SUPPLEMENTARY MATERIALS

A Diagnostic Breast Cancer Biomarker Study: Application of the PRoBE Design Principles

A diagnostic marker is used in people with signs or symptoms to aid in assessing if they have a condition. For example, women with breast lesions that are suspicious for cancer typically undergo diagnostic biopsy procedures. Yet most lesions prove to be negative for cancer. The Early Detection Research Network seeks to develop diagnostic biomarkers to reduce the number of these unnecessary biopsy examinations without reducing the number of women with invasive cancer that receive biopsy examinations. The motivating concept is that there exist biomarkers relating to specific unknown benign conditions and that these biomarkers could be used to identify which subsets of women with positive mammograms are most likely cancer-free. Alternatively, a marker may exist that is common to all invasive cancers but is absent in a subset of women without cancer. A collaborative effort is underway to create a repository of blood specimens for evaluating such biomarkers. Briefly, blood specimens and clinical data are collected from women before they are subjected to biopsy examination. Later, by use of pathology reports that indicate the presence or absence of invasive cancer, the condition of interest, specimens are selected at random for inclusion in the reference set. We describe the PRoBE design of this study in detail below by use of the items pertaining to its four key aspects.

Clinical Context

All women undergoing an initial diagnostic biopsy of a breast lesion at any of four tertiary health centers are being enrolled in the study. Women are excluded if they are younger than 18 years of age, have a history of cancer (except basal or squamous cell skin carcinoma), or are pregnant or nursing. Blood is drawn preoperatively. Subjects complete a questionnaire administered by the study coordinator after blood is drawn. Demographic information and medical and family histories are obtained.

The pathologist's reading of the biopsy results is the primary outcome variable. Treatment and data regarding disease outcome are also collected.

The primary group of case patients is composed of women with invasive cancer. No stratification for disease stage or histology is used. A secondary group of case patients is composed of women with ductal carcinoma in situ. The primary control group is women with benign nonproliferative conditions. Two other control groups are included—women with benign proliferative breast conditions and asymptomatic women undergoing routine screening mammography who are assigned a Breast Imaging Reporting and Data System score of 1 or 2. This last group of normal women do not fit in the main prospective study and are not considered part of the PRoBE design. Their purpose is to give preliminary evidence for biomarker positivity rates in the population of healthy control subjects, which is information that could be useful in deciding whether or not to study the marker as a premammographic screen if it is found to perform well in this study. These women are not part of the target population for the current study and the setting for their

specimen collection differs from it. Moreover, their outcome is essentially known at the time of specimen collection because almost all women in the mammography screening clinic are cancer free. Therefore, comparisons of this group of control subjects with case patients will need to be interpreted cautiously.

The four centers are to contribute equal numbers of subjects from each case and control group. That is, case patients and control subjects are matched on study center. In addition, case patients and control subjects are frequency matched on age and race. Subjects are selected randomly from among those enrolled according to these specifications by the Early Detection Research Network data coordinating center. Once selected, serum and plasma samples will be sent to the central repository. The study has been approved by Internal Review Boards at each of the four participating institutions. All women enrolled in the cohort study provide written informed consent.

Performance Criteria

The priority for diagnostic biomarkers in women elected for biopsy examination is to maintain the detection of almost all women with invasive cancer. Therefore, a high true-positive rate for invasive breast cancer is required. We set the minimally acceptable true-positive rate at 98%. We have not set separate target levels for different subtypes of invasive cancer because all invasive cancers are serious. It would be minimally beneficial to reduce the number of unnecessary biopsy examinations among women with benign nonproliferative disease by 25%. This value was not based on a formal cost–benefit analysis but rather on an informal consensus among study investigators. Therefore, the maximally acceptable false-positive rate is 75% (ie, the minimally acceptable specificity

= 25%). Minimally acceptable performance criteria were not set for the secondary case and control groups because positivity rates in these subgroups were considered of secondary interest and no specific hypotheses were to be tested. Because there are no existing clinical factors or tests used to counsel women in this setting, the study is not comparative in nature.

