Optimal Cut-point and Its Corresponding Youden Index to Discriminate Individuals Using Pooled Blood Samples

Enrique F. Schisterman, Neil J. Perkins, Aiyi Liu, and Howard Bondell

Abstract: Costs can hamper the evaluation of the effectiveness of new biomarkers. Analysis of smaller numbers of pooled specimens has been shown to be a useful cost-cutting technique. The Youden index *(J)*, a function of sensitivity *(q)* and specificity *(p)*, is a commonly used measure of overall diagnostic effectiveness. More importantly, *J* is the maximum vertical distance or difference between the ROC curve and the diagonal or chance line; it occurs at the cut-point that optimizes the biomarker's differentiating ability when equal weight is given to sensitivity and specificity. Using the additive property of the gamma and normal distributions, we present a method to estimate the Youden index and the optimal cut-point, and extend its applications to pooled samples. We study the effect of pooling when only a fixed number of individuals are available for testing, and pooling is carried out to save on the number of assays. We measure loss of information by the change in root mean squared error of the estimates of the optimal cut-point and the Youden index, and we study the extent of this loss via a simulation study. In conclusion, pooling can result in a substantial cost reduction while preserving the effectiveness of estimators, especially when the pool size is not very large.

(*Epidemiology* 2005;16: 73–81)

The current emphasis on early detection and prevention of chronic and acute diseases has led to development of new and sophisticated biomarkers. However, costs of evaluation of their effectiveness can sometimes limit the feasibility of such testing. For example, the interleukin-6 biomarker of inflammation has been suggested to have potential discriminatory ability for myocardial infarction. However, the cost of a single assay can be so high (20 or more times greater than storage and technician costs) that financial considerations will hinder attempts to evaluate the usefulness of the biomarker. Analysis of results based on smaller numbers of pooled specimens has been shown to be a useful cost-cutting technique, especially with microarray experiments.[1–5]

By pooling (ie, physically combining individual specimens), the amount of information per assay is increased while the number of assays needed to evaluate this information decreases.[4,6] We assumed that measurement of the samples being pooled adequately represents the average of the individual unpooled sample. Conveniently, this is often the case, as most tests are expressed per unit of volume.

Receiver operating characteristics (ROC) curves are used in biomedical research to evaluate the effectiveness of biomarkers for distinguishing individuals with disease from those without.[7–10] The Youden index *(J)*, a function of sensitivity *(q)* and specificity *(p)*, is a commonly used measure of overall diagnostic effectiveness.[7–13] This index ranges between 0 and 1, with values close to 1 indicating that the biomarker's effectiveness is relatively large and values close to 0 indicating limited effectiveness. Figure 1 shows that *J* is the maximum vertical distance or difference between the ROC curve and the diagonal or chance line. *J* is defined by

$$J = \text{maximum } \{\text{sensitivity}(c) + \text{specificity}(c) - 1\} \qquad (1)$$

over all cut-points c, $-\infty < c < \infty$. If risk of disease is a monotonically increasing function of the marker level, sensitivity decreases and specificity increases with rising c. Thus, there is a penalty, decreased specificity for increasing sensitivity too far. *J* occurs at the optimal cut-point for calling a patient diseased, maximizing the number of correctly classified individuals.[11–14] On the other hand, the consequences of a positive or negative test result (ie, intervention) may be quite different and the loss from missing a case may be greater than from overcalling a control. Then, a differential

Supplemental material for this article is available with the online version of the Journal at www.epidem.com.

Submitted 12 September 2004; final version accepted September 21, 2004.

From the Division of Epidemiology, Statistics and Prevention Research, National Institute of Child Health and Human Development, National Institutes of Health, DHHS, Bethesda, Maryland; >Department of Mathematics and Statistics, American University, Washington, DC; ‡Department of Statistics, Rutgers University, Piscataway, New Jersey.

Supported by Intramural resources from NICHD.

Correspondence: Enrique F. Schisterman, Division of Epidemiology, Statistics and Prevention Research, National Institute of Child Health and Human Development (NICHD), 6100 Executive Blvd., Bethesda, MD 20852. E-mail: schistee@mail.nih.gov.

ISSN: 1044-3983/05/1601-0073
DOI: 10.1097/01.ede.0000147512.81966.ba

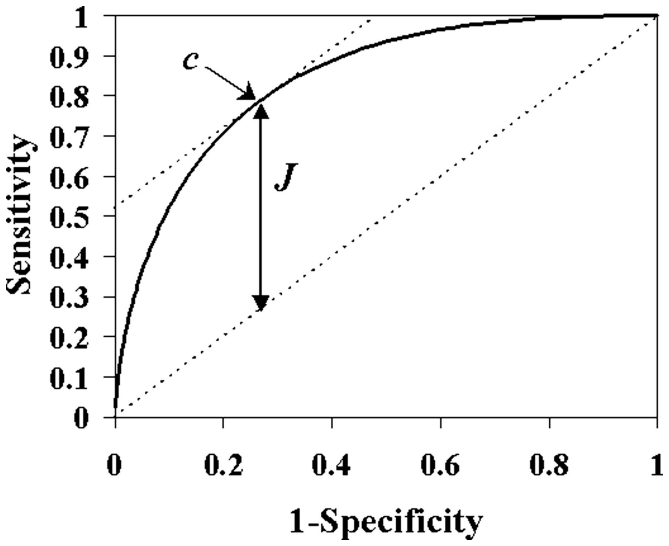

FIGURE 1. Receiver Operating Characteristic (ROC) curve of the Interleukin-6 data depicting Youden index *(J)* and optimal cut-point (*c*). Sensitivity *(q)* and Specificity *(p)* are both functions of some cut-point, with J = maximum $\{q(c) + p(c) - 1\}$ occurring at the optimum cut-point, *c*.

weighting is needed to optimize J (Appendix A.1, available with the electronic version of the article).

