Generating Longitudinal Screening Algorithms Using Novel Biomarkers for Disease

Martin W. McIntosh,2 Nicole Urban, and Beth Karlan
Division of Public Health Sciences, Fred Hutchinson Cancer Research Center, Seattle, Washington 98109-1024 [M. W. M., N. S.], and Division of Gynecologic Oncology, Cedars-Sinai Medical Center and University of California Los Angeles School of Medicine, Los Angeles, California 90048 [B. K.]

Abstract
Recent advances in molecular technology are leading to the discovery of new tumor biomarkers that may be useful for cancer screening and early diagnosis. Translating a potential screening biomarker from the laboratory to its use in patient care may require an algorithm or screening rule for its application. An algorithm that can detect the smallest deviation from a defined norm is likely to achieve the highest sensitivity, but any practical screening algorithm must do so with subject controls on test specificity to avoid false-positive results, and unnecessary patient alarm and risk. Longitudinal algorithms that make use of previous tumor marker values and trends are likely to obtain improvements over single threshold rules. Thus far, a few longitudinal screening algorithms have been proposed (e.g., using serial prostate-specific antigen values for the detection of prostate cancer and serial CA125 values for the detection of ovarian cancer), but these algorithms are not appropriate for novel tumor marker discoveries, because they rely on unverifiable assumptions that may not translate to the behavior of the new marker. The algorithm presented here is motivated by: (a) the need to develop an algorithm for early detection using novel markers; (b) the practical demands on data and specimen availability; and (c) the need to be robust enough to accommodate a wide range of tumor growth behavior. We use Parametric Empirical Bayes statistical theory to model the trajectory of markers over time in a cohort of asymptomatic healthy subjects, and use the estimated trajectory to produce person-specific thresholds that depend on the screening history of each person. The thresholds are chosen to give the person (or population) a specified false-positive rate. The resulting algorithm is simple and can be represented in a simple graph or a chart. The statistical analysis needed to generate the algorithm can be found in nearly every basic statistical package. The algorithm is highly robust and can detect a wide range of tumor behaviors. The Parametric Empirical Bayes screening algorithm should take a central role when evaluating marker discoveries for use in screening. The algorithm is particularly useful when screening with a new marker of which the behavior in the preclinical period is not well known.

Introduction
The most common tumor marker screening algorithm is the ST3 rule, which gives a positive (likely to have disease) outcome for any subject of which the marker concentration deviates from a common population-wide threshold. However, when tumor marker behavior is heterogeneous across individuals, a better approach is to use a longitudinal algorithm to tailor the threshold to the individual screening history. Complex longitudinal algorithms, such as those using PSA values to detect prostate cancer (1–3) or CA125 values to diagnose ovarian cancer (4), have been proposed, but those algorithms are not appropriate for general use, for new tumor marker discoveries, or for other diseases. These algorithms are computationally intensive and require substantial modeling of the tumor marker growth during the preclinical period. When modeled adequately, these algorithms yield high sensitivity, but the specimens needed to observe and define the preclinical trajectory are so rare that these approaches are impractical for general use. The marker trajectory can be observed only by having access to several serially collected specimens taken from several cases during their preclinical period. Such observations are extremely rare for many diseases, and not even stored blood specimens collected as part of large longitudinal studies may be sufficient. For example, the ovarian cancer screening study by Jacobs et al. (5) randomized half of nearly 22,000 postmenopausal women to three annual primary screens with CA125 and secondary screen with ultrasound. Only 6 ovarian cancer cases were found during screening, with another 10 reported within a 7-year follow-up. However, most of the cases (the 10 found during follow-up) will not have had serum samples taken during their early preclinical period. Specimens from the cases found during screening were collected at annual intervals, not frequent enough to observe the early preclinical behavior of a disease such as ovarian cancer where the median preclinical duration is thought to be <18 months. Moreover, in diseases where stored specimens are sufficient, the storage conditions cannot anticipate the processing requirements of all of the future tumor markers.

Results.

1 The abbreviations used are: ST, single threshold; PSA, prostate-specific antigen; PEB, Parametric Empirical Bayes; SEER, Surveillance, Epidemiology, and End Results.
An alternative to those data- and computation-intensive algorithms is presented here. This algorithm is particularly appropriate for screening with new markers or in new populations. Our approach generates screening rules by focusing its attention on the trajectory of the markers in healthy women, for whom data are plentiful. The algorithm then determines if a general, rather than a specific (i.e., exponential or linear) change in marker behavior has occurred. The algorithm may be calibrated to work using any specificity, chosen either for the population or for each individual. The required specificity, along with the screening history of the subject, are used together to compute a person-specific threshold that is used to assess the next screening event. In this way the behavior of the marker is characterized and used to reflect the disease process or tumor growth and progression. Computing the threshold is very simple and involves computing the sample average of the subject screening history. Then, the screening threshold can be read off of a graph or looked up in a simple table. Our approach is based on PEB statistical methods, and we refer to our method simply as the PEB screening algorithm or simply the PEB rule. A full technical description of the algorithm derivation and theoretical properties is given by McIntosh and Urban (6), and is outlined below enough for the reader to understand its properties and implement it.

