

Dynamic Thresholds and a Summary ROC Curve: Assessing the Prognostic Accuracy of Longitudinal Markers

P. Saha-Chaudhuri and P. J. Heagerty

October 2, 2012

Abstract

Cancer patients, chronic kidney disease (CDK) patients, and subjects infected with HIV are commonly monitored over time using biomarkers that represent key health status indicators. Furthermore, biomarkers are frequently used to guide initiation of new treatments or to inform changes in intervention strategies. Since key medical decisions can be made on the basis of a longitudinal biomarker it is important to evaluate the potential accuracy associated with longitudinal monitoring. We introduce a summary receiver operating characteristic (ROC) curve that displays the overall sensitivity associated with a time-dependent threshold that controls specificity. The proposed statistical methods are similar to concepts considered in disease screening, yet our methods are novel in choosing a potentially time-dependent threshold to define a positive test, and our methods allow time-specific control of the false-positive rate. Finally, the proposed summary ROC curve is a natural averaging of time-dependent incident/dynamic ROC curves proposed by Heagerty and Zheng (2005) and therefore provides a single summary of net error rates that can be achieved in the longitudinal setting.

1 Introduction

Clinical tests and/or biomarkers are commonly used to guide medical decisions. Diagnostic markers provide evidence regarding a current disease state such as the presence of cancer, while prognostic markers imply a high risk for future transition to a poorer health status such as disease progression or death. In many practical settings, patients are monitored sequentially and clinical tests or markers are repeatedly evaluated on the same subject over a period of time (days, months) with the goal of updating prognosis and correctly classifying subjects with different predicted risk profiles. In therapeutic settings, biomarkers may be monitored over time for prompt identification of an adverse response to the treatment so that the affected patients can have prompt changes to their treatment plan. Serial marker measurements are thought to add “value” over a single baseline measurement and it is frequently of interest to see if the serial marker measurements add to the prognostic ability of an existing risk model. For example, in a recent study of heart failure patients, cardiologists were interested in examining the prognostic utility of *longitudinal* TnI measurements (Xue et al., 2011). Risk stratification for various diseases, such as cardiovascular diseases and cancers, remains suboptimal even after the introduction of various risk scores. Therefore, the need for gathering “more data, preferably from serial measurements in large populations” has become apparent (Koenig, 2010).

Biomedical investigators continue to develop new candidate longitudinal markers, but statistical methods are needed to summarize the potential of such markers for improving clinical decision making. The literature on estimation of risk associated with a time-dependent marker for a survival outcome is extremely rich and continues to grow, but methods for assessing the prognostic accuracy of a longitudinal marker are relatively new and remain incomplete. The use of a time-dependent marker for prediction is inherently complex due to the potential endogeneity of the marker process (Jewell and Kalbfleisch, 1996, Jewell and Nielsen, 1993, Mammen and Nielsen, 2007). An excellent review of the issues are discussed by Fisher and Lin (1999). However, recently there has been work that proposes joint model-

ing of the longitudinal marker process and the survival time with the objective of estimating prediction accuracy parametrically based on these joint models (Rizopoulos, 2011).

In this manuscript, we show how the existing methods of time-dependent predictive accuracy assessment of Heagerty and Zheng (2005) can be suitably adapted to evaluate *overall* sensitivity and specificity or “the true positive and true negative rates associated with an entire body of testing” (Thompson, 2003). In the literature, the overall accuracy of a continuous marker for a survival outcome has been dealt with in either of two ways: use of a fixed disease status over time or use of a fixed marker threshold over time. The existing approaches that treat the disease status over time as static (Murtaugh, 1995), may not be suitable in many clinical applications. In a prospective longitudinal setting, the disease status is subject to change and the analytical methods should accommodate such dynamic nature of the disease. Similarly, as a marker process continues to evolve over time among both the diseased and non-diseased subgroups (Ito et al., 2003, Zheng and Heagerty, 2004, Deslandes and Chevret, 2007, Ye et al., 2008, Saha and Heagerty, 2010), use of a fixed threshold over time (DeLong et al., 1985, Parker and DeLong, 2003) may not be suitable (Thompson, 2003). When an optimal marker threshold is not a priori defined, the accuracy of all possible thresholds may be of interest. Furthermore, the trajectory of time-dependent AUC (Heagerty and Zheng, 2005, Saha and Heagerty, 2010, Saha-Chaudhuri and Heagerty, 2012) may not provide enough relevant information since variation in thresholds controlling the false positive rate is not shown. Our proposal is similar in spirit to the marker comparison approach of Emir et al. (1998) and Emir et al. (2000) in that we propose use of a time-dependent marker threshold that corresponds to a fixed specificity over time to define test-positive at each outcome time and we reweight the sensitivity over time to achieve overall accuracy measures. Unlike Zheng and Heagerty (2004), we thus provide an overall measure of accuracy for a longitudinal marker. We explicitly account for censoring and use the distribution of survival times as natural weights. The resulting estimator of net (or average) sensitivity can then be interpreted as an overall measure of marker sensitivity. This way, we are able to provide an estimator of overall ROC curve that corresponds to all levels of longitudinally controlled specificity and also meaning-

fully incorporate the information contained in the distribution of survival times in order to summarize performance over the full follow-up time period.

