12-19-2003

Semiparametric Estimation of Time-Dependent: ROC Curves for Longitudinal Marker Data

Yingye Zheng
Fred Hutchinson Cancer Research Center, yzheng@fhcrc.org

Patrick Heagerty
University of Washington, heagerty@u.washington.edu

1. Introduction

Recent developments in molecular technology such as gene-expression microarrays and proteomics have led to a surge in research aimed at discovering markers useful for disease screening or for providing patient prognosis. We consider a marker to be any measurement that has the potential to signal the onset or progression of disease. Examples of potential markers include prostate specific antigen (PSA) as an indicator of prostate cancer, or pulmonary function measures such as forced expiratory volume (FEV1) as an indicator of functional decline prior to death among cystic fibrosis patients. Longitudinal marker analysis seeks to evaluate whether changes in a marker process correlate with a key clinical event time such as disease onset or death. Ideally a marker process would provide definitive and early information regarding disease onset. However, in practice most biological measurements contain substantial variability which complicates their use in guiding health care decisions. Characterizing the accuracy of time-varying markers for disease detection is therefore a critical step in the marker discovery process.

Retrospective longitudinal studies can be useful in a key phase of marker development (Pepe, Etzioni, Feng, Potter, Thompson, Thornquist, Winget and Yasui 2001). With this design the capacity of a marker to detect preclinical disease can be evaluated as a function of time prior to the clinical occurrence of signs and symptoms. The scientific goal is to identify markers that exhibit high discriminatory ability at relatively early stages of disease. Traditionally in medical diagnostic research the receiver operating characteristic (ROC) curve is used to summarize the accuracy or discrimination potential of a marker measurement. (Hanley 1989; Zweig and Campbell 1993). However, to characterize early disease detection the cross-sectional concepts of sensitivity and specificity that are typically displayed in an ROC curve need to be extended to incorporate both the time-varying nature of a marker and the clinical onset time of the disease. In this article we consider new semiparametric statistical methods for estimating time-dependent sensitivity, specificity, and ROC curves that adopt weak distributional assumptions. In addition, we develop the asymptotic distribution theory that allows inference for either time-dependent sensitivity and specificity, or for time-dependent ROC curves. Our methods also allow accuracy summaries to depend on covariates, and thus can be used to evaluate whether the accuracy of a marker varies across patient subgroups.

In the classic diagnostic testing setting, disease status is represented by a binary variable D_i for subject i. Furthermore, the relationship between D_i and a continuous test result Y_i for a given subgroup defined by a covariate vector $\mathbf{Z}_i$ is displayed through a conditional ROC curve. A conditional ROC curve displays the full spectrum of values for sensitivity, or the "true positive rate" (TP), $P(Y_i < c|D_i = 1, \mathbf{Z}_i)$, and 1 minus the specificity, or the "false positive rate" (FP), $1 - P(Y_i \geq c|D_i = 0, \mathbf{Z}_i)$ by considering all possible test threshold values c. Therefore, an ROC curve is a plot of TP versus FP for $c \in (-\infty, \infty)$. In our definitions for TP and FP we assume that a low marker value is indicative of disease, but parallel definitions can be adopted when high marker values are associated with disease. When tests are measured longitudinally, we can define TP and FP as time-dependent functions

$$TP(s, c|t, \mathbf{Z}_i) = P[Y_i(s) < c|D_i(t) = 1; T_i > s, \mathbf{Z}_i] \quad (1.1)$$

$$FP(s, c|t, \mathbf{Z}_i) = P[Y_i(s) < c|D_i(t) = 0; T_i > s, \mathbf{Z}_i]. \quad (1.2)$$

Here $Y_i(s)$ indicates a marker measured at time s, with $s \geq t_0$, the baseline time, and $D_i(t)$ is a binary variable indicating the disease status of subject i at time t. We recognize that there can be two useful time-varying "case" definitions to obtain $D_i(t) = 1$ based on the disease time T_i: if we are interested in the incidence of the disease at time t, then we can define $D_i(t) = 1$ when $T_i = t$; if we are interested in the prevalence of the disease by time t, we can define $D_i(t) = 1$ when $T_i \leq t$.

In this manuscript we focus on incident ROC curves defined as a plot of $[TP^{\mathbb{I}}(s, c|t, \mathbf{Z}), FP^{\mathbb{C}}(s, c|t^\star, \mathbf{Z})]$ for all possible c, where

$$TP^{\mathbb{I}}(s, c|t, \mathbf{Z}_i) = P[Y_i(s) < c|T_i = t; \mathbf{Z}_i, T_i > s] \quad (1.3)$$

$$FP^{\mathbb{C}}(s, c|t^\star, \mathbf{Z}_i) = P[Y_i(s) < c|T_i > t^\star; \mathbf{Z}_i, T_i > s] \ . \quad (1.4)$$

Here, we use $t^\star$ in the definition of FP to denote as controls those subjects that do not experience the disease prior to a fixed follow-up time. This definition is useful when some subjects are thought to avoid disease and may reasonably be identified as the "long term survivors" characterized by $T_i > t^\star$. Incident ROC curves are most useful in a retrospective study when the timing of diagnosis for diseased patients is certain.

Incorporating the time dimension in ROC analysis has recently been discussed by a number of authors (Etzioni et al. 1999; Slate and Turnbull 2000; Heagerty, Lumley and Pepe 2000; Cai and Pepe 2002). Non-parametric methods that characterize accuracy using disease prevalence for the case definition, $D_i(t) = 1(T_i \leq t)$, is given in Heagerty et al. (2000). Other authors have used the incident case definition, $D_i(t) = 1(T_i = t)$, including Etzioni et al. (1999), Slate and Turnbull (2000), and Cai and Pepe (2002). There are two general approaches to characterizing accuracy and to calculating covariate specific ROC curves. First, direct regression approaches for ROC analysis have been proposed (see Pepe (1998) and Pepe (2003) for reviews). These methods use generalized linear model concepts to characterize the shape of the ROC curve and to allow covariates to directly impact accuracy. For longitudinal data Etzioni et al. (1999) illustrate the ROC-GLM approach for evaluating time-dependent accuracy. Despite recent work that permits flexibility in the direct ROC regression approach (Cai and Pepe 2002; Li, Tiwari and Wells 1999), the methods have not been extended to allow covariates to impact the link function, and thus potentially place restrictions on how the shape of the ROC curve varies with covariates. The semi-parametric methods that we introduce in section 2 permit the shape of the ROC curve to change smoothly with covariates.

A second general approach to estimating covariate specific ROC curves is based on modelling of the marker distribution conditional on disease status and covariates. Given estimates of the marker distribution for cases and controls an induced covariate specific ROC can be calculated (Tosteson and Begg 1988). For longitudinal markers Etzioni et al. (1999) and Slate and Turnbull (2000) discuss a parametric regression approach that uses linear mixed models to characterize the longitudinal marker distribution for diseased and non-diseased subjects. In this approach the disease onset time, T_i, is an additional covariate for the cases only. Our proposal is to use smooth semi-parametric regression quantile methods to model the marker distribution as a function of disease status, covariates, and disease onset time for the cases. By using flexible regression quantile methods we can relax the distributional assumptions adopted by previous proposals.

In section 2 we describe the estimation and inference procedures. We outline the asymptotic distribution theory for the proposed estimators in section 3. We use pulmonary function data from cystic fibrosis patients to illustrate the method and to compare semi-parametric analysis results to

parametric results.

2. Semiparametric Regression Quantile Method

2.1 Notation

In many studies subjects are classified either as disease cases or as controls on the basis of the observed event time. For example, control subjects may be defined as individuals who after $t^\star$ time units remain free of disease, while cases are individuals with an observed event time, $T_i < t^\star$. Here we consider a case-control study setting, where there are n_D cases and $n_{\bar{D}}$ controls among n subjects (i.e., $n = n_D + n_{\bar{D}}$). Let Y_{ik} be the continuous random variable representing the marker value obtained from the ith case at the kth visit time s_{ik}, with $i = 1, \ldots, n_D$, and $k = 1, \ldots, K_i$. For cases let $\mathbf{Z}_{ik}^T = vec(T_i, s_{ik}, \mathbf{X}_i)$, or $\mathbf{Z}_D$, denote a vector of covariates associated with Y_{ik}, where $\mathbf{X}_i$ denotes a vector of covariates that do not change with time. The total number of observations from the case population is $N_D = \sum_{i=1}^{n_D} K_i$. Similarly, denote Y_{jl} as the marker value obtained from the jth control at the lth visit time s_{jl}, with $j = n_D + 1, \ldots, n_D + n_{\bar{D}}$, and $l = 1, \ldots, L_j$. For controls let $\mathbf{Z}_{jl}$ or $\mathbf{Z}_{\bar{D}}$ denote a vector of covariates associated with Y_{jl}. Since there are no disease diagnosis times for controls, $\mathbf{Z}_{jl}^T = vec(s_{jl}, \mathbf{X}_j)$ with $\mathbf{X}_j$ as the time-independent covariate vector for the jth control. The total number of observations for the control group is $N_{\bar{D}} = \sum_{j=1}^{n_{\bar{D}}} L_j$. We thus have $N = N_D + N_{\bar{D}}$ observations from the two populations.

2.2 Semiparametric Estimation of Regression Quantiles

We propose to use semiparametric regression quantile estimation in order to construct time-dependent ROC curves. Similar to the methods of Heagerty and Pepe (1999), we characterize the conditional distribution of $[Y_{ik}|\mathbf{Z}_{ik}]$ as from a location-scale family. Specifically, we represent a case marker value as

$$Y_{ik} = \mu_D(\mathbf{Z}_{ik}) + \sigma_D(\mathbf{Z}_{ik})\epsilon_D(\mathbf{Z}_{ik}) \tag{2.1}$$

where μ_D and σ_D are the location and scale functions, and the baseline distribution function for ϵ_D is

$$F_{0,\mathbf{z}_D}(\epsilon) = P[\epsilon_D(\mathbf{Z}_{ik}) \leq \epsilon|\mathbf{Z}_{ik}]. \tag{2.2}$$

We characterize the conditional distribution of the marker measured from the control population in the same way, with $\mu_{\bar{D}}$ and $\sigma_{\bar{D}}$ representing the location and scale functions, and with the baseline distribution function denoted

$$G_{0,\mathbf{z}_{\bar{D}}}(\epsilon) = P[\epsilon_{\bar{D}}(\mathbf{Z}_{jl}) \leq \epsilon | \mathbf{Z}_{jl}]. \tag{2.3}$$

As in Heagerty and Pepe (1999), for continuous $\mathbf{Z}_D$, we model $\mu_D(\mathbf{Z}_{ik})$ and $\sigma_D(\mathbf{Z}_{ik})$ as smooth functions of $\mathbf{Z}$ parametrically using regression splines:

$$\mu_D(\mathbf{Z}_{ik}) = \sum_{p=1}^{P} \beta_p R_p(\mathbf{Z}_{ik}) \tag{2.4}$$

$$log\sigma_D(\mathbf{Z}_{ik}) = \sum_{q=1}^{Q} \gamma_q S_q(\mathbf{Z}_{ik}) \ ; \tag{2.5}$$

where $R_p(\mathbf{Z}_{ik})$ and $S_q(\mathbf{Z}_{ik})$ are regression spline basis functions. Heagerty and Pepe (1999) used a quasi-likelihood method for estimation. For example, to estimate $\hat{\beta}_p$ and $\hat{\gamma}_p$ (for $p = 1, \ldots, P$) associated with the case population, we can simultaneously solve the following estimating equations:

$$\sum_{i=1}^{n_D} \sum_{k=1}^{K_i} R(\mathbf{Z}_{ik})^T [Y_{ik} - \mu_D(\mathbf{Z}_{ik})] / \sigma_D^2(\mathbf{Z}_{ik}) = 0 \tag{2.6}$$

$$\sum_{i=1}^{n_D} \sum_{k=1}^{K_i} S(\mathbf{Z}_{ik})^T \{[Y_{ik} - \mu_D(\mathbf{Z}_{ik})]^2 - \sigma_D^2(\mathbf{Z}_{ik})\} / \sigma_D^2(\mathbf{Z}_{ik}) = 0 \tag{2.7}$$

Estimation for $\mu_{\bar{D}}$ and $\sigma_{\bar{D}}$ follows in a parallel fashion.