The PEB screening rule uses PEB statistical theory to infer the individual normal marker levels of each subject accounting for the population heterogeneity. The estimated normal levels of the individual are then used to produce a threshold tailored to his/her normal behavior. This normal behavior is revealed over time by event-free screening. The PEB screening rule has the surprising property of giving the vast majority of all of the women a threshold far lower than the single threshold rule but maintains the same population-wide specificity. Thus, in theory, cancers can be detected at an earlier stage without increasing the burden on the healthy population. In can be shown technically that the PEB rule uniformly outperforms the single threshold rule in that for any level of specificity, the PEB rule achieves a higher sensitivity (6).

The following two sections describe the behavior the PEB algorithm requires of a tumor marker and the data needed to implement the algorithm. We also contrast the PEB requirements with the existing longitudinal algorithms for CA125 and PSA mentioned above. We do this to point out that the PEB assumptions are typically less restrictive but also to point out that those other approaches may be considered when their more stringent requirements are met. The third section below describes how to estimate the parameters needed for the PEB algorithm and presents the algebraic expression to determine the threshold. An informal discussion of the technical aspects of the PEB algorithm follows. Finally, we discuss aspects of implementing the PEB rule that should be considered before applying the model to an actual population.

Materials and Methods
Marker Behavior Required of the PEB Screening Algorithm. A PEB screening algorithm can be generated for any marker meeting a modest set of requirements. Specifically: (a) marker concentrations in healthy subjects are continuous; (b) marker concentrations tend to either remain unchanged or grow in a known direction (elevate or decrease) with the onset of cancer or other pathologic/ disorder; (c) covariates (fixed and time varying) that describe large differences in marker behavior in healthy subjects are available and accounted for (see “Discussion”); and (d) the serial variation of the marker is not extreme over time in healthy subjects. These assumptions appear widely applicable to many tumor markers (7, 8). Importantly, note that no specific assumptions are made about the growth of a marker at cancer onset other than that it elevates (or declines) with cellular transformation.

The first assumption suggests that the marker concentrations, with some transformation, will have a normal distribution in healthy subjects. The second assumption allows us to know in which direction to look for deviation in the marker when disease develops. The final two assumptions allow us to model the trajectory over time for healthy women and without substantial mathematical modeling. Each of these assumptions can be relaxed for the general PEB rule but not for the simpler form presented by our report in detail. See McIntosh and Urban (6) for the most general presentation.

The complex longitudinal algorithms also require these assumptions, but in those applications they are even more specific and restrictive. Those algorithms require a specific (i.e., log) transformation to normality (restrictive form of 1), a constant linear marker growth after cancer onset (restrictive form of 2), and have the same restrictions with serial correlation. Moreover, the transformation to normality must be on the same scale that gives the marker a linear growth. These restrictions are violated, for example, with any marker that is normal on the raw scale and grows with an additive exponential model. It is also violated if the tumor growth model meets all of the assumptions, but the growth is not constant over the entire preclinical period. There is no evidence that the same uniform growth model should hold for every tumor type (9) much less for every potential marker, of which the growth may not follow the tumor volume exactly. However, if those requirements are met, and the data are available to estimate it, those complex algorithms may find higher sensitivity than the PEB algorithm presented here.

Specimen Requirements to Implement the PEB Algorithm. Novel tumor markers are typically evaluated by measuring their concentrations in stored specimens. Most commonly, the stored specimens are of two types: specimens from healthy control subjects and specimens from clinically presenting cases. A second easily obtained category of specimens is serially collected samples, measured in the same control individual at two time points, typically separated by several months. These serially measured samples from control subjects are necessary to implement the PEB algorithm. Sequential measurements on healthy controls, while not as abundant as cross-sectional specimens from healthy controls, can be obtained in substantial abundance over a short period of time. For technical and practical reasons, the duration between specimen collections should not be too frequent so that the serial correlation assumption above is not violated and not more frequent than the proposed screening interval. The duration between intervals is likely to be both marker- and disease-dependent, but a large set of markers for ovarian cancer have found that specimens taken at 1-month intervals were sufficiently spaced to satisfy these requirements.

