BIOMETRICS 56, 337–344
June 2000

Time-Dependent ROC Curves for Censored Survival Data and a Diagnostic Marker

Patrick J. Heagerty,[1,2,*] Thomas Lumley,[1] and Margaret S. Pepe[2]

[1]Department of Biostatistics, University of Washington,
Seattle, Washington 98195, U.S.A.
[2]Division of Public Health Sciences, Fred Hutchinson Cancer Research Center,
Seattle, Washington 98109, U.S.A.
*email: heagerty@biostat.washington.edu

SUMMARY. ROC curves are a popular method for displaying sensitivity and specificity of a continuous diagnostic marker, X, for a binary disease variable, D. However, many disease outcomes are time dependent, $D(t)$, and ROC curves that vary as a function of time may be more appropriate. A common example of a time-dependent variable is vital status, where $D(t) = 1$ if a patient has died prior to time t and zero otherwise. We propose summarizing the discrimination potential of a marker X, measured at baseline ($t = 0$), by calculating ROC curves for cumulative disease or death incidence by time t, which we denote as ROC(t). A typical complexity with survival data is that observations may be censored. Two ROC curve estimators are proposed that can accommodate censored data. A simple estimator is based on using the Kaplan–Meier estimator for each possible subset $X > c$. However, this estimator does not guarantee the necessary condition that sensitivity and specificity are monotone in X. An alternative estimator that does guarantee monotonicity is based on a nearest neighbor estimator for the bivariate distribution function of (X, T), where T represents survival time (Akritas, M. J., 1994, *Annals of Statistics* **22,** 1299–1327). We present an example where ROC(t) is used to compare a standard and a modified flow cytometry measurement for predicting survival after detection of breast cancer and an example where the ROC(t) curve displays the impact of modifying eligibility criteria for sample size and power in HIV prevention trials.

KEY WORDS: Accuracy; Discrimination; Kaplan–Meier estimator; Kernel smoothing; Sensitivity; Specificity.

1. Introduction

Over the past decade, tumor characteristics quantified through cytometric analysis have proven useful for establishing prognosis in several human carcinomas. For breast cancer patients, the percent of cells in the synthesis phase of the cell cycle, or S-phase, has been shown to correlate with survival (Sigurdsson et al., 1990). Recently developed techniques allow S-phase measurements to target malignant epithelial cells within tumor samples and therefore may provide more accurate S-phase assessment and improved prognostic potential. A key scientific question is whether the new targeted measurement can more accurately discriminate between women that succumb to disease and those women that survive. Accuracy summaries such as sensitivity and specificity are well established for simple binary (disease) variables with either discrete or continuous marker measurements. The goal of this manuscript is to extend the concepts of sensitivity and specificity to time-dependent binary variables such as vital status, allowing characterization of diagnostic accuracy for censored survival outcomes.

For test results defined on continuous scales, receiver operator characteristic (ROC) curves are standard summaries of accuracy. If X denotes the diagnostic test or marker, with higher values more indicative of disease, and D is a binary indicator of disease status, then the ROC curve for X is a plot of the sensitivity associated with the dichotomized test $X > c$ versus (1 − specificity) for all possible threshold values c, i.e., the ROC curve is the monotone increasing function in $[0, 1]$, $[\{P(X > c \mid D = 0), P(X > c \mid D = 1)\}, c \in (-\infty, \infty)]$. This function characterizes the diagnostic potential of a continuous test by summarizing all of the possible trade-offs between sensitivity and specificity. The higher the ROC curve is in the quadrant $[0, 1] \times [0, 1]$, the better is its capacity for discriminating diseased from nondiseased subjects.

In diagnostic medicine, ROC curves are recognized as having several attractive features. First, an ROC curve describes the inherent discrimination capacity of a test without linking it to any specific threshold. Second, ROC curves are particularly useful for comparing the discriminatory capacity of different diagnostic markers. They provide a valid approach to comparison even when markers are on completely different measurement scales. Another attribute is that they do not depend on disease prevalence and hence can be estimated from case–control studies. Last, the area under the ROC curve can

be interpreted as the probability that the test result from a randomly chosen diseased individual exceeds that for a randomly chosen nondiseased individual and is often used to summarize the ROC curve. General discussions of ROC analysis can be found in Swets and Pickett (1982), Hanley (1989), Begg (1991), Zweig and Campbell (1993), and Pepe, Leisenring, and Rutter (1999).

ROC curves for continuous diagnostic tests can be estimated nonparametrically using empirical estimates of the survivor functions, $S_0(c) = P(X > c \mid D = 0)$ and $S_1(c) = P(X > c \mid D = 1)$. Smooth nonparametric and semiparametric estimators have been proposed by Zhou, Hall, and Shapiro (1997) and by Metz, Herman, and Shen (1998), respectively.

Disease status is considered a fixed characteristic of a study subject in classic ROC analysis. In this paper, we consider settings where disease status can change with time. Subjects are initially nondiseased but can succumb to disease during the course of the study. The question to be addressed is how well a diagnostic marker measured at baseline can distinguish between subjects who become diseased and subjects who do not in a follow-up interval $[0, t]$. A common complexity in longitudinal studies is that the disease onset time may be censored.

