The optimal ratio of cases to controls for estimating the classification accuracy of a biomarker

HOLLY JANES∗†
Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health,
615 North Wolfe Street, Baltimore, MD 21205, USA
hjanes@jhsph.edu

MARGARET PEPE
Department of Biostatistics, University of Washington, and Division of Public Health Sciences,
Fred Hutchinson Cancer Research Center, 1100 Fairview Avenue North,
Seattle, WA 98195, USA

SUMMARY
The case–control design is frequently used to study the discriminatory accuracy of a screening or diagnostic biomarker. Yet, the appropriate ratio in which to sample cases and controls has never been determined. It is common for researchers to sample equal numbers of cases and controls, a strategy that can be optimal for studies of association. However, considerations are quite different when the biomarker is to be used for classification. In this paper, we provide an expression for the optimal case–control ratio, when the accuracy of the biomarker is quantified by the receiver operating characteristic (ROC) curve. We show how it can be integrated with choosing the overall sample size to yield an efficient study design with specified power and type-I error. We also derive the optimal case–control ratios for estimating the area under the ROC curve and the area under part of the ROC curve. Our methods are applied to a study of a new marker for adenocarcinoma in patients with Barrett’s esophagus.

Keywords: Case–control design; Efficiency; Power; ROC curve; Sample size; Sensitivity; Specificity.

1. INTRODUCTION
Recent scientific and technological innovations have produced an abundance of potential screening and diagnostic biomarkers. Screening markers, such as those for various cancers and for cardiovascular disease, are being investigated with the goal of detecting disease at an early stage, when treatment is often more effective. Accurate diagnosis of disease, on the other hand, is a fundamental first step in treating symptomatic patients. Efforts are underway to develop diagnostic tests for recently defined conditions, and faster, more accurate tests for long-established diseases. Note that the tasks of disease screening

∗To whom correspondence should be addressed.
†Work conducted while at the University of Washington, Department of Biostatistics, Seattle, WA, USA.
and diagnosis are fundamentally the same: the goal is to classify individuals as either diseased or non-diseased.

We use the term ‘biomarker’ to refer to a generic medical test, biochemical or otherwise, for screening or diagnosis. Rigorous evaluation of biomarkers is essential, in order to guarantee that the tests that are developed are sufficiently accurate and beneficial to the patient. In this paper, we are concerned with the efficient design of studies to evaluate biomarkers.

The case–control study is commonly employed (Pepe et al., 2001). A fixed number of cases and controls are sampled in order to determine how well the biomarker distinguishes between these two groups. The appropriate ratio in which to sample cases and controls has never been established. It is common to simply sample equal numbers of cases and controls (Etzioni et al., 1999), or, when controls are easier to obtain, to sample more controls than cases (Janes et al., 2005). In this paper, we derive the optimal case–control ratio, which enables us to design a future study with optimal case–control allocation, thereby increasing efficiency and reducing the cost of the study.

We consider continuous biomarkers with diagnostic accuracy evaluated using the receiver operating characteristic (ROC) curve. The ROC curve plays a key role in the evaluation of biomarker accuracy (Baker, 2003; Hanley, 1989; Begg, 1991; Zhou et al., 2002; Pepe, 2003). It displays the trade-off between false-positive and false-negative error rates associated with rules which classify individuals as ‘positive’ (negative) if their biomarker values are above (below) a threshold, for all possible thresholds. For population screening markers, false-positive errors are especially troublesome, and our interest is estimation of the true-positive fraction (TPF) for a fixed, low, false-positive fraction (FPF). In other cases, false-negative errors are more important, and thus we estimate the FPF corresponding to an acceptably high TPF. We focus here on the first scenario, but note that the second can be handled in a similar fashion.

We define the optimal case–control ratio as that for which the variance of the ROC estimator is minimized. In Section 2, we derive the optimal ratio and show how it can be combined with choosing the overall sample size. A non-parametric estimation method is proposed in Section 3. In Section 4, we consider how the optimal ratio changes as a function of the diagnostic accuracy of the biomarker in a classic distributional case. Section 5 demonstrates that efficiency loss associated with sub-optimal case–control allocation can be substantial. In Section 6, we derive the optimal case–control ratios for estimating two other popular summary measures, the area under the ROC curve (AUC) and the area under part of the ROC curve (pAUC). Finally, we illustrate these methods using data from a study of a marker for adenocarcinoma in patients with Barrett’s esophagus.