The rest of the article is organized as follows. We introduce the notation and the concept of a summary survival ROC curve and then discuss estimation and inference in section 2. We study the proposed approach via simulation in section 3. We apply the summary survival ROC curve to a well-known dataset from the Multicenter AIDS Cohort Study in section 4. We conclude with a discussion (section 5).

2 Estimation

2.1 Notation

Let n denote the total number of subjects in the cohort under study and at risk at $t = 0$. Let T_i denote the event time and C_i denote the (independent) censoring time for subject i . We assume that subjects are independent. For each subject, we observe the follow-up time $Z_i = \min(T_i, C_i)$ and the censoring indicator $\delta_i = \mathbb{1}(T_i \leq C_i)$. Let $R_i(t) = \mathbb{1}(Z_i \geq t)$ denote the at-risk indicator for subject i and $\mathcal{R}(t) = \{i : Z_i \geq t\}$ denotes the riskset at t . Let m_t denote the number of subjects who experienced an event at t , while n_t denotes the number of subjects surviving beyond time t . Further, a baseline marker is denoted by M and a time-dependent marker process is denoted by $\{M(t) : t > 0\}$. We assume that the markers may be a single covariate or a model score generated through any regression model like the Cox model. We assume that higher marker values are more indicative of disease and therefore a shorter survival time. We focus on *incident* case and *dynamic* control definition of Saha and Heagerty (2010) and define the measures of time-dependent predictive accuracy as follows:

$$\text{TP}^{\mathbb{I}}(c(t), t) = P(M(t) > c(t) \mid T = t)$$

$$\text{FP}^{\mathbb{D}}(c(t), t) = P(M(t) > c(t) \mid T \geq t).$$

2.2 Summary survival ROC curve for a marker

Consider a dynamic false positive (FP) criterion:

$$c^p(t) : p = P[M(t) > c^p(t) | T > t].$$

When such a dynamic FP criterion is used repeatedly over time for a time-dependent marker or a model score, we may be interested in the overall predictive accuracy of the score. An important question, therefore, is: if this dynamic criteria is used repeatedly, how many cases will test positive at the appropriate times? Or, conversely, how many cases will be missed? We can answer this question by looking at the overall fraction of cases that test positive when they are about to fail. At time t , a subject can experience the event of interest with probability $P(T = t)$. If we use the test-positive threshold $c^p(t)$ to produce a false positive fraction equal to p , the $TP[c^p(t), t]$ proportion of the cases who fail at t will test positive. Hence, $TP[c^p(t), t] \times P(T = t)$ proportion of the subjects will correctly test positive at t . To get the overall proportion of cases that test positive when they are about to fail, we simply sum over (integrate) the failure times and label the proportion as Total True Positive (TTP). Notationally, we define TTP at $FP = p$ as:

$$\begin{aligned} \text{TTP}(p) &= \int_t \text{TP}[c^p(t), t] P(T = t) dt \\ &= \int_t P[M(t) > c^p(t) | T = t] P(T = t) dt. \end{aligned}$$

An ROC curve can be obtained by plotting (FP, TTP) pairs for different FP thresholds $p \in (0, 1)$ and is labeled as Summary Survival ROC curve.

In practice we would typically restrict attention to a fixed follow-up period $(0, \tau)$. The TTP can be modified to account for finite follow-up:

$$\text{TTP}^\tau(p) = \int_0^\tau \text{TP}[c^p(t), t] w^\tau(t) dt$$

where $w^\tau(t) = P(T = t | T \leq \tau)$. This restricted TTP summarizes the predictive accuracy of the marker over a finite interval of time $(0, \tau)$.

Note that the constituents of a summary survival ROC curve that involve both time and marker distribution are: the test-positive threshold $c^p(t)$ for a given FP fraction p and time t ; and the TP corresponding to this threshold. The marginal $P(T = t)$ simply involves the failure time distribution and can be estimated non-parametrically, for example, via a nearest neighbor approach (Akritas, 1994).

At a given time t , non-parametric estimation of FP is straight-forward using the empirical distribution function of the marker among the subjects that remain event-free at time t (Heagerty and Zheng, 2005), and $c^p(t)$ can be estimated by the empirical marker quantiles among the controls at time t :

$$\begin{aligned} \widehat{P}[M(t) > c | T > t] &= \frac{\sum_i \mathbb{I}\{M_i(t) > c, T_i > t\}}{\sum_i \mathbb{I}\{T_i > t\}} = \widehat{S}_{M(t)|T>t}(c) \\ \widehat{c}^p(t) &= \widehat{S}_{M(t)|T>t}^{-1}(p) = \inf\{x : \widehat{S}_{M(t)|T>t}(x) \geq p\}. \end{aligned}$$

