Lifetime Data Anal (2008) 14:86–113
DOI 10.1007/s10985-007-9073-x

Evaluating the ROC performance of markers for future events

**Margaret S. Pepe · Yingye Zheng · Yuying Jin ·
Ying Huang · Chirag R. Parikh · Wayne C. Levy**

Received: 9 November 2007 / Accepted: 14 November 2007 / Published online: 7 December 2007

Abstract Receiver operating characteristic (ROC) curves play a central role in the evaluation of biomarkers and tests for disease diagnosis. Predictors for event time outcomes can also be evaluated with ROC curves, but the time lag between marker measurement and event time must be acknowledged. We discuss different definitions of time-dependent ROC curves in the context of real applications. Several approaches have been proposed for estimation. We contrast retrospective versus prospective methods in regards to assumptions and flexibility, including their capacities to incorporate censored data, competing risks and different sampling schemes. Applications to two datasets are presented.

M. S. Pepe (✉) · Y. Zheng
Biostatistics and Biomathematics, Fred Hutchinson Cancer Research Center,
1100 Fairview Ave N., M2-B500, Seattle, WA 98109, USA
e-mail: mspepe@u.washington.edu

Y. Zheng
e-mail: yzheng@fhcrc.org

Y. Jin · Y. Huang
Department of Biostatistics, University of Washington, Box 357232, F-600 HSB, Seattle,
WA 98195, USA

Y. Jin
e-mail: jinyu@u.washington.edu

Y. Huang
e-mail: ying@u.washington.edu

C. R. Parikh
Section of Nephrology, Department of Internal Medicine, Yale University School of Medicine,
P.O. Box 208029, New Haven, CT 06520-8029, USA
e-mail: chirag.parikh@yale.edu

W. C. Levy
University of Washington Medical Center,
1959 N.E. Pacific Street, Campus Box 356422 Seattle, WA 98195-6422, USA
e-mail: levywc@u.washington.edu

Keywords Prediction · Diagnostic test · Prognosis · Sensitivity · Specificity

1 Introduction

Techniques for analyzing event time data are now routinely applied in biomedical research. In particular, regression models such as Cox regression are often fit to data in order to study the effects of predictors on risk of a future event. However, Sam Wieand was amongst those who recognized that risk models do not address the potential of a predictor to distinguish between those who will have an event and those who will not (Emir et al. 1998). Indeed a marker can have a large effect on risk yet perform poorly as a discriminator.

For a binary outcome variable the classification accuracy of a marker is typically quantified with the true and false positive fractions. The former, TPF = P(marker positive|outcome positive), is the probability of correctly classifying a subject with positive outcome and the latter, FPF = P(marker positive|outcome negative), is the probability of incorrectly classifying a subject with negative outcome. The benefit of true positive classifications is gained at the cost of false positive classifications. When the predictive marker, Y, is continuous, thresholding criteria, $Y > c$, are used to define marker positivity. The ROC curve is the standard summary of classification performance for continuous markers (Baker 2003). It plots the TPF versus FPF for all thresholds c. When multiple predictors are involved, the marker is naturally defined as the risk function, $Y =$ P[outcome positive|predictors], or equivalently as any monotone increasing function of it. It is the optimal combination of predictors for classification (McIntosh and Pepe 2002), a result that follows from the Neyman-Pearson lemma. See Zheng et al. (2006) for related results pertaining to event time outcomes.

In this paper we consider marker performance for outcomes that are not simply binary but that are event time outcomes. A recent paper in the *New England Journal of Medicine* on biomarkers for cardiovascular events (Wang et al. 2006) stated that "standard methods do not exist for deriving ROC curves for time-to-event data." In fact several approaches have been proposed. However the literature is scattered and a standard approach has not emerged. Here, we review existing methods and discuss issues that lead to preference for one method over another.

2 Applications

We describe two applications that exemplify key issues in evaluating the performance of markers for event time outcomes.

2.1 Markers of acute kidney injury

Patients who undergo major cardiac surgery are at high risk of suffering kidney injury due to interruption of blood flow to the kidneys during surgery. Although monitoring of rise in serum creatinine is the standard approach to detecting acute kidney injury (AKI), it has an important drawback. The serum creatinine increase is typically delayed by 1 – 3 days due to several non-kidney factors that affect its production and equilibration.

Biomarkers are sought to detect AKI earlier than serum creatinine so that appropriate treatment can be promptly initiated. Two such markers measured in urine are currently under investigation in a multicenter study of 1,800 patients undergoing major cardiac surgery. Urine samples are taken at various intervals after surgery, frozen and stored. At the end of the study these stored specimens will be assayed for the new markers. Serum creatinine is monitored in these patients as part of their routine clinical care. An AKI event occurs when the creatinine level increases by 25% over preoperative levels and is sustained for 24 h. Patients with AKI are classified as having a severe event if the level reaches $>200\%$ of preoperative serum creatinine during the course of their clinical care. Otherwise the event is classified as mild.

Approximately 80% of patients recover from surgery without AKI or other devastating events and are discharged 3–5 days after surgery. We expect that 20% will experience an AKI event, 12% mild and 8% severe. An additional small group of patients, $<1\%$, will die from complications associated with their disease or surgery without meeting criteria for AKI. We call these non-AKI deaths.

Some questions of interest in this study are: to determine the numbers of patients for whom the diagnosis of AKI can be advanced with the new markers, by how long and at what cost in terms of false diagnoses. The analysis needs to accommodate competing risk events due to non-AKI deaths, the varying degrees of severity of AKI and the longitudinal nature of the biomarkers. The data will not be censored as all subjects are followed closely until discharge or death.

2.2 Seattle Heart Failure study

More than 5 million people in the United States have heart failure, a medical condition where the heart fails to maintain adequate circulation. However, outcomes are highly variable and annual mortality estimates range from 5% to 75%. The Seattle Heart Failure (SHF) study (Levy et al. 2006) combined data from 6 cohorts of patients. One cohort, the PRAISE study, (Packer et al. 1996) was used to develop a 'risk of mortality' model with Cox regression that included easily obtainable baseline information relating to clinical status, therapy and laboratory parameters. The other cohorts were used to evaluate the accuracy of this model.

Here we consider evaluating the performance of the linear predictor obtained from PRAISE, the SHF score, to discriminate people who die in the first 2 years from those who do not. We use the largest independent validation cohort (Val-heft) considered by Levy et al. This cohort of 5,010 patients was derived from a randomized trial of the angiotensin-receptor blocker valsartan (Cohn and Tognoni 2001). Characteristics of these subjects have been reported. There were 979 deaths and the mean followup was 1.9 years. Survival estimates at 1, 2 and 3 years were 91%, 81.6% and 71.7%, respectively. These numbers agree well with the survival rates predicted by the PRAISE risk model (Levy et al. 2006).

In this application the outcome, death, is a classic event outcome qualified only by the time at which it occurs but not by severity. There is censoring due almost entirely to gradual enrollment into the study over time and not due to loss to followup. There are no competing risk events and the marker, i.e., the risk score, is measured only at the baseline enrollment time.

3 Definitions of ROC for event time outcomes

3.1 Time dependent TPF

Consider first the simplest scenario where the marker Y is binary and measured at a baseline time $t = 0$. Cases are subjects with the event of interest and let T denote the event time for a case. One definition of time-dependent true positive fraction is

$$\mathrm{TPF}(t) = \mathrm{Prob}(Y = 1|T = t).$$

This definition allows the sensitivity of the marker to depend on the time that the event occurs. Certainly in the Kidney Biomarkers study, baseline (day 0) biomarkers are likely to be more sensitive to AKI diagnosed on day 1 rather than AKI diagnosed on day 3 since it is more likely that the kidney injury was not present at baseline if it was not picked up by serum creatinine until day 3. In the heart failure study, it is likely that people who die early have more extreme levels of baseline risk factors than subjects who die later. Therefore the TPF associated with the risk score is likely to be higher for earlier events than for later events. In the context of cancer screening, biomarker levels tend to be higher for subjects with larger subclinical tumor and those are likely to manifest clinically sooner. Thus again, the TPF is likely to be a decreasing function of t.

Now suppose that the marker is measured at time s, such as in longitudinal studies or when there is no natural baseline time. We write

$$\mathrm{TPF}(t) = \mathrm{Prob}(Y(s) = 1|T = t + s)$$

the sensitivity of the marker to events that occur t units after Y is measured. In some applications this sensitivity could depend not only on the time lag, t, but also on s the absolute time of measurement. For example, if s denotes age or time after intervention then the TPF could depend on t and s, and we would write

$$\mathrm{TPF}(t, s) = \mathrm{Prob}(Y(s) = 1|T = t + s).$$

Heagerty and Zheng (2005) introduced a taxonomy for time dependent measures of accuracy. The $\mathrm{TPF}(t)$ defined above is the *incident* true positive fraction and is the version adopted by most methodologists (Heagerty and Zheng 2005; Etzioni et al. 1999; Cai et al. 2006; Zheng and Heagerty 2004; Song and Zhou in press). An alternative version is the cumulative TPF:

$$\mathrm{TPF}_c(t, s) = \mathrm{Prob}(Y(s) = 1|s < T <= t + s),$$

which evaluates sensitivity for events that occur throughout the time interval $(s, s+t]$ as opposed to events that occur at t time units after Y is measured. This definition is used in many applied papers (e.g., Wang et al. 2006) in part because it is easily estimated empirically. Specifically in uncensored data an estimate is the simple proportion of subjects with events in the time interval who have positive marker values t time units prior to their events. Estimators that accommodate censoring have been developed (Heagerty et al. 2000; Delong et al. 1985; Parker and Delong 2003; Zheng and Heagerty 2007; Song and Zhou in press). Cai et al. 2006 focused on the incident TPF but noted that the cumulative TPF, which may be of interest clinically, can be calculated directly from it as $\text{TPF}_c(t, s) = \int(\text{TPF}(u, s)dF_T(u))/[F_T(t+s) - F_T(s)]$ where F_T is the cumulative distribution of the event time. On the other hand, estimating the incident on the basis of a cumulative TPF estimate is more difficult, in our opinion, because differentiation is harder numerically than is integration. In addition, by lumping all events in $(s, s+t]$ together, the cumulative TPF does not distinguish between sensitivity to events that occur early versus late in the interval. Moreover, a series of cumulative TPFs indexed by t shows redundant information in the sense that $\text{TPF}_c(t_2, s)$ includes events in $\text{TPF}_c(t_1, s)$ if $t_1 < t_2$, while a series of incident TPFs show essentially different information in each. We focus on estimating the incident TPF in the remainder of this article.

