Frank Samuelson[1] / Craig Abbey[2]

Using Relative Statistics and Approximate Disease Prevalence to Compare Screening Tests

[1] Division of Imaging, Diagnostics and Software Reliability, US Food and Drug Administration, 10903 New Hampshire Ave Building 62, Room 3102, Silver Spring, MD 20903-1058, USA, E-mail: frank.samuelson@fda.hhs.gov. http://orcid.org/0000-0002-1130-0303.

[2] Department of Psychological and Brain Sciences, University of California Santa Barbara, Santa Barbara, CA, USA. http://orcid.org/0000-0003-4608-9402.

Abstract:
Schatzkin et al. and other authors demonstrated that the ratios of some conditional statistics such as the true positive fraction are equal to the ratios of unconditional statistics, such as disease detection rates, and therefore we can calculate these ratios between two screening tests on the same population even if negative test patients are not followed with a reference procedure and the true and false negative rates are unknown. We demonstrate that this same property applies to an expected utility metric. We also demonstrate that while simple estimates of relative specificities and relative areas under ROC curves (AUC) do depend on the unknown negative rates, we can write these ratios in terms of disease prevalence, and the dependence of these ratios on a posited prevalence is often weak particularly if that prevalence is small or the performance of the two screening tests is similar. Therefore we can estimate relative specificity or AUC with little loss of accuracy, if we use an approximate value of disease prevalence.

Keywords: diagnostic, screening, ratios, prevalence, relative statistics

DOI: 10.1515/ijb-2016-0017

1 Introduction

Some large, well-funded clinical trials of diagnostic tests follow all patients in the study to determine whether those patients were truly diseased and whether the diagnostic tests were correct. These studies will report sensitivities, specificities, and sometimes areas under the ROC curve (AUC). However, there are also many large, observational, screening studies that do not follow all patients. In these studies only patients found positive by the diagnostic test to go through further confirmatory diagnoses or reference procedures. For examples see Friedewald et al. [1] or Brem et al. [2]. This happens in clinical practice because the confirmatory gold-standard procedures are frequently invasive, costly, or time consuming. In these large observational clinical studies authors may report only detection and recall rates. Here the detection rate is defined as the fraction of all patients who have the disease and are called positive by the diagnostic test. The recall rate is the fraction of patients who test positively. To calculate detection and recall rates only patients found positive by the diagnostic test need to be followed.

In the notation of Table 1, the number of true positives t and false positives f are measured in these observational clinical studies, but the number of true and false negatives, v and u, in the sample are unknown. The numbers of patients who are truly diseased $n = t + u$ or not diseased $m = f + v$ are also unknown. The detection rate can be calculated as $d = t/N$ where $N = t + f + v + u$ is the total number of patients in the study. The recall rate is estimated as $r = (t + f)/N$. Typical estimates of conditional probabilities such as the true positive rate (TPR) or sensitivity, t/n, or the true negative rate (TNR) or specificity, v/m, are not possible. The usual estimate of the prevalence of disease in the population, $\hat{\pi} = n/N$,[1]can not be calculated.

Table 1 A contingency table from a hypothetical study. This table gives the number of patients who have disease, the number who do not, and how many of each of those tested positively. In many clinical studies only t and f are known.

		Diagnostic Test		
		Positive	Negative	Sum
Patient state	Diseased	t, true positives	u, false negatives	n
	Not Diseased	f, false positives	v, true negatives	m

Frank Samuelson is the corresponding author.

Typically we use measures of sensitivity and specificity to determine which of two diagnostic tests is superior. Many observational studies report results for two diagnostic tests on the same set of patients or on the same patient population. Though we can not calculate an absolute sensitivity or specificity of either diagnostic from these studies, we can calculate relative sensitivities or false positive rates between two tests.