Heagerty and Zheng (2005) estimated incident TP using a Cox Model and riskset reweighting of the marker:

$$\begin{aligned} \widehat{\text{TP}}(c, t) &= \sum_{i \in \mathcal{R}(t)} \mathbb{I}\{M_i(t) > c\} \times \pi_i[\widehat{\gamma}(t), t] \\ &= \sum_{i \in \mathcal{R}(t)} \mathbb{I}\{M_i(t) > c\} \times \frac{\exp[M_i(t) \cdot \widehat{\gamma}(t)]}{W(t)} \end{aligned}$$

where $\gamma(t)$ is the log hazard ratio associated with the marker: $\lambda(t|M(t)) = \lambda_0(t) \exp[M(t) \times \gamma(t)]$ and $W(t) = \sum_{i \in \mathcal{R}(t)} \exp[M_i(t) \cdot \widehat{\gamma}(t)]$ is the normalizing constant. We adopt this estimator of TP for estimation of TTP. Suppose, $t_1 < t_2 < \dots < t_k$ are observed event times. Then TTP can be

estimated as

$$\widehat{\text{TTP}}(p) = \sum_{i=1}^k \left\{ \sum_{j \in \mathcal{R}(t)} \mathbb{1}\{M_j(t) > \widehat{c}^p(t_i)\} \times \pi_j[\widehat{\gamma}(t_i), t_i] \right\} \widehat{P}(T = t_i).$$

2.3 Comparison of markers

We now turn to the comparison of prognostic scores or markers, considering in particular the comparison of two markers that we label as marker A and marker B. We use subscripts A and B to identify the marker-specific TTP. In a case-control setting, marker comparison is done in two dimensions - TP and FP. Various measures, such as, absolute difference, odds ratio and relative probabilities may be used to compare $(\text{FP}_A, \text{TP}_A)$ with $(\text{FP}_B, \text{TP}_B)$. Among these metrics, relative probability $\left(\frac{\text{FP}_A}{\text{FP}_B}, \frac{\text{TP}_A}{\text{TP}_B}\right)$ is usually preferred for ease of interpretation and inference. In the present context, we fix the FP for both the markers and compare predictive accuracy in one dimension only. We compare TTP_A with TTP_B at a given FP fraction of p via relative probability of total true positive identified by each marker.

$$\widehat{\text{rTTP}}(p) = \frac{\widehat{\text{TTP}}_A(p)}{\widehat{\text{TTP}}_B(p)}.$$

2.4 Asymptotic Properties

Let us first assume that $c^p(t)$ and $\gamma(t)$ are known rather than estimated. Further, we condition on the observed event times. To show the asymptotic normality of proposed estimator, we first replace $W(t)$ with its expectation $\mathcal{W}(t)$ and rewrite TTP as:

$$\begin{aligned}
\widehat{\text{TTP}}(p) &= \sum_{i=1}^k \left\{ \sum_{j \in \mathcal{R}(t)} \mathbb{1}\{M_j(t) > c^p(t_i)\} \times \pi_j[\gamma(t_i), t_i] \right\} P(T = t_i) \\
&= \sum_{j=1}^n \sum_{i: j \in \mathcal{R}(t_i)} \mathbb{1}\{M_j(t_i) > c^p(t_i)\} \times \pi_j[\gamma(t_i), t_i] \times P(T = t_i) \\
&= \sum_{j=1}^n \sum_{i: j \in \mathcal{R}(t_i)} \mathbb{1}\{M_j(t_i) > c^p(t_i)\} \times \frac{\exp[M_j(t_i) \cdot \gamma(t_i)]}{\mathcal{W}(t_i)} \times P(T = t_i) \\
&= \sum_{j=1}^n \sum_{i: j \in \mathcal{R}(t_i)} \mathbb{1}\{M_j(t_i) > c^p(t_i)\} \times \exp[M_j(t_i) \cdot \gamma(t_i)] \times \underbrace{\frac{P(T = t_i)}{\mathcal{W}(t_i)}} \\
&= \sum_{j=1}^n \sum_{i: j \in \mathcal{R}(t_i)} \mathbb{1}\{M_j(t_i) > c^p(t_i)\} \times \exp[M_j(t_i) \cdot \gamma(t_i)] \times h(t_i).
\end{aligned}$$

Conditioning on the event times, $\widehat{\text{TTP}}(p)$ is a sum of contribution from independent subjects. Hence, CLT will hold for standardized $\widehat{\text{TTP}}(p)$. A variance estimator can be obtained by summing over the variance contribution from individual subjects and assuming a parametric distribution of the marker, e.g. normal. The parameters of the marker distribution may be estimated empirically among the appropriate subset of subjects. In practical applications, $\mathcal{W}(t)$, $c^p(t)$ and $\gamma(t)$ are replaced by their consistent estimators and the CLT will continue to hold for $\widehat{\text{TTP}}(p)$. However, the estimation of variance is not straight-forward in this situation and hence, a bootstrap variance estimator may alternatively be used that will also allow relaxing the parametric assumption for the marker.

For an unpaired design, asymptotic normality of $\log[\widehat{\text{TTP}}(p)]$ can be proved by invoking the asymptotic normality of TTP_A and TTP_B and their independence. For a paired design, the correlation between the markers needs to be taken into account. Asymptotic normality will hold for this scenario as well. A variance estimator may be obtained by assuming bivariate normality of the markers. Alternatively, bootstrap may be used to estimate the variance of the estimator.

