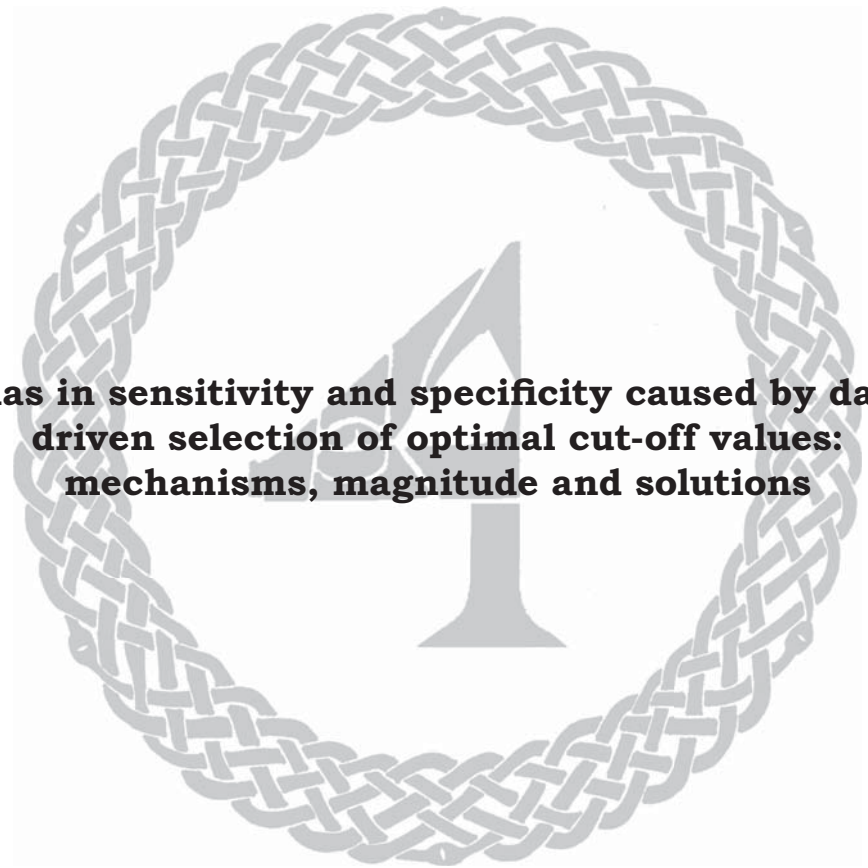
Systematic reviews of diagnostic test accuracy

Leeflang, M.M.G.

Publication date
2008



Bias in sensitivity and specificity caused by data driven selection of optimal cut-off values: mechanisms, magnitude and solutions

Mariska M.G. Leeflang, Karel G.M. Moons, Johannes B. Reitsma, Aeilko H. Zwinderman

Clin Chem. 2008; 54(4):729-37

Abstract

Background: Optimal cut-off values for continuous test results are often derived in a data-driven way. This may however lead to overoptimistic measures of diagnostic accuracy.

Aim of study: To determine the magnitude of bias in sensitivity and specificity associated with data-driven selection of cut-off values and to examine potential solutions to reduce this bias.

Methods: Simulation study using different sample sizes, distributions and prevalences. We compared data-driven estimates of accuracy based on the Youden index with the true values, and calculated the median bias. Three alternative approaches (assuming specific distribution, leave-one-out, smoothed ROC) were examined for their ability to reduce this bias.

Results: The magnitude of bias caused by data-driven optimization of cut-off values was inversely related to sample size. If the true value of sensitivity and specificity are 84%, estimates in studies with a total sample size of 40 will be around 90%. If sample size increases to 200, estimates will be 86%. The distribution of the test results had little impact on the amount of bias if sample size was held constant. More robust methods of optimizing cut-off values were less prone to bias, but the performance deteriorated if the underlying assumptions were not met.

Discussion: Data-driven selection of the optimal cut-off value can lead to overoptimistic estimates of sensitivity and specificity, especially in small studies. Alternative methods can reduce this bias, but finding robust estimates of cut-off values and accuracy requires considerable sample sizes.

