A Proposed Design and Analysis for Comparing Digital and Analog Mammography: Special Receiver Operating Characteristic Methods for Cancer Screening

Stuart G. BAKER and Paul F. PINSKY

Because randomized trials have shown a reduction in breast cancer mortality, analog mammography for the early detection of breast cancer has gained widespread use. Recently, several manufacturers have developed digital mammography, which promises great advantages in the storage and transmission of images. We were asked to design a study to compare the two types of mammography in terms of their performance for the early detection of breast cancer. A standard measure of mammography performance is the receiver operating characteristic (ROC) curve, which is a plot of false- and true-positive rates for each ordered classification of the mammography images. Methods for study design and data analysis based on ROC curves have been well developed for diagnostic tests, particularly in radiology. But for comparing the performance of mammography for the early detection of breast cancer among asymptomatic women, special considerations motivate new designs and methodology. First, digital mammography may cost substantially more than analog mammography. If this is the case, then the standard paired design, in which each subject undergoes both types of mammography, may be more expensive than necessary. To reduce costs, we propose a partial testing design, in which all subjects undergo analog mammography and those recommended for biopsy and a random sample not recommended for biopsy also undergo digital mammography. Second, the false-positive rate for analog mammography, defined as the rate of unnecessary biopsy, is near 1%. A standard ROC analysis that compares areas under the entire ROC curve would summarize performance over false-positive rates that are not relevant for evaluating the performance of cancer screening. As a more appropriate alternative, we propose basing inference on the areas under the small part of the ROC curves near the false-positive rates corresponding to a biopsy recommendation. Third, the vast majority of screened subjects are not biopsied, and so have an unknown cancer state at the time of screening. To make inference about the performance of a cancer screening test, the standard approach is to follow subjects not biopsied for some period, usually 1 year, and assume that those who developed cancer were missed on screening and those who did not develop cancer were cancer-free at screening. Unfortunately, this follow-up period can greatly lengthen the duration of the study. To compare the performance of digital and analog mammography without the need for a follow-up period, we propose estimating the *ratio* of areas under the ROC curves near the small false-positive rates associated with a biopsy recommendation. To compute sample sizes, our null hypothesis is that the ratio of partial ROC areas is 1, and our two possible alternative hypotheses are ratios of 1.6 and 2, both indicating superior performance for digital mammography. We assume a breast cancer prevalence of .003 and specify various parameters for the shapes of the ROC curves and their dependence. For a two-sided type I error of .05 and a power of .9, a standard paired design would require that 22,000 subjects undergo both analog and digital mammography. For the same type I error and power, the proposed partial testing design would require that 35,000 subjects undergo analog mammography and 10,000 subjects undergo both analog and digital mammography. Compared to the paired design, the reduction in the cost per subject is 23% if digital mammography costs four times as much as analog mammography and 41% if digital mammography costs 10 times as much as analog mammography.

KEY WORDS: Cancer screening; Diagnostic testing; Double sampling; Permutation test; Sample size; Verification bias; Sensitivity; Specificity.

1. INTRODUCTION

Because results from randomized clinical trials of analog, also known as film screen, mammography have shown reductions in breast cancer mortality (e.g., Kerlikowske, Grady, Rubin, Sandrock, and Ernster 1995), analog mammography has gained widespread use in the U.S. According to a 1992 U.S. survey, 36% of women age 50 or older received a mammography within the past year and 50% received one within the past 3 years (Anderson and May 1995). Several manufacturers have recently developed mammographic machines with digital image receptors. According to Winfeld et al. (1994), digital mammography offers four potential advantages: (1) improved image quality, (2) digital image processing for improved lesion contrast; (3) computer-aided diagnosis for enhanced radiological interpretation; and (4) teleradiology for facilitated radiological consultation." In addition, digital mammography promises improved storage and retrieval of images.

We were asked to design a study to compare the performance of digital and analog mammography in early detection of breast cancer. Although a randomized trial with a mortality endpoint would be more informative, the anticipated duration of 10 years makes such a trial impractical. For breast cancer screening to reduce mortality, the cancer must be detected early, and the early intervention must be effective. Given the similarity in the screening modalities, the early interventions should be similar, which provides some justification for comparing performance in early detection.

Performance for disease detection is typically summarized using a receiver operating characteristic (ROC) curve, which plots true- and false-positive rates for various cutpoints of the test result (e.g., Swets 1988). Compared with an evaluation at a single cutpoint, an ROC formulation avoids the arbitrariness of selecting a particular cutpoint, avoids the difficulties

Stuart G. Baker is Mathematical Statistician, Biometry Research Group, and Paul F. Pinsky is Mathematical Statistician, Early Detection Research Group, Division of Cancer Prevention, National Cancer Institute, Bethesda, MD 20892 (E-mail: *sb16i@nih.gov*). The authors thank Constantine Gatsonis, John Gohagan, Philip Prorok, William Sacks, Dan Sullivan, Robert Wagner, and the reviewers for helpful comments.

of making joint inferences for true- and false-positive rates, and provides more data. Study designs and analyses involving ROC curves have been well developed for comparing the performance of diagnostic tests. But for comparing the performance of digital and analog mammography, special considerations motivate the development of new designs and analyses. This article is organized as follows. In Section 2 we propose the partial testing design for reducing costs when digital mammography is more expensive than analog mammography. In Section 3 we propose a test statistic that is based on the area under the relevant part of the ROC curve and that eliminates the need for the standard follow-up period. In Section 4 we compute sample size.

