Published in final edited form as:
Med Decis Making. 2010 ; 30(1): 113–122. doi:10.1177/0272989X09341753.

An equivalent relative utility metric for evaluating screening mammography

Craig K. Abbey[1,2], **Miguel P. Eckstein**[1], and **John M. Boone**[2,3]
[1]Dept. of Psychology, University of California, Santa Barbara, CA.
[2]Dept. of Biomedical Engineering, University of California, Davis, CA.
[3]Dept. of Radiology, UC Davis Medical Center, Sacramento, CA

Abstract

Comparative studies of performance in screening mammography are often ambiguous. A new method will frequently show a higher sensitivity or detection rate than an existing standard with a concomitant increase in false positives or recalls. We propose an equivalent relative utility (ERU) metric based on signal detection theory to quantify screening performance in such comparisons. The metric is defined as the relative utility, as defined in classical signal detection theory, needed to make two systems equivalent. ERU avoids the problem of requiring a predefined putative relative utility, which has limited application of utility theory in ROC analysis. The metric can be readily estimated from recall and detection rates commonly reported in comparative clinical studies. An important practical advantage of ERU is that in prevalence matched populations, the measure can be estimated without an independent estimate of disease prevalence. Thus estimating ERU does not require a study with long term follow up to find cases of missed disease. The approach is applicable to any comparative screening study that reports results in terms of recall and detection rates, although we focus exclusively on screening mammography in this work. We derive the ERU from the definition of utility given in classical treatments of signal detection theory. We also investigate reasonable values of relative utility in screening mammography for use in interpreting ERU using data from a large clinical study. As examples of application of ERU, we re-analyze two recently published reports using recall and detection rates in screening mammography.

1. Introduction

Breast cancer screening involves a high volume of examinations of asymptomatic women for disease with low prevalence in this population. While screening mammography is now generally established as beneficial [1–3], the exam has nontrivial false-positive and false negative rates. This has lead to substantial efforts to improve screening mammography through a variety of approaches. Large scale studies evaluating new methods in screening mammography typically report endpoints of recall and detection rates, and/or sensitivity and false-positive rate. Because of the low prevalence of disease, accurate estimation of these summary statistics in the screening arena requires large samples with many thousands of patients.

A more fundamental problem is that results of comparative studies in screening mammography are often ambiguous. An improvement in detection rate or sensitivity often comes with concomitant increase in recall or false-positive rate. In principle, there is a rigorous and well known solution to the question of defining optimal performance. According to classical signal detection theory, the optimal system maximizes the expected utility of the decisions [4]. When screening mammography is considered as a binary decision (recall or no recall), the expected utility is based on the frequency of the four decision outcomes (true positive, false positive,

true negative, and false negative) weighted by the utility of each outcome. Utility theory and its relation to receiver operating characteristic (ROC) analysis is well documented [4–7], and is generally used to theoretically identify the optimal operating point on an ROC curve. Some approaches based on utility theory have been developed to analyze ROC data [8–10]. However, utility theory is rarely used in practical settings because there is little consensus on what the weighting of different decision outcomes should be [11,12].

Here we present a method to evaluate screening performance based on the notion of equivalent relative utility (ERU). The approach is intended for large clinical population studies where typical endpoints are recall and detection rates or sensitivity and specificity. Surprisingly, when estimated from recall and detection rates, the ERU does not require an estimate of disease prevalence. Essentially, prevalence is already factored into the recall and detection rates appropriately. Disease prevalence can be difficult to measure in a clinical population because it requires counting all patients that had disease at the time of screening, not just those that could be detected by the screening procedure. This requires tracking the patient population for at least one or two years after the study is completed. Not requiring a separate estimate of prevalence allows our approach to avoid the difficult and time-consuming problem of long-term follow-up to find cases of missed disease. The measure can be computed as soon as recall and detection rates for two or more screening methods have been obtained.

We derive the method below, along with a Bayesian approach to performing inference on the results. We then turn to previously published studies to better understand interpretation of the ERU in the context of screening mammography.

2. Method

In this section we will show how measurements of recall and detection rates for two screening methods can be used to determine the ERU– the relative utility of correct and incorrect decisions that is needed to make the two methods have an equal decision-theoretic utility. We begin with a brief review of utility analysis for binary decision processes, then define the ERU measure, and show how it can be estimated from recall and cancer detection rates in matched populations without a separate estimate of disease prevalence.

2.1 Utility analysis of binary decisions

When screening mammography is regarded as a binary decision – with exam results dichotomized into categories of follow-up or no follow-up – utility is determined by probabilities of the four possible outcomes. These are true positives (TP), where patients with disease are assigned to follow-up; false positives (FP), where patients who do not have disease are assigned to follow-up; true negatives (TN), where patients without disease are not assigned to follow-up; and false negatives (FN) where patients with disease are not assigned to follow-up. The basis for defining utility in binary decisions such as this is to determine a utility value for each of the outcomes, and then compute the total expected utility

$$U = U_{\mathrm{TP}} P(\mathrm{TP}) + U_{FN} P(\mathrm{FN}) + U_{\mathrm{TN}} P(\mathrm{TN}) + U_{\mathrm{FP}} P(\mathrm{FP}), \quad (1)$$

where $P(\cdots)$ indicates the probability of the outcome.

The various outcome probabilities can be decomposed into the TP and FP rates as well as the prevalence of the disease in the population, π. The true positive rate, R_{TP}, is defined as the conditional probability of a positive finding given that disease is present, and the false positive rate, R_{FP}, is the conditional probability of a positive finding given that disease is absent. These terms can be used to rewrite each of the outcome probabilities. Specifically, we see that P(TP)

$= R_{TP}\pi$, $P(FN) = (1 - R_{TP})\pi$, $P(TN) = (1 - R_{FP})(1 - \pi)$, and $P(FP) = R_{FP}(1 - \pi)$. This reparameterization makes explicit the connection to ROC analysis where R_{TP} is plotted as a function of R_{FP} for a diagnostic test.

Rearrangement of terms in Equation 1 with substitution of terms involving R_{TP}, R_{FP}, and π results in the following iso-utility equation,

$$R_{TP} = \frac{(U_{TN} - U_{FP})(1 - \pi)}{(U_{TP} - U_{FN})\pi} R_{FP} + \frac{U - U_{FN}\pi - U_{TN}(1 - \pi)}{(U_{TP} - U_{FN})\pi}. \quad (2)$$

The important feature of Equation 2 is that for a fixed value of U, iso-utility curves in R_{TP} and R_{FP} come in the form of a line with positive slope (under the reasonable assumption that correct decisions have greater utility than incorrect ones). This means that every pair, (R_{TP},R_{FP}), which satisfies Equation 2 for a given U has equal utility. As we shall see, the slope of this line will play an important role in defining the ERU measure.

