Published in final edited form as:

Acad Radiol. 2008 November ; 15(11): 1446–1457. doi:10.1016/j.acra.2008.04.022.

Prevalence Scaling: Applications to an Intelligent Workstation for the Diagnosis of Breast Cancer

Karla Horsch, PhD, Maryellen L. Giger, PhD, and Charles E. Metz, PhD

Department of Radiology, The University of Chicago, Chicago, IL

Introduction

Intelligent workstations for computer aided diagnosis (CAD) promise to reduce the biopsy rate of benign lesions while maintaining high sensitivity [1,2]. Such workstations provide an estimate of a lesion's probability of malignancy (PM), often obtained from the output of a neural network classifier that has been trained on some database. By Bayes' rule, a lesion's probability of malignancy depends upon the prevalence of cancer in the population from which it was drawn: the greater the cancer prevalence in the population, the greater the probability that a case sampled randomly from that population with any given image characteristics will prove to be cancerous. The prevalence inherent in the computer-estimated PM may not be the prevalence encountered by the user in the clinic. For example, when the PM has been estimated using a Bayesian neural network, the prevalence inherent in the computer-estimated PM is the prevalence in the classifier's training database. Moreover, the prevalence of cancer encountered clinically by the user depends on the particular clinical setting in which that user is working. For example, the prevalence of breast cancer encountered by breast imaging radiologists working in breast centers to which patients are referred after being diagnosed elsewhere can be expected to be higher than that encountered by general radiologists. The reported prevalence of cancer in the diagnostic work-up population is about 3%¹ [3]. To our knowledge, data concerning differences in prevalence among various types of clinical environments have not been published. In addition, many of the observer studies performed to quantify the benefit of CAD have used only biopsy-proven cases, for which the prevalence of cancer may vary between 10–30% [4]. Researchers must therefore choose from among many values of prevalence when developing, testing and clinically translating a diagnostic computer aid.

Because estimates of probability of malignancy are prevalence dependent, the radiologist's computer-aided performance potentially depends upon the prevalence that is inherent in the computer output. For example, when the prevalence inherent in the computer-estimated PM differs from the prevalence that is clinically relevant to the user, the computer's estimate of PM may prove to be confusing to its user, limiting the user's ability to employ the computer aid effectively. In this situation, it is interesting to consider the possibility of allowing each user to modify the computer-estimated PM to reflect the prevalence that is clinically relevant

Corresponding Author: Karla Horsch, PhD, Department of Radiology, MC2026, The University of Chicago, 5841 South Maryland Ave, Chicago, IL 60637, kjhorsch@uchicago.edu, Phone: 773-834-5099, FAX: 773-702-0371.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

¹At the time of this writing, the Breast Cancer Surveillance Consortium reported on its website that out of 714,984 diagnostic breast exams, the number of cancers diagnosed within one year of an exam with an abnormal interpretation was 21,365.

to that user. Questions of whether and how the radiologist's computer-aided performance depends upon the prevalence inherent in the computer output can be answered only through observer studies and are not addressed in this paper. Instead, we seek here to better understand the degree to which clinically relevant changes in prevalence affect the computer PM output as well as the computer's optimal operating point, as defined by maximizing the expected utility of computer-aided decisions [5,6]. If clinically relevant changes in prevalence have "large" effects on the computer output and, thus, on patient-management outcomes, then the issue of prevalence cannot be ignored and observer studies that investigate the effect of changes in prevalence on radiologist performance are needed.

In order to investigate the effect of changes in prevalence on our computer's PM output, we considered a scaling transformation based on Bayes' rule that rescales the PM reflecting one prevalence to the PM reflecting another. We will use this transformation to show how changes in prevalence affect the PM-estimate histograms and computer performance of our multi-modality intelligent workstation for breast lesions. In order to investigate the effect of changes in prevalence on our computer classifier's "optimal" sensitivity and specificity, as determined by utility maximization, we use the conventional binormal model [7] to represent the receiver operating characteristic (ROC) curves of our mammographic and sonographic computer classifiers.