The PEB algorithm can be additionally improved if individual characteristics that can predict the marker behavior in healthy subjects are known. For example, the ovarian cancer tumor marker CA125 is known to have substantial differences based on the menopausal status of a woman or the presence of benign gynecologic conditions. These conditions may not contribute to the ovarian cancer risk substantially, but they may cause statistically significant changes in the CA125 values. Characteristics believed to influence the behavior of the mark-
ers in healthy subjects should be recorded and evaluated before implementing the PEB algorithm. The algorithm presented below will assume it is estimated separately for women having similar values of such covariates, but extensions can be made to include covariates in the PEB rule formally (see Ref. 6). Omitting fixed covariates when estimating the PEB rule will not invalidate the algorithm, but it will be less efficient. However, time varying covariates that predict marker behavior must be accommodated (e.g., age for the prostate cancer marker PSA). These points hold for all of the screening algorithms not only the PEB approach. The precision of the PEB algorithm will depend on the number of serial controls available to estimate the parameters, but only a modest number are needed to implement the algorithm.

The specimen needs of the PEB algorithm are far less demanding than those required by the more complex longitudinal screening approaches. Those algorithms have great demands on the rarest type of specimen: serial observations in cancer cases obtained at high frequency during the preclinical period. Such specimens are extremely difficult to come by and typically suggest that implementing the algorithms for new markers or diseases would rely on untested assumptions. Thus, to use those methods for all but highly prevalent and slow growing diseases requires, like the early work by Skates (10, 11), a substantial amount of time to accumulate the information demanding than those required by the more complex longitudinal screening approaches. The specimen needs of the PEB algorithm are far less demanding than those required by the more complex longitudinal screening approaches.

How to Estimate the Parameters and Implement the PEB Algorithm. Presently, we assume that the marker concentration has been measured serially on a large number, N, of healthy control subjects. For clarity exposition we assume: (a) the control subjects have contributed exactly two marker observations; (b) all of the subjects are relatively homogeneous in regard to covariates that predict the normal marker levels (see section above); and (c) the markers are measured on a scale giving them a normal distribution in controls. The first and second assumptions are made only to reduce notation burden, and in practice are not needed, although relaxing them will require more detailed modeling (see Ref. 12 for details about how to measure these parameters in a more general setting and Ref. 6 on how to extend those to include time dependent and fixed covariates more formally). The final assumption (c) is necessary for our method to function properly, but it does not limit the generality of our method for any marker having continuous measurements, because then the existence of a transformation to normality is assured. Transformation can be done by some mathematical transformation such as log, or by simply translating the quantiles of the empirical marker distribution to that of the standard normal distribution.

The PEB algorithm requires that we examine the marker concentration mean, denoted \( \mu \), and decompose its variance into its total, within, and between components, denoted \( \sigma^2 \), \( \tau^2 \), respectively. Note that by definition \( V = \sigma^2 + \tau^2 \), and so in actuality we need to estimate only three parameters. If subject \( i \) donates two sequential marker concentrations we denote the first and second level with \( Y_{i1} \) and \( Y_{i2} \), respectively. Our parameters are defined formally as follows: \( \mu = \text{mean}[Y_i] \), the mean of all marker measurements, \( V = \text{Var}[Y_i] \), the variance of the markers in the population, \( \sigma^2 = \frac{1}{2} \text{Var}[Y_{i2} - Y_{i1}] \) the marker variance within a subject, and \( \tau^2 = V - \sigma^2 \), the between subject heterogeneity of the markers. Decomposing the marker variance into its within- and between-components allows us to characterize the relative subject heterogeneity in the population by the quantity \( B_i = \tau^2 / V \), which is often called the intraclass correlation (the reason for the subscript will become apparent below). Describing the most efficient manner to estimate these quantities is not in the scope of this article, but the expressions above do suggest simple ways to do so (see Ref. 12 for more efficient methods that can accommodate different numbers of observations per subject).

The PEB rule uses these quantities to determine the threshold for a subject from his/her screening history. Suppose a subject with \( n \) historical screens with an average concentration denoted by \( \bar{y}_n \) is about to undergo her next screen, and we wish the screen to operate at specificity \( \alpha \). The threshold for the next screen is given by the PEB algorithm to be equal to:

\[
\text{Threshold} = \mu + (\bar{y} - \mu)B_n + z_\alpha \sqrt{B_nV_n}
\]

where

\[
B_n = \frac{\tau^2}{\sigma^2/n + \tau^2}
\]

and \( z_\alpha \) is the \( \alpha \) quantile of a standard normal distribution (i.e., \( z_{0.05} = 1.96 \) when \( \alpha = 0.975 \)). Note that in the expression above, only the mean \( \bar{y} \) and shrinkage factor \( B_n \) need to be recomputed from time to time, because all of the other quantities remain the same throughout the screening study. An important special case is an initial screen of a subject, when \( n = 0 \). At the initial screen \( B_n = 0 \), and so the threshold becomes

\[
\text{Threshold} = \mu + z_\alpha \sqrt{V}, \text{ which equals the ST rule.}
\]

The PEB rule can also be represented in a graph or in a table, thus requiring that we only compute an average to determine the threshold. We demonstrate this below.

