UCLA
Research Reports

Title

Time-Varying Effect Modeling with Longitudinal Data Truncated by Death: Conditional Models, Interpretations and Inference

Permalink

https://escholarship.org/uc/item/1zz0p2d2

Authors

Estes, Jason P.
Nguyen, Danh V.
Dalrymple, Lorien S.
et al.

Publication Date

2015-01-27

Peer reviewed

eScholarship.org Powered by the California Digital Library
University of California

Time-Varying Effect Modeling with Longitudinal Data Truncated by Death: Conditional Models, Interpretations and Inference

Jason P. Estes, Danh V. Nguyen, Lorien S. Dalrymple, Yi Mu and Damla Şentürk

Abstract

Recent studies found that infection-related hospitalization was associated with increased risk of cardiovascular (CV) events, such as myocardial infarction and stroke in the dialysis population. In this work, we develop time-varying effects modeling tools in order to examine the CV outcome risk trajectories during the time periods before and after an initial infection-related hospitalization. For this, we propose partly conditional and fully conditional partially linear generalized varying coefficient models (PL-GVCMs) for modeling time-varying effects in longitudinal data with substantial follow-up truncation by death. Unconditional models that implicitly target an immortal population is not a relevant target of inference in applications involving a population with high mortality, like the dialysis population. A partly conditional model characterizes the outcome trajectory for the dynamic cohort of survivors, where each point in the longitudinal trajectory represents a snapshot of the population relationships among subjects who are alive at that time point. In contrast, a fully conditional approach models the time-varying effects of the population stratified by the actual time of death, where the mean response characterizes individual trends in each cohort stratum. We compare and contrast partly and fully conditional PL-GVCMs in our aforementioned application using hospitalization data from the United States Renal Data System. For inference, we develop generalized likelihood ratio tests. Simulation studies examine the efficacy of estimation and inference procedures.

Key words: Cardiovascular outcomes; End stage renal disease; Fully conditional model; Infection; Partially linear generalized varying coefficient models; Time-varying effects; United States Renal Data System

---

Jason P. Estes is Ph.D. Student, Department of Biostatistics, University of California, Los Angeles, CA 90095. Danh V. Nguyen is Professor, Department of Medicine, University of California, Irvine, CA 92868. Lorien S. Dalrymple is Associate Professor, Division of Nephrology, Department of Medicine, University of California, Sacramento, CA 95691. Yi Mu is Ph.D. Student, Graduate Group in Epidemiology, University of California, Davis, CA 95616. Damla Şentürk (Email: *dsenturk@ucla.edu*) is Associate Professor, Department of Biostatistics, University of California, Los Angeles, CA 90095. This publication was made possible by grants from the National Institute of Diabetes and Digestive and Kidney Diseases (R01 DK092232 and K23 DK093584), National Center for Advancing Translational Sciences (UL1 TR000153, UL1 TR000002), and Dialysis Clinic Inc. The interpretation and reporting of the data presented here are the responsibility of the authors and in no way should be seen as an official policy or interpretation of the United States government.

1 Introduction

As of 2011, more than 430,000 adults in the United States were on dialysis, a life-sustaining treatment (USRDS, 2013). Annual mortality for patients on dialysis treatment is approximately 20-25% with an overall 5-year survival lower than that associated with most malignancies. Cardiovascular (CV) disease and infection remain the leading causes of mortality and hospitalization in patients on dialysis (USRDS, 2013). Characterizing the time-varying effects of risk factors on outcomes of patients with chronic diseases, such as CV events and infection in patients on dialysis, is important for exploring more effective approaches to disease management. For instance, more effective CV risk reduction strategies will first require understanding the time-dynamic changes in patients' CV outcome trajectories over time to allow for identification of timeframes of increased CV risk (probability). This depends on the interplay of time on dialysis (vintage); time since sentinel events such as infection; and patients' baseline characteristics, including baseline co-existing illnesses. Our recent studies (Dalrymple et al., 2011; Mohammed et al., 2012; 2013; Estes et al., 2014) found that infection or infection-related hospitalization was associated with increased risk of CV outcomes (e.g., myocardial infarction stroke, transient ischemic attack) in older patients on dialysis.

