Statistical Methods in Medical Research 2011; **20**: 541–550

A limitation of the diagnostic-odds ratio in determining an optimal cut-off value for a continuous diagnostic test

Dankmar Böhning Applied Statistics, School of Biological Sciences, University of Reading, UK, **Heinz Holling** Statistics and Quantitative Methods, Faculty of Psychology and Sport Science, University of Münster, Germany and **Valentin Patilea** Centre de Mathématiques–IRMAR, Institut National des Sciences Appliquées (INSA) de Rennes, France

The article considers the diagnostic odds ratio, a special summarising function of specificity and sensitivity for a given diagnostic test, which has been suggested as a measure of diagnostic discriminatory power. In the situation of a continuous diagnostic test a cut-off value has to be chosen and it is a common practice to choose the cut-off value on the basis of the maximised diagnostic odds ratio. We show that this strategy is not to be recommended since it might easily lead to cut-off values on the boundary of the parameter range. This is illustrated by means of some examples. The source of the deficient behaviour of the diagnostic odds ratio lies in the convexity of the log-diagnostic odds ratio as a function of the cut-off value. This can easily be seen in practice by plotting a non-parametric estimate of the log-DOR against the cut-off value. In fact, it is shown for the case of a normal distributed diseased and a normal distributed non-diseased population with equal variances that the log-DOR is a convex function of the cut-off value. It is also shown that these problems are not present for the Youden index, which appears to be a better choice.

1 Introduction and background

We are interested in the *diagnostic test accuracy* of a diagnostic test B for diagnosing the presence of a specific condition. A typical setting is described as follows. The outcome of B is binary where $B = 1$ indicates the presence of the condition (test is positive) and $B = 0$ indicates the absence of the condition. Here 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 accuracy are considered: the *sensitivity* defined as $S^+ = Pr(\text{test positive}|\text{diseased}) = (1 - \beta)$ and the *specificity* defined as $S^- = Pr(\text{test negative}|\text{non-diseased}) = (1 - \alpha)$. The sensitivity measures the capability of the diagnostic test to recognise a diseased person correctly, whereas the specificity measures the capability of diagnosing a healthy person correctly. Consequently, β is the error probability of falsely classifying a diseased person as healthy and α is the error probability of falsely classifying a healthy person as diseased. Ideally, α and β should

Address for correspondence: Dankmar Böhning, Applied Statistics, School of Biological Sciences, University of Reading, UK.
E-mail: d.a.w.bohning@reading.ac.uk

10.1177/0962280210374532

be small if not zero at all. In some situations, sensitivity has precedence over specificity, for instance to assess the clinical utility of prognostic biomarkers in cancer. Then, the focus will be on constructing a diagnostic test with a desired sensitivity. In the absence of a clear preference, however, one turns to summary measures. A natural summary measure of sensitivity and specificity is Youden's index J[1] defined as

$$J = S^+ + S^- - 1 = 1 - (\alpha + \beta). \tag{1}$$

For a more general discussion see Pepe[2] or Greiner.[3] In addition, Le[4] summarises a number of positive properties of the Youden index. We mention also the possibility of weighted Youden index $J = pS^+ + (1-p)S^-$ to incorporate potential different roles of sensitivity and specificity by means of a weight p.

As an alternative the *diagnostic odds ratio* (DOR) has been suggested and utilised frequently in the literature. The diagnostic odds ratio as a single indicator of diagnostic performance, as proposed and recommended for example by Glas *et al.*[5] is defined as

$$D = \frac{S^+}{1-S^+} \times \frac{S^-}{1-S^-}. \tag{2}$$

Note that (2) can be written as the ratio of the odds $\frac{S^+}{1-S^+}$ for diagnosing a diseased person as diseased to the odds $\frac{1-S^-}{S^-}$ for diagnosing a healthy person as diseased.