Schatzkin et al. [3] demonstrated that estimates of ratios of the conditional positive rates (the true positive rate and the false positive rate) of diagnostic tests X and Y on the same sample can be calculated from the measured positive values. Indeed the relative sensitivity or relative true positive rate estimate,

$$\mathrm{rSens} = \mathrm{rTPR} = \frac{\mathrm{TPR}_X}{\mathrm{TPR}_Y} = \frac{t_X/n}{t_Y/n} = \frac{t_X}{t_Y}, \tag{1}$$

is the ratio of two estimated detection rates, $(t_X/N)/(t_Y/N) = (t_X/t_Y)$, where t_X is the number of true positives of test X. (Note that estimating ratios in this manner is slightly biased, particularly when t_X is small.) The relative sensitivity of two tests on different samples from the same population can be estimated as

$$\mathrm{rSens} = \mathrm{rTPR} = \frac{\mathrm{TPR}_X}{\mathrm{TPR}_Y} = \frac{t_X/n_X}{t_Y/n_Y} = \frac{t_X/(N_X\pi)}{t_Y/(N_Y\pi)} = \frac{t_X}{t_Y}\frac{N_Y}{N_X} \tag{2}$$

which is also the ratio of two estimated detection rates where N_X is the total number of patients in the study of test X. This estimate may be more variable than a ratio that was calculated on same set of patients (eq. (1)), because the actual fraction of diseased patients in the two samples may be different, though the expected fractions are the same. Other authors [4, 5] have developed approaches for performing inference on estimates of ratios of statistics from screening data without disease confirmation.

Schatzkin et al. [3] also showed that the relative false positive rate is estimated as

$$\mathrm{rFPR} = \frac{\mathrm{FPR}_X}{\mathrm{FPR}_Y} = \frac{f_X/m_X}{f_Y/m_Y} = \frac{f_X}{N_X(1-\pi)}\frac{N_Y(1-\pi)}{f_Y} = \frac{f_X}{N_X}\frac{N_Y}{f_Y}. \tag{3}$$

This ratio is calculable using only the numbers of false positives and numbers of patients, which are measured in observational studies.

With these ratios we can compare diagnostic tests. If we were to perform a study, and the relative sensitivity of X to Y were 1.1, and rFPR were 0.9, then we would judge test X to be about 10% better on both diseased and non-diseased patients, indicating that it was an overall superior test. However if the ratios of both positive rates were greater than 1, then superiority would be ambiguous. Therefore before performing the study we may want to select a single study endpoint that is less ambiguous, such as AUC, on which we would perform our inference. This paper examines the calculation of relative AUC, expected utility, and specificity in scenarios where the number of false negatives is unmeasured, and the prevalence may only be known approximately.

Abbey et al. [7] defined the metric “expected utility”. We can write it as

$$\mathrm{EU} = \mathrm{TPR} - \frac{1-\pi}{\pi U_r}\mathrm{FPR}. \tag{4}$$

Expected utility is a measure proportional to the total utility (benefits minus costs) of performing screening with a particular diagnostic test in the intended population. Calculated from the positive rates it shows how utility changes as a function of the decision threshold used on the diagnostic. U_r is the relative utility, which can be thought of as a ratio of the utility benefit for making a correct decision when a patient has the disease to that of making a correct decision when a patient does not have the disease. Abbey et al. [6] and [7] examined large clinical studies and estimated U_r to be approximately 162 for breast cancer screening. This indicates that the average benefit of finding a breast cancer is considered to be about 162 times greater than the benefit of calling a normal patient negative. Another quantity, the “recall-corrected detection rate” [7], is proportional to expected utility and can be written as

$$\mathrm{RCDR} = d - \frac{r}{1+U_r}. \tag{5}$$

This metric is the detection rate penalized by a utility-dependent fraction of recalls.

In Section 2 we demonstrate that ratios of expected utility, have the same property as the positive rates. Those relative utilities can be estimated independently of the number or fraction of diseased patients in the sample.

In Section 3 we demonstrate that for some metrics which depend upon the number or fraction of diseased patients, such as specificity and AUC, the ratios of such metrics typically depend only weakly on the prevalence of disease in the population, particularly if the prevalence is low and the difference in performance of diagnostic tests is small. Therefore we can use approximate values of prevalence to estimate ratios of specificities and AUCs with little loss of accuracy for many screening studies of diagnostic tests.