A challenge in the next step to elucidate the time-varying effect of infection on patients' CV outcome trajectories over time, from the start of dialysis, is the high mortality in this population. In chronic diseases and/or geriatric populations with high mortality, such as the older dialysis population, many individuals' follow-up times are truncated by death. For example, our study to assess the time-varying effect of infection on CV outcomes uses longitudinal hospitalization data from the United States Renal Data System Annual Data Report (USRDS) for patients aged 65 and older who newly initiated dialysis between January 1, 2000 and December 31, 2007 without a prior history of renal transplant. The follow-up on 80% of the patients through the end of 2009 have been truncated by death. Under such high level of mortality

and when death is related to the outcome variable, one must be careful in selecting statistical models that have useful targets of inference. For instance, information from an estimate of the CV outcome trajectory based on an *unconditional* model, ignoring truncation by death (which implicitly assumes an immortal population), would be of limited practical use.

Thus, a primary focus of the current paper is to develop *conditional* modeling approaches for handling truncation by death. More precisely, for the first time, we will present developments for partially linear generalized varying coefficient models (PL-GVCMs) to model time-varying effects, where the expected outcome trajectory is modeled by conditioning on (a) the dynamic cohort of survivors (*"partly conditional"* approach) and (b) the actual death time (*"fully conditional"* approach). Second, we will apply these conditional PL-GVCM approaches to assess the time-varying effect of infection on CV outcome trajectory. And in this process, we will contrast the targets/goals of inference (i.e., their interpretations) for partly and fully conditional models to provide practical guidance on their applications in the context of longitudinal data with substantial truncation by death. Third, we will present studies evaluating the proposed estimation methods as well as efficacy of generalized likelihood ratio tests (GLRTs) on the varying coefficient functions in the presence of follow-up truncation by death.

We now provide a summary of the relevant literature and an introductory illustration of the partly conditional, fully conditional and unconditional targets of inference for time-varying effects. Standard varying coefficient models (VCMs; Cleveland et al., 1991; Hastie and Tibshirani, 1993) for continuous outcomes and generalized varying coefficient models (GVCMs) for generalized outcomes, including binary and count data (Cai et al., 2000; Zhang, 2004; Qu and Li, 2006; Senturk and Mueller, 2009; Senturk et al., 2013), have been adapted for analyzing longitudinal data (Hoover et al., 1998; Wu and Chiang, 2000; Fan and Zhang, 2000; Chiang et al., 2001; Huang et al., 2002; 2004; Senturk and Mueller, 2010; Senturk and Nguyen, 2011 and references therein). Lu (2008) proposed PL-GVCMs where some regression coefficients

vary with time and others remain constant. PL-GVCM is an extension of the partially linear varying coefficient models (Zhang et al., 2002; Xia et al., 2004; Ahmad et al., 2005; Fan and Huang, 2005) for generalized outcomes, where the covariates are cross-sectional. In our current work, we consider the following PL-GVCM, containing both cross-sectional and longitudinal predictors, necessary for our application:

$$g[E\{Y(t)|X,U(t)\}] = \sum_{r=1}^{p} \beta_r X_r + \sum_{s=1}^{q} \alpha_s(t) U_s(t), \tag{1}$$

where $Y(t)$ is the outcome trajectory, $g(\cdot)$ is a known link function, $X = (X_1, \ldots, X_p)^{\mathrm{T}}$ is the vector of baseline covariates, and $U(t) = \{U_1(t), \ldots, U_q(t)\}^{\mathrm{T}}$ is the vector of longitudinal covariates. The coefficients, $\beta = (\beta_1, \ldots, \beta_p)^{\mathrm{T}}$, describe constant effects corresponding to baseline factors and the time-varying regression coefficients, $\alpha(\cdot) = \{\alpha_1(\cdot), \ldots, \alpha_q(\cdot)\}^{\mathrm{T}}$, capture the dynamic effects of the longitudinal predictors.