With equal weight given to errors of sensitivity and specificity, J can be determined graphically by plotting f_x and f_y, the probability density functions of the cases and controls, respectively, for a continuously distributed biomarker (Fig. 2). J is the difference of the area under f_x and f_y to the right of the cut-point, with negative area when $f_y > f_x$. This area is identical to the difference of the area under f_y and f_x to the left of the cut-point. With unequal weights, in a ratio R, J can be seen as the difference between f_y and Rf_x.

For diagnostic purposes and decision-making, health practitioners dichotomize continuous biomarkers into healthy and diseased patients. The optimal cut-point, the value used to separate these groups, and J occur at an intersection between the probability density functions of cases and controls such that

$$f_x(c) = f_y(c), \tag{2}$$

$$f_x(c + \varepsilon) > f_y(c + \varepsilon)$$

for some small $\epsilon > 0$. This is true when the mean of the cases is greater than that of the controls. The second criterion is necessary when multiple intersections exist (for the proof, see Appendix A.1). From now on, c will denote the optimal cut-point resulting in J.

Due to the high costs entailed by some biomarkers, several authors have proposed the use of pooling, and have evaluated ROC curve analysis when dealing with such data.[6,15] However, the effect of pooling on J and c has yet to be explored.

In this paper, we extend the work of Faraggi et al[6] and Liu and Schisterman[15] to evaluation of c and J under various distributional assumptions. We examine the effect of pooling on the efficiency of c and J estimation. In the circumstances when a fixed number of individuals are available for testing and samples are pooled to reduce the number of assays (thus lowering the costs), a loss of information is expected. We measure this loss of information by the change in root mean squared error (RMSE) of the estimate of c and J, and examine the extent of this loss via a simulation study.

We also examine the situation where the number of assays available is fixed and pooling is used to increase the information per assay. This procedure may improve the accuracy of the estimate and is specifically applicable in cases where assaying cost significantly exceeds the cost of obtaining samples.

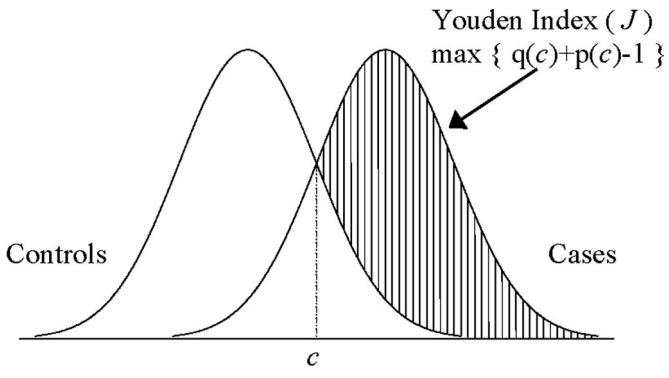

FIGURE 2. Graphical representation of the Youden index and the optimal cut-point (*c*) under normality and equal variance assumptions.

Inference on Youden Index and Optimal Cut-point Based on Pooled Data: Normal Assumptions

Assume that the responses of a specific biomarker are normally distributed, such that the cases (X), or true positives, have mean μ_x and variance σ_{x^2}, and the controls (Y), or true negatives, have mean μ_y and variance σ_{y^2}, and $\mu_x > \mu_y$. For $\mu_x < \mu_y$, one may simply switch the cases with controls in the following analysis. Under these assumptions, sensitivity ($q(c)$) and specificity ($p(c)$) can be written as

$$q(c) = P(X \geq c) = \Phi((\mu_x - c)/\sigma_x), \tag{3}$$

$$p(c) = P(Y \leq c) = \Phi((c - \mu_y)/\sigma_y), \tag{4}$$

for a given cut-point c, where Φ denotes the standard normal distribution function. Accordingly, test measurements falling

below c are negative results and those at or above c are positive.

As stated in the previous section, the optimal cut-point (and thus J) occurs at an intersection of the probability density functions of cases and controls. The number of intersections is a function of the variances of the cases and controls. One simple case is that of equal variance in cases and controls, $\sigma_{x^2} = \sigma_{y^2}$, where only one intersection exists and c is simply the midpoint between the means, $(\mu_y - \mu_x)/2$. In the case of unequal variance, the intersections can be found by the following quadratic equation:

$$c_{1,2} = \frac{\mu_y(b^2 - 1) - a \pm b\sqrt{a^2 + (b^2 - 1)\sigma_y^2 \ln(b^2)}}{b^2 - 1} \quad (5)$$

where $a = \mu_x - \mu_y$ and $b = \sigma_x/\sigma_y$. To find J, let us first order the intersections, $c_1 < c_2$. If $b > 1$, then J occurs at c_2; alternatively, if $b < 1$, then J occurs at c_1. Now, using Eq 3 for sensitivity and Eq 4 for specificity, J can be found from the appropriate cut-point.

