Bentele, and Härter (2007)

Study	First author	$1 - x_i^D/n_i^D$	n_i^D	$1 - x_i^H/n_i^H$	n_i^H
1	Gual, 2002	0.844	64	0.901	191
2	Contel, 1999	0.900	10	0.815	178
3	Gómez, 2001	0.891	46	0.932	454
4	Rumpf, 2002	0.779	281	0.810	3270
5	Gordon, 2001	0.903	754	0.849	6200
6	Taj, 1998	0.667	75	0.692	26
7	Daeppen, 2000	0.727	77	0.929	255
8	MacKenzie, 1996	0.929	28	0.938	211
9	Neumann, 2004	0.765	260	0.825	1667
10	Kokotailo, 2004	0.909	88	0.598	214
11	Bradley, 1998	0.867	105	0.699	156
12	Philpot, 2003	0.667	18	0.964	110
13	Skipsey, 1997	0.968	31	0.686	51
14	Hiro, 1996	0.913	23	0.786	70
15	Piccinelli, 1997	0.843	70	0.900	412
16	Bradley, 2003	0.697	89	0.859	304

Table 3. Diagnostic value of the AUDIT in different settings at specific cut-off of 8 taken from the meta-analysis by Berner, Kriston, Benteler, and Härter (2007)

Setting	*k*		SENS (95% CI)	SPEC (95% CI)	LR^+ (95% CI)	LR^- (95% CI)	DOR (95% CI)
Primary care	9	pooled estimate	NA	NA	6.20 (4.78–7.86)	0.22 (0.15–0.33)	29.01 (16.31–51.58)
		heterogeneity (I^2)	87.7%	92%	91.7%	89.8%	89%
Inpatient	1	estimate	0.93 (0.76–0.99)	0.94 (0.90–0.97)	15.07 (8.81–25.77)	0.08 (0.02–0.29)	198.00 (42.29–927.13)
Emergency department	1	estimate	0.77 (0.71–0.82)	0.83 (0.81–0.84)	4.38 (3.87–4.96)	0.28 (0.23–0.35)	15.43 (11.27–21.11)
Young patients	1	estimate	0.91 (0.83–0.96)	0.60 (0.53–0.66)	2.26 (1.90–2.70)	0.15 (0.08–0.30)	14.88 (6.85–32.35)
Elderly patients	2	pooled estimate	NA	NA	6.97 (1.10–41.78)	0.25 (0.14–0.44)	24.21 (7.25–80.86)
		heterogeneity (I^2)	74%	97.1%	91.8%	50.7%	61.4%
Mentally ill patients	1	estimate	0.97 (0.83–1.00)	0.69 (0.54–0.81)	3.08 (2.05–4.65)	0.05 (0.01–0.33)	65.63 (8.21–524.41)
Highly selected	1	pooled estimate	0.91 (0.72–0.99)	0.79 (0.67–0.87)	4.26 (2.67–6.79)	0.11 (0.03–0.42)	38.50 (8.10–182.98)
All studies	16	pooled estimate	NA	NA	5.38 (4.35–6.65)	0.21 (0.16–0.28)	28.65 (18.90–43.42)
		heterogeneity (I^2)	83.4%	93.7%	93.7%	83.9%	84.9%

Notes. AUDIT = Alcohol Use Disorders Identification Test; k = number of studies; SENS = sensitivity; CI = confidence interval; SPEC = specificity; LR^+ = positive likelihood ratio; LR^- = negative likelihood ratio; DOR = diagnostic odds ratio; NA = not applicable because of large heterogeneity.

Figure 4. Log-sensitivity versus log-specificity for the AUDIT meta-analysis using specific cut-offs of Berner, Kriston, Bentele, and Härter (2007). The line corresponds to the unweighted regression of log-sensitivity on log-specificity.

for the subjects with a positive diagnosis than for subjects with a negative diagnosis.

Heterogeneity for the likelihood ratios and diagnostic odds ratio was tested by Cochran's Q test. The amount of heterogeneity for all measures was assessed by $I^2 = 100 \cdot (\chi^2 - df)/\chi^2$. Regardless of heterogeneity, positive and negative likelihood ratios as well as diagnostic odds ratios were pooled using a random-effects model based on the DerSimonian and Laird (1986) approach.