2 Ratios of expected utilities

Even if we did not measure the number of true or false negatives in an observational clinical study, we can estimate the true and false positive rates as

$$\mathrm{TPR} = \frac{t}{N\pi} \tag{6}$$

$$\mathrm{FPR} = \frac{f}{N(1-\pi)} \tag{7}$$

if we know the disease prevalence π. Then we can write the ratio of two estimated expected utility (eq. (4)) expressions for two tests X and Y as

$$\mathrm{rEU} = \frac{\mathrm{EU}_X}{\mathrm{EU}_Y} = \frac{\frac{t_X}{N_X\pi} - \frac{1-\pi}{\pi U_r}\frac{f_X}{N_X(1-\pi)}}{\frac{t_Y}{N_Y\pi} - \frac{1-\pi}{\pi U_r}\frac{f_Y}{N_Y(1-\pi)}} = \frac{t_X - f_X/U_r}{t_Y - f_Y/U_r}\cdot\frac{N_Y}{N_X}. \tag{8}$$

This relative EU value, like relative TPR and FPR, can be calculated from a study where the only patients who test positively have confirmed diagnoses, as described in the introduction. Specifically it depends only upon the numbers of true and false positives and sample sizes. Because expected utility is proportional to the total utility of performing a diagnostic test, ratios of total utility must also be independent of disease prevalence.

Using our previous definitions of recall and detection rates, $d = t/N$ and $r = (t+f)/N$, we can show that the ratio of two recall-corrected detection rates eq.(5) is

$$\frac{\mathrm{RCDR}_X}{\mathrm{RCDR}_Y} = \frac{\frac{t_X}{N_X} - \frac{t_X+f_X}{N_X(1+U_r)}}{\frac{t_Y}{N_Y} - \frac{t_Y+f_Y}{N_Y(1+U_r)}} = \frac{t_X(1+U_r) - t_X - f_X}{t_Y(1+U_r) - t_Y - f_Y}\cdot\frac{N_Y}{N_X} = \frac{t_X - f_X/U_r}{t_Y - f_Y/U_r}\cdot\frac{N_Y}{N_X},$$

which is the same as the ratio of two expected utilities in eq. (8),

$$\frac{\mathrm{RCDR}_X}{\mathrm{RCDR}_Y} = \frac{\mathrm{EU}_X}{\mathrm{EU}_Y}.$$

3 Estimating ratios of metrics with weak dependence on prevalence

Schatzkin et al. [3] showed that are a number of statistics, such as TPR and FPR, whose ratios we can estimate independently of the total number or prevalence of diseased patients. Therefore ratios of these statistics can be estimated from studies where there is no confirmation of disease status for negatively tested patients. In addition we show below that there are some statistics, such as specificity and AUC, whose ratios we can estimate with little loss of accuracy without knowing the number of diseased patients.

Frequently we can do the following.

(1) Write an estimate of our statistic Z in terms of the measured true and false positives and the true unknown prevalence, π.

(2) Show that the dependence of the estimated ratio of two of these statistics, Z_X/Z_Y, on the common disease prevalence π is weak for the scenario of interest.

(3) Use an approximate, commonly accepted, value of prevalence π_p in our estimate of the ratio. The use of this approximate value of prevalence π_p instead of the true population value will bias our estimates of Z_X/Z_Y. However for screening tests, the prevalence of disease is small, and we expect that the magnitude of an error in our posited prevalence can not be much larger than the true value of prevalence itself. Therefore the bias in our estimates of Z_X/Z_Y due to our uncertainty in π should be small.

(4) Perform a sensitivity analysis to demonstrate empirically that the uncertainty in the value of π_p leads to an absolute bias in the estimated ratio that is below some tolerated limit.

For step 4 a typical, tolerated limit is the statistical standard error of the ratio itself. If we are willing to accept a value for a statistic that is within two standard errors of a measured value, then in general we will accept an absolute systematic bias below one standard error, as it increases the total root-mean-squared error less than 40%. In this paper we do not discuss how to calculate this statistical standard error, because it depends on the design, experimental factors, and correlations in the study. Here we just assume that variance calculations will be done properly, and other texts deal with these calculations.