When data on both cases and controls are available, appropriate estimates for μ_x, σ_x^2, μ_y, σ_y^2 can be calculated. These parameter estimates, when substituted into Eq 3, Eq 4, and Eq 5, yield the estimate $\hat{c}$ and subsequently $\hat{J}$.

Suppose that the data available are in the form of pooled samples, obtained as follows. First, individuals of similar disease status (ie, cases with cases, controls with controls) are randomly placed into groups of size g. Then, grouped individual specimens are combined as pooled samples and are tested as single observations. Assuming that the specimens are measured per unit of volume, a pool's measurement is then considered as the average of the member's measurements. This has been shown to be a reasonable assumption.[4]

Consider the instance where there are n and m pooled observations available of cases and controls, respectively, with groups of size g. Let XP_i, $i = 1,...,n$, denote cases, and YP_j, $j = 1,...,m$ denote controls such that

$$XP_i \sim N(\mu_{xp}, \sigma_{xp}^2)$$

$$YP_j \sim N(\mu_{yp}, \sigma_{yp}^2)$$

Consequently, from the additive property of the normal distribution, we have $\mu_{xp} = \mu_x$, $\sigma^2\text{xp} = \sigma^2\text{x}/g$, $\mu_{yp} = \mu_y$, $\sigma^2\text{yp} = \sigma^2\text{y}/g$.

Using the usual notation, let $\overline{xp}$, S_{xp}, $\overline{yp}$ and S_{yp} denote the standard estimates of μ_{xp}, σ_{xp}, μ_{yp}, and σ_{yp}, respectively. The parameters of the unpooled distributions can then be estimated accordingly by $\mu_x = \overline{xp}$, $\sigma_x = \sqrt{g}\, S_{xp}$, $\mu_y = \overline{yp}$ and $\sigma_y = \sqrt{g}S_{yp}$. Substituting these estimates for the parameters in the above equations yields the estimates $\hat{c}$ and $\hat{J}$ based on pooled data.

Inference on Youden Index and Optimal Cut-point Based on Pooled Data: Gamma Assumptions

The assumption of normality is often not justifiable in practice. Some biomarkers are skewed right and are best represented by some form of the gamma distribution. Suppose that the responses to a given biomarker follow a gamma distribution such that cases are $g\text{amma}(\alpha_X, \beta_X)$ and controls are $g\text{amma}(\alpha_Y, \beta_Y)$. Based on these distributional assumptions, sensitivity ($q(c)$) and specificity ($p(c)$) can be calculated as shown in Appendix A.2 (available with the electronic version of the article).

As with the normal case, J is realized at an intersection of the probability density functions. When case and control responses follow a gamma distribution, a single intersection frequently exists, the location of which defines the optimal cut-point c. Some special cases are

1. $\alpha_x = \alpha_y = \alpha$, then

$$c = \alpha \times \ln\left(\frac{\beta_x}{\beta_y}\right)\left(\frac{1}{\beta_y} - \frac{1}{\beta_x}\right)^{-1},$$

2. $\beta_x = \beta_y = \beta$, then

$$c = \beta \times \left(\frac{\Gamma(\alpha_x)}{\Gamma(\alpha_y)}\right)^{\frac{1}{\alpha_x - \alpha_y}}.$$

Otherwise, the intersection must be found numerically. When 2 intersections exist, c is located by the previous criteria (Eq 2). Now, J can be calculated at c by substituting Eq 6 and Eq 7 (Appendix A.2) for sensitivity and specificity in Eq 1.

Suppose again that pooled specimens are available with the pooling process being the same as defined earlier. Again, we let XP_i, $i = 1,...,n$ and YP_j, $j = 1,...,m$ denote the pooled observations of cases and controls, respectively. Using the additive property of the gamma distribution, we have

$$XP_i \sim Gamma(\alpha_{XP}, \beta_{XP}), \qquad i = 1,\ldots,n$$

and

$$YP_j \sim Gamma(\alpha_{YP}, \beta_{YP}), \qquad j = 1,\ldots,m,$$

each having a pooling size g. We will continue to assume that the measure of a pooled observation is the mean of the g unpooled measures. Consequently, $\alpha_{XP} = g\times\ \alpha_X$, $\alpha_{YP} = g\times\ \alpha_Y$, $\beta_{XP} = \beta_X/g$ and $\beta_{YP} = \beta_Y/g$.

The maximum likelihood estimates $\hat{\alpha}_{PX}$, $\hat{\beta}_{PX}$, $\hat{\alpha}_{PY}$ and $\hat{\beta}_{PY}$ can be obtained numerically from the observations on the pooled specimens. Using these estimators and the association between the distributions of pooled and unpooled observa-

tions, estimates of the unpooled distribution parameters can be obtained by $\hat{\alpha}_X = \hat{\alpha}_{PX}/g$, $\hat{\beta}_X = g\times \hat{\beta}_{PX}$, $\hat{\alpha}_Y = \hat{\alpha}_{PY}/g$, and $\hat{\beta}_Y = g\times \hat{\beta}_{PY}$. The estimates $\hat{c}$ and $\hat{J}$ can now be obtained by substituting these estimates for the parameters and following the steps outlined previously.