In Section 2, we define time-dependent ROC curves for such disease incidence settings. Two ROC curve estimators are proposed, both of which can accommodate censoring of the time-dependent disease data. The first is a simple estimator based on Kaplan–Meier survivor function methods. However, we show that, with censored data, this simple estimator can yield nonmonotone sensitivity or specificity functions. The second method is based on a valid bivariate survivor function estimator and always yields monotone ROC curves. This estimator has the additional advantage that it allows the censoring process to depend on X. Since ascertainment of disease status can be highly dependent on the results of a screening test (and is known as ascertainment or verification bias; Begg [1987]), this flexibility is potentially important in many applications. The new methods are applied to breast cancer mortality data in Section 3.1, where we compare two prognostic markers measured at the time of cancer diagnosis. A second example in Section 3.2 demonstrates a novel application of ROC methods to design methodology for clinical trials. We show that time-dependent ROC curves can assist in developing eligibility criteria for clinical trials and apply the approach to the design of an HIV prevention study. Concluding remarks and a brief discussion are presented in Section 4.

2. Estimation

In this section, we outline two approaches for the estimation of ROC curves with a time-dependent disease variable, or more generally a failure time, for data obtained in a prospective cohort study. The first approach is based on direct use of Bayes' theorem and the Kaplan–Meier estimator (Kaplan and Meier, 1958) but may lead to nonmonotone sensitivity and specificity functions. An alternative estimator is introduced that guarantees monotonicity.

2.1 *Definitions*

Let T_i denote failure time and X_i the covariate value for subject i. Let C_i denote the censoring time, $Z_i = \min(T_i, C_i)$ the follow-up time, and δ_i a censoring indicator with $\delta_i = 1$ if $T_i \leq C_i$ and $\delta_i = 0$ if $T_i > C_i$. We use the counting process $D_i(t) = 1$ if $T_i \leq t$ and $D_i(t) = 0$ if $T_i > t$ to denote failure (disease) status at any time t with $D_i(t) = 1$ indicating that subject i has had an event prior to time t.

Recall that ROC curves display the relationship between a covariate X_i and a binary disease variable D_i by plotting estimates of the sensitivity, $P(X > c \mid D = 1)$, and one minus the specificity, $1 - P(X \leq c \mid D = 0)$ for all possible values c. When disease status is time dependent, consider sensitivity and specificity as time-dependent functions and define them as

$$\text{sensitivity}(c, t) = P\{X > c \mid D(t) = 1\}$$
$$\text{specificity}(c, t) = P\{X \leq c \mid D(t) = 0\}.$$

Using these definitions, we can define the corresponding ROC curve for any time t, ROC(t).

2.2 *Using the Kaplan–Meier Estimator*

We can use Bayes' theorem to rewrite the sensitivity and the specificity as

$$P\{X > c \mid D(t) = 1\} = \frac{\{1 - S(t \mid X > c)\}P(X > c)}{\{1 - S(t)\}}$$
$$P\{X \leq c \mid D(t) = 0\} = \frac{S(t \mid X \leq c)P(X \leq c)}{S(t)},$$

where $S(t)$ is the survival function $S(t) = P(T > t)$ and $S(t \mid X > c)$ is the conditional survival function for the subset defined by $X > c$.

A widely used nonparametric estimate of $S(t)$ is given by Kaplan and Meier (1958). Define $\mathcal{T}_n$ to be the unique values of Z_i for observed events, $\delta_i = 1$. The Kaplan–Meier (KM) estimator is defined as

$$\hat{S}_{KM}(t) = \prod_{s \in \mathcal{T}_n,\ s \leq t} \left\{ 1 - \frac{\sum_j \mathbf{1}(Z_j = s)\delta_j}{\sum_j \mathbf{1}(Z_j \geq s)} \right\}.$$

The KM estimator uses all of the information in the data, including censored observations, to estimate the survival function.

A simple estimator for sensitivity and specificity at time t is then given by combining the KM estimator and the empirical distribution function of the marker covariate, X, as

$$\hat{P}_{KM}\{X > c \mid D(t) = 1\} = \frac{\{1 - \hat{S}_{KM}(t \mid X > c)\}\{1 - \hat{F}_X(c)\}}{\{1 - \hat{S}_{KM}(t)\}}$$
$$\hat{P}_{KM}\{X \leq c \mid D(t) = 0\} = \frac{\hat{S}_{KM}(t \mid X \leq c)\hat{F}_X(c)}{\hat{S}_{KM}(t)},$$

where $\hat{F}_X(c) = \Sigma\, \mathbf{1}(X_i \leq c)/n$.

One problem with this simple estimator is that it does not guarantee that sensitivity (or specificity) is monotone. By definition, we have $P\{X > c \mid D(t) = 1\} \geq P\{X > c' \mid D(t) = 1\}$ for $c' > c$. However, the estimates produced via Bayes' theorem and Kaplan–Meier may violate this monotonicity since the quadrant probability estimator

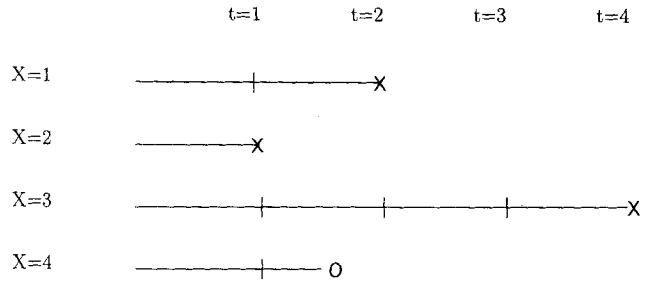

Figure 1. Hypothetical data used to illustrate the potential difficulty in direct use of the Kaplan–Meier estimator for estimating a bivariate distribution function. The symbol × indicates an observed failure time and ○ denotes a censored time.