In the following sections we show that we can do the above steps for specificity and a simple estimate of AUC.

3.1 Specificity

As Schatzkin et al. [3] points out, the specificity and ratios of specificity estimators

$$\mathrm{rTNR} = \mathrm{rSpec} = \frac{v_X/m}{v_Y/m} = \frac{v_X}{v_Y}$$

depend upon the number of true negatives, which we do not know. While the specificites do not contribute any statistical inference beyond what the false positive rates provide, we still may wish to estimate their ratios for the purposes of communicating the differences between two tests to others.

We can rewrite the estimate of specificity in terms of the prevalence,

$$\mathrm{TNR} = \mathrm{Spec} = 1 - \frac{f}{N(1-\pi)}, \tag{9}$$

For screening tests π is typically small compared to one. Therefore using a first order Taylor approximation we can write

$$\mathrm{TNR} = \mathrm{Spec} \approx 1 - \frac{f}{N}(1+\pi). \tag{10}$$

If π is small, then its contribution to the specificity is small.

If we use for π a value of prevalence inferred from other published studies on similar populations of patients, π_p, then there will be a bias in our estimate of specificity because this is not the true prevalence. However, we expect that this bias can not be much larger than the true value of prevalence. So if prevalence is small, any bias due to its mis-estimation in eq. (10) is also small.

From eq. (9) the relative specificity of two tests X and Y can be written

$$\mathrm{rTNR} = \mathrm{rSpec} = \frac{1 - f_X/N_X - \pi}{1 - f_Y/N_Y - \pi}. \tag{11}$$

A Taylor series expansion of this equation about $\pi = 0$ is

$$\mathrm{rSpec} = 1 + \frac{z_X - z_Y}{z_Y}\left(1 + \frac{\pi}{z_Y} + \left(\frac{\pi}{z_Y}\right)^2 + \ldots\right) \tag{12}$$

where $z_X = 1 - f_X/N_X$ and $z_Y = 1 - f_Y/N_Y$. From this expression we see that if the approximate prevalence is much less than the fraction of all patients who are not false positives, i.e. $\pi \ll 1 - f_Y/N_Y$ or $n_Y \ll t_Y + u_Y + v_Y$, then the relative specificity will be very weakly dependent upon the value of that prevalence. This is frequently the case in clinical screening studies of diagnostic tests. Typically the rate of disease in the tested population or sample is much smaller than the true negative rate.

If we let δ be the error in our posited prevalence, $\pi_p = \pi + \delta$, and if we use π_p in expression (14) instead of the true population value π, then the first order bias in that expression is

$$\frac{z_X - z_Y}{z_Y^2}\delta. \tag{13}$$

This expression comes from the first π dependent term in eq. (12) and implies that the bias in the ratio of two estimated specificities will be very small if the difference in false positive rates between the two tests X and Y is much smaller than z_Y and if the error in our prevalence is small, which is probably the case when the prevalence itself is small. If this bias is small, then we can make estimates of the relative specificity without knowing the number of true negatives. An example of a small bias in relative specificity is given in Section 4.

3.2 AUC

Baker and Pinsky [8] demonstrated that ratios of a partial area under an empirical ROC curve below a recall threshold can also be calculated without knowledge of the total number or prevalence π of actually diseased patients. Shaw et al. [9] demonstrated the construction of ROC curves using only data from patients who were called positive by a diagnostic test. Here we demonstrate that under certain conditions the ratios of total estimated areas under parametric ROC curves are weakly dependent upon an estimated prevalence.

Frequently published observational clinical studies do not report ROC data for the diagnostic tests under investigation. They present only a single detection and single recall measurement for each diagnostic test. If we write estimators of the true and false positive rates in terms of the number of positives and an unknown prevalence, as in eqs (6) and (7), then we can estimate an approximate AUC value for the diagnostic. If that TPR, FPR point lies upon the ROC curve of the screening diagnostic, and we assume a parametric form of the diagnostic test's ROC curve described by a single parameter, then we can choose the parametric ROC curve that passes through that point as an estimate of an ROC curve and the area under that curve as an estimate of AUC.