Now, we suppose that the diagnostic procedure is providing a continuous outcome or an ordered categorical outcome, which we denote as T. For example, a psychological test is used (potentially among other procedures) to determine a certain condition such as the presence of dementia in an elderly person. Often these diagnostic tests deliver a score and a cut-off value c is used to decide about the presence or absence of the condition. Note that T and the binary test result variable B are connected via $B = \mathbb{I}_{\{T>c\}}$, where $\mathbb{I}_S$ denotes the indicator function for a set S defined as $\mathbb{I}_S(s) = 1$ if $s \in S$ and 0 otherwise. This situation is illustrated in Figure 1(a) for a continuous outcome T which is normally distributed in the two populations.

In the healthy population we have a normal distribution with mean $\mu_H = 0$ and variance $\sigma_H^2 = 1$, in the diseased population we have a normal distribution with mean $\mu_D = 2$ and variance $\sigma_D^2 = 4$. A cut-off value c determines sensitivity as $S^+(c) = 1 - \Phi\left(\frac{c-\mu_D}{\sigma_D}\right)$ and specificity as $S^-(c) = \Phi\left(\frac{c-\mu_H}{\sigma_H}\right)$, 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, if c is shifted to the right, the sensitivity decreases whereas the specificity increases, and *vice versa* if c is shifted to the left. Consequently, the discriminatory power of the diagnostic test becomes a function of the cut-off value and the question of the choice of the cut-off value c for the diagnostic test arises.[2,3] Ideally, as Leeflang *et al.*[6] point out, the optimal choice for this cut-off is ultimately determined by the consequences associated with false-positive and false-negative test results (see also the comments by Greiner *et al.*[7]). For example, if preference is on sensitivity, a desired value for sensitivity will determine the cut-off uniquely and no further consideration is necessary. But Leeflang *et al.*[6] also continue saying that in the