3 Simulation

To demonstrate the validity of the summary survival ROC method introduced here, we conducted a simulation study. We assumed a standard normal marker M . The survival time given the marker $T|M = m$ was exponentially distributed with rate $\lambda = \exp(\beta m)$ such that the proportional hazard assumption holds for $T|M = m$. A value of $\beta > 0$ was chosen to ensure that a higher marker value is indicative of poor survival. For each of $m = 500$ simulated data sets, we generated (M, T) values for $n = 500$ subjects. Additionally, for each subject, an independent censoring time was also generated as exponential with rate λ such that 20% of the survival times were censored. The observed survival time was the minimum of the event time and censoring time. A censoring indicator was also generated. We restricted the follow-up time to τ such that $P(T \leq \tau) = 0.8$, to ensure that there were at least 100 subjects present in the riskset at all observed event times so that the quantiles could be estimated reasonably well. We used bootstrap to estimate the variance of the TTP. For each simulated data, $k = 200$ bootstrap replicates were generated by randomly resampling $n_b = 500$ subjects with replacement. For each simulation, we estimated the TTP at FP = 0.01 – 0.99. A range of values were chosen for β and here we show the results for $\beta = 0.9$. We present the true and estimated TTP, relative bias of the estimates, the Monte-Carlo SD, and empirical coverage of the confidence interval (nominal= 95.0) for a subset of FP values in Table 1. Additionally, we report the true and estimated area under the summary survival ROC curve (AUC). We note that the estimated ROC curve is in agreement with the true ROC curve. The relative bias is less than 2.6% for all FP values. The coverage based on the bootstrap variances is generally in agreement with the nominal level. The performance of the proposed estimator improved with a larger sample size (results not shown).

4 Example

In this section, we apply summary survival ROC curve to data from the Multicenter AIDS Cohort Study (MACS) (Kaslow et al., 1987). 5622 homosexual and bisexual men were enrolled

in the study and 3426 of them were sero-negative at the baseline. Between 1984 and 1996, 479 of these men became sero-positive. We analyzed a subset of 438 sero-positive subjects for whom the dates of sero-conversion were known to within ± 4.5 months. These subjects had an average of 13 measurements per person (3807 total observations). We evaluate the ability of percent CD4 lymphocyte (henceforth, CD4) and percent CD8 lymphocyte (henceforth, CD8) measures as predictors of progression to an AIDS diagnosis or death (whichever comes first). We use the 1987 CDC definition of AIDS, which relies on the symptoms rather than CD4 lymphocyte counts to define AIDS. Under this definition 176 sero-converters developed AIDS during the study period and 34 subjects died before the AIDS diagnosis.

The objective of the present analysis is to use a time-dependent marker based on CD4 and CD8 and judge the overall predictive accuracy of the marker via a summary survival ROC curve. We define a time-dependent composite marker based on sum and difference of the CD4 and sum and difference of CD8 from last two visits. The composite marker is relevant in the following sense. A large reduction in CD4 and CD8 around the time of sero-conversion is expected to be more indicative of a poor prognosis. However, if a subject's lymphocytes were stable (even though low) as reflected by the sum of the CD4 between these two visits, that would indicate a better prognosis. To capture both these aspects with a marker, we used a linear predictor from a Cox proportional hazard model with the four variables (sum and difference of CD4 between last two visit, and sum and difference CD8 between last two visits).

We display the composite longitudinal marker for controls (left hand panel) and cases (right hand panel) in Figure 1. The time-dependent marker threshold $c^p(t)$ for $FP = p = 0.1$ is also plotted in both the panels. The TTP at $FP = 0.1$ can be thought of as the proportion of cases (AIDS or death) who had a marker value above the solid, blue curve (10th) in the right hand panel. For this curve, the FP at each time point is fixed at 0.1. Contrast this with a time-invariant marker threshold (horizontal, dot-dashed, magenta line at marker = -2.00). This time-invariant marker threshold renders an overall FP of 0.1. However, when this threshold is used at all times, the time-specific FP varies widely from this marginal FP rate at each

time-point. For example, at $t = 20$ months the time-specific FP using this threshold of -2.00 is 0.02 while at $t = 100$ months the time-specific FP equals to 0.22.

A summary ROC curve along with 1-year and 10-year ROC curve is plotted in Figure 2. Comparing the 1-year and 10-year ROC curves, we see that the accuracy of the composite marker diminishes over time. For example, if a marker threshold is chosen to ensure that $FP = 0.2$, 87% of the cases are correctly identified at 1 year, but less than 40% of the cases are correctly identified at 10 year. Overall, about 56% of the cases are correctly identified if this threshold is used. The AUC for the summary survival ROC curve is 0.756. The AUCs for 1-year and 10-year ROC curves are 0.921 and 0.616 respectively.

We also compare the predictive accuracy of the composite marker with two other markers - one that is a linear combination of sum and difference of CD4 and the other is a linear combination of sum and difference of CD8 (Figure 3). We see that the marker based on CD8 is considerably less accurate. For example, at $FP = 0.2$, less than 40% of the cases will test positive using the CD8-based marker. However, the CD4-based marker is equivalent to the composite marker in terms of predictive accuracy, as seen from the summary survival ROC curves for the two markers. The AUC for the CD4-based marker is 0.751 and that for the CD8-based marker is 0.628.