Simulation Study

To fully explore the effects of pooling on the estimates $\hat{c}$ and $\hat{J}$, we conducted simulation studies by generating data (cases and controls) from either normal or gamma distributions. Since $\hat{c}$ and $\hat{J}$ are a function of the intersection of the probability density functions for cases and controls, the parameters selected represent a wide variety of distributional conditions (normal and gamma) exemplified by different levels of separation ($J = 0.2, 0.4, 0.6, 0.8$). While simulations at all J levels are presented, analysis will focus primarily on J of 0.6 and 0.8, or the "useful", better diagnostic biomarker levels. Our simulations were limited to pooled size of 2 or 4 because pooling sizes of 5 and above result in a loss of identifiable skewness, due to the central limit theorem. A summary of our investigation is presented in Tables 1-4 We considered 2 common general conditions regarding availability of samples in an experimental setting.[6] The first involves fixing the number of study subjects ($N=M=$ 40,100,200), and the second fixes the number of assays ($n = m =$ 40,100,200). We generated 2000 individual samples from each set of parameters. Percent bias and relative root mean squared error (RMSE) were then determined by comparing estimates to the true c and J (calculated using the true parameter values) as follows:

$$\%\ Bias(\hat{c}) = \frac{\hat{c} - c}{c}\ 100\%$$

where $\hat{c}$ is the estimated optimal cut-point. %Bias($\hat{J}$) was calculated in the same manner; and

$$Relative\ \sqrt{MSE(\hat{c})} = \frac{\sqrt{MSE(g = 1,2,4)}}{\sqrt{MSE(g = 1)}}$$

where the $\sqrt{MSE(g = 1)}$ is the $\sqrt{MSE}$ for unpooled data, and $\sqrt{MSE(g = 2,4)}$ is $\sqrt{MSE}$ for pooled data of size 2 or 4.

The first condition, when the number of subjects available is fixed, looks at the degradation of the estimate $\hat{c}$ as pooling size increases ($g = 1,2,4$) resulting in a decrease in the number of tested samples—$n = N/g$. For instance, 40 control unpooled specimens are converted to 20-pooled specimens with each specimen consisting of a randomly chosen pair of controls ($g = 2$) or are converted to 10-pooled specimens with each specimen consisting of randomly chosen tetrad of controls ($g = 4$). The same procedure is applied to the case population.

Under normality assumptions (Table 1), the percent bias in the estimate of the optimal cut-point was negligible on all levels of discrimination and pooling, even for small sample sizes. As expected, the relative RMSE was inversely associated with the pooled size. No considerable distinction could be made between the RMSE from un-pooled data ($g = 1$) and pooled data ($g = 2$), $J = 0.6, 0.8$. However, for $g = 4$, the relative loss of efficiency is 3 times that of pairs. This is the effect of central limit theorem and is to be expected when cutting the sample by 75%.

In the gamma case (Table 2), the percent bias and relative RMSE increase in magnitude as g increases and, consequently, n and m decreases. The increase in bias due to pooling is negligible for all $J = 0.4$, 0.6, and 0.8. Relative RMSE increase for $g = 2$ are on par with that of the normal

TABLE 1. Bias, as Percent of *c*, and Root Mean Square Error (RMSE) Relative to Unpooled, $g = 1$, of the Optimal Cut-Point, as Pooling Increases and the Number of Specimens Available* Is Fixed—The Normal Case†

			N = M = 40			N = M = 100			N = M = 200		
J	c		g = 1	g = 2	g = 4	g = 1	g = 2	g = 4	g = 1	g = 2	g = 4
0.2	0.25	% Bias	−3.2	5.1	−2.8	2.8	3.9	0.4	−0.4	0.0	2.4
		Rel. RMSE	1.00	1.18	1.34	1.00	1.28	1.58	1.00	1.33	1.64
0.4	0.52	% Bias	0.6	−0.2	−0.4	0.2	0.2	−0.2	−0.2	0.6	−0.2
		Rel. RMSE	1.00	1.18	1.35	1.00	1.21	1.46	1.00	1.18	1.44
0.6	0.84	% Bias	0.1	0.8	0.0	−0.1	0.1	0.0	0.2	0.1	0.0
		Rel. RMSE	1.00	1.03	1.15	1.00	1.04	1.11	1.00	1.04	1.10
0.8	1.28	% Bias	0.4	0.2	0.1	−0.2	−0.2	0.2	−0.2	0.0	−0.2
		Rel. RMSE	1.00	1.08	1.34	1.00	1.07	1.21	1.00	1.04	1.17

*N and M are the number of cases and controls, n and m are the number of pooled samples of size g, such that n = N/g and m = M/g.
†Cases (X) are distributed normal (μ_x,1) and controls (Y) are normal (0,1). The levels of J are achieved by $\mu_x = 0.51$, 1.05, 1.68, and 2.56, respectively.