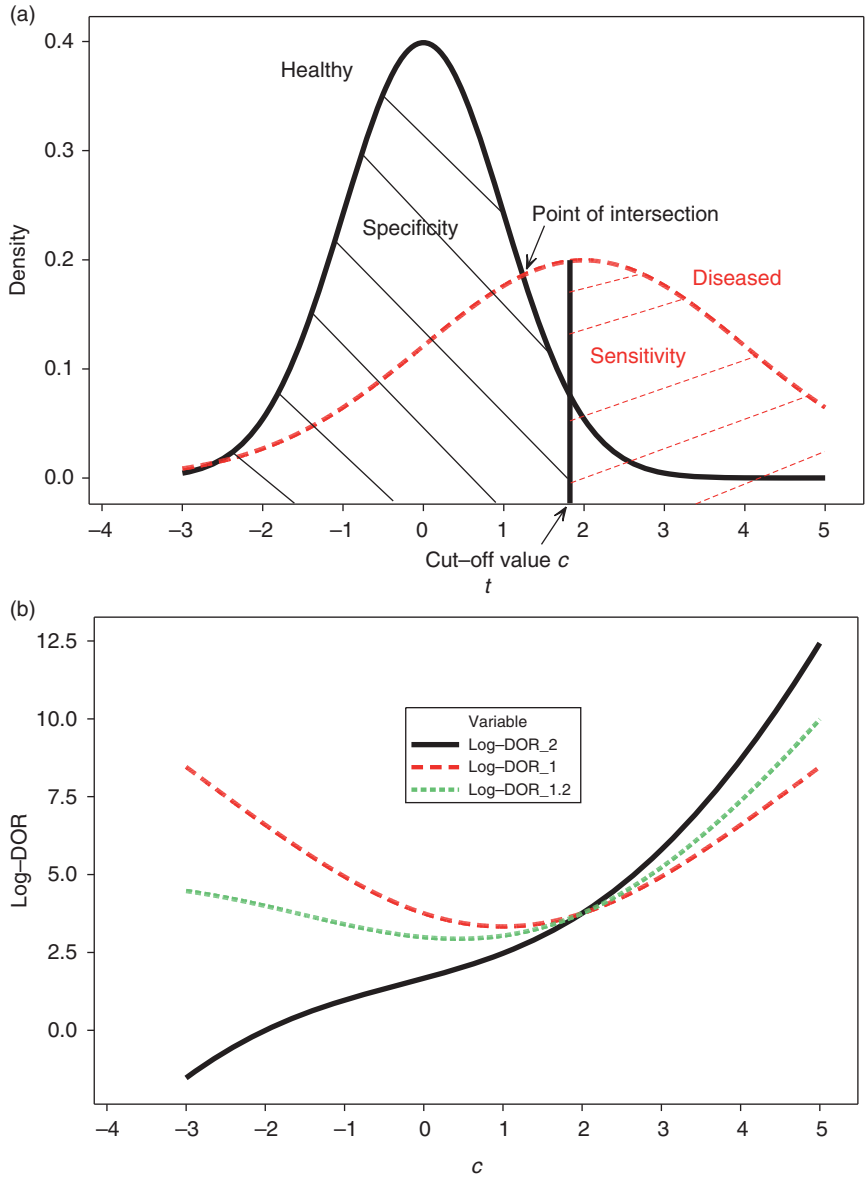

Figure 1 Diagnostic situation illustrated with two normal distributions: (a) one has mean 0 and variance 1 (healthy population), the other has mean 2 and variance 4 (diseased population) and $\log D(c)$ as a function of the cut-off value, c, for three scenarios of two normal distributions; (b) one has always mean 0 and variance 1 (healthy population), the other has mean 2 and varies in the variance (diseased population): $\sigma^2 = 1, 1.2^2$ and 2^2.

Table 1 Performance of various PHQ-9 cut-off scores in detecting major depression (following Lotrakul *et al*.[9])

Cut-off	Sensitivity	Specificity	Youden index
6	0.95	0.48	0.43
7	0.95	0.55	0.50
8	0.89	0.65	0.54
9	0.84	0.77	0.61
10	0.74	0.85	0.59
11	0.68	0.89	0.57
12	0.68	0.90	0.58
13	0.63	0.94	0.57
14	0.47	0.96	0.43
15	0.37	0.97	0.34

early phases of test development preference for choosing the optimal cutoff is often a criterion that weighs both sensitivity and specificity equally. Similarly, Magder and Fix[8] argue that when the test is used for research purposes and treatment decisions are not based on the test results, the choice of the cut point should be based on scientific (statistical) considerations.

Indeed, it has become common practice to choose the cut-off value to maximise some function of sensitivity and specificity, often the Youden index. To illustrate we choose one example out of many. Lotrakul *et al.*[9] consider choosing an optimal cut-off for the Thai version of the Patient Health Questionnaire (PHQ-9), which has been developed as a screening tool for major depression in primary care patients. Sensitivity and specificity were estimated in a diagnostic study involving 279 patients for different cut-off values using the Mini International Neuropsychiatric Interview and and the Hamilton Rating Scale for Depressions as gold standards. Lotrakul *et al.*[9] consider different cut-off values and determine associated sensitivities and specificities. The values are reproduced here as Table 1. The authors recommend as choice of the cut-off value the score-value 9, which evidently is the largest value for the Youden index. See also Table 1 for details. It is also common practice to report a number of measures associated with a diagnostic test. In fact, Fisher *et al.*[10] develop a reader's guide to the interpretation of diagnostic test properties. They recommend for enhancing the critical appraisal on articles on diagnostic tests to report several measures of test accuracy including besides sensitivity, specificity, ROC, likelihood ratios also the diagnostic odds ratio. In fact, in applied papers it is now common practice to report the DOR besides other values. However, it is often less clear which measure has been used to determine the cut-off value. In fact, the original paper by Glas *et al.*[5] suggested the DOR as a single indicator of test performance to facilitate the formal meta-analysis of studies on diagnostic test performance. Magder and Fix[8] suggest to use the *precision* of the DOR as a criterion for choosing the cut-off value. The DOR was clearly suggested as a measure of discriminatory power and, consequently, it is not surprising that practitioners understand it in this way. The DOR is increasingly used in the applied field as criterion for choosing the cut-off value optimally. An example is Wei *et al.*[11] who use the DOR for a diagnostic test on bone metastasis in prostate cancer. There is also software available to analyse data from diagnostic test accuracy,