5 Discussion

This article introduces a summary for time-dependent ROC curves that is useful in characterizing the overall predictive accuracy of a marker or a model score, when interest is in prediction of a censored survival time. When longitudinal markers are used to discriminate *incident events* from non-events, an overall measure of predictive accuracy of the marker may be desirable. The total true positive (TTP) at a given FP thresholds can be viewed as the total yield over time among the cases as detected by a corresponding marker threshold. The resulting ROC curve can be obtained as usual by plotting (FP, TTP) pairs by varying the threshold and may be used to compare the overall accuracy of two correlated markers. Left truncation

can also be accommodated to include the subjects in relevant risksets.

As mentioned before, this approach of summarizing the TP over a fixed FP fractions is different from the existing approach of summarizing both TP and FP over a given marker threshold (Emir et al., 1998, DeLong et al., 1985, Parker and DeLong, 2003). Referring to Figure 1, a fixed marker threshold would correspond to a horizontal line in both the panels (e.g., dot-dashed, magenta line). Though this may be useful in some cases, a time-invariant threshold fails to capture the time-dependent nature of the marker itself as was seen in Figure 1 and in other applications (Ito et al., 2003, Zheng and Heagerty, 2004, Deslandes and Chevret, 2007, Ye et al., 2008). We see that the increasing longitudinal trend of the marker is preserved by a time-dependent threshold. The concept is similar to covariate-adjusted ROC curve of Janes and Pepe (2009) that has been recently introduced in the context of a retrospective study.

If we are alternatively interested in obtaining a summary false positive, we can apply similar techniques to fix the TP at desired level and obtain a Total False Positive (TFP):

$$\begin{aligned}\tilde{c}^p(t) : p &= P(M(t) > \tilde{c}^p(t) | T = t) \\ \text{TFP}(p) &= \int_t P[M(t) > \tilde{c}^p(t)] \cdot \frac{P(T > t)}{\mathbb{E}(T)} dt\end{aligned}$$

The weights are adjusted by dividing with $\mathbb{E}(T)$ to ensure that $\text{TFP}(p) \in (0, 1)$. Note however, the two ROC curves would be different. The ROC curve for (FP, TTP) would represent the overall TP when FP is fixed while the ROC curve for (TFP, TP) would correspond to the overall FP when TP is fixed.

A number of aspects warrant additional research. First of all, the censoring time was assumed to be independent of the failure time. Relaxation of this assumption to allow conditional independence between the censoring time and failure time given the marker would be useful in practice. Further, accommodating discrete failure times or interval-censored time would be important since this type of data arise frequently in practice, especially, within the context of screening. Also, adjustment for competing risks would be useful and can be done

easily via cause-specific hazards and cumulative incidence based on our earlier work (Saha and Heagerty, 2010). We suggest using bootstrap approach for variance estimation. However, the issue of variance estimators based on different distributional assumption should be further investigated.

References

- Akritis, M. G. (1994). Nearest neighbor estimation of a bivariate distribution under random censoring. *Annals of Statistics*, 22 : 1299 – 1327.
- DeLong, E. R., Vernon, W. B., and Bollinger, R. R. (1985). Sensitivity and specificity of a monitoring test. *Biometrics*, 41(4): 947 – 958.
- Deslandes, E. and Chevret, S. (2007). Assessing surrogacy from the joint modelling of multi-variate longitudinal data and survival: Application to clinical trial data on chronic lymphocytic leukaemia. *Statistics in Medicine*, 26(30):5411–5421.
- Emir, B., Wieand, S., Jung, S., and Ying, Z. (2000). Comparison of diagnostic markers with repeated measurements: a non-parametric ROC curve approach. *Statistics in Medicine*, 19: 511 – 523.
- Emir, B., Wieand, S., Su, J. Q., and Cha, S. (1998). Analysis of repeated markers used to predict progression of cancer. *Statistics in Medicine*, 17: 2563 – 2578.
- Fisher, L. D. and Lin, D. Y. (1999). Time-dependent covariates in the Cox proportional-hazards regression model. *Annual Review of Public Health*, 20: 145 – 157.
- Heagerty, P. J. and Zheng, Y. (2005). Survival model predictive accuracy and ROC curves. *Biometrics*, 61 : 92 – 105.
- Ito, K., Yamamoto, T., Ohi, M., Takechi, H., Kurokawa, K., Suzuki, K., and Yamanaka, H. (2003). Natural history of PSA increase with and without prostate cancer. *Urology*, 62(1): 64 – 69.