Similar to Heagerty and Pepe (1999), the baseline distribution functions F_0 and G_0 are left unspecified, and can be estimated empirically based on the standardized residual $\epsilon_{ikD} = \epsilon_D(\mathbf{Z}_{ik}) = [Y_{ik} - \mu_D(\mathbf{Z}_{ik})]/\sigma_D(\mathbf{Z}_{ik})$ and $\epsilon_{jl\bar{D}} = \epsilon_{\bar{D}}(\mathbf{Z}_{jl}) = [Y_{jl} - \mu_{\bar{D}}(\mathbf{Z}_{jl})]/\sigma_{\bar{D}}(\mathbf{Z}_{jl})$. If the distributions of the standardized residuals are independent of covariates, the natural estimators for F_0 and G_0 are the empirical distribution functions of the forms

$$\hat{F}_0(\epsilon) = \sum_{i=1}^{n_D} \sum_{k=1}^{K_i} \frac{1}{N_D} \mathbf{1}(\hat{\epsilon}_{ikD} \leq \epsilon) \tag{2.8}$$

and

$$\hat{G}_0(\epsilon) = \sum_{j=1}^{n_{\bar{D}}} \sum_{l=1}^{L_j} \frac{1}{N_{\bar{D}}} \mathbf{1}(\hat{\epsilon}_{jl\bar{D}} \leq \epsilon), \tag{2.9}$$

where $\hat{\epsilon}_{ikD} = [Y_{ik} - \hat{\mu}_D(\mathbf{Z}_{ik})]/\hat{\sigma}_D(\mathbf{Z}_{ik})$ and $\hat{\epsilon}_{jl\bar{D}} = [Y_{jl} - \hat{\mu}_{\bar{D}}(\mathbf{Z}_{jl})]/\hat{\sigma}_{\bar{D}}(\mathbf{Z}_{jl})$. However, more general methods can be adopted when the baseline distribution functions also vary with covariates $\mathbf{Z}_D$ or $\mathbf{Z}_{\bar{D}}$.

In this manuscript we consider the symmetrized nearest neighbor (SNN) estimator (Yang 1981) for the conditional baseline distribution function, which is defined as a weighted sum

$$\hat{F}_0(\epsilon|\mathbf{Z}, a_n) = \sum_i^{n_D} \sum_k^{K_i} \frac{1}{W(\mathbf{Z})} w_{a_n}(\mathbf{Z}, \mathbf{Z}_{ik}) \mathbf{1}(\hat{\epsilon}_{ikD} \leq \epsilon) \tag{2.10}$$

where $W(\mathbf{Z}) = \sum w_{a_n}(\mathbf{Z}, \mathbf{Z}_{ik})$ and a_n is a bandwidth parameter. A general form for w_{a_n} can be assumed with

$$w_{a_n}(\mathbf{Z}, \mathbf{Z}_{ik}) = K\left[\frac{H_n(\mathbf{Z}) - H_n(\mathbf{Z}_{ik})}{a_n}\right] \tag{2.11}$$

where $K(\cdot)$ is a continuous and bounded probability kernel on [-1,+1], $H_n(\cdot)$ is the empirical distribution function of $\mathbf{Z}$. An SNN estimator for $G_0(\epsilon|\mathbf{Z}, a_n)$ can be similarly defined. In the following, we use the notation $F_{0,\mathbf{z}_D}$, $G_{0,\mathbf{z}_{\bar{D}}}$ and $F_0(\epsilon|\mathbf{Z} = \mathbf{z}_D, a_n)$, $G_0(\epsilon|\mathbf{Z} = \mathbf{z}_{\bar{D}}, a_n)$ interchangeably.

2.3 Semiparametric Regression-Quantile-Based Estimator of ROC Curves

We now construct an estimator for a time-dependent ROC curve based on the semiparametric regression quantile estimators. We estimate an ROC curve for a given covariate value $\mathbf{Z}$ by first modeling the marker as functions of covariates $\mathbf{Z}_D$ for the cases or $\mathbf{Z}_{\bar{D}}$ for controls. By using the semiparametric regression quantile methods to model the marker, we can derive a semiparametric estimator for conditional ROC curves. For convenience, we define a positive test as a marker value less than a certain threshold, and thus the true positive rate (TP) is defined in terms of the cumulative distribution function $F(c)$ instead of survival function $1 - F(c)$. For any false positive rate $p \in [0, 1]$, our proposed estimator for the ROC curve is

$$\begin{aligned} \widehat{ROC}_{\mathbf{z}}(p) &= \hat{F}_{0,\mathbf{z}_D}\left[\frac{\hat{\mu}_{\bar{D}}(\mathbf{z}_{\bar{D}}) - \hat{\mu}_D(\mathbf{z}_D)}{\hat{\sigma}_D(\mathbf{z}_D)} + \hat{G}^{-1}_{0,\mathbf{z}_{\bar{D}}}(p)\frac{\hat{\sigma}_{\bar{D}}(\mathbf{z}_{\bar{D}})}{\hat{\sigma}_D(\mathbf{z}_D)}\right] \\ &\equiv \hat{F}_{0,\mathbf{z}_D}\left[\hat{\alpha}_0 + \hat{G}^{-1}_{0,\mathbf{z}_{\bar{D}}}(p)\hat{\alpha}_1\right] \end{aligned} \tag{2.12}$$

where $F_{0,\mathbf{z}_D}$, $G_{0,\mathbf{z}_{\bar{D}}}$, $\mu_D(\mathbf{z}_D)$, $\mu_{\bar{D}}(\mathbf{z}_{\bar{D}})$, $\sigma_D(\mathbf{z}_D)$ and $\sigma_{\bar{D}}(\mathbf{z}_{\bar{D}})$ are defined as in section 2.2. $\hat{G}^{-1}_{0,\mathbf{z}_{\bar{D}}}(p)$ is the conditional empirical quantile function, with $\hat{G}^{-1}_{0,\mathbf{z}_{\bar{D}}}(p) = inf[\epsilon : \hat{G}_{0,\mathbf{z}_{\bar{D}}}(\epsilon) \geq p]$.

3. Asymptotic Distribution Theory for ROC Curve Estimators

In this section we derive the large sample properties of our semiparametric ROC estimators. We first consider an ROC curve estimator that assumes the baseline distribution functions do not vary with

covariates **Z**. We then consider an ROC curve estimator based on SNN estimation of the baseline distribution function, assuming dependence on covariates. Proofs for the main theorems are detailed in the appendix.

Hsieh and Turnbull (1996) studied the asymptotic properties of a nonparametric ROC estimator of the form $F[G^{-1}(p)]$, and a semiparametric ROC estimator where the distributions of F and G are normal. In the simplest situation where there are no covariates and each subject contributes a single observation, our estimators reduce to that of Hsieh and Turnbull (1996).

3.1 Baseline Distribution Functions Are Independent of Covariates

When baseline functions are independent of covariates, we use equation 2.8 and equation 2.9 to estimate the baseline distribution functions F_0 and G_0.

For asymptotic results we assume the following regularity conditions:

- F_0, G_0 are continuous with continuous densities f_0, g_0, respectively.
- The slope of the ROC curve $f_0[\alpha_0 + G_0^{-1}(p)\alpha_1]/g_0[G_0^{-1}(p)]$, is bounded on any interval (a, b), with $0 < a < b < 1$.
- $n_D/n \to \lambda > 0$ as $n \to \infty$.
- The number of observations per subject is relatively small with respect to n, the total number of subjects. i.e., we assume $N_D/n_D \to c_D$ and $N_{\bar{D}}/n_{\bar{D}} \to c_{\bar{D}}$.

Theorem 1 (Consistency of ROC curve estimator) *Under the conditions above,*

$$\sup_{0\le p\le 1} |\widehat{ROC}_{\boldsymbol{z}}(p) - ROC_{\boldsymbol{z}}(p)| \to 0 \ a.s. \ as \ n \to \infty.$$

Theorem 2 (Asymptotic normality) *Under the conditions above, there exists a probability space on which one can define sequences of two independent zero-mean Gaussian variables* $[\, U_1(p),\ U_2(p), 0 \le p \le 1\,]$ *respectively such that for* $0 < a < b < 1$,

$$\sqrt{n}[\widehat{ROC}_{\boldsymbol{z}}(p) - ROC_{\boldsymbol{z}}(p)] \to \Psi(p), a.s. \ as \ n \to \infty \quad uniformly \ on \ [a,b].$$

with

$$\Psi(p) = \frac{1}{c_D\sqrt{\lambda}} U_1(p) + \frac{1}{c_{\bar{D}}\sqrt{1-\lambda}} \alpha_1 \frac{f_0[\alpha_0 + G_0^{-1}(p)\alpha_1]}{g_0[G_0^{-1}(p)]} U_2(p)$$

We specify the forms for $U_1(p)$ and $U_2(p)$ in the appendix.

3.2 Baseline Distribution Functions Are Smooth Functions of Covariates

In this section we provide asymptotic results for the situation in which baseline functions are smooth functions of covariates. We use equation (2.10) to estimate the baseline distribution functions $F_{0,\mathbf{z}}$ and use a similar procedure for $G_{0,\mathbf{z}}$.

In addition to the assumptions in section 3.1, we also require the following:

- $K(x)$ is twice continuously differentiable probability kernel which vanishes outside some finite interval.
- $(a_n)_n$ is a sequence of bandwidths converging to zero, and $na_n^3 \to \infty$ as $n \to \infty$.
- $\mathbf{Z}$ has continuous distribution function $H(\cdot)$.

Theorem 3 (Consistency) *Under the conditions above,*

$$\sup_{0\le p\le 1} |\widehat{ROC}_z(p) - ROC_z(p)| \to 0 \ a.s. \ as \ n \to \infty.$$

Theorem 4 (Asymptotic normality of SSN ROC estimator) *Under the conditions above, there exist a probability space on which one can define sequences of two independent zero-mean Gaussian random variables* $[\, U_1^\star(p),\ U_2^\star(p),\ 0 \le p \le 1\,]$ *such that for* $0 < a < b < 1$,

$$\sqrt{na_n}[\widehat{ROC}_{\mathbf{z}}(p) - ROC_{\mathbf{z}}(p)] \to \Psi(p), a.s. \ as \ n \to \infty \ \ uniformly \ on \ [a,b].$$

with

$$\Psi(p) = \frac{1}{c_D\sqrt{\lambda}} U_1^\star(p) + \frac{1}{c_{\bar{D}}\sqrt{1-\lambda}} \alpha_1 \frac{f_{0,z}[\alpha_0 + G_{0,z}^{-1}(p)\alpha_1]}{g_{0,z}[G_{0,z}^{-1}(p)]} U_2^\star(p)$$

We specify the forms for $U_1^\star(p)$ and $U_2^\star(p)$ in the appendix.

3.3 Bandwidth Selection

Methods that guide the selection of smoothing parameters facilitate optimal estimation by balancing bias and variance. In practice one can either employ cross-validation to select a bandwidth that minimizes a pre-specified loss function, or adopt a data-driven procedure to estimate the asymptotically optimal bandwidth. Ideally if the statistical objective is to estimate conditional ROC curves

then a criterion focusing directly on key aspects of the ROC curve would be desirable. However, our methodology is "indirect" in that we first model the marker distribution separately for cases and controls, and then construct the induced conditional ROC curves. Given our approach, we adopt an optimal bandwidth selection strategy that is based on the asymptotic distribution of conditional quantile estimators for the cases and the controls.

For either cases or controls a given percentile $p \in [0,1]$ defines a quantile function $\epsilon_p(x) = F^{-1}(p|x)$. A theoretical bandwidth that minimizes the asymptotic integrated mean square error (IMSE) is of the form $a_n = (b/n)^{1/5}$, with

$$b_{opt}(p) = \left(\int \frac{\sigma^2[\epsilon_p(x)|x]}{\{f[\epsilon_p(x)]|x\}^2} dx \right) / \left(\mu_2 \int \left\{ \frac{\omega[\epsilon_p(x)|x]}{f[\epsilon_p(x)|x]} \right\}^2 dx \right) \tag{3.1}$$

where $\sigma^2[\epsilon_p(x)|x]$ is the variance of the estimator of $F[\epsilon_p(x)|x, a_n]$, $\omega[\epsilon_p(x)|x]$ denotes the second derivative of the smoothed conditional empirical process evaluated at $\epsilon_p(x)$, and $f[\epsilon_p(x)|x, a_n]$ is the conditional density function for ϵ_{ij} evaluated at $\epsilon_p(x)$. Ducharme, Gannoun, Guertin and Jéquier (1995) give a detailed discussion of estimation for the necessary components in equation (3.1) when analyzing independent data. For dependent data we simply modify the estimate of $\sigma^2[\epsilon_p(x)|x]$ using a robust variance estimator (see the appendix for details). Since b_{opt} involves unknown quantities such as $\sigma^2[\epsilon_p(x)|x]$ and $\omega[\epsilon_p(x)|x]$, in applications we simply use a plug-in estimator by first choosing an arbitrary a_n^* of order $n^{1/5}$, and then estimating the terms in $b_{opt}(p)$ based on the preliminary estimates using a_n^*.