which allows to choose the DOR as a cut-off value optimising criterion.[12] Hence, given these practices we feel that it is appropriate to take a deeper look into the behaviour of the DOR as a function of the cut-off value. If the DOR in the data set on depression in Thai primary care patients is plotted against the PHQ-9 cut-off score a U-shape pattern arises. This shape of the function occurs in many data sets on diagnostic testing and is more the rule than the exception. In Figure 1(b) the graph of the log-DOR as function of the cut-off is plotted for various situations of normal distributions. If the variances are identical or similar a U-shaped pattern arises. Consequently, optimising values for the DOR will have the tendency to occur near the boundary of the observed data leading to implausible and statistical inferior cut-off values as will be seen below. The major point of the paper is a finding that implies that this U-shape pattern, the log-convexity of the DOR, is not an accidental property but rather a systematic feature of the DOR. Hence, it out-rules the DOR as a criterion for selecting the cut-off value in a continuous test.

2 The convexity result for the DOR

We now come to the general result and consider the situation that the diagnostic test has the same variance $\sigma_D^2 = \sigma_H^2 = \sigma^2$ in the diseased and the non-diseased population. Without limitation of generality we set $\sigma^2 = 1$, $\mu_D = \mu$, $\mu_H = 0$. Hence, the following result is proved under the assumption of normality with equal variances in the two populations of healthy and diseased individuals.

Theorem 1. *Let* $\Phi(\cdot)$ *be the cumulative distribution function of the standard normal distribution. Also, let*

$$D(c) = \frac{S^+(c)}{1 - S^+(c)} \times \frac{S^-(c)}{1 - S^-(c)} = \frac{1 - \Phi(c - \mu)}{\Phi(c - \mu)} \times \frac{\Phi(c)}{1 - \Phi(c)}.$$

Then:

$$D(c) > D(\mu/2), \quad \text{for all } 0 \le c \le \mu, \text{ but } c \ne \mu/2, \tag{3}$$

$$\frac{\mathrm{d}^2}{\mathrm{d}c^2} \log D(c) > 0 \text{ for all } c \in [0, \mu]. \tag{4}$$

The theorem says that $D(\cdot)$ is actually *minimised* at $\hat{c} = \mu/2$ and that $\log D(\cdot)$ is *convex*. As a consequence, points maximising the $D(c)$ will be on the boundary of the parameter space $[0, \mu]$, leading to useless cut-off values. In conclusion, the DOR is not useful as a criterion for maximising discriminatory power.

The proof is not given here for the sake of brevity, but is provided completely at

`http://www.personal.reading.ac.uk/~sns05dab/proof.pdf`.

In it's first part, it will not use the assumption of normality and the proof is done without reference to normality. At a particular stage, however, normality is used and we will draw the attention of the reader to this point.

Finally, we would like to mention that these problems do not exist for the Youden index. The following results are fairly known[13,14] but we summarise here for completeness and correctness what can be said in the normal case with *different* variances for the healthy and diseased population. We do so since the results are frequently loosely stated, in particular, the regularity condition (7) is often ignored. It is easy to see that in the normal case we have for Youden's index that

$$J(c) = S^+(c) + S^-(c) - 1 = 1 - \Phi\left(\frac{c-\mu_D}{\sigma_D}\right) + \Phi\left(\frac{c-\mu_H}{\sigma_H}\right) - 1$$

$$= -\Phi\left(\frac{c-\mu_D}{\sigma_D}\right) + \Phi\left(\frac{c-\mu_H}{\sigma_H}\right).$$