Materials and Methods

Databases

Three retrospectively collected databases of breast lesions were used in this study. The first is a mammographic database of 319 lesions (cancer prevalence 57%), whereas the second is a sonographic database of 358 lesions (cancer prevalence 19%). The third is an independent database of 97 cases with both mammograms and breast sonograms (cancer prevalence 37%). These databases have been described in detail elsewhere [8]. The mammographic and sonographic databases were used to train our mammographic and sonographic classifiers, respectively, and to generate the mammographic and sonographic PM histograms that are displayed on our intelligent workstation, whereas the independent multi-modality database was used to obtain unbiased estimates of the classification performance of both classifiers.

Computer classifiers

Our mammographic and sonographic CAD systems that classify breast mass lesions have been described in detail elsewhere (mammographic classifier: [9,10]; sonographic classifier: [11–13]). On the basis of a manually defined lesion center, both systems automatically segment the mass lesion from the normal tissue and automatically extract features. Four features are extracted by the mammographic workstation quantifying spiculation, margin-sharpness, texture and shape in some computer-defined neighborhood of the lesion [9,10]. The sonographic workstation also extracts four features, quantifying lesion shape, margin, texture and posterior acoustic behavior [11,12]. The mammographic classifier merges its four features by using a Bayesian neural network with four hidden nodes, whereas a Bayesian neural network with three hidden nodes is used to merge the four sonographic features. The mammographic and sonographic Bayesian neural networks were trained on the mammographic and sonographic database, respectively.

These mammographic and sonographic computer classifiers were used to determine computer PM estimates for all of the cases in all three databases. Because Bayesian neural networks were used, the prevalence inherent in the computer-estimated PM is the prevalence in the classifier's training database (57% and 19% for the mammographic and sonographic classifiers, respectively). The re-substitution AUC achieved by the mammographic and

sonographic classifiers when they were tested on their own training databases was 0.89 ± 0.02 for both classifiers, whereas AUC values of 0.81 ± 0.04 and 0.93 ± 0.03 were achieved by the mammographic and sonographic classifiers, respectively, when they were tested on the independent multi-modality database.

Multi-Modality Intelligent Workstation

Our intelligent workstation [8] displays for a given lesion, the lesion images together with the numerical values of probability of malignancy (PM) as determined by the mammographic and sonographic classifiers. These numerical values are, however, meaningless without information concerning the computer's performance. The user is therefore also provided with a performance table giving the mammographic and sonographic classifiers' sensitivity and specificity at various PM thresholds. In addition, the computer's PM estimates are also shown relative to the PM histograms of malignant and benign lesions from the appropriate training database (for examples of histograms, see Figs. 3, 4 and 5). This graphical display allows for an intuitive interpretation of the computer-estimated PM output. When the PM output for a particular lesion is in an interval for which there is a high concentration of malignant training database PM values and low concentration of benign training database PM values, the user can have more confidence that the lesion is malignant.

Kinds of Prevalence

In this paper, we find it convenient to differentiate among various kinds of prevalence. *Clinical prevalence* (denoted by η_{clin}) is the prevalence of cancer in the relevant clinical population, which for us is the prevalence of breast cancer that is encountered in diagnostic breast imaging clinics: approximately 3% [3]. In contrast, *computer prevalence* (denoted by η_{comp}) is the prevalence of cancer inherent in the computer estimate of the PM — i.e., the prevalence for which the computer's estimates of PM are most accurately calibrated. This is often the prevalence of cancer in the database that was used to train the automated classifier. The third type of prevalence, which we call the *scaled prevalence* (denoted by η_{scal}), will be described next.