The Biomarker Test

Blood (28 mL) is drawn from each subject preoperatively and collected in four 7-mL tubes, two red-top tubes for serum and 2 CPT (B-D) tubes for plasma and white blood cells. Blood is spun within 5 hours of collection. Red top tubes are centrifuged at 3000 x g for 10 minutes at 4°C, and serum is removed by pipetting. Serum (in four 1-mL aliquots) is stored locally at –80°C. The CPT are processed according to manufacturers instructions. Samples that are selected for the case–control set are sent to an Early Detection Research Network reference laboratory, divided into 200-µL aliquots, labeled with codes to preserve blinding for future biomarker assays, and sent to a long-term storage and distribution center that is maintained by the National Cancer Institute at Fredrick, Maryland.

Currently no specific biomarkers are proposed for evaluation. The reference set of specimens is simply being constructed for future evaluation of biomarkers. Labeling of specimen aliquots at the reference laboratory will ensure that future assays are blinded to outcome status and to all other patient-related information. Procedures for retrieving and processing specimens will be specified along with assay procedures by investigators in the future before biomarker evaluation begins.

Study Size

The key requirement in this study is to continue to detect almost all cancer (ie, we fix the true-positive rate TPR_0 at 98%). The minimally acceptable value for the corresponding false-positive rate, FPR_0 is 75%. We assume that a candidate biomarker is continuous and we require that there is a 90% chance that the study will yield a positive conclusion if the false-positive rate is as low as 0.50. This value implies that 50% of unnecessary biopsy examinations are avoided by use of the biomarker test (ie, the alternative false-positive rate FPR_1 is 0.50). In addition, we require that the biomarker threshold estimated from this study will, with high probability, ensure the actual true-positive rate associated with that threshold is at least 97%.

Because no pilot data are available, we calculate sample sizes under some assumptions. The calculations are detailed below. Samples from 300 case patients with invasive cancer and from 100 control subjects with benign non-proliferative disease will be stored. In addition, we will store samples for 100 subjects in each of the secondary case and control groups, described above. We anticipate that 25% of women undergoing biopsy procedures will turn out to have invasive cancer. Therefore, to accrue a total of 300 such women, we expect to enroll 300 women at each of the four study sites. Enrollment will stop when sufficient numbers of case patients and control subjects are enrolled.

We do not plan to terminate any future studies early that use this specimen repository. Most of the 300 case patients will be required to estimate the biomarker threshold that yields 98% true-positive rate, and the marker will be useful even if only

25% of control biomarker values lie below it. Therefore, it is unlikely that we will be able to rule out a marker on the basis of analyses of early data. Moreover, by analyzing specimens from all subjects, we leave open the possibility of combining the study marker with additional markers in the future. Note that such a combination would need further evaluation in a separate study because development of the combination would constitute a discovery exercise, which is outside of the PROBE design.