TABLE 2. Bias, as Percent of *c,* and Root Mean Square Error (RMSE) Relative to Unpooled, $g = 1$, of the Optimal Cut-Point, as Pooling Increases and the Number of Specimens Available* Is Fixed—The Gamma Case†

			N = M = 40			N = M = 100			N = M = 200		
J	**c**		**g = 1**	**g = 2**	**g = 4**	**g = 1**	**g = 2**	**g = 4**	**g = 1**	**g = 2**	**g = 4**
0.2	1.12	% Bias	8.0	15.3	25.0	4.2	7.9	12.8	2.2	3.8	7.3
		Rel. RMSE	1.00	1.37	1.74	1.00	1.27	1.69	1.00	1.29	1.78
0.4	1.79	% Bias	1.4	2.9	5.0	0.6	1.1	2.2	0.1	0.6	1.1
		Rel. RMSE	1.00	1.22	1.47	1.00	1.21	1.52	1.00	1.33	1.68
0.6	2.45	% Bias	0.0	0.5	1.9	0.0	0.1	0.9	−0.1	0.2	0.4
		Rel. RMSE	1.00	1.12	1.25	1.00	1.11	1.24	1.00	1.06	1.21
0.8	3.42	% Bias	0.3	0.4	1.5	−0.2	0.4	0.6	0.1	0.1	0.1
		Rel. RMSE	1.00	1.12	1.43	1.00	1.08	1.24	1.00	1.05	1.25

*N and M are the number of cases and controls, n and m are the number of pooled samples of size g, such that $n = N/g$ and $m = M/g$.
†Cases (X) are distributed gamma ($\alpha_x = 2.5$, β_x) and controls (Y) are gamma ($\alpha_y = 1.5$, $\beta_y = 1$). The levels of J are achieved by $\beta_x = 0.79$, 1.22, 1.97, and 3.82, respectively.

TABLE 3. Bias, as Percent of *c,* and Root Mean Square Error (RMSE) Relative to Unpooled, $g = 1$, of the Optimal Cut-Point, as Pooling Increases and the Number of Pooled Samples* Is Fixed—The Normal Case†

			n = m = 40			n = m = 100			n = m = 200		
J	**c**		**g = 1**	**g = 2**	**g = 4**	**g = 1**	**g = 2**	**g = 4**	**g = 1**	**g = 2**	**g = 4**
0.2	0.25	% Bias	−1.6	0.8	5.9	−1.2	−2.8	−1.6	0.8	−1.6	−0.4
		Rel. RMSE	1.00	0.89	0.84	1.00	0.92	0.88	1.00	0.96	0.93
0.4	0.52	% Bias	−1.0	0.2	−0.2	0.6	0.4	0.4	0.2	−0.2	0.4
		Rel. RMSE	1.00	0.89	0.76	1.00	0.85	0.78	1.00	0.86	0.77
0.6	0.84	% Bias	0.2	−0.2	0.1	0.0	−0.1	0.0	0.0	0.1	0.0
		Rel. RMSE	1.00	0.75	0.53	1.00	0.73	0.55	1.00	0.72	0.55
0.8	1.28	% Bias	−0.3	−0.2	0.01	−0.1	0.1	−0.1	0.1	−0.2	0.1
		Rel. RMSE	1.00	0.75	0.60	1.00	0.73	0.59	1.00	0.75	0.58

*n, m are the number of assays performed. N, M are the total number of cases and controls necessary for n and m pooled groups of size g, $N = n \times g$ and $M = m \times g$.
†Cases (X) are distributed normal (μ_x,1) and Controls (Y) normal (0,1). The levels of J are achieved by $\mu_x = 0.51$, 1.05, 1.68, and 2.56, respectively.

case, but $g = 4$ results are consistently 10% higher than the normal tetrads. The positive bias for both estimates based on the unpooled as well as on the pooled data greatly attenuates as sample size is increased. This is a result of using maximum likelihood estimators to estimate the optimal cut-point under small samples. Moreover, the bias is largely reduced for $J = 0.4$, 0.6, and 0.8 even for small sample size, which are actually the markers of scientific interest.

Biomarkers with poor distinguishing ability (eg, $J = 0.2$), also behave poorly under pooling. For example, when 40 unpooled samples are pooled in pairs, the RMSE increases by 37% for the gamma case. More generally, this relationship is true for both normal and gamma cases.

Under the second condition, when the number of assays to be performed is fixed, pooling effectively increases the overall sample size and the amount of information, via an increase in $N(N = n \cdot g)$ (Tables 3 and 4).

Again, bias remains unaffected, less than 1% bias for all levels of pooling for "useful" J. As pooling size increases, there is a consistent reduction in RMSE. For the normal case, as the level of pooling increases ($g = 1,2,4$), the RMSE for "useful" J substantially decreases (about half for pools of 4). Likewise, under gamma assumptions, as the level of pooling increases, $g = 1,2,4$, the benefits in RMSE are substantial (40% decrease for pools of 4). Pooling when $J = 0.2$ and 0.4 reveals a less dramatic benefit in RMSE.

These methods provide a useful tool for making inferences about unpooled samples when assays are based on pooled specimens. This is more clearly seen through use of an example, as illustrated below.