To motivate the notion of scaled prevalence, we begin with the Bayesian expression for the posterior probability of malignancy,

$$p(\text{malignant}|x) = \frac{\left(\frac{\eta}{1-\eta}\right)R(x)}{\left(\frac{\eta}{1-\eta}\right)R(x) + 1}, \quad (1)$$

where x is the test result (for example, a feature vector), p is the posterior probability of malignancy given x , η is the prior probability of malignancy (e.g., prevalence in the population to which the case in question belongs) and R is the likelihood ratio $p(x | \text{malignant})/p(x | \text{not malignant})$. The computer PM output is an estimate of the posterior probability of malignancy when the prior probability of malignancy η is the computer prevalence. How good this approximation is depends upon how well the computer's neural network approximates the probability of malignancy given the feature vector [14]. For general η , the posterior probability of malignancy in Eqn. 1 is approximated by the computer PM estimates that *would have been obtained had the computer prevalence been η* (for example, if the training database prevalence had been η). The Bayesian expression given by Eqn. 1 can therefore be used to scale the computer's PM estimates to reflect a prevalence other than the inherent computer prevalence. We call this new prevalence the *scaled prevalence*.

In the next section, we will discuss the transformation that scales the original computer-estimated PM, reflecting the computer prevalence, to the scaled PM, reflecting the scaled-prevalence. The scaled prevalence can take on any value between 0 and 1. The value assigned to the scaled prevalence then defines a “scale” for the resulting PM output. What is the “best” value to assign the scaled prevalence is an open question. A natural choice is to let the scaled prevalence equal the clinical prevalence, in which case the resulting PM output represents the posterior probability of malignancy encountered in the clinic. Indeed, one of the motivations of the research presented in this paper is to allow the user to scale the computer-estimated PM output to reflect different clinical prevalence. However, in general, it is not known whether the prevalence reflected in the computer-estimated PM output affects radiologist performance, and if so, what is the best prevalence (the best scaled prevalence) by which to scale PM output. Thus, we will not make any assumptions concerning the value of the scaled-prevalence. In particular, although the scaled prevalence *can* equal the clinical prevalence, we will not assume that it does.

Prevalence Scaling

To scale the computer’s PM to reflect another prevalence, Bayes’ rule can be used to obtain the transformation

$$p' = \frac{\kappa p}{(\kappa - 1)p + 1}, \quad (2)$$

where p is the computer PM, p' is the scaled PM, and κ is the ratio of the scaled odds to the computer odds—i.e., the ratio of the (prior) odds of malignancy when the prevalence is η_{scal} to the odds of malignancy when the prevalence is η_{comp} :

$$\kappa = \frac{\eta_{\text{scal}} / (1 - \eta_{\text{scal}})}{\eta_{\text{comp}} / (1 - \eta_{\text{comp}})}. \quad (3)$$

Note that Eqn. 2 can, of course, be used to scale the threshold used on the computer-estimated PM output, as the threshold on PM is itself a PM. Figure 1 shows the results of prevalence scaling for various values of odds ratio. Using the prevalence-scaling transformation given by Eqn. 2, the PM estimates provided by an intelligent workstation can be scaled to reflect whatever prevalence is most appropriate.

It is important to note that prevalence scaling is a monotonic transformation and therefore does not affect the computer classifier’s ROC curve. Prevalence scaling does change the threshold at which any classifier achieves a particular sensitivity and specificity pair on its ROC, however.

We will study how scaling computer PM estimates to reflect another prevalence affects the PM histograms and classification performance, as measured by the pairs of TFP and FPF achieved at various thresholds, of our multi-modality intelligent workstation. In choosing which prevalences to employ in our study, we are motivated by the following observations: 1) a prevalence of approximately 50% is often the prevalence that is used to train, and therefore, is inherent in, laboratory-designed classifier PM estimates; 2) the prevalence of malignancy among biopsied cases can vary according to practice in a range from approximately 10% to 30% [4]; and 3) the prevalence of cancer in the diagnostic population is about 3% [3]. Therefore, we consider prevalences of 50%, 20% and 3%.