Table 3 shows the main results of the meta-analysis of 16 studies using the optimal cut-off value ("specific cut-off"). Because of heterogeneity, pooling of sensitivities and specificities was not appropriate. The pooled positive likelihood ratio resulted in a value of 5.38 and 0.21 for the pooled negative likelihood ratio, respectively. The diagnostic odds ratio was 28.65.

The overall variance explained by study settings was 24.4%. Since setting could not account for heterogeneity Berner et al. (2007) conclude that heterogeneity has to be explained by other factors such as population-specific variables, methodological issues, or other effects.

One source of heterogeneity might be varying thresholds. In order to test this assumption, sensitivity is plotted against specificity for the data reported by Berner et al. (2007). As Figure 4 shows, there is a clear negative trend. When plotting the sum of specificity and sensitivity against their difference (see Figure 5), the effect of the cut-off value variation is very much attenuated.

Estimation of $\alpha + \beta$

Now 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 $i = 1, \ldots, k$. Note that we only assume that the sum $\alpha_i + \beta_i = \lambda$ is independent of the study, whereas the study-specific error rates α_i and β_i

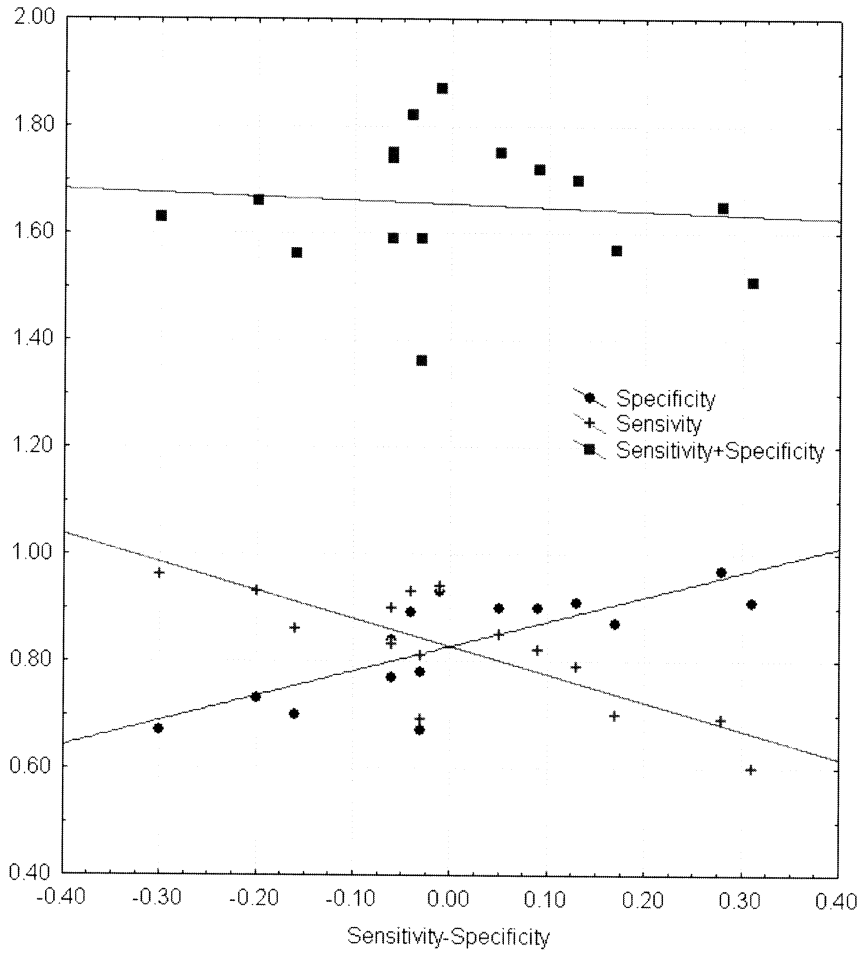

Figure 5. Sensitivity, specificity, and their sum versus their difference for the AUDIT meta-analysis using specific cut-offs of Berner, Kriston, Bentele, and Härter (2007).

might vary between studies. We want to derive a good estimator for λ, where "good" will be specified as follows. A simple, pooled estimator of the form

$$\frac{\Sigma_i x_i^D}{\Sigma_i n_i^D} + \frac{\Sigma_i x_i^H}{\Sigma_i n_i^H}$$

should be avoided as it might be confounded by study effects. In general, it is state of the art in meta-analysis to investigate between-study heterogeneity in the measure of interest, since such potential heterogeneity could lead to valuable hints in identifying important covariates and moderator variables. Such heterogeneity could be caused by covariates such as difference of persons involved in the studies or differences in the diagnostic procedures.