2. THE OPTIMAL CASE–CONTROL RATIO

Suppose we are planning a case–control study with \( n_D \) cases and \( n_D \) controls to evaluate the diagnostic accuracy of a biomarker, \( Y \). Let \( Y_D \) and \( Y_D \) denote diseased and non-diseased observations with survivor functions \( S_D(y) = P[Y_D > y] \) and \( S_D(y) = P[Y_D > y] \) and densities \( f_D(y) \) and \( f_D(y) \). We begin by assuming that the study has a fixed overall sample size, \( N \). We seek to determine the fraction of cases in this total sample size in order to most efficiently estimate the ROC curve at a fixed acceptable FPF.

The ROC curve is a plot of the TPF (sensitivity) versus the FPF \((1 - \text{specificity})\) for the set of rules that define ‘test positive’ as ‘\( Y > c \)’, where \( c \) varies from \(-\infty\) to \( \infty \). It can be written as (Pepe, 2003, p. 69)

\[
\text{ROC}(t) = S_D(S_D^{-1}(t)), \quad t \in [0, 1].
\] (2.1)

We use the empirical non-parametric ROC estimator, which estimates the constituent survivor functions empirically:

\[
\hat{\text{ROC}}(t) = \hat{S}_D(\hat{S}_D^{-1}(t)), \quad t \in (0, 1).
\]
Its asymptotic variance, which performs well even in small samples, is given by (Greenhouse and Mantel, 1950; Hsieh and Turnbull, 1996; Pepe, 2003, p. 101)

\[ V(t) = \text{Var}(\widehat{\text{ROC}}(t)) = \frac{\text{ROC}(t)(1 - \text{ROC}(t))}{n_D} + (\text{ROC}'(t))^2 \cdot \frac{t(1-t)}{n_D}, \]

where \( \text{ROC}'(t) = \frac{\partial}{\partial t} \text{ROC}(t) \). This variance is minimized when

\[ \frac{n_D}{n_D} = \rho_{\text{opt}}(t) = \frac{1}{\text{ROC}'(t)} \cdot \sqrt{\frac{\text{ROC}(t)(1 - \text{ROC}(t))}{t(1-t)}}. \]

The form of \( \rho_{\text{opt}}(t) \) is completely general and interesting in several respects. The ratio depends on \( t \), the FPF of interest, and on the slope of the ROC curve at \( t \). A steeper slope at \( t \) implies a lower optimal ratio since, if a small change in \( t \) has a large effect on \( \text{ROC}(t) \), more emphasis should be placed on estimating the threshold accurately, which necessitates sampling more controls. The ratio also depends on \( \sqrt{\frac{\text{ROC}(t)(1 - \text{ROC}(t))}{t(1-t)}} \); this function can be greater or less than 1, and we explore its form further in Section 4. Interestingly, the only case in which the optimal ratio is equal to 1 for all \( t \) is when \( \text{ROC}(t) = t \). This result is proven in the Appendix.

The optimal ratio, \( \rho_{\text{opt}}(t) \), is that which minimizes the variance of \( \widehat{\text{ROC}}(t) \) given the total sample size, \( N \). This is equivalent to minimizing the size of the study for fixed \( V(t) \).

The optimal case–control ratio can be combined with existing sample size calculations (Zhou et al., 2002, Chapter 6; Pepe, 2003, Chapter 8) to determine the basis of case–control pilot data. Having chosen the target FPF \( = v_0 \) against the alternative \( H_1: \text{ROC}(t) = v_1 \), where \( t \) is the fixed FPF, \( v_0 \) is the minimal TPF of interest, and \( v_1 \) is the TPF alternative, require that (Pepe, 2003, p. 223)

\[ \frac{1}{V(t)} = \frac{(\Phi^{-1}(1 - \alpha) + \Phi^{-1}(1 - \beta))^2}{(v_0 - v_1)^2}, \]

where \( V(t) \) is the asymptotic variance of \( \widehat{\text{ROC}}(t) \),

\[ V(t) = \frac{t(1-t)}{n_D} + (\text{ROC}'(t))^2 \cdot \frac{t(1-t)}{n_D}. \]

These expressions, together with the optimal ratio, yield the desired values of \( n_D \) and \( n_{\bar{D}} \). Observe that no additional parameters are involved in the optimal ratio beyond those required for the standard sample size calculation.