Expected Utility

The optimal operating point on the ROC curve of a diagnostic test used in clinical practice is the (FPF, TPF) pair that maximizes the expected utility [5,6] given by

$$E[U] = \eta_{clin}(U_{TP} \bullet TPF + U_{FN} \bullet FNF) + (1 - \eta_{clin})(U_{FP} \bullet FPF + U_{TN} \bullet TNF), \quad (4)$$

where η_{clin} is the clinical prevalence in the patient population of interest, whereas U_{TN} , U_{TN} , U_{TN} , and U_{TN} are the utilities associated with true negative (TN), false positive (FP), true positive (TP) and false negative (FN) decisions, respectively. [We use the word “utility” here in a general sense that connotes quantitative considerations of benefit and cost.] This expected utility reflects the clinical prevalence because we are interested in the operating point that should be used in the clinic from an expected-utility perspective. One can show [5] that this optimal operating point occurs where the slope of the ROC curve, and therefore the likelihood ratio, equals

$$R_{opt} = \left(\frac{1 - \eta_{clin}}{\eta_{clin}} \right) \left(\frac{U_{TN} - U_{FP}}{U_{TP} - U_{FN}} \right) = \frac{RUD}{k_{clin}}, \quad (5)$$

where the ratio of utility differences (RUD) and the pre-test odds (k_{clin}) are given by

$$RUD = \frac{U_{TN} - U_{FP}}{U_{TP} - U_{FN}} \quad \text{and} \quad k_{clin} = \frac{\eta_{clin}}{1 - \eta_{clin}}, \quad (6)$$

respectively. Equation 5 implies that at the optimal operating point, the post-test odds, which is the product of the likelihood ratio and the pre-test odds, is a constant given by the RUD:

$$\text{post - test odds} = R_{opt} k_{clin} = RUD \quad (7)$$

The likelihood ratio at the optimal operating point given in Eqn. 5 can be substituted into Eqn. 1 to determine the “optimal threshold” on the theoretical PM after it has been transformed to the scaled prevalence η_{scal} via Eqns. 2 and 3. This “scaled-optimal” PM threshold yields the optimal operating point on the ROC curve and is given by

$$\tau_{scal} = \frac{\kappa RUD}{\kappa RUD + 1}, \quad (8)$$

where RUD is given by Eqn. 6 and the scaled-to-clinical odds ratio is given by

$$\tilde{\kappa} = \frac{\eta_{scal} / (1 - \eta_{scal})}{\eta_{clin} / (1 - \eta_{clin})}. \quad (9)$$

(Note the difference between the scaled-to-clinical odds ratio (Eqn. 9), which is used to determine the scaled-optimal PM threshold, and the scaled-to-computer odds ratio (Eqn. 3),

which is used in the prevalence scaling transformation.) It is clear that when the scaled prevalence is chosen to equal the clinical prevalence (meaning that the PM has been scaled such that the prevalence reflected in the PM is the clinical prevalence and that $\kappa = 1$), the optimal PM threshold is prevalence-independent because then it depends only on RUD:

$$\tau_{clin} = \frac{RUD}{RUD+1}. \quad (10)$$

We will call the optimal threshold given above in Eqn. 10 the clinically-optimal threshold (i.e., the optimal threshold on the PM after scaling to the clinical prevalence). That the clinically-optimal threshold is prevalence independent means that, in practices with different clinical prevalences, the same optimal threshold should be used on PM estimates that have been scaled differently to reflect the different clinical prevalences in the different practices. For example, consider a diagnostic task for which the prevalence of disease is 60% (η_1 in the upper panel of Fig. 2) in one clinic and 20% (η_2 in the upper panel of Fig. 2) in another. The pre-test odds of disease are then $0.6/(1.0-0.6) = 1.5$ in the first clinic and $0.2/(1.0-0.2) = 0.25$ in the second. However, post-test odds are given by the product of likelihood ratio and the pre-test odds (See Eqn. 7), so to maintain the same post-test odds, one must operate at a smaller critical likelihood ratio (i.e., at an operating point closer to the upper right-hand corner in ROC space) in the first clinic than in the second. To summarize: the optimal (FPF, TPF) pair depends only on the clinical prevalence; the scaled-optimal PM threshold depends on both the clinical prevalence and the scaled prevalence, which in general may not equal the clinical prevalence; and the clinically-optimal PM threshold is prevalence independent (Fig. 2).