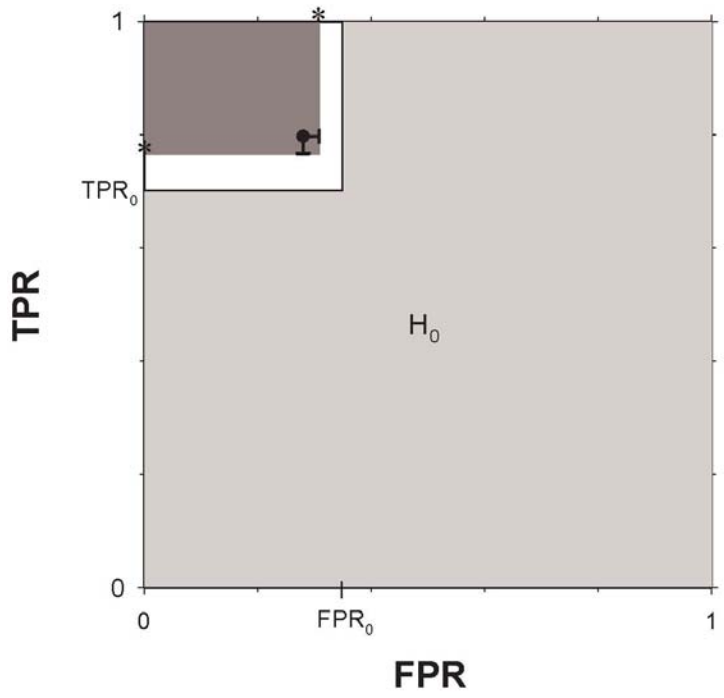
Sample Size Calculations

True-Positive and False-Positive rates for a Dichotomous Biomarker

The result of a dichotomous biomarker is either positive or negative. Recall that minimally acceptable values are set for the false-positive rate (FPR_0) and the true-positive rate (TPR_0). When data will become available from the study, the analysis will report joint confidence intervals for FPR and TPR with upper limit FPR_H for false-positive rate and lower limit TPR_L for true-positive rate. A positive conclusion will be drawn if

$$FPR_H \leq FPR_0 \text{ and } TPR_L \geq TPR_0.$$

An example of such a confidence region is shown with the hatched box in Supplementary Figure 1. A positive conclusion is drawn in that example because the confidence limits are well within the performance region in which the false-positive rate is below the minimally acceptable level of $FPR_0 = 0.35$ and true-positive rate is above the minimally acceptable level of $TPR_0 = 0.70$.



Supplementary Figure 1. A one-sided rectangular confidence region for the false-positive rate (FPR) and true-positive rate (TPR) of the exercise stress test calculated with data from the Coronary Artery Surgery Study (1). Performance meets the minimally acceptable criteria of FPR being at most 0.35 and TPR being at least 0.70. The points indicated with asterisks are the joint upper and lower 95% confidence limits for FPR and TPR, respectively. Values of FPR and TPR that lie in the shaded region, denoted by the null hypothesis, H_0, are not acceptable.

At the design stage, one must specify anticipated levels of performance for the alternative false-positive (FPR_1) and true-positive (TPR_1) rates. We require a high chance of drawing a positive conclusion (ie, the power is set at some level, $1-\beta$, such as 90%). To achieve that chance of drawing a positive conclusion one should choose sample sizes

$$n_{\text{cases}} = \frac{\left\{ Z^{1-\alpha^*}\sqrt{TPR_0(1-TPR_0)} + Z^{1-\beta^*}\sqrt{TPR_1(1-TPR_1)} \right\}^2}{(TPR_1 - TPR_0)^2}$$

and

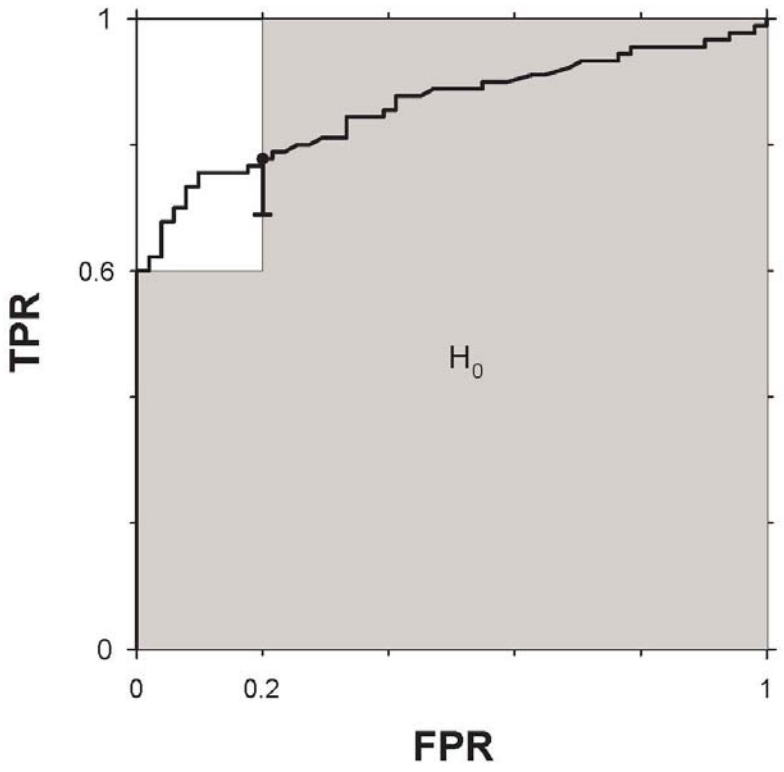
$$n_{\text{controls}} = \frac{\left\{ Z^{1-\alpha^*}\sqrt{FPR_0(1-FPR_0)} + Z^{1-\beta^*}\sqrt{FPR_1(1-FPR_1)} \right\}^2}{(FPR_1 - FPR_0)^2},$$