$\hat{P}(X > c, T > t) = \hat{S}_{KM}(t \mid X_i > c)\{1 - \hat{F}_X(c)\}$ may not produce a valid bivariate distribution. Figure 1 presents hypothetical data that illustrate the potential difficulty. In this example, we obtain conditional survival estimates of $\hat{S}_{KM}(3 \mid X_i > 0) = 3/8$ and $\hat{S}_{KM}(3 \mid X_i > 1) = 2/3$. Combining these with $\hat{F}_X(0) = 0$ and $\hat{F}_X(1) = 1/4$ yields $\hat{P}(T > 3, X > 0) = (3/8) \times 1 = 3/8$ and $\hat{P}(T > 3, X > 1) = (2/3) \times (3/4) = 1/2$. These estimates imply that $\hat{P}(T > 3, 1 \geq X > 0) = 3/8 - 1/2 = -1/8$. The problem can be attributed to the fact that, as the conditioning set $X > c$ changes, the redistribution to the right (Efron, 1967) of the probability mass associated with censored observations also changes. In this example, the KM estimator distributes mass for the censored observation $(X_i = 4, T_i > 1)$ to those observed times greater than the censored time. When the subset $X > 0$ is considered, mass is distributed between $t = 2$ and $t = 4$, but when the subset $X > 1$ is considered, mass from the censored observation is distributed entirely to $t = 4$ since the point $(X_i = 1, T_i = 2)$ is no longer included in the subset. The changing redistribution can lead to inconsistencies that produce negative probability mass and therefore produce ROC curves that are not monotone.

A second potential problem with the KM-based ROC estimator is that the conditional Kaplan–Meier estimator $\hat{S}_{KM}(t \mid X > c)$ assumes that the censoring process does not depend on X. This assumption may be violated in practice when the intensity of follow-up efforts are influenced by the baseline diagnostic marker measurements.

2.3 *Using Nearest Neighbor Estimation of the Bivariate Distribution*

A valid ROC solution can be provided by using an estimator of the bivariate distribution function, $F(c,t) = P(X \leq c, T \leq t)$, or equivalently $S(c,t) = P(X > c, T > t)$, provided by Akritas (1994). This estimator is based on the representation $S(c,t) = \int_c^\infty S(t \mid X = s) dF_X(s)$, where $F_X(s)$ is the distribution function for X. As shown by Akritas (1994), an estimator can be provided by

$$\hat{S}_{\lambda_n}(c,t) = \frac{1}{n} \sum_i \hat{S}_{\lambda_n}(t \mid X = X_i) \mathbf{1}(X_i > c),$$

where $\hat{S}_{\lambda_n}(t \mid X = X_i)$ is a suitable estimator of the conditional survival function characterized by a parameter λ_n. Unless X is discrete and there are sufficient observations at each value of X_i, some smoothing is required to estimate $S(t \mid X = X_i)$.

Define the weighted Kaplan–Meier estimator as

$$\hat{S}_{\lambda_n}(t \mid X = X_i) = \prod_{s \in \mathcal{T}_n,\ s \leq t} \left\{ 1 - \frac{\sum_j K_{\lambda_n}(X_j, X_i) \mathbf{1}(Z_j = s) \delta_j}{\sum_j K_{\lambda_n}(X_j, X_i) \mathbf{1}(Z_j \geq s)} \right\},$$

where $K_{\lambda_n}(X_j, X_i)$ is a kernel function that depends on a smoothing parameter λ_n. Akritas (1994) uses a 0/1 nearest neighbor kernel, $K_{\lambda_n}(X_i, X_j) = \mathbf{1}\{-\lambda_n < \hat{F}_X(X_i) - \hat{F}_X(X_j) < \lambda_n\}$, where $2\lambda_n \in (0,1)$ represents the percentage of observations that is included in each neighborhood (except for the boundaries in the distribution of X). Other kernel choices are also possible; however, using the nearest neighbors implies that the resulting ROC estimates are invariant to monotone transformations of the marker variable. For the nearest neighbor kernel, Akritas (1994) presents bounds on the sequence of smoothing parameters λ_n that are sufficient to yield the weak consistency of the bivariate distribution function estimator. Using $\lambda_n = O(n^{-1/3})$ satisfies these conditions and can be used to guide the choice of λ_n in practice. Finally, Akritas (1994) shows that the nearest neighbor estimator (NNE) is a semiparametric efficient estimator.

The resulting estimates of sensitivity and specificity are given by

$$\hat{P}_{\lambda_n}\{X > c \mid D(t) = 1\} = \frac{[\{1 - \hat{F}_X(c)\} - \hat{S}_{\lambda_n}(c,t)]}{\{1 - \hat{S}_{\lambda_n}(t)\}} \quad (1)$$

$$\hat{P}_{\lambda_n}\{X \leq c \mid D(t) = 0\} = 1 - \frac{\hat{S}_{\lambda_n}(c,t)}{\hat{S}_{\lambda_n}(t)}, \quad (2)$$

where $\hat{S}_{\lambda_n}(t) = \hat{S}_{\lambda_n}(-\infty, t)$. Observe that the numerator in (1) is $n^{-1} \Sigma_i \mathbf{1}(X_i > c)\{1 - S_{\lambda_n}(t \mid X = X_i)\}$, which is monotone decreasing in c, and that the numerator in (2) is similarly monotone increasing in c. This is in contrast to the simple KM-based estimator given in Section 2.2.