Results

Example: Using CA 125 to Screen a Population of High-risk Women. Here we demonstrate the PEB rule when using CA 125 to screen a population of women for ovarian cancer and implement the algorithm with a specificity of 98%. Several publications have found that log(CA 125), to the base e, has a highly normally distributed distribution (4, 8, 10–12), and so here we generate the PEB rule using log(CA 125). The behavior of log(CA 125) may depend heavily on the population and the type of assay used (13). Here we use the population investigated by Crump et al. (8), which is a population of women at high risk for ovarian cancer and screened as part of an ongoing screening program at the Gilda Radner Ovarian Cancer Detection Program at Cedars-Sinai Medical Center in Los Angeles, California (14). The following behavior of log(CA 125) has been estimated in that population: \( \mu = 2.27, V = 0.30, \tau^2 = 0.21, \sigma^2 = 0.09, \) and \( B_n = 0.21/0.30 = 0.68 \). We will demonstrate the algorithm when operated with 98% specificity; \( z_{0.05} = 2.05 \).

Fig. 1 represents the PEB screening rule graphically, and Table 1 shows the PEB rule in tabular form. For convenience, either can be used in place of the formal PEB expression above. Fig. 1 represents the screening rule using log (CA 125; left axis) and raw CA 125 (right axis). Note that the PEB rule is linear only on the transformed scale, and so we note that when summarizing the screening history mean of the log(CA 125) values must be used. This is not the same as first computing the average of the raw CA 125 then taking the log. Doing the latter will give a threshold that is higher than it should be.

The lines in Fig. 1 show the serum concentration thresholds leading to a positive screen for a woman based on her screening history. The horizontal axis in Fig. 1 represents the numeric average of her historic log(CA 125), and each line in the figure is used to determine the threshold based on the value
and the amount of screening history available for each woman. To determine the threshold for a woman undergoing screening, a clinician will: (a) compute the average of her previous log (CA 125) values; (b) locate the line in graph or the table column corresponding to the number of historical screens used in step 1; and (c) if using the table, find the appropriate row by identifying the row in column 1 that matches the average computed from step 1 or for the graph, find the point on that line directly above the value computed in step 1, and then read directly across to the vertical axis to find the threshold. For best accuracy, the PEB rule should be computed every time or represented in a table, like Table 1. The following example demonstrates how the PEB rule works on a hypothetical screening history of a woman.

Consider a woman whose first 3-year experience with screening found CA 125 concentrations equal to 8, 16, and 12, in her first, second, and third screen, respectively. The natural log of these concentrations is 2.1, 2.8, and 2.5, respectively. At the initial screen, when no screening history is available, the PEB rule is mathematically equivalent to the single threshold rule, and so a threshold of 2.27 \( \exp(3.39) \) on the raw scale; see column 2 of Table 1, and the horizontal line in Fig. 1. The first screen result of 2.1, or 8 on the raw scale, is used to determine the threshold to evaluate her second screen. The third column, row 21, of Table 1 shows the second screen is abnormal if it exceeds 2.98 or 20 on the raw scale. Note that this threshold is far lower than the single threshold rule used at the initial screen. The third screen uses the average of the first two screens 2.4 = (2.1 + 2.8)/2 to compute the next threshold. Table 1, row 24 column 4, shows that the third screen should use a threshold of 3.13 or 23 units/ml on the raw scale. At the fourth screen, where the average history equals 2.5 = (2.1 + 2.8 + 2.5)/3, Table 1 shows that a woman with a log(CA 125) exceeding 3.19 or 24 units/ml should get a positive screen. Table 1 will need enough columns to accommodate number of screens that will be accounted for in clinical practice, but the total number of columns can be limited, because the effect of screening history tails off over time. For example, in Table 1 we see that the screening rules change less drastically from 9 to 10 previous screens than from the initial to the first. The table may be truncated past the point when differences become inconsequential.

**Estimating the Performance of the PEB Algorithm in a Population.** A comprehensive ovarian cancer microsimulation model has been used to evaluate the behavior of the PEB algorithm in a large cohort of women in the United States (15). Our microsimulation model evaluates the performance of the PEB rule when used in a cohort of 1 million United States women who begin screening at age 50 and stop when they turn 85 years of age. The microsimulation model is explained in detail in Urban et al. (16), and described briefly below. These simulations are quite comprehensive. The life histories (cancer incidence, cancer death, treatment, and other causes of death) are modeled to reflect the United States population.