For example if we assume a power-law ROC model ([10, 11]) the theoretical relationship between the true positive rate p_t and the false positive rate p_f is $p_t = p_f^\beta$, or

$$\beta = \log p_t / \log p_f. \tag{14}$$

The area under an ROC curve is the integral of p_t over p_f,

$$\mathrm{AUC}_P = \int_0^1 p_t \cdot \mathrm{d}p_f = \int_0^1 p_f^\beta \cdot \mathrm{d}p_f = \frac{1}{1+\beta}. \tag{15}$$

Our ROC curve estimate should pass through the estimated (TPR, FPR) point, where TPR and FPR may be estimated from eqs (6) and (7). Inserting TPR for p_t and FPR for p_f into our equation for β eq. (14), and putting that β into eq. (15), we get our parametric estimate of the AUC,

$$\mathrm{AUC}_P = \frac{\log \mathrm{FPR}}{\log \mathrm{TPR} + \log \mathrm{FPR}} = \frac{\log \frac{f}{N(1-\pi)}}{\log \frac{t}{N\pi} + \log \frac{f}{N(1-\pi)}}. \tag{16}$$

Using the same methodology, if we assume a parametric, equal-variance, bi-normal ROC model, then an estimate of AUC is

$$\mathrm{AUC}_N = \Phi\left(\frac{\Phi^{-1}(1-\mathrm{FPR}) - \Phi^{-1}(1-\mathrm{TPR})}{\sqrt{2}}\right) \tag{17}$$

where Φ is the standard cumulative normal distribution.

In general we do not recommend estimating values of AUC from a single estimated true positive rate and a single estimated false positive rate. Estimates of AUC derived from multiple decision points with different FPR and TPR values [10] rely on fewer assumptions about the distribution of the data. However, if a recall, detection or sensitivity, specificity pair is all the data available, e.g. in a meta-analysis, then such an estimated AUC may still be informative or useful for comparisons with other studies.

If we let $w_X = \log \mathrm{FPR}_X + \log \mathrm{TPR}_X$, then an estimator of the ratio of two power-law AUC estimators (eq. (16)) for two diagnostic tests X and Y is

$$\mathrm{rAUC}_P = \frac{\log \mathrm{FPR}_X}{\log \mathrm{FPR}_Y} \cdot \frac{w_Y}{w_X} \tag{18}$$

If we assume that π is small, and our posited estimate of prevalence is R times larger than the true prevalence, $\pi_p = R\pi$, then the bias in our estimate of rAUC_P due to this posited prevalence is approximately

$$\frac{\log \mathrm{FPR}_X}{\log \mathrm{FPR}_Y} \cdot \frac{w_Y - w_X}{w_X^2} \log R. \tag{19}$$

We give a short proof of this approximation in the appendix.

Here we see that the bias in the relative AUC depends upon the logarithm of the fractional error in the assumed prevalence. If this error is not far from unity, then the logarithm of it should be close to zero. Likewise this bias is small when the difference between w_Y and w_X is small. In the Example section we analyze a study where this bias in rAUC_P is small compared to the statistical uncertainty of the statistic.

3.3 Discussion

The errors in the estimates of relative specificity and relative AUC depend upon the true value of prevalence, the error in our assumed prevalence value, the positive rates, and the difference of the test results themselves. The dependence upon the true value of π is not evident in approximate expressions (16) or (22), because these expressions assume a π that is not large. Indeed when π is small, the errors in these ratios depend little on true value π itself and depend more on the error in π (δ or R) and the difference in performance of the tests. When π is large these approximate expressions are not as accurate and the actual errors of the ratios have a stronger dependence on π.

If the difference in positive rates between the tests is small, then the error in our posited prevalence can be quite large without affecting the estimates of relative specificty or relative AUC. This is frequently the case if the test X is an incremental improvement over test Y.