TABLE 4. Bias, as Percent of *c,* and Root Mean Square Error (RMSE) Relative to Unpooled, $g = 1$, of the Optimal Cut-Point, as Pooling Increases and the Number of Pooled Samples* Is Fixed—The Gamma Case†

			n = m = 40			n = m = 100			n = m = 200		
J	c		g = 1	g = 2	g = 4	g = 1	g = 2	g = 4	g = 1	g = 2	g = 4
0.2	1.12	% Bias	9.9	8.3	8.4	3.9	4.1	3.3	2.1	2.0	1.3
		Rel. RMSE	1.00	0.60	0.55	1.00	0.84	0.73	1.00	0.85	0.75
0.4	1.79	% Bias	0.8	1.8	1.7	−0.1	0.7	0.8	0.0	0.3	0.3
		Rel. RMSE	1.00	0.84	0.80	1.00	0.85	0.79	1.00	0.91	0.82
0.6	2.45	% Bias	0.3	0.3	0.7	−0.1	0.1	0.1	0.0	0.1	0.1
		Rel. RMSE	1.00	0.75	0.59	1.00	0.75	0.60	1.00	0.77	0.62
0.8	3.42	% Bias	−0.1	0.0	0.3	0.0	0.0	0.1	−0.1	0.0	0.1
		Rel. RMSE	1.00	0.76	0.61	1.00	0.73	0.58	1.00	0.76	0.58

*n, m are the number of assays performed. N, M are the total number of cases and controls necessary for n and m pooled groups of size g, $N = n \times g$ and $M = m \times g$.

†Cases (X) are distributed gamma ($\alpha_x = 2.5$, β_x) and Controls (Y) gamma ($\alpha_y = 1.5$, $\beta_y = 1$). The levels of J are achieved by $\beta_x = 0.79$, 1.22, 1.97, and 3.82, respectively.

Example

Evidence shows that inflammation may play a contributing role in the development of coronary heart disease (CHD). Interleukin-6 has been linked with the presence of infections in the vessel wall and with atherosclerosis.[16,17] Moreover, epidemiologic data show that infection in remote sites in the etiology of CHD.

Individual measurements of interleukin-6 on 80 volunteers were obtained at Cedars-Sinai Medical Center. Forty individuals who recently (within 2 weeks from the event) survived a myocardial infarction (MI) were defined as cases, after being confirmed by rest electrocardiogram (ECG) and laboratory measurements; the remaining 40 subjects served as controls. The controls had a normal rest ECG, were free of symptoms and had no previous cardiovascular procedures or MIs. In addition, the blood specimens were randomly pooled in groups of 2 and 4, for the cases and the controls separately, and remeasured. Faraggi et al[6] have shown, using the same data, that for interleukin-6 the assumption that the pooled sample measurements are the equivalent of the average of the individual cases is justified. Due to the costs involved such confirmatory evidence for the averaging assumption will generally not be available.

Distributional assumptions were also tested and found to fit well with gamma assumptions, confirming the findings of Faraggi and coauthors.[6] The mean (± SD) in the control and case unpooled samples, respectively, were 1.85 (±1.37) and 4.29 (±2.18). Youden index and cut-point were estimated using the method described previously under gamma assumptions. Table 5 shows that the Youden index was approximately 0.5 for unpooled and pooled data. More importantly, the optimal cut-point was estimated to be 2.41 for unpooled data and was not very much affected by pooling, as shown in Figure 1 and 3. A 95% bootstrapped confidence interval based on unpooled data was estimated to be 1.8 to 3.6, containing both estimates (2.06 [$g = 2$] and 2.70 [$g = 4$]) based on pooled data, despite the small number of specimens.

DISCUSSION

In this paper, we have presented a method to estimate the Youden index and the optimal cut-point and extended its applications to pooled samples. We extend the work of

TABLE 5. Interleukin-6 Example: Youden Index and Optimal Cut-Point for Individual and Pooled Data under Gamma Assumptions*

Method	Estimated Youden Index ($\hat{J}$)	Estimated Cut-point ($\hat{c}$)	Estimated Sensitivity	Estimated Specificity
Gamma assumption, individual	0.52 (0.38–0.67)	2.41 (1.77–3.58)	0.79 (0.65–0.93)	0.73 (0.58–0.85)
Gamma assumption, pooled $g = 2$	0.54 (0.40–0.75)	2.07 (1.71–4.06)	0.84 (0.64–0.96)	0.71 (0.47–0.86)
Gamma assumption, pooled $g = 4$	0.46 (0.30–0.76)	2.70 (1.65–4.38)	0.64 (0.26–0.85)	0.82 (0.39–0.97)

*Numbers in parenthesis are 95% confidence intervals. FPR and FNR can be easily derived by FPR = 1-Specificity and FNR = 1- Sensitivity.

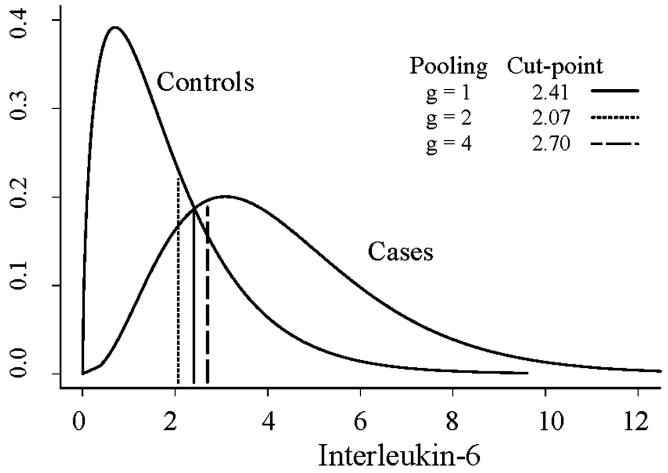

FIGURE 3. Optimal cut-point for Interleukin-6 under unpooled and pooled assumptions.