2.2 Relative utility and equivalent relative utility

We will follow Wagner *et al*. [12] in defining the relative utility as the difference between correct and incorrect decisions when the patient has disease divided by this difference when the patient is not diseased,

$$U_{Rel} = \frac{(U_{TP} - U_{FN})}{(U_{TN} - U_{FP})}. \quad (3)$$

Note that the numerator and denominator in Equation 3 could be switched and the result would still provide a reasonable definition of relative utility. However, we believe Equation 3 is more intuitive for screening applications, since the utility of correctly diagnosing patients with disease is generally considered to be higher than correctly diagnosing normal patients, and thus relative utility should be large ($U_{Rel} \gg 1$). Using this definition of relative utility, the iso-utility line in Equation 3 is given by

$$R_{TP} = \frac{Q_\pi}{U_{Rel}} R_{FP} + \text{const.}, \quad (4)$$

where Q_π is an odds ratio based on disease prevalence, $Q_\pi = (1 - \pi) / \pi$. At typical estimates of prevalence in breast cancer screening ($\pi \approx 0.5\%$), this ratio is roughly 200.

Figure 1 summarizes the standard relationship between utility and the operating point of a diagnostic test [6]. The ROC curve specifies a set of possible operating points, with the utility of each point governed by Equation 2. Traditionally, utility has been used to derive the optimal operating point of the ROC curve, which is seen in Figure 1 to be a point on the ROC curve which is tangent to the iso-utility lines. The slope of the iso-utility line – and hence the tangent point on the ROC curve – is highly dependent on the relative utility. This has been considered a limitation of utility analysis, since there is no universally agreed upon value for this quantity [11]. The ERU metric we propose essentially uses these same concepts for the purpose of comparing two screening systems in a way that does not require an *a-priori* established relative utility.

Thus far we have considered the process adopting a given utility structure and have seen the consequences in terms of the true positive and false positive rates that have equal utility. Now

we reverse the situation and start with the operating points of two diagnostic tests and then derive the relative utility that makes them both fall on an iso-utility line. Let us imagine a situation where we sought to compare two well characterized diagnostic tests. Test 1 has an operating point ($R_{TP,1}$,$R_{FP,1}$) and Test 2 has operating point ($R_{TP,2}$,$R_{FP,2}$). The slope of the line connecting these two operating points is given by

$$\text{Slope(test1,test2)}=\frac{(R_{TP,1}-R_{TP,2})}{(R_{FP,1}-R_{FP,2})}. \tag{5}$$

We define the ERU between Test 1 and Test 2 as the relative utility needed for Test 1 and Test 2 to lie on an iso-utility line. Therefore the ERU can be found by equating the slopes in Equation 4 and Equation 5, and solving for the relative utility,

$$\text{ERU}=\frac{Q_\pi}{\text{Slope(test1,test2)}}. \tag{6}$$

2.3 Interpretation of ERU

We propose ERU as a measure for comparing screening methods. We therefore need to be able to interpret the result and say one method is better, assuming statistical significance is achieved. As a first step, we consider the case when one method increases the true-positive rate and reduces the false-positive rate. In this case one test is clearly superior to the other. We also note that a test with a lower recall rate and simultaneously higher detection rate will always have higher sensitivity and a lower false-positive rate. In this case, the ERU will be negative because the slope between the two operating points is negative. Thus the interpretation of a negative ERU is that one method is superior, and it would take a fundamentally flawed utility structure to make them equivalent. The superior method should be readily apparent from sensitivity/specificity data or alternatively recall/detection rate data.

Now let us assume without loss of generality that test 1 has the lower false positive rate and a lower true positive rate as well, as shown in Figure 2. We can say that for a putative relative utility greater than the ERU, test 2 is superior since it would reside on a better iso-utility line. Conversely, for a putative relative utility less than the ERU, test 1 is superior since it would then reside on a better iso-cost line. While the ultimate judgment of the systems still requires definition of the appropriate relative utility for interpretation, the ERU itself can be readily estimated without it. We will discuss this issue further in Section 3.

2.4 Determination of ERU from recall and detection rates

We have shown how ERU can be determined from true-positive and false-positive rates. However, as mentioned in the introduction, it is often easier to acquire recall and cancer detection rates in practical studies. In this section we describe how these measures can be used to find the ERU. As we shall see, a surprising result of using recall and detection rates to determine the ERU is that explicit reference to disease prevalence cancels, and thus a separate estimate of disease prevalence for the population is not required.

The cancer detection rate, R_D, is simply the probability of a true-positive outcome, and therefore it is related to the true positive rate by

$$R_D=\pi R_{TP}. \tag{7}$$

The recall rate, R_R, is the rate of true-positive and false positive outcomes, and is therefore related to the true-positive and false-positive rates by

$$R_R = \pi R_{TP} + (1 - \pi) R_{FP}. \tag{8}$$

From the recall and detection rates we can solve for the true-positive and false-positive rates by

$$R_{TP} = \frac{R_D}{\pi}, \text{ and } R_{FP} = \frac{R_R - R_D}{(1 - \pi)}. \tag{9}$$

Let us now assume a situation similar to Equation 5, except that now we have cancer-detection and recall rate data for two tests instead of true-positive and false-positive rate data. Let $(R_{D,1}, R_{R,1})$ be the detection rate and recall rate for test 1 and $(R_{D,2}, R_{R,2})$ be the corresponding measures for test 2. Using Equation 9 to determine $(R_{TP,1}, R_{FP,1})$ and $(R_{TP,2}, R_{FP,2})$, and then using these to determine the slope in Equation 5 yields

$$\text{Slope(test1,test2)} = \frac{(R_{D,1} - R_{D,2})/\pi}{((R_{R,1} - R_{D,1}) - (R_{R,2} - R_{D,2}))/(1 - \pi)}. \tag{10}$$

Substituting this into Equation 6 specifies the ERU between test 1 and test 2 as

$$\text{ERU} = \frac{(R_{R,1} - R_{D,1}) - (R_{R,2} - R_{D,2})}{(R_{D,1} - R_{D,2})}, \tag{11}$$

Equation 11 shows that the ERU is defined by the difference in the rate of false positive recalls $(R_R - R_D)$ divided by the difference in the rate of detected cancers. Cancelling terms yields a final form in terms of the difference between recall and detection rates,

$$\text{ERU} = \frac{R_{R,1} - R_{R,2}}{R_{D,1} - R_{D,2}} - 1. \tag{12}$$

Note that in Equation 11 and Equation 12, all prevalence terms have canceled and so there is no need to know π explicitly in order to determine the ERU. Viewed another way, the prevalence dependence of ERU is built implicitly into the recall and cancer-detection rates, and thus does not require separate measurement.