Urban et al. (15) used the microsimulation model and the PEB algorithm to estimate the performance of screening expected when using a screening protocol recommended by Skates et al. (11), who recommend two levels of positivity for a screen: positive screen for extreme elevations and early recall for modest ones. We used a false-positive rate of 2% to define the extreme and 15% for modest (i.e., 13% of healthy women were recalled early). With annual screening, this configuration of the PEB algorithm detects 70% of cancers before their clinical diagnosis, whereas the ST rule finds only 46% of them. Importantly, the PEB rule finds >50% of all cancers at early stage compared to 30% by the ST rule and 20% without screening. The PEB rule found an expected 31% drop in cause-specific mortality compared with 18% using the ST rule. This simulation study used \( B_1 = 0.6 \), which is lower than what is now reported in general risk populations with modern assays (12). Using a greater value would yield even greater improvement.

The predictions of Urban et al. (15) are very close to those of Skates and Pauler (4), who predict 60% of cancers detected at early stage using a complex longitudinal algorithm and more favorable assumptions. Skates and Pauler (4) assume a larger \( B_1 \), a longer duration of early stage disease, and recall 15% of
women early. Other aspects of their simulation performed are not described with enough detail to make our simulations comparable.

The simulations summarized above are specific to ovarian cancer and CA 125. Here we use a comprehensive ovarian cancer microsimulation model to evaluate the performance of screening. The simulation aged 1 million women from age 50 to death (nearly 1 in 64), and 77% of them died from their disease. The rows of Table 2 summarize their expected experience if they had instead been screened between ages 50 and 80 using the PEB algorithm. Note that the PEB algorithm used here is operating at a specificity of 98% and does not perform early recall. Thus, these results should not be compared to those of Skates and Pauler (4).

Row 1 reports the fraction of all cases that would have been detected earlier, row 2 measures the average number of years earlier that all cases would have been detected, row 3 represents the fraction of cases that would have been detected at an earlier stage, row 4 represents the absolute reduction (from 77% of cancer cases dying of ovarian cancer, and row 5 approximates the mean years extended for all cases. We emphasize that the performance in Table 2 averages the experience of all of the cases, not just those that were screen detected.

The performance of screening improves as $B_1$ increases and, indeed, shows that the PEB rule still achieves better performance than the ST rule even when $B_1$ is high (a relative 8% improvement). This leads to a substantial improvement in cancer mortality reductions depend primarily on a stage shift, this 42% find one with the PEB rule (a 20% increase). Because mortality reductions depend primarily on a stage shift, this leads to a substantial improvement in cancer mortality reduction from 19% to 26% (a 40% increase). We may also draw the

<table>
<thead>
<tr>
<th>y; mean of historical log(CA 125)</th>
<th>Screen</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1</td>
<td>3.39</td>
</tr>
<tr>
<td>0.2</td>
<td>3.39</td>
</tr>
<tr>
<td>0.3</td>
<td>3.39</td>
</tr>
<tr>
<td>0.4</td>
<td>3.39</td>
</tr>
<tr>
<td>0.5</td>
<td>3.39</td>
</tr>
<tr>
<td>0.6</td>
<td>3.39</td>
</tr>
<tr>
<td>0.7</td>
<td>3.39</td>
</tr>
<tr>
<td>0.8</td>
<td>3.39</td>
</tr>
<tr>
<td>0.9</td>
<td>3.39</td>
</tr>
<tr>
<td>1.0</td>
<td>3.39</td>
</tr>
<tr>
<td>1.1</td>
<td>3.39</td>
</tr>
<tr>
<td>1.2</td>
<td>3.39</td>
</tr>
<tr>
<td>1.3</td>
<td>3.39</td>
</tr>
<tr>
<td>1.4</td>
<td>3.39</td>
</tr>
<tr>
<td>1.5</td>
<td>3.39</td>
</tr>
<tr>
<td>1.6</td>
<td>3.39</td>
</tr>
<tr>
<td>1.7</td>
<td>3.39</td>
</tr>
<tr>
<td>1.8</td>
<td>3.39</td>
</tr>
<tr>
<td>1.9</td>
<td>3.39</td>
</tr>
<tr>
<td>2.0</td>
<td>3.39</td>
</tr>
<tr>
<td>2.1</td>
<td>3.39</td>
</tr>
<tr>
<td>2.2</td>
<td>3.39</td>
</tr>
<tr>
<td>2.3</td>
<td>3.39</td>
</tr>
<tr>
<td>2.4</td>
<td>3.39</td>
</tr>
<tr>
<td>2.5</td>
<td>3.39</td>
</tr>
<tr>
<td>2.6</td>
<td>3.39</td>
</tr>
<tr>
<td>2.7</td>
<td>3.39</td>
</tr>
<tr>
<td>2.8</td>
<td>3.39</td>
</tr>
<tr>
<td>2.9</td>
<td>3.39</td>
</tr>
<tr>
<td>3.0</td>
<td>3.39</td>
</tr>
<tr>
<td>3.1</td>
<td>3.39</td>
</tr>
<tr>
<td>3.2</td>
<td>3.39</td>
</tr>
<tr>
<td>3.3</td>
<td>3.39</td>
</tr>
<tr>
<td>3.4</td>
<td>3.39</td>
</tr>
<tr>
<td>3.5</td>
<td>3.39</td>
</tr>
</tbody>
</table>
Conclusion that the PEB algorithm is more cost effective, because the algorithm achieves its performance without increasing the number of screening events.