3. Estimating the optimal ratio using pilot data

In this section, we propose a non-parametric method for estimating the optimal case–control ratio on the basis of case–control pilot data. Having chosen the target FPF \( t \), the task is to estimate \( \text{ROC}(t) \) and \( \text{ROC}'(t) \). We estimate \( \text{ROC}(t) \) non-parametrically using the empirical estimator, or a smooth version (Peng and Zhou, 2004). To estimate \( \text{ROC}'(t) \), we exploit its relationship with the likelihood ratio function,

\[ LR(y) = \frac{f_D(y)}{f_{\bar{D}}(y)}, \]

evaluated at \( y = S_{\bar{D}}^{-1}(t) \):

\[ \text{ROC}'(t) = \frac{\partial}{\partial t} (S_D(S_{\bar{D}}^{-1}(t))) = \frac{f_D(S_{\bar{D}}^{-1}(t))}{f_{\bar{D}}(S_{\bar{D}}^{-1}(t))} = LR(S_{\bar{D}}^{-1}(t)). \]

We estimate \( \text{ROC}'(t) \) using a ratio of kernel density estimates for \( f_D \) and \( f_{\bar{D}} \), evaluated at \( \hat{S}_{\bar{D}}^{-1}(t) \), an empirical or parametric estimate of the \( (1 - t) \) quantile for \( Y_{\bar{D}} \). In Section 7, we illustrate this estimation method using a Barrett’s esophagus dataset.

Clearly, the estimates of the parameters involved in the optimal case–control ratio have uncertainty. With a small pilot study, this uncertainty may be substantial, and the ROC slope estimator in particular
may be sensitive to the choice of kernel or bandwidth. In practice, we recommend that one err on the conservative side and consider a range of plausible estimates consistent with the pilot data.

4. THE OPTIMAL RATIO IN A CLASSIC DISTRIBUTIONAL SETTING

The binormal ROC curve is the ROC curve induced by normal marker distributions, \( Y_D \sim N(0, 1) \) and \( Y_{\bar{D}} \sim N(\mu_D, \sigma_D^2) \), where without loss of generality we can assume \( Y_D \) has mean 0 and variance 1. However, it is more general; this is the ROC curve for any marker, \( Y \), for which there exists a monotone function that transforms both \( Y_D \) and \( Y_{\bar{D}} \) to normality (Swets and Pickett, 1982; Swets, 1986; Hanley, 1988, 1996; Pepe, 2003, pp. 84–85). The binormal ROC is

\[
\text{ROC}(t) = \Phi\left( \frac{\mu_D}{\sigma_D} + \frac{1}{\sigma_D} \Phi^{-1}(t) \right),
\]

where \( \Phi \) is the standard normal cumulative distribution function. Using (2.3), the corresponding optimal ratio is

\[
\rho_{\text{opt}}(t) = \sigma_D \cdot \frac{\phi(\Phi^{-1}(t))}{\phi\left( \frac{\mu_D}{\sigma_D} + \frac{1}{\sigma_D} \Phi^{-1}(t) \right)} \cdot \sqrt{\frac{\Phi\left( \frac{\mu_D}{\sigma_D} + \frac{1}{\sigma_D} \Phi^{-1}(t) \right) \left( 1 - \Phi\left( \frac{\mu_D}{\sigma_D} + \frac{1}{\sigma_D} \Phi^{-1}(t) \right) \right)}{t(1-t)}}, \quad (4.1)
\]

where \( \phi \) is the standard normal density function. We emphasize that this is the optimal ratio for estimating \( \text{ROC}(t) \) non-parametrically, as is typical in practice. Here, the binormal model simply provides a context in which to illustrate how the height and steepness of the ROC curve affect the optimal ratio.

A variety of binormal ROC curves, indexed by \( \mu_D \), are shown in Figure 1. The parameter \( \sigma_D \) is fixed at unity, guaranteeing a concave ROC curve. We show in Figure 2 the optimal ratio as a function of \( \mu_D \).