where $\sqrt{(1-\alpha)} = 1-\alpha^*$ and $\sqrt{(1-\beta)} = 1-\beta^*$. These formulas are based on large sample theoretical results as described previously (2). In practice, they only provide reasonable starting points for choices of sample sizes. One can generate data from simulated studies to see what the actual power will be to draw positive conclusions and adjust sample sizes accordingly. A simulation program is available from the Stata archive (3) (see the DABS website at `www.fhcrc.org/science/labs/pepe/dabs/software.html`).

Continuous Biomarker: False-Positive Rate Specified

An attribute of continuous biomarkers is that one can fix the false-positive rate (or true-positive rate) by choosing an appropriate biomarker threshold for defining a positive result. If one fixes the false-positive rate at a minimally acceptable value (FPR_0), then the task is to estimate the corresponding true-positive rate and to determine if it is at or above the minimally acceptable null value, TPR_0. That is, one calculates a confidence interval for the true-positive rate that corresponds to the false-positive rate (FPR_0), and then a positive conclusion is drawn if the lower limit is TPR_0 or greater. The true-positive rate corresponding to FPR_0 is otherwise known as the value of the receiver operating characteristic (ROC) curve at FPR_0, written as ROC (FPR_0). In the example shown in Supplementary Figure 2, FPR_0 is 0.2, the minimally acceptable true-positive rate of TPR_0

is 0.6, the marker conclusively detects more than 60% of case patients and the null hypothesis is rejected.



Supplementary Figure 2. Empirical receiver operating characteristic (ROC) curve for CA-19-9 as a biomarker for pancreatic cancer (4). The null hypothesis is that at the threshold corresponding to false-positive rate (FPR_0 = 0.2), the true-positive rate (TPR) does not exceed 0.6 for the CA 19-9 marker of pancreatic cancer. The lower 95% confidence limit for ROC(0.2) is shown. The shaded region denoted by H_0 displays the set of unacceptable false- and true-positive rates for the biomarker.

Now consider designing a study to determine if ROC(FPR_0)≥ TPR_0. Note that the width of the confidence interval depends on the variance of the estimated ROC(FPR_0), which in turn depends on the derivative (or slope) of the ROC curve, written as $r(FPR_0)$. If pilot data are available, an estimate of the slope can be calculated. If pilot data are not available, some assumptions need to be made for the purposes of sample size calculations (2). If after some unspecified transformation, biomarkers have normal distributions in case patients and in control subjects with variance (case biomarker)/variance(control biomarker) $= 1/b$, then the slope is calculated as:

$$r(FPR_0) = b\frac{\phi\{\Phi^{-1}(TPR_1)\}}{\phi\{\Phi^{-1}(FPR_0)\}},$$

where ϕ and Φ^{-1} are the standard normal density and quantile functions, respectively. If the case to control ratio is κ, the required number of case patients is