4.1 Introduction

Diagnostic accuracy is the amount of agreement between the results of an index test (the test under evaluation) and the reference standard (the best available method to determine the presence or absence of the disease of interest). Commonly used accuracy measures are sensitivity (the proportion of those with the target condition who have a positive index test result) and specificity (the proportion of those without the target condition who have a negative index test result). In case of a continuous or ordinal test, the ROC curve is an informative way to present the sensitivity versus 1–specificity for each possible cut-off value of the index test[1,2]. In situations where higher test results are more indicative of the presence of disease, lowering the cut-off value will increase sensitivity, while specificity decreases. For clinical purposes in order to link actions to test results, one threshold or cut-off value is used. The optimal choice of this cut-off value is ultimately determined by the consequences associated with false positive and false negative test results[3].

In early phases of test development, when the exact role of the index test is not fully defined and thus the consequences of incorrect test results are not yet determined, a criterion that equally weighs both sensitivity and specificity is often preferred to choose the optimal cut-off value. Such a criterion is the Youden index, which is defined by sensitivity + specificity – 1 [4,5]. The optimal cut-off value that maximizes the Youden index is often determined in a "data-driven" way. This means that the sensitivities and specificities across all possible cut-off values within the range of test results are calculated from the data at hand, and the cut-off value that leads to the highest Youden index is then selected.

This data-driven selection of optimal cut-off values is prone to bias, meaning that it systematically leads to overestimation of sensitivity and specificity of the test under study. Because chance variation plays a larger role in smaller studies, it means that the observed ROC curve from a single small study will deviate more from the true underlying ROC curve than the observed ROC curve from a large study (see Figure 4.1). These fluctuations occur in both directions leading to both underestimation and overestimation in relation to the true sensitivity and specificity. In small studies, an increase in sensitivity by taking a lower threshold will not directly lead to a decrease in specificity. Because the data-driven approach specifically selects the cut-off value with the highest sum of sensitivity and specificity (i.e. closest to the top left corner of the ROC plot), it is generally a point above the true underlying ROC curve. Data-driven selection of cut-off values for continuous test results in studies with low sample size may therefore lead to overoptimistic estimates of sensitivity and specificity. Because small sample sizes (<200) are common in diagnostic studies[6], overestimation of diagnostic accuracy by data-driven selection of cut-off values can be a serious and prevalent problem.

This potential for bias associated with data-driven selection of the optimal Youden index has been recognized before, both in diagnostic and prognostic studies[7-13].

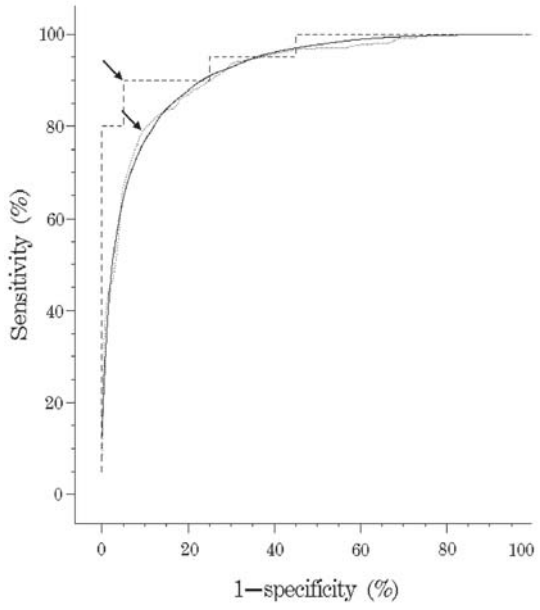


Figure 4.1. ROC curves from three single studies in which the test results have been generated from the same underlying distribution, but with varying sample size.
Disease prevalence was 50% in all three studies. The dashed line is based on a single study with a total sample size of 40 patients; the dotted line on a study with 1000 patients; the solid line is the true ROC curve belonging to a study with an infinite number of patients. The data-driven maximum Youden indices for the two empirical datasets are pointed by arrows: the upper arrow points at the optimal cut-off value in the population with 40 patients and the lower arrow points at the optimal cut-off value for sample size 1000. The true optimal sensitivity and specificity are both 84%.