![Fig. 1. A range of binormal ROC curves, ROC(t) = \Phi(\mu_D + \Phi^{-1}(t)), indexed by \mu_D.](image-url)
For several values of $t$, the FPF of interest. Observe that $\rho_{opt}(t)$ is not a monotone function of $\mu_D$: it decreases, then increases. This is primarily due to the first component of the optimal ratio, $\frac{1}{\text{ROC}(t)}$, as illustrated in Figures 2(a) and (b), which display $\frac{1}{\text{ROC}(t)}$ and $S = \sqrt{\frac{\text{ROC}(t)(1-\text{ROC}(t))}{t(1-t)}}$ as functions of $\mu_D$. The optimal ratio decreases and then increases with $\mu_D$ due to the fact that, for a fixed $t$, the slopes of the ROC curves increase and then decrease. The ‘turning point’, the value of $\mu_D$ at which the slopes begin to increase, is lower for larger $t$. The factor $S$ generally has the effect of amplifying $\frac{1}{\text{ROC}(t)}$. Hence, we can view the optimal ratio as simply being inversely proportional to the slope of the ROC curve.

The association shown here between $\rho_{opt}(t)$ and $\mu_D$ is valid for all concave binormal models. We have found that it also holds for other concave ROC models, such as those induced by exponential or logistic models for $Y_D$ and $\overline{Y_D}$. 
5. Efficiency Loss Associated with Sub-optimal Case–Control Allocation

We assess efficiency loss associated with a sub-optimal case–control ratio by comparing the variance of $\hat{\text{ROC}}(t)$ with the optimal case–control ratio to the variance with a sub-optimal ratio, assuming a constant total sample size, $N$. Letting $n^o_D$ and $n_D$ denote the optimal and sub-optimal numbers of cases and $V^o(t)$ and $V(t)$ the corresponding variances of $\hat{\text{ROC}}(t)$, the relative efficiency is

$$\frac{V^o(t)}{V(t)} = \frac{\frac{\text{ROC}(t)(1-\text{ROC}(t))}{n^o_D} + (\text{ROC}'(t))^2 \cdot \frac{t(1-t)}{N-n^o_D}}{\frac{\text{ROC}(t)(1-\text{ROC}(t))}{n_D} + (\text{ROC}'(t))^2 \cdot \frac{t(1-t)}{N-n_D}}.$$ 

Using the optimal ratio expression, (2.3), this simplifies to

$$\frac{V^o(t)}{V(t)} = \gamma \left\{ 1 + \frac{\text{ROC}'(t)}{\text{ROC}(t)(1-\text{ROC}(t))} \right\} \times \frac{\text{ROC}(t)(1-\text{ROC}(t)) + \text{ROC}'(t) \sqrt{\text{ROC}(t)(1-\text{ROC}(t))} t(1-t)}{\text{ROC}(t)(1-\text{ROC}(t)) + \frac{1}{1-\gamma} \cdot (\text{ROC}'(t))^2 \cdot t(1-t)},$$

where $\gamma = \frac{n_D}{N}$. Observe that this expression is invariant with respect to $N$, and depends only on $\text{ROC}(t)$, $\text{ROC}'(t)$, and $\gamma$. Figure 3 shows the loss of efficiency associated with sub-optimal case–control allocation under the binormal model, when $\mu_D = 0.75, 1.50, 2.25$ (see the corresponding ROC curves in Figure 1). The plot of $\frac{V^o(t)}{V(t)}$ versus $\gamma = \frac{n_D}{N}$ is shown in Figure 3 for a low FPF of $t = 0.01$, and a moderate value, $t = 0.25$. The relative efficiency is equal to 1 when the percent cases is $\frac{n^o_D}{N}$, and decreases monotonically on either side. Efficiency losses can be substantial. For example, when $\mu_D = 2.25$ and $t = 0.01$, 25% cases is optimal; with 50% cases, efficiency is just 80%. When $t = 0.25$, 60% cases is optimal; with only 40% cases, efficiency is 86%.

Fig. 3. The relative efficiency of a design with an optimal case–control ratio relative to a sub-optimal design, as a function of the percent cases in the sub-optimal design. Lines are drawn for three binormal models, where $\text{ROC}(t) = \Phi(\mu_D + \Phi^{-1}(t))$, and $\mu_D$ takes the values 0.75, 1.50, 2.25. In (a), the FPF is 0.01, and in (b), the FPF is 0.25.
6. THE OPTIMAL CASE–CONTROL RATIOS FOR ESTIMATING THE AUC AND pAUC