We remark that computer estimates of PM only approximate the Bayesian posterior probability of malignancy given by Eqn. 1, so the optimal threshold for the actual computer estimates only approximates the optimal threshold for the theoretical PM. Empirical determination of the computer-output threshold that produces any particular (FPF, TPF) pair on an ROC curve is difficult [15], so we will assume for this investigation that the “empirical” optimal threshold is a good approximation of the “theoretical” optimal threshold.

To determine the optimal (FPF, TPF) pair for a given classifier, we will assume the conventional binormal model [7] for the ROC curve of that classifier and find the combination of FPF and TPF that provides the curve slope given in Eqn. 5 [16]. The optimal (FPF, TPF) pair is then determined by the values of the binormal parameters a and b , together with choices of clinical prevalence and RUD value. In order to focus upon how changes in prevalence within a diagnostic breast-imaging population in the neighborhood of 3% [3] affect the optimal (FPF, TPF) pair, we will consider clinical prevalences in the range from 1% to 5%. Because determining values for the terms U_{TN} , U_{TN} , U_{TN} , and U_{TN} in Eqn. 4 is difficult and beyond the scope of this paper (for a discussion, see [17]), we instead use the relation between RUD and the clinically-optimal PM threshold to choose values of RUD. First, we will consider the value of RUD that is implied (via Eqn. 10) by a clinically-optimal threshold of 2%, because the American College of Radiology (ACR) recommends considering biopsy for all lesions with an assessed probability of malignancy greater than 2%² [18]. The RUD values determined from clinically-optimal thresholds of 1% and 0.5% will be investigated also. We will use the binormal parameters of our mammographic

²In particular, a BI-RADS category 3 (probably benign findings—initial short-term follow-up suggested) is given to those findings having a probability of malignancy of less than 2%, while the BI-RADS category 4 (suspicious abnormality—biopsy should be considered) contain findings with greater probability of malignancy. BI-RADS categories 1, 2, 3, 4 and 5 (and rarely 0) are used in the diagnostic setting (3% cancer prevalence), while BI-RADS categories 0, 1 and 2 tend to be used in the screening setting.

($a=1.1330$ and $b=0.8089$) and sonographic classifiers ($a=3.2897$ and $b=2.0281$), as estimated by LABROC4 [19,20] from our independent multimodality database of 97 cases.

Results

The mammographic and sonographic fractional occurrence histograms (malignant and benign histograms normalized separately such that each has unit total area) that our intelligent workstation displays are shown for scaled prevalences of 50%, 20% and 3% in Figs. 3, 4 and 5, respectively. For scaled prevalences of 20% and 3%, the PM estimates move toward the lower end of the probability scale, so it is useful also to show a “zoomed” version of the histograms for PM estimates between 0 and 0.20 (panels c and d in Figs. 4 and 5). The performance of the mammographic and sonographic classifiers for scaled prevalences of 50%, 20% and 3% (computed using our training databases) is reported also in Table 1. Note that when the scaled prevalence equals 3%, the clinical prevalence of cancer in the diagnostic population, a PM threshold of 0.02 corresponds to TPF and FPF of 0.80 and 0.23, respectively, for mammography and to TPF and FPF of 0.90 and 0.30, respectively, for sonography. To operate at the same combination of sensitivity and specificity after modifying the prevalence to 20% or 50% would require setting the PM threshold at approximately 0.14 or 0.40, respectively.