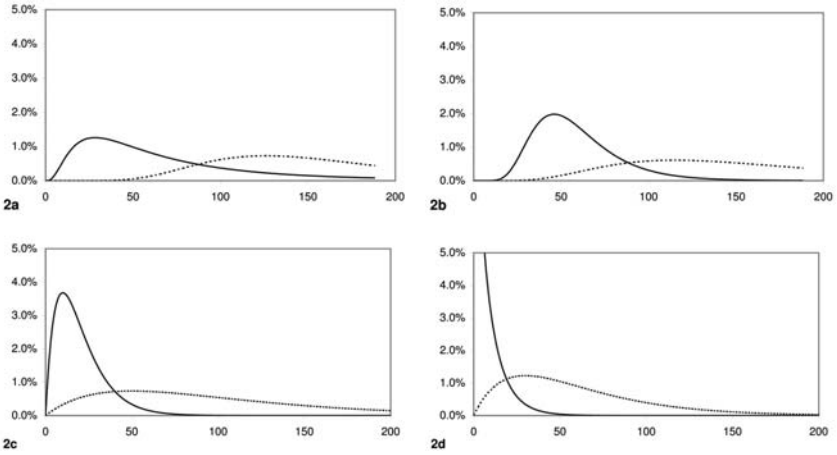


Figure 4.2. Alternative distributions.
Alternative distributions of generating non-normal test results in our simulations: two Lognormal distributions (4.2a and 4.2b) and two Gamma distributions (4.2c and 4.2d). The solid lines reflect the distribution within the patients without the disease and the dotted lines in the patients with the disease. On the X-axis the test result and on the Y-axis the percentage of patients with that test result.

These publications have been rather technical, without offering clear guidance or solutions for practice. We therefore performed a series of simulations to document the magnitude of overestimation of sensitivity and specificity under a range of conditions and examined the possible role of alternative ways of estimating the sensitivity and specificity, using the Youden index. Based on these simulations we will be able to inform readers when to be aware of this bias and advice researchers how to reduce the potential for this bias in future studies.

4.2 Methods

4.2.1 Simulation of data sets

Continuous index test results for individuals with and without the disease were simulated based on a specific distribution, sample size and disease prevalence. The true values of optimal Youden index and cut-off and the corresponding true maximum sensitivities and specificities, were calculated from the true underlying distribution of test results among individuals with the disease and those without the disease.

To examine the impact of sample size, disease prevalence, amount of spread in index test results, and their underlying distribution on the amount of bias, we varied these parameters across scenarios. Sample sizes varied from 20 to 1000 patients; prevalence from 5% to 95%; standard deviations from 5 to 20; and test results were generated from an underlying Normal distribution and two non-symmetrical distribution Lognormal and Gamma distribution (see Figure 4.2).

All analyses were carried out using SAS for Windows, version 9.1.3 (SAS Institute).

4.2.2 Data driven estimation of sensitivity and specificity

Data driven estimates of diagnostic accuracy associated with the optimal cut-off value were determined in each simulated data set and compared with their true values. Each simulation scenario was replicated 2000 times to determine the median magnitude of bias (difference between each data driven estimate of both sensitivity and specificity and their true values) and the number of times (%) sensitivity and specificity were overestimated.

4.2.3 Potential solutions to reduce overestimation

Three alternative methods were examined whether they can reduce the magnitude of bias: (a) using sample characteristics and assuming a specific underlying distribution, (b) leave-one-out cross-validation, (c) robust fitting of ROC-curves. These methods were applied to two scenarios with a true underlying distribution of the index test results that was a Normal distribution, two scenarios with a true underlying Lognormal distribution and two scenarios with a true underlying Gamma distribution (see Figure 4.2). Within each simulated data set we compared the data

driven estimate with the estimates of the potential solutions to examine the effectiveness of the solutions in reducing the bias.

Deriving optimal cut-off point from assumed distributions
Sample characteristics describing the central tendency and shape of distribution of test results can be used to estimate the optimal cut-off value. By assuming a specific underlying distribution (e.g. a Normal distribution) for the test results in the patients with the disease, these sample characteristics (descriptives like mean and SD) can be used to calculate the cumulative proportion of diseased patients who will have an index test result equal to or above that cut-off value, e.g. an estimate of true sensitivity. Similarly, using the observed mean and SD of the non-diseased patients, the proportion of patients without the disease and with an index test result below each possible cut-off value can be calculated. This equals the specificity of that test. The Gamma distribution is characterized by a shape and a scale parameter, that, just like the mean and SD, describe the variation in test results in individuals with and without the disease within a sample. The lower the shape parameter, the more skewed the distribution is. The lower the scale parameter, the less spread the results are (just like a smaller standard deviation in Normal distributions). We estimated the shape and scale parameters of a Gamma distribution, based on the sampled data, using the Univariate Procedure. The cumulative Gamma distribution was then used to calculate sensitivity and specificity.