Note that $b_{opt}(p)$ is a function of p and therefore may vary for different percentile values. In ROC analysis we seek to estimate the entire distribution function (all quantiles) rather than focus on a single percentile value, p. However, not all values for the false positive rate are scientifically important, and in general interest primarily targets the true positive rate (sensitivity) that corresponds to low false positive rates (1-specificity). Therefore, for controls we consider $b_{opt}(p)$ calculated for a range of values such as $0.75 < p < 0.95$, and for cases we consider a slightly wider range such as $0.60 < p < 0.95$. For cases and controls, we then evaluate the range of optimal bandwidths and for each group we choose a single value which is nearly optimal over the percentiles of interest. Further work is warranted toward developing an optimal bandwidth selection algorithm that is specially tailored for ROC analysis.

4. Example: Cystic Fibrosis Data

4.1 Study Description

The cystic fibrosis (CF) data we analyze come from the U.S. Cystic Fibrosis Foundation (CFF) National Patient Registry, a database of patients with CF seen at CFF-accredited care centers. The registry has been updated annually since 1966 and contains longitudinal measures of participant health status. Registry information available from 1990-1998 provides 9 years of follow-up on 21,138 patients for a total of N = 171,306 observations. During the follow-up period, 8.52% of the patients died and thus most of the longitudinal marker information comes from patients who remain alive through the end of the observation period.

In the subsequent analysis we focus on a standardized pulmonary function measurement known as FEV1 percent-predicted. This outcome measures the volume of air that a subject expires in 1 second, and standardizes by dividing the raw volume by the age and gender specific population reference value. Thus, an FEV1 of 100 indicates that a subject has 100% of the expected lung function for a healthy child of his/her age. Currently, children are registered for lung transplantation if their FEV1 drops below a value of 30 (Davis 1997). One clinical question is whether this is an accurate decision threshold, and whether an age-specific criterion might be warranted. Specifically, what is the accuracy of the decision guideline? Does this identify most children who would otherwise die in the near future, and does it capture few of the children who are likely to survive? To examine the prognostic value of FEV1 at various ages we characterize age-specific sensitivity and specificity. By estimating conditional quantiles we can compute the threshold value that correctly identifies a given percent of children who do progress to death (i.e. the sensitivity), and similarly the threshold value that incorrectly identifies a given percent of children who do not progress to death (i.e. the false positive rate, or 1-specificity). Furthermore, by creating age-specific ROC curves we can determine whether controlling the false positive rate at a low value such as 10% leads to adequate sensitivity, and whether the threshold required to achieve good specificity suggests using a decision cut-off value that depends on age.

To address the scientific questions we conduct a case-control analysis. A case-control study for this application has several advantages. First, the sample size of the original CF data is considerable

and therefore a full cohort analysis would be computationally demanding. Second, the dataset is well suited for a case-control study since the outcome of interest, death, is rare and the exposure of interest is commonly available. Therefore, we should be able to estimate an effect comparable to that obtained from a full cohort analysis without much loss of statistical efficiency. Finally, as our analysis is essentially retrospective and only requires the measurement time (age) and the time of death we circumvent the issue of defining an appropriate time origin which would be required for proper prospective analysis.

In the CF sample we have 21,138 patients, of whom 1,496 died within the follow up period. For each of the observed cases, we assign two matched controls by selecting subjects who had been followed for more than 5 years and who are still alive at the end of the study. Furthermore, for the controls we exclude any observations that were obtained during the last five years before their final study visit so that all measurements included for analysis were taken at least 5 years before death. Thus the controls are defined here as those subjects who are 'known' to be healthy at the time FEV1 is measured; 'known' because death did not occur for at least 5 years. Controls are matched to cases by age at study entry. The resulting analysis data contain 1,496 cases and 2,992 control subjects with a total of 15,594 pulmonary function measurements.

The top panel of Figure 1 shows the association between FEV1 and subject age at measurement time. We use smoothed curves to describe FEV1 as function of age. As expected, the trends are for FEV1 to decrease with age for both cases and controls. This is consistent with the fact that decline of pulmonary function with increasing age is a recognized consequence of the disease process. The data suggest that the distribution of FEV1 is clearly different for cases as compared to controls, and this discrepancy may also vary with age. The association between FEV1 and time relative to death for the cases is displayed in the bottom panel of Figure 1. We see that FEV1 tends to be lower at times close to the time of death compared to times further from death.

In order to estimate ROC curves and to determine whether they vary with age and with time before death for cases, we fit a series of regression models. For comparison, models are fit both parametrically using the method described by Etzioni et al. (1999), and using a semi-parametric regression quantile method introduced in section 2.3. Although we focus on the current FEV1 value as the marker, other

derived outcome measures that are a scalar function of the current and/or past response process could be candidate markers. For example, the change in FEV1, $Y^*_{ik}(s) = Y_{ik}(s) - Y_{ik}(s-1)$, may represent a marker that reflects failing pulmonary function. However, recent epidemiologic work suggests that the current level of FEV1 rather than change in FEV1 is predictive of death (Liou, Adler, FitzSimmons, Cahill, Hibbs and Marshall 2001). Below we briefly summarize the models for FEV1.

4.2 **Parametric estimation of ROC curves**

For the parametric approach, we use a linear mixed model for FEV1. For cases, we consider two covariates for the kth measurement of subject i: age at which FEV1 is recorded ($s_{ik} = age_{ik}$), and years before death ($T_i - s_{ik} = yearsBD_{ik}$). Denote $\mathbf{Z}_{ik} = (s_{ik},\ T_i - s_{ik})^T = (age_{ik},\ yearsBD_{ik})^T$, the model takes the form

$$E[FEV1_{ik}|\mathbf{Z}_{ik}, b_i] = b_{0,i} + b_{1,i} \cdot age_{ik} + \beta_0^D + \beta_1^D \cdot age_{ik} + \beta_2^D \cdot yearsBD_{ik} + \beta_3^D \cdot age_{ik} \cdot yearsBD_{ik},$$

For the lth measurement of control subject j, denote $\mathbf{Z}_{jl} = (s_{jl}) = (age_{jl})$. We assume

$$E[FEV1_{jl}|\mathbf{Z}_{jl}, b_j] = b_{0,j} + b_{1,j} \cdot age_{jl} + \beta_0^{\bar{D}} + \beta_1^{\bar{D}} \cdot age_{jl} \ .$$

The models for both cases and controls assume a random intercept and a random slope for age. Maximum likelihood estimates for the linear mixed models provide estimates for the means (μ_D, and $\mu_{\bar{D}}$) and standard deviations (σ_D and $\sigma_{\bar{D}}$). Assuming normality for both the case and control populations leads to an induced ROC model:

$$ROC_{\mathbf{Z}}(p) = \Phi\left[\frac{\mu_{\bar{D}}(\mathbf{Z}_{\bar{D}}) - \mu_D(\mathbf{Z}_D)}{\sigma_D(\mathbf{Z}_D)} + \frac{\sigma_{\bar{D}}(\mathbf{Z}_{\bar{D}})}{\sigma_D(\mathbf{Z}_D)}\Phi^{-1}(p)\right],$$

with Φ denotes the cumulative standard normal distribution. Table 1 summarizes the parameter estimates for the linear mixed models.

Figure 2 displays the induced ROC curves at various times before death for FEV1 measured at age 10, 15 and 20. At all ages, we see that the discrimination is better when FEV1 is measured at times closer to death. For example, at age 15, with a threshold of 39.17 for defining "test positive" (90% of the controls are "test negative"), 60.3% of subjects who subsequently die at age 15 (the same year) are "test positive"; However, among those who subsequently die at age 20 (thus FEV1 is measured 5

years prior to death), only 30.6% are "test positive". Using this approach, there does not appear to be a trend in accuracy with age (Table 2). For example, if the false positive rate is controlled at 10% them when FEV1 is measured at the year of death (the time at which FEV1 is the most accurate), 57% of the children at age 10 are identified as "test positive", whereas for FEV1 measured at age 20, 59% of those patients who died at age 20 were identified as "test positive".

4.3 Semiparametric Regression Quantile Estimation of ROC curves

Separate models are now considered using the semiparametric regression quantile method to characterize the marker distributions for cases and controls. For controls, we model the mean function as

$$\mu_{\bar{D}}(Z_{\bar{D}}) = \beta_0^{\bar{D}} + (\boldsymbol{\beta}^{\bar{D}})^T \mathcal{B}(age)$$

and the logarithm of the standard deviation as

$$log\sigma_{\bar{D}}(Z_{\bar{D}}) = \gamma_0^{\bar{D}} + (\gamma^{\bar{D}})^T \mathcal{B}(age)$$

where $\mathcal{B}(age)$ is a natural spline basis with knots at 10 and 20 years. For the baseline distribution G_0 we chose either to use $\hat{G}_0(\epsilon_{\bar{D}})$, the empirical distribution function of the standardized residuals, or use the locally weighted empirical distribution function $\hat{G}_0(\epsilon_{\bar{D}}|age, a_n)$, where we consider distance based on age.

For cases, we model the mean function as

$$\mu_D(\mathbf{Z}_D) = \beta_0^D + (\boldsymbol{\beta}^D)^T [\mathcal{B}(age) \cdot \mathcal{B}(yearBD)]$$

and the logarithm of the standard deviation as

$$log\sigma_D(Z) = \gamma_0^D + (\gamma^D)^T [\mathcal{B}(age) + \mathcal{B}(yearBD)]$$

where $\mathcal{B}(age)$ is a natural spline basis with knots at 10 and 20, and $\mathcal{B}(yearBD)$ is a natural spline basis with knots at 2 and 4 years. For baseline distribution F_0 we chose either to use $\hat{F}_0(\epsilon_D)$, the empirical distribution function of the standardized residuals, or use the locally weighted empirical distribution function, where we consider distance based on either age only, $\hat{F}_0(\epsilon_D|age, a_n)$, or years before death only, $\hat{F}_0(\epsilon_D|yearBD, a_n)$. In our application, we use the Epanechnikov kernel $K(x) =$

$\frac{3}{4}(1-x^2)I\{|x| \leq 1\}$ and we select a bandwidth such that about 20% of the observations are included for estimation at each unique value of $\mathbf{Z}$. This bandwidth is selected based on a data-driven optimal bandwidth selection algorithm described in section 3.3, and the estimated optimal bandwidths do not appear to vary much when $0.75 \leq p \leq 0.95$. For example, the bandwidth estimates for the cases given by (3.1) are $\widehat{b}_{opt}(0.75) = 0.134$ and $\widehat{b}_{opt}(0.95) = 0.143$.

For controls, we observe a trend for the estimated mean FEV1 function $\hat{\mu}_{\bar{D}}$ to decrease with increasing age. The standard deviation function, however, does not appear to vary much between 10 to 30 years of age. We also examine plots (not shown here) of the baseline distribution function $\hat{G}_0(\epsilon|age)$, and the estimated percentiles of FEV1 as function of age based on the semiparametric quasi-likelihood method: $\hat{\mu}_{\bar{D}}(age)+\hat{\sigma}_{\bar{D}}(age)\hat{G}_0(\epsilon_{\bar{D}}|age)$. Both the baseline distribution and equivalently the estimated percentiles appear to depend on age.

For cases, we examine the estimated mean $\hat{\mu}_D$ and the estimated standard deviation $\hat{\sigma}_D$ as a function of year before death for various ages (Figure 3). We see that the mean function at age 10 is generally higher than that at age 20, similar to the trend observed for controls. For each age, FEV1 level appears to increase as the time relative to death increases from 0 to four years, then gradually levels off. The standard deviation function follows a similar pattern, but to a lesser extent. We also check plots (Figure 4) that show the baseline distribution function as a function of age, $\hat{F}_0\{\epsilon|age\}$, and as a function of years before death, $\hat{F}_0\{\epsilon|yearBD\}$. For both situations, higher percentiles appear to vary with covariates more than lower percentiles. This suggests that a variable baseline distribution estimator may be more appropriate than a constant baseline assumption for these data.

We compare the empirical distribution of the standardized residuals with a standard normal distribution, for cases, the empirical percentile is generally higher than the Gaussian percentile in the left part of the distribution (below 30%), but it tends to be lower than the Gaussian percentile between 30% and 90%. QQ-plot (not shown) also reveals right skewness. This implies that the normal assumption may not be plausible for these data. In contrast, for controls the empirical percentile is in good agreement with the Gaussian percentile, and the QQ-plot reveals little skewness.