The optimal (FPF, TPF) pairs for the mammographic and sonographic classifiers evaluated on the independent database ($AUC = 0.81$, $a=1.13$, $b=0.81$ for mammography; $AUC = 0.93$, $a=3.29$, $b=2.03$ for sonography) were determined by using the conventional binormal model to fit the ROC curves with our LABROC4 algorithm [19,20] and are shown in Fig. 6 on their respective ROC curves for prevalences of 1%, 3% and 5% and for clinically-optimal thresholds of 0.005, 0.01 and 0.02. Because increasing clinical prevalence causes a decrease in the slope of the ROC at the optimal operating point, it is to be expected that increasing clinical prevalence will cause the optimal operating point to move toward the upper right hand corner in ROC space, and this trend can be seen in Fig. 6. The numerical values of TPF and FPF at the optimal operating points for our mammographic and sonographic classifiers are given in Table 2 for a clinically-optimal PM threshold of 0.02 and for various values of clinical prevalence.

For the higher-performing sonographic classifier, either an increase or a decrease of 2% from the clinical prevalence of 3% results in relatively small changes in optimal TPF and FPF for all three PM thresholds. However, for the lower performing mammographic classifier, an increase or decrease of 2% in clinical prevalence results in much larger absolute changes in optimal TPF and FPF.

Discussion

This paper has used Bayes' rule to explore the dependence of CAD output and optimal operating points on prevalence. In so doing, we have assumed that the computer output provides good estimates of the actual probabilities of malignancy, at least for some particular clinical or scaled prevalence. Any error in the computer's PM estimates propagates to estimates of scaled PM and, thus, to the optimal operating point. Another source of error in estimating the optimal operating point is uncertainty in estimation of the values of the binormal parameters of the ROC curve for each classifier. We have not attempted to study the impact of these errors at this time. It is important also to note that whenever we have used Bayes' rule, we have assumed that although prevalence changes, everything else remains the same. In particular, we have assumed that even if the prevalence of cancer changes with a change in clinical practice, the spectrum of case difficulty and the radiologist's ability to differentiate malignant and benign lesions do not.

In our analysis of optimal operating point, we did not attempt to assign values to utilities associated with various diagnostic decisions. Instead, we began with the PM threshold of 0.02 that is recommended by the ACR and used Bayes' rule to determine the ratio of utility differences that defines the ROC curve slope at the optimal operating point. It is interesting to note that, for a clinical prevalence of 3% and an optimal PM threshold of 0.02, Bayes' rule implies that the optimal TPF on our mammographic classifier's ROC curve is around 0.80³, a rather low sensitivity for diagnostic breast imaging. One possible solution to this seeming paradox might be to optimize utility with constraints imposed upon the minimum TPF allowed.

Questions concerning the prevalence inherent in the computer output are largely a matter of human/computer interface. Prevalence scaling does not affect the computer's classification performance — e.g., its ROC curve or the sensitivity and specificity at its optimal operating point — but potentially does affect the computer-aided radiologist's performance, in part because the sensitivity and specificity of the computer's scaled PM estimates will change unless the PM threshold is modified accordingly. Thus, when a computer aid based on estimates of the probability of malignancy is translated to the clinic, the question of which scaled prevalence to have the computer report becomes important.

On the other hand, the clinical prevalence that exists in a particular clinical setting does affect the computer's optimal operating point — i.e., optimal combination of sensitivity and specificity or, equivalently, TPF and FPF. How sensitive the optimal operating point is to changes in clinical prevalence is a potentially important question in the development of computer classifiers. This question is related to the shape of the ROC curve at the optimal operating point. We found, not surprisingly, that assuming optimal thresholds of 2% and less, the optimal operating point our lower performing mammographic classifier was more sensitive to changes in clinical prevalence than was our higher performing sonographic classifier.