2. DESIGN

In the standard paired design for comparing the performance of diagnostic tests, each subject undergoes both tests (Hanley and McNeil 1983). In a study comparing the performance of digital and analog mammography, the cost of digital mammography may be substantially greater than the cost of analog mammography, particularly if most of the cost of analog mammography is borne by insurance. In that case, one can substantially reduce costs by implementing the following extension of the partial testing design of Baker, Connor, and Kessler (1998). All subjects undergo analog mammography. A radiologist reads the analog mammogram, obtains any additional views or tests results, and provides one of the following ratings: (0) no biopsy recommendation, (1) biopsy recommendation with a likelihood of cancer below 20%, (2) a biopsy recommendation with a likelihood of cancer of 20–80%, and (3) a biopsy recommendation with likelihood of cancer greater than 80%. All subjects with ratings (1), (2), or (3) are selected to undergo digital mammography, and a subject with rating (0) is selected to undergo digital mammography with probability f. As discussed in Section 4, the choice of f depends on the relative cost of digital versus analog mammography. Depending on the screening center and the particular case, digital mammography may be performed on the same day as the analog mammography or on another day. To avoid bias, a different radiologist reads the analog and digital mammograms from the same subject (Metz 1989). The radiologist reading the digital mammogram is blinded to the analog mammogram, the rating associated with analog mammography, and to any additional tests prompted by the analog mammography. The radiologist reads the digital mammogram and obtains any additional views or tests results, then provides one of the same four ratings as used with analog mammography.

Although this is the first proposed two-stage design for comparing ROC curves, two-stage designs have been applied in other situations (see the references in Baker et al. 1998). Mathematically, we could include any number of rating categories, but in practice five or six categories is usually the limit (Metz 1986). The proposed rating system differs from the standard BI-RADS score (American College of Radiology 1995) by providing more information about subjects recommended for biopsy, which is of primary interest, and no information about subjects not recommended for biopsy. In addition, it avoids the problems of comparing ROC curves based on BI-RADS (Pepe, Urban, Rutter, and Longton 1997).

3. TEST STATISTIC

We develop the test statistic as follows. First, we define the partial ROC area and discuss its computation. Second, we show how using the ratio of partial ROC areas as a test statistic avoids the need for follow-up of subjects not biopsied. Third, we modify the statistic to incorporate both planned missing data from the partial testing design and unplanned missing data from subjects not undergoing a recommended biopsy.

3.1 Partial Receiver-Operating Characteristic Areas

In comparing the performance of diagnostic tests, a standard procedure is to compare areas under the entire ROC curves (e.g., Hanley and McNeil 1983). But most of the ROC curve is not relevant for comparing the performance of cancer screening tests, because the false-positive rate must be very small for cancer screening to be acceptable (e.g., Lillienfeld 1974). But if the false-positive rate is too small, then the true-positive rate will be unacceptably low (e.g., Baker 2000). These considerations motivate basing inference on the area under the ROC curve in an interval between two small false-positive rates, which we call the *partial ROC area* (see also McClish 1989; Thompson and Zucchini 1989). In particular, we base inference on the partial ROC area in a small interval of false-positive rates bounded above by the false-positive rate associated with a biopsy recommendation.

One way to estimate the partial ROC area is to first fit a parsimonious model to the entire ROC curve, then use the estimates to compute the partial ROC area. But a good fit to the entire ROC curve does not guarantee a good fit to the small part of the ROC curve used to compute the partial ROC area. Instead, we propose estimating the partial ROC area using a modification of the trapezoidal rule of Bamber (1975). Let random variable A with realization a denote the ordered categories for the analog classification, where $A = 0$ denotes no biopsy recommendation and $A = 1, 2, 3$, denote the rating categories in Section 2 corresponding to a biopsy recommendation with increasing likelihoods of cancer. Similarly, let random variable D with realization d denote the ordered categories for the digital classification, where $D = 0$ denotes no biopsy recommendation and $D = 1, 2, 3$ denotes the rating categories corresponding to a biopsy recommendation with increasing likelihoods of cancer. Let random variable C with realization c denote the cancer state at time of mammography, with $C = 0$ if absent and $C = 1$ if present.

The basic parameters are $\alpha_{ca} = \Pr(A = a \mid C = c)$ and $\beta_{cd} = \Pr(D = d \mid C = c)$. False- and true-positive rates are sums of the basic parameters. For analog mammography, the false- and true-positive rates associated with category $i = 1, 2, 3$ are $\Pr(A \geq i \mid C = c) = \Sigma_{a=i}^{3} \alpha_{ca}$ for $c = 0$ and 1. A similar formula applies to the false- and true-positive rates associated with digital mammography. For analog and digital mammography, the set of false- and true-positive rates constituting a full ROC curve is

$$\mathrm{ROC}_A = \left\{ (0,0), (\alpha_{03}, \alpha_{13}), \left(\sum_{a=2}^{3} \alpha_{0a}, \sum_{a=2}^{3} \alpha_{1a} \right), \right.$$
$$\left. (\alpha_0, \alpha_1), (1,1) \right\},$$

and

$$\mathrm{ROC}_D = \left\{(0,0), (\beta_{03},\beta_{13}), \left(\sum_{a=2}^{3}\beta_{0a}, \sum_{a=2}^{3}\beta_{1a}\right), (\beta_0,\beta_1), (1,1)\right\},$$

$$\text{where } \alpha_c = \alpha_{c3}+\alpha_{c2}+\alpha_{c1} \text{ and } \beta_c = \beta_{c3}+\beta_{c2}+\beta_{c1}. \tag{1}$$