In some settings exact event times are unknown but interval censored versions are known instead. For example, testing for a subclinical condition may be done monthly. Then T must be considered a discrete variable, indicating time intervals. With uncensored data, the incident TPF is easily calculated as a proportion.

3.2 The FPF and ROC

In the classic setting with binary outcomes, the false positive fraction, $\text{FPF} = P(Y = 1|D = 0)$, is the fraction of controls that test positive. Who are the controls when the outcome is a failure time? One sensible definition is to consider controls to be those individuals for whom a positive test is an error. A natural control group emerges in some scenarios. In the Kidney Biomarkers study, controls are those 80% of patients who recover from surgery and are discharged without experiencing AKI or other devastating events. In cancer screening, ideal controls would be individuals who would never be diagnosed with life threatening cancer in their lifetimes in the absence of screening. In practice lifetime follow up is rarely available for all subjects, so subjects without cancer at the time of analysis are often considered an approximate control group (Etzioni et al. 1999).

Definition of control status is more problematic when all subjects eventually have the event of interest. Since everybody dies, the Seattle Heart Failure study does not have a natural control group. One possibility is to choose a large landmark time point, τ, and define controls as subjects with $T > \tau$. The optimal choice for τ should be context dependent. For example, if the intention is to monitor individuals at time intervals of length, δ, considering that intervention will be adequate if administered at time δ before the event, then the choice $\tau = 2\delta$ would be sufficient. Subjects for whom $T > 2\delta$, do not need to test positive because they can still be tested and treated adequately with future monitoring. In the Seattle Heart Failure study, however, the optimal choice of τ is not clear. Moreover, limitations of the data may limit the possibilities. We choose $\tau = 2$ years in part because few subjects were followed beyond 2.5 years. We acknowledge that with this choice of τ, the controls are better described as a reference group against which to compare subjects with earlier events, rather than as a true control group.

Heagerty and Zheng (2005) call the FPF defined above,

$$\text{FPF}(s) = \text{Prob}(Y(s) = 1|T > s + \tau),$$

the *static* false positive fraction. An alternative is to allow the FPF to vary with the time lag t since marker measurement $\text{FPF}_d(t, s) = \text{Prob}(Y(s) = 1|T > s + t)$, the proportion of positive tests among subjects without events by t time units after the marker measurement time, is called the *dynamic* FPF (Heagerty and Zheng 2005; Zheng and Heagerty 2007). This quantity can sometimes misrepresent the accuracy of a biomarker. Consider that subjects with an event shortly after the time lag t are counted as dynamic controls at t. A positive test for such a subject is counted against the biomarker, as a false positive. Yet perhaps it should count in favor of the test's ability to flag future events.

For a continuous marker $Y(s)$, the time dependent ROC curve compares cases with events at time t to controls. In particular the proportion of cases with marker values exceeding $c(s)$, the $(1 - f)$ quantile of $Y(s)$ in controls, is displayed versus f, the false positive fraction. Another practical problem with dynamic FPFs is that because the control groups vary with time so too does the x-axis of the corresponding ROC curves. It therefore becomes more difficult to interpret trends over time in time-dependent ROC curves. With a static control group, time trends in ROCs relate to trends in the detection of events. However, such trends may be due to a combination of changing control groups and changing detection properties when ROC curves use dynamic controls. Indeed consider that when using dynamic FPFs, even if the TPF associated with a specific thresholding rule, $I(Y(s) \geq c)$, is constant over time, the ROC curves will appear to increase with larger t as the control groups drop subjects with larger values of $Y(s)$ who have events. We will focus on the static FPF in the remainder of this paper. For applications like the Kidney Biomarker study, where there is a natural control group that is not defined solely by time, we will assign a fictitious event time larger than τ to controls, so that we can use uniform notation.

If F denotes the cdf for $Y(s)$ in the control group, the time-dependent ROC curve is written mathematically as

$$\text{ROC}_{t,s}(f) = \text{Prob}(Y(s) \geq c(s)|T = s + t)$$

where $c(s) = F^{-1}(1 - f)$, $f \in (0, 1)$. That is $ROC_{t,s}(f)$ is the TPF(t, s) corresponding to an FPF$(s) = f$.

We emphasize that the marker at time s, $Y(s)$, may be a function of marker history up to time s, and is not necessarily the value of a single measurement at time s. Moreover, the distribution of $Y(s)$ may vary with s, for example when the time scale s is age or time after an intervention, so the threshold $c(s)$ may depend on s. In some applications discrimination achieved with the marker may depend on the absolute time scale s as well as on the time lag t, and our notation allows this level of generality.

3.3 Censoring and competing risk events

Censoring is often but not always an issue in prospective studies with event time outcomes. It arises in the Seattle Heart Failure study but not in the Kidney Biomarker study. Censoring is a nuisance in the data and clearly should not impact on the definitions for TPF and FPF. The common simple practice of including all subjects without events in the FPF calculation (Wang et al. 2006) is flawed in part because some of the included censored subjects may have events, thus contaminating the control group. Proper accommodation of censoring is discussed in Sect. 4.

On the other hand, competing risk events are real phenomena that occur in the population and should therefore impact on (TPF, FPF) definitions. Should subjects with competing risk events be considered cases or controls? In the Kidney Biomarker study, it is possible that the biomarkers will be predictive of competing risk events. One might consider them a second case group and evaluate the markers in them separately. That is, two separate ROC curves could be estimated: one of primary interest that compares subjects with AKI events to the controls and one of secondary interest that compares subjects with competing risks to controls. An alternative is to include them in the control group. This would only be warranted if flagging such subjects as positive would lead to clinically erroneous decisions. Even then, it might still be of interest to compare them with the other controls so that one can interpret the overall false positive fractions in terms of the components from each type of control group.

4 Estimation from data

Approaches to estimating biomarker performance parameters can be classified broadly as prospective or retrospective. Each class has its own strengths.

4.1 Retrospective methods

Consider the simplest setting where $Y(s)$ is a binary marker and there is no censoring. We write the data for controls as $\{Y_j(s_{jk}), s_{jk}; k = 1, \ldots, n_j; j = 1, \ldots, n_{\bar{D}}\}$ and the data for cases as $\{Y_i(s_{ik}), s_{ik}; k = 1, \ldots, n_i; T_i; i = 1, \ldots, n_D\}$. Leisenring et al. proposed simple binary regression methods for this setting (Leisenring et al. 1997). One can model FPF(s) as a parametric function of s using the control data to fit model parameters. The jth control contributes n_j data records of the form $(Y_j(s_{jk}), s_{jk})$. Similarly TPF(t, s) can be modeled as a parametric function of (t, s) and fit with data for cases. Each case contributes n_i data records of the form $(Y_i(s_{ik}), s_{ik}, t_{ik})$ where the time lag is $t_{ik} = T_i - s_{ik}$. The time lag varies across data records depending on the biomarker measurement time s_{ik}. Standard errors of parameter estimates are based on sandwich variance estimates to account for correlation between records from the same subject. In their application to a new test for cytomegalovirus infection in bone marrow transplant patients, although $Y(s)$ varied over time for individuals, the distribution of $Y(s)$ did not vary with s, the time since transplant. They therefore reported the overall FPF and the monotone decreasing function TPF(t) modeled as

$$\text{TPF}(t) = g(\alpha + \beta\eta(t)) \tag{1}$$

where g^{-1}, the link function, was logistic and $\eta(t)$ was a set of polynomial basis functions.

For a continuous marker, again in the absence of censoring, Etzioni et al. (1999) extended the binary regression approach. To simplify notation we suppose, as in Etzioni et al., that the time dependent ROC curves do not depend on s. They modeled

$$\text{ROC}_t(f) = g(h(f) + \beta\eta(t)) \tag{2}$$

where g^{-1} is the link function and $g(h(f))$ is the baseline ROC curve at $t = 0$. They implemented the method on data from a prostate cancer study, estimating the distribution of $Y(s)$ nonparametrically in controls and using a parametric form for h, namely $h(f) = a_0 + a_1\Phi^{-1}(f)$. Cai and Pepe (2002) allowed nonparametric baseline function h.

Cai et al. (2006) offers the most comprehensive of existing retrospective approaches, encompassing previous methods and extending them to censored failure time data. With binary markers, functions to be estimated are FPF(s) and TPF(t, s). Uncensored subjects enter into the analysis as in the Leisenring et al. approach, as a case if $0 < T - s \leq \tau$ and as a control if $T - s > \tau$. Censored subjects, censored at X, enter as either a control if $X - s > \tau$ or otherwise as weighted averages of cases and controls. In the latter case observe that since $X - s \leq \tau$

$$\begin{aligned}\text{Prob}(Y(s) = 1|T > X) = {} & \text{FPF}(s)P(T > s + \tau|T > X)\\ & + \int_{X-s}^{\tau} \text{TPF}(s, t)dP(T \leq t + s|T > X).\end{aligned}$$

Therefore the "likelihood contribution" for $Y(s)$ is a weighted average of FPF(s) and TPF(s, t) for time lags t in $(X - s, \tau)$. The weights are easily determined by estimating the distribution of T with standard failure time methods. Note that if competing risk events exist the distributions should be estimated with cumulative incidence methods (Kalbfleisch and Prentice 1980, p. 169) rather than treating them as censoring events.