Despite the aforementioned rich literature on modeling time-varying effects, limited works have dealt with the consequences of longitudinal data truncated by death. In particular, when death is related to the outcome variable, the statistical modeling requires careful consideration of the relevant targets of inference. For instance, methods based on imputation from the nonignorable dropout literature targeting an unconditional mean trajectory model, specifically $\mu \equiv E\{Y(t)|X,U(t)\}$, would have limited relevance because the imputation of longitudinal data after death implicitly assumes a population where nobody dies. Alternatively, a relevant target of inference is to condition on the cohort of individuals still alive at time t (i.e., all individuals with death time S, where $S > t$), and target the *partly* conditional mean trajectory $\mu_P \equiv E\{Y(t)|X,U(t),S>t\}$ (Estes et al., 2014). A second relevant target of inference in the presence of substantial truncation by death is to target the *fully* conditional mean trajectory, $\mu_F \equiv E\{Y(t)|X,U(t),S=t\}$, which conditions on the actual death time (i.e., $S = t$). We note that the ideas of conditioning on the survival and on actual death time, namely partly and fully conditionals, were introduced by Kurland and Heagerty (2005) and Kurland et al. (2009)

for standard generalized linear models; the current work develops these ideas for time-varying effects.

As a prelude to the more general conditional models considered in this paper, we first illustrate the difference between partly and fully conditional models using a simple GVCM, where about 3 of 4 subjects die during follow-up (detailed in supplemental Appendix A). Figure 1 displays the partly conditional and fully conditional (along with the unconditional) estimates of the varying coefficient function targets. The partly conditional model, which conditions on the cohort alive at time t (years), characterizes time-varying regression relationships for the *dynamic* cohort of survivors. It is relevant to addressing questions such as, *"What is the expected CV outcome risk trajectory during the first two years of dialysis among patients who survive at least two years on dialysis?"* (The partly conditional CV outcome trajectories for the time periods before and after infection can then be compared, for instance.) In Figure 1, it can be seen that the partly conditional trajectory diverges from the unconditional trajectory around year 3 because, by then, the cohort of individuals still alive have critically changed/declined; this reflects the fact that $\mu_P(t) \neq \mu(t)$, particularly for high level of mortality during follow-up. In contrast, a fully conditional model is conditioned on a specific time of death t and, thus, the inferential interest focuses on the time-varying trajectory for the strata of patients who died at time t. Typically, a series of fully conditional models, conditioned on a sequence of death times (as illustrated in Figure 1 for death times $t = 3, 4$, and 5 years) are estimated to compare trends in the expected outcome trajectories for the death stratum. For our data application, the fully conditional model approach will allow us *to compare the CV risk trajectories before and after infection for a series of dialysis patient cohorts who die around 1, 2, and 3 years etc.*

The paper is organized as follows. Conditional PL-GVCM models formulation, estimation, and generalized likelihood ratio tests (GLRTs) for analyzing time-varying effects of infection on CV outcome risk using USRDS data are described in Section 2. Section 3 provides modeling

results and interpretations, followed by simulation studies in Section 4 and a discussion in Section 5.

2 Partly and Fully Conditional Time-Varying Effect Modeling

2.1 Model Specification: Conditional PL-GVCM

As introduced in Section 1, our primary interest is to determine the course of CV risk over time, from the start of dialysis, and assess how the CV risk trajectory changes over time after a pivotal infection-related hospitalization. To specify the conditional PL-GVCMs for this purpose, let S_i be the death time and t_i be the overall follow-up time index of patient i. We divide the time axis, t_i, into two parts, t_{0i} and t_{1i}, to track the follow-up time before and after the first pivotal infection-related hospitalization, respectively. Also, let Z_i mark the time of the first infection-related hospitalization. Thus, for patients who experienced a pivotal initial infection-related hospitalization during follow up, note that $t_i = Z_i + t_{1i}$ after infection, and for patients who do not experience a pivotal infection-related hospitalization during follow up and for those who do experience infection, before their initial infection, $t_i = t_{0i}$. To study the time-varying CV event probability (risk), we model the binary indicator of having a CV event within a 3 month follow-up interval. Since the probability of having more than one CV event in a three month interval is less than 0.1% in our data, we use a binary (rather than a count) outcome in our modeling. Hence, let $Y_i(t_i, t_{0i}, t_{1i})$ be the indicator of a CV event for subject i in a 3 month time interval centered around a fixed value of t_{0i} or t_{1i}. The proposed partly conditional PL-GVCM targets the CV risk, conditioned on being alive:

$$\mu_{i,P} \equiv \mu_{i,P}(t_i, t_{0i}, t_{1i}) = E\{Y_i(t_i, t_{0i}, t_{1i}) | Z_i, X_i, \mathbb{I}_{\mathrm{G}_i}(t_i), S_i > t_i\}, \tag{2}$$

where $\mathbb{I}_{\mathrm{G}_i}(t_i)$ denotes a time-varying indicator of infection-related hospitalization *prior* to time t_i; Z_i is the vintage till first infection-related hospitalization if patient i has at least one infection-related hospitalization; $X_i = (X_{2i}, \ldots, X_{pi})^{\mathrm{T}}$ are baseline covariates. We use the

logit link function, denoted $g(\mu_{i,P}) = \log\{\mu_{i,P}/(1-\mu_{i,P})\}$, to connect the partly conditional mean to the time-varying effects of the covariates:

$$g(\mu_{i,P}) = \alpha_{0,P}(t_{0i})\{1 - \mathbb{I}_{\mathrm{G}_i}(t_i)\} + \alpha_{1,P}(t_{1i})\mathbb{I}_{\mathrm{G}_i}(t_i) + \beta_{1,P}Z_i\mathbb{I}_{\mathrm{G}_i}(t_i) + \sum_{r=2}^{p}\beta_{r,P}X_{ri}, \qquad (3)$$

where $\alpha_{0,P}(t_{0i})$ captures the vintage-varying effects; $\alpha_{1,P}(t_{1i})$ captures the time-varying effects after the initial infection-related hospitalization; the coefficients $\{\beta_{r,P}\}_{r=1}^{p}$, correspond to the effects of vintage prior to the first infection and baseline covariates. The supports for the varying coefficient functions in (3) are: $t_{0i} \in [0, T_{0i}]$, $t_{1i} \in [0, T_{1i}]$, $T_{0i} \leq T, T_{1i} \leq T$, where T is the maximum study follow-up duration; $T = 5$ years in our USRDS data application.

The time-varying indicator, $\mathbb{I}_{\mathrm{G}_i}(t_i)$, in the PL-GVCM (3) allows for a natural transition between the model components before and after the pivotal initial infection-related hospitalization. That is, for the time period *before* the initial infection-related hospitalization among patients with infection-related hospitalization(s) and for the entire follow-up time period among patients with no infection-related hospitalization, the CV risk model is $\mu_{i,P} = g^{-1}\{\alpha_{0,P}(t_{0i}) + \sum_{r=2}^{p}\beta_{r,P}X_{ri}\}$. For patients with at least one infection-related hospitalization, we see from (3) that the CV risk model *after* the initial infection-related hospitalization transitions to $\mu_{i,P} = g^{-1}\{\alpha_{1,P}(t_{1i}) + \beta_{1,P}Z_i + \sum_{r=2}^{p}\beta_{r,P}X_{ri}\}$. Note that this model appropriately accounts for vintage till the initial infection-related hospitalization, namely Z_i.

In contrast, for the fully conditional model, instead of conditioning on survival status, we condition on time of death. For this, we partition the overall follow-up time into disjoint 3 months intervals/bins, where the left endpoint of the first bin is 0, and the right endpoint of the last bin is T. Denote the jth death bin by D_j. The CV risk within bin D_j is

$$\mu_{ij,F} \equiv \mu_{ij,F}(t_i, t_{0i}, t_{1i}) = E\{Y_i(t_i, t_{0i}, t_{1i})|Z_i, X_i, \mathbb{I}_{\mathrm{G}_i}(t_i), S_i \in D_j\}, \qquad (4)$$

and the fully conditional PL-GVCM for the CV risk is

$$g(\mu_{ij,F}) = \alpha_{0j,F}(t_{0i})\{1 - \mathbb{I}_{\mathrm{G}_i}(t_i)\} + \alpha_{1j,F}(t_{1i})\mathbb{I}_{\mathrm{G}_i}(t_i) + \beta_{1j,F}Z_i\mathbb{I}_{\mathrm{G}_i}(t_i) + \sum_{r=2}^{p}\beta_{rj,F}X_{ri}. \qquad (5)$$

The parameters and varying coefficient functions in (5) above are analogously defined as in the partly conditional model (3).