Without loss of generality, consider Figure 1 with a higher ROC curve for digital than analog mammography. The relevant part of the ROC curve is near the false-positive rate associated with a biopsy recommendation, which is the point on each ROC curve corresponding to rating category (1) and the point farthest to the right on each ROC curve in the expanded view. We do not consider any part of the ROC curve that extends beyond the maximum false-positive rate associated with a biopsy recommendation because (a) the extension would be based on the part of the ROC curve extending to (1,1), where the adjacent points are so far apart that interpolation could be misleading, and (b) estimation would require data from a follow-up period, as discussed in Section 3.2. Therefore, the maximum false-positive rate of interest is the smaller of the false-positive rates for digital analog mammography associated with rating category (1), $\beta_0 = \min(\alpha_0, \beta_0)$. In Figure 1 this corresponds to the false-positive rate of the point farthest to the right on the upper ROC curve in the expanded view. For the minimum false-positive rate of interest, to keep the interval small we select the larger of the false-positive rates for digital and analog mammography associated with rating category (2), $\alpha_{02}+\alpha_{03} = \max(\alpha_{02}+\alpha_{03}, \beta_{02}+\beta_{03})$. In Figure 1 this corresponds to the false-positive rate of the point second from the right on the lower ROC curve in the expanded view. We connect points by lines and compute areas under trapezoids, using linear interpolation where necessary. For Figure 1, the partial ROC areas between the dashed vertical lines are

$$\begin{aligned}\mathrm{Area}_A &= w(\alpha_{13}+\alpha_{12}) + w\alpha_{1s}/2,\\ \mathrm{Area}_D &= w(\beta_{13}+\beta_{12}+\beta_{11}-\beta_{1s}) + w\beta_{1s}/2,\end{aligned} \tag{2}$$

where $w = \beta_{03}+\beta_{02}+\beta_{01}-(\alpha_{03}+\alpha_{02})$ is the width of the interval, $\alpha_{1s} = w\,\alpha_{11}/\alpha_{01}$ is the height of the top triangle in the interval for the ROC curve for analog mammography, and $\beta_{1s} = w\,\beta_{11}/\beta_{01}$is the height of the top triangle in the interval for the ROC curve for digital mammography. The formulas in (2) can be easily modified if the ROC curve for digital mammography is lower than the ROC curve for analog mammography or if a different lower bound on the interval is selected.

3.2 Ratios of Partial Receiver Operating Characteristic Areas

One difficulty with using ROC methods for comparing the performance of cancer screening tests is that the cancer state at the time of screening is unknown if there is no biopsy. To circumvent this difficulty, the standard approach is to follow subjects not biopsied for a certain period, typically 1 year for breast cancer screening, and assume that those who developed cancer were missed on screening and those who did not develop cancer were cancer-free at the time of screening. Under the standard approach, one obtains data x_{cad}, the number of subjects with $A = a$ and $D = d$ in cancer state c. With these data one can estimate α_{ca} and β_{cd} by x_{ca+}/x_{c++} and x_{c+d}/x_{c++}, where the subscript "+" denotes summation over the corresponding index. Although the parameter estimates can be substantially biased (Day 1985), the estimated differences between ROC areas will have little bias (calculations not shown). Therefore, the main drawback of the standard approach is the need for a follow-up period.

Alternatively, to circumvent the aforementioned difficulty and also obviate a follow-up period, we propose the ratio of estimated partial ROC areas as a test statistic. With no follow-up, x_{c00} is not observed, although the total number with no

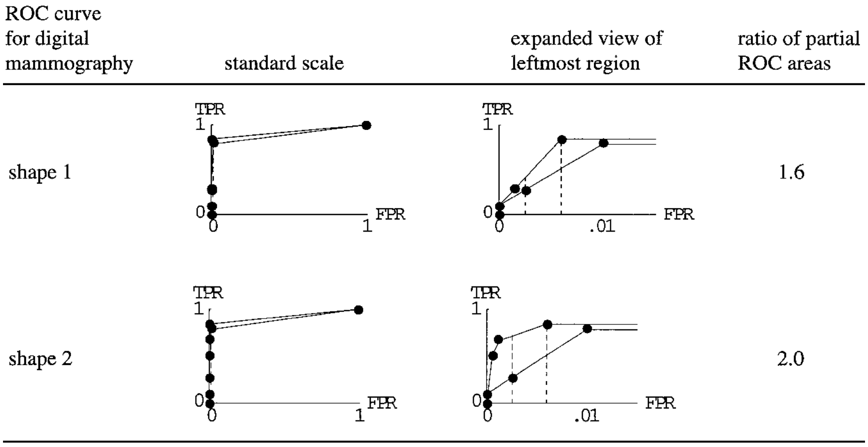

Figure 1. Shapes of the ROC Curves Under Two Possible Alternative Hypotheses Assuming a Prevalence of .003. In each plot the lower ROC curve corresponds to analog mammography and the upper ROC curve corresponds to digital mammography. The large dots plot the false- and true-positive rates associated with the rating categories. The partial ROC area for digital (analog) mammography is the area under the ROC curve for digital (analog) mammography between the dashed vertical lines. Without data from a follow-up period, one cannot estimate the ROC curves, but can estimate the ratio of partial ROC areas. See the text for details.

biopsy recommendation, x_{+00}, is observed. What makes this a particularly difficult estimation problem is that the gold standard is missing in *all* subjects with a particular test result–unlike in the study of Baker (1995), for example, where the gold standard was missing in only some of the subjects with a particular test result.