Another important advantage of the NNE estimator is that the censoring process is allowed to depend on the diagnostic marker X (Akritas, 1994). This results since only local Kaplan–Meier estimators are used in each possible neighborhood of $X = x$. Since in screening studies follow-up will often be more intense for subjects with marker values that appear more indicative of disease, this flexibility in the censoring mechanism is likely to be important in practice.

2.4 *Estimation of Standard Errors*

Akritas (1994) gives a formula for the variance of the smoothed survival estimator that could be used to calculate standard errors for the sensitivity and specificity. However, results of van der Vaart and Wellner (1996, Theorems 3.6.2 and 3.9.11) together with the functional central limit theorem proved by Akritas (1994) imply that the bootstrap can be

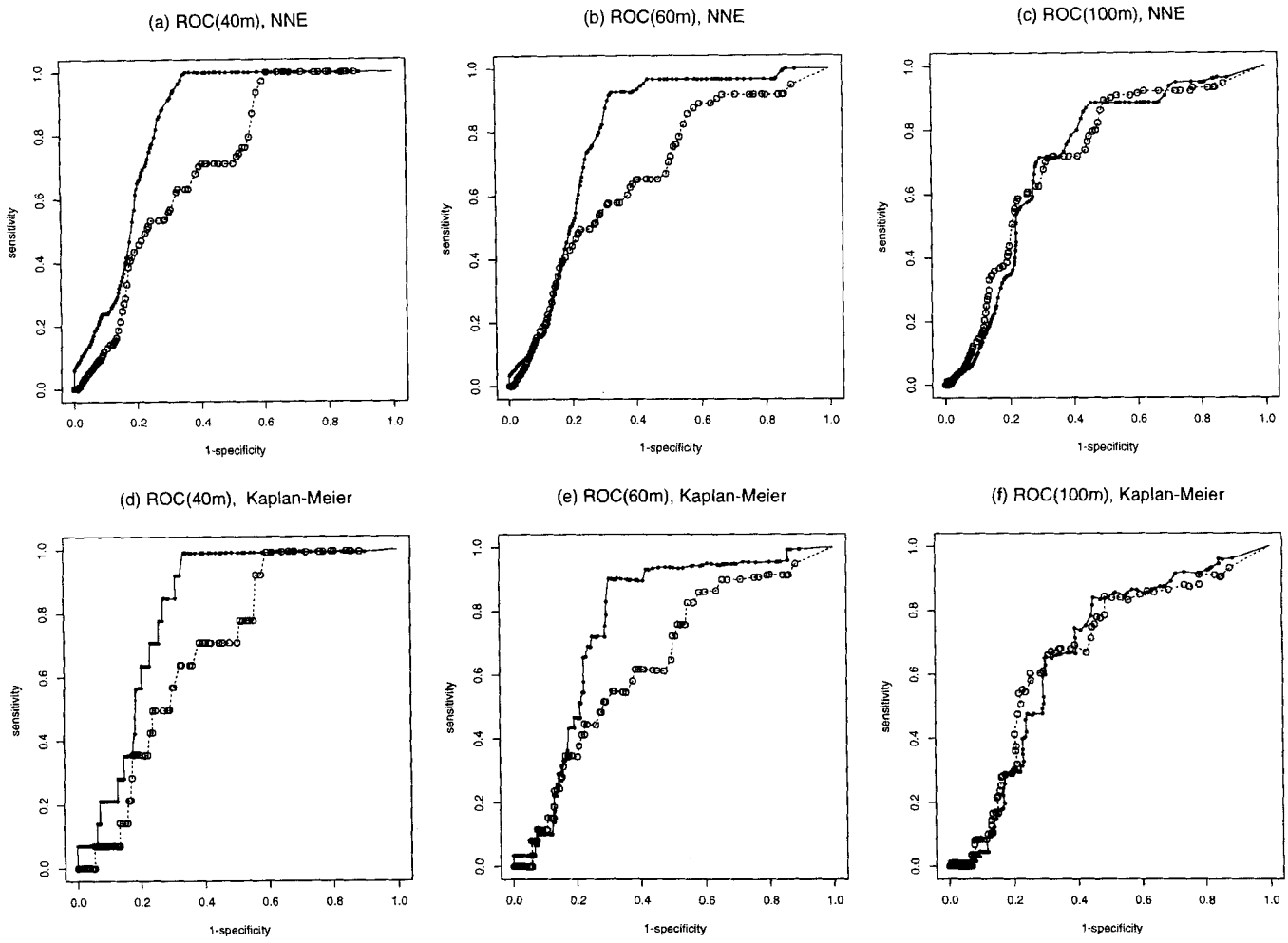

Figure 2. ROC curves for percent S-phase (%S) measurements and survival after breast cancer diagnosis. Solid lines and • represent the new %S measurements while dashed lines and ○ represent the standard %S measurements. Panels (**a**)–(**c**) show the ROC curves for 40, 60, and 100 months using the nearest neighbor estimator (NNE) and panels (**d**)–(**f**) show the estimates based on direct use of the Kaplan–Meier estimator.

used for asymptotically valid confidence limits at times t, where $P(T > t)$ is bounded away from zero. The bootstrap also allows confidence bands to be calculated for the whole ROC(t) curve or for any other differentiable function of $S(c, t)$. In the example presented in Section 3.1, we use the bootstrap to provide a confidence interval for the comparison of two ROC curves.