ROC curves at various times before death for 15 year olds are displayed in Figure 5. We compare the ROC curves that assume F_0 does not depend on covariate (panel (a)), those that assume F_0

depends on age (panel (b)), and those that assume F_0 depends on the year before death (panel (c)). Similar to the findings with the parametric method, we observe better discrimination between cases and controls when FEV1 is measured at times closer to death across all ages and regardless of the specific assumptions on F_0. However, the ROC curves at 0, 1, 3, 5 years before death are considerably different depending on the assumption about F_0: the ROC curves at various times before death are relatively closer to each other if we use the empirical distribution of the standardized residual to estimate F_0, in contrast to the ROC curves obtained by letting F_0 vary with the time relative to death. This can also be seen from Table 3, which lists the estimated sensitivities when specificity equals 0.9. For example, at age 15, with an FEV1 value of 41 as the threshold for defining "test positive", 90% of the controls (those who lived at least beyond 20 years and are known to be alive by the end of the study) are "test negative". When F_0 is assumed to be constant, the fraction of subjects who are test positive are 74%, 66%, 49%, and 36% among those CF patients who die at age 15, 16, 18, and 20, respectively. However, if F_0 depends on the time relative to death, then with the same threshold, the corresponding true positive fractions become 80%, 69%, 50%, and 35%, which for 0 and 1 year prior to death are meaningfully different from the estimates obtained based on the other assumptions. Furthermore, comparing Table 3 with Table 2, we see there are also discrepancies between the estimates from the parametric method and those from the semi-parametric method. In order to estimate an ROC curve, we need to carefully characterize all components of the distribution, including the mean, the standard deviation, and the baseline distribution function. Misspecification of any of the three components may result in biased estimates of sensitivity and specificity.

The next logical step is to chose a model that best describes the baseline distribution. Developing a formal model selection procedure in the semi-parametric setting appears to be difficult. Instead, we suggest employing graphical summaries to assess whether the baseline distribution appears to vary with covariates. For example, we can utilize QQ-plots to compare the distribution of the standardized residuals across different values of a covariate. In Figure 6, we plot the quantiles of residuals at k years prior to death against the quantiles of residuals at $k+1, \ldots, 8$ years prior to death for cases. Many QQ-plots in the figure appear to be curved, indicating that the baseline distribution may not be constant over years prior to death. In contrast, when we examine the baseline distribution across

different ranges of age most of the QQ-plots are close to diagonal lines, and this is especially true for controls. Thus, the assumption that G_0 does not depend on age may be plausible for the data from controls.

We also fit the same models using a Gaussian kernel function instead of the Epanechnikov's kernel function to estimate the baseline distributions. The estimated sensitivities are similar to those in Table 3, indicating that the choice of the kernel function does not have substantial impact on the estimation.

5. Summary

In this article, we introduce an approach for constructing time-dependent ROC curves that is based on the semi-parametric regression quantile method for longitudinal data studied by Heagerty and Pepe (1999). We characterize the reference distribution of a key clinical measurement for healthy and diseased populations using a location-scale family. Quasi-likelihood methods are used to estimate conditional mean and standard deviation functions. The empirical distribution, or a weighted empirical distribution is used to characterize the shape of the marker distribion. We detail the asymptotic theory for ROC estimators under two situations: where the baseline distribution is constant; and where the baseline distribution is allowed to depend on covariates. For the second case we modify the theoretical results from the conditional empirical process literature for the independent situation (Stute 1984; Stute 1986) to account for the repeated measurements. Finally we use CF data to assemble a case-control study and build ROC curves that assess how well the distribution of FEV1 for cases at various times prior to death is separated from that of the controls. We compare the results from our new methodology with that of a parametric method (Etzioni et al. 1999). Our results indicate that specification of all three features of a distribution: mean; standard deviation; and baseline function; makes a significant impact on the resulting ROC estimates. Compared with the parametric approach, our method offers greater flexibility by having separate model choices for each of the key distributional aspects. For independent data a parametric approach can be made more flexible by adopting Box-Cox transformation methods for quantile estimation proposed by Cole and Green (1992). Heagerty and Pepe (1999) discuss use of these methods for longitudinal measurements. However, one of the main limitations of the Box-Cox approach is the necessary correct model

specification (i.e. normality) for consistency of estimates, while our semiparametric method provides generally consistent quantile estimates.

In our application, we have used kernel weight functions that require specification of a bandwidth. Further work is warranted exploring appropriate data-driven optimal bandwidth selection procedures tailored for ultimate ROC analysis. In addition, although we have detailed the large sample theory for semiparametric ROC estimators the performance of our theoretical results should be evaluated in small samples, and perhaps compared with alternative bootstrap inference methods. Finally, a methodological issue illustrated in the cystic fibrosis analysis is the potential sensitivity of ROC results to the selection of the baseline distribution model. Although we propose graphical methods to assess the appropriateness of model assumptions, more formal model comparison methods would be useful.

Appendix A. Large Sample Properties for Proposed Estimators

A.1 Proof of Theorem 1

Lemma 1 *Let* $\hat{F}_0(\epsilon) = \sum_{i=1}^{n_D} \sum_{k=1}^{K_i} \frac{1}{N_D} \mathbf{1}(\hat{\epsilon}_{ik} \leq \epsilon)$ *as in section 2.2, then*

$$P[\sup_x |\hat{F}_0(x) - F_0(x)| \to 0] = 1.$$

The lemma is a straightforward extension of the Glivenko-Cantelli Theorem to weakly dependent, identically distributed random variables. We omit the proof here.

Lemma 2 *Let* $\hat{G}_0(\epsilon) = \sum_{j=1}^{n_{\bar{D}}} \sum_{l=1}^{L_j} \frac{1}{N_{\bar{D}}} \mathbf{1}(\hat{\epsilon}_{jl} \leq \epsilon)$ *and* $\hat{G}_0^{-1}(p) = inf[\epsilon : \hat{G}_0(\epsilon) \geq p]$ *as in section 2.2.*

$$P[|\hat{G}_0^{-1}(p) - G_0^{-1}(p)| \to 0] = 1.$$

The lemma can be established the same way as for showing the convergence of quantile for the independent data (Shorack and Wellner 1986). For additional details regarding lemmas 1 and 2 please see Zheng (2002).

Proof. For $\hat{\alpha}_0$ and $\hat{\alpha}_1$ as defined in equation (2.12),

$$\begin{aligned}
& \sup_{0\leq p\leq 1} |\widehat{ROC}_{\mathbf{z}}(p) - ROC_{\mathbf{z}}(p)| \\
= & \sup_{0\leq p\leq 1} \left| \hat{F}_0 \left[\hat{\alpha}_0 + \hat{G}_0^{-1}(p)\hat{\alpha}_1 \right] - F_0 \left[\hat{\alpha}_0 + \hat{G}_0^{-1}(p)\hat{\alpha}_1 \right] + F_0 \left[\hat{\alpha}_0 + \hat{G}_0^{-1}(p)\hat{\alpha}_1 \right] - F_0 \left[\alpha_0 + G_0^{-1}(p)\alpha_1 \right] \right| \\
\leq & \sup_{0\leq p\leq 1} \left| \hat{F}_0 \left[\hat{\alpha}_0 + \hat{G}_0^{-1}(p)\hat{\alpha}_1 \right] - F_0 \left[\hat{\alpha}_0 + \hat{G}_0^{-1}(p)\hat{\alpha}_1 \right] \right| + \sup_{0\leq p\leq 1} \left| F_0 \left[\hat{\alpha}_0 + \hat{G}_0^{-1}(p)\hat{\alpha}_1 \right] - F_0 \left[\alpha_0 + G_0^{-1}(p)\alpha_1 \right] \right| \\
= & I_1 + I_2
\end{aligned}$$

for I_1, since $\hat{G}_0^{-1}(p) \to G^{-1}(p)$ by Lemma 2, $\hat{\alpha}_1 \to \alpha_1$, and $\hat{\alpha}_0 \to \alpha_0$, then by the Slutsky theorem $\hat{\alpha}_0 + \hat{G}_0^{-1}(p)\hat{\alpha}_1 \to \alpha_0 + G^{-1}(p)\alpha_1$ in probability. Following Lemma 1, we have $I_1 \to 0$ in probability. In addition, $I_2 \to 0$ in probability by the continuity of F_0. This proves the Theorem. □

A.2 **Proof of Theorem 2**

Proof.

$$\begin{aligned}
&\sqrt{n}[\widehat{ROC}_{\mathbf{Z}}(p) - ROC_{\mathbf{Z}}(p)] \\
= \;& \sqrt{n}\left\{\hat{F}_0\left[\hat{\alpha}_0 + \hat{G}_0^{-1}(p)\hat{\alpha}_1\right] - F_0\left[\hat{\alpha}_0 + \hat{G}_0^{-1}(p)\hat{\alpha}_1\right]\right\} + \sqrt{n}\left\{F_0\left[\hat{\alpha}_0 + \hat{G}_0^{-1}(p)\hat{\alpha}_1\right] - F_0\left[\alpha_0 + G_0^{-1}(p)\alpha_1\right]\right\} \\
= \;& W_1 + W_2,
\end{aligned}$$

we first approximate W_1 with a sum of i.i.d. terms: $\tilde{W}_1 = \frac{1}{\sqrt{\bar{\lambda}}c_D}\sqrt{n_D}\frac{1}{n_D}\sum_i^{n_D}\xi_{iD}$, with

$$\xi_{iD} = \sum_{k=1}^{K_i}\left\{\mathbf{1}[\epsilon_{ik} < \alpha_0 + G_0^{-1}(p)\alpha_1] - F_0[\alpha_0 + G_0^{-1}(p)\alpha_1]\right\}$$

By applying the Central limit theorem, we have for any $0 < p < 1$, $\tilde{W}_1 \to_D \frac{1}{\sqrt{\bar{\lambda}}c_D}U_1(p)$,where $U_1(p)$ is a zero-mean normal distribution whose variance σ_D^2 can be estimated by $\hat{\sigma}_D^2 = \frac{1}{n_D}\sum_i^{n_D}\hat{\xi}_{iD}^2$. $\hat{\xi}_{iD}^2$ are obtained from ξ_{iD} by replacing ϵ_{ik}, F_0, $G_0^{-1}(p)$, α_0 and α_1 with $\hat{\epsilon}_{ik}$, $\hat{F}_0$, $\hat{G}_0^{-1}(p)$, $\hat{\alpha}_0$ and $\hat{\alpha}_1$.

For W_2, we first take first order Taylor series expansion:

$$W_2 = \sqrt{n_D + n_{\bar{D}}}f_0\left[\alpha_0 + G_0^{-1}(p)\alpha_1\right]\left\{\hat{\alpha}_1\left[\hat{G}_0^{-1}(p) - G_0^{-1}(p)\right] + \left[\hat{\alpha}_0 + \hat{\alpha}_1 G_0^{-1}(p) - \alpha_0 - \alpha_1 G_0^{-1}(p)\right]\right\} + o_p(1)$$

Now, $\sqrt{n_{\bar{D}}}\{\hat{G}_0[G_0^{-1}(p)] - G_0[G_0^{-1}(p)]\} = \sqrt{n_{\bar{D}}}\frac{n_{\bar{D}}}{N_{\bar{D}}}\frac{1}{n_{\bar{D}}}\sum_j^{n_{\bar{D}}}\xi_{j\bar{D}}$ is a sum of i.i.d terms, with $\xi_{j\bar{D}} = \sum_{l=1}^{L_j}\left\{\mathbf{1}\left[\epsilon_{jl} < G_0^{-1}(p)\right] - p\right\}$, again by applying the Central limit theorem, we have for any $0 < p < 1$, $\sqrt{n_{\bar{D}}}\{\hat{G}_0[G_0^{-1}(p)] - G_0[G_0^{-1}(p)]\}$ converge to a zero-mean normal distribution $U_{2a}(p)$ whose variance can be estimated by $\hat{\sigma}_{\bar{D}}^2 = \sum_j^{n_{\bar{D}}}\hat{\xi}_{j\bar{D}}^2$. Let h be a mapping such that $h(y) = G_0^{-1}(y)$, and $\dot{h}^{-1} = g_0[G_0^{-1}(p)]$ is continuous. Following Cramér's theorem, we have

$$\sqrt{n_{\bar{D}}}\left[\hat{G}_0^{-1}(p) - G_0^{-1}(p)\right] \to_d \frac{1}{g_0[G_0^{-1}(p)]}U_{2a}(p)$$