2.5 Estimation of ERU from measured detection and recall rates

Estimation of the ERU consists of replacing recall and detection rates in Equation 12 with sample estimates $\hat{R}_D$ and $\hat{R}_R$. Let N_1 be the total sample size (i.e. the total number of patients evaluated by method 1), and let $N_{R,1}$ be the number of these patients recalled for follow-up and $N_{D,1}$ be the number of patients with detected cancer. The estimated recall and detection rates are determined from the sample proportions

$$\widehat{R}_{R,1} = \frac{N_{R,1}}{N_1}, \text{ and } \widehat{R}_{D,1} = \frac{N_{D,1}}{N_1}. \tag{13}$$

An analogous procedure is used to produce $\hat{R}_{D,2}$, and $\hat{R}_{R,2}$. These estimates of recall and detection rates can be used in Equation 12 to produce an estimate of the ERU.

However, as we shall see, ERU is difficult to estimate precisely, and hence it is probably more useful to describe it in terms of confidence intervals than a point estimate. In Appendix 1 we describe a posterior sampling method for computing Bayesian confidence intervals on ERU estimates computed from observed proportions as in Equation 13. For relative utilities within the confidence interval, the data is indeterminate for which system is optimal.

2.6 Limitations of the approach

As a final step in presenting the general methodology for evaluating and interpreting the ERU, we review some important limitations of the approach. The first is the critical assumption of equal disease prevalence in the two (or more) cohorts being evaluated. This issue is endemic to comparisons of recall and detection rates in general since these are dependent on disease prevalence. As an example, consider two hypothetical tests with identical true-positive and false-positive rates of 70% and 5% respectively. Now assume that test 1 is evaluated in a cohort with a disease prevalence of 5/1000, and test 2 is evaluated in a cohort with a prevalence of 7/1000. The recall and detection rates will be (5.33%,3.5/1000) for test 1 and (5.46%,4.6/1000) for test 2. This results in a negative ERU suggesting that test 2 is superior. Thus differences in underlying prevalence can bias the ERU. The ERU measure, as defined here, is only appropriate for comparison in cohorts that have been selected in a way that does not lead to a systematic mismatch in prevalence.

A second important issue is to recognize that the analysis here is based on decision utility, which is not equivalent to a cost/benefit analysis. A screening modality that has a favorable ERU may still be prohibitively costly.

A third potential limitation of the analysis is that it strongly links the screening modality with an operating point. A suboptimal operating point used in one modality may result in a poor ERU in the comparison. This may be the consequence of an unfamiliar new method leading to the adoption of an overly strict or lax decision criterion. For criterion-free analysis, ROC type studies are more appropriate. The ERU is most appropriate for use analyzing clinical performance when the operating point as well as the technology or methodology is under evaluation.

3. Retrospective analysis of published data

In this section we reanalyze previously published data from three studies to better understand and interpret the ERU in clinical data sets related to screening mammography. The large retrospective study of screening mammography by Barlow and colleagues [13] is used to get a rough estimate of the operating relative utility, and we compare this to the rationale used by Wagner *et al*. [12] in arriving at a putative relative utility of 150. We then consider two case studies for application of the ERU measure. The first is a recent study of Hambly *et al*. [14], comparing full-field digital mammography (FFDM) to standard screen-film mammography on a large clinical population in Ireland. This case study closely fits the two-modality comparison paradigm we have used above to develop the approach. The second is a comparison of recall and detection rates in screening mammography amongst a group of practicing radiologists by Gur *et al*. [15]. This serves as an example of how the ERU concept can be adapted to analyze other types of data.

3.1 Determining reasonable values of ERU

As mentioned previously, the ERU measure does not specify what the appropriate relative utility should be, but interpreting the ERU does. In this section we devote some effort to getting

a rough sense of what an appropriate relative utility for screening mammography should be. Wagner and colleagues [12] have addressed this issue briefly for the purpose of showing how reader skill can influence the operating point on an ROC curve. They argue that for mammography to have any advantage over simply not screening at all, the relative utility must be at least 50. Positing a definite benefit to the exam, they settle on a lower limit of 150. Wagner *et al*. also argue for an upper limit based on exam cost compared to the value in quality-adjusted life years and arrive at a value of 500.

To investigate this issue in a direct data-driven fashion, we analyzed screening mammography data from a large recently published study by Barlow and colleagues [13]. The study investigated screening performance on 469,512 patients from three breast cancer registries participating in the Breast Cancer Surveillance Consortium. The purpose of the study was to investigate the sources of variability amongst radiologists, but for the purposes here we will only consider the aggregate data. They considered seven possible outcomes from a screening mammography exam. These essentially followed the BI-RADS interpretation codes (0 to 5) with code 3 split into two groups (3A and 3B) depending on whether immediate work-up was recommended. The codes were ordered by the increasing likelihood of cancer (1, 2, 3A, 3B, 0, 4, 5), which was determined by biopsy or a one-year follow-up interval.

In Figure 3A we show the results of fitting binormal ROC curves to the aggregate Barlow *et al*. data using well tested software (ROCKIT) for categorical rating data [16]. A global fit is obtained by fitting to all seven possible outcome categories. A close inspection of Figure 3A suggests that the binormal fit appears to slightly underestimate sensitivity in the neighborhood of a 10% false positive rate. We find this to be due to fitting the two very low false-positive rate categories corresponding to BI-RADS Codes 4 and 5, which are rare and considered by many to be outside the purview of a screening exam [17]. To reduce the influence of these points, we also fit a binormal ROC curve by combining the 0, 4, and 5 codes into a single category leaving a total of 5 categories. We refer to this as the local fit, and Figure 3A shows that it does fit the observed data near the 10% false positive rate more closely.