- Janes, H. and Pepe, M. S. (2009). Adjusting for covariate effects on classification accuracy using the covariate-adjusted receiver operating characteristic curve. *Biometrika*, 96(2): 371 – 382.
- Jewell, N. P. and Kalbfleisch, J. D. (1996). Marker process in survival analysis. *Lifetime Data Analysis*, 2: 15 – 29.
- Jewell, N. P. and Nielsen, J. P. (1993). A framework for consistent prediction rules based on markers. *Biometrika*, 80(1): 153 – 164.
- Kaslow, R. A., Ostrow, D. G., Detels, R., et al. (1987). The multicenter AIDS cohort study: Rationale, organization and selected characteristics of the participants. *American Journal of Epidemiology*, 126(2): 310 – 318.
- Koenig, W. (2010). Integrating biomarkers: The new frontier? *Scandinavian Journal of Clinical & Laboratory Investigation*, 70: 117 – 123.
- Mammen, E. and Nielsen, J. P. (2007). A general approach to the predictability issue in survival analysis with application. *Biometrika*, 94(4): 873 – 892.
- Murtaugh, P. A. (1995). ROC curves with multiple marker measurements. *Biometrics*, 51(4): 1514 – 1522.
- Parker, C. B. and DeLong, E. R. (2003). ROC methodology within a monitoring framework. *Statistics in Medicine*, 22: 3473 – 3488.
- Rizopoulos, D. (2011). Dynamic predictions and prospective accuracy in joint models for longitudinal and time-to-event data. *Biometrics*, 67(3):819–829.
- Saha, P. and Heagerty, P. J. (2010). Time-dependent predictive accuracy in the presence of competing risks. *Biometrics*, 66(4): 999 – 1011.
- Saha-Chaudhuri, P. and Heagerty, P. J. (2012). Non-parametric estimation of time-dependent predictive accuracy curve. *In press, Biostatistics*.

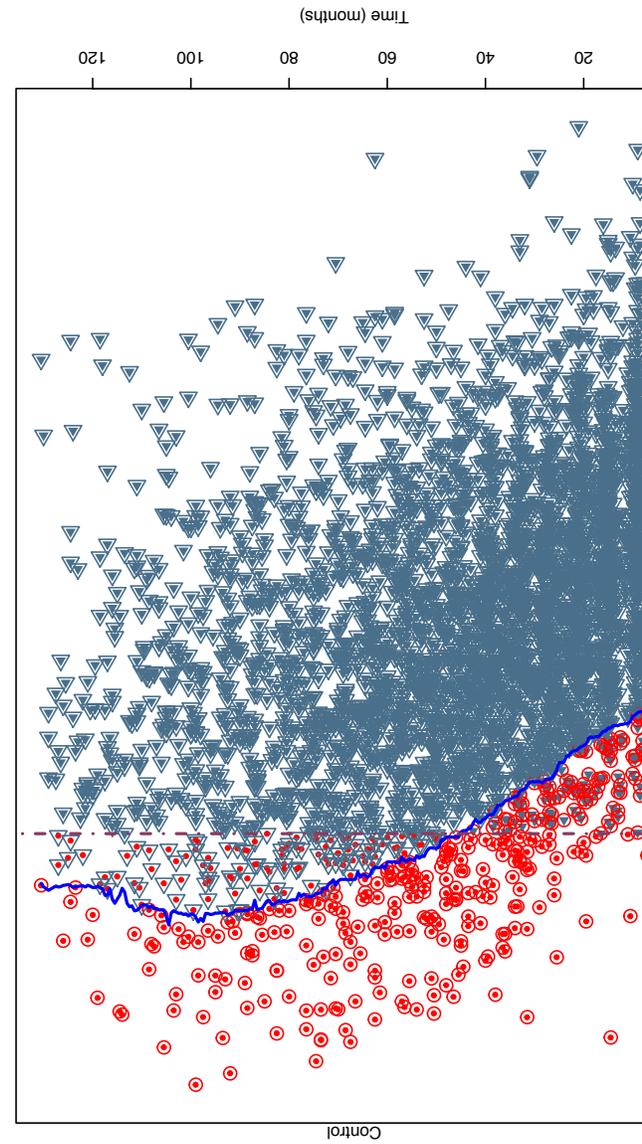
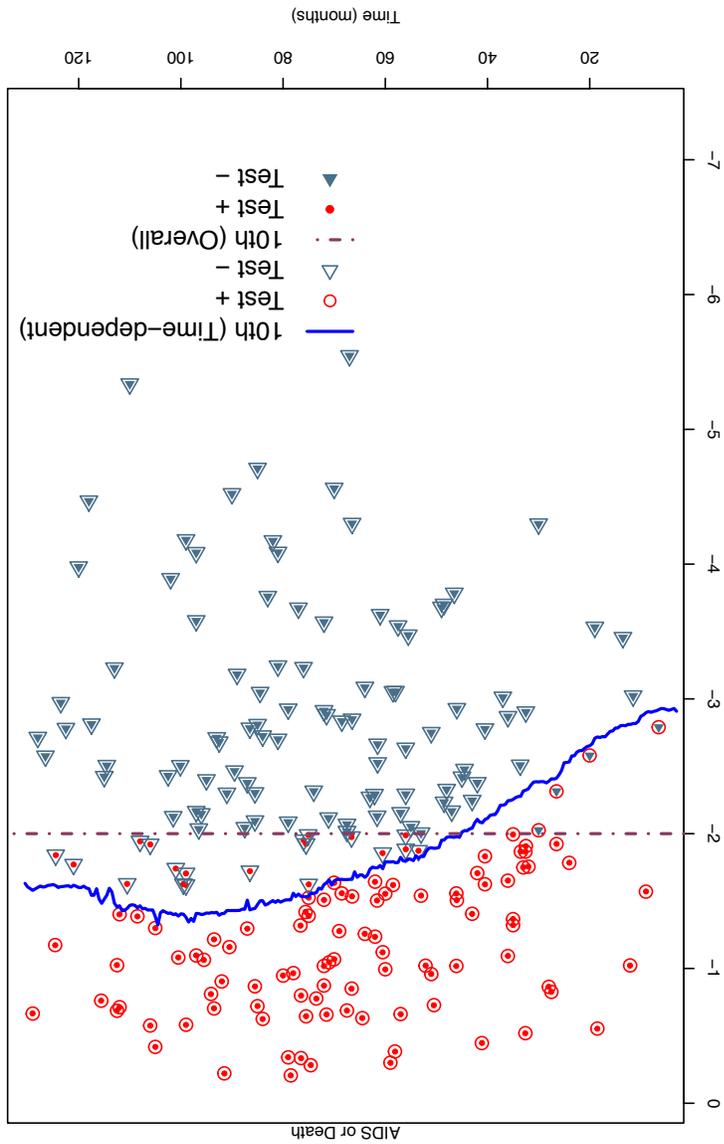
- Thompson, M. L. (2003). Assessing the diagnostic accuracy of a sequence of tests. *Biostatistics*, 4(3):341–351.
- Xue, Y., Clopton, P., Peacock, W. F., and Maisel, A. S. (2011). Serial changes in high-sensitive troponin i predict outcome in patients with decompensated heart failure. *European Journal of Heart Failure*, 13(1): 37 – 42.
- Ye, W., Lin, X., and Taylor, J. M. G. (2008). Semiparametric modeling of longitudinal measurements and time-to-event data-a two-stage regression calibration approach. *Biometrics*, 64(4):1238–1246.
- Zheng, Y. and Heagerty, P. (2004). Semiparametric estimation of time-dependent ROC curves for the longitudinal marker data. *Biostatistics*, 5: 615 – 632.