For continuous biomarkers, Cai et al. adopt the ROC-GLM model (2) with nonparametric baseline ROC curve h. Similar to Etzioni et al, the approach is nonparametric with respect to the distribution of $Y(s)$ in controls as well. It can be implemented by replacing each biomarker record, $Y(s)$, with a series of P binary variable records of the form $I(Y(s) > c_p)$, corresponding to biomarker thresholds $c_1, \ldots, c_P$. The algorithm for binary markers is then applied with a series of FPFs, $\{\text{FPF}_1(s), \text{FPF}_2(s), \ldots, \text{FPF}_P(s)\}$, corresponding to the thresholds estimated in this approach. In addition a series of intercepts in (1), $\{\alpha_1, \alpha_2, \ldots, \alpha_P\}$ that correspond to the P thresholds are estimated. These are interpreted as $\{h(\text{FPF}_1(s)), \ldots, h(\text{FPF}_P(s))\}$. See the appendix for details.

4.2 Prospective methods

Risk regression techniques are well established for modeling event time data and they naturally accommodate censoring. After fitting a prospective model one can combine it with observed predictor distributions to calculate TPF and FPF parameters.

Heagerty and Zheng (2005) employ a Cox model for a baseline marker, Y:

$$\lambda(t) = \lambda_0(t)\exp(\gamma(t)Y),$$

where the regression parameter γ may depend on t. Fitting the model to a simple random sample $\{(Y_i, T_i), i = 1, \ldots, n\}$, they note that for a binary marker and denoting the risk set at t, by $R(t)$,

$$\widehat{\text{TPF}}(t) = \frac{\sum_{i \in R(t)} Y_i \exp(\widehat{\gamma}(t)Y_i)}{\sum_{i \in R(t)} \exp(\widehat{\gamma}(t)Y_i)}$$

is a consistent estimate of TPF(t). This follows from the observation (Xu and O'Quigley 2000) that under the Cox model, the distribution of $Y\exp(\gamma(t)Y)$ for subjects in the risk set $R(t)$ is equal to the conditional distribution of Y given $T = t$. To estimate FPF, they employ the empirical estimate in the controls in the risk set at τ:

$$\widehat{\text{FPF}} = \sum_{i \in R(\tau)} Y_i / n(\tau)$$

where $n(\tau)$ is the size of the risk set. With continuous biomarkers let the empirical distribution of Y for subjects in the risk set at τ be $\widehat{F}_\tau$, then

$$\widehat{\text{ROC}}_t(f) \equiv \frac{\sum_{i \in R(t)} I(Y_i \geq \widehat{F}_\tau^{-1}(1-f))\exp\{\widehat{\gamma}(t)Y_i\}}{\sum_{i \in R(t)} \exp\{\widehat{\gamma}(t)Y_i\}}$$

Song and Zhou (in press) employ the same data structure but a simplified model with non-time dependent parameter $\gamma(t) = \gamma$. They use Bayes' theorem to write TPF(t) and FPF for a binary marker:

$$\begin{aligned}
&\text{TPF}(t) \\
&\quad = P(Y=1)P(T=t|Y=1)/P(T=t) \\
&\quad = P(Y{=}1)\frac{\lambda_0(t)\exp(\gamma)\exp(-\Lambda_0(t)\exp(\gamma))}{\lambda_0(t)\exp(\gamma)\exp(-\Lambda_0(t)\exp(\gamma))P(Y{=}1){+}\lambda_0(t)\exp(-\Lambda_0(t))P(Y{=}0)} \\
&\quad = \text{logit}^{-1}\{\text{logit}P(Y=1) + \gamma + \Lambda_0(t)(1-\exp(\gamma))\},
\end{aligned}$$

where $\text{logit}^{-1}(x) = \exp(x)/(1 + \exp(x))$ and Λ_0 is the cumulative baseline hazard function;

$$\begin{aligned}\text{FPF} &= P(Y = 1)P(T > \tau|Y = 1)/P(T > \tau)\\ &= \text{logit}^{-1}\{\text{logit}P(Y = 1) + \Lambda_0(t)(1 - \exp(\gamma))\}.\end{aligned}$$

Observe that if $\gamma = 0$ then $\text{TPF}(t) = \text{FPF} = P(Y = 1)$, which is an intuitive result. Under our convention that larger Y is associated with larger hazard rate, $\gamma > 0$, and we have that larger baseline hazard leads to smaller TPF and FPF. On the other hand, if events are rare, i.e., $\Lambda_0(\tau) \approx 0$, we have FPF$\approx P(Y = 1)$, the proportion of positive markers in the population at baseline and $\text{TPF}(t) \approx \text{logit}^{-1}\{\text{logit}P(Y = 1) + \gamma\}$, which does not depend on t.

With continuous marker, integrals over the distribution of Y enter into the TPF and FPF expressions corresponding to the thresholded marker, $I[Y \geq y]$:

$$\begin{aligned}\text{TPF}(t) = 1 - F_{D,t}(y) &\equiv \frac{\int_y^\infty \exp(\gamma Y)\exp\{-\Lambda_0(t)\exp(\gamma Y)\}dF(Y)}{\int_{-\infty}^\infty \exp(\gamma Y)\exp\{-\Lambda_0(t)\exp(\gamma Y)\}dF(Y)}\\ \text{FPF} = 1 - F_\tau(y) &= \frac{\int_y^\infty \exp\{-\Lambda_0(\tau)\exp(\gamma Y)\}dF(Y)}{\int_{-\infty}^\infty \exp\{-\Lambda_0(\tau)\exp(\gamma Y)\}dF(Y)}.\end{aligned}$$

Song and Zhou substitute $\widehat{\gamma}$, $\widehat{\Lambda}_0$ and the empirical distribution of Y, $\widehat{F}$, into the above expressions to estimate TPF(t) and FPF. The ROC curve estimator is calculated as $\widehat{\text{ROC}}_t(f) = 1 - \widehat{F}_{D,t}(\widehat{F}_\tau^{-1}(1 - f))$.

Song and Zhou's method has two advantages over the Heagerty and Zheng approach. First, it was shown to be more efficient in simulation studies (Song and Zhou in press). This is likely due to its employment of the maximum partial likelihood estimators (mple) for γ and Λ_0 and so the corresponding estimators of (TPF, FPF) are also mple. In contrast the estimator of TPF(t) employed by Heagerty and Zheng is not the mple. Moreover their empirical estimator of FPF does not utilize the structure conferred by the Cox-model. Song and Zhou utilize this structure in estimating FPF. The second advantage concerns censoring. Heagerty and Zheng's methods cannot allow censoring to depend on Y. Subjects at risk at t must be representative of the "at risk" population in regards to the predictor distribution. The Song and Zhou approach only utilizes the baseline marker distribution and parameters of the risk model. The latter are consistently estimated under standard censoring assumptions that allow follow-up to depend on modeled predictors. Hence Song and Zhou's approach is valid even if censoring depends on the marker Y. However, Song and Zhou's method is only valid when the proportional hazards assumption is satisfied whereas Heagerty and Zheng extend their approach to allow estimation under nonproportional hazards.

4.3 Comparisons of attributes

Among the retrospective methods, Cai et al. (2006) is the most comprehensive. Other retrospective methods can be viewed as special cases of Cai's. Therefore we compare it with the two prospective approaches, Song and Zhou's method and Heagerty and Zheng's. We use the notation Cai, S+Z, H+Z for the three methods below.

4.3.1 Perspectives

The true and false positive fractions are defined as retrospective quantities in the sense that they concern the distribution of Y conditional on outcome. In our opinion a retrospective analysis seems like the more natural and direct approach to estimating them. Moreover, parameters relating to t in the retrospective approach, i.e., β in (1) and (2), directly quantify how performance varies with t. Inference about these parameters is straightforward with the retrospective approach. In contrast parameters in the prospective models do not directly quantify the time effect on biomarker performance.

4.3.2 Modeling assumptions

The modeling assumptions required by Cai are very mild. The method is nonparametric with respect to the distribution of Y in controls and semiparametric in regards to the distribution of Y in cases. In particular, they do not specify a distributional form for $\text{Prob}(Y|T)$, but model only the effect of T on this distribution with a parametric form.

The prospective methods are similarly mild in their assumptions. They use a semiparametric model for $\text{Prob}(T|Y)$ but leave the distribution of Y in the cohort unspecified.

4.3.3 Censoring

Censoring that is independent of Y is accommodated by all methods. However, censoring that depends on Y is only accommodated by the S+Z method at this point. Extension of the other two methods to the more general setting of conditionally independent censoring is possible though not completely trivial (Xu and O'Quigley 2000; Cai et al. 2006).

Interestingly the problem of verification biased sampling that is well studied in diagnostic test evaluation (Pepe 2003, Chapt. 7) is entirely analogous to predictor dependent censoring. Verification biased sampling occurs when the result of the diagnostic test is used to select subjects for ascertainment of their true disease status. The resulting bias in naive estimates of (TPF, FPF) is called verification bias. Corrected estimates (Begg and Greenes 1983) are calculated by using naive estimators of positive and negative predictive values, which are unbiased, and putting these together with raw frequencies of positive and negative tests via Bayes theorem. Analogously, when follow-up for the event time outcome depends on the predictor, one can use estimates of prospective parameters and the baseline distribution of predictors to calculate TPF and FPF via Bayes theorem. Viewed in this manner, the S+Z method extends verification bias correction methods to event time data.

4.3.4 Competing risk events

Although not specifically addressed by any of the methods as proposed, they can all be extended to accommodate competing risk events. Hazard functions are replaced with cause specific hazard functions and survivor functions are replaced with the probability of not having an event (of any type, neither events of interest nor competing risk events). One can estimate a separate TPF(t, s) function for competing risk events. In Cai's method a separate model is stipulated. In the prospective approaches, separate cause specific hazard models would be employed. A detailed study of these methods is warranted but is outside the scope of this review paper.