Figure 3B shows the results of our utility analysis on the Barlow *et al*. data. The relative utility necessary for a given operating point to be optimal is found by dividing the prevalence odds ratio, Q_π, by the slope of the ROC curve, similar to the computation in Equation 6 and depiction in Figure 1. In these data, the prevalence is 0.512%, resulting in $Q_\pi = 194.5$. The relative utilities for the global and local fits are somewhat divergent over the range of false-positive rates tested. At a false positive rate of 10%, the global fit suggests a relative utility of 117, whereas the local fit is as high as high as 222. The large difference between the two estimates shows the strong dependency on the shape of the ROC curve. Nonetheless, the two estimates appear to bracket the lower bound of 150 as suggested by Wagner *et al*., and thus provide some additional quantitative data to support their rationale.

3.2 Case Study 1: Assessment of full-field digital mammography compared to conventional screen film mammography

A recently reported study by Hambly *et al*. [14] serves as an example of how ERU can be used in the comparison of two different screening methods. Their study compared recall and detection rates of standard screen-film mammography (SFM) to full field digital mammography (FFDM) in the Irish National Breast Screening Program from 2005 to 2007. A total of 163,031 patients were evaluated for the study with 26,593 (16%) screened with FFDM and 136,438 (84%) screened with standard SFM. Within each screening method, mammograms were partitioned into two groups depending on whether it was the patient's first screening exam or a subsequent exam. Disease prevalence was generally higher in initial screening exams, which comprised a slightly larger proportion of the SFM cases compared to FFDM cases (30.6% and 27.6% respectively). Stratifying the data into these two groups helped match

prevalence across comparisons. The recall and detection rates for each group reported by Hambly *et al*. are given in Table 1. In both groups, FFDM raised the cancer detection rate along with a simultaneous increase in the recall rate. Thus arguing that FFDM offers a benefit over SFM requires some justification that the additional detections are worth the additional recalls.

Table 1 also gives the results of the ERU analysis. For the initial scan patients, the ERU estimated from Equation 12 is 78.9, but it is important to realize that this estimate has low precision. A 90% confidence interval computed using the methods in Appendix 1 (with 100,000 Monte-Carlo samples) ranged from 6.6 to over 1000, and the posterior probability that the ERU exceeds the nominal value of 150 is 0.48. Hence the ERU assessment is equivocal for the initial scan data using the nominal relative utility of 150. In the subsequent exam group, the ERU is more definitive. Here the estimated ERU is 4.1 with a 90% confidence interval ranging from 1.8 to 28.8. The posterior probability that the ERU is greater than 150 is 0.03, suggesting a significant benefit to screening mammography from FFDM.

3.3 Case Study 2: Comparisons of recall and detection rates between individual radiologists

A study by Gur and colleagues [15] evaluated the recall and detection rates of 10 radiologists at a university medical center. They find that the relationship between recall and detection across radiologists is reasonably well fit by a regression line indicating a significant positive association between recall and detection rates ($\rho = 0.76$, $p < 0.01$). They argue that programs to reduce recall rates may produce a similar reduction in detection rates. This argument assumes that the linear model determined from a sample of radiologists will also apply to an individual who has been induced to adopt a lower recall rate. However, we can also frame the issue as trying to determine whether the radiologists with higher recall rates are giving better or worse performance than those with lower recall rates. In this case, the fitted line serves as an aggregate measure of the tradeoff, and permits an evaluation of ERU.

Figure 4 plots the Gur *et al*. data along with the best fit line from their publication. When a line describes the relationship between recall and detection rates (i.e. $R_D = aR_R + b$), then Equation 12 shows that any two points on the line have an ERU value directly related to the inverse of the slope,

$$\mathrm{ERU} = \frac{1}{a} - 1.. \tag{14}$$

Therefore the slope of the fitted line is equivalent to a particular value of the ERU, which characterizes the tradeoff between recall and detection in an aggregate sense. The ERU determined from the slope of the line fitted to the Gur *et al*. data is 46.3, which is well below the putative relative utility derived from the Barlow data. For a relative utility of 150, the slope of the recall/detection line is 0.066, which falls just outside the 95% confidence interval on the observed slope (0.068 to 0.378). Thus the slope predicted by a relative utility of 150 is not supported by the Gur *et al*. data. A line of this slope with a least-squares fitted intercept is plotted on Figure 4 for comparison. Generally, radiologists with lower recall and detection rates operate at a correspondingly lower overall utility under a putative relative utility of 150. A notable exception to this is the radiologist with the lowest recall rate, who operates at almost the same total utility as the radiologist with the highest recall rate.

These findings reinforce the conclusions of Gur *et al*. who argue that programs and policies that cause radiologists to reduce their recall rate may be reducing detection rates at an unacceptable rate. To the extent that the model they find for a sample of radiologists can be used to model the effect of such policies on an individual, lowering recall rates lowers the overall utility of the exam for the putative relative utility derived from the Barlow data. We

contend that utility based arguments such as this should be part of the careful evaluation of such policies that Gur *et al*. suggest.

4. Summary and Conclusions

We have introduced equivalent relative utility as a measure of diagnostic performance for comparing two or more screening methods. The approach is grounded in signal-detection theory as it is used in classical ROC analysis, where relative utility specifies the tradeoff between errors and is used to define the optimal operating point on an ROC curve. However, for the purpose of comparing two screening methods, the ERU measure gives the relative utility needed for the two systems to be considered equal. In this way the approach avoids having to decide *a-priori* what constitutes an acceptable relative utility, which has limited the use of utility in ROC studies. In comparisons of two screening methods, investigators can simply report the ERU and let the community decide whether one method clearly constitutes an improvement.

We show that the ERU is readily computed from commonly reported recall and detection rates, and we describe a technical but relatively simple posterior sampling approach towards determining a Bayesian confidence interval on the estimate. Surprisingly, when estimated in this way, the ERU does not require a separate estimate of disease prevalence within the target population. This has important practical consequences in a clinical setting because it means that the measure can be computed after the last recalled patient has been evaluated. Long term follow-up to find cases of missed disease is not necessary.