The motivation for using the ratio of partial ROC areas to obviate a follow-up period comes from Schatzkin, Connor, Taylor, and Bunnag (1987), who showed that one can estimate ratios of false- and true-positive rates for two tests even when the disease state is unknown in subjects negative on both tests. A key insight is to reparameterize $\alpha_{ca} = \Pr(A = a, C = c)/\Pr(C = c) = \alpha^*_{ca}/\gamma_c$ and $\beta_{cd} = \Pr(D = d, C = c)/\Pr(C = c) = \beta^*_{ca}/\gamma_c$. Although we do not observe x_{c00}, we observe x_{ca+} for $a > 0$, because a biopsy is recommended for all rating categories greater than 0. In addition, because we know the total number of subjects x_{+++}, for $a > 0$, we can estimate α^*_{ac} by x_{ca+}/x_{+++}. Similarly for $d > 0$,we can estimate β^*_{cd} by x_{c+d}/x_{c++}. Because we cannot estimate γ_c without follow-up data, we cannot estimate the partial ROC areas, Area_D and Area_A, without follow-up data. But even without follow-up data, we can estimate the ratio of partial ROC areas, $\text{Area}_D/\text{Area}_A$, because $\gamma_0\gamma_1$ cancels from the numerator and denominator! Note that to estimate the ratio of partial ROC areas without follow-up data, the partial ROC areas cannot be based on false-positive rates greater than the false-positive rate associated with category (1). Otherwise, the ratio of partial ROC areas would depend on $\Pr(A = 0, C = c)$, which requires follow-up data for estimation.

3.3 Partially Observed Data

Estimation is also complicated by two types of partially observed data. First, under the proposed partial testing design, not all subjects undergo digital mammography. Second, some subjects recommended for biopsy do not undergo the biopsy.

Let z_0 denote the number of subjects with $a = 0$ not selected for digital mammography under the partial testing design and hence not biopsied. Let y_{ad} denote the number of subjects with $A = a$ and $D = d$, excluding $a = d = 0$, who did not undergo the recommended biopsy. To easily incorporate these data into our test statistic, we reparameterize along the lines of Zhou (1998) by letting $\psi_a = \Pr(A = a)$, $\lambda_{d\,|\,a} = \Pr(D = d\,|\,A = a)$, and $\phi_{c\,|\,ad} = \Pr(C = c\,|\,A = a, D = d)$ for all a and d excluding the unidentifiable case where $a = d = 0$. We can then write

$$\alpha^*_{ca} = \sum_{d=0}^{k} \phi_{c\,|\,ad}\lambda_{d\,|\,a}\psi_a \quad \text{for } a > 0,$$

and

$$\beta^*_{cd} = \sum_{a=0}^{k} \phi_{c\,|\,ad}\lambda_{d\,|\,a}\psi_a \quad \text{for } d > 0. \tag{3}$$

We also define π_{ad} as the probability of undergoing a recommended biopsy, which we assume depends only on the rating categories for analog and digital mammography.

The model is saturated. With k categories associated with biopsy recommendation (we specified $k = 3$), there are $3\,k^2 - 3$ independent parameters: $k^2 - 1$ parameters in $[\phi_{c\,|\,ad}]$, $k(k-1)$ in $\{\lambda_{d\,|\,a}\}$, $k - 1$ in $\{\psi_a\}$, and $k^2 - 1$ in $\{\pi_{ad}\}$. There are also $3\,k^2 - 3$ independent cell counts: $2\,k^2 - 2$ in $\{x_{cad}\}$ and $k^2 - 1$ in $\{y_{ad}\}$. The value of z is fixed by the design parameter f. The maximum likelihood estimates (see the Appendix) are the obvious proportions based on all of the available data,

$$\widehat{\phi}_{c\,|\,ad} = x_{cad}/x_{+ad}, \tag{4}$$

$$\widehat{\lambda}_{d\,|\,a} = (x_{+ad} + y_{ad})/(x_{+a+} + y_{a+}), \tag{5}$$

and

$$\widehat{\psi}_a = \begin{cases} (x_{+0+} + y_{0+} + z_0)/N, & \text{if } a = 0 \\ (x_{+a+} + y_{a+})/N, & \text{if } a = 1, 2, \ldots, k, \end{cases} \tag{6}$$

where $N = x_{+++} + y_{++} + z_0$. Substituting (4), (5), and (6) into (3) gives

$$\widehat{\alpha}^*_{ca} = \sum_{d=0}^{k} \frac{x_{cad}}{\widehat{\pi}_{ad}N} \quad \text{for } a > 0,$$

and

$$\widehat{\beta}^*_{cd} = \sum_{a=1}^{k} \frac{x_{cad}}{\widehat{\pi}_{ad}N} + \frac{x_{c0d}}{f\,\widehat{\pi}_{0d}N}, \quad \text{for } d > 0, \tag{7}$$

where $\widehat{\pi}_{ad} = x_{+ad}/(x_{+ad} + y_{ad})$ and $f = (x_{+0+} + y_{0+})/(x_{+0+} + y_{0+} + z_0)$, which is fixed by design.

3.4 Inference

Based on the preceding discussion, our test statistic is

$$\widehat{\theta} = \log(\widehat{\text{Area}}_D/\widehat{\text{Area}}_A), \tag{8}$$

where the partial ROC areas are estimated by substituting (7) into (2). One approach to computing p values and confidence intervals is to assume that the data $\{x_{cad}, y_{ad}\}$ follow a multinomial distribution. Using the multinomial Poission transformation (Baker 1994) and the delta method, the asymptotic variance of $\widehat{\theta}$ is

$$\text{var}(\widehat{\theta}) = \sum_{c=0}^{1}\sum_{a=0}^{k}\sum_{d=0}^{k} \delta_{ad}\left(\frac{\partial\widehat{\theta}}{\partial x_{cad}}\right)^2 x_{cad} + \sum_{a=0}^{k}\sum_{d=0}^{k} \delta_{ad}\left(\frac{\partial\widehat{\theta}}{\partial y_{ad}}\right)^2 y_{ad}, \tag{9}$$

where $\delta_{ad} = 0$ if $a = d = 0$ and 1 otherwise, and the derivatives are computed symbolically. In the examples that we considered, results from (9) were similar to bootstrap calculations.