3. Examples

3.1 *An Improved Flow Cytometry Measurement?*

Flow cytometry is a technique for measuring various cellular characteristics of tumors. One such cytometry measure is the percent of cells that are in the DNA synthesis phase of the cell cycle, or S-phase. In breast cancer, this measure is known to correlate with the average survival time after primary diagnosis (Sigurdsson et al., 1990). Typically, flow cytometry is based on a tumor tissue sample that contains a mixture of cell types. It is possible to sort the cells that contain the protein cytocarotine by a process referred to as gating, making the cytometry measurement specific to the epithelial cells. Since breast cancer is an epithelial cancer, this gated measurement may provide improved prognostic potential over the standard S-phase measurement (the ungated measurement).

Data that we analyze are from a study conducted at the Fred Hutchinson Cancer Research Center in Seattle, Washington. Newly diagnosed cases of breast cancer identified through the Cancer Surveillance System of western Washington were recruited into a prospective study between 1983 and 1992. A total of 1292 women aged 20–44 were enrolled and a subset of $n = 253$ women whose tumor sample had a diploid DNA index were selected for both gated and ungated S-phase measurement. For our analysis, the follow-up time was censored at May 1, 1997. The median follow-up was 62 months, with a total of 44 deaths observed among the subset of 253 women. Among the diploid women, the standard ungated S-phase measurements range from 0.0 to 44% with a median of 2.2% and quartiles of 0.8 and 4.2%. The gated

measurements range from 0.0 to 50% with a median of 3.3% and quartiles of 1.1 and 7.8%.

A goal of the study is to assess the capacity of the gated measurement compared to the standard ungated measurement for discriminating between surviving and dying women. Time-dependent ROC curves provide a graphical display for addressing this question by characterizing the distribution of markers among women that die by time t, $[X \mid D(t) = 1]$, relative to the distribution of the marker among women that survive, $[X \mid D(t) = 0]$. For any fixed time t and specificity level, we can compare the sensitivity of the gated and ungated S-phase measurements for detecting women that will die by time t. Figure 2 constructs ROC curves at $t = 40$, $t = 60$, and $t = 100$. Panels (d)–(f) use the Kaplan–Meier estimator directly as described in Section 2.2 while panels (a)–(c) use the nearest neighbor estimator (NNE) of Section 2.3 with a span of $\lambda_n = 5\%$. A relatively small span was chosen to yield only moderate smoothing in order to facilitate comparison of the NNE estimator and the simple KM estimator. The ROC curves in (e) and (f) have monotonicity violations since the simple KM method does not guarantee a valid estimate of the bivariate distribution $F(c, t)$. The NNE estimates are very similar to the KM estimates but ensure that the sensitivity and specificity estimates are monotone.

At $t = 40$ and $t = 60$, the ROC curve for the new measurement dominates the ROC curve for the standard measurement. This implies that, for any fixed specificity, the gated measurement is a more sensitive marker. For example, using the gated measurement with a threshold of 5.4% for defining screen positivity, at 60 months, 71% of the surviving women are screen negative, $\hat{P}\{X_1 \leq 5.4 \mid D(60) = 0\} = 0.71$, and 82% of the women that died are screen positive, $\hat{P}\{X_1 > 5.4 \mid D(60) = 1\} = 0.82$. The threshold for the ungated measurement with the same specificity is 3.5%, $\hat{P}\{X_2 \leq 3.5 \mid D(60) = 0\} = 0.71$. However, the associated sensitivity is only 54%, $\hat{P}\{X_2 > 3.5 \mid D(60) = 1\} = 0.54$. At later times, the difference between the two ROC curves diminishes. By $t = 100$, the ROC curve for the new measure is nearly identical to the ROC curve for the standard measurement. Therefore, the gated S-phase measurement appears to be more sensitive at identifying women that die within the first 60 months but does not appear to provide improved discrimination for cumulative mortality at later follow-up times.

Using the bootstrap discussed in Section 2.4, we can test whether the ROC curve for the gated measurement is significantly different from the ROC curve for the ungated measurement. One statistic that is commonly used for comparing two ROC curves is the area under the curve (Hanley and McNeil, 1982; Delong, Delong, and Clarke-Pearson, 1988). The area under the curve at time t can be interpreted as the probability that X measured on a random case $(D(t) = 1)$ exceeds that for a random control $(D(t) = 0)$. Using the NNE estimator at $t = 60$, the area under the ROC for the gated measurement is 0.80 and the area under the ROC curve for the ungated measurement is 0.68. We sampled the $n = 253$ observations with replacement for 1000 bootstrap samples. For each sample, we computed the area under the ROC curve for the gated measurement, X_1, and the ungated measurement, X_2. The 95% confidence interval for the difference in the areas is (0.03, 0.26), indicating a significant difference at significance level $\alpha = 0.05$. At $t = 40$, the area under the ROC for X_1 is 0.83 and the area under the ROC for X_2 is 0.70. A bootstrap confidence interval for the difference in areas is $(-0.002, 0.26)$. At $t = 100$, the two areas are nearly identical: 0.72 for X_1 and 0.70 for X_2.

3.2 *Impact of Modified Eligibility Criteria for an HIV Trial*

In HIV prevention trials that use seroincidence as the primary outcome, the sample size or power is determined by the total number of incident infections observed (Fleming and Harrington, 1991). In the United States, HIV incidence is quite low (1–2%/year for gay men), and a large number of subjects may be required for any definitive clinical trial. If high-risk subsets of the population are identifiable, then smaller clinical trials may be feasible.