Summary of the Ovarian Cancer Microsimulation Model.
For ovarian cancer and CA 125, screening is done in a multimodal manner where CA 125 is the first stage of a screening procedure; a positive CA 125 screen is used to refer women to additional diagnostic work up using transvaginal sonography. All of the women are assigned age at clinical diagnosis with ovarian cancer, stage of the disease at clinical diagnosis, and age at death because of ovarian cancer to match United States experience (SEER cancer data). We also assign each subject an age at death attributable to other causes to match the United States Census Bureau life tables. The microsimulation model then infers a preclinical duration backward in time from the age of clinical diagnosis to its cancer onset. The preclinical duration is assigned a value that depends on the age and the cancer stage at the time of detection. The cancer marker is then assigned growth behavior over the preclinical period in a manner similar to Skates and Pauler et al. (4) and Pauler et al. (12), log (CA 125) has linear growth with random (person-specific) slope. The cancer is given four sequential stages over the preclinical period representing four progressive FIGO cancer stages, with time spent in each stage based on the opinions of experts. Thus, the observable quantities for women when not screened match the United States SEER registry and United States Census tables, and their unobservable quantities match currently understood behavior of the tumor biology.

An annual screening regimen is imposed on the life histories of these women. We calibrate the PEB algorithm to operate at false-positive rate = 0.01, and the transvaginal sonography component of the screen is assumed 100% sensitive for all tumors older than 6 months and 80% sensitive overall (17, 18). If detected earlier by screening, the simulation model uses the age and stage at the screen detection to reestablish the cause-specific life expectancy. The primary benefit of screening comes when a stage shift is achieved. Women diagnosed at an earlier stage are given extended cause-specific life expectancy, and women without a stage shift are not given any benefit from screening. If the new ovarian cancer age of death extends past her assigned other cause of death then she does not die of the disease, and cancer mortality is reduced. Thus, women can be unaffected by screening (if not detected or detected still at advanced stage), have life extended but still die of the disease (if new ovarian cancer death does not extend past other cause of death), or have life extended and die of another cause. Only in this latter case has ovarian cancer-specific mortality been reduced. The model compares the life histories under screening and without screening, and compares the rates of early detection, stage shifts, lead-time, and mortality effects.

Discussion
The usual method for screening with a tumor marker, the ST rule, gives all women the same population-wide threshold for positivity. This threshold is determined by finding a population-wide reference range, or quantile, and defining a positive screen when the marker exceeds that threshold. Using a α quantile gives a rule with that specificity. A marker transformed to a normal distribution has μ + zα/2σ as its threshold. However, this single reference range is inefficient for any marker that behaves heterogeneously in the population (i.e., for any marker having B1 > 0). Heterogeneity means that natural or normal marker concentrations vary between individuals, and thus an implication of heterogeneity is that even though the single threshold rule gives the population a specificity α, subject-specific specificity can vary; women with high (low) natural levels have specificity that is lower (higher) than α.

The ST rule accounts for heterogeneity and attempts to give each woman the same selected level of specificity. This is done by using PEB statistical methods to estimate the normal marker concentration of each person then determining the threshold defined by deviations from his/her normal levels, rather than population-wide deviations. Without showing the details, it can be shown (6) that the PEB rule: (a) always has higher sensitivity than the ST rule with the same specificity; and (b) has the same overall population-wide specificity as the ST rule.