Because the classifications for each radiologist may be not be independent, there may be overdispersion. But overdispersion likely will have only a small effect on inference, because most of the repeated classifications by the same radiologists will occur in classifications related to no biopsy recommendation, which is not used in the computation of our test statistic. Nevertheless, to avoid making any distributional assumption, one could compute p values and confidence intervals using a permutation test (see Campbell 1994; Venkatraman and Begg 1996). A permutation test for these data could be based on the strong null hypothesis that in the interval of interest, the ROC curve for digital mammography is identical to the ROC curve for analog mammography. Because the interval is small, we think that this would be similar to a test of the weak null hypothesis that the partial ROC areas are identical. Under the strong null hypothesis, one could generate $\{x_{cad}, x_{cda}\}$ from Binomial$(.5, x_{cad} + x_{cda})$ and $\{y_{ad}, y_{da}\}$ from Binomial$(.5, y_{ad} + y_{da})$, and use the observed values for x_{caa} and y_{aa}. Substituting the simulated values into (7) and (8) gives the permutation distribution, which can be used to compute p values and confidence intervals.

4. SAMPLE SIZE

4.1 Formulas

For the partial testing design, we need to compute two sample sizes: $n(f)$, the number of analog mammograms, and $m(f)$,the number of digital mammograms, where f is the fraction of subjects selected for digital mammography following no biopsy recommendation for analog mammography. Let $\theta_{H_0}$and θ_{H_A}denote the expected values of the statistics under the null and alternative hypotheses. Also let $\sigma^2_{H_0}(f)$ and $\sigma^2_{H_A}(f)$ denote the variances under the null and alternative hypotheses when the sample size is 1. Because it is difficult to predict the amount of overdispersion, we assume that the counts follow a multinomial distribution, recognizing that overdispersion would inflate the sample size.

The sample size for the number of subjects undergoing analog mammography for a test based on $\widehat{\theta}$ with a two-sided type I error of .05 and a power of .9 is

$$n(f) = \frac{(1.96\sigma_{H_0}(f) + 1.64\sigma_{H_A}(f))^2}{(\theta_{H_0} - \theta_{H_A})^2}. \tag{10}$$

The sample size for the number who undergo digital mammography is the product of $n(f)$ and the probability of selection,

$$m(f) = n(f)\ (\omega + f(1-\omega)), \tag{11}$$

where $\omega = \Pr(A > 0) = \gamma_0\alpha_0 + \gamma_1\alpha_1$ is the probability of no biopsy recommendation following analog mammography. The total cost of the study equals a fixed cost that does not depend on sample size plus a variable cost, $c(f) = n(f) + m(f)\,r$, where r is the cost per subject of a digital mammogram divided by the cost per subject of an analog mammogram. Compared with a paired design with $f = 1$ and $n(1) = m(1)$, as f gets smaller, n increases because $\sigma_{H_0}(f)$ and $\sigma_{H_A}(f)$ increase, and m decreases because the fraction selected for digital mammography is approximately proportional to f. If digital mammography is more expensive per subject than analog mammography, then decreasing f reduces variable costs. The reduction in variable costs associated with a partial testing design relative to a paired design is $1 - c(f)/c(1)$.

4.2 Calculations

For computing the sample size, it is convenient to specify hypotheses in terms of the ratio of partial ROC areas, $\exp(\theta)$. To apply the sample size formula in (10), our null hypothesis is that the ratio of partial ROC areas equals 1. We considered two possible alternative hypotheses: a ratio of 1.6 and a ratio of 2.0. We obtained these alternative hypotheses by specifying a shape for the ROC curve for analog mammography and two possible shapes for the ROC curve associated with digital mammography (see Fig. 1). For sample size calculations based on a binormal model, Obuchowski and McClish (1997) also needed to specify the shape for the ROC curve under the alternative hypotheses.

The shapes were specified in the following manner. Based on Table 1, we anticipate that the false- and true-positive rates for analog mammography associated with a biopsy recommendation are $\alpha_0 = .010$ and $\alpha_1 = .80$. For digital mammography

Table 1. Parameter Estimates for Sample Size Calculations

False-positive rate associated with a biopsy recommendation for analog

The estimates are the fraction of subjects undergoing analog mammography who received an unnecessary biopsy—namely, a biopsy in which no cancer was detected.

α_0	Thurfjell, Lernevall, and Taube (1994)	Table 1	.005
	Sickles (1997)	Table 3	.015, .006
	Elmove et al. (1998)	Tables 1 and 3	0.13
	Sample size calculations		.010

True-positive rate associated with a biopsy recommendation for analog

The estimates are based on a mathematical model or on data collected after following subjects not biopsied.

α_1	Walter and Day (1983)	Table 2	.82, .99
	Alexander (1989)	Table V	.74
	Bird (1990)	Letter	.87
	Chen, Duffy, and Tabar (1996)	Table 2	1.00
	Kerlikowske, Grady, Barclay, Sickles, and Emster (1996)	Table 2	.94
	Kerlikowske et al. (1998)	Table 6	.75
	Sample size calculation		.80

Pr(biopsy recommendation for digital | biopsy recommendation for analog, no cancer; null hypothesis of the same performance for digital and analog)

The estimate comes from a study in which two radiologists read the same analog mammograms in subjects with benign breast cancer. The estimate is the fraction of times that the second radiologist recommended biopsy given that the first radiologist recommended biopsy, averaged over both radiologists.

q_0	Thurfjell et al. (1994)	Table 1	.78
	Sample size calculation		.78

Pr(biopsy recommendation for digital | biopsy recommendation for analog, cancer; null hypothesis of the same performance for digital and analog)

The estimate comes from a study in which two radiologists read the same analog mammograms in subjects with breast cancer. The estimate is the fraction of times that the second radiologist recommended biopsy given that the first radiologist recommended biopsy, averaged over both radiologists.

q_1	Bird, Wallace, and Yankaskas (1994)	Table 1	.99
	Thurfjell et al. (1994)	Table 1	.85
	Kerlikowske et al. (1998)	Table 6	.89
	Sample size calculation		.95

under the alternative hypotheses, we specify the false- and true-positive rates associated with a biopsy recommendation as $\beta_0 = .006$ and $\beta_1 = .8$. Let $u_{ca} = \Pr(A = a \mid A > 0,\ C = c)$ and $v_{cd} = \Pr(D = d \mid D > 0,\ C = c)$. Based roughly on Table 1 of Kerlikowske et al. (1998), we set $u_{01} = .75$, $u_{02} = .25$, $u_{03} = 0$, $u_{11} = .66$, $u_{12} = .22$, and $u_{13} = .12$. To create shape 1, we set $v_{cd} = u_{cd}$; to create shape 2, we set $v_{01} = .8$, $v_{02} = .1$, and $v_{03} = 1$.