Instead, we prefer the following estimator:

$$\frac{\sum_{i=1}^{k} w_i \left(\frac{x_i^D}{n_i^D} + \frac{x_i^H}{n_i^H} \right)}{\sum_{i=1}^{k} w_i}, \qquad (1)$$

with $w_i \geq 0$ for all i = 1, . . ., k. Considering

$$\frac{x_i^D}{n_i^D} + \frac{x_i^H}{n_i^H} = \frac{(n_i^D x_i^H + n_i^H x_i^D)/n_i^+}{n_i^H n_i^D / n_i^+}$$

for all i, with $n_i^+ = n_i^D + n_i^H$, we arrive at the Mantel-Haenszel-type estimator by taking sums before ratios:

$$\hat{\lambda}_{MH} = \frac{\sum_{i=1}^{k} (n_i^D x_i^H + n_i^H x_i^D)/n_i^+}{\sum_{i=1}^{k} (n_i^H n_i^D)/n_i^+} \qquad (2)$$

$\hat{\lambda}_{MH}$ is a weighted estimator for the sum of the error rates of the form (1) with $w_i = (n_i^D n_i^H)/n_i^+$, and, since the weights are nonrandom, it is *unbiased*. Furthermore, its variance is

$$Var(\hat{\lambda}_{MH}) = \frac{\sum_{i=1}^{k} (n_i^{D^2} n_i^H \beta_i(1-\beta_i) + n_i^{H^2} n_i^D \alpha_i(1-\alpha_i))/n_i^{+^2}}{(\sum_{i=1}^{k} n_i^H n_i^D / n_i^+)^2} \qquad (3)$$

The error rates α_i and β_i are simply estimated by x_i^D/n_i^D and

x_i^H/n_i^H, respectively, so that $\beta_i(1-\beta_i)$ can be estimated by $x_i^H/n_i^H(1-x_i^H/n_i^H)$ and $\alpha_i(1-\alpha_i)$ by $x_i^D/n_i^D(1-x_i^D/n_i^D)$, an estimator of the variance can be obtained

$$\hat{Var}(\hat{\lambda}_{MH}) = \frac{\sum_{i=1}^{k}(n_i^{D^2}x_i^H(1-x_i^H/n_i^H) + n_i^{H^2}x_i^D(1-x_i^D))/n_i^{+^2}}{(\sum_{i=1}^{k} n_i^H n_i^D/n_i^+)^2} \quad (4)$$

Note that not only is the expression given in (2) 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 that the variance-estimator given in (4) has this property as well.

Let us consider (2) for the data from the AUDIT-meta-analysis. We find here that $\hat{\lambda}_{MH} = 0.327$ with estimated variance of $\hat{Var}(\hat{\lambda}_{MH}) = 0.00884$ and associated 95% confidence interval limits 0.310–0.345. This rather small confidence interval also illustrates the increase in efficiency achieved by the Mantel-Haenszel-estimator if homogeneity holds.

As mentioned above, Berner et al. (2007) conducted a further meta-analysis of 17 studies using an identical cut-off of 8. Here, a relationship between sensitivity and specificity might not be expected. Nevertheless, as Figure 6 shows, a negative correlation between both kinds of validity measures occurs as well. Other characteristics than varying thresholds, e.g., study settings, may have been responsible for this heterogeneity. Again, the sum of sensitivity and specificity is quite constant as in the case of the meta-analysis reported above. Thus, the measure $\hat{\lambda}_{MH}$ seems to be an appropriate coefficient for integrating validities for these studies as well.

Here, $\hat{\lambda}_{MH} = 0.416$ with estimated variance of $\hat{Var}(\hat{\lambda}_{MH}) = 0.00011$ and associated confidence interval limits range from 0.396–0.437. Since suboptimal cut-off scores have been used the estimate sum of error rates is higher. This result enforces the argument to integrate studies with different cut-offs using the newly developed measure.