4.3.5 Sampling

The prospective methods were proposed for cohort studies where data on a random sample from the population are obtained. However, they can be generalized. In brief, if the sampling method allows calculation of estimates of the hazard function and of the population distribution of the predictor, the two prospective approaches can be applied. Case-cohort studies and nested case-control designs where controls at T are a random sample from the population at risk at T can therefore be accommodated. An alternative case-control design where controls are a random sample from the population of controls with $T > \tau$, do not give rise to estimates of the hazard function, hence they are not accommodated. In contrast, retrospective methods naturally accommodate the latter case-control study design, assuming censoring does not depend on Y. They also directly accommodate cohort, case-cohort and case-'risk set control' designs under the same censoring assumption.

4.3.6 Longitudinal biomarkers

Cai et al. (2006) developed the retrospective method in the general context where marker data are collected longitudinally over time. An implicit assumption is that marker data at s are missing at random conditional on subsequent event data. The prospective methods can be generalized to accommodate longitudinal data using marginal regression models (Zheng and Heagerty 2007). Specifically, each marker measurement $Y(s)$ generates a data record with time origin for T reset to s. That is the event or censoring time associated with $Y(s)$ is $T - s$ or $X - s$, respectively. By allowing the corresponding baseline hazard and regression coefficients to depend on s, TPF and FPF can be written as functions of s and t.

4.3.7 Covariates

Various factors can affect the marker distribution and/or performance of the marker as a predictor of events. We call these factors covariates. One class is disease specific covariates, i.e., characteristic of the disease. For example, the severity of the AKI event might affect the capacity of a biomarker to predict it. For example, PSA may be a better predictor of one type of prostate cancer than another. Disease specific covariates associated only with cases can be modeled as part of the TPF function with the retrospective analysis approach. Such covariates are not generally accommodated by the prospective methods. However, if the covariate is discrete, events can be classified and treated as competing risks. For example, severe AKI events and mild AKI events could be considered different competing risk event types and TPF estimates could be calculated for the two event types.

Other covariates apply to controls as well as to cases. For example, study site in a multicenter study or characteristics of the subject or tester might influence the marker or its performance. Cai et al. (2006) and Song and Zhou (in press) describe how to incorporate such covariates into the analysis. We refer the reader to those papers for details.

4.3.8 Comparing markers

Retrospective methods can include multiple markers in the context of regression models for TPF and FPF in a fashion similar to that described in chapter 3 of Pepe (2003). The models specify parameters that relate to differences in performance between markers and comparative inference can therefore be made. See Sect. 6.4.3 and 6.4.4 of Pepe (2003) for illustrations including illustration with prostate cancer screening data. Currently prospective methods have no capacity for doing this. Comparing relative risks does not answer the question (Emir et al. 1998).

4.3.9 Combining markers

The methods discussed in this paper are not concerned with how to combine predictors together. However, once a combination is defined, the methods discussed in this paper can be used to evaluate the performance of the combination using an independent test dataset. The Seattle Heart Failure study fits exactly this paradigm. The combination score derived from one cohort is evaluated on an independent cohort in the next section.

5 Data analyses

5.1 Seattle Heart Failure study

For computational ease we extracted a random sample of $n = 1{,}000$ observations from the Val-heft trial. Controls are defined as subjects alive at 2 years after enrollment into the trial. Figure 1 shows the Kaplan–Meier survivor function over (0, 2) years. There were 165 deaths observed and 375 subjects were censored in this time period.

The remaining 460 subjects observed alive at 2 years are known controls. In addition, some unknown proportion of those censored prior to 2 years are controls. Figure 2 displays the marker, the SHF score measured at baseline, in the known controls and for comparison in the cases. Interestingly, earlier deaths do not appear to have higher scores ($p = 0.75$ according to linear regression of SHF score on event time for cases).

Assuming that censoring does not depend on the baseline SHF score, one could estimate crude ROC curves by categorizing cases on the basis of their failure times and comparing their SHF scores with the distribution for known controls. The crude curves in Fig. 3a were calculated as empirical ROC curves for controls versus 3 groups of

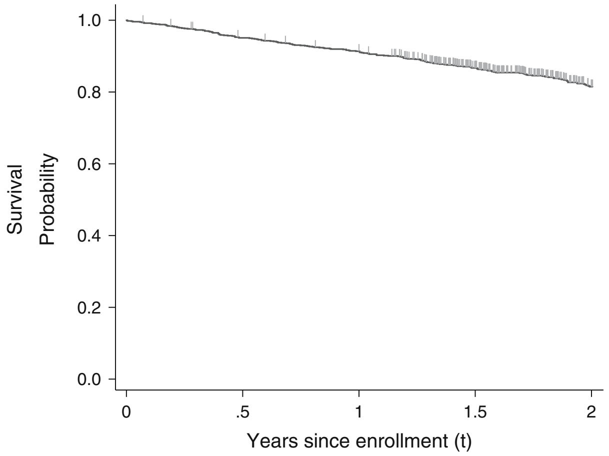

Fig. 1 Kaplan–Meier survival estimates for 1000 subjects enrolled in the Val-heft trial. 460 subjects remain alive and uncensored at $\tau = 2$ years

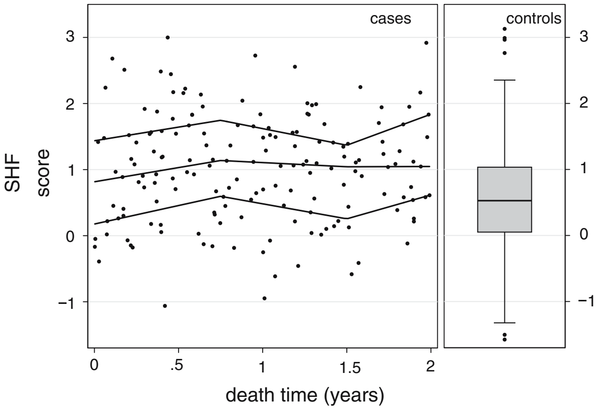

Fig. 2 SHF scores measured at enrollment in cases (left panel) as a function of their death time T and a box-plot of the SHF score distribution in known controls (right panel). The median, 25th and 75th percentile curves displayed in the left panel were modelled as linear splines with knots at 0.75 and 1.5 years and estimated using quantile regression methods (Koenker and Bassett 1978)

cases formed by categorizing T into intervals (0.25,0.75], (0.75,1.25] and (1.25,1.75]. The median death times for the 3 groups of cases were 0.47, 1.01 and 1.46 years, respectively. The corresponding curves approximate ROC curves for subjects who died at 0.5, 1.0 and 1.5 years after baseline. The Heagerty and Zheng curves in Fig. 3c use the same known controls for the FPF axis, but their Cox-model based estimate of TPF(t) on the vertical axis. Cai's method (Fig. 3b) was implemented using a logistic model with the effect of t on logit$\{\text{ROC}_t(f)\}$ modelled as a linear spline with knot at $t = 1$ year, logit$\{\text{ROC}_t(f)\} = h(f) + \beta_1 t + \beta_2(t-1)I[t > 1]$. We estimated $\widehat{\beta_1} = 0.271$ and $\widehat{\beta_2} = -0.233$ without the censored observations. A Wald test for

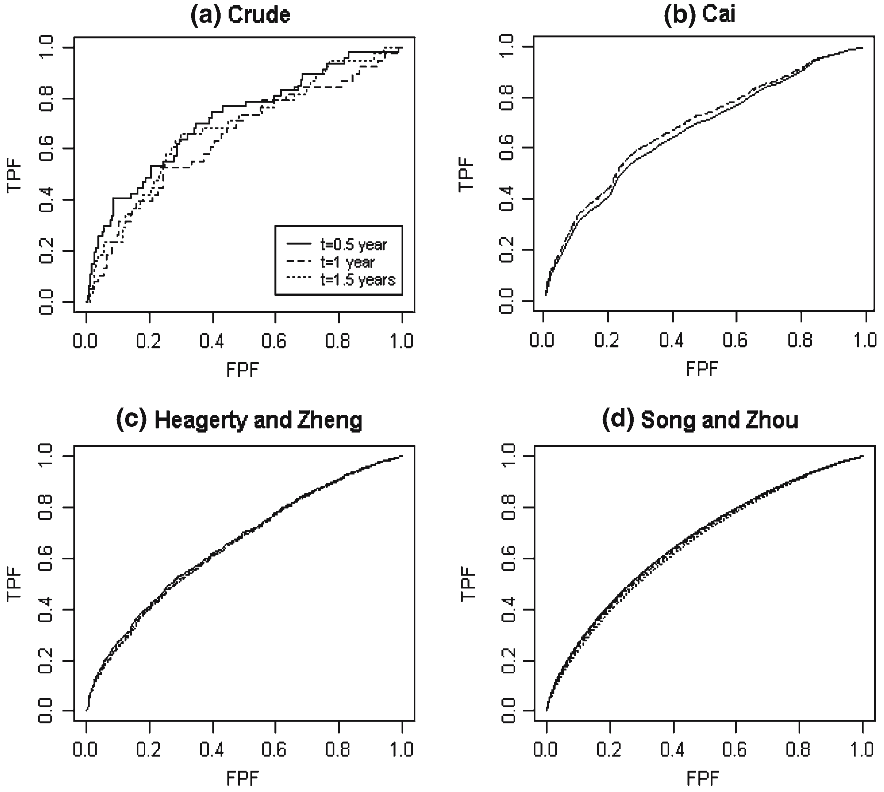

Fig. 3 ROC curves calculated with the Seattle Heart Failure data using 4 methods: (**a**) categorizing T and comparing with known controls only; (**b**) Cai's retrospective method with logit$\{\mathrm{ROC}_t(f)\} \equiv h(f) + \beta_1 t + \beta_2(t-1)I[t>1]$, nonparametric h; (**c**) Heagerty and Zheng with proportional hazards model; and (**d**) Song and Zhou with a proportional hazards model

$H_0 : \beta_1 = \beta_2 = 0$; was not significant (p = 0.79) indicating that ROC curves did not vary with time. The same conclusion was reached after including observations censored by 2 years(p = 0.83). The SHF-score is equally sensitive to events that occur later versus earlier in the 2 year time interval. The Song and Zhou curves (Fig. 3d) also indicate very little variation in the ROC curves with t.