Conclusions

By use of prevalence scaling, computer PM estimates can be transformed to reflect whatever prevalence future research deems to be most appropriate in a particular clinical setting. Relatively small changes in clinical prevalence can have large effects on the computer classifier's optimal operating point.

Bibliography

1. Giger, ML.; Huo, Z.; Vyborny, CJ.; Lan, L.; Nishikawa, RM.; Rosenbrough, I. Results of an observer study with an intelligent mammographic workstation for CAD. In: Peitgen, H-O., editor. *Digital Mammography: IWDM 2002*. Berlin: Springer; 2002. p. 297-303.
2. Giger, ML.; Huo, Z.; Lan, L.; Vyborny, CJ. Intelligent search workstation for computer aided diagnosis. In: Lemke, HU.; Inamura, K.; Kunio, D.; Vannier, MW.; Farman, AG., editors. *Proceedings of Computer Assisted Radiology and Surgery (CARS) 2000*. Philadelphia, Pa: Elsevier; 2000. p. 822-827.
3. Web site of the Breast Cancer Surveillance Consortium. [Accessed February 26, 2008]. <http://breastscreening.cancer.gov/data/benchmarks/diagnostic>
4. Newstead, G. Personal correspondence. Sept 22. 2006
5. Metz CE. Basic principles of ROC analysis. *Semin Nucl Med* 1978;8:283–298. [PubMed: 112681]

³When thresholding is used on the PM estimates of cases in the mammographic training database, the TPF is 0.80 (Table 1). The TPF is 0.78 when determined from the binormal ROC curve using the binormal parameters that we estimated for our mammographic classifier evaluated on an independent multi-modality database (Table 2).

6. Green, DM.; Swets, JA. Signal Detection Theory and Psychophysics. Huntington, NY: Krieger; 1974.
7. Dorfman DD, Alf E. Maximum-likelihood estimation of parameters of signal-detection theory and determination of confidence intervals — rating-method data. *J Math Psychol* 1969;6:487–496.
8. Horsch K, Giger ML, Vyborny CJ, Lan L, Mendelson EB, Hendrick RE. Multi-modality computer-aided diagnosis for the classification of breast lesions: observer study results on an independent clinical dataset. *Radiology* 2006;240:357–368. [PubMed: 16864666]
9. Huo Z, Giger ML, Vyborny CJ, Bick U, Lu P, Wolverton DE, Schmidt RA. Analysis of spiculation in the computerized classification of mammographic masses. *Med Phys* 1995;22:1569–1579. [PubMed: 8551981]
10. Huo Z, Giger ML, Vyborny CJ, Wolverton DE, Schmidt RA, Doi K. Automated computerized classification of malignant and benign masses on digitized mammograms. *Acad Radiol* 1998;5:155–168. [PubMed: 9522881]
11. Giger ML, Al-Hallaq H, Huo Z, Moran C, Wolverton DE, Chan CW, Zhong W. Computerized analysis of lesions in US images of the breast. *Acad Radiol* 1999;6:665–674. [PubMed: 10894069]
12. Horsch K, Giger ML, Venta LA, Vyborny CJ. Automatic segmentation of breast lesions on ultrasound. *Med Phys* 2001;28:1652–1659. [PubMed: 11548934]
13. Horsch K, Giger ML, Venta LA, Vyborny CJ. Computerized diagnosis of breast lesions on ultrasound. *Med Phys* 2002;29:157–164. [PubMed: 11865987]
14. Kupinski MA, Edwards DC, Giger ML, Metz CE. Ideal observer approximation using Bayesian classification neural networks. *IEEE Trans Med Imaging* 2001;20:886–899. [PubMed: 11585206]
15. Pesce, LL.; Horsch, K.; Metz, CE. Estimation of the relationship between test-result values and operating points on “proper” binormal ROC curves. Presented at MIPS XII (12th meeting of the Medical Image Perception Society); Iowa City, Iowa, USA. October 2007;
16. 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]
17. 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]
18. American College of Radiology. Breast Imaging Reporting and Data System Atlas (BI-RADS Atlas). Reston, Va: American College of Radiology; 2003.
19. LABROC4. Website of Kurt Rossman Laboratories, Department of Radiology, University of Chicago; [Accessed June 10, 2004]. http://www-radiology.uchicago.edu/krl/roc_soft.htm
20. Metz CE, Herman BA, Shen J-H. Maximum-likelihood estimation of ROC curves from continuously-distributed data. *Statist Med* 1998;17:1033–1053.