Three retrospective analyses of published data sets have been presented as a way to better understand and interpret ERU in the context of screening mammography. The large screening evaluation described by Barlow et al. [13] is used to get a rough idea of the relative utility being used implicitly at clinical operating points. We find this to be consistent with a value of 150, which is the lower limit suggested by Wagner *et al*. [12] on the basis of a general assessment of mammographic effectiveness. A recently completed comparison of screen-film and full-field digital mammography in Ireland by Hambly *et al*. [14] is used to demonstrate the ERU approach on a clinical dataset. In that study, digital mammography was found to increase the detection rate in both first-time and subsequent exams with a simultaneous increase in recall rate as well. Because of the smaller sample size, the ERU was indeterminate for the first-time exams. For the subsequent-exam data, ERU was found to be approximately 4.1 with a 90% confidence interval ranging from 1.8 to 28.8. At the putative relative utility implied from the Barlow study, FFDM would be considered to have significantly greater utility than the screen film standard. In the Gur *et al*. study of individual radiologists [15], the line they fit to recall and detection rate data implies an ERU of 46 with the putative relative utility derived from the Barlow study predicting a significantly lower slope. Therefore, by this criterion radiologists with higher recall rates are generally operating at higher utility.

In both of the comparative studies, the ERU supports the conclusions of the original authors using the putative relative utility of 150. However, the ERU itself is noncommittal, and these conclusions could change under a different relative utility. Thus the ERU measure allows the field to gauge performance as it sees fit while remaining rooted in rigorous signal detection theory. Therefore, we believe that it is an important summary metric for clinical comparisons of screening mammography systems, and can be readily applied to other screening procedures as well.

Acknowledgments

This work was supported in part by grants R01-EB002138 and R01-CA118294 from the NIH. The authors are grateful to Dr. Sheng Zheng for assistance with the ROCKIT code. In particular, we gratefully acknowledge Prof. David Gur for supplying the original data used in his publication [15].

References

1. Humphrey LL, Helfand M, Chan BK, Woolf SH. Breast cancer screening: a summary of the evidence for the U.S. Preventive Services Task Force. Ann Intern Med 2002;137(5 Part 1):347–360. [PubMed: 12204020]
2. Fletcher SW, Elmore JG. Clinical practice: Mammographic screening for breast cancer. N Engl J Med 2003;348(17):1672–1680. [PubMed: 12711743]
3. Berry DA, Cronin KA, Plevritis SK, Fryback DG, Clarke L, Zelen M, Mandelblatt JS, et al. Effect of screening and adjuvant therapy on mortality from breast cancer. N Engl J Med 2005;353(17):1784–1792. [PubMed: 16251534]
4. Green, DM.; Swets, JA. Signal Detection Theory and Psychophysics. New York: John Wiley & Sons, Inc; 1966.
5. Swets, JA.; Pickett, RM. Evaluation of Diagnostic Systems: Methods from Signal Detection Theory. New York: Academic; 1982.
6. Metz CE. Basic principles of ROC analysis. Semin. Nucl. Med 1978;8:283–298. [PubMed: 112681]
7. Somoza E, Soutullo-Esperon L, Mossman D. Evaluation and optimization of diagnostic tests using receiver operating characteristic analysis and information theory. Int. J. Biomed. Comput 1989;24:153. [PubMed: 2509379]
8. Halpern EJ, Alpert M, Krieger AM, Metz CE, Maidment AD. Comparisons of ROC curves on the basis of optimal operating points. Acad Radiol 1996;3:245–253. [PubMed: 8796672]
9. Edwards, DC.; Metz, CE. A utility-based performance metric for ROC analysis of N-class classification tasks. In: Jiang, Y.; Sahiner, B., editors. Proceedings SPIE; 2007. p. 1-10.
10. Edwards, DC.; Metz, CE. Optimality of a utility-based performance metric for ROC analysis. In: Sahiner, B.; Manning, D., editors. Proceedings SPIE; 2008. p. 1-10.
11. Metz CE. ROC analysis in medical imaging: a tutorial review of the literature. Radiol. Phys. Technol 2008;1:2–12.
12. Wagner RF, Beam CA, Beiden SV. Reader Variability in Mammography and Its Implications for Expected Utility over the Population of Readers and Cases. Med. Decis. Making 2004;24:561–572. [PubMed: 15534338]
13. Barlow WE, Chi C, Carney PA, Taplin SH, D'Orsi C, Cutter G, et al. Accuracy of Screening Mammography Interpretation by Characteristics of Radiologists. J. Natl. Cancer Inst 2004;96(24):1840–1850. [PubMed: 15601640]
14. Hambly N, Phelan N, Hargaden G, O'Doherty A, Flanagan F. Krupinski EA. Impact of Digital Mammography in Breast Cancer Screening: Initial Experience in a National Breast Screening Program. Intl. Workshop in Digital Mmammography 2008;5116:55–60. LCNS 1008.
15. Gur D, Sumkin JH, Hardesty LA, Clearfield RJ, Cohen CS, Ganott MA, et al. Recall and Detection Rates in Screening Mammography. Cancer 2004;100(8):1590–1594. [PubMed: 15073844]
16. Metz CE, Herman BA, Shen J-H. Maximum-likelihood estimation of ROC curves from continuously-distributed data. Stat. Med 1998;17:1033. [PubMed: 9612889]
17. American College of Radiology (ACR). Breast Imaging Reporting and Data System Atlas (BI-RADS Atlas). Reston, Va: American College of Radiology; 2003.
18. Gelman, A.; Carlin, JB.; Stern, HS.; Rubin, DB. Bayesian Data Analysis. New York: Chapman & Hall Publishers; 1995.
19. Johnson, NL.; Kotz, S. Distributions in Statistics: Continuous Multivariate Distributions. Vol. vol. 4. New York: Wiley; 1972.

Appendix 1

In this appendix we describe a Monte-Carlo approach [18] for computing Bayesian confidence intervals on the estimated ERU. We will assume two matched patient cohorts of defined size from which the number of recalled patients (N_R) and patients with detected cancers (N_D) are drawn. The method can be easily adapted to other experimental designs such as a fixed total number of patients with individuals selected randomly for each modality.

For simplicity in the analysis below, we will subdivide each cohort of N patients into three groups: cases with detected cancers (N_D), false-positive recall cases in which patients were recalled but found not to have disease ($N_{FR} = N_R - N_D$), and screening negative cases ($N_{Neg} = N - N_R$). The total number of cases is the sum of these three groups. In a given sample, we consider these 3 numbers to be drawn from a multinomial distribution

$$P(N_D, N_{FR}, N_{Neg} | R_D, R_{FR}, R_{Neg}) = \frac{N!}{N_D! N_{FR}! N_{Neg}!} R_D^{N_D} R_{FR}^{N_{FR}} R_{Neg}^{N_{Neg}}. \quad \text{(A1)}$$

where R_D, R_{FR}, and R_{Neg} are the probabilistic rates associated with each patient subdivision. If we assume a noninformative uniform prior on all possible combinations of rates that satisfy $R_D + R_{FR} + R_{Neg} = 1$, the posterior distribution is

$$P(R_D, R_{FR}, R_{Neg} | N_D, N_{FR}, N_{Neg}) = \frac{(N+2)!}{N_D! N_{FR}! N_{Neg}!} R_D^{N_D} R_{FR}^{N_{FR}} R_{Neg}^{N_{Neg}}, \quad \text{(A2)}$$

which is a Dirichlet distribution [18,19] with parameters ($N_D + 1, N_{FR} + 1, N_{Neg} + 1$).