4 Example

As an example we use a study from Rose et al. [12]. In this study recall rates and detection rates in breast cancer screening were compared for two diagnostic tests. The first test was digital mammography (DM, or test Y) and the second test was digital mammography combined with digital breast tomosynthesis (DM+DBT, or test X). In this study the patients on whom the two tests were applied were not all the same, but we can assume that the two tests were applied on the same population with the same prevalence of disease.

Rose et al. give the number of recalls and detections for each reader who participated in the study for both DM+DBT and DM. From these numbers we calculated estimates of the average ratios of sensitivity, false positive rate, and expected utility between these two tests. These values and their estimated errors are given in Table 2. These estimated values do not depend on the number or prevalence of truly diseased patients.

We calculate all these relative statistics in this section to demonstrate their values and errors in an applicable study, not to infer anything from the study. When analyzing a study for purposes of diagnostic inference, we might only calculate the predetermined endpoint of the study and its error. But if we were to calculate multiple diagnostic metrics for inference, as is quite common in applications, it would be important to report p-values and confidence intervals that account for the testing of these multiple metrics [13].

We performed a bootstrap [14] across the six readers to estimate the uncertainty in our ratios. Because of the variability among readers, this method gave estimates of variance larger than those found in the original Rose et al. paper, whose error estimates are binomial based on the patient sample size. Because cases are nested within readers, our method should be accurate. The bootstrap distributions were not normal, so bootstrap or other [4] confidence intervals should be used.

The top half of Table 3 gives the estimates of the average relative specificities and their errors for several different possible prevalence values. It also gives the estimates of the difference or bias in these ratios that we presented in Section 3. The bottom half of the table gives the same information for relative AUC.

Estimates of the prevalence of breast cancer in the U.S. population differ significantly. The Digital Mammographic Imaging Screening Trial [15] gives an estimate as high as 0.0079. Using Breast Cancer Surveillance Consortium data [16] we estimated a value of prevalence around 0.005. The detection rate for DM+DBT in the Rose et al. study was 0.00537, so our working estimate of prevalence must be greater than that. For a reasonable range of prevalence we use 0.0055 to 0.0080 to calculate our estimated performance ratios. In addition we include a substantial overestimate of 0.0110.

Table 2 This table gives the relative sensitivity (rSens, eq. (2)), false positive rate (rFPR, eq. (3)), and expected utility (rEU, eq. (8)) between digital mammography with tomosynthesis and digital mammography alone using data from Rose et al. [12]. The estimated standard errors for each ratio follow the values.

rSens	Std. Err.	rFPR	Std. Err.	rEU	Std. Err.
1.33	0.27	0.59	0.06	1.44	0.33

Table 3 This table gives the ratios of statistics for digital mammography with tomosynthesis to digital mammography alone for various assumed values of prevalence π_p using data from Rose et al. [12]. The ratios are specificities (rSpec, 11]), areas under a power-law ROC curve (rAUC$_P$, 18]), and areas under an equal-variance bi-normal ROC curve (rAUC$_N$). The estimated standard errors of these ratios are in the third and seventh columns. The fifth column gives the actual differences between the estimated values and the center value estimated at $\pi_p = 0.008$. The fourth column gives estimates of these differences using the bias expressions (13) and (19).

Prev.		Standard	Difference	Measured		
π_p	rSpec	error	estimate	difference		
0.0055	1.03728	0.0060	-0.00010	-0.00010		
0.0080	1.03739	0.0060	0.0	0.0		
0.0110	1.03751	0.0060	0.00012	0.00012		
	rAUC$_P$				rAUC$_N$	Std. Err.
0.0055	1.102	0.08	-0.0119	-0.0088	1.08	0.04
0.0080	1.111	0.07	0.0	0.0	1.11	0.05
0.0110	1.117	0.07	0.0067	0.0064	1.14	0.05

In Table 3 we see that the rSpec and rAUC values depend weakly upon the value of π. For all possible values of prevalence, the relative specificity is almost constant. Over the 100% change in prevalence in the table, rAUC$_P$ changes by only 1.5%. Within the reasonable range of prevalence (0.0055–0.008) the change in all ratios is less than the estimated standard error of the ratios. Therefore we can make estimates of the relative increase in specificity or AUC values with useful accuracy using only approximate estimates of the prevalence of disease.