Leave-one-out cross-validation
In the leave-one-out cross validation a single subject is removed from the study population and used in the validation process. In the remaining (n-1) subjects the cut-off value is determined in a data-driven way, as described above. Thereafter, the resulting cut-off is applied to the single subject that did not take part in this process. This subject is then classified as either true positive, false positive, false negative, true negative depending on whether the subject is classified as having or not having the disease and whether its test result is below or above the cut-off value. This process is repeated for all patients in the data set and the resulting 2-by-2 table based on all subjects is used to determine sensitivity and specificity corresponding to the cut-off value which was derived in the n-1 patients.

Robust ROC curve fitting
In the robust ROC fitting approach a smooth, non-parametric curve is fitted through the observed data points plotted in ROC through a smoothing procedure which is included in SAS software (Loess Procedure). The point on the fitted curve with the highest Youden index was used to obtain estimates of sensitivity and specificity.

4.2.4 Empirical evidence from published diagnostic reviews
From a set of 28 published systematic reviews, used in a previously published meta-epidemiological project, we selected those reviews that reported on continuous test results and included both studies with and without a pre-specified cut-off value. We then compared the summary diagnostic odds ratio between those two groups to ex-

amine whether the diagnostic accuracy was higher (overestimated) in studies with data driven selection of cut-off values than in studies using pre-specified cut-off values. The diagnostic odds ratio is an overall measure of accuracy combining both sensitivity and specificity: [sens/(1-sens)]\[(1-spec)/spec]. Further details about this set of systematic reviews and the applied statistical methods can be found in an earlier publication[14].

4.3 Results

4.3.1 Simulation of data sets

In the basic scenario, index test results were generated from a Normal distribution with a mean value of 100 (SD=10) for persons without the disease and a mean value of 120 (SD=10) for persons with the disease, leading to a true maximum Youden index of 0.68, a true optimal cut-off value of 110 and true values of sensitivity and specificity of both 84%. These true values will only alter if the underlying distribution changes (like the difference in means between diseased and non-diseased or the spread of test results), but are not affected by changes in sample size or disease prevalence.

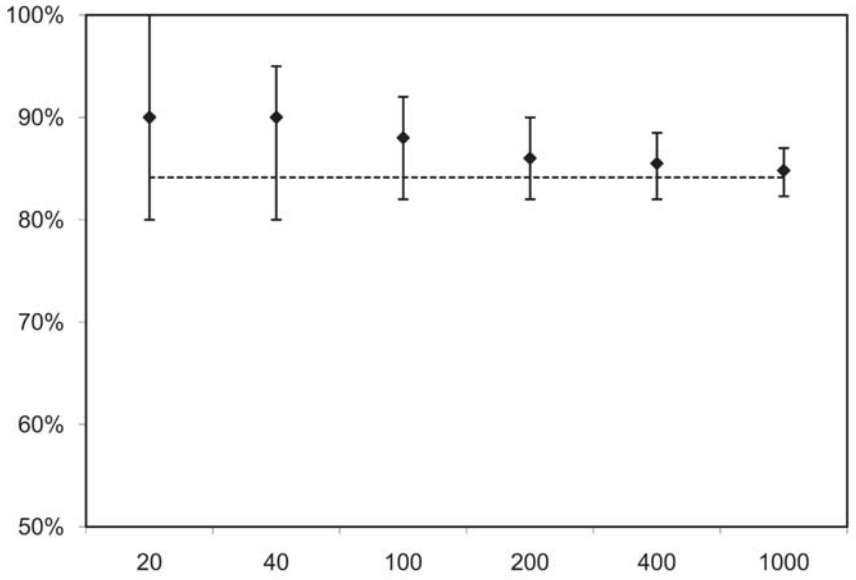


Figure 4.3. Effect of sample size on data-driven estimates of sensitivity.
The median sensitivity across 2000 simulations together with the 25th and 75th percentiles are shown. On the X-axis: the number of persons with the disease (disease prevalence was 50%). Data based on Normally distributed test results with mean=100 and SD=10 for non-diseased and mean=120 and SD=10 for diseased. The dotted line represents the true value of sensitivity. The results for specificity were similar.

4.3.2 Data driven overestimation of sensitivity and specificity

The effect of sample size

The amount of bias in the data-driven estimates was inversely related to sample size (Figure 4.3). At a total sample size of 40, the median sensitivity and specificity were both 90% (interquartile range 80 to 95%), while their true values were both 84%. Both measures were overestimated in 74% of the simulations. In sixty percent of the simulations, estimates of sensitivity and specificity exceeded 89%, while their true value was 84%. When the total sample size was 200, sensitivity was overestimated in 62% and specificity in 60% of all simulations, while their median values were approaching their true values (86% (interquartile range 82 to 89%) in stead of 84%).