To aid interpretation, we compute the average false- and true-positive rates in the relevant interval of the ROC curve (Table 2). This requires that we specify a prevalence. In the HIP study of analog mammography and breast self-examination, 54 cancers were detected among 20,000 initial screens (Baker and Chu 1990, table 1). A true-positive rate of .8 implies a prevalence of .003. The average true-positive rate for analog mammography in Table 2 is smaller than that in Table 1, because the average false-positive rate is also smaller.

To specify the variance in (10), we also need to specify a dependence between the classifications for analog and digital mammography. For simplicity, we write $\Pr(A = a,\ D = d \mid A > 0,\ D > 0,\ c) = u_{ca}v_{cd}$ and specify the dependence in terms of the false- and true-positive rates associated with a biopsy recommendation. Let $q_c = \Pr(D > 0 \mid A > 0, C = c)$ and $p_c = \Pr(D > 0 \,|_c A = 0, C = c)$. Because $\beta_c = \alpha_c q_c + (1 - \alpha_c)p_c$, we need only specify q_c along with α_c and β_c, and compute $p_c = (\beta_c - \alpha_c q_c)/(1 - \alpha_c)$. To ensure that $0 \le p_c \le 1$, we require that $\beta_c/\alpha_c \ge q_c \ge (\alpha_c + \beta_c - 1)/\alpha_c$. Based on Table 1, we specify $q_0 = .78$ and $q_1 = .95$ under the null hypothesis. This corresponds to correlations of .71 among subjects without cancer and .46 among subjects with cancer. For the alternative hypothesis, we assume less dependence and consider two cases that satisfy the constraint on q_c. In the low correlation case, $q_0 = .35$ and $q_1 = .85$, which correspond to correlations of .46 among subjects without cancer and .24 among subjects with cancer. In the high correlation case, $q_0 = .50$ and $q_1 = .95$, which correspond to correlations of .65 among subjects without cancer and .39 among subjects with cancer.

Anticipating that some subjects with rating category (1) will not undergo a biopsy, we set $\pi_{01} = \pi_{10} = .5$, $\pi_{11} = .75$, and $\pi_{ad} = 1$ for $a > 1$ or $d > 1$. Using these parameters, we write the expected counts as

$$x^*_{cad} = \begin{cases} N\,\gamma_c \alpha_c (1 - q_c)\, u_{ca} \pi_{a0} \\ \qquad \text{if } a = 1, 2, \ldots, k \text{ and } d = 0 \\ N\,\gamma_c (1 - \alpha_c) p_c v_{cd} f \pi_{0d} \\ \qquad \text{if } a = 0 \text{ and } d = 0, 1, 2, \ldots, k \\ N \gamma_c \alpha_c q_c u_{ca} v_{cd} \pi_{ad} \\ \qquad \text{if } a = 1, 2, \ldots, k \text{ and } d = 1, 2, \ldots, k \end{cases} \tag{12}$$

and

$$y^*_{ad} = x^*_{+ad}\big(1 - \pi_{ad}\big)/\pi_{ad}. \tag{13}$$

To compute θ_{H_A} and θ_{H_0}, we substituted (12) and (13) into (7) and then (8). As a check, f cancels from (7), so θ_{H_A} does not depend on f, and θ_{H_0} equals 0. To compute $\sigma_{H_0}(f)$ and $\sigma_{H_A}(f)$, we substituted (12) and (13) into (9).

4.3 Results

The main determinant of sample size was the ratio of partial ROC areas under the alternative hypothesis. In contrast, the correlation had relatively little effect (Table 3). For the partial testing design, we found that for the shapes and correlations under investigation, $f = .2$ yielded a near-optimal reduction in variable costs relative to the variable costs for the paired design. The reduction in variable costs was substantial, 23%–39% when the ratio of the variable cost of digital mammography to the variable cost of analog mammography was 4, and 41%–53% when the ratio of variable costs was 10.

The variable cost associated with analog mammography should include the cost of recruitment and any cost associated with obtaining informed consent. Nevertheless, if most of the cost of performing analog mammography is borne by insurance, then cost ratios of 4–10 are reasonable. In this investigation, the largest sample sizes were $n = m = 22{,}000$ under a paired design and $n = 45{,}000$ and $m = 10{,}000$ under a partial testing design. Because the partial testing design involves more subjects than the paired design, the accrual period will likely be longer, which should be weighed against the potential cost savings.

Table 2. Average False- and True-Positive Rates for Sample Size Calculations

			Average value in interval of ROC curve			
			For analog mammography		For digital mammography	
Hypothesis	ROC curve for digital mammography	Ratio of partial ROC areas	False-positive rate	True-positive rate	False-positive rate	True-positive rate
Null		1.0	.014	.54	.014	.54
Alternative	Shape 1	1.6	.012	.40	.012	.63
	Shape 2	2.0	.012	.40	.012	.79

NOTE: The prevalence of breast cancer is set at .003.