2.2 Estimation

Estimation procedures for partially linear VCMs and PL-GVCMs usually contain several main steps, where the regression coefficients of the linear part are targeted first, followed by estimation of the varying coefficient functions (VCFs) using coefficient estimates of the linear part from the initial step (Zhang et al., 2002; Xia et al., 2004; Fan and Huang, 2005). For example, Lu (2008) proposed local quasi-likelihood for estimation of the α_s's first, then targeting the β_r's via maximum likelihood using the estimated VCFs, followed by re-estimation of the VCFs using the estimated β_r's. To fit the proposed conditional PL-GVCMs, we will extend the method of Lu (2008) to the context of longitudinal data and allow for longitudinal covariates where follow-up is truncated by death. The proposed 3-step estimation algorithm is provided next for the partly conditional model. We note that for the fully conditional PL-GVCM, this estimation method is applied to data from death bin D_j, instead of the entire cohort.

2.2.1 Step 1: Initial Estimation of $\alpha_{0,P}(t_0)$ and $\alpha_{1,P}(t_1)$

We begin by partitioning each patient's follow-up period into disjoint 3-month intervals after initiation of dialysis and after the initial infection-related hospitalization if the patient has at least one infection-related hospitalization. Let N_{0i} denote the number of 3-month intervals in the ith patient's follow-up after initiation of dialysis until the initial infection-related hospitalization or to the end of follow-up (for a patient without an infection-related hospitalization). Similarly, let N_{1i} be the number of 3-month intervals since the initial infection-related hospitalization to the end of follow-up for patient i. Further, define t_{0ik} and $t_{1ik'}$ to be the midpoints of the kth and k'th 3-month time intervals since initiation of dialysis and since the initial infection-related hospitalization, respectively. We define the binary response variable

$Y_{0,ik} \equiv Y_i(t_i = t_{0i} = t_{0ik}) = 1$, if the ith patient had at least one CV event in the kth 3-month interval after initiation of dialysis. Similarly, $Y_{1,ik'} \equiv Y_i(t_i = Z_i + t_{1i}, t_{1i} = t_{1ik'}) = 1$ if the ith patient had at least one CV event in the k'th 3-month interval after the initial infection-related hospitalization. Hence, the available data is $\{(t_{0ik}, t_{1ik'}, X_{ri}, Z_i,\ Y_{0,ik}, Y_{1,ik'}) : i = 1, \ldots, n;\ k = 1, \ldots, N_{0i};\ k' = 1, \ldots, N_{1i}\}$, where n is the total number of subjects.

The first step of the estimation algorithm targets the VCFs, $\alpha_{0,P}(t_0)$ and $\alpha_{1,P}(t_1)$, via local maximum likelihood (ML). Assuming that the VCFs have continuous second derivatives, we approximate each function locally by $\alpha_{0,P}(t_0) \approx c_0 + c_1(t_0 - s_0)$ and $\alpha_{1,P}(t_1) \approx d_0 + d_1(t_1 - s_0)$ for t_0 and t_1 in the neighborhood of the fixed time point s_0. Maximizing the local log-likelihood $\ell_1(\mathbf{c})$, defined by

$$\begin{aligned} \ell_1(\mathbf{c}) &= \frac{1}{\sum_{i=1}^n N_i} \sum_{i=1}^n \Bigg(\sum_{k=1}^{N_{0i}} \ell\left[g^{-1}\left\{ c_0 + c_1(t_{0ik} - s_0) + \sum_{r=2}^p b_r X_{ri} \right\}, Y_{0,ik} \right] K_h(t_{0ik} - s_0) \\ &+ \sum_{k'=1}^{N_{1i}} \ell\left[g^{-1}\left\{ d_0 + d_1(t_{1ik'} - s_0) + b_1 Z_i + \sum_{r=2}^p b_r X_{ri} \right\}, Y_{1,ik'} \right] K_h(t_{1ik'} - s_0) \Bigg), \end{aligned} \tag{6}$$