The effect of disease prevalence

A prevalence of 50% is the most efficient prevalence to ensure that combined uncertainty in both sensitivity and specificity is smallest. This was also reflected in our results. Lowering the prevalence (conditional on the same total sample size) leads to fewer individuals with the disease, larger fluctuation in sensitivity by chance and therefore more room for overestimation of sensitivity. The opposite occurs for specificity. The median absolute bias at a prevalence of 10% was 5.9% for sensitivity and 3.6% for specificity. At a prevalence of 90%, the median absolute bias was 2.2% for sensitivity and 6.7% for specificity (results not shown).

Overlap in test results between populations with and without the disease

The spread and overlap in test results between populations with and without the disease determines the absolute size of sensitivity and specificity. A smaller standard deviation (less spread) while the difference in mean values between the populations remains the same, will lead to less overlap in test results between diseased and non-diseased. Thus, sensitivity and specificity will increase, leaving less room for overestimation (ceiling effect): sensitivity cannot exceed 100%. On the other hand, if we allow the standard deviations to change without changing sensitivity and specificity, then the amount of bias did not vary (results not shown).

The effect of underlying distributions

The underlying distribution of the simulated test results by comparing scenarios based on a Normal, Lognormal or Gamma distribution had little impact on the average amount of bias (see Figures 4.4 and 4.5). However, the amount of bias could vary substantially within a specific distribution based on the actual values of the parameters of that distribution. For example, one of the Lognormal distributions resulted in 60% of the simulations with an overestimation of sensitivity that was more than 5% points, while the other Lognormal distribution resulted in such an overestimation in 35% of the simulations.

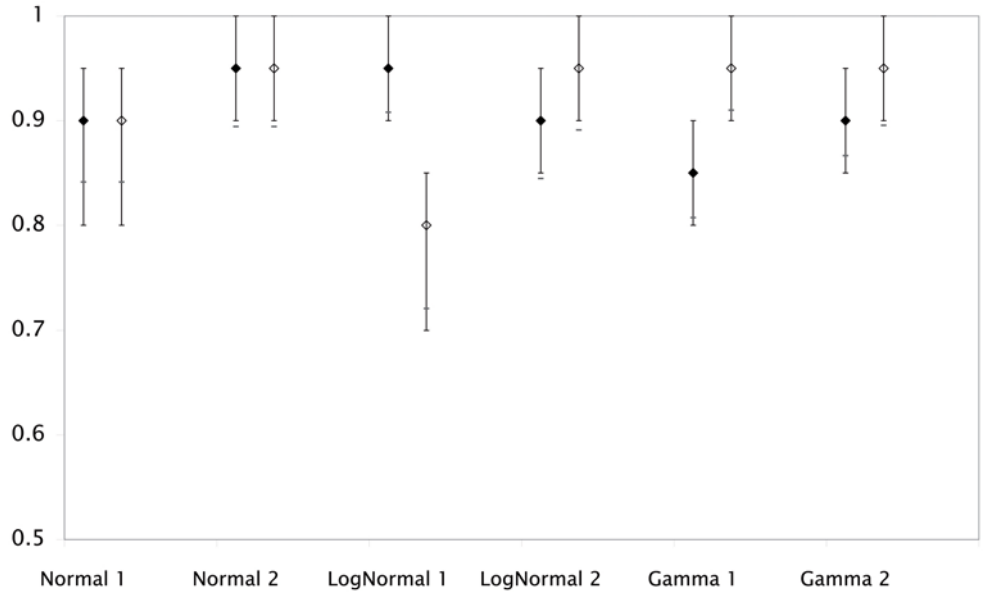


Figure 4.4. The effect of the underlying distribution on sensitivity and specificity.
The closed diamonds are the median data driven values of the sensitivities and the open diamonds are the median values of the specificities. Also shown are the data-driven 25th and 75th percentiles and the true values (dashes). Prevalence was in all situations 50% and total sample size was 40. Normal distribution 1: mean(SD) diseased = 120(10) and mean(SD) non-diseased = 100(10). Normal distribution 2: mean(SD) diseased = 122.5(10) and mean(SD) non-diseased = 97.5(10). The Lognormal and Gamma distributions are shown in Figure 4.2.