The ROC curve estimates shown in Table 1 are consistent with each other. However there are considerable differences amongst the methods in terms of precision, as quantified by widths of confidence intervals derived from quantiles of their bootstrap distributions. The crude ROC curves have largest variance. Confidence intervals based on the Cai method are narrower. Inclusion of the censored data improves them but not by alot. Amongst the prospective methods, as expected Song and Zhou's method is more efficient than Heagerty and Zheng's. Both prospective methods yield narrower confidence intervals than those calculated with Cai's method. At this point we do not have an explanation. Further work will be needed to determine if this is a general phenomenon.

Table 1 Comparison of estimated ROC curves calculated from The Seattle Heart Failure study data

		Crude	Cai[a]	Cai-cens[b]	H+Z[c]	S+Z[d]
$f = 0.2$	$t = 0.5$	0.489(0.319, 0.633)	0.406(0.297, 0.506)	0.410(0.313, 0.514)	0.410(0.329, 0.487)	0.418(0.335, 0.486)
	$t = 1.0$	0.395(0.256, 0.581)	0.439(0.272, 0.584)	0.443(0.284, 0.587)	0.399(0.322, 0.467)	0.407(0.327, 0.466)
	$t = 1.5$	0.421(0.274, 0.606)	0.444(0.314, 0.546)	0.441(0.341, 0.553)	0.401(0.333, 0.472)	0.392(0.319, 0.449)
$f = 0.5$	$t = 0.5$	0.766(0.654, 0.878)	0.698(0.585, 0.787)	0.707(0.615, 0.795)	0.702(0.626, 0.753)	0.723(0.652, 0.771)
	$t = 1.0$	0.737(0.538, 0.864)	0.726(0.549, 0.820)	0.734(0.586, 0.836)	0.692(0.619, 0.743)	0.715(0.646, 0.760)
	$t = 1.5$	0.737(0.593, 0.870)	0.730(0.630, 0.796)	0.732(0.653, 0.813)	0.693(0.624, 0.737)	0.705(0.639, 0.748)
$f = 0.8$	$t = 0.5$	0.936(0.860, 1.000)	0.897(0.826, 0.950)	0.901(0.836, 0.940)	0.911(0.876, 0.937)	0.918(0.885, 0.937)
	$t = 1.0$	0.842(0.710, 0.957)	0.909(0.827, 0.964)	0.912(0.830, 0.956)	0.907(0.873, 0.934)	0.915(0.883, 0.934)
	$t = 1.5$	0.947(0.857, 1.000)	0.910(0.860, 0.958)	0.912(0.863, 0.946)	0.908(0.871, 0.934)	0.911(0.879, 0.930)

[a] Cai’s method using only known controls for the FPF

[b] Cai’s method including censored observations in (0,2) years

[c] Heagerty and Zheng’s method

[d] Song and Zhou’s method

95% confidence intervals in parentheses are based on the same 200 bootstrapped samples

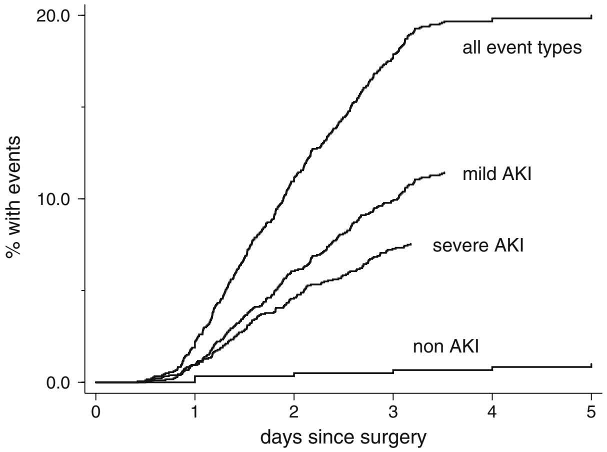

Fig. 4 Cumulation distributions of event times in the Kidney Biomarker study

5.2 Kidney Biomarkers study

Recall the objective of the Kidney Biomarkers study is to evaluate biomarkers for early diagnosis of acute kidney injury (AKI) in subjects undergoing major cardiac surgery. The outcome, AKI, is defined in terms of elevation in serum creatinine levels, and AKI may be mild or severe in nature. We seek biomarkers that can detect kidney injury in advance of the serum creatinine AKI response. This study is in progress and data will not be available for some time. We have simulated data that approximates the study design, as described in Appendix B. These data are available on the DABS website (www.fhcrc.org/science/labs/pepe/dabs/).

Of the 1,800 subjects in the study, 342 had AKI events, 136 severe AKI and 206 mild AKI. In addition, 18 patients died from causes seemingly unrelated to kidney damage. Figure 4 shows the distributions of event times. Consider the biomarker measured from the first postoperative urine sample that we call the baseline biomarker. Its distribution is displayed in Fig. 5 for the 4 patient groups. We see that compared with controls the AKI groups have generally higher baseline biomarker values, with the severe AKI group being more removed from controls than are the mild AKI values. The baseline biomarker in patients who die from non-AKI events does not appear to differ from controls. Formal comparisons between the groups based on the Wilcoxon rank sum statistic yield $p-$values: $p < 0.001$ for mild AKI versus controls; $p < 0.001$ for severe AKI versus controls; $p = 0.038$ for severe AKI versus mild AKI; and $p = 0.23$ for non-AKI deaths versus controls. Note that the Wilcoxon rank sum statistic is a simple function of the nonparametric area under the ROC curve that compares two groups (Pepe 2003, p. 103).

Consider now the capacity of the baseline biomarker to diagnose AKI events before their clinical diagnosis with serum creatinine. In the future treatment may be initiated on the basis of biomarkers. Clinicians will need to balance the potential benefits of treatment for subjects who would have an AKI in the absence of treatment against the false positives, those who would not have an AKI but are flagged as positive by

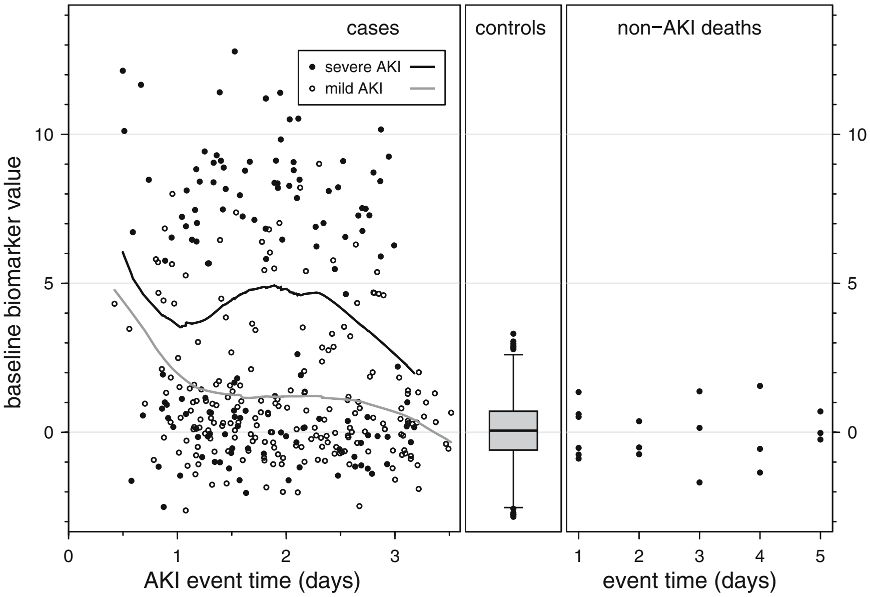

Fig. 5 Baseline AKI biomarker distributions. Lowess curves for biomarkers in severe and mild AKI subgroups are shown

the biomarker. For potentially toxic or expensive treatments the false positive rate should be low. For non-toxic treatments, such as additional hydration, a large false positive rate would be acceptable. A range of treatment options will be considered in the future, so a range of false positive rates are of interest here. The crude ROC curves for the baseline biomarker in Fig. 6 were calculated by categorizing the event time axis as early=(.25,1.5] and medium = (1.5,3]. ROC curves comparing baseline biomarker values in controls with those of subjects with severe AKI events in each of the time intervals are shown in Fig. 6 left panel while corresponding curves for subjects with mild AKI events are in the right panel.

We implemented Cai's method for the baseline marker with the following time-dependent ROC curve model

$$\text{logit}\{\text{ROC}_t(f)\} = h_0(f) + \beta_1 t + \beta_2(t - 1.5)I[t > 1.5].$$

That is, we used a logistic link function, nonparametric baseline ROC curve and modelled event time effects as a linear spline with one knot at $t = 1.5$. Separate models were fit for mild and severe AKI events, although we note that a model including both could have been fit by including interactions with 'event type' in the above ROC-GLM formulation.