In addition, denote V_D as the variance-covariance matrix for $\boldsymbol{\beta}_D$ (the parameters for μ_D) and $\boldsymbol{\gamma}_D$ (the parameters for σ_D). Similarly, denote $V_{\bar{D}}$ as the variance-covariance matrix for $\boldsymbol{\beta}_{\bar{D}}$ (the parameters for $\mu_{\bar{D}}$) and $\boldsymbol{\gamma}_{\bar{D}}$ (the parameters for $\sigma_{\bar{D}}$), $\boldsymbol{\theta} = (\boldsymbol{\beta}_D, \boldsymbol{\gamma}_D, \boldsymbol{\beta}_{\bar{D}}, \boldsymbol{\gamma}_{\bar{D}})$, we have

$$\sqrt{n}\begin{bmatrix} \hat{\boldsymbol{\beta}}_D - \boldsymbol{\beta}_D \\ \hat{\boldsymbol{\gamma}}_D - \boldsymbol{\gamma}_D \\ \hat{\boldsymbol{\beta}}_{\bar{D}} - \boldsymbol{\beta}_{\bar{D}} \\ \hat{\boldsymbol{\gamma}}_{\bar{D}} - \boldsymbol{\gamma}_{\bar{D}} \end{bmatrix} \to_d \mathcal{N}(0, \begin{bmatrix} \frac{1}{\lambda}V_D & 0 \\ 0 & \frac{1}{1-\lambda}V_{\bar{D}} \end{bmatrix}) \equiv \mathcal{N}(0, \boldsymbol{\Sigma}),$$

Let $g(\boldsymbol{\theta}) = \alpha_0 + G_0^{-1}(p)\alpha_1$, and let $g'(\boldsymbol{\theta})$ denote its derivative, by Cramér's device, we have

$$\sqrt{n_D + n_{\bar{D}}}\left[\hat{\alpha}_0 + \hat{\alpha}_1 G_0^{-1}(p) - \alpha_0 - \alpha_1 G_0^{-1}(p)\right] \rightarrow_D \mathcal{N}(0, g'(\boldsymbol{\theta})\Sigma g'(\boldsymbol{\theta})^T) \equiv U_{2b}(p)$$

Thus for any $0 < p < 1$, we have

$$\begin{aligned} W_2 \quad &\rightarrow_d \quad \frac{1}{c_{\bar{D}}\sqrt{1-\lambda}}\alpha_1 \frac{f_0[\alpha_0 + G_0^{-1}(p)\alpha_1]}{g_0[G^{-1}(p)]} U_{2a}(p) + f_0[\alpha_0 + G_0^{-1}(p)\alpha_1] U_{2b}(p) \\ &\equiv \quad \frac{\alpha_1}{c_{\bar{D}}\sqrt{1-\lambda}} \frac{f_0[\alpha_0 + G_0^{-1}(p)\alpha_1]}{g_0[G^{-1}(p)]} U_2(p) \end{aligned}$$

Since W_1 and W_2 are independent,

$$W_1 + W_2 \rightarrow_d \Psi(p) = \frac{1}{c_D\sqrt{\lambda}} U_1(p) + \frac{\alpha_1}{c_{\bar{D}}\sqrt{1-\lambda}} \frac{f_0[\alpha_0 + G_0^{-1}(p)\alpha_1]}{g_0[G^{-1}(p)]} U_2(p)$$

□.

A.3 **Proof of Theorem 3**

Without loss of generality, we now assume the covariate vector $\mathbf{Z}$ is a scalar in the proof. The following lemmas, whose proofs are omitted, follow from the results for conditional empirical processes and conditional quantile processes (Stute 1986).

Lemma 3 *Let* $\hat{F}_0(\epsilon|Z = z, a_n) = \sum_i^{n_D} \sum_k^{K_i} \frac{1}{W(z)} w_{a_n}(z, Z_{ik}) \mathbf{1}(\hat{\epsilon}_{ik} \leq \epsilon)$, *where* $W(z) = \sum w_{a_n}(z, Z_{ik})$ *with a general form of* w_{a_n}*:*

$$w_{a_n}(z, Z_{ik}) = K\left[\frac{H_n(z) - H_n(Z_{ik})}{a_n}\right]$$

we have

$$P\left[\sup_\epsilon |\hat{F}_0(\epsilon|z, a_n) - F_0(\epsilon|z, a_n)| \rightarrow 0\right] = 1.$$

Lemma 4 *Let* $\hat{G}_0(\epsilon|Z = z, a_n) = \sum_j^{n_{\bar{D}}} \sum_l^{L_j} \frac{1}{W(z)} w_{a_n}(z, Z_{jl}) \mathbf{1}(\hat{\epsilon}_{jl} \leq \epsilon)$ *and* $\hat{G}_{0,z}^{-1}(p) = inf[\epsilon : \hat{G}_0(\epsilon|Z = z, a_n) \geq p]$, *then*

$$P\left[\sup_z |\hat{G}_{0,z}^{-1}(p) - G_{0,z}^{-1}(p)| \rightarrow 0\right] = 1.$$

Proof of Theorem 3.

The theorem is easily shown in view of lemma 3 and lemma 4.

A.4 **Proof of Theorem 4**

Lemma 5 *Under the assumptions*

A. $a_n \to 0$ *such that* $na_n^3 \to \infty$ *and* $na_n^5 \to 0$,

B. *there exists a distribution function of* (Z, ϵ), *say, M, with uniform marginals, and*

C. $\sup_{\|t-s\|<\delta} |F(t|z) - F(s|z)| = o[(ln\delta^{-1})^{-1}]$ *as* $\delta \to 0$, *uniformly in a neighborhood of* z,

we have for $\bar{F}_0(\epsilon|Z = z, a_n) = a_n^{-1} \int \mathbf{1}(u \le \epsilon) K\left[\frac{H(z)-H(v)}{a_n}\right] M(dv, du)$,

$$\sqrt{n_D a_n}[\hat{F}_0(\epsilon|Z = z, a_n) - \bar{F}_0(\epsilon|Z = z, a_n)] \to_d \mathcal{N}(0, \sigma_D^2)$$

for μ*-almost all* $\epsilon \in \mathcal{R}$, *where* M *is a distributional function of* (Z, ϵ) *and* σ_D^2 *can be consistently estimated by*

$$\begin{aligned}\hat{\sigma}_D^2 &= \frac{1}{n_D}\sum_{i=1}^{n_D}\sum_{k=1}^{K_i}\sum_{l=1}^{K_i}\frac{1}{a_n c_D^2}\left\{I(\epsilon_{ik} \le \epsilon)K\left[\frac{H_n(Z) - H_n(Z_{ik})}{a_n}\right]\right. \\ &\quad \left. I(\epsilon_{il} \le \epsilon)K\left[\frac{H_n(Z) - H_n(Z_{il})}{a_n}\right]\right\}\end{aligned}$$

Remarks: The assumptions are the same as for the independent case given by Stute (1986). For assumption B, Let H denote the distribution function of Z, and L be the distribution function of ϵ, we have

$$M(Z, \epsilon) = C[H(Z), L(\epsilon)]$$

i.e., M, a distribution function on $[0, 1]^2$ with uniform marginals, can be obtained by finding a transformation function C on $[H(Z), L(\epsilon)]$. Thus we reduce our investigation to some uniform random function.

Assumption C is satisfied whenever F_0 is continuous of some order. It entails the equicontinuity of F_0 in a neighborhood of z.

Proof.

The proof for independent observations is given by Stute (1986). We now extend his results to the correlated data situation to show that $\beta_n = \sqrt{n_D a_n}[\hat{F}_0(\epsilon|Z = z, a_n) - \bar{F}_0(\epsilon|Z = z, a_n)] \to_d \mathcal{N}(0, \sigma_D^2)$.

Let

$$\hat{F}_0^*(\epsilon|Z=z,a_n) = a_n^{-1}\sum_i^{n_D}\sum_k^{K_i} w_{a_n}(z,Z_{ik})\mathbf{1}(\hat{\epsilon}_{ik}\le\epsilon).$$

now, $\hat{F}_0^*(\epsilon|Z=z,a_n) = \hat{F}_0(\epsilon|Z=z,a_n)(N_D a_n)^{-1}W(\mathbf{Z}) \equiv \hat{F}_0(\epsilon|Z=z,a_n)f_n(z)$, and Stute (1986) shows that $(n_D a_n)^{1/2}[f_n(z)-1] \to_p 0$, and thus under the smoothness assumptions

$$\beta_n = \sqrt{n_D a_n}[\hat{F}_0^*(\epsilon|Z=z,a_n) - \bar{F}_0(\epsilon|Z=z,a_n)] + o_p(1) \text{ uniformly in } \epsilon.$$

In what follows we only work with $\hat{F}_0^*(\epsilon|Z=z,a_n)$, of which the asymptotic properties are the same as those of $\hat{F}_0(\epsilon|Z=z,a_n)$.

Let M_n denote the bivariate empirical d.f. of the sample $(Z_1,\epsilon_1),\cdots,(Z_{N_D},\epsilon_{N_D})$, we can write $\hat{F}_0^*(\epsilon|Z=z,a_n)$ in the form of $\hat{F}_0^*(\epsilon|Z=z,a_n) = a_n^{-1}\int \mathbf{1}(u\le\epsilon)K\left[\frac{H_n(z)-H_n(v)}{a_n}\right]M_n(dv,du)$. Because of the heavy dependence of the weights for summands, we can not apply a central limit theorem to n_D dependent r.v.s directly. So the first goal of the proof is to approximate the SNN estimator with a quantity that is the sum of n_D independent random variables. The asymptotic approximation is essentially the same as the method used when the ϵ_i's are independent, details of which were given by Stute (1984) and Stute (1986). In the following, we only outline the approximation procedure, omitting further details that can be found in Stute (1984).

Assuming K is twice differentiable, we start by applying Taylor expansion to $\hat{F}_0^*(\epsilon|Z=z,a_n)$:

$$\begin{aligned}
\hat{F}_0^*(\epsilon|Z=z,a_n) &= a_n^{-1}\int \mathbf{1}(u\le\epsilon)K\left[\frac{H(z)-H(v)}{a_n}\right]M_n(dv,du)\\
&+ a_n^{-2}[H_n(z)-H_n(v)-H(z)+H(v)]\int \mathbf{1}(u\le\epsilon)K'\left[\frac{H_n(z)-H_n(v)}{a_n}\right]M_n(dv,du)\\
&+ a_n^{-3}\int \mathbf{1}(u\le\epsilon)[H_n(z)-H_n(v)-H(z)+H(v)]^2K''(\Delta)M_n(dv,du)/2\\
&\equiv I_1+I_2+I_3,
\end{aligned}$$

where Δ is on the line segment between $a_n^{-1}[H_n(z)-H_n(v)]$ and $a_n^{-1}[H(z)-H(v)]$. Follows from Lemma 1 in Stute (1984), we have $(n_D a_n)^{1/2}I_3 \to_p 0$ as $n_D\to\infty$. Furthermore, Stute (1984) yields that $(na_n)^{1/2}I_2$ is asymptotically equivalent to

$$-n_D^{1/2}a_n^{-1/2}F_0(\epsilon|Z=z)\int K\left[\frac{H(z)-H(v)}{a_n}\right]n_D^{1/2}[H_n(dv)-H(dv)].$$

thus,

$$\begin{aligned}
\beta_n &= \sqrt{n_D a_n} a_n^{-1} \int \mathbf{1}(u \leq \epsilon) K\left[\frac{H(z) - H(v)}{a_n}\right] M_n(dv, du) \\
&- n_D^{1/2} a_n^{-1/2} F_0(\epsilon | Z = z) \int K\left[\frac{H(z) - H(v)}{a_n}\right] n_D^{1/2} [H_n(dv) - H(dv)] \\
&- \sqrt{n_D a_n} a_n^{-1} \int \mathbf{1}(u \leq \epsilon) K\left[\frac{H(z) - H(v)}{a_n}\right] M(dv, du) \\
&= (\frac{n_D}{a_n})^{1/2} \int [\mathbf{1}(u \leq \epsilon) - F_0(\epsilon | Z = z)] K\left[\frac{H(z) - H(v)}{a_n}\right] [M_n(dv, du) - M(dv, du)]
\end{aligned}$$

Furthermore, $(\frac{n_D}{a_n})^{1/2} \int [\mathbf{1}(u \leq \epsilon) - F_0(\epsilon | Z = z)] K\left[\frac{H(z) - H(v)}{a_n}\right] M(dv, du)$ is asymptotically negligible. Thus, asymptotically, $\beta_n = n_D^{1/2} \sum_{i=1}^{n_D} \xi_i$ with

$$\xi_i = \sum_{k=1}^{K_i} \xi_{ik} = \sum_{k=1}^{K_i} \int \frac{1}{\sqrt{a_n}} [\mathbf{1}(u \leq \epsilon) - F_0(\epsilon | Z = z)] K\left[\frac{H(z) - H(v)}{a_n}\right] M_{ik}(dv, du)$$

which is a standardized sum of n_D $i.i.d$ random variables. To apply the Central Limit Theorem, we need to check that $E(\xi_i^2) < \infty$. Now, $E(\xi_i^2) = E\left(\sum_{k=1}^{K_i} \sum_{l=1}^{K_i} \xi_{ik} \xi_{il}\right) \leq \sum_{k=1}^{K_i} \sum_{l=1}^{K_i} \left(E\xi_{ik}^2 E\xi_{il}^2\right)^{1/2}$, with

$$\begin{aligned}
E\xi_{ik}^2 &= a_n^{-1} \int [\mathbf{1}(u \leq \epsilon) - F_0(\epsilon | Z = z)]^2 K^2\left[\frac{H(z) - H(v)}{a_n}\right] M(dv, du) \\
&= a_n^{-1} \int E\left\{[\mathbf{1}(u \leq \epsilon) - F_0(\epsilon | Z = z)]^2 | Z = v\right\} K^2\left[\frac{H(z) - H(v)}{a_n}\right] H(dv) \\
&\to h(z) \int K^2(s) ds < \infty.
\end{aligned}$$

Hence $E(\xi_i^2) < \infty$ since K_i is small relative to n. It then follows from the Central Limit Theorem that $\beta_n \to_d \mathcal{N}(0, \sigma_D^2)$. A consistent estimator for σ_D^2 is $\hat{\sigma}_D^2 = \frac{1}{n_D} \sum_{i=1}^{n_D} \sum_{k=1}^{K_i} \sum_{l=1}^{K_i} \hat{\xi}_{ik} \hat{\xi}_{il}$.