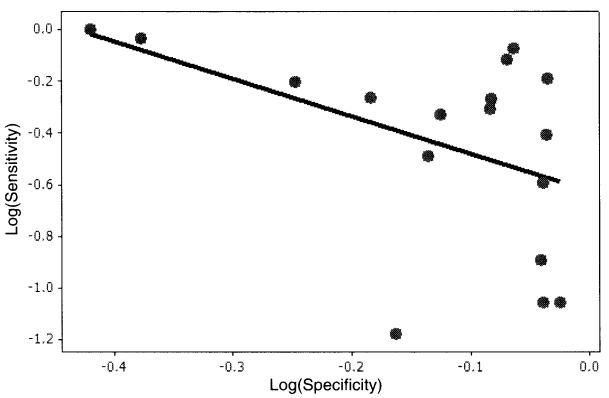

Figure 6. Log-sensitivity versus log-specificity for the AUDIT-meta-analysis using an identical cut-off of Berner, Kriston, Bentele, and Härter (2007). The line corresponds to the unweighted regression of log-sensitivity on log-specificity.

Discussion

Recently, two alternative approaches, the bivariate random-effects meta-analysis (van Houwelingen, Zwinderman, & Stijnen, 1993; Reitsma et al., 2005) and the summary receiver operating characteristic (ROC) model (Rutter & Gatsonis, 2001) have been reviewed by Harbord, Deeks, Egger, Whiting, and Sterne (2007). The suggestion had already been made to use techniques based upon the ROC (Midgette, Stukel, & Littenberg, 1993) to cope with the issue of different cut-off values. If c ranges from $-\infty$ to $+\infty$ (see Figure 1), a continuum of $(1-\alpha, \beta)$-pairs arises and, if plotted in a $(1-\alpha) \times \beta$-diagram, the ROC-curve occurs. The ROC-analysis has the advantage that the cut-off value problem is incorporated. Two ROC-curves can be directly compared, and if one lies uniformly above the other, the discriminatory power can be uniformly evaluated. The use of the ROC-analysis, therefore, has been recommended repeatedly (Irwig, Macaskill, Glasziou, & Fahey, 1995; Sutton, Abrams, Jones, Sheldon, & Song, 2000). However, disadvantages of this technique are that pairs arising from different studies might not lie on any smooth curve (as the theory might suggest). Then, smoothing and fitting techniques of parametric (McCullagh, 1980; Tosteson & Begg, 1988) or semi- and nonparametric kinds (Zou, Hall, & Shapiro, 1997) might need to be used (Greiner, 2003; Pepe, 2000). Furthermore, 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, a desire from the practical side is often to have a summary measure available for each diagnostic test to be compared. Thus, measures like the Area Under the Curve (AUC) have been suggested and discussed (Greiner, 2003).

A different measure for the discriminatory power is frequently suggested: the sum of log-odds of sensitivity and log-odds of specificity. This suggestion goes back to Hasselblad and Hedges (1995) and has found its entry in the literature including guidelines such as Irwig et al. (1994) or Deville et al. (2002) under the term *diagnostic odds ratio*. This approach is used rather frequently in meta-analysis (e.g., Cruciani et al., 2002). It was also used in the meta-analysis of Berner et al. (2007) and plays a core part in estimating the *summary receiver operating characteristic* (SROC; Dukic & Gatsonis, 2003; Sutton et al., 2000). 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 (1966), Hasselblad and Hedges (1995), and others that this measure is remarkably constant in the vicinity of the optimal cut-off

value. "... the sum of the logits is almost, but not quite, invariant under the choices of the cutpoint c when the test scores are normally distributed" (Hasselblad and Hedges 1995, p. 169) They write further: "If the two distributions are logistic with equal variances, the sum of the logarithm of the odds of sensitivity and the logarithm of the odds of specificity is independent of the cutpoint c." Hence, the question arises why seek for an alternative to the DOR? For one, the DOR is an odds ratio and, thus, has similar interpretational difficulties as the odds ratio itself, whereas the Youden index is based upon risk and is easy to interpret. Second, if there is either almost perfect diagnostic performance for sensitivity or specificity, the DOR will be dominated by the perfect measure almost independently of the value of the other. Hence, the DOR seems adequate only if both diagnostic measures, sensitivity and specificity, are imperfect. This is different for the Youden index. Third, limited empirical experience supports that the Youden index is less sensitive to the choice of the cut-off value, in particular, when the variances of the two populations are different.