Song and Zhou's method was also applied. We included only subjects with severe events and controls in estimating ROC curves corresponding to severe AKI versus controls and only mild AKI versus controls in the second set of analyses. Follow up was technically terminated at 5 days which is the end of the observation time. Separate models and analyses were used for mild and severe cases. Figure 6 displays estimated

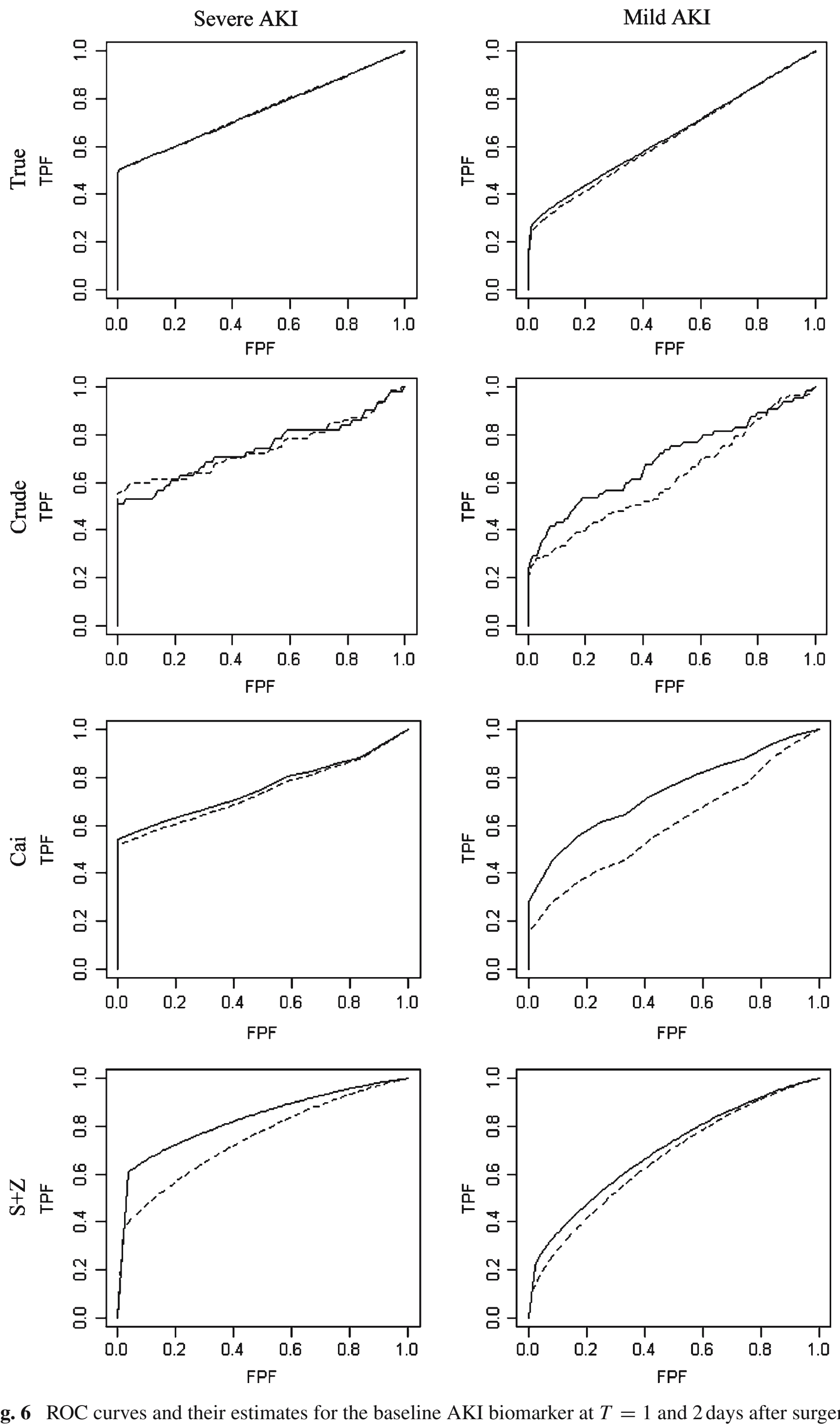

Fig. 6 ROC curves and their estimates for the baseline AKI biomarker at $T = 1$ and 2 days after surgery

Table 2 Comparison of estimated ROC curves for the baseline biomarker of acute kidney injury

		Severe AKI				Mild AKI			
		True	Crude	Cai	S+Z[a]	True	Crude	Cai	S+Z[a]
$f = 0.05$	$T = 1$	0.525	0.529	0.565	0.619	0.315	0.354	0.385	0.278
	$T = 2$	0.524	0.600	0.541	0.414	0.289	0.283	0.226	0.199
$f = 0.20$	$T = 1$	0.599	0.608	0.631	0.722	0.436	0.538	0.577	0.475
	$T = 2$	0.599	0.613	0.608	0.568	0.413	0.398	0.384	0.414
$f = 0.50$	$T = 1$	0.751	0.745	0.755	0.861	0.644	0.754	0.770	0.741
	$T = 2$	0.752	0.725	0.736	0.783	0.634	0.584	0.605	0.710
$f = 0.80$	$T = 1$	0.897	0.843	0.875	0.958	0.858	0.892	0.917	0.923
	$T = 2$	0.901	0.863	0.864	0.934	0.857	0.867	0.837	0.914

[a] Song and Zhou's method

Here T is the time after surgery that AKI was diagnosed

ROC curves at $T = 1$ and 2 days. Since the data are simulated, we were also able to calculate the true time-dependent ROC curves by simulating a very large data set, and selecting cases of each severity with events in the interval $[T - .01, T + .01]$ and controls, and calculating the empirical ROC curves. Table 2 displays results.

We see for example that allowing a 20% false positive rate the baseline marker detects 59.9% of subjects who develop severe AKI 2 days after surgery and 41.3% of those who develop mild AKI at 2 days after surgery. It detects a slightly higher fraction, 43.6%, of those that develop mild AKI at one day after surgery. The true ROC curves rise steeply on the left and turn sharply linear at approximately TPF = 0.5 for severe AKI and at TPF = 0.25 for mild AKI. The nonparametric nature of the baseline ROC curve allows the curves calculated with Cai's method to follow this shape. Moreover, the curves estimated with Cai's method are similar to the crude nonparametric curves, i.e., they follow the raw data rather well. On the other hand, the Song and Zhou estimates are not close to the crude ROC curves. Presumably this is because the proportional hazards assumption does not hold. The results suggest that the Song and Zhou approach should be generalized to allow non proportional hazards models.

Turning now to the longitudinal biomarker data, we first explored if in controls the biomarker distribution varied with s, time from surgery. It appears to be stable over time in accordance with our simulation model. Figure 7 is similar to the display of biomarker distributions in Fig. 5 except that all biomarker measurements are displayed, and marginalized over time for the controls. The time axis t for cases is time from marker measurement to AKI event. Each case has multiple observations, (Y, t), corresponding to the various measurements prior to his event. Note that the time axis here differs from that used for the baseline marker in the earlier analysis where $t = T$. Here, the analysis acknowledges that the baseline marker is measured at some time in the interval (0,0.25), not at 0. Therefore t, time from measurement of the baseline marker to event, is not the same as the event time, T, even for the baseline measurement.

ROC curves were fit using these longitudinal data with the same methods as described earlier. Results are shown in Fig. 8 and Table 3. We conclude that with the new urine biomarker when allowing a 20% FPF, 94.2% of subjects with severe AKI events

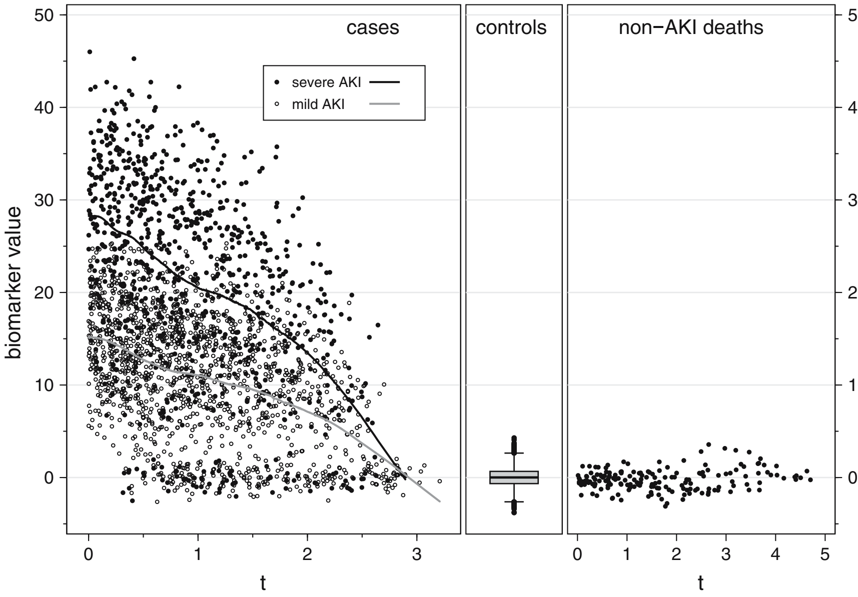

Fig. 7 Biomarker distributions in cases as a function of the time lag between marker measurement and event time, $t = T - s$, and in controls

Table 3 Comparison of estimated ROC curves for the biomarker of acute kidney injury

		Severe AKI				Mild AKI			
		True	Crude	Cai	S+Z[a]	True	Crude	Cai	S+Z[a]
$f = 0.05$	$t = 1$	0.932	0.923	0.933	0.645	0.874	0.887	0.880	0.564
	$t = 2$	0.854	0.839	0.804	0.325	0.743	0.696	0.665	0.323
$f = 0.20$	$t = 1$	0.942	0.933	0.942	0.717	0.897	0.919	0.917	0.663
	$t = 2$	0.876	0.850	0.828	0.461	0.789	0.753	0.748	0.475
$f = 0.50$	$t = 1$	0.965	0.957	0.962	0.839	0.936	0.950	0.953	0.817
	$t = 2$	0.923	0.894	0.882	0.693	0.868	0.853	0.846	0.715
$f = 0.8$	$t = 1$	0.986	0.980	0.983	0.943	0.975	0.979	0.983	0.939
	$t = 2$	0.968	0.950	0.944	0.891	0.947	0.948	0.940	0.905

[a] Song and Zhou's method

Here t is the time interval in days prior to clincial diagnosis of kidney injury that the biomarker was measured. Longitudinal biomarker data are included

can be detected 1 day prior to their clinical diagnosis, and 87.6% can be detected 2 days prior. The corresponding numbers for subjects with mild AKI are 89.7% and 78.9%. Contrast these with the much smaller proportions that could be detected using only the baseline biomarker. In regards to estimating the time-dependent ROC curves, the Song and Zhou method appears to underestimate. The underestimation is particularly problematic at smaller FPFs. Presumably the proportional hazards assumption again fails. Cai's method does a much better job of estimation here. It is close to the nonparametric 'crude' curves, but does not require choosing time intervals about t to estimate the ROC curve at t.