$\hat{\xi}_{ik}$ can be obtained by substituting the theoretical terms with their empirical counterparts.

Follows Corollary 2 of Stute (1986), we have for μ-almost all $0 < z < 1$, when $n a_n^5 \to 0$, and under the additional assumption that for each ϵ $F_0(\epsilon | \cdot)$ is twice continuously differentiable in a neighborhood of z,

$$(n_D a_n)^{1/2} [\hat{F}_0(\epsilon | Z = z, a_n) - F_0(\epsilon | Z = z, a_n)] \to_d N(0, \sigma_D^2)$$

Remark when we select optimal bandwidth with a_n^{opt}, $na_n^5 \to c$, $c > 0$, the limit process of the conditional empirical is a noncentered Gaussian process, with some 'bias' term of the form

$$\frac{\sqrt{c}}{2} F''(\epsilon|z) \int u^2 K(u) du$$

Lemma 6 (Asymptotic normality of the conditional quantile process) *Define a conditional quantile function as*

$$\hat{G}_0^{-1}(p|z) = inf\{\epsilon \in \mathcal{R} : \hat{G}_0(\epsilon|z) \geq p\},$$

which is an estimator for the p quantile of $G_0(\cdot|z)$ for $0 < p < 1$. for such p, write $\epsilon_p = G_0^{-1}(p|z)$. Under the above assumptions, if $g_0(\epsilon_p|z) = (\partial/\partial\epsilon)G_0(\epsilon|z) > 0$ at $\epsilon = \epsilon_p$ and F is continuous we have for almost all z

$$(n_{\bar{D}} a_n)^{1/2}[\hat{G}_0^{-1}(p|Z = z, a_n) - G_0^{-1}(p|Z = z, a_n)] \to_d \frac{1}{g_0(\epsilon_p|z)} \mathcal{N}(0, \sigma_{\bar{D}}^2)$$

where

$$\sigma_{\bar{D}}^2 = \frac{1}{n_{\bar{D}}} \sum_{j=1}^{n_{\bar{D}}} \sum_{l=1}^{L_j} \sum_{m}^{L_j} \xi_{jl} \xi_{jm}$$

with

$$\xi_{jl} = \frac{1}{\sqrt{a_n} c_{\bar{D}}} \left[\mathbf{1}(\epsilon_{jl} \leq \epsilon_p) - G_0(\epsilon_p|Z = z) \right] K \left[\frac{H(z) - H(z_{jl})}{a_n} \right]$$

We omit the proof here.

Proof of Theorem 4

$$\begin{aligned}
& \sqrt{na_n}[\widehat{ROC}_z(p) - ROC_z(p)] \\
= \; & \sqrt{na_n} \left\{ \hat{F}_{0,\mathbf{z}} \left[\hat{\alpha}_0 + \hat{G}_{0,\mathbf{z}}^{-1}(p)\hat{\alpha}_1 \right] - F_{0,\mathbf{z}} \left[\hat{\alpha}_0 + \hat{G}_{0,\mathbf{z}}^{-1}(p)\hat{\alpha}_1 \right] \right\} \\
+ \; & \sqrt{na_n} \left\{ F_{0,\mathbf{z}} \left[\hat{\alpha}_0 + \hat{G}_{0,\mathbf{z}}^{-1}(p)\hat{\alpha}_1 \right] - F_{0,\mathbf{z}} \left[\alpha_0 + G_{0,\mathbf{z}}^{-1}(p)\alpha_1 \right] \right\} \\
= \; & W_1 + W_2
\end{aligned}$$

Now, by lemma 5, it can be shown that $W_1 \to_d \frac{1}{c_D \sqrt{\lambda}} U_1^\star(p)$. Here $U_1^\star(p)$ is a zero-mean normal distribution with variance σ_D^2 for any $0 < p < 1$. A consistent estimator of σ_D^2 is

$$\hat{\sigma}_D^2 = \frac{1}{n_D} \sum_{i=1}^{n_D} \sum_{k=1}^{K_i} \sum_{l=1}^{K_i} \hat{\xi}_{ik} \hat{\xi}_{il}$$

$\hat{\xi}_{ik}$ can be obtained by substituting the theoretical terms with their empirical counterparts, i.e.,

$$\hat{\xi}_{ik} = \frac{1}{\sqrt{a_n}} \left\{ \mathbf{1}[\hat{\epsilon}_{ik} \leq \hat{\alpha}_0 + \hat{G}_0^{-1}(p)\hat{\alpha}_1] - \hat{F}_0[\hat{\alpha}_0 + \hat{G}_0^{-1}(p)\hat{\alpha}_1 | Z = z] \right\} K \left[\frac{H_n(z) - H_n(z_{ik})}{a_n} \right]$$

By the fact that $a_n \to 0$ and lemma 6, it is easy to show that for any $0 < p < 1$, we have $W_2 \to_d \frac{1}{c_{\bar{D}}\sqrt{1-\lambda}} \alpha_1 \frac{f_0[\alpha_0 + G_0^{-1}(p)\alpha_1]}{g_{0,z}(\epsilon_p)} U_2^\star(p)$, where $U_2^\star(p)$ is a zero-mean normal distribution with variance $\sigma_{\bar{D}}^2$ for any $0 < p < 1$. A consistent estimator of $\sigma_{\bar{D}}^2$ is

$$\hat{\sigma}_{\bar{D}}^2 = \frac{1}{n_{\bar{D}}} \sum_{j=1}^{n_{\bar{D}}} \sum_{l=1}^{L_j} \sum_{m=1}^{L_j} \hat{\xi}_{jl} \hat{\xi}_{jm}$$

$\hat{\xi}_{jl}$ can be obtained by substituting the theoretical terms with their empirical counterparts.

Since W_1 and W_2 are independent,

$$W_1 + W_2 \to_d \Psi(p) \equiv \frac{1}{c_D\sqrt{\lambda}} U_1^\star(p) + \frac{1}{c_{\bar{D}}\sqrt{1-\lambda}} \alpha_1 \frac{f_{0,z}[\alpha_0 + G_{0,\mathbf{z}}^{-1}(p)\alpha_1]}{g_{0,z}(\epsilon_p)} U_2^\star(p)$$

□.

REFERENCES

Cai, T., and Pepe, M. S. (2002), "Semiparametric receiver operating characteristic analysis to evaluate biomarkers for disease," *Journal of the American Statistical Association*, 97(460), 1099–1107.

Cole, T. J., and Green, P. J. (1992), "Smoothing reference centile curves: The LMS method and penalized likelihood," *Statistics in Medicine*, 11, 1305–1319.

Davis, P. B. (1997), "The decline and fall of pulmonary function in cystic fibrosis: New models, new lessons," *The Journal of Pediatrics*, 131, 789–790.

Ducharme, G. R., Gannoun, A., Guertin, M.-C., and Jéquier, J.-C. (1995), "Reference values obtained by kernel-based estimation of quantile regressions," *Biometrics*, 51, 1105–1116.

Etzioni, R., Pepe, M., Longton, G., Hu, C., and Goodman, G. (1999), "Incorporating the time dimension in receiver operating characteristic curves: A case study of prostate cancer," *Medical Decision Making*, 19, 242–251.

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

Heagerty, P. J., Lumley, T., and Pepe, M. S. (2000), "Time-dependent ROC curves for censored survival data and a diagnostic marker," *Biometrics*, 56(2), 337–344.

Heagerty, P. J., and Pepe, M. S. (1999), "Semiparametric estimation of regression quantiles with application to standardizing weight for height and age in US children," *Applied Statistics*, 48, 533–551.

Hsieh, F., and Turnbull, B. W. (1996), "Nonparametric and semiparametric estimation of the receiver operating characteristic curve," *The Annals of Statistics*, 24, 25–40.

Li, G., Tiwari, R. C., and Wells, M. T. (1999), "Semiparametric inference for a quantile comparison function with applications to receiver operating characteristic curves," *Biometrika*, 86, 487–502.

Liou, T., Adler, F., FitzSimmons, S., Cahill, B., Hibbs, J., and Marshall, B. (2001), "Predictive 5-year survivorship model of cystic fibrosis," *American Journal of Epidemiology*, 153, 245–352.

Pepe, M. S. (1998), "Three approaches to regression analysis of receiver operating characteristic curves for continuous test results," *Biometrics*, 54, 124–135.

Pepe, M. S. (2003), *The Statistical Evaluation of Medical Tests for Classification and Prediction* Oxford University Press.

Pepe, M. S., Etzioni, R., Feng, Z., Potter, J., Thompson, M. L., Thornquist, M., Winget, M., and Yasui, Y. (2001), "Phases of biomarker development for early detection of cancer," *Journal of the National Cancer Institute*, 93(14), 1054–1061.

Shorack, G. R., and Wellner, J. A. (1986), *Empirical processes with applications to statistics* John Wiley & Sons.

Slate, E. H., and Turnbull, B. W. (2000), "Statistical models for longitudinal biomarkers of disease onset," *Statistics in Medicine*, 19(4), 617–637.

Stute, W. (1984), "Asymptotic normality of nearest neighbor regression function estimates," *The Annals of Statistics*, 12, 917–926.

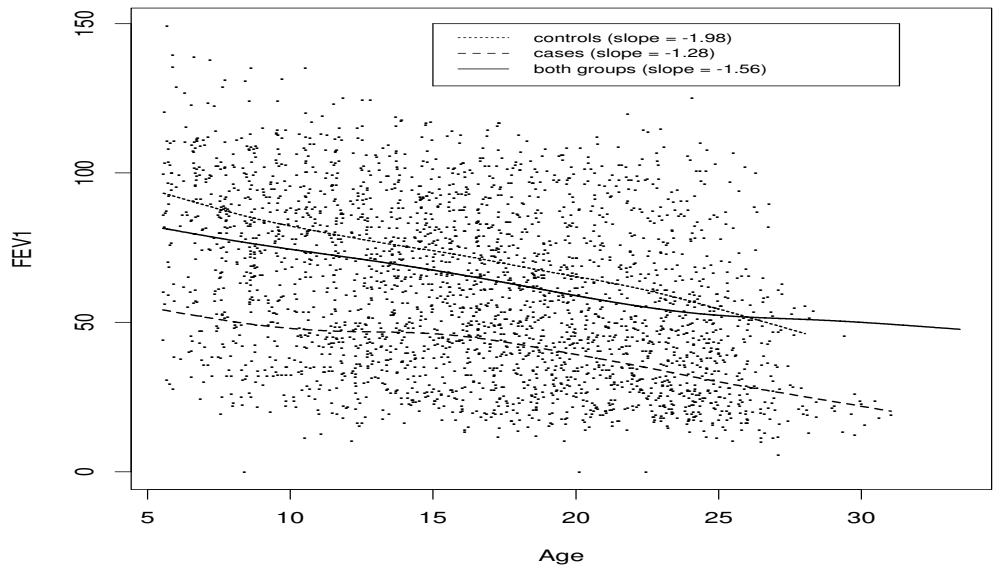
Stute, W. (1986), "Conditional empirical processes," *The Annals of Statistics*, 14, 638–647.

Tosteson, A. N., and Begg, C. B. (1988), "A general regression methodology for ROC curve estimation," *Medical Decision Making*, 8, 204–215.