Faraggi et al[6] and Liu and Schisterman[15] to the cut-point, c, and Youden Index, J, under various distributional assumptions. We have shown that pooling is a statistically viable cost-saving approach, through a reduction in the number of assays required, especially with pool sizes of 2 and 4.

Most other statistical methods currently available for the analysis of biomarkers deal with comparison of proportions between cases and controls and power analysis, eg, for a genotype.[4,19] Our methods are specific for continuous data, where finding the optimal cut-point an important issue.

Relation Between Youden Index and The Likelihood Ratio

It is of interest to note that, since the Youden index of a continuous biomarker is a function of sensitivity and specificity, its relation to the likelihood ratio positive and negative may be useful. Graphically, the likelihood ratio positive (LR+) is the slope ($q/(1-p)$) of the line through the origin and a point on the ROC curve, while the likelihood ratio negative (LR−) is the slope ($(1-q)/p$) of the line through (1,1) and the same point on the ROC curve. The product of the likelihood ratios $[q(1-q)/p(1-p)]$ is the slope of the angle bisector. The Youden index, J, is the point at which the product of the two-likelihood ratio is equal to 1 or when the tangent to the ROC curve is parallel to the chance line (Fig. 1). Also, confidence intervals for c and J can be easily obtained using bootstrap methods and statistical software that is currently available.[18]

Pooling Assumptions

Correct implementation of the method developed in this paper requires assumptions, if the researcher sees only the pooled data. The first assumption is that the value obtained from a pooled assay can be considered to be the average of the individual values of the pooled specimens. There is both a biologic and a methodological aspect to this assumption. Biologically, this assumption can be deemed reasonable based on expert knowledge of the biomarker. If, for example, because of the molecular structure of the biomarker, pooling blood samples might yield a statistic other than the average (eg, maximum), then this methodology is inappropriate for the evaluation of the optimal cut-point and the Youden index. On the other hand, when this assumption is reasonable biologically, differences between the pooled sample and the average of individual specimens is due to "random measurement error," defined as the random variability that led to inaccuracy in the estimation of the true mean value. For instance, if the volume of the individual specimens to be pooled is not equal, the pooled sample will result in a weighted average of the volume per value of the biomarker. Therefore, for normally distributed biomarkers, the addition of mean zero measurement error and variance σ_ε^2 will affect the estimates of $\hat{c}$ one of 3 different ways depending on the ratio between σ_X^2/σ_Y^2. If $\sigma_X^2/\sigma_Y^2 = 1$, then $\hat{c}$ will remain unbiased, because the location where the 2 distributions intercept would remain unchanged (Fig. 2). If $\sigma_X^2/\sigma_Y^2 > 1$, then $\hat{c}$ will be positively biased and similarly if $\sigma_X^2/\sigma_Y^2 < 1$ then $\hat{c}$ will be negatively biased. For biomarkers that follow a gamma distribution, measurement error will always cause a positive bias in $\hat{c}$. This is due to the dependent relationship between the mean and variance of gamma distributions. Also, measurement error always results in an attenuation of $\hat{J}$. Since J is a measure of differentiation between cases and controls, it is intuitive that when error is introduced the ability to differentiate decreases.

Distributional Assumptions

The second assumption is that the unpooled biomarkers follow a known parametric distribution. A more formal evaluation of distributional assumption would be possible using a moment-based estimating-equation approach to deal with situations where likelihood functions based on pooled data are difficult to work with. We outlined the method to obtain estimates and test statistics of the parameters of interest in the general setting. We demonstrated the approach on the family of distributions generated by the Box-Cox transformation model, and, in the process, construct tests for goodness of fit based on the pooled data. Nevertheless, in our experience, the researcher will often develop some sense of both these assumptions during the early stages of the biomarker development by means of a validation study.

Pooling Size

Pooling sizes of 5 and above, while fiscally attractive, are prone to 2 difficulties. The first is a consequence of the central limit theorem; averages tend to be more normally distributed as sample size increases. Identifying a biomarker's un-pooled distribution is difficult because the central

limit theorem hinders our ability to distinguish between a skewed and a symmetric distribution. The second difficulty arises only when a fixed number of subjects are reduced to an unreasonably small sample size due to pooling and rendering the parameter estimation unreliable. For instance, in the example presented above, we had 40 cases and 40 controls contributing blood samples. If $g = 10$, then we are left with 8 assays (4 cases and 4 controls) on which to estimate the means and standard deviations necessary for $\hat{c}$ and $\hat{J}$.

Implications

This method is relevant to studies of markers for early detection and prevention of disease and for studies of markers of exposure and disease in molecular epidemiology when, for example, deciding whether a biomarker is worth pursuing further or is ready for a study. Furthermore, once this method is applied and a biomarker demonstrates discriminatory ability, the optimal cut-point can be used in clinical practice to classify patients as healthy or diseased, after proper validation.

In summary, we showed that estimating c and J under pooling is a cost-effective, statistically sound approach for evaluating biomarkers. Such estimation has potential applications for research and clinical practice and for hypothesis development.