6.1 Estimating the AUC

The AUC is a popular summary index for the ROC curve. It has several drawbacks, including the fact that it treats false-positive and false-negative errors equally and summarizes the entire ROC curve when frequently only a portion is of interest (McClish, 1989; Thompson and Zuccini, 1989; Pepe, 2003, Section 4.3). It is nevertheless commonly used and has some desirable mathematical properties (Bamber, 1975; Hanley and McNeil, 1982; Pepe, 2003, Section 4.3). We derived the optimal case–control ratio for estimating the AUC empirically. The asymptotic variance of the empirical AUC (DeLong et al., 1988) is

$$\text{Var}(\hat{\text{AUC}}) = \frac{\text{Var}(S_{\bar{D}}(Y_D))}{n_D} + \frac{\text{Var}(S_D(Y_{\bar{D}}))}{n_{\bar{D}}}.$$  

With $N$ fixed, this is minimized when

$$\frac{n_D}{n_{\bar{D}}} = \rho_{\text{opt}}^{\text{AUC}} = \sqrt{\frac{\text{Var}(S_{\bar{D}}(Y_D))}{\text{Var}(S_D(Y_{\bar{D}}))}}. \quad (6.1)$$

This result was first pointed out by Hanley and Hajian-Tilaki (1997).

Observe that $\rho_{\text{opt}}^{\text{AUC}}$ depends on the variances of two random variables, called placement values (Hanley and Hajian-Tilaki, 1997; DeLong et al., 1988; Pepe, 2003, pp. 105–106; Pepe and Cai, 2004). $S_{\bar{D}}(Y_D)$ is the placement value of a case observation in the control population, and $S_D(Y_{\bar{D}})$ is the placement value of a control observation in the case population. For good discrimination, $S_{\bar{D}}(Y_D)$ should take small values and $S_D(Y_{\bar{D}})$ large values. When estimating the AUC, more cases than controls are required if the case placement values are more variable than the control placement values. To estimate this optimal ratio using pilot data, we estimate the placement values empirically (or perhaps parametrically) and then calculate the sample variances.

In contrast to estimation of the ROC curve, there are many settings where equal numbers of cases and controls are optimal for estimating the AUC. All that is required is that the case and control placement values have the same variance. If there exists a monotone increasing transformation, $g$, such that $g(Y_D)$ and $g(Y_{\bar{D}})$ follow a symmetric location-scale distribution with the same scale parameter, the height of the ROC curve and the value of the AUC are irrelevant: sampling equal numbers of cases and controls is optimal (see Appendix). One such example is when ROC($t$) = $t$ (in this case, the case and control placement values actually have the same distribution). This result also implies that a case–control ratio of 1 is optimal when $g(Y_D)$ and $g(Y_{\bar{D}})$ are both normally distributed for some function $g$ with the same scale and different location parameters. This includes all the scenarios shown in Figure 1.

We emphasize that the optimal ratio for estimating the AUC may be entirely different than the optimal ratio for estimating ROC($t$), a consequence of the fact that these are two different measures of classification accuracy.

Calculating $\rho_{\text{opt}}^{\text{AUC}}$ may be combined with choosing the overall sample size, as was done for the ROC curve. Briefly, when AUC$_0$ is a minimally acceptable value for the AUC, in order to achieve power $1 - \beta$ and type-I error $\alpha$ when the AUC is equal to AUC$_1$, we require that (Pepe, 2003, p. 226)

$$n_D = (\text{Var}(S_{\bar{D}}(Y_D)) + \sqrt{\text{Var}(S_{\bar{D}}(Y_D)) \cdot \text{Var}(S_D(Y_{\bar{D}}))}) \times \left(\frac{\Phi^{-1}(1 - \alpha) + \Phi^{-1}(1 - \beta)}{AUC_1 - AUC_0}\right)^2. \quad (6.2)$$

This expression and (6.1) determine the sample sizes.
6.2 Estimating the pAUC

The partial area under the curve, or pAUC, is the area under a specified region of the ROC curve,

\[
pAUC_{t_0,t_1} = \int_{t_0}^{t_1} \text{ROC}(t)\,dt.
\]