Data from 2 cohorts, ($N_{D,1}, N_{FR,1}, N_{Neg,1}$) and ($N_{D,2}, N_{FR,2}, N_{Neg,2}$) define two posterior Dirichlet distributions according to Equation A2. The posterior sampling approach [18] proceeds by drawing an independent sample from each posterior distribution, and using the sampled rates to determine a sampled ERU. This process is repeated many times to make a large set of sampled ERU values from which to construct one- or two-sided confidence intervals. Let $\left(R_{D,1}^{(m)}, R_{FR,1}^{(m)}, R_{Neg,1}^{(m)}\right)$ and $\left(R_{D,2}^{(m)}, R_{FR,2}^{(m)}, R_{Neg,2}^{(m)}\right)$ for $m = 1, \cdots, M$, be posterior samples for cohort 1 and 2 respectively. Note that a Dirichlet distribution can be sampled using independent draws from a Gamma distribution, normalized by their sum. The sample ERU is then given by

$$\mathrm{ERU}^{(m)} = \frac{R_{FR,1}^{(m)} - R_{FR,2}^{(m)}}{R_{D,1}^{(m)} - R_{D,2}^{(m)}}. \quad \text{(A3)}$$

Note that $R_{FR} = R_R - R_D$, and hence Equation A3 directly corresponds to Equation 11 from the text. Confidence intervals and other inferences can then be constructed from these samples. For example, a 95% confidence interval can be estimated by finding a range of ERU values that contain 95% of the posterior samples.

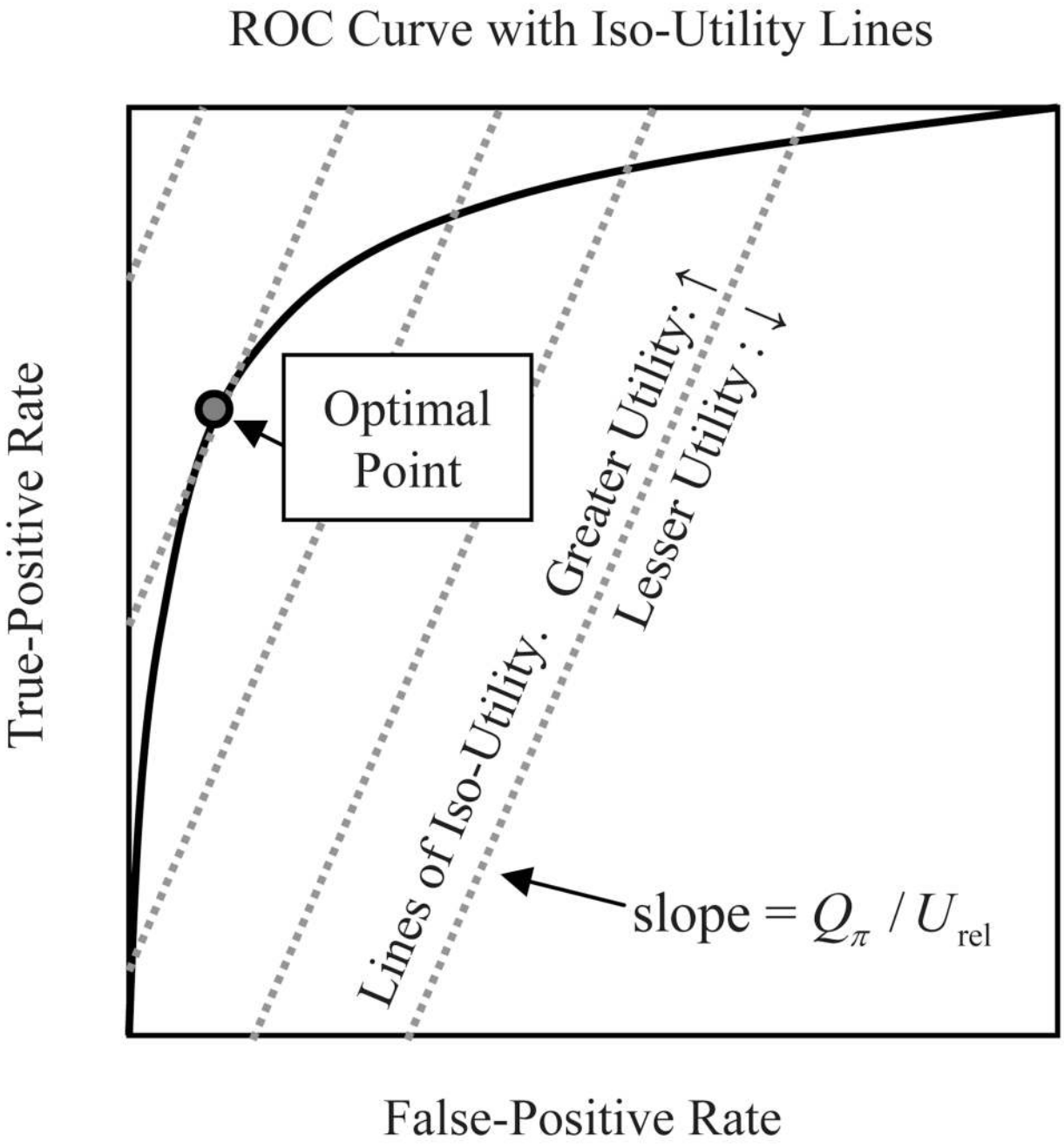

Figure 1. Utility in the ROC domain

A generic ROC curve is shown along with lines of iso-utility as defined in Equation 4. The iso-lines partition the domain into regions of higher utility (upper left) from lower utility (lower right). The relative utility (U_{Rel}) and the odds ratio for disease prevalence (Q_π) define the slope of iso-utility lines. The optimal operating point on the ROC curve is tangent to an iso-utility line.