Taking derivatives with respect to c we obtain

$$J'(c) = -\frac{1}{\sigma_D}\phi\left(\frac{c-\mu_D}{\sigma_D}\right) + \frac{1}{\sigma_H}\phi\left(\frac{c-\mu_H}{\sigma_H}\right),$$

where $\phi(\cdot) = \Phi'(\cdot)$. Setting the derivative $J'(c)$ to zero provides the result that the optimal cut-off value $\hat{c}$, which maximises the Youden index, is found as the point of intersection of the two normal curves, lying between the means μ_H and μ_D:

$$\frac{1}{\sigma_D}\phi\left(\frac{\hat{c}-\mu_D}{\sigma_D}\right) = \frac{1}{\sigma_H}\phi\left(\frac{\hat{c}-\mu_H}{\sigma_H}\right). \tag{5}$$

Hence, the cut-off value needs to be moved to the point of intersection, depicted in Figure 1, between the two means. Note that there is a second point of intersection when the two normals have different variances which can be excluded. Analytically, $\hat{c}$ is found to be the solution of the quadratic equation in c

$$\left(\frac{1}{\sigma_H^2} - \frac{1}{\sigma_D^2}\right)c^2 + 2\left(\frac{\mu_D}{\sigma_D^2} - \frac{\mu_H}{\sigma_H^2}\right)c + \left(\frac{\mu_H^2}{\sigma_H^2} - \frac{\mu_D^2}{\sigma_D^2}\right) + 2\log\left(\frac{\sigma_H}{\sigma_D}\right) = 0 \tag{6}$$

which lies between μ_H and μ_D.

Furthermore, we find for the second derivative that

$$J''(c) = \frac{1}{\sigma_D}\frac{c-\mu_D}{\sigma_D^2}\phi\left(\frac{c-\mu_D}{\sigma_D}\right) - \frac{1}{\sigma_H}\frac{c-\mu_H}{\sigma_H^2}\phi\left(\frac{c-\mu_H}{\sigma_H}\right) < 0$$

for $c \in (\mu_H, \mu_D)$, since $(c-\mu_D) < 0$ and $(c-\mu_H) > 0$ for $c \in (\mu_H, \mu_D)$. Hence, $J(c)$ is a strictly concave function on (μ_H, μ_D). We summarise this result without further proof (which is available from the authors) in the following theorem.

Theorem 2. *Let* $\delta = \mu_D - \mu_H$ *and* $\rho = \sigma_D/\sigma_H$. *If* $T \sim N(\mu_D, \sigma_D^2)$ *in the diseased population and* $T \sim N(\mu_H, \sigma_H^2)$ *in the non-diseased population, then:*

1. $J(c)$ *is a strictly concave on* (μ_H, μ_D);
2. *The cut-off value* $\hat{c}$ *maximising the Youden index* $J(c)$ *is found as the point of intersection of the two normal curves (a solution of (6)) which lies between* μ_H *and* μ_D *if*

$$\delta^2 > \max\{-2\sigma_D^2 \log \rho, 2\sigma_H^2 \log \rho\}\ ; \tag{7}$$

3. *If both normals have the same variance, then* $\hat{c} = (\mu_D + \mu_H)/2$.

Note the importance of the condition (7). If condition (7) fails to hold the maximum is one of the end points of the interval $[\mu_H, \mu_D]$. The regularity condition will hold in most situations of practical interest and it will fail to hold only if the ratio ρ differs largely from unity. Note that a more general discussion of estimation of the Youden index including the consideration of non-parametric techniques has been provided by Fluss *et al.*[15]