In Table 3 we also see that our approximate bias expressions 13 and 19 for rSpec and rAUC$_P$ give reasonably good estimates of the difference between two relative estimates with different prevalences. Therefore we can be confident that the conditions for the insensitivity of the ratios to the prevalence are those conditions that make those bias expressions small.

Table 3 also shows that the relative AUC values are very similar for the power-law ROC model and the binormal ROC model. The differences between these models are less than the estimated standard errors. Though eqs (16) and (17) for estimating AUC appear very different, our results are not sensitive to the choice of equation.

Using a posited prevalence to estimate the *absolute* values of some statistics would lead to much larger errors than using that prevalence to estimate their ratios. Obviously over a 100% change in prevalence, an absolute sensitivity estimate (6)) would change by 100%, whereas its ratio is independent of prevalence and does not change. For the Rose et al. data set the value of AUC$_P$ (16)) changes by about 25% with our 100% prevalence change.

These estimators of rSpec and rAUC can be used on populations with significantly higher prevalences than the prevalence of breast cancer in screening in this example. Indeed π_p could be 10%, if that is much less than the true negative rate and the positive rates of the two tests are similar. Because the dependence on π_p can be quite complicated, we recommend that users test how sensitive their relative estimates are to values and errors of the posited prevalence.

5 Conclusions

Like relative sensitivities and false positive rates [3], we have shown that we can calculate an estimate of the ratio or percent increase between two expected utilities for two screening tests from the same population, even if negative-test patients in the study do not undergo disease confirmation with a reference procedure, and the true and false negative rates are unknown. Using resampling or other methods we can perform uncertainty estimation and inference on this ratio.

Additionally the ratios of some other statistics, such as specificity and AUC, can be written in terms of the number of true and false positives and an approximately known disease prevalence. We showed that if the approximate prevalence and its uncertainty are small, and if the performance of the two tests are not substantially different, then those ratios are frequently insensitive to the value of that prevalence in screening studies. Therefore we can use approximate values of prevalence known from published literature to obtain estimates of percent increases in specificity or AUC with little loss of accuracy for many purposes such as meta-analyses.

We have found that the requirements for these ratios being insensitive to the prevalence occur frequently in clinical screening studies. However, before publishing results of estimates of relative specificity and AUC using this technique, we recommend testing how sensitive your estimates are to your assumed prevalence and ROC model, as we did in Table 3. If you find that reasonably modifying the assumed prevalence leads to changes in the inference on the relative statistic, then these approximate methods may not be appropriate for your data set, or your finding is probably not significant.

Acknowledgement

This work was supported in part by NIH Grant R21 EB018939.

Appendix

In this appendix we show how we calculated expression (19), the approximate bias in the ratio of the power-law AUC estimate, rAUC$_P$.

In 16) we do not know the true population value of disease prevalence π, so we use our posited value $\pi_p = R\pi$ in place of it. Therefore our estimate of AUC$_P$ using our posited prevalence will be

$$\mathrm{AUC}_P^* = \frac{\log \frac{f}{N} - \log(1-R\pi)}{\log \frac{t}{N\pi} + \log \frac{f}{N} - \log(1-R\pi) - \log R} \tag{20}$$

If π is not large and R is not much different from 1, then $\log(1-R\pi)$ and $\log(1-\pi)$ will be small and approximately equal, and eq. (20) differs from eq. (16) primarily by a factor of $\log R$. We can write

$$\mathrm{AUC}_P^* \approx \frac{\log \mathrm{FPR}}{\log \mathrm{TPR} + \log \mathrm{FPR} - \log R} = \frac{\log \mathrm{FPR}}{w - \log R}$$

where we define $w = \log \mathrm{TPR} + \log \mathrm{FPR}$.