With the expression given in (2), we introduced an estimator of the form

$$\frac{\sum_{i=1}^{k} w_i \left(\frac{x_i^D}{n_i^D} + \frac{x_i^H}{n_i^H} \right)}{\sum_{i=1}^{k} w_i} \qquad (5)$$

and suggested the Mantel-Haenszel weights $(n_i^D n_i^H)/n_i^+$. Alternatives should be mentioned such as the weights $(n_i^D n_i^H)$. The Mantel-Haenszel weights and the latter weights have the tendency to give larger weight to studies with a large sample size. However, the weights $(n_i^D n_i^H)$ give relative larger weight to studies with a large sample size. They also fail to provide the most efficient estimator in special settings such as the following. It was mentioned that expression (2) is unbiased and less affected by zeros. In the balanced case, for example $n_i^D = n_i^H$ for all i, (1) is also the estimator with smallest variance (under all estimators of equation (5)) if sensitivity and specificity are identical between studies, in other words, if $\alpha_i = \alpha$ and $\beta_i = \beta$ for all i. In this case, (1) can be written as

$$\hat{\lambda}_{MH} = \frac{\sum_{i=1}^{k} x_i^D}{\sum_{i=1}^{k} n_i} + \frac{\sum_{i=1}^{k} x_i^H}{\sum_{i=1}^{k} n_i} = \frac{\sum_{i=1}^{k} (x_i^D + x_i^H)}{\sum_{i=1}^{k} n_i}.$$

The measure *sum of the error rates* was chosen since it should be 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. Residual heterogeneity might be explained by the inclusion of covariates, which might be modeled by means of a generalized linear model. We have mentioned that frequently the ROC (or SROC in meta-analysis) is used. To achieve a summary measure for a given ROC, the AUC is used. It is mentioned by Greiner (2003, p. 112), 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. Indeed, this point becomes quite clear if we consider Figure 7.

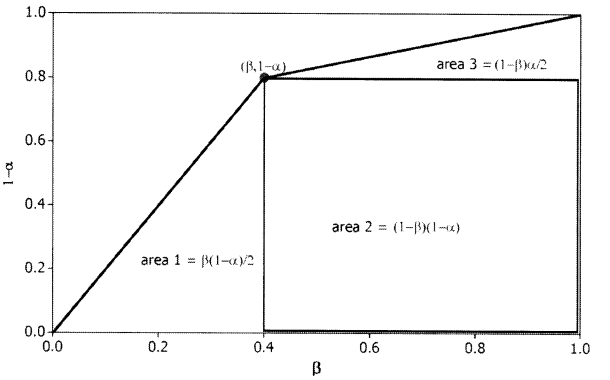

Figure 7. Equivalence of the Youden index and the area under the ROC curve.

The ROC curve is defined by connecting the points (0,0) with $(\beta,(1-\alpha))$ and (1,1). The area under the curve consists of two triangles and a rectangle, each having area $\beta(1-\alpha)/2$, $(1-\beta)\alpha$ and $(1-\beta)(1-\alpha)$, respectively. The sum of these areas is

$$\beta(1-\alpha)/2+(1-\beta)\alpha/2+(1-\beta)(1-\alpha) = \frac{(1-\beta)+(1-\alpha)}{2},$$

the average of sensitivity and specificity. As a result, there is a closer connection between the proposed measure and the SROC-approach than might be expected at first.

References

References marked with an asterisk (*) indicate studies included in the meta-analysis (see Table 2).

Babor, T.F., de la Fuente, J.R., Saunders, J., & Grant, M. (1989). International perspectives. From clinical research to secondary prevention: International collaboration in the development of the Alcohol Use Disorders Identification Test (AUDIT). *Alcohol Health and Research World, 13*, 371–374.

Berner, M.M., Kriston, L., Bentele, M., & Härter, M. (2007). The Alcohol Use Disorders Identification Test for detecting at-risk drinking: A systematic review and meta-analysis. *Journal of Studies on Alcohol and Drugs, 68*, 1–13.

*Bradley, K.A., Bush, K.R., McDonell, M.B., Malone, T., & Fihn, S.D. (1998). For the Ambulatory Care Quality Improvement Project (ACQUIP). Screening for problem drinking: Compar-

ison of CAGE and AUDIT. *Journal of General Internal Medicine, 13,* 379–388.