Example: Screening for hearing impairment in newborn babies. We would like to illustrate the results from Theorem 1 and 2 with an example. Pepe[2] (p. 72) discusses data on screening for hearing impairment in newborn babies going back to Norton *et al.*[16] The complete data set contains 5058 observations and three passive hearing tests as well as the gold standard and other covariate information. As diagnostic test we have considered the values from the DPOAE 65 at 2 kHz. Ignoring the clustered structure of the data (right and left ear of baby) we have constructed nonparametric estimates of the functions $J(c)$ and $\log D(c)$ and the graphs are provided in Figure 2(a) and (b). The Youden index shows reasonably concave pattern over a wide range of the central sample space. Indeed, if we compute estimates for the two populations we find $\hat{\mu}_H = -8.917$, $\hat{\sigma}_H = 7.781$, $\hat{\mu}_D = -4.872$, $\hat{\sigma}_D = 8.581$ with condition (7) being fulfilled (with population parameters replaced by sample estimates). The estimate of ρ is $\hat{\rho} = 1.10281$ which is close to unity. This will lead to a solution for c between -8.917 and -4.872, not far from the midpoint between these two values. For the diagnostic odds ratio the situation is completely different, as shown in Figure 2(b). The $\log \hat{D}(c)$ appears to be convex (as it is suggested by Theorem 2 for the equal variance case) for a wide sample range. Maximising the diagnostic odds ratio will lead to values of c near the boundary of the sample space.

3 Discussion

We have shown a convex property for the $\log D(c)$ as a function of c in the case of normal distributions with equal variances. Now, as illustrated in the example of Section 2 this behaviour occurs in many situations and we have included the result to point out a structural failure in discriminatory power of the DOR. A reviewer remarked that our result would be more powerful if it would include the unequal variance case. Now, we know the convexity of the log-DOR does not hold for unequal variances