It can be used to summarize the performance of the marker over an acceptable range of FPFs (McClish, 1989; Thompson and Zuccini, 1989). The non-parametric estimator of the pAUC (Dodd and Pepe, 2003) is

\[
\hat{pAUC}_{t_0,t_1} = \frac{1}{n_D n_{\tilde{D}}} \sum_{i=1}^{n_D} \sum_{j=1}^{n_{\tilde{D}}} I(Y_{Di} > Y_{Dj}, Y_{Dj} \in (\hat{q}_1, \hat{q}_0)),
\]

where \( \hat{q}_i = \hat{S}_{D}^{-1}(t_i) \). This is approximately asymptotically equivalent to (Dodd, 2001)

\[
\frac{1}{n_D n_{\tilde{D}}} \sum_{i=1}^{n_D} \sum_{j=1}^{n_{\tilde{D}}} I(Y_{Dj} \in (q_1, q_0)) \cdot S_{D}(Y_{Dj}) + I(Y_{Di} \in (q_1, q_0)) \cdot S_{\tilde{D}}(Y_{Di}),
\]

which has asymptotic variance

\[
\frac{\text{Var}(S^R_{\tilde{D}}(Y_D))}{n_{\tilde{D}}} + \frac{\text{Var}(S^R_D(Y_D))}{n_D},
\]

where \( S^R_{\tilde{D}}(Y_D) = P(Y_D > Y_{\tilde{D}}, Y_{\tilde{D}} \in (q_1, q_0)|Y_D) \) and \( S^R_D(Y_D) = P(Y_D > Y_D, Y_D \in (q_1, q_0)|Y_D) \) are restricted placement values (Dodd and Pepe, 2003). For fixed \( N \), the variance is minimized when

\[
\frac{n_D}{n_{\tilde{D}}} = \rho^{pAUC}_{\text{opt}} = \sqrt{\frac{\text{Var}(S^R_{\tilde{D}}(Y_D))}{\text{Var}(S^R_D(Y_D))}},
\]

Observe that this optimal ratio depends on the FPF region of interest, and will be greater than unity if restricted case placement values are more variable than those for controls. The ratio will equal 1 if and only if the restricted placement values have equal variance. This will occur much less frequently than will equal variance of the unrestricted placement values. For example, a disease that (after some transformation) causes a simple shift in a symmetric location-scale distribution need not have unrestricted placement values with equal variance.

Note that the variance in (6.3) used to derive \( \rho^{pAUC}_{\text{opt}} \) holds only approximately, since it does not include the variance due to estimating the two quantiles, \( q_1 \) and \( q_0 \) (Dodd and Pepe, 2003). This second component tends to be small. Nevertheless, in practice, one might check using simulations that \( \rho^{pAUC}_{\text{opt}} \) truly minimizes the variance of the estimator.

Estimating \( \rho^{pAUC}_{\text{opt}} \) is straightforward, using sample variances of the empirical (or parametrically estimated) restricted placement values. Calculation of this optimal ratio can also be combined with sample size calculations. When pAUC\(_0\) and pAUC\(_1\) are the null and alternative values for pAUC\(_{t_0,t_1}\), we require that

\[
n_D = \left( \text{Var}(S^R_{\tilde{D}}(Y_D)) + \sqrt{\text{Var}(S^R_{\tilde{D}}(Y_D)) \text{Var}(S^R_D(Y_D))} \right) \times \left( \Phi^{-1}(1 - \alpha) + \Phi^{-1}(1 - \beta) \right) \frac{\text{pAUC}_1 - \text{pAUC}_0}{\text{pAUC}_1 - \text{pAUC}_0}
\]

in order to achieve power \( 1 - \beta \) and type-I error \( \alpha \).
Barrett’s esophagus is a condition that develops in some patients with gastroesophageal reflux disease, in which the normal lining of the esophagus becomes replaced with a red-cell lining called ‘Barrett’s esophagus’. Certain cells within this new lining progress to become a cancer of the lower esophagus, or adenocarcinoma.

The current study concerns a new potential diagnostic marker for adenocarcinoma in Barrett’s esophagus patients which we call Y. A pilot study of 71 cases and 55 controls have marker measurements taken from endoscopy tissue. Our goal is to estimate the numbers of cases and controls that are needed in a future validation study of the marker.

Since work-up procedures following a positive screening test will require the patient to undergo invasive repeat endoscopies, it is important that the false-positive rate be low. The investigator seeks a marker for adenocarcinoma that has a FPF of no more 10%. We calculate the sample sizes for estimating the ROC curve at FPFs of 5 and 10%. According to the pilot data, the empirical ROC estimates are $\hat{\text{ROC}}(0.05) = 0.28$ and $\hat{\text{ROC}}(0.10) = 0.37$ (see Figure 4[a]). We also estimate $\hat{S}_D^{-1}(t)$ empirically for $t = 0.05$ and $t = 0.10$, and evaluate the kernel density estimates for $f_D$ and $f_{D_r}$ at these thresholds (see Figure 4[b]). The final slope estimates are 3.36 and 0.96.