The operating characteristics of a trial are determined by the total number of participants and the expected number of infections by the end of the trial, $P\{D(t') = 1\}$, where $D(t)$ denotes infection status and t' denotes the maximum follow-up time. By modifying trial eligibility criteria, it may be possible to retain a large fraction of the incident infections yet screen out many participants that do not ultimately become infected. Such a modification can preserve the statistical power of a trial yet reduce the operational burden by decreasing the total sample size. Eligibility modifications can naturally be denoted as a restriction to the subpopulation satisfying $X > c$, where X is a subject characteristic and c is an eligibility threshold. Examples of characteristics that might be used in HIV prevention research include numbers of sexual contacts or frequency of intravenous drug use. By increasing the inclusion threshold, c, we potentially exclude incident infections and therefore may sacrifice power. The impact on power can be summarized by the proportion of infections that would still be observed if the criterion, $X > c$, were included for eligibility. This proportion is $P\{X > c \mid D(t) = 1\}$ or the sensitivity of the criterion, $X > c$. The practical impact of setting a higher inclusion criterion can be summarized by the reduction in the total sample size. For a rare outcome such as HIV, the sample size effectively corresponds to the number of trial participants that remain uninfected. We can assess the reduction in sample size resulting from eligibility modification, $X > c$, by calculating the fraction of subjects that continue to be infection free at study termination who satisfy the criterion $P\{X > c \mid D(t') = 0\}$ or 1 − specificity.

We analyze data from the Vaccine Preparedness Study (VPS) (Koblin et al., 1998) to illustrate the use of ROC curves for graphically displaying the potential impact of eligibility modification on power and sample size. In this study, $n = 3257$ HIV-negative gay men were followed prospectively for three semiannual visits. Baseline measurements included a risk assessment where factors known to correlate with HIV acquisition were measured, such as the total number of male partners in the last 6 months. Follow-up visits included HIV testing and counseling as well as risk assessment questionnaires. A total of 72 incident infections were observed during the 18 months of follow-up. Rates of follow-up were high with only 11% of the subjects not completing the 18 month visit. The primary goal of VPS was to obtain information necessary for the planning of HIV vaccine efficacy trials.

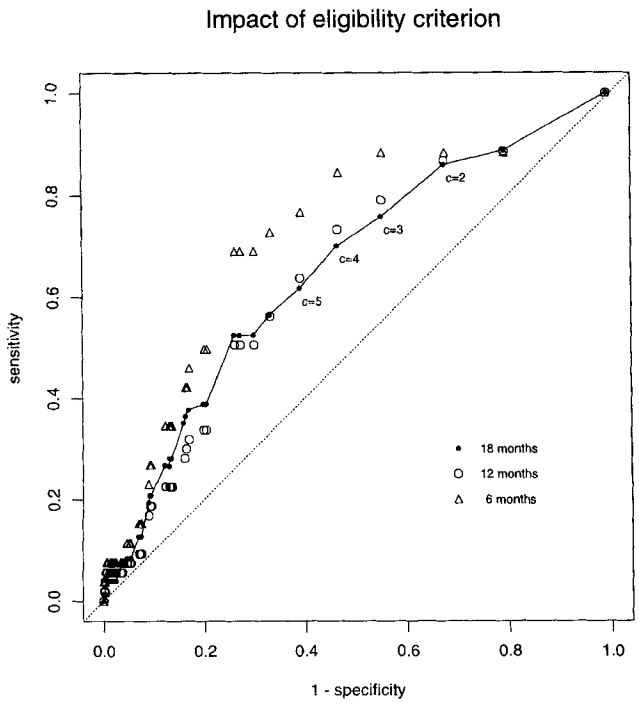

Figure 3. ROC curves displaying the impact of a modified eligibility criterion on an HIV prevention trial.

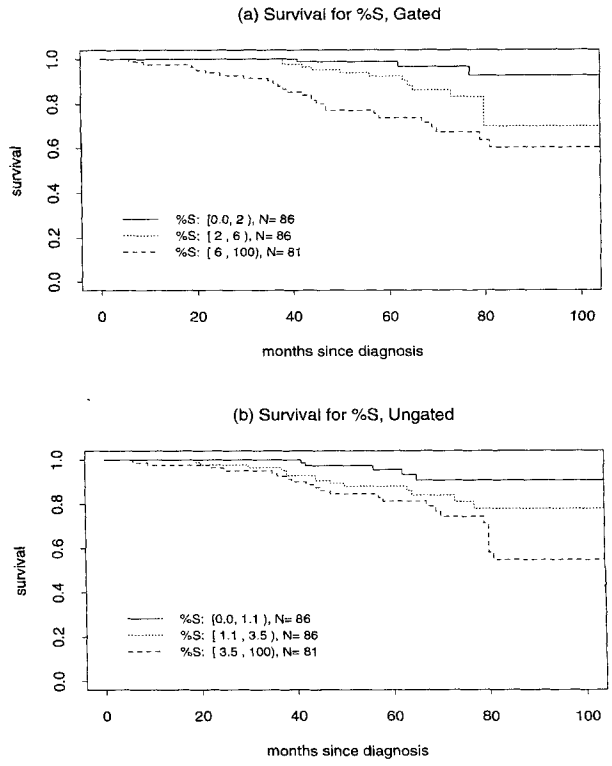

Figure 4. Kaplan–Meier curves for survival after breast cancer diagnosis. Panel (**a**) shows survival curves for groups of women based on tertiles of percent S-phase (%S) using the new measurement method, i.e., gated, and panel (**b**) shows survival for groups of women based on tertiles of the standard S-phase measurement, i.e., ungated.