*Contel, M., Gual, A., & Colom, J. (1999). Alcohol Use Disorders Identification Test (AUDIT): Translation and validation to Catalan and Spanish (Spanish). *Adicciones, 11*, 337–347.

Cooper, H., & Hedges, L. (Eds.). (1994). *The handbook of research synthesis*. New York: Russell Sage Foundation.

Cruciani, M., Marcati, P., Malena, M., Bosco, O., Serpelloni, G., & Mengoli, C. (2002). Meta-analysis of diagnostic procedures for *pneumocystis carninii* pneumonia in HIV-1-infected patients. *European Respiratory Journal, 20*, 982–989.

*Daeppen, J.-B., Yersin, B., Landry, U., Pecoud, A., & Decrey, H. (2000). Reliability and validity of the Alcohol Use Disorders Identification Test (AUDIT) imbedded within a general health risk screening questionnaire: Results of a survey in 332 primary care patients. *Alcoholism: Clinical and Experimental Research, 24*, 659–665.

DerSimonian, R., & Laird, N. (1986). Meta-analysis in clinical trials. *Controlled Clinical Trials, 7*, 177–188.

Deville, W.L., Buntinx, F., Bouter, L.M., Montori, V.M., de Vet, H.C.W., van der Windt, D.A.W.M. et al. (2002). Conducting systematic reviews of diagnostic studies: Didactic guidelines. *BMC Medical Research Methodology, 2*, 9.

Dukic, V., & Gatsonis, C. (2003). Meta-analysis of diagnostic test accuracy assessment studies with varying number of thresholds. *Biometrics, 59*, 936–946.

Edwards, J.H. (1966). Some taxonomic implications of a curious feature of the bivariate normal surface. *British Journal of Prevention and Social Medicine, 20*, 42.

*Gómez, A., Conde, M., Aguiar, J.A., Santana, J.M., Jorrin, A., & Betancor, P. (2001). Diagnostic usefulness of Alcohol Use Disorders Identification Test (AUDIT) for detecting hazardous alcohol consumption in primary care settings (Spanish). *Medicina Clínica, 116*, 121–124.

*Gordon, A.J., Maisto, S.A., McNeil, M., Kraemer, K.L., Conigliaro, R.L., Kelley, M.E. et al. (2001). Three questions can detect hazardous drinkers. *Journal of Family Practice, 50*, 313–320.

Greiner, M. (2003). *Serodiagnostische Tests* [Serodiagnostic tests]. Berlin: Springer-Verlag.

*Gual, A., Segura, L., Contel, M., Heather, N., & Colom, J. (2002). AUDIT-3 and AUDIT-4: Effectiveness of two short forms of the Alcohol Use Disorders Identification Test. *Alcohol and Alcoholism, 37*, 591–596.

Haddock, K.C., Rindskopf, D., & Shadish, W.R. Jr. (1998). Using odds ratios as effect sizes for meta-analysis of dichotomous data: A primer on methods and issues. *Psychological Methods, 3*, 339–353.

Harbord, R.M., Deeks, J.J., Egger, M., Whiting, P., & Sterne, J.A.C. (2007). A unification of models for meta-analysis of diagnostic accuracy studies. *Biostatistics, 8*, 239–251.

Hasselblad, V., & Hedges, L.V. (1995). Meta-analysis of screening and diagnostic tests. *Psychological Bulletin, 117*, 167–178.

Higgins, J.P., Thompson, S.G., Deeks, J.J., & Altmann, D.G. (2003). Measuring inconsistency in meta-analyses. *British Medical Journal, 327*, 557–560.

*Hiro, H., & Shima, S. (1996). Availability of the Alcohol Use Disorders Identification Test (AUDIT) for a complete health examination in Japan. *Japanese Journal of Alcoholism. Studies and Drug Dependence, 31*, 437–450.

Hunter, J.E., & Schmidt, F.L. (2004). *Methods of meta-analysis: Correcting error and bias in research findings* (2nd ed.). Thousand Oaks, CA: Sage.

Irwig, L., Tosteson, A.N., Gatsonis, C., Lau, J., Colditz, G., Chalmers, T.C. et al. (1994). Guidelines for meta-analyses evaluating diagnostic tests. *Annals of Internal Medicine, 120*, 667–676.