We next calculate the optimal sample sizes. Although not ideal, the marker will be considered minimally useful clinically if it detects 15% of cancers when the FPF is 0.05 ($v_0 = 0.15$), or 25% of cancers when the FPF is 0.10 ($v_0 = 0.25$). The observed TPFs at $t = 0.05$ and $t = 0.10$ are used as the anticipated alternative values. In order to have 90% power and 5% type-I error, we require $(n_D, n_{D_r}) = (262, 427)$ and $(n_D, n_{D_r}) = (235, 140)$ at $t = 0.05$ and $t = 0.10$, respectively (see Table 1). The optimal ratios at these two FPFs are 0.61 and 1.68. Interestingly, more cases are required than controls at $t = 0.10$, while fewer cases than controls are required at $t = 0.05$. The projected variance of $\hat{\text{ROC}}(t)$ in the future study can also be calculated using (2.2) (see Table 1). Observe that this variance depends on the estimates of $\text{ROC}(t)$ and $\text{ROC}'(t)$ and the sample sizes.

We also calculate the sample sizes that are required to estimate the AUC from $t = 0.0$ to $t = 0.1$. Using sample variances of empirical restricted placement values, we calculate that the optimal case–control ratio for estimating the pAUC is 0.34. This case–control ratio is much smaller than those pertaining to

Fig. 4. Barrett’s esophagus data: (a) The empirical ROC curve for $Y$, for $t \in [0, 0.3]$; (b) Kernel density estimates of $Y$. 
estimation of the ROC curve at our two FPFs of interest. In order to calculate the raw sample sizes required, we use a null pAUC of 0.01 (which corresponds to the binormal ROC curve passing through the null values for ROC[0.05] and ROC[0.10]) and an alternative pAUC of 0.02 (the observed non-parametric estimate). In order to have 90% power and 5% type-I error, we require \( n_D = 314 \) cases and \( n_{\bar{D}} = 928 \) controls (see Table 1). The large sample size is a consequence of the fact that the null and alternative pAUC values are close together. We find that the choice of summary measure greatly affects the optimal case–control ratio and required sample sizes.

8. DISCUSSION

Case–control studies have long been used in epidemiology to evaluate associations between disease and exposure. Odds ratios are typically used to quantify association. It has been demonstrated, however, that appropriate evaluation of case–control biomarker studies is fundamentally different from that of case–control association studies (Pepe et al., 2004). The odds ratio does not characterize classification accuracy. Accordingly, sample size calculations for case–control biomarker studies must be based on estimating parameters that relate to classification accuracy. Zhou et al. (2002, Chapter 6) and Pepe (2003, Chapter 8) document sample size calculations that fix the case–control ratio arbitrarily. In this paper, we develop methods that select the ratio optimally.

We have shown the optimal case–control ratio for estimating the ROC curve at a given FPF \( t \) to be a simple function of \( t \) and the ROC curve and its derivative at \( t \). In general, the optimal ratio is simply inversely proportional to the slope of the ROC curve at \( t \), and the use of a sub-optimal case–control ratio can be associated with substantial loss of efficiency. We also derived the optimal case–control allocations for estimating the AUC and pAUC.

The sample size and the optimal ratio are highly dependent on the accuracy parameter chosen as the basis of the study design. This was made clear in the example of Section 7, where we observed that estimation of the pAUC required a much larger sample size than did estimation of the ROC curve at a fixed point \( (N = 1242 \text{ versus } N = 375) \). We have also found in practice that estimating the AUC requires a much smaller sample size than does estimating a specific point on the ROC curve. Such small studies, however, will not have the appropriate power to draw adequate conclusions about specific points on the ROC curve. In most situations, the ROC curve is of more interest since it has more clinical meaning than AUC. This relationship between parameter and sample size is not surprising. Moreover, it reinforces the notion that the parameter choice should be based on careful consideration of the study objectives and consequences of study results.

In order to estimate the optimal ratio and the required sample sizes, estimates of the parameters are needed. These estimates incorporate uncertainty into the calculations. Unfortunately, this is so with any sample size calculation. For example, in order to design a clinical trial to evaluate treatment efficacy, we require an estimate of \( \sigma^2 \), the variance of the outcome. This estimate is often highly uncertain, but is the best that we can do with the available information. Similarly, calculating the case–control ratio for estimating ROC(\( t \)), as well as determining the raw sample sizes, requires an estimate of the slope of the
ROC curve at $t$. We have proposed a non-parametric method for estimating the slope which relies on kernel density estimates, and which can be sensitive to the choice of kernel and bandwidth. We have provided suggestions for estimating the ROC slope given limited pilot data. However, there are other existing approaches to estimate the slope; one is to use the form induced by a parametric form for the ROC curve. (Interestingly, this leads to quite a different optimal ratio estimate in our example.) A thorough study of methods to estimate the slope of the ROC curve is warranted.