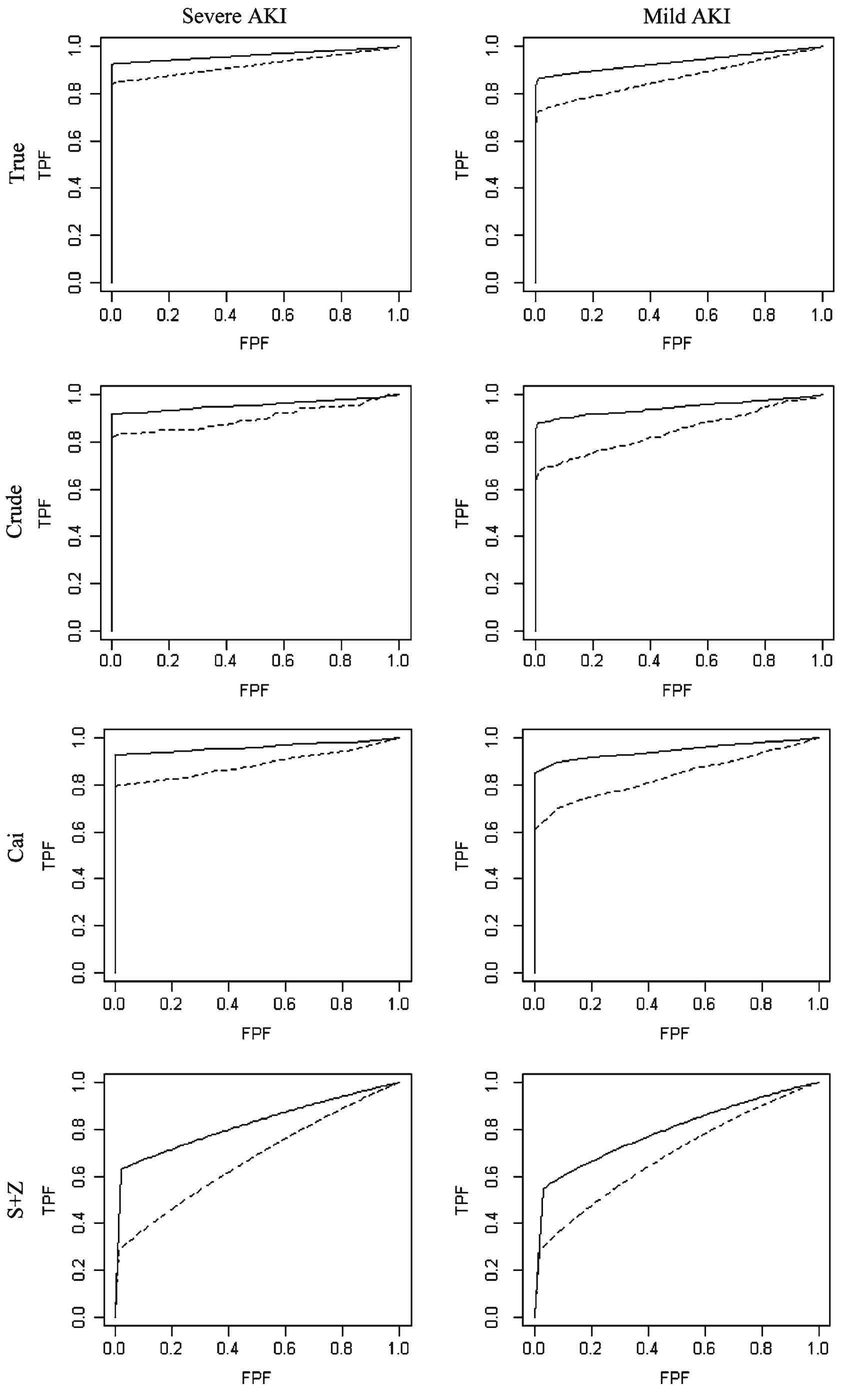

Fig. 8 ROC curves for the longitudinally measured AKI biomarker measured at 1 and 2 days prior to clinical diagnosis of AKI with serum creatinine

6 Discussion

Sam Wieand made many contributions to the fields of biostatistics and oncology. One of his legacies is the promotion of sound approaches to evaluating predictors for diagnostic and prognostic purposes. He recognized that relative risks alone are inadequate and he promoted the use of ROC curves instead. Since his landmark 1989 paper with Gail, James and James (Wieand et al. 1989), ROC analysis methodology has progressed considerably. Yet ROC analysis methods are not well developed for the analysis of censored failure time data, another topic of great interest to Sam Wieand. Our paper is an effort to summarize the current state of this field. These methods should be used in practice and some directions for further work are apparent.

The focus of this paper has been on estimating time dependent ROC curves. Methods for estimating summary indices such as the area under the time-dependent ROC curve (AUC), were not discussed, although they have been developed (Antolini et al. 2005; Chambless and Diao 2006). Although the AUC is a popular summary index, it has been widely criticized as clinically irrelevant (Cook 2007; Baker 2003). Sam Wieand himself suggested using instead the partial AUC to summarize predictor performance over a restricted range of false positive (or true positive) fractions (Wieand et al. 1989). It would be interesting and useful to develop methodology for inference about time-dependent partial AUC as a summary of the performance of a marker for predicting event time outcomes.

Appendix A: Implementation of Cai's procedure

Use Y_i, T_i, X_i, δ_i to denote the marker measurement, event time, observation time, and censoring indicator for the ith subject.

(I) First suppose that the marker Y_i is binary, measured only at baseline time 0. Each subject is classified as a case with event time $T_i \leq X_i, T_i \leq \tau$, as a control with observation time $X_i > \tau$, or as a censored observation of unknown outcome status with $T_i > X_i, X_i \leq \tau$. Writing the TPF and FPF models as

$$\begin{aligned} \text{TPF}(t) &= g(\alpha + \beta\eta(t)) \\ \text{FPF} &= f \end{aligned}$$

with $\psi = (\alpha, \beta, f)$, a two-step approach can be used for estimating ψ.

(1) We calculate $\tilde{\alpha}$ and $\tilde{\beta}$, the initial estimates of α and β, by fitting a binary generalized linear model (GLM) to the data for cases with outcome variable Y and covariate $\eta(T)$. The initial estimate of f, $\tilde{f}$, is calculated as the proportion of controls with $Y = 1$.

(2) Next, we include censored observations. Let t_j, $j = 1, \ldots, J$ be the set of unique event time in the dataset and define $t_0 = 0$. Let $p_i(\psi)$ be the probability that $Y_i = 1$ conditional on the group label of subject i (i.e., case, control, or censored with unknown status), the likelihood of the data is

$$\prod_{i=1}^{n}\{p_i(\psi)\}^{Y_i}\{1-p_i(\psi)\}^{1-Y_i}.$$

We estimate ψ by maximizing the approximated score equation,

$$\sum_{i=1}^{n}\frac{\frac{\partial\hat{p}_i(\psi)}{\partial\psi}}{\hat{p}_i(\psi)\{1-\hat{p}_i(\psi)\}}\{Y_i-\hat{p}_i(\psi)\}=0, \tag{3}$$

where $\hat{p}_i(\psi)$ is

$$\begin{array}{rl} \text{TPF}(X_i) & \text{if } X_i\le\tau, \delta_i=1,\\ \dfrac{-\sum_{t_j\in(X_i,\tau)}\text{TPF}(t_j)\Delta\hat{S}(t_j)+\text{FPF}\hat{S}(\tau)}{\hat{S}(X_i)} & \text{if } X_i\le\tau, \delta_i=0,\\ \text{FPF} & \text{if } X_i>\tau, \end{array}$$

where $\hat{S}(t)=\hat{P}(T>t)$ is the Kaplan–Meier estimate of $P(T>t)$ and $\Delta\hat{S}(t_j)=\hat{S}(t_{j-1})-\hat{S}(t_j)$.

Equation 3 can be solved using a Newton–Raphson algorithm with the weight function

$$\frac{\frac{\partial\hat{p}_i(\psi)}{\partial\psi}}{\hat{p}_i(\psi)\{1-\hat{p}_i(\psi)\}}$$

evaluated at $\tilde{\psi}$.

(II) To analyze data in which biomarkers are measured at multiple time points, each subject has a data record for each biomarker measurement time: $s_{ik}, Y_i(s_{ik}), X_i, \delta_i, k=1,\ldots,K_i$. The time origin for this record is reset to 0 at s_{ik} since we are concerned with t = time until event after biomarker measurement. Thus replace X_i with X_i-s_{ik} and the analysis proceeds as before based on this augmented dataset. Note that the measurement time s_{ik} may be included as a covariate. That is, if the distribution of the biomarker can vary with s, the models for FPF and TPF(t) may be extended to include s as a covariate.

(III) Finally, suppose that the biomarker Y is continuous. This is accommodated by replacing each record in the dataset with P records, each corresponding to a different cutpoint c_p, $p=1,\ldots,P$, and replacing the continuous marker $Y_i(s_{ik})$ with the dichotomous version $I[Y_i(s_{ik})>c_p]$. The FPF and TPF models include factor variables for the cutpoint such that for a record with cutpoint c_p, the corresponding TPF(t) and FPF are

$$\begin{aligned}\text{TPF}_{c_p}(t)&=g(\alpha(c_p)+\beta\eta(t))\\ \text{FPF}_{c_p}&=f(c_p).\end{aligned}$$

Ideally the cutpoints used are estimated quantiles of the biomarker in controls (possibly depending on s through regression quantile techniques). This implies that they represent points corresponding to specific FPF points on the x-axis of the ROC curve.