Yang, S.-S. (1981), "Linear combination of concomitants of order statistics with application to testing and estimation," *Annals of the Institute of Statistical Mathematics*, 33, 463–470.

Zheng, Y. (2002), *Semiparametric Methods for Longitudinal Diagnostic Accuracy* PhD Thesis. Univerisity of Washington, Seattle.

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



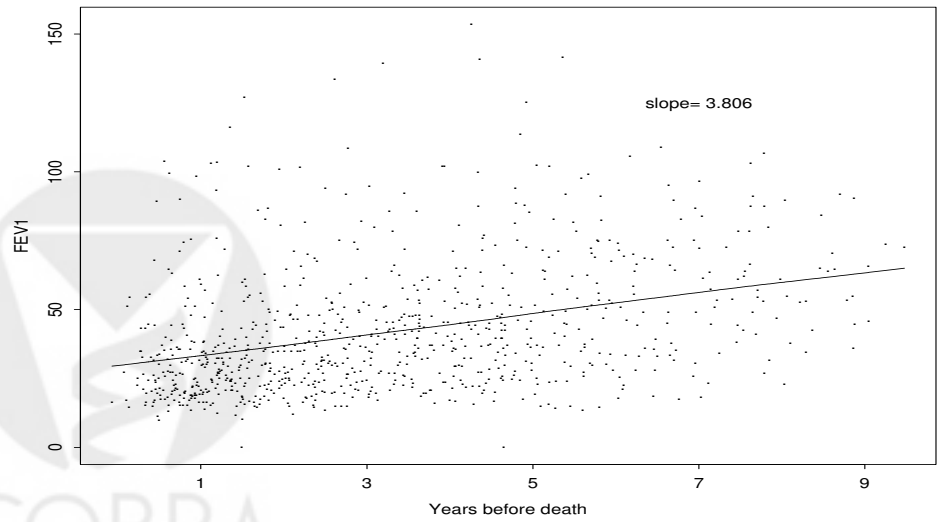


Figure 1: (a) FEV1 versus age at measurement for CF patients. Separate lines are fitted for controls, cases and all patients with smoothing splines. (b)FEV1 versus time relative to death for CF patients.

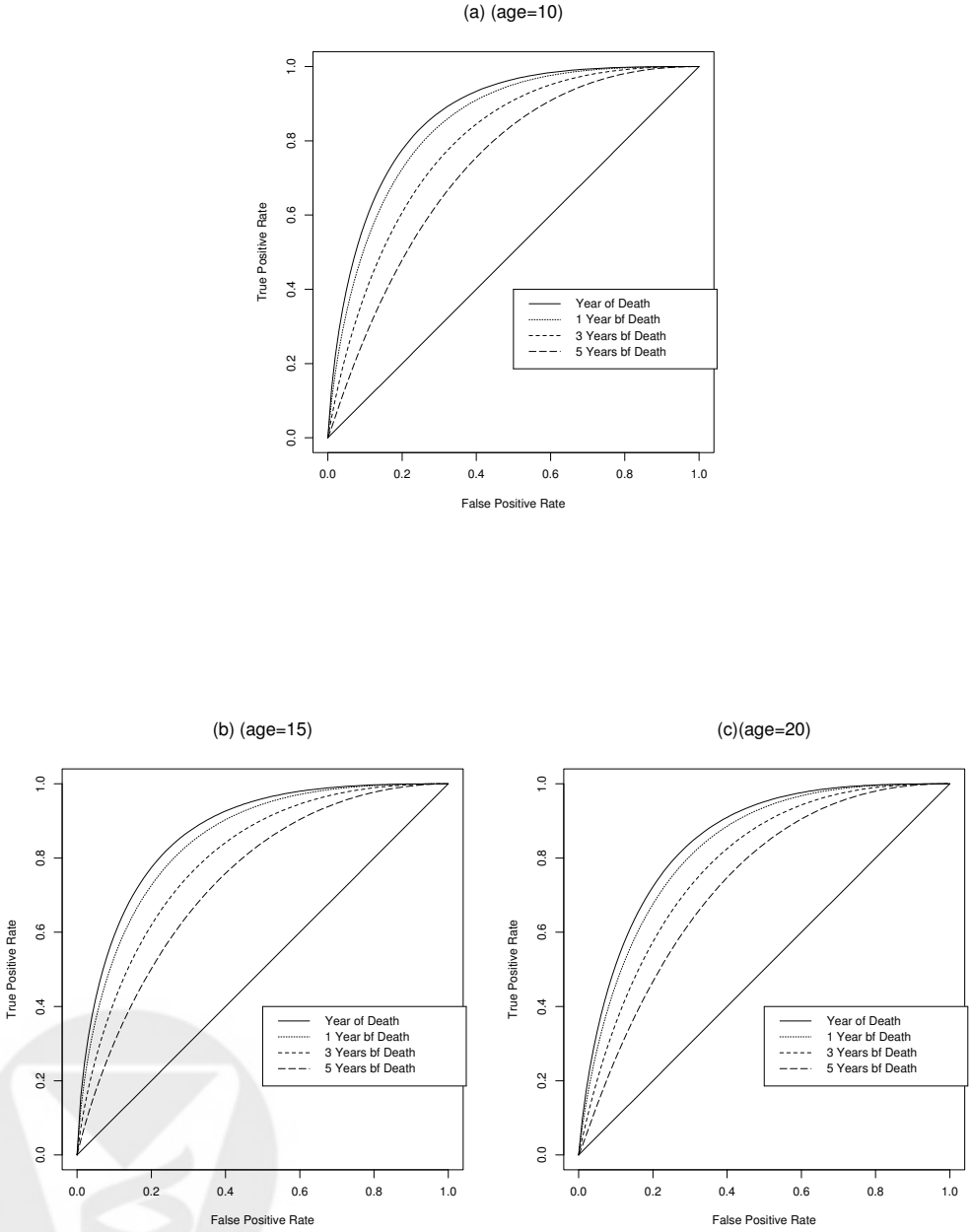


Figure 2: ROC curves for FEV1 measured at age 10, 15, and 20 at 0, 1, 3 and 5 years prior to death using parametric RQ method. Panel(a)-(c) show the ROC curves at different ages. The diagonal line in each plot is included for reference.

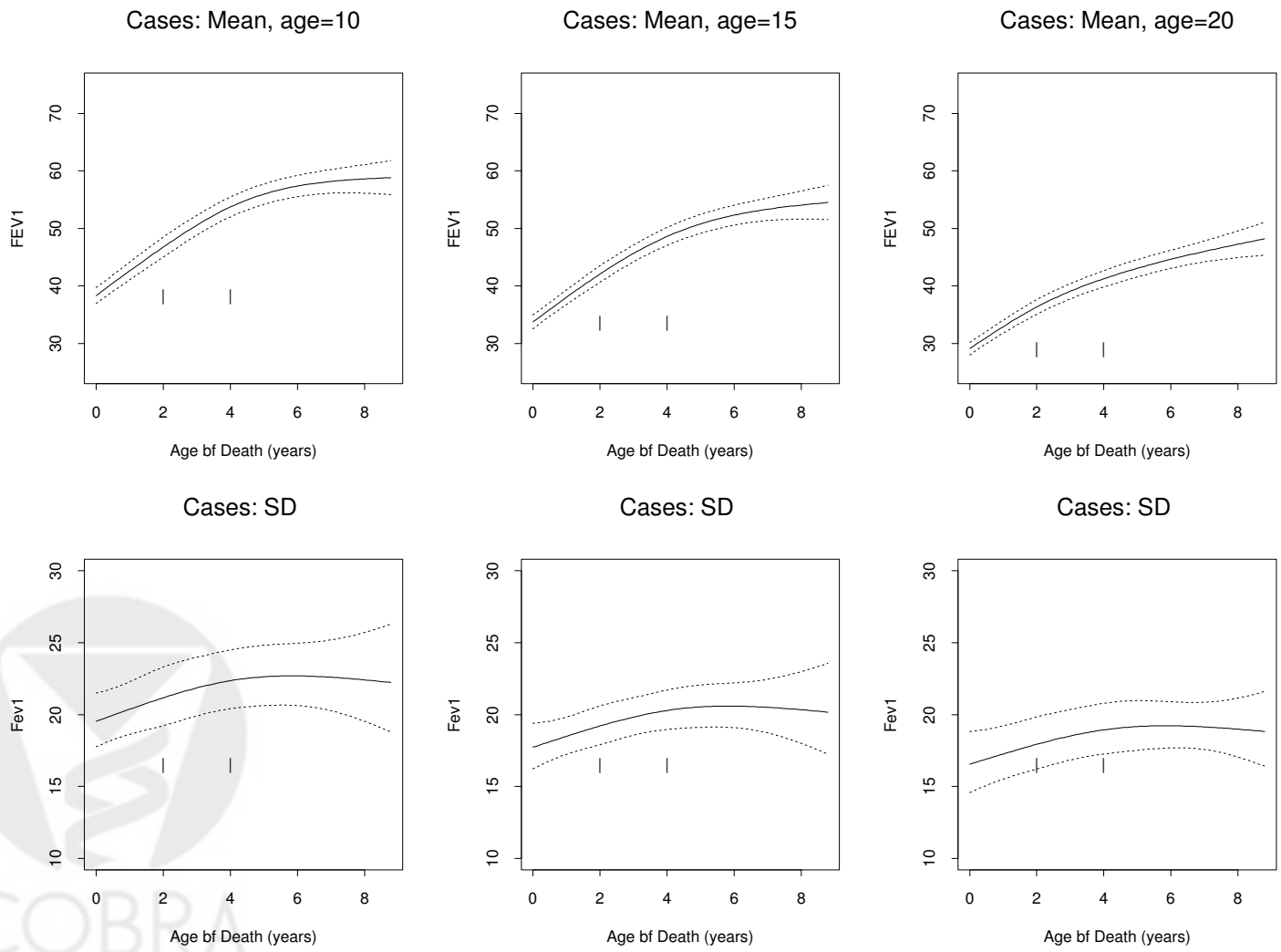
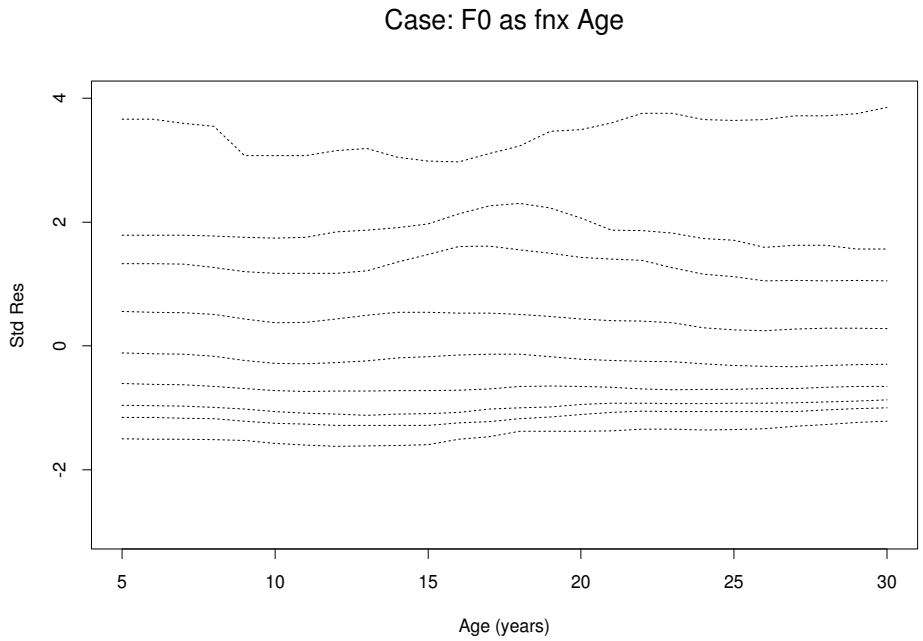


Figure 3: Mean (top panels) and standard deviation (bottom panels) of FEV1 as a function of years before death for cases at various ages: the functions are estimated with natural regression splines with knots placed at locations denoted by the vertical tick marks. Dotted lines are the pointwise 95% confidence intervals.