REFERENCES

1. Farrington C. Estimating prevalence by group testing using generalized linear models. *Stat Med*. 1992;11:1591–1597.
2. Tu X, Litvak E, Pagano M. On the informativeness and accuracy of pooled testing in estimating prevalence of a rare disease: application to HIV screening. *Biometrika*. 1995;82:287–297.
3. Barcellos L, Klitz W, Field L, et al. Association mapping of disease loci, by use of a pooled DNA genomic screen. *Am J Hum Genet*. 1995;61: 734–747.
4. Weinberg CR, Umbach DM. Using pooled exposure assessment to improve efficiency in case-control studies. *Biometrics*. 1999;55:718–726.
5. Kemdziorski CM, Zhang Y, Lan H, Attie AD. The efficiency of pooling mRNA in micro array experiments. *Biostatistics*. 2003;4:465–477.
6. Faraggi D, Reiser B, Schisterman EF. ROC curve analysis for biomarkers based on pooled assessments. *Stat Med*. 2003;22:2515–2527.
7. Zou KH, Hall WJ, Shapiro DE. Smooth non-parametric receiver operating characteristic (ROC) curves for continuous diagnostic tests. *Stat Med*. 1997;16:2143–2156.
8. Zweig MH, Campbell G. Receiver operator characteristic (ROC) plots; a fundamental evaluation tool in clinical medicine. *Clin Chem*. 1993; 39:561–577.
9. Goddard MJ, Hinbery I. Receiver operator characteristic (ROC) curves and non-normal data: an empirical study. *Stat Med*. 1990;9:325–337.
10. Wieand S, Gail MH, James BR, James KL. A family of non-parametric statistics for comparing diagnostic markers with paired or unpaired data. *Biometrika*. 1989;76:585–592.
11. Youden WJ. An index for rating diagnostic tests. *Cancer*. 1950;3:32–35.
12. Barkan N. Statistical inference on r*specificity + sensitivity. Doctoral Dissertation 2001. Haifa University.
13. Bamber DC. The area above the ordinal dominance graph and the area below the receiver operating characteristic graph. *J Math Psychol*. 1975;12:387–415.
14. Hilden J, Glasziou P. Regret graphs, diagnostic uncertainty and Youden's Index. *Stat Med*. 1996;15:969–986.
15. Liu A, Schisterman EF. Comparison of diagnostic accuracy of biomarkers with pooled assessments. *Biom J*. 2003;45:631–644.
16. Chilton RJ. Recent discoveries in assessment of coronary heart disease: impact of vascular mechanisms on development of atherosclerosis. *J Am Osteopath Assoc*. 2001;101:S1–S5.
17. Yudkin JS, Kumari M, Humphries SE, Mohamed-Ali V. Inflammation, obesity, stress and coronary heart disease: is interleukin-6 the link? *Atherosclerosis*. 2000;148:209–214.
18. Carpenter J, Bithell J. Bootstrap confidence intervals: when, which, what? A practical guide for medical statisticians. *Stat Med*. 2000;19: 1141–1164.
19. Peng X, Wood CL, Blalock EM, et al. Statistical implications of pooling RNA samples for microarray experiments. *BMC Bioinformatics*. 2003; 24:4–26.

Appendix 1

Assume that cases, X, and controls, Y, are represented by continuous unimodal distributions, and $\mu_y < \mu_x$. Let c_0 be some cut-point and $c_i(I = 1,2)$ be the i^{th} intersection of the probability density functions denoted by f. Youden index (J) is found by

$$J = \max[q + p - 1]$$

$$= \max[P_x(X > c) + P_y(Y < c) - 1]$$

$$= \max[P_y(Y < c) - P_x(X < c)]$$

$$= \max\left[\int_{-\infty}^{c} (f_y(t) - f_x(t))dt\right]$$

The intervals for which $f_y > f_x$ and $f_y < f_x$ are determined by the variances of the distributions. Assuming $\sigma_{x^2} > \sigma_{y^2}$ could result in 1 or 2 intersections. The 2-intersection case follows

$$f_y < f_x \qquad (-\infty, c_1)$$

$$f_y > f_x \qquad (c_1, c_2)$$

$$f_y < f_x \qquad (c_2, \infty)$$

For c_0 in $(-\infty, c_1)$

$$\int_{-\infty}^{c_1} (f_y(t) - f_x(t))dt = \int_{-\infty}^{c_0} (f_y(t) - f_x(t))dt + \int_{c_0}^{c_1} (f_y(t) - f_x(t))dt$$

$$\Rightarrow \int_{-\infty}^{c_1} (f_y(t) - f_x(t))dt > \int_{-\infty}^{c_0} (f_y(t) - f_x(t))dt$$

Similarly, for c_0 in (c_1, c_2)

$$\int_{-\infty}^{c_2} (f_y(t) - f_x(t)dt > \int_{-\infty}^{c_0} (f_y(t) - f_x(t))dt$$

And, for c_0 in (c_2, ∞)

$$\int_{-\infty}^{c_2} (f_y(t) - f_x(t))dt > \int_{-\infty}^{c_0} (f_y(t) - f_x(t))dt$$

Therefore

$$J = \int_{-\infty}^{c_2} (f_y(t) - f_x(t))dt$$

A similar argument proves that when a single intersection exists, the intersection is the cut-point for *J*. For the case where $\sigma_{x^2} < \sigma_{y^2}$, this approach yields c_1 as the optimal cut-point used for *J*.

Note: Using Figure 1 as a reference, it can be seen that moving the cut-point to the right would result in a loss in shaded area (Youden index). Since Youden index can be represented by the area between the 2 curves to either the right or left of the cut-point, moving the cut-point to the left also result in a decrease.