PEB refers to class of statistical procedures for estimating a group of individual means when the individuals are drawn from a common population (19). Casella (20) and Efron and Morris (21) give introductory descriptions of PEB. The PEB estimator of the normal marker mean of a subject is given by μ + (ȳ − μ)Bn, where ȳ is the average of her historical screens. Note that the PEB estimator is part of the threshold expression above and that the PEB estimate is also a function of the shrinkage factor, Bn = τ2/(σ2 + τ2). The shrinkage factor ranges from zero (when n = 0) to one (when n is large), and the PEB estimator concomitantly ranges from population mean μ to the sample average ȳ of the individual. The PEB estimate comes into closer agreement with the individual history when history is plentiful, but with small n the PEB estimator is closer to the population average. Thus, the PEB estimator allows the individual history to carry a greater voice when that subject has a substantial history but anticipates regression to the mean for subjects with little screening history.
The PEB estimator has two properties that make it useful for screening: (a) it is unbiased for estimating an individual mean level; and (b) it is more precise (has a smaller variance) than the usual estimator does \( \bar{y} \). In particular, the variance of the PEB estimator is given by \( (\sigma^2/n)B_n \), which is always < \( \sigma^2/n \), the variance of \( \bar{y} \). The PEB estimator has added precision compared with the sample average, because it uses information shared among the group. This is often called “borrowing of strength” in meta-analysis. Combining these two properties with the variability of a new observation (\( \sigma^2 \)), the threshold expression above follows. We note here that generating screening rules using the sample average, \( \bar{y} \), instead of the PEB estimator can lead to screening rules that do worse than the single threshold rule and can never do better than the PEB screening rule (see Ref. 6 for derivation and additional discussion).

Two special cases of the PEB rule are when there is no screening history, \( n = 0 \) (and so \( B_n = 0 \)) and when screening history is plentiful, \( n = \infty \) (and so \( B_n = 1 \)). With the former, the PEB rule and the ST rule coincide, and with the latter the PEB threshold becomes \( \bar{y} + z_{0.025} \frac{\sigma}{\sqrt{n}} \). Note that the latter gives a person a threshold that looks like the ST rule except that it uses only the within woman variance, as if the threshold was determined by person-specific information only. Therefore, the PEB rule can detect marker elevations that are \( \sqrt{1 - B_n} \times 100\% \) the size that the ST rule needs (\( \sqrt{1 - \bar{B}} = \sqrt{\frac{\sigma^2}{n}} \)), the ratio of deviations of the PEB rule and the ST rule), and because the PEB rule can detect smaller elevations, it will have a higher sensitivity (6).

In between the two extremes, when screening history \( n \) is small to modest, the PEB rule is a compromise between the population-wide rule and the limiting individual rule. However, Fig. 1 shows that the performance of the limiting rule does not require a large screening history, and large benefits can follow after only a small number of screens. In Fig. 1 the second screening event has a dramatic adjustment to a woman’s threshold, with each subsequent screen providing comparatively smaller gains. Indeed, the screening rules at the fourth screens are practically equivalent to the limiting rule for the value of \( B \) used in our example.

The reason why the PEB rule is able to increase sensitivity without changing the population specificity can be seen by examining Fig. 1 more closely. Fig. 1 shows that the PEB thresholds (the lines) diverge from the single threshold rule (horizontal line), depending on the mean screening history; for our example, women with a mean screening history \( \bar{y} \approx 2.6 \), approximately, are given lower thresholds than the ST rule, and other women have their levels elevated from the ST rule. However, the average women has \( \bar{y} \approx 2.06 \), far under this level, and so the majority of women will have their threshold lowered if the PEB rule is used. Indeed at the second screen \( \sim 84\% \) of all of the women will have a lower threshold, and on subsequent screen that fraction increases to \( 88\%, 89\%, \) and \( 95\% \), \( n = 2, 3, \infty \), respectively. That is, \( >95\% \) of all women can have their cancer detected when at lower levels than a ST rule can. The reduction in the threshold can be dramatic, with average women receiving a cutoff of \( \sim 17 \) on the raw scale, nearly half the ST cutoff. The exact fraction that gets a lower threshold is marker dependent and is determined by the intra-class correlation \( B_i \); higher values imply more women with lower thresholds. However, for any value of \( B_i > 0 \), more than half of the subjects will be given lower thresholds.

Of course, Fig. 1 shows too that a small minority of women have higher thresholds with the PEB rule, and at first it may appear that the PEB rule may not benefit, and may even harm, these women. Additional consideration shows that these women benefit as well, but in a different way. The PEB rule works by controlling the specificity of the screen for each woman. Compare this with the ST rule, which controls only the population specificity. Overall, the PEB and ST rules will produce the same rate of false positives, but with the PEB rule each woman will share this burden equally. For ovarian cancer screening with the CA125 ST rule it is only a small fraction of all women, who are “unlucky” enough to have naturally high CA125 concentrations, that experiences the false positives. In our example it can be shown that \( 5\% \) of all women are responsible for \( >80\% \) of all false positives in the population. This burden is great enough that many women may forego screening because of it. The PEB rule automatically adjusts the thresholds so that women share a controlled chance of experiencing a false-positive test, and all share the same ability to detect comparable elevations. This switching from a population to an individual perspective for screening permits a dramatic increase in sensitivity without any increased burden on false-positive rates.