Figure 1.
Prevalence-scaling transformation for various values of the odds ratio κ , given by Eqn. 3.

Clinically-optimal Threshold: Scaled and Clinical Prevalence Equal

The optimal threshold is *independent* of clinical prevalence.

- Two different clinical prevalence
- Two different distributions related by monotonic transformation
- Same ROC curve

The optimal operating point is *dependent* on clinical prevalence.

Scaled-optimal Threshold: Scaled and Clinical Prevalence Unequal

Figure 2. Diagrams depicting the relationship between the scaled and clinical prevalence on one hand and the optimal operating point and optimal threshold on the other. For both sets of diagrams, the two distributions shown are related by a monotonic prevalence-scaling transformation and therefore produce the same ROC curve.

Figure 3.

Assuming a prevalence of 50%, fractional occurrence histograms of the (a) mammographic and (b) sonographic PM estimates for the mammographic and sonographic training databases, respectively.

Figure 4.

Assuming a scaled prevalence of 20%, fractional occurrence histograms of the (a) mammographic and (b) sonographic PM estimates for the mammographic and sonographic training databases, respectively. Also shown are fractional occurrence histograms for the (c) mammographic and (d) sonographic PMs in the range from zero to 0.2.

Figure 5. Assuming a scaled prevalence of 3%, fractional occurrence histograms of the (a) mammographic and (b) sonographic PM estimates for the mammographic and sonographic training databases, respectively. Also shown are fractional occurrence histograms for the (c) mammographic and (d) sonographic PMs in the range from zero to 0.2.

Figure 6. Dependence of optimal operating points on clinical prevalence for clinically-optimal thresholds of 0.005, 0.01, 0.02.

Performance table of the mammographic and sonographic classifiers, showing PM thresholds for scaled prevalences of 50, 20 and 3% at various pairs of TPF and FPF. Table entries were computed by thresholding the PM estimates of cases in the training databases.

Table 1

PM threshold as a function of scaled prevalence			Mammographic classifier performance		Sonographic classifier performance	
50%	20%	3%	TPF	FPF	TPF	FPF
0.15	0.042	0.0054	0.96	0.79	1.00	0.56
0.20	0.059	0.0077	0.95	0.62	0.99	0.49
0.25	0.077	0.010	0.93	0.52	0.97	0.43
0.30	0.097	0.013	0.88	0.39	0.96	0.37
0.35	0.12	0.016	0.84	0.30	0.91	0.34
0.40	0.14	0.020	0.80	0.33	0.90	0.30

Table 2

Clinically-optimal pairs of TPF and FPF, assuming a clinically-optimal threshold of 0.02 (a threshold of 0.02 on PM estimates that have been scaled to reflect clinical prevalence). Pairs of TPF and FPF were computed using the binormal parameters associated to our mammographic and sonographic classifiers evaluated on an independent multi-modality database.

Clinical prevalence	Mammography		Sonography	
	Clinically-optimal operating point TPF	Clinically-optimal operating point FPF	Clinically-optimal operating point TPF	Clinically-optimal operating point FPF
5%	0.51	0.09	0.86	0.14
4%	0.68	0.21	0.93	0.19
3%	0.78	0.33	0.95	0.22
2%	0.84	0.44	0.97	0.24
1%	0.89	0.54	0.97	0.25