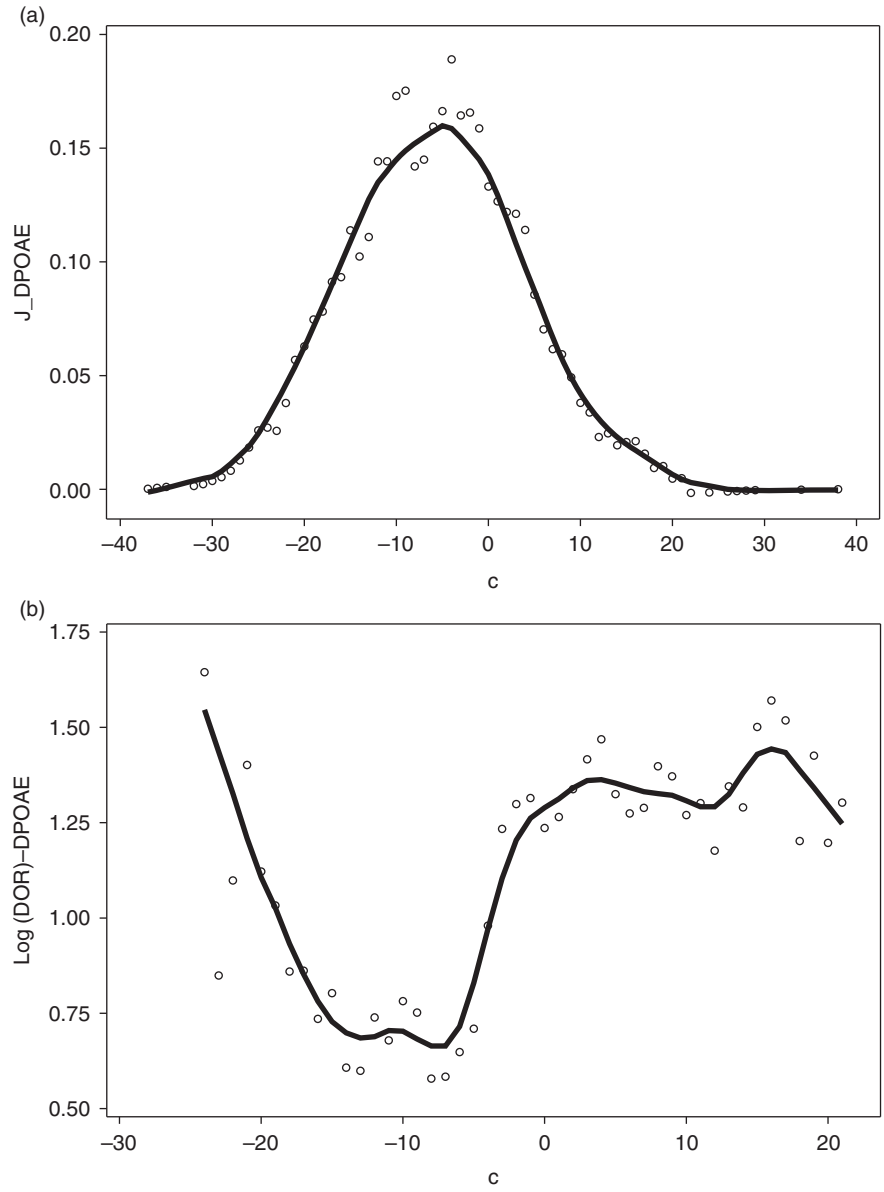

Figure 2 $\hat{J}(c)$ (dots) (a) and $\log \hat{D}(c)$ (dots) (b) with LOWESS smoother (solid line) for the data on screening for hearing impairment of newborn babies provided in Pepe[2]; screening test is the DPOAE.

as Figure 1(b) shows. Other patterns than convexity arise, which are neither suitable for using the DOR as an optimising criterion. So, it is not worthwhile looking for a generalisation of the convexity property — it does not generalise. But it is just this strange behaviour of the DOR that outrules it's usefulness as a selection criterion for a cut-off. In practice it is easy to construct a plot of a non-parametric estimate of the log $D(c)$ against

the cut-off value c as done in Section 2 to support the failure of the DOR in choosing an optimal cut-off value. In a similar flavour, Pepe *et al.*[17] point out the limitations of the odds ratio for a diagnostic, prognostic, or screening marker. In summary, the authors state that odds ratios do not characterise the discriminatory capacity of a marker.

It was pointed out by Hasselblad and Hedges[18] that the diagnostic odds ratio experiences very little change over a wide range of the cut-off value, a fact already noted by Edwards.[19] In fact, they point out that if both distributions are logistic with equal variances, the DOR is invariant with respect to the cut-off value. This is a beneficial property if it is desired to combine results from different studies (as it is in meta-analysis). Deeks[20] mentions in the context of systematic reviews of diagnostic and screening tests that the diagnostic odds ratio often is reasonably constant regardless of the diagnostic threshold. However, in the context of finding a measure, which maximises discriminatory power (in separating diseased and non-diseased populations) it is less advantageous.

Frequently, the consequences of false-positive and false-negative decisions are needed to be incorporated in the choice of the threshold value.[21,22] For example, the consequences of a false-negative decision might be more severe than a false-positive leading to attaching more weight to sensitivity than to specificity. This could be achieved and incorporated into an index by generalising the Youden index to $J(c) = wS^+(c) + (1 - w)S^-(c)$ where w is a number between 0 and 1. Here, clearly the difficulty lies in the specification of the number w, similar to the specification of a trade-off function balancing harm and benefit for the patient.

Often it is desirable to include the disease prevalence into the determination of the cut-off value, in particular if a population screening is the application of the diagnostic test. Again, the Youden index can be generalised to take the disease prevalence p into account:[22] $J(c) = pS^+(c) + (1 - p)S^-(c)$. The prevalence weight p will not change the concavity property of $J(.)$. Hence, the positive behaviour of the Youden index is retained. Of course, the optimal cut-off value will depend now on the prevalence value of p. In any of the above generalisations of Youden's index, the result provided in Theorem 2 will still hold.

Finally, it might be argued that instead of investigating a summary measure of sensitivity and specificity it might be more fertile to focus on curve modelling techniques such as Receiver Operating Curves as discussed in Pepe[2] or Greiner.[3] Pepe[2] points out that the receiver operating characteristic (ROC) curve is currently the best-developed statistical tool for describing the performance of diagnostic tests. ROC curves have been developed in signal detection theory. Later their potential for medical diagnostic testing was recognised.[2] The clear benefit lies in the direct modelling of the dependency of specificity and sensitivity from the cut-off value and offers a global characterisation of the behaviour of the continuous test outcome. If, however, a cut-off value needs to be determined the question of optimising an optimality criterion (such as the Euclidean distance of the ROC curve to the upper left corner of the unit square) arises as well.