 McIntosh and Urban (6) investigate the conditions sufficient for the PEB algorithm to always meet or exceed the performance of the ST. Particular attention is paid to the performance in subjects after one or many false negative screens, in whom our algorithm at first appears to pose a potential problem; could a sequence of false negative screens, by biasing the screening history, mask the evidence of disease at the next screen and delay detection? McIntosh and Urban (6) show that the algorithm is very robust to false negatives and give conditions under which the PEB rule out performs the ST rule even for subjects who experience one or more false negatives. The conditions are quite general and verifiable for each marker; as long as \( B_i \) is not small or if the screening history \( \bar{y}_n \) contains one or more unbiased screens, then overall, subjects with false negatives will fare better with the PEB rule even for diseases that progress very slowly. See McIntosh and Urban (6) for specific details.

The PEB algorithm requires the markers to have a normal distribution. If no convenient mathematical transformation can achieve this then a nonparametric transformation can be made. The simplest way to do this is to tabulate the quantiles of the observed marker values then convert them to the corresponding quantile of a standard normal distribution. Thus, a marker at the 97.5% quantile in the test population would then be given a value of 1.96, and the median value would be converted to 0.0, and so forth. This gives the required behavior but may also require larger amounts of data so that the conversion is not too sparse. If formal inclusion of covariates is needed, this will require quantile regression methods and, thus, will add computational complexity.

The PEB rule presented here does not contain any covariates, but if they predict the normal marker levels of a population, then covariates must be accommodated by the screening rule. Fixed covariates, denoted \( X \), can be accommodated simply by replacing \( \mu \) in the PEB rule with a covariate adjusted mean \( \mu + X \beta \), but note now that the between-person heterogeneity \( \tau^2 \) is now a model-adjusted quantity. Pauker et al. (12) shows how to estimate these quantities specifically for a biomarker. Accommodation of time-dependent covariates (e.g., age at screen) may also be made but requires models beyond the scope of this manuscript. See McIntosh and Urban (6) for details.

The amount of reduction gained by an individual screen depends on the \( B_i \) of the marker, with smaller values implying smaller elevations are detectable, and more screens are needed to approach the limiting rule. This suggests strongly that a
marker value of $B_1$ is an important determinant of how the marker will perform when used in a cancer-screening program and, thus, should be reported as a routine part of marker assessment. McIntosh and Urban (6) give an example of a hypothetical marker having lower sensitivity than CA 125 in a ST rule, but its larger $B_1$ gives it greater sensitivity when used longitudinally.

The data requirements of the PEB method are practical for the data available and robust to a wide variety of marker behaviors. The specifics of the model are outlined in McIntosh and Urban (6) and should be consulted before applying this method to novel tumor markers. There are a few considerations that should be made before using the PEB rule in a population such as outlined in that report. Specifically, additional investigation should be made if the intraclass correlation $B_1$ is very small, the screening interval is very short compared to the preclinical period, and the marker grows at much less than a linear rate. If all of these conditions hold, the screening rule should be modified slightly, for example, by not using the PEB rule until at least two or three historical screens are available.

The report has omitted any consideration of the sample size needed to estimate the model parameters. This was done to improve clarity. Of course in practice larger numbers of samples will improve the precision of the estimates, but we make no recommendation regarding the number needed to evaluate the PEB algorithm other than to state that using it to investigate the potential performance of a marker needs less data than when implementing it in an actual population. When beginning a screening program we recommend initially evaluating the model parameters on stored sera, then continually evaluating the behavior of the markers in healthy subjects until a substantial number have accumulated to be confident that it can proceed without monitoring. As new data become available, the PEB algorithm (the transformation and the intraclass correlation estimates) may be updated as needed. Because the PEB rule needs only healthy subjects, limits on data availability will not take long to overcome and can allow more rapid development and evaluation of new tumor marker discoveries.

References

Generating Longitudinal Screening Algorithms Using Novel Biomarkers for Disease

Martin W. McIntosh, Nicole Urban and Beth Karlan

*Cancer Epidemiol Biomarkers Prev* 2002;11:159-166.

**Updated version**
Access the most recent version of this article at:
http://cebp.aacrjournals.org/content/11/2/159

**Cited articles**
This article cites 15 articles, 3 of which you can access for free at:
http://cebp.aacrjournals.org/content/11/2/159.full.html#ref-list-1

**Citing articles**
This article has been cited by 18 HighWire-hosted articles. Access the articles at:
/content/11/2/159.full.html#related-urls

**E-mail alerts**
Sign up to receive free email-alerts related to this article or journal.

**Reprints and Subscriptions**
To order reprints of this article or to subscribe to the journal, contact the AACR Publications Department at pubs@aacr.org.

**Permissions**
To request permission to re-use all or part of this article, contact the AACR Publications Department at permissions@aacr.org.