Table 3. Sample Sizes

Alternative hypothesis			Paired design		Partial testing design			
			Number of mammograms		Number of mammograms		Reduction in variable costs if ratio of digital to analog costs equals	
ROC curve for digital mammography	Ratio of partial ROC areas	Correlation	Analog	Digital	Analog	Digital	4	10
Shape 1	1.6	Low	22,000	22,000	45,000	10,000	23%	41%
	1.6	High	17,000	17,000	29,000	6,000	39%	53%
Shape 2	2.0	Low	9,000	11,000	17,000	4,000	27%	44%
	2.0	High	7,000	12,000	12,000	3,000	37%	52%

NOTE: Under the partial testing design, all subjects recommended for biopsy after analog mammography and a random sample of 20% of subjects not recommended for biopsy after analog mammography are selected to undergo digital mammography. The null hypothesis is that the ratio of partial ROC areas is 1. The two-sided type 1 error is .05, the power is .9, and the prevalence of breast cancer is set at .003.

5. DISCUSSION

To compare the performance of digital and analog mammography in early detection of breast cancer, we make three recommendations for design and analysis. First, to reduce costs when digital mammography costs more than analog mammography, we propose the partial testing design. Second, to make the results more relevant for screening, we recommend computing the area under the part of the ROC curve corresponding to small false-positive rates. Third, to avoid the need to follow subjects not biopsied, we propose basing inference on the ratio of partial ROC areas. Although our example involved a test of the superiority of digital versus analog mammography, the methodology could also be used to test equivalence.

One possible criticism of our method its the reliance on the trapezoidal rule for computing partial ROC areas. If the true underlying curve were concave, then trapezoidal rule would underestimate the partial ROC area, because it assumes that a straight line connects adjacent points. The fundamental problem is that data are missing between observed cutpoints, so assumptions are necessary. As a sensitivity analysis, we also investigated the sample size when a concave curve rather than a straight line connects the points. In particular, instead of computing a trapezoid, which we split into a rectangular area plus a triangular area, we computed the same rectangular area but, to account for concavity, added an area 20% larger than that of the triangle. The resulting sample size was 18,000 instead of 22,000. The reason for the smaller sample size is that the increase in the area over that of triangle has a larger effect on the higher ROC curve, and hence the ratio of partial ROC areas is larger.

An alternative to the trapezoidal approach is to fit a parametric model to a large part of each ROC curve and use the resulting parameter estimates to estimate the partial ROC areas. Even if rating category (0) were split into subcategories, there still would be a concern that a good fit to a large part of the ROC curve would not give a good fit in the small region of interest. Also, as discussed in Section 3.2, fitting a model to the part of the ROC curve to the right of the false-positive rate associated with a biopsy recommendation would require follow-up data.

APPENDIX: LIKELIHOOD

In formulating the likelihood, we introduce random variables S and B, such that $S = 1$ if a subject is selected for digital mammography and 0 otherwise and $B = 1$ if a subject is biopsied and 0 otherwise. By design, the probability of selection for digital mammography is

$$f_a = \Pr(S = 1 \mid A = a) = \begin{cases} f & \text{if } a = 0 \\ 1 & \text{if } a = 1, 2, \ldots, k. \end{cases} \tag{A.1}$$

Using this notation, the probability of undergoing a recommended biopsy can be written as

$$\pi_{ad} = \Pr\bigl(B = 1 \mid S = 1, A = a, D = d\bigr), \tag{A.2}$$

where $\pi_{00} = 0$ because subjects classified as (0) on both analog and digital mammography do not receive a biopsy. We set $S = 1$, because subjects not selected for digital mammography are not biopsied. The probabilities of biopsy and selection for digital mammography, no biopsy and selection for digital mammography, and no selection for digital mammography are

$$\Pr\bigl(B = 1, S = 1, C = c, A = a, D = d\bigr) = \pi_{ad} phi_{c\mid ad} \lambda_{d\mid a} \psi_a f_a, \tag{A.3}$$

$$\Pr(B = 0, S = 1, A = a, D = d) = (1 - \pi_{ad}) \lambda_{d\mid a} \psi_a f_a, \tag{A.4}$$

$$\Pr(S = 0, A = 0) = (1 - f)\psi_0. \tag{A.5}$$

Thus likelihood kernel is

$$L = \prod_{c=0}^{1} \prod_{a=0}^{k} \prod_{d=0}^{k} (\pi_{ad} \phi_{c\mid ad} \lambda_{d\mid a} \psi_a f_a)^{x_{cad}} \prod_{a=0}^{k} \prod_{d=0}^{k} \bigl((1 - \pi_{ad}) \lambda_{d\mid a} \psi_a f_a\bigr)^{y_{ad}} ((1 - f)\psi_0)^{z_0}.$$

[Received August 1999. Revised November 2000.]

REFERENCES

Alexander, F. E. (1989), “Estimation of Sojourn Time Distributions and False Negative Rates in Screening Programmes Which use Two Modalities,” *Statistics in Medicine*, 8, 743–755.

American College of Radiology (1995), *Breast Imaging Reporting and Data System (BI-RADS)* (2nd ed.), Reston, VA: American College of Radiology.

Anderson, L. M., and May, D. S. (1995), “Has the Use of Cervical, Breast, and Colorectal Cancer Screening Increased in the United States?” *American Journal of Public Health*, 85, 840–842.

Baker, S. G. (1994), "The Multinomial-Poisson Transformation," *The Statistician*, 43, 495–504.

———. (1995), "Evaluating Multiple Diagnostic Test With Partial Verification," *Biometrics*, 51, 330–337.

———. (2000), "Identifying Combinations of Cancer Biomarkers for Further Study as Triggers of Early Intervention," *Biometrics*, 56, 1082–1087.