6 Tables and Figures

Table 1: Estimated and true TTP based on a marker and time under the proportional hazard assumption based on a sample size of $N = 500$. The marker is normally distributed and the time is exponentially distributed conditional on the marker. The estimates are averaged over 500 simulated datasets and for each dataset 200 bootstrap replicates were considered.

FP	TTP	Estimate	Relative Bias (%)	MCSD	Coverage ¹
0.01	0.053	0.052	-0.026	0.005	0.910
0.02	0.091	0.091	-0.006	0.008	0.936
0.03	0.124	0.124	0.001	0.010	0.938
0.04	0.153	0.154	0.004	0.011	0.944
0.05	0.180	0.181	0.006	0.012	0.944
0.10	0.294	0.297	0.009	0.016	0.940
0.20	0.463	0.467	0.008	0.019	0.940
0.30	0.592	0.595	0.006	0.019	0.938
0.40	0.695	0.697	0.004	0.017	0.932
0.50	0.779	0.781	0.003	0.015	0.936
0.60	0.847	0.849	0.002	0.012	0.926
0.70	0.903	0.904	0.001	0.009	0.928
0.80	0.948	0.948	0.000	0.006	0.938
0.90	0.981	0.981	0.000	0.003	0.936
AUC	0.704	0.706	0.006	0.011	0.932