Acknowledgements

Part of this work was funded by the *German Research Foundation* (GZ: Ho1286/7-1). We are grateful to the Editor as well as two reviewers for their very helpful comments.

References

1 Youden D. Index for rating diagnostic tests. *Cancer* 1950; **3**: 32–5.
2 Pepe MS. *The statistical evaluation of medical tests for classification and prediction*. University Press, Oxford; 2003.
3 Greiner M. *Serodiagnostische tests*. Springer, Berlin; 2003.
4 Le CT. A solution for the most basic optimisation problem associated with an ROC curve. *Statistical Methods in Medical Research* 2006; **15**: 571–84.
5 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.
6 Leeflang MMG, Moons KGM, Reitsma JB, Zwinderman AH. Bias in sensitivity and specificity caused by data-driven selection of optimal cutoff values: mechanisms, magnitude, and solutions. *Clinical Chemistry* 2008; **54**: 729–37.
7 Greiner M, Pfeiffer DU, Smith, RD. Principles and practical applications of the receiver operating characteristic (ROC) for diagnostic test. *Preventive Veterinary Medicine* 2000; **45**: 23–41.
8 Magder LS, Fix AD. Optimal choice of a cut point for a quantitative diagnostic test performed for research purposes. *Journal of Clinical Epidemiology* 2003; **56**: 956–62.
9 Lotrakul M, Sumrithe S, Saipanish R. Reliability and validity of the Thai version of the PHQ-9. *BMC Psychiatry* 2008; **8**: 46.
10 Fischer JE, Bachmann LM, Jaeschke R. A readers' guide to the interpretation of diagnostic test properties: clinical example of sepsis. *Intensive Care Medicine* 2003; **29**: 1043–51.
11 Wei LH, Chiu JS, Chang SY, Wang YF. Predicting bone metastasis in prostate cancer patients: value of prostate specific antigen. *Tzu Chi Medical Journal* 2008; **20**: 291–95.
12 Brasil P. The DiagnosisMed Package. Available at: http://cran.r-project.org/web/packages/-DiagnosisMed/DiagnosisMed.pdf, (accessed 9 March 2010).
13 Schisterman EF, Perkins N. Confidence intervals for the Youden index and corresponding optimal cut-point. *Communications in Statistics – Simulation and Computation* 2007; **36**: 549–63.
14 Perkins N, Schisterman EF. The Youden index and the optimal cut-point corrected for measurement error. *Biometrical Journal* 2005; **47**: 428–41.
15 Fluss R, Faraggi D, Reiser B. Estimation of the Youden index and its associated cutoff point. *Biometrical Journal* 2005; **47**: 458–72.
16 Norton SJ, Gorga MP, Widen JE, *et al.* Identification of neonatal hearing impairment: Evaluation of transient evoked optoacoustic emission, distortion product optoacoustic emission, and auditory brain stem response test performance. *Ear and Hearing* 2000; **21**: 508–28.
17 Pepe MS, Janes H, Longton G, Leisenring W, Newcomb P. Limitations of the odds ratio in gauging the performance of a diagnostic, prognostic, or screening marker. *American Journal of Epidemiology* 2004; **159**: 882–90.
18 Hasselblad V, Hedges LV. Meta-analysis of screening and diagnostic tests. *Psychological Bulletin* 1995; **117**: 167–78.
19 Edwards JH. Some taxonomic implications of a curious feature of the bivariate normal surface. *British Journal of Prevention and Social Medicine* 1966; **20**: 42.
20 Deeks JJ. Systematic reviews of evaluations of diagnostic and screening tests. *British Medical Journal* 2001; **323**: 157–62.
21 Hunink M, Glasziou P, Siegel J, *et al.* *Decision making in health and medicine – integrating evidence and values*. University Press, Cambridge; 2001.
22 Hand DJ, Krzanowski W. *ROC curves for continuous data*. CRC Press, Boca Raton; 2009.