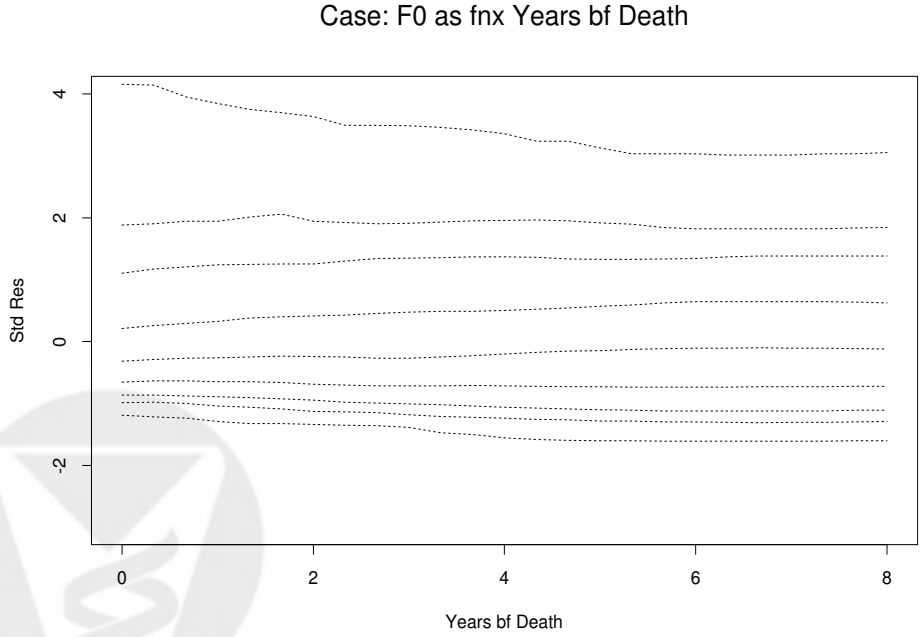


Figure 4: (a) Estimated quantiles of F_0 based on the kernel estimation method. F_0 is allowed to depend on age.(b) Estimated quantiles of F_0 based on the kernel estimation method. F_0 is allowed to depend on years before death.

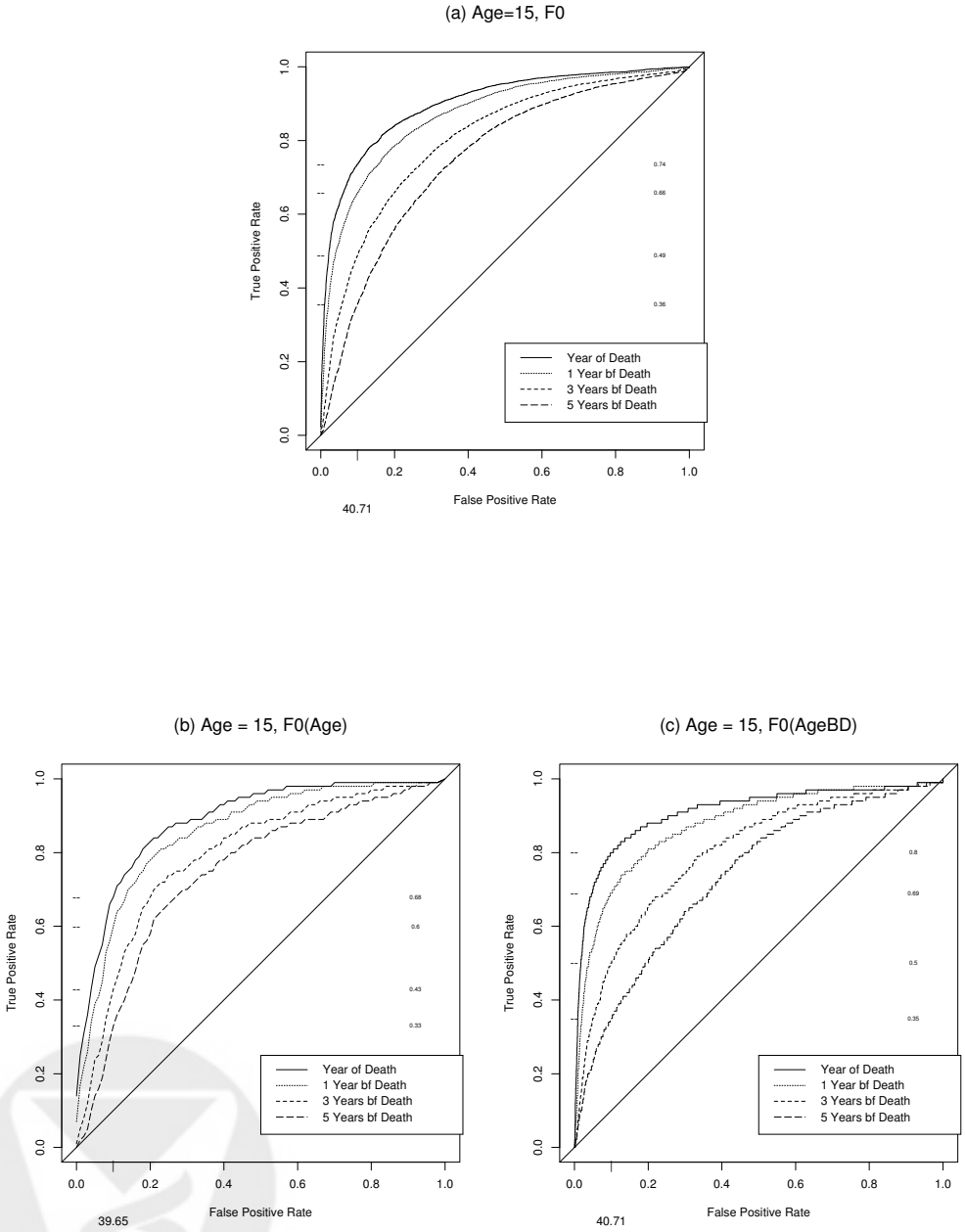


Figure 5: ROC curves for FEV1 measured at age 15 at 0, 1, 3 and 5 years prior to death using semiparametric RQ method. Panel (a) shows the ROC curves assuming F_0 does not depend on covariate. Panel (b) shows the ROC curves assuming F_0 depends on age. Panel (c) shows the ROC curves assuming F_0 depends on years before death. The diagonal line in each plot is included for reference.

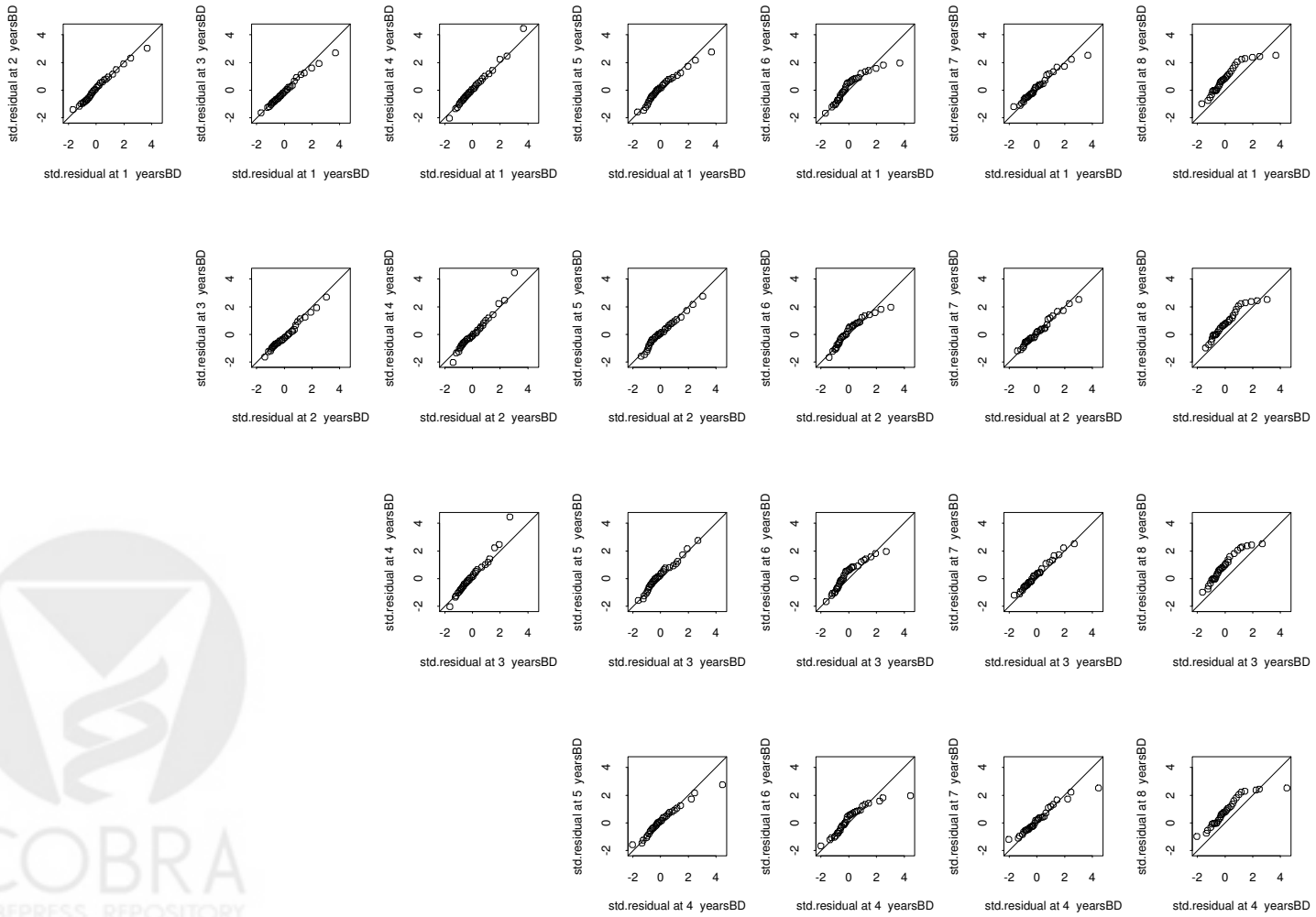


Figure 6: QQ-plots of standardized residuals $\epsilon_D(\mathbf{Z}_{ik}) = [Y_{ik} - \mu_D(\mathbf{Z}_{ik})]/\sigma_D(\mathbf{Z}_{ik})$ at different times prior to death for cases.

Table 1: Estimation based on linear mixed models for CF data.

	Fixed effect				Random effect		
	$\hat{\beta}_0$ (SE)	$\hat{\beta}_1$ (SE)	$\hat{\beta}_2$(SE)	$\hat{\beta}_3$(SE)	$\hat{\sigma}_{int}$	$\hat{\sigma}_{age}$	$\hat{\sigma}$ †
Control	92.41 (1.46)	-1.34(0.06)			38.17	1.34	8.68
Case	43.55 (1.44)	-0.62 (0.06)	4.20(0.25)	-0.09 (0.01)	25.19	0.93	12.48

†: $\hat{\sigma}$ is the estimated standard deviation for the residuals.

Table 2: Estimated sensitivities at specificity = 0.9 based on linear mixed models

		years relative to death			
age	threshold	0	1	3	5
age = 10	41.29	0.573	0.511	0.389	0.276
age = 15	39.17	0.603	0.542	0.421	0.306
age = 20	35.25	0.594	0.538	0.424	0.317

Table 3: Estimated sensitivities with 95% confidence interval at specificity = 0.9 based on the semi-parametric regression quantile method with Epanechnikov's kernel.

		years relative to death			
	threshold	0	1	3	5
age =10					
F_0 †	53.33	0.83 (0.81-0.85)	0.77(0.75-0.79)	0.65(0.62-0.68)	0.56(0.53-0.58)
$F_0(age)$ ‡	55.04	0.86(0.80-0.92)	0.81(0.75-0.87)	0.70(0.63-0.77)	0.62(0.52-0.72)
$F_0(ageBD)$ † ‡	53.33	0.87(0.84-0.90)	0.79(0.77-0.81)	0.64(0.61-0.67)	0.50(0.46-0.54)
age =15					
F_0	40.71	0.74(0.71-0.76)	0.66(0.63-0.68)	0.49(0.46-0.52)	0.36(0.33-0.38)
$F_0(age)$	39.65	0.68(0.60-0.76)	0.60(0.53-0.67)	0.43(0.35-0.51)	0.33(0.25-0.41)
$F_0(ageBD)$	40.71	0.80(0.76-0.84)	0.69(0.66-0.72)	0.50(0.46-0.54)	0.35(0.30-0.40)
age =20					
F_0	30.69	0.64(0.62-0.67)	0.55(0.52-0.58)	0.38(0.35-0.41)	0.27(0.24-0.30)
$F_0(age)$	27.17	0.55(0.46-0.64)	0.42(0.34-0.50)	0.24(0.16-0.32)	0.15(0.10-0.20)
$F_0(ageBD)$	30.69	0.72(0.68-0.76)	0.57(0.52-0.62)	0.39(0.33-0.45)	0.29(0.24-0.34)

†: F_0 does not depend on covariates.
‡: F_0 depends on age. Estimated with Epanechnikov's kernel.
† ‡ : F_0 depends on years before death. Estimated with Epanechnikov's kernel.