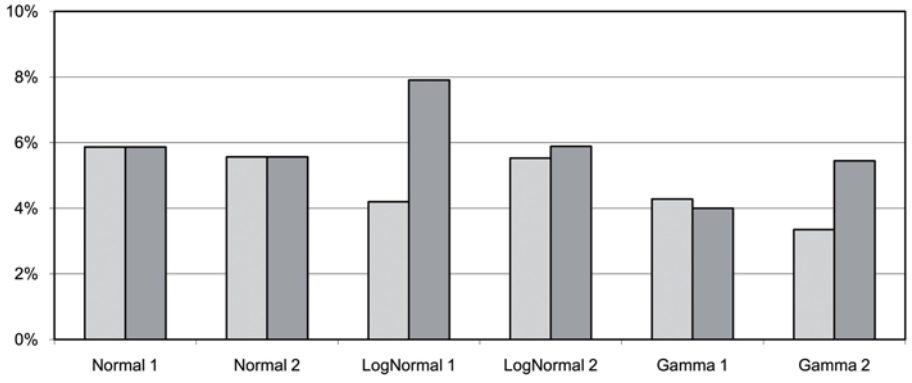


Figure 4.5. Effect of the underlying distribution on the absolute amount of bias in sensitivity (light) and specificity (dark).
Prevalence was in all situations 50% and total sample size was 40. On the Y-axis the absolute bias in % points above the true value. Normal distribution 1: mean(SD) diseased = 120(10) and mean(SD) non-diseased = 100(10). Normal distribution 2: mean(SD) diseased = 122.5(10) and mean(SD) non-diseased = 97.5(10). The Lognormal and Gamma distributions are shown in Figure 4.2.

4.3.3 Potential solutions to reduce bias

Deriving optimal cut-off point from assumed distributions

Using the estimated mean and standard deviation from a data set and then calculating the true optimal cut-off value by assuming a Normal distribution, decreases the amount of bias when the underlying distribution was indeed Normal. In one of the scenarios with a true underlying Normal distribution of the index test results, median sensitivity and specificity following this strategy were both 85%, while their true value was 84%. This is a difference of only 1% point (see Figure 4.6).

When the underlying distribution is a Gamma or Lognormal one, this same procedure leads to a systematic underestimation of sensitivity and overestimation of specificity, which was sometimes worse than the uncorrected, data-driven results. In these situations, the median estimated sensitivity was 2-13% points lower than the true sensitivity (see Figure 4.6). The difference between the median estimated specificity and the underlying true specificity was 7 or 8% points.

The Gamma distribution is more flexible in approximating various distributions and led to less bias in all scenarios than the data-driven method. The median estimated sensitivity varied from 2% points below to 3% points above the true sensitivity. The median estimated specificity varied from 1% points to 4% points above the true specificity.

Because we sometimes observed that results for sensitivity and specificity were in the opposite direction (overestimation in one and underestimation in the other parameter), we summed the absolute value of the bias in sensitivity and specificity. When we assumed the underlying distributions to be Normal, the total absolute value of bias was 59% points (summed absolute bias of all sensitivities in all five studied scenarios was 28% points, summed bias of all specificities in all five sce-

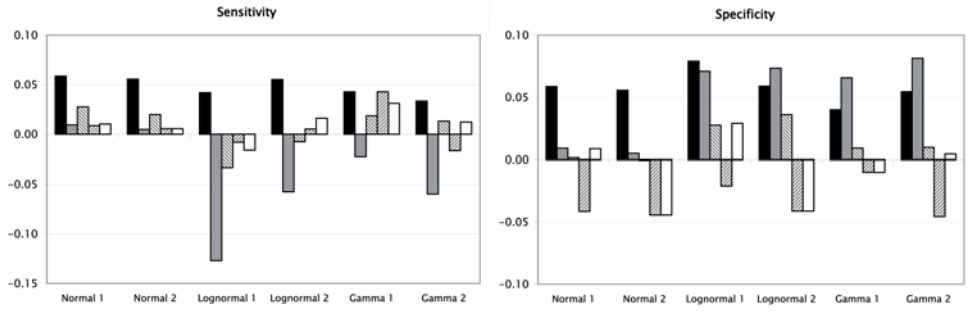


Figure 4.6. Figures 4.6a, bias in sensitivity, and 4.6b, bias in specificity.
Median amount of bias of the data-driven and alternative approaches in relation to the true value. On the Y-axis the absolute amount of bias in % points above or under the true value. Disease prevalence was 50% in all situations and total sample size was 40. Black bars, data-driven analysis; grey bars, assuming a Normal distribution; top-left-to-bottom-right striped bars, assuming a Gamma distribution; top-right-to-bottom-left striped bars, leave-one out validation; white bars, robust ROC fitting.

narios was 31% points). When we assumed underlying Gamma distributions, the total absolute value of bias was 20% points (sum of absolute bias in sensitivity was 12% points and in specificity 8% points).