Data from the VPS can be used to assess whether restriction of the enrollment criteria may be used to obtain more efficient study designs for prevention research. Figure 3 considers the impact of using the reported number of male partners at baseline as an eligibility criterion. For increasing values of this covariate, we plot an estimate of the percent of the infected participants (through time t) that exceed this criterion, $P\{X > c \mid D(t) = 1\}$ or sensitivity, and the percent of the participants that remain infection free (through time t) that would satisfy the criterion $P\{X > c \mid D(t) = 0\}$ or 1− specificity. Since very few observations were censored prior to 18 months, we used the simple Kaplan–Meier estimator to obtain estimates of sensitivity and specificity at $t = 6$, $t = 12$, and $t = 18$ months. We see that the baseline number of sexual partners is better at discriminating those infections that occur within the first 6 months than those that occur through 12 or 18 months. Since the number of sexual partners varies over time, we should expect a stronger association between the number of sexual partners at baseline and early incident infections rather than with those that occur later in time.

If a proposed trial were planned to be 18 months long, then we can use the ROC curve for cumulative infections though $t = 18$ to summarize the potential impact of an eligibility modification. From Figure 3, we see that an eligibility criterion of more than two male partners would include 86% of the infected participants and only 67% of the remaining uninfected participants. If power considerations required, e.g., 75 incident infections in the control arm, then without eligibility modification, 2500 men in each arm would need to be studied for 18 months if the HIV incidence was 2%/year (with perfect follow-up). Adoption of the eligibility criterion $X > c$, where X represents the reported number of male partners in the previous 6 months, would require $2500 \times 0.67/0.86 = 1948$ men in each study arm, a 22% reduction in sample size. Here the ratio $R = (1 - \text{specificity})/\text{sensitivity}$ denotes the overall reduction due to first restricting to the subcohort with $X > c$ (approximately 1 − specificity when $P\{D(t) = 1\}$ is small) and then increasing the sample size drawn from this subcohort to recover the required total infections (1/sensitivity). Alternatively, using more than four partners leads to inclusion of 70% of the infected participants and inclusion of only 53% of the uninfected participants, for a reduction of $(1 - 0.53/0.70) = 24\%$. These calculations assume that the treatment effect does not depend on the variable used for restriction. In practice, the relative costs of recruitment versus follow-up and potential impacts on scientific generalizability would need to be weighed when considering adoption of any criterion $X > c$.

4. Discussion

We have suggested the use of ROC curves for evaluating and comparing the prognostic capacity of diagnostic markers measured at baseline for disease outcomes that are naturally treated as a counting process, $D(t)$, where $D(t) = 0$ at times prior to disease onset and $D(t) = 1$ after disease onset. In our first example, the time-dependent outcome was mortality,

and in our second example, it was the time until HIV infection. ROC curves are already standard for characterizing diagnostic accuracy when the outcome is not time dependent or when the marker and disease status variables are measured at the same time. Extension of ROC methods to the time-dependent setting where a time lag exists between the measurement of the marker and the onset of disease is natural.

Other approaches to summarizing the prognostic value of a marker measurement are also feasible. For example, survivor functions associated with interval categories of X at baseline, $X \in \mathcal{G}_j$, might be used. Figure 4 shows the Kaplan–Meier curves for survival after diagnosis with breast cancer for the data analyzed in Section 3.1. The covariate groupings, $\mathcal{G}_j$, are defined by the tertiles of the gated S-phase measurement in panel (a) and by the tertiles of the ungated S-phase measurement in panel (b). A difficulty with this approach is that the categorizations of X are often somewhat arbitrary and may not yield range categories that are comparable for different markers. ROC curves do not require arbitrary classification of covariates and do provide a common scale for comparison of different markers. It should also be noted that the role of ROC curves is typically in the development of screening or diagnostic modalities where performance over a range of possible threshold values for X needs to be assessed. Such studies precede definitive evaluation of well-defined screening criteria (with a fixed threshold for screen positivity) wherein actions associated with screen positivity and related costs would be important outcomes.

We introduce methods for estimating ROC curves when disease incidence is considered time-dependent. A simple estimator based on Bayes' theorem and Kaplan–Meier methods turned out to have some serious shortcomings when used with censored observations. This motivated us to develop an alternative approach based on a valid bivariate survivor function estimator. The estimator that we use is attractive since it is semiparametric efficient and allows the censoring process to depend on the marker (Akritas, 1994). We used bootstrap techniques for confidence interval estimation and hypothesis testing. However, it is also possible to use asymptotic distribution theory for inference, building on the results for $\hat{S}_{\lambda_n}(c,t)$ provided by Akritas (1994). We note that the ROC curve at time t, as a function of the (1 − specificity) domain $(0,1)$, can be written as $\text{sensitivity}(f) = S_1\{S_0^{-1}(f)\}$, where S_1 and S_0 denote the survivor functions for X in the diseased and nondiseased populations, respectively (Pepe, 1997). Therefore, distribution theory for ROC estimates can be derived from the joint distribution of the functions $\hat{S}_1(c) = \hat{P}\{X > c \mid D(t) = 1\}$ and $\hat{S}_0(c) = \hat{P}\{X > c \mid D(t) = 0\}$, which are obtained from $\hat{S}_{\lambda_n}(c,t)$. Our estimators assume that the data are derived from a cohort study where sampling does not depend on the disease outcome $D(t)$. Development of methods that could be used when sampling does depend on the data, such as with case–control or case–cohort studies, would also be of interest.