¹SD-based. Nominal: 95.0



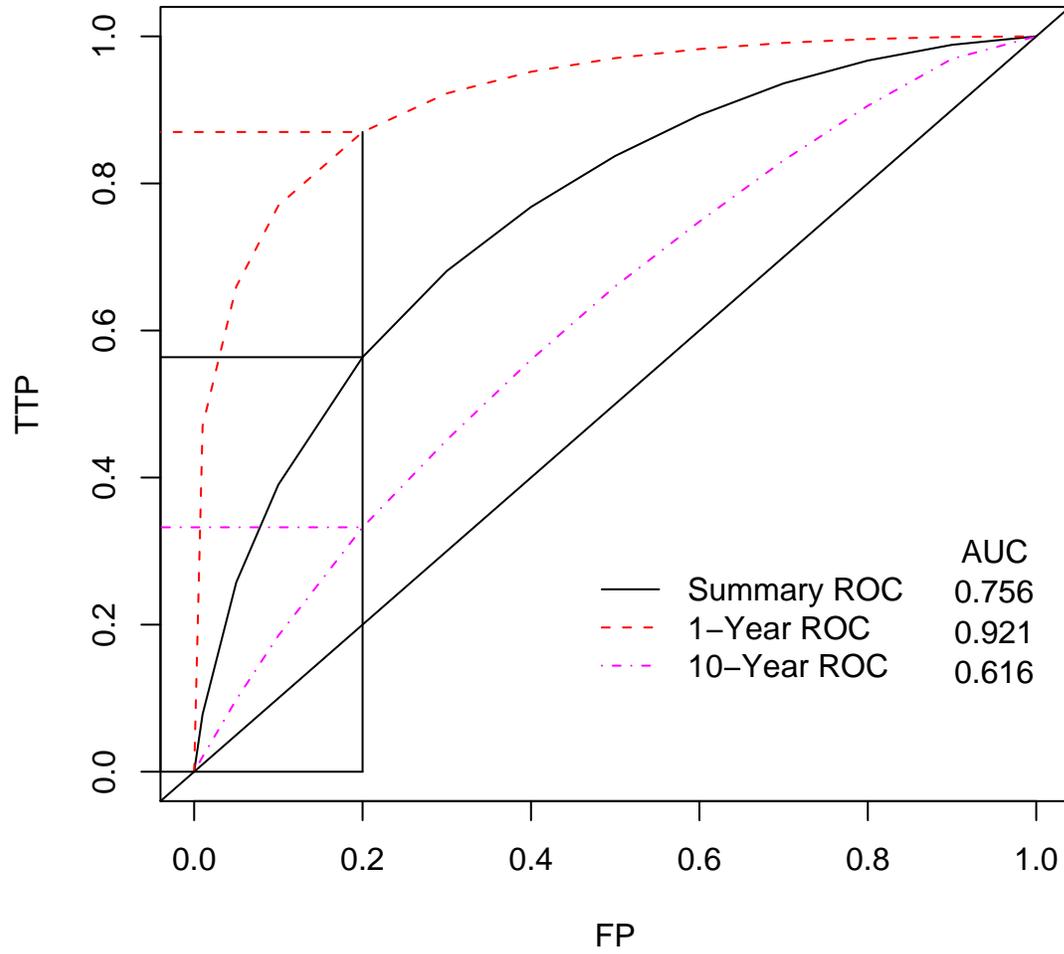


Figure 2: Summary survival ROC curve and 1-year and 10-year ROC curves for a composite marker from MACS data.

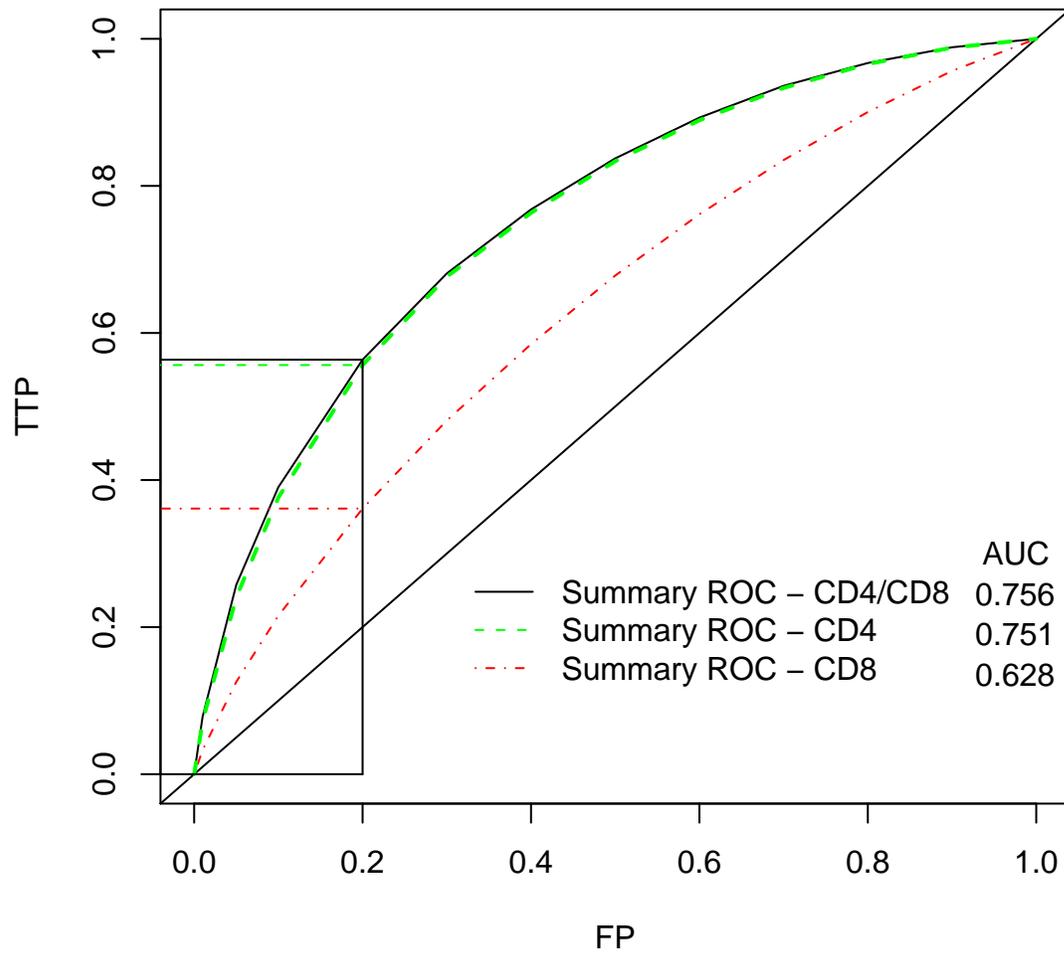


Figure 3: Summary survival ROC curve for three markers from MACS data.