Leave-one-out cross-validation
The leave-one-out cross validation resulted in less bias in sensitivity, the estimated values were 2% lower to 4% higher than their true values. Specificity was a marginally underestimated (1 to 5% lower than their true value) (see Figure 4.6). The sum of the absolute value of the bias in sensitivity was 9% points and in specificity 20% points (total bias of 29% points).

Robust ROC curve fitting
Robust fitting of ROC curves also resulted in less bias in both sensitivity and specificity. Difference between true and estimated sensitivity ranged from minus 2% points to 3 % points and difference between true and estimated specificity ranged from minus 4% points to 3% points (see Figure 4.6). The sum of the absolute value of the bias in sensitivity was 9% points and in specificity 14% points (total bias of 23% points).

4.3.4 Empirical evidence from published diagnostic reviews

A total of seven systematic reviews evaluated a test producing continuous test results and five of these reviews included both studies with a pre-specified cut-off value and studies with a data driven cut-off value. The diagnostic odds ratio on average was 1.71 (95% confidence interval: 1.04 to 2.82; *P*=0.03) times higher in studies with a data-driven cut-off value compared to studies with a pre-specified cut-off value. Translating this results to sensitivity and specificity, it means that if a study with a pre-specified cut-off value would estimate sensitivity and specificity both at 84% (=diagnostic odds ratio of 28), a study using data-driven selection would find estimates of sensitivity and specificity of 87.4% , corresponding to a diagnostic odds ratio of 48 (=28 times 1.71).

4.4 Discussion

Our simulation study showed that data-driven selection of the optimal cut-off values for a continuous test by using the Youden index led to overestimated estimates of sensitivity and specificity. The amount of bias in sensitivity and specificity was predominantly dependent on total sample size. A typical value for the absolute amount of bias in studies with a sample size of 40 was 5% points occurring in both sensitivity and specificity.

The amount of bias becomes smaller by increasing sample size. Overestimation of more than 5% was present in 27% of the simulations if the total sample size was 200 compared to 60% of the studies with sample size of 40. The underlying distri-

butions had little or no effect on the amount of bias. This can be explained by the non-parametric way of data driven selection of the optimal cut-off value. The absolute magnitude of the true sensitivity and specificity did have an effect: the nearer the true values approached 100%, the less room there was for overestimation.

In this study, we have only reported the effect of optimizing cut-off values on sensitivity and specificity, although we also examined the effects on likelihood ratios and diagnostic odds ratios (results not reported). These effects were in line with the results on sensitivity and specificity, which is not surprising because they are direct functions of sensitivity and specificity. This potential for bias was confirmed in our empirical data, as the diagnostic odds ratio in studies with data-driven cut-off values was significantly higher than in studies with pre-specified values.

We applied three alternative and more robust methods for determining the sensitivity and specificity associated with the optimal cut-off value to examine whether these methods were less prone to bias. In general, these methods resulted in lower estimates of sensitivities and specificities, sometimes even producing too conservative estimates (see Figure 4. 6). As expected, the performance of the method which assumes that the underlying distribution was Normal deteriorated considerably if this assumption was not met. Because it is difficult to examine in a small sample whether it is reasonable to assume a Normal underlying distribution, we do not recommend this method in general. Assuming a Gamma distribution is a more flexible approach, as it can mimic various shapes of distribution and therefore this method performed consistently well across our simulations. The smooth ROC fitting can be viewed as a distribution-free method, meaning that it would perform consistently irrespective of the true underlying distribution. The leave-one-out approach is a traditional way of cross validating the results in regression analyses to reduce the impact of over fitting. In our situation, the leave-one-out approach produced indeed lower estimates than the data-driven method. However, sometimes the estimates form the leave-one-out approach became too conservative, especially for specificity. We do not have an explanation for this. Bootstrapping would have been a slightly different approach based on the same principle of cross-validation. Therefore, we expect similar results with this method as with the leave-one-out approach.