$$n_{\text{cases}} = \theta\frac{\{TPR_1(1-TPR_1) + \kappa r^2(FPR_0)FPR_0(1-FPR_0)\}}{(TPR_1 - TPR_0)^2},$$

where $\theta = \{\Phi^{-1}(1-\alpha) + \Phi^{-1}(1-\beta)\}^2$ depends on the confidence level, $(1-\alpha)$, and the power, $(1-\beta)$. The ratio of case patients to control subjects that minimizes the total number of samples, $n_{\text{cases}} + n_{\text{controls}}$, is as described previously (5),

$$\kappa = \frac{n_{\text{cases}}}{n_{\text{controls}}} = \frac{1}{r(FPR_0)}\sqrt{\frac{TPR_1(1-TPR_1)}{FPR_0(1-FPR_0)}}.$$

Data from the control subjects provide an estimate of the biomarker threshold that yields the minimally acceptable false-positive rate (FPR_0). We will also use the data to calculate a confidence interval for the actual false-positive rate associated with that threshold. To

be sure (with power $1-\beta$) that the actual false-positive rate associated with the threshold is no more than $FPR_0 + \varepsilon$, a control sample size of

$$n_{\text{controls}} \geq \left\{ \frac{\Phi^{-1}(1-\beta)}{\varepsilon} \right\}^2 FPR_0(1-FPR_0)$$

is required. We choose the size of the control sample so that it is at least as large as n_{cases}/κ and that it also satisfies the above inequality.

The sample size formulas presented are again approximations that are based on large sample statistical theory. Simulation studies are the best approach to calculating required sample sizes. Software on the DABS website will perform simulation studies in Stata (3). They can also provide information about how varying the case-control ratio from the optimal value affects power.

Continuous Biomarker: True-Positive Rate Specificied

In some studies one will specify the minimally acceptable true-positive rate, TPR_0, and the goal is to estimate the corresponding false-positive rate. The diagnostic biomarkers for breast cancer study provides an illustrative example. In this example, we provide the corresponding sample size formulas that are entirely analogous to those in the previous section. If an estimate of the slope of ROC^{-1} (TPR_0) is not available from pilot data it is calculated as:

$$r(1-TPR_0) = \left\{ b \frac{\phi(\Phi^{-1}[1-TPR_0])}{\phi(\Phi^{-1}[1-FPR_1])} \right\}^{-1}.$$

The required number of control subjects is

$$n_{\text{controls}} = \theta \frac{\{FPR_1(1-FPR_1) + \kappa^{-1} r^2 (1-TPR_0)(1-TPR_0)TPR_0\}}{(FPR_1 - FPR_0)^2}.$$

The ratio of case patients to control subjects that minimizes the total number of samples is shown by Janes and Pepe (3) to be

$$\kappa = \frac{n_{\text{cases}}}{n_{\text{controls}}} = \left\{ \frac{1}{r(1-TPR_0)} \sqrt{\frac{FPR_1(1-FPR_1)}{TPR_0(1-TPR_0)}} \right\}^{-1}.$$

Data from the case patients provide an estimate of the biomarker threshold that yields the minimally acceptable true positive rate TPR_0. To be sure (with power $1-\beta$) that the actual true-positive rate associated with the threshold is no less than $TPR_0 - \varepsilon$, the number of case patients should also satisfy

$$n_{\text{cases}} \geq \left\{ \frac{\Phi^{-1}(1-\beta)}{\varepsilon} \right\}^2 TPR_0(1-TPR_0).$$

Alternatively, if n_{cases} is fixed, we calculate the probability that the actual true-positive rate associated with $\widehat{\eta}$ the estimated threshold, $TPR(\widehat{\eta}) = \text{Prob}(Y_D > \widehat{\eta})$, is at least $TPR_0 - \varepsilon$,

$$\text{Prob}(TPR(\widehat{\eta}) > TPR_0 - \varepsilon) = \Phi \left\{ \frac{\varepsilon \sqrt{n_{\text{cases}}}}{\sqrt{TPR_0(1-TPR_0)}} \right\}.$$