Let $\psi = (\alpha(c_1), \ldots, \alpha(c_P), \beta, f(c_1), \ldots, f(c_P))$, we estimate ψ by maximizing the approximated score equation analogous to Eq. 3,

$$\sum_{i=1}^{n}\sum_{k=1}^{K_i} \frac{\frac{\partial \hat{p}_{ik}(c_p,\psi)}{\partial \psi}}{\hat{p}_{ik}(c_p,\psi)\{1-\hat{p}_{ik}(c_p,\psi)\}}\{I[Y_i(s_{ik}) > c_p] - \hat{p}_{ik}(c_p,\psi)\} = 0,$$

where $\hat{p}_{ik}(c_p, \psi)$ is

$$\begin{array}{rl} \mathrm{TPF}_{c_p}(X_i - s_{ik}), & \text{if } X_i - s_{ik} \le \tau, \delta_i = 1, \\ \dfrac{-\sum_{t_j \in (X_i - s_{ik}, \tau)} \mathrm{TPF}_{c_p}(t_j)\Delta\hat{S}(t_j) + \mathrm{FPF}_{c_p}\hat{S}(\tau)}{\hat{S}(X_i - s_{ik})} & \text{if } X_i - s_{ik} \le \tau, \delta_i = 0, \\ \mathrm{FPF}_{c_p} & \text{if } X_i - s_{ik} > \tau. \end{array}$$

Appendix B: Generation of simulated Kidney Biomarker data

(i) Notation

$$\begin{aligned} n &= \text{ total sample size } = 1,800 \\ i &= \text{ subject index} \\ k &= k\text{th specimen sample} \\ s_{ik} &= \text{ time of the } k\text{th specimen for the } i\text{th subject} \end{aligned}$$

(ii) Sampling times (s_{ik})

Patients should have a urine sample taken approximately every 6 h for 5 days after surgery. The timing is often delayed. Generate

$$s_{ik} = 0.25k + \varepsilon_{ik}$$

$k = 1, \ldots, 20$ and $\varepsilon \sim$ uniform (0,0.25). These *potential* sampling times are modified (below) depending on patient status.

(iii) Patient subgroups

Controls

A random set of 1,440 patients were assigned control status. We simulated their being discharged on day 3 (30%), day 4 (40%) and day 5 (30%) by dropping measurement times s_{ik} exceeding days 3 and 4 for random subsets of 30% and 40% of controls, respectively.

Non-AKI deaths

18 patients were assigned non-AKI death status. Measurement times after day 1 were dropped for 6 of the patients simulating that the event occurred on day 1. Similarly by dropping measurement times after days 2, 3, and 4 for sets of 3 patients each, we simulated events at day 2 for 3 patients, at day 3 for 3 patients, at day 4 for 3 patients and at day 5 for 3 patients.

AKI events (T)

All remaining 342 patients had an AKI event. Of these, we assigned 206 severity status mild and 135 severe. An unobserved latent event time E was generated as follows for patients with severe AKI:

$$E^{sev} \sim \text{ uniform } (0, 0.25) \text{ with probability } 0.6$$
$$E^{sev} \sim \text{ uniform } (0.25, 1.25) \text{ with probability } 0.4$$

That is, E^{sev}, the true latent time of AKI, was uniformly distributed between 0 and 0.25 in 60% of severe patients and uniformly distributed between 0.25 and 1.25 in 40%. Corresponding true time of AKI in patients who had mild AKI was such that

$$E^{\text{mild}} \sim \text{ uniform } (0, 0.25) \text{ with probability } 0.4$$
$$E^{\text{mild}} \sim \text{ uniform } (0.25, 2.25) \text{ with probability } 0.6$$

The time, T, of *clinical* diagnosis of AKI with serum creatinine was generated as

$$T = E + V \text{ where } V \sim \text{ uniform } (0, 2.75)$$

(iv) Biomarker values

Controls and non-AKI deaths

Biomarker values are normally distributed with mean 0 and variance 1 with no trend over time. An auto-regressive structure was simulated:

$$Y_{i,1} \sim N(0, 1)$$
$$Y_{i,k} = \alpha Y_{i,k-1} + \sqrt{1 - \alpha^2}\ \ \varepsilon_{ik}, \quad k \geq 2$$

where ε_{ik} is independent $N(0, 1)$ error and the autoregressive correlation is determined by α. We chose $\alpha = 0.8$.

Cases

In cases, biomarker values are generated as for controls up to the time of their (unobserved) event time E. Let s_{ik^*} be the time of the first measurement after E. We generated

$$Y_{i,k^*} = \Delta + \alpha Y_{i,k^*-1} + \sqrt{1-\alpha^2}\ \ \varepsilon_{ik^*}$$

where $\Delta = \mu + \delta$, δ independent normally distributed with mean 0 and standards deviation 2 and μ, the mean of Y_{ik^*} depends on severity of AKI.

$$\mu = 8 \text{ for subjects with severe AKI}$$
$$\mu = \ 4 \text{ for subjects with mild AKI.}$$

For later measurement times,

$$Y_{i,k} = \alpha Y_{i,k-1} + \sqrt{1-\alpha^2}\ \ \varepsilon_{ik}, \quad k > k^*.$$

References

Antolini L, Boracchi P, Biganzoli E (2005) A time dependent discrimination index for survival data. Stat Med 24:3927–3944

Baker SG (2003) The central role of receiver operating characteristic (ROC) curves in evaluating tests for the early detection of cancer. J Natl Cancer Inst 95:511–515

Begg CB, Greenes RA (1983) Assessment of diagnostic tests when disease verification is subject to selection bias. Biometrics 39:207–215

Cai T, Pepe MS, Zheng Y, Lumley T, Jenny NS (2006) The sensitivity and specificity of markers for event times. Biostatistics 7:182–197

Cai T, Pepe MS (2002) Semi-parametric ROC analysis to evaluate biomarkers for disease. J Am Stat Assoc 97:1099–1107

Chambless LE, Diao G (2006) Estimation of time-dependent area under the ROC curve for long-term risk prediction. Stat Med 25:3474–3486

Cohn JN, Tognoni G (2001) A randomized trial of the angiotensin-receptor blocker valsartan in chronic heart failure. N Engl J Med 345:1667–1675

Cook NR (2007) Use and misuse of the receiver operating characteristic curve in risk prediction. Circulation 115:928–935

Delong ER, Vernon WB, Bollinger RR (1985) Sensitivity and specificity of a monitoring test. Biometrics 41:947–958

Emir B, Wieand S, Su JQ, Cha S (1998) Analysis of repeated markers used to predict progression of cancer. Stat Med 17:2563–2578

Etzioni R, Pepe M, Longton G, Hu C, Goodman G (1999) Incorporating the time dimension in receiver operating characteristic curves: a case study of prostate cancer. Med Decis Making 19:242–251

Heagerty PJ, Zheng Y (2005) Survival model predictive accuracy and ROC curves. Biometrics 61:92–105

Heagerty PJ, Lumley T, Pepe MS (2000) Time-dependent ROC curves for censored survival data and a diagnostic marker. Biometrics 56:337–344

Kalbfleisch JD, Prentice RL (1980) The statistical analysis of failure time data. Wiley, New York

Koenker R, Bassett G (1978) Regression quantiles. Econometrica 46:33–50

Leisenring W, Pepe MS, Longton G (1997) A marginal regression modelling framework for evaluating medical diagnostic tests. Stat Med 16:1263–1281

Levy WC, Mozaffarian D, Linker DT, Sutradhar SC, Anker SD, Cropp AB, Anand I, Maggioni A, Burton P, Sullivan MD, Pitt B, Poole-Wilson PA, Mann DL, Packer M (2006) The Seattle Heart Failure model: prediction of survival in heart failure. Circulation 113:1424–1433

McIntosh M, Pepe MS (2002) Combining several screening tests: optimality of the risk score. Biometrics 58:657–664

Packer M, O'Connor CM, Ghali JK, Pressler ML, Carson PE, Belkin RN, Miller AB, Neuberg GW, Frid D, Wertheimer JH, Cropp AB, DeMets DL, for the Prospective Randomized Amlodipine Survival Evaluation Study Group (1996) Effect of amlodipine on morbidity and mortality in severe chronic heart failure. New Eng J Med 335:1107–1114

Parker CB, Delong ER (2003) ROC methodology within a monitoring framework. Stat Med 22:3473–3488

Pepe MS (2003) The statistical evaluation of medical tests for classification and prediction. Oxford University Press, New York

Song X, Zhou XH (in press) A semiparametric approach for the covariate specific ROC curve with survival outcome. Stat Sinca.

Wang TJ, Gona P, Larson MG, Tofler GH, Levy D, Newton-Cheh C, Jacques PF, Rifai N, Selhub J, Robins SJ, Benjamin EJ, D'Agostino RB, Vasan RS (2006) Multiple biomarkers for the prediction of first major cardiovascular events and death. N Engl J Med 355:2631–2639

Wieand S, Gail MH, James BR, James KL (1989) A family of nonparametric statistics for comparing diagnostic markers with paired or unpaired data. Biometrika 76:585–592

Xu R, O'Quigley J (2000) Proportional hazards estimate of the conditional survival function. J Roy Stat Soc Ser B 62:667–680

Zheng Y, Heagerty PJ (2004) Semiparametric estimation of time-dependent ROC curves for longitudinal marker data. Biostatistics 5:615–632

Zheng Y, Heagerty PJ (2007) Prospective accuracy for longitudinal markers. Biometrics 63:332–341

Zheng Y, Cai T, Feng Z (2006) Application of the time-dependent ROC curves for prognostic accuracy with multiple biomarkers. Biometrics 62:279–287