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

*Kokotailo, P.K., Egan, J., Gangnon, R., Brown, D., Mundt, M., & Fleming, M. (2004). Validity of the Alcohol Use Disorders Identification Test in college students. *Alcoholism: Clinical and Experimental Research, 28*, 914–920.

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

*MacKenzie, D., Langa, A., & Brown, T.M. (1996). Identifying hazardous or harmful alcohol use in medical admissions: A comparison of AUDIT, CAGE and brief MAST. *Alcohol and Alcoholism, 31*, 591–599.

McCullagh, P. (1980). Regression models for ordinal data. *Journal of the Royal Statistical Society, Series B, 42*, 109–142.

Midgette, A.S., Stukel, T.A., & Littenberg, B. (1993). A meta-analytic method for summarizing diagnostic test performances: Receiver-operating-characteristic-summary point estimates. *Medical Decision Making, 13*, 253–257.

*Neumann, T., Neuner, B., Gentilello, L.M., Weiss-Gerlach, E., Mentz, H., & Rettig, J.S. (2004). Gender differences in the performance of a computerized version of the Alcohol Use Disorders Identification Test in subcritically injured patients who are admitted to the emergency department. *Alcoholism: Clinical and Experimental Research, 28*, 1693–1701.

Pepe, M.S. (2000). Receiver operating characteristic methodology. *Journal of the American Statistical Association, 95*, 308–311.

*Philpot, M. Pearson, N., Petratou, V., Dayanandan, R., Silverman, M., & Marshall, J. (2003). Screening for problem drinking in older people referred to a mental health service: A comparison of CAGE and AUDIT. *Aging and Mental Health, 7*, 171–175.

*Piccinelli, M., Tessari, E., Bortolomasi, M., Piasere, O., Semenzi, M., Garzotto, N. et al. (1997). Efficacy of the Alcohol Use Disorders Identification Test as a screening tool for hazardous alcohol intake and related disorders in primary care: A validity study. *British Medical Journal, 314,* 420–424.

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

*Rumpf, H.J., Hapke, U., Meyer, C., & John, U. (2002). Screening for alcohol use disorders and at-risk drinking in the general population: Psychometric performance of three questionnaires. *Alcohol and Alcoholism, 37*, 261–268.

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

Sánchez-Meca, J., Marín-Martínez, F., & Chacón-Moscoso, S. (2003). Effect-size indices for dichotomized outcomes in meta-analysis. *Psychological Methods, 8*, 448–467.

Schulze, R., Holling, H., & Böhning, D. (Eds.). (2003). *Meta-analysis. New developments and applications in medical and social sciences*. Seattle, WA: Hogrefe & Huber.

Shields, A.L., & Caruso, J.C. (2003). Reliability generalization of the Alcohol Use Disorders Identification Test. *Educational and Psychological Measurement, 63*, 404–413.

*Skipsey, K., Burleson, J.A., & Kranzler, H.R. (1997). Utility of the AUDIT for identification of hazardous or harmful drinking in drug-dependent patients. *Drug and Alcohol Dependence, 45*, 157–163.

Sutton, A.J., Abrams, K.R., Jones, D.R., Sheldon, T.A., & Song, F. (2000). *Methods for meta-analysis in medical research*. New York: Wiley.

*Taj, N., Devera-Sales, A., & Vinson, D.C. (1998). Screening for problem drinking: Does a single question work? *Journal of Family Practice, 46*, 328–335.

Tosteson, A., & Begg, C. (1988). A general regression methodology for ROC curve estimation. *Medical Decision Making, 8*, 204–215.

Van Houwelingen, H.C., Zwinderman, K.H., & Stijnen, T. (1993). A bivariate approach to meta-analysis. *Statistics in Medicine, 12*, 2273–2284.

Youden, D. (1950). Index for rating diagnostic tests. *Cancer, 3*, 32–35.

Zou, K., Hall, W., & Shapiro, D. (1997). Smooth nonparametric ROC curves for continuous diagnostic tests. *Statistics in Medicine, 16*, 2143–2156.

Prof. Dr. Heinz Holling

Lehrstuhl für Statistik und Methoden
Fachbereich Psychologie und Sportwissenschaft
Westfälische Wilhelms-Universität Münster
Fliednerstr. 21
D-48149 Münster
Germany
Tel. +49 251 83-39419
Fax +49 251 83-39469
E-mail holling@psy.uni-muenster.de