Baker, S. G., and Chu, K. C. (1990), "Evaluating Screening for the Early Detection and Treatment of Cancer Without Using a Randomized Control Group," *Journal of the American Statistical Association*, 85, 321–327.

Baker, S. G., Connor, R. J., and Kessler, L. (1998), "The Partial Testing Design: A Less Costly Way to Test Equivalence for Sensitivity and Specificity," *Statistics in Medicine*, 17, 2219–2232.

Bamber, D. (1975), "The Area Above the Ordinal Dominance Graph and the Area Below the Receiver Operating Characteristic Graph," *Journal of Mathematical Psychology*, 12, 387–415.

Bird, R. E. (1990), "Professional Quality Assurance for Mammography Screening Program," (letter), *Radiology*, 177, 587.

Bird, R. E., Wallace, R. W., and Yankaskas, B. C. (1992), "Analysis of Cancers Missed at Screening Mammography," *Radiology*, 184, 613–617.

Campbell, G. (1994), "Advances in Statistical Methodology for the Evaluation of Diagnostic and Laboratory Tests," *Statistics in Medicine*, 13, 499–508.

Chen, H. H., Duffy, S. W., and Tabar, L. (1996), "A Markov Chain Method to Estimate the Tumour Progression Rate from Preclinical to Clinical Phase, Sensitivity and Positive Predictive Value for Mammography in Breast Cancer Screening," *The Statistician*, 45, 307–317.

Day, N. E. (1985), "Estimating the Sensitivity of a Screening Test," *Journal of Epidemiology and Community Health*, 39, 364–366.

Elmore, J. G., Barton, M. B., Moceri, V. M., Polk, S., Arena, P. J., and Fletcher, S. W. (1998), "Ten-Year Risk of False Positive Screening Mammograms and Clinical Breast Examinations," *The New England Journal of Medicine*, 338, 1089–1096.

Hanley, J .A., and McNeil, B. J. (1983), "A Method of Comparing the Areas Under Receiver Operating Characteristic Curves Derived from the Same Cases," *Radiology*, 143, 29–36.

Kerlikowske, K., Grady, D., Rubin, S., Sandrock, C., and Ernster, V. (1995), "Efficacy of Screening Mammography. A Meta-Analysis," *Journal of the American Medical Association*, 273, 149–154.

Kerlikowske, K., Grady, D., Barclay, J., Sickles, E., and Emster, V. (1996), "Effect of Age, Breast Density and Family History on the Sensitivity of First Screening Mammography," *Journal of the American Medical Association*, 276, 33–38.

Kerlikowske, K., Grady, D., Barclay, J., Frankel, S. D., Ominsky, S. H., Sickles, E. A., and Ernster, V. (1998), "Variability and Accuracy in Mammographic Interpretation Using the American College of Radiology Breast Imaging Reporting and Data System," *Journal of the National Cancer Institute*, 90, 1801–1809.

Lillienfeld, A. M. (1974), "Some Limitations and Problems of Screening for Cancer," *Cancer*, 35, Supp., 1720–1724.

McClish, D. K. (1989), "Analyzing a Portion of the ROC Curve," *Medical Decision-Making*, 9, 190–195.

Metz, C. E. (1986), "ROC Methodology in Radiological Imaging," *Investigative Radiology*, 21, 720–733.

———. (1989), "Some Practical Issues of Experimental Design and Data Analysis in Radiological ROC Studies," *Investigative Radiology*, 24, 234–245.

Obuchowski, N. A., and McClish, D. K. (1997), "Sample Size Determination for Diagnostic Accuracy Studies Involving Binormal ROC Curve Indices," *Statistics in Medicine*, 16, 1529–1542.

Pepe, M. S., Urban, N. U., Rutter, C., and Longton, G. (1997), "Design of a Study to Improve Accuracy in Reading Mammograms," *Journal of Clinical Epidemiology*, 50, 1327–1338.

Schatzkin, A., Connor, R. J., Taylor, P. R., and Bunnag, B. (1987), "Comparing New and Old Screening Tests When a Reference Procedure Cannot be Performed on All Screenees. Example of Automated Cytometry for Early Detection of Cervical Cancer," *American Journal of Epidemiology*, 125, 672–678.

Sickles, E. A. (1997), "Breast Cancer Screening Outcomes in Women Age 40–49: Clinical Experience With Service Screening Using Mammography," *Monographs of the National Cancer Institute*, 22, 99–104.

Swets, J. A. (1988), " Measuring the Accuracy of Diagnostic Systems," *Science*, 240, 1285–1293.

Thompson, M. L., and Zucchini, W. (1989), "On the Statistical Analysis of ROC Curves," *Statistics in Medicine*, 8, 1277–1290.

Thurfjell, E. L., Lernevall, K. A., and Taube, A. A. S. (1994), "Benefit of Independent Double Reading in a Population-Based Mammography Screening Program," *Radiology*, 191, 241–244.

Venkatraman, E. S., and Begg, C. B. (1996), " Procedure for Comparing Receiver Operating Characteristic Curves from a Paired Experiment," *Biometrika*, 83, 835–848.

Walter, S. D., and Day, N. E. (1983), "Estimation of the Duration of a Pre-Clinical Disease State Using Screening Data," *American Journal of Epidemiology*, 118, 865–886.

Winfield, D., Silbger, M., Brown, G. S., Clarke, L., Dwyer, S., Yaffer, M., and Shtern, F. (1994), "Technology Transfer in Digital Mammography. Report of the Joint National Cancer Institute–National Aeronautics and Space Administrations Workshop of May 19–20, 1993," *Investigative Radiology*, 29, 507–515.

Zhou, X. H. (1998), "Comparing Correlated Areas Under the ROC Curves of Two Diagnostic Tests in the Presence of Verification Bias," *Biometrics*, 54, 453–470.