Another approach that will reduce the problem of overestimation is using a pre-specified cut-off value. However, in the early phase of test evaluation, there may be little indication about the likely value of the optimal cut-off value. Other more complex solutions to generate less biased results, but still use the actual data of a study have been described. These involve the reporting of a confidence interval around the 'true' cut-off value and a Bayesian method to smooth the steps in an ROC curve. Details can be found here[5,15].

Readers of diagnostic studies should be aware of the potential for bias when optimal cut-off values have been derived in a data-driven way, especially if the sample size was small. Defining a small study is rather arbitrary and depends on the amount

of bias you are willing to accept. Our results show that there is probability of 27% that sensitivity and specificity will be overestimated more than 5% points in a study with a sample size 200. A rule of thumb would be that a diagnostic study should have at least 100 individuals without the disease as well as 100 individuals with the disease before a cut-off value can be reliably estimated from the data. The problem is however, that most diagnostic studies will not have these numbers[6]. Another problem both clinicians and laboratory professionals may encounter, is that not only the amount of bias will increase if sample sizes get smaller, also the confidence interval around the estimate of the optimal cut-off value and of both sensitivity and specificity will increase. Even if bias is reduced by using more robust methods, uncertainty about the true optimal cut-off value and its corresponding diagnostic accuracy will remain.

In conclusion, researchers and readers of diagnostic studies should be aware of over optimistic measures of diagnostic accuracy when the results have been generated by a data-driven approach in a small study. Several methods exist that can reduce the amount of this bias, but it is important to stress that finding robust estimates of cut-off values and their associated measures of accuracy require studies of considerable sample size. In smaller studies, researchers may present a scatter graph showing the distribution of all test results in the non-diseased and the diseased individuals. In addition they can draw the empirical ROC curve and a robust (smoothed) ROC curve, but refrain from selecting the most outlying point closest to the top left corner (=maximum Youden).

References

1. Shapiro DE. The interpretation of diagnostic tests. Stat Meth Med Res. 1999: 8(11);113–34.
2. Obuchowski NA, Lieber ML, Wians FH Jr. ROC curves in Clinical Chemistry: uses, misuses, and possible solutions. Clin Chem. 2004; 50(7):1118–25.
3. Greiner M, Pfeiffer D, Smith RD. Principles and practical application of the receiver–operating characteristic analysis for diagnostic tests. Prev Vet Med. 2000: 45(1–2);23–41.
4. Youden WJ. Index for rating diagnostic tests. Cancer. 1950: 3(1);32–5.
5. Fluss R, Faraggi D, Reiser B. Estimation of the Youden index and its associated cutoff point. Biometrical J. 2005; 47(4):458–72.
6. Bachmann LM, Puhan MA, ter Riet G, Bossuyt PM. Sample sizes of studies on diagnostic accuracy: literature survey. BMJ. 2006; 332(7550):1127–9.
7. Linnet K, Brandt E. Assessing diagnostic tests once an optimal cutoff point has been selected. Clin Chem. 1986; 32(7):1341–46.
8. Le CT. A solution for the most basic optimization problem associated with an ROC curve. Stat Methods Med Res. 2006; 15(6):571–84.
9. Ewald B. Post hoc choice of cut points introduced bias to diagnostic research. J Clin Epidemiol. 2006; 59(8):798–801.
10. Jund J, Rabilloud M, Wallon M, Ecochard R. Methods to estimate the optimal threshold for normally or log–normally distributed biological tests. Med Decis Making. 2005; 25(4): 406–15.
11. Perkins NJ, Schisterman EF. The Youden index and the optimal cut–point correctment for measurement error. Biometrical J. 2005; 47(7):428–441.
12. Schisterman EF, Perkins NJ, Liu A, Bondell H. Optimal cut–point and its corresponding Youden Index to discriminate individuals using pooled blood samples. Epidemiology. 2005; 16(1):73–81.
13. Altman DG, Lausen B, Sauerbrei W, Schumacher M. Dangers of using "optimal" cutpoints in the evaluation of prognostic factors. J Natl Cancer Inst. 1994; 86(11):829–35.
14. Leeflang M, Reitsma J, Scholten R, Rutjes A, Di Nisio M, Deeks J, Bossuyt P. Impact of adjustment for quality on results of meta–analyses of diagnostic accuracy. Clin Chem. 2007; 53(2):164–72.
15. Gail MH, Green SB. A generalization of the one–sided two–sample Kolmogorov–Smirnov statistic for evaluating diagnostic tests. Biometrics 1976; 32(3):561–570.