If we define

$$S = \frac{\log \mathrm{FPR}_X}{\log \mathrm{FPR}_Y}$$

then we can write the ratio of two biased AUC estimates as

$$\begin{aligned} \mathrm{rAUC}_P^* &\approx S\frac{w_Y - \log R}{w_X - \log R} \approx S\frac{w_Y - \log R}{w_X}\left(1 + \frac{\log R}{w_X}\right) \\ &\approx S\frac{w_Y}{w_X} + S\frac{w_Y - w_X}{w_X^2}\log R = \mathrm{rAUC}_P + S\frac{w_Y - w_X}{w_X^2}\log R \end{aligned}$$

Here we have assumed that $\log R$ is smaller than w_X. The first term is the ratio of two AUC estimates calculated with the population prevalence which is eq. (18). The term on the right is expression (19) and is the very approximate bias in our estimate of that ratio if our posited prevalence is incorrect by a factor of R.

Notes

[1] In this work most variables represent measured values or estimated values. The exceptions are the values of π and p, which may be considered true population values.

References

[1] Friedewald SM, Rafferty EA, Rose SL, Durand MA, Plecha DM, Greenberg JS, et al. Breast cancer screening using tomosynthesis in combination with digital mammography. J Am Med Assoc. 2014;311:2499–2507.

[2] Brem RF, Tabár L, Duffy SW, Inciardi MF, Guingrich JA, Hashimoto BE, et al. Assessing improvement in detection of breast cancer with three-dimensional automated breast ultrasound in women with dense breast tissue: The SomoInsight study. Radiologist. 2015;274:663–73.

[3] Schatzkin A., Connor R. J., Taylor P. R., Bunnag B.. Comparing new and old screening tests when a confirmatory procedure cannot be performed on all screens. American Journal of Epidemiology. 1987;125:672–678.

[4] Cheng H, Macaluso M.. Comparison of the accuracy of two tests with a confirmatory procedure limited to positive results. Epidemiology. 1997;8:104–6.

[5] Pepe MS, Alonzo TA.. Comparing disease screening tests when true disease status is ascertained only for screen positives. Biostatistics. 2001;2:249–60.

[6] Abbey CK, Eckstein MP, Boone JM.. Estimating the relative utility of screening mammography. Med. Decision Making 2013;33.

[7] Abbey CK, Eckstein MP, Boone JM.. An equivalent relative utility metric for evaluating screening mammography. Medical Decision Making. 2010;30:113–22.

[8] Baker SG, Pinsky PF.. A proposed design and analysis for comparing digital and analog mammography: special receiver operating characteristic methods for cancer screening. J Am Stat Assoc. 2001;96:421–8.

[9] Shaw PA, Pepe MS, Alonzo TA, Etzioni R.. Methods for assessing improvement in specificity when a biomarker is combined with a standard screening test. Stat Biopharma Res. 2009;1:18–25.

[10] Egan JP. Signal detection theory and ROC analysis. New York: Academic Press, 1975.

[11] Samuelson FW, He X.. A comparison of semi-parametricROC models on observer data. SPIE J Med Imaging. 2014;1:031004.

[12] Rose SL, Tidwell AL, Bujnoch LJ, Kushwaha AC, Nordmann AS, Sexton R.. Implementation of breast tomosynthesis in a routine screening practice: an observational study. Am J Roentgenol. 2013;200:1401–8.

[13] Benjamini Y, Yekutieli D.. False discovery rate-adjusted multiple confidence intervals for selected parameters. J Am Stat Assoc. 2005;100:71–81.

[14] Davison AC, Hinkley DV.. Bootstrap methods and their applications. Cambridge: Cambridge University Press, 1997.

[15] Pisano ED, Gatsonis C, Hendrick E, Yaffe M, Baum JK, Acharyya S, et al. Diagnostic performance of digital versus film mammography for breast-cancer screening. N Engl J Med. 2005;353:1773–83.

[16] Rosenberg RD, Yankaskas BC, Abraham LA, Sickles EA, Lehman CD, Geller BM, et al. Performance benchmarks for screening mammography. Radiology. 2006;241:55–66.