Acknowledgements

We would like to thank Peggy Porter, Jeri Glogovac, and Janet Daling from the Fred Hutchinson Cancer Research Center for permission to use the breast cancer cytometry and mortality data. Funding for this research was provided by National Institutes of Health contract N01 AI45200 and NIH grant R01 GM54438.

Résumé

Les courbes ROC sont une méthode répandue pour représenter la sensibilité et la spécificité d'un marqueur diagnostic continu, X, d'une variable de maladie binaire, D. Cependant, les critères de morbidité dépendent souvent du temps, $D(t)$, et des courbes ROC qui varient en fonction du temps peuvent être mieux adaptées. Un exemple usuel de variable dépendant du temps est l'état vital, où $D(t) = 1$ si le patient est décédé avant le temps t, et 0 sinon. Nous proposons de résumer l'aptitude à discriminer d'un marqueur X, mesuré au départ ($t = 0$), en calculant les courbes ROC pour l'incidence cumulée de maladie ou de décès avant le temps t, notées ROC(t). Une complication typique avec les données de survie est que les observations peuvent être censurées. Nous proposons deux estimateurs de courbes ROC qui prennent en compte la censure. Un estimateur simple est basé sur l'utilisation de l'estimateur de Kaplan-Meier pour chacun des sous-ensembles possibles $X > c$. Cependant, cet estimateur ne garantit pas la condition nécessaire que la sensibilité et la spécificité sont monotones en X. Un autre estimateur qui garantit bien la monotonie est basé sur l'estimateur du plus proche voisin pour une fonction de distribution bivariée de (X,T), où T représente le temps de survie (Akritas, 1994). Nous présentons un exemple où ROC(t) est utilisé pour comparer des mesures de flux cytométrique standards et modifiées pour prédire la survie après détection d'un cancer du sein, et un exemple où la courbe ROC(t) montre l'impact d'une modification des critères d'éligibilité sur la taille de l'échantillon et la puissance dans des essais de prévention HIV.

References

Akritas, M. G. (1994). Nearest neighbor estimation of a bivariate distribution under random censoring. *Annals of Statistics* **22,** 1299–1327.

Begg, C. G. (1987). Biases in the assessment of diagnostic tests. *Statistics in Medicine* **6,** 411–423.

Begg, C. G. (1991). Advances in statistical methodology for diagnostic medicine in the 1980's. *Statistics in Medicine* **10,** 1887–1895.

DeLong, E. R., DeLong, D. M., and Clarke-Pearson, D. L. (1988). Comparing the areas under two or more receiver operating characteristic curves: A nonparametric approach. *Biometrics* **44,** 837–845.

Efron, B. (1967). The two-sample problem with censored data. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability,* Volume 4, L. Le Cam and J. Neyman (eds), 831–853. Berkeley: University of California Press.

Fleming, T. R. and Harrington, D. P. (1991). *Counting Processes and Survival Analysis.* New York: Wiley.

Hanley, J. A. (1989). Receiver operating characteristic (ROC) methodology: The state of the art. *Critical Reviews in Diagnostic Imaging* **29,** 307–335.

Hanley, J. A. and McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* **143,** 29–36.

Kaplan, E. L. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association* **53,** 457–481.

Koblin, B., Heagerty, P., Sheon, A., Metzger, D., Buchbinder, S., Celum, C., Critchlow, C., Douglas, J., Gross, M., Marmor, M., Mayer, K., and Seage, G. (1998). Readiness of high-risk populations in the HIV network for prevention trials to participate in HIV-1 vaccine efficacy trials in the United States. *AIDS* **42,** 785–793.

Metz, C. E., Herman, B. A., and Shen, J. (1998). Maximum likelihood estimation of receiver operating characteristic (ROC) curves from continuously distributed data. *Statistics in Medicine* **17,** 1033–1053.

Pepe, M. S. (1997). A regression modelling framework for ROC curves in medical diagnostic testing. *Biometrika* **84,** 595–608.

Pepe, M. S., Leisenring, W., and Rutter, C. (1999). Evaluating diagnostic tests in public health. In *Handbook of Biostatistics,* Volume 18, C. R. Rau and P. K. Sen (eds), 397–422. New York: Elsevier Scientific.

Sigurdsson, H., Baldetorp, B., Borg, A., Dalberg, M., Fernö, M., Killander, D., and Olsson, H. (1990). Indicators of prognosis in node-negative breast cancer. *New England Journal of Medicine* **322,** 1045–53.

Swets, J. A. and Pickett, R. M. (1982). *Evaluation of Diagnostic Systems: Methods from Signal Detection Theory.* New York: Academic Press.

van der Vaart, A. W. and Wellner, J. A. (1996). *Weak Convergence and Empirical Processes with Applications to Statistics.* New York: Springer.

Zhou, K. H., Hall, W. J., and Shapiro, D. E. (1997). Smooth non-parametric receiver operating characteristic (ROC) curves for continuous diagnostic tests. *Statistics in Medicine* **16,** 2143–2156.

Zweig, M. H. and Campbell, G. (1993). Receiver-operator characteristic plots: A fundamental evaluation tool in clinical medicine. *Clinical Chemistry* **39,** 561–577.

Received December 1998. *Revised July* 1999.
Accepted August 1999.