In calculating the optimal ratio for estimating ROC$(t)$, we have assumed that the goal is to estimate the TPF at a given FPF = $t$. However, in some situations, a priority is to avoid false-negative errors, and our interest lies in calculating the FPF at a fixed large TPF = $v$. We note that this can easily be handled using the methods we have described by replacing $t$ with ROC$^{-1}(v)$ in all the expressions.

We have proposed methods for choosing the case–control ratio and overall sample size in a biomarker study. In some studies, only a fixed number of cases are available, and the real question is how many controls should be sampled. In such circumstances, efficiency cannot be maximized, but the number of controls can be calculated as a function of the desired power and type-I error, using standard sample size formulas.

Our methods can be extended in a number of ways. First, we have sought to minimize the variance of the estimator for fixed overall sample size, or equivalently, minimized the sample size for fixed variance. An alternative approach would be to assign different costs to cases and controls, and to minimize the total cost of the study, while fixing the variance. We also focused on minimizing the asymptotic variance of the estimator. While this is sufficient for relatively large biomarker studies, it may not be adequate for small studies. We suggest using small-sample simulation studies with various case–control ratios, and selecting that which minimizes the variance of ROC$(t)$. Our large-sample results could provide reasonable starting values for such simulations. Finally, in later phase biomarker studies, the task is to compare the new biomarker to the best available markers. Our methods may be extended to derive the optimal case–control ratio and raw sample sizes needed to compare two markers. We leave these questions for future research.

**Acknowledgments**

This work was supported by R01GM 54438 and U01CA 086368. Conflict of Interest: None declared.

**Appendix**

In terms of estimating ROC$(t)$ at a fixed $t$ of interest, the only case in which $\rho_{opt}(t) = 1$ for all $t$ is when ROC$(t) = t$. To see this, note that $\rho_{opt}(t) = 1$ if and only if ROC$(t) = \sqrt{\text{ROC}(t)(1-\text{ROC}(t))}$. Letting ROC$(t) = y$ and integrating both sides, we have arcsin$(2y-1) = \arcsin(2t-1) + c$, where $c$ is a constant. The condition $y(0) = 0$ implies that $c = 0$, which yields the desired result.

In estimating the AUC, if there exists a monotone increasing transformation, $g$, such that $g(Y_D)$ and $g(Y_D) + \mu$ have survivor function $S$, where $S$ is symmetric, $\rho_{opt}^{\text{AUC}} = \sqrt{\text{Var}(S_D(Y_D)) / \text{Var}(S_D(Y_D))} = 1$. Assume without loss of generality that $S$ is symmetric about zero, and hence $S(y) = 1 - S(-y)$ and $f(y) = f(-y)$, where $f$ is the density corresponding to $S$. Using $S_D(Y_D) = S(g(Y_D))$, $S_D(Y_D) = S(g(Y_D) + \mu)$, and $E(S_D(Y_D)) = P(g(Y_D) > g(Y_D)) = P(Y_D > Y_D)$, we have

\[
E(S_D(Y_D)^2) = E(S(g(Y_D))^2) = \int S(y)^2 f(y + \mu)dy
= \int (1 - S(-y))^2 f(-y - \mu) dy = \int (1 - S(z + \mu))^2 f(z) dz
= 1 - 2 \left[ 1 - P(Y_D > Y_D) \right] + E(S_D(Y_D)^2),
\]
and so

$$\text{Var}(S_D(Y_D)) = E(S_D(Y_D)^2) - (E(S_D(Y_D)))^2$$
$$= -1 + 2P(Y_D > Y_D) + E(S_D(Y_D)^2) - (P(Y_D > Y_D))^2$$
$$= E(S_D(Y_D)^2) - (1 - P(Y_D > Y_D))^2$$
$$= \text{Var}(S_D(Y_D)).$$

REFERENCES


[Received March 2, 2005; first revision November 28, 2005; second revision January 17, 2006; accepted for publication January 17, 2006]