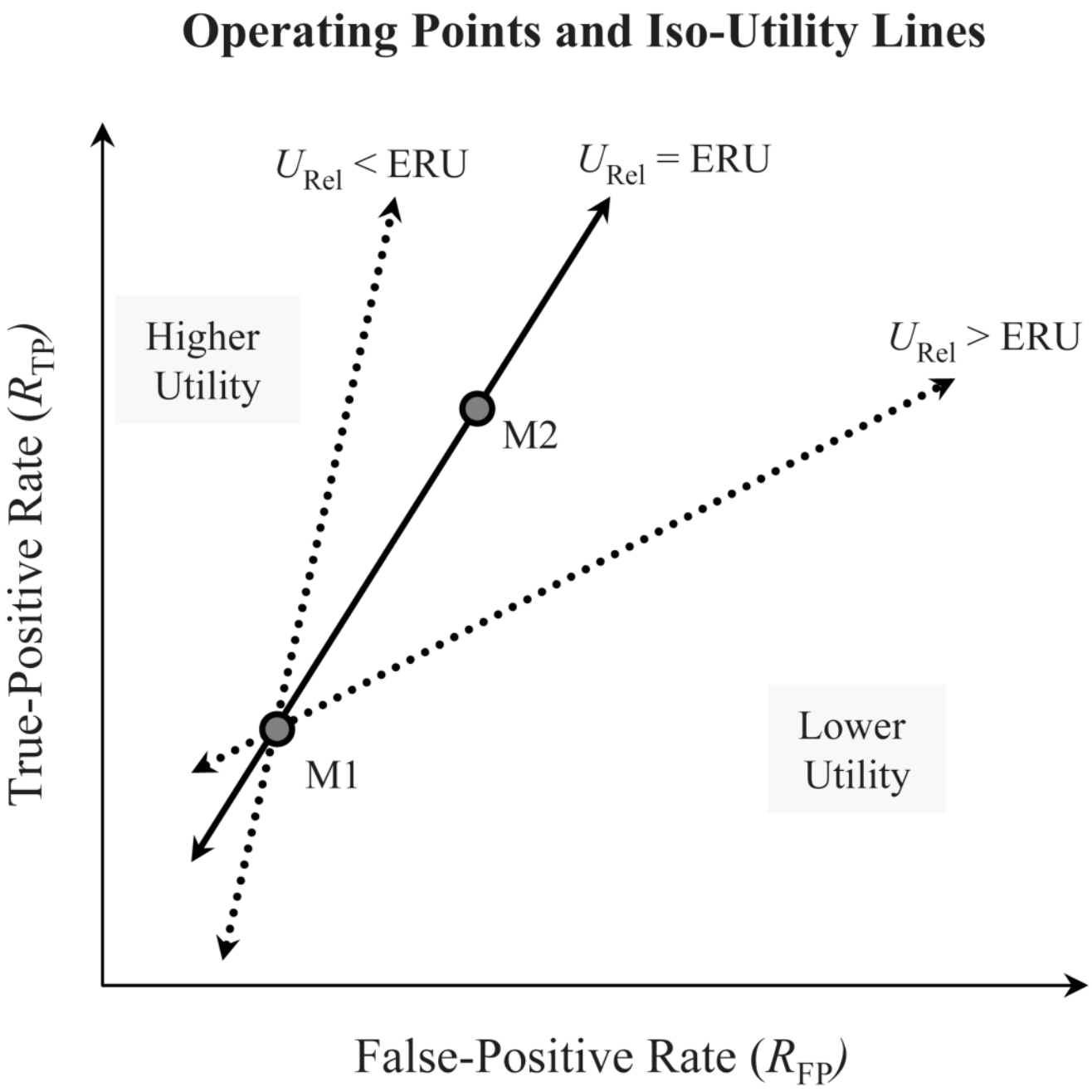

Figure 2. Interpretation of Equivalent Relative Utility (ERU)

This schematic figure shows the operating points of two hypothetical screening modalities (M1 and M2) in terms of true-positive and false-positive rates. Lines represent points of equal utility with higher utility to the upper right and lower utility to the lower left. Different lines represent different values of the putative relative utility (U_{Rel}) according to Equation 4. When U_{Rel} is equal to the ERU, the iso-utility line passes through both operating points. If U_{Rel} is less than the ERU, Modality 2 falls on the lower utility side of the iso-line and Modality 1 is superior. Conversely, if U_{Rel} is greater than the ERU, Modality 2 is superior.