Sample Size for the Diagnostic Breast Cancer Study

The key requirement in this study is to maintain detection of almost all cancer, at $TPR_0 = 98\%$. The minimally acceptable value for the corresponding false positive rate is $FPR_0 = 75\%$. We assume that a candidate biomarker will be continuous and require that there is a 90% chance the study will yield a positive conclusion if the false-positive rate

corresponding to $TPR_0 = 98\%$ is as low as 0.50; that is, if 50% of unnecessary biopsy examinations are avoided with the biomarker, then $FPR_1 = 0.50$. In addition, we require that the biomarker threshold estimated from this study will, with high probability, ensure the actual true-positive rate associated with that threshold is at least 97%.

Because no pilot data are available, we calculate sample sizes under some assumptions. If, after some unknown transformation, the biomarker is normally distributed in case patients and in control subjects with the same variance, then the slope of the anticipated ROC^{-1} curve is:

$$r(1 - TPR_0) = \left\{ \frac{\phi(\Phi^{-1}[1 - 0.98])}{\phi(\Phi^{-1}[1 - 0.50])} \right\} = \left(\frac{.0484}{.3987} \right)^{-1} = 8.24.$$

The one-sided confidence level $\alpha = 0.05$ and power $\beta = 0.90$ yields:

$$\theta = (1.64 + 1.28)^2 = 8.526.$$

We choose a case–control ratio of 3, which is close to the optimal value of 2.4, which minimizes the total number of samples assayed:

$$\kappa = \left\{ \frac{1}{8.24} \sqrt{\frac{0.5 \times 0.5}{0.02 \times 0.98}} \right\}^{-1} = 2.4.$$

This yields

$$n_{\text{controls}} = (8.526) \left\{ \frac{(0.5)(0.5) + (1/3)(8.24)^2(0.02)(0.98)}{(0.25)^2} \right\} = 94$$

and $n_{\text{cases}} = 3 \times 94 = 282$. Rounding up, we propose to include 300 invasive cancers for the primary case group and 100 benign proliferative control subjects for the primary control group. We note that the biomarker threshold recommended from this study will be the value corresponding to the second percentile in case patients, we expect that 98% of subjects in the target clinical population with cancer will then continue to be subjected

to biopsy examination (ie, the true-positive rate = 98%). Because this threshold will be based on 300 case patients, we calculate that with 89% certainty the actual population true-positive rate associated with the study threshold will be at least 97%.

Remarks

We do not detail sample size calculations for comparative studies in this section. They have been described previously (2). Correlations between biomarkers enter into these calculations and have a large impact on them. Pilot data are therefore highly desirable. In the absence of pilot data, the prudent approach is to assume that biomarkers are statistically independent and the results provided in this section are easily adapted.

References

1. Weiner DA, Ryan TJ, McCabe CH, Kennedy JW, Schloss M, Tristani F, et al. Exercise stress testing. Correlations among history of angina, ST-segment response and prevalence of coronary-artery disease in the Coronary Artery Surgery Study (CASS). N Engl J Med 1979;301(5):230-5.
2. Pepe MS. The Statistical Evaluation of Medical Tests for Classification and Prediction: Oxford University Press; 2003.
3. Stata Statistical Software Release 10.0. In. College Station, TX: StataCorp 2007.
4. Wieand S, Gail MH, James BR, James KL. A family of nonparametric statistics for comparing diagnostic markers with paired or unpaired data. Biometrika 1989;76(3):585-592.
5. Janes H, Pepe M. The optimal ratio of cases to controls for estimating the classification accuracy of a biomarker. Biostatistics 2006;7(3):456-68.