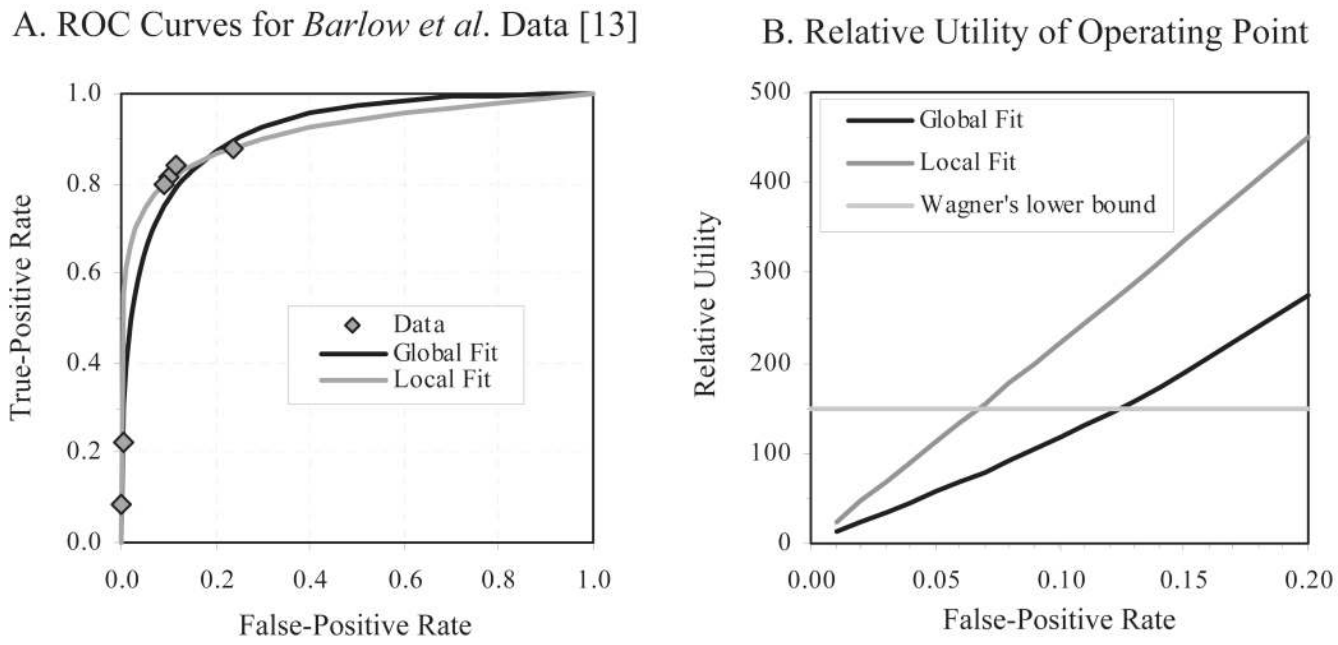

Figure 3. Utility Analysis of the Barlow *et al*. Mammography Data

ROC curves fit to the aggregate data (A) for a 7-point scale (Global Fit) or a reduced 5-point scale (Local Fit) have similar A_z values (0.920 global and 0.916 local). The relative utility (B) over a limited range of false-positive rates is computed from the slope of the ROC curves using a prevalence of 0.512%. Relative utilities near the 10% False-positive rate bracket the Wagner *et al*.[12] estimate of 150.

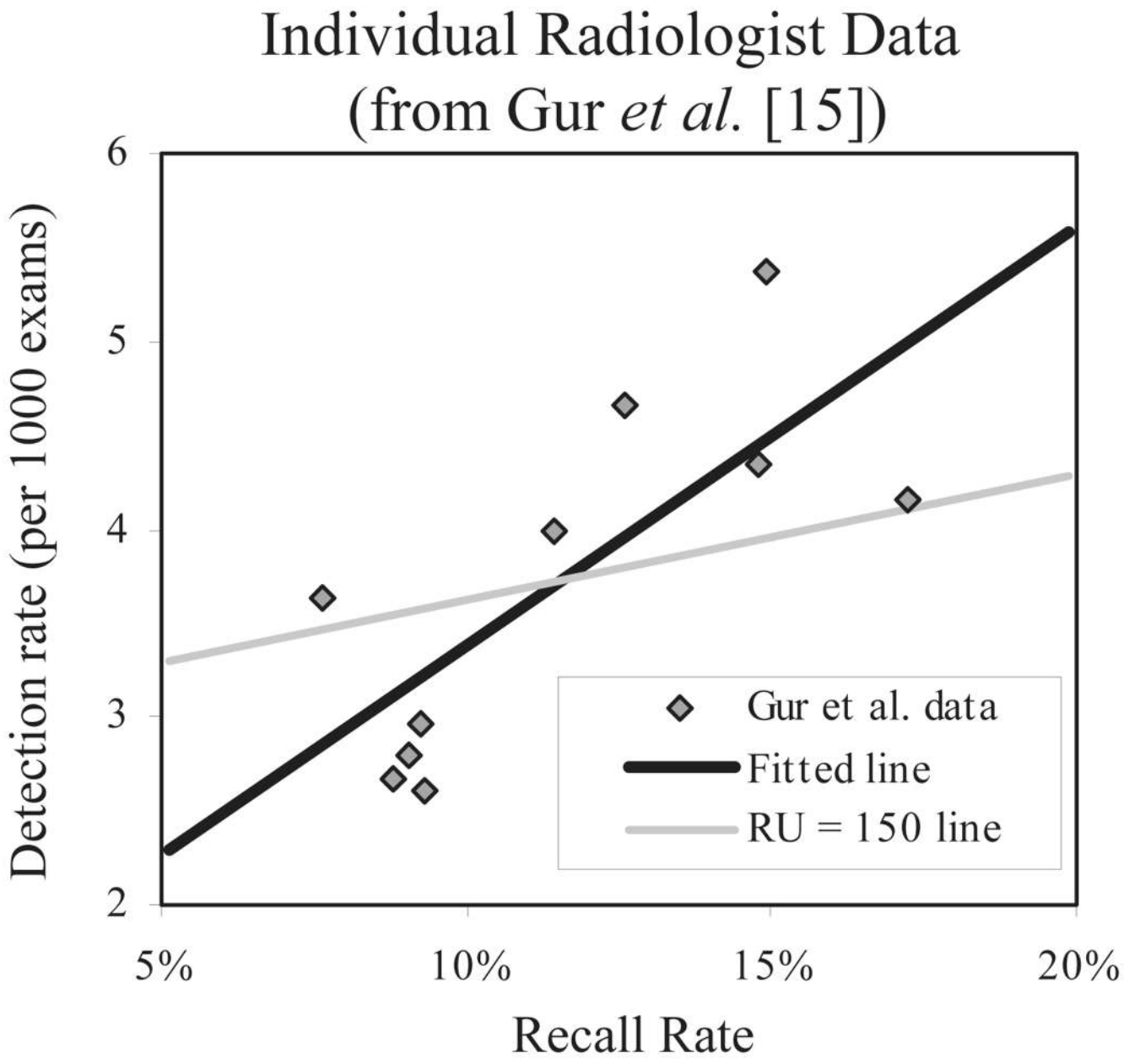

Figure 4. Equivalent Relative Utility Analysis of the Gur *et al*. Mammography Data
Recall and detection rates are plotted along with the best fit line from the Gur *et al*. publication, yielding an ERU of 46. We also plot a line with a slope corresponding to an ERU of 150 and least squares fitted intercept.

Table 1

Equivalent Relative Utility Analysis of Hambly *et al.* Data [14]

The data is partitioned into Initial and Subsequent Exam cohorts. For each cohort, the number of patients, recall rate, and detection rate data for screen film mammography (SFM) and full-field digital mammography (FFDM) exams is taken directly from the Hambly *et al.* publication. The equivalent relative utility (ERU) comparing SFM and FFDM is given along with the posterior probability of an ERU that is greater than or equal to the putative value of 150.

Exam	Screening Modality	N	Recall Rate (%)	Detection Rate (/1000)	ERU	$p(\text{ERU} \geq 150)$
Initial Exam	SFM	41,744	5.65	7.02	---	---
	FFDM	7,351	7.17	7.21	78.9	0.48
Subsequent Exam	SFM	94,694	2.16	4.79	---	---
	FFDM	19,242	2.72	5.87	4.1	0.03