provides the initial local ML estimators for the VCFs, namely $\widehat{\alpha}_{0,P}(t_0) = \widehat{c}_0$ and $\widehat{\alpha}_{1,P}(t_1) = \widehat{d}_0$. In the above local log-likelihood, $K_h(\cdot) = K(\cdot/h)/h$, $K(\cdot)$ denotes a kernel function and h is the bandwidth; $\mathbf{c} \equiv (c_0, c_1, d_0, d_1, b_1, \ldots, b_p)^{\mathrm{T}}$; $N_i = N_{0i} + N_{1i}$ and $\ell(\cdot, \cdot)$ denotes the log-likelihood function. Note that the local likelihood only includes data from subjects who are still alive at t_0 and t_1. We also point out that the formulation for the local log-likelihood $\ell_1(c)$ in (6) tacitly utilizes working independence for repeated values within a subject. This is consistent with Kurland and Heagerty (2005), which found that standard likelihood based methods will not target the partly conditional mean, and generalized estimating equations with independence weights provides unbiased estimation in a generalized linear model of longitudinal data.

The maximization can be implemented using the Newton-Raphson algorithm. For this, let $\widehat{p}_{0,ik} = g^{-1}\{\widehat{c}_0 + \widehat{c}_1(t_{0ik} - s_0) + \widehat{\tau}_i\}$ and $\widehat{p}_{1,ik'} = g^{-1}\{\widehat{d}_0 + \widehat{d}_1(t_{1ik'} - s_0) + \widehat{b}_1 Z_i + \widehat{\tau}_i\}$, where $\widehat{\tau}_i = \sum_{r=2}^p \widehat{b}_r X_{ri}$. Also, define $\{\kappa_{v,ij} \equiv K_h(t_{vij} - s_0)\}_{j=1}^{N_{vi}}$, $\{\widetilde{p}_{v,ij} \equiv \widehat{p}_{v,ij}(1 - \widehat{p}_{v,ij})\}_{j=1}^{N_{vi}}$, and

$\{\widetilde{\kappa}_{v,ij} \equiv \kappa_{v,ij}\widetilde{p}_{v,ij}\}_{j=1}^{N_{vi}}$, for $v = 0, 1$. Then the Newton-Raphson update at iteration $m+1$ is given by

$$\widehat{\mathbf{c}}_{m+1} = \widehat{\mathbf{c}}_m + \left\{\sum_{i=1}^{n} \mathcal{X}_{1i}^{\mathrm{T}} W_{1i}(\widehat{\mathbf{c}}_m)\mathcal{X}_{1i}\right\}^{-1} \sum_{i=1}^{n} \mathcal{X}_{1i}^{\mathrm{T}} W_{2i}\widetilde{Y}_i(\widehat{\mathbf{c}}_m),$$

where

$$\mathcal{X}_{1i} = \left[\begin{array}{cc|cc|ccc} 1 & (t_{0i1}-s_0) & 0 & 0 & 0 & X_{2i} & \ldots & X_{pi} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & & \vdots \\ 1 & (t_{0iN_{0i}}-s_0) & 0 & 0 & 0 & X_{2i} & \ldots & X_{pi} \\ \hline 0 & 0 & 1 & (t_{1i1}-s_0) & Z_i & X_{2i} & \ldots & X_{pi} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 1 & (t_{1iN_{1i}}-s_0) & Z_i & X_{2i} & \ldots & X_{pi} \end{array}\right] \tag{7}$$

is the predictor matrix of size $N_i \times (p+4)$, $W_{1i}(\widehat{\mathbf{c}}_m) = \mathrm{diag}\{\widetilde{\kappa}_{0,i1}, \ldots, \widetilde{\kappa}_{0,iN_{0i}}, \widetilde{\kappa}_{1,i1}, \ldots, \widetilde{\kappa}_{1,iN_{1i}}\}$, $W_{2i} = \mathrm{diag}\{\kappa_{0,i1}, \ldots, \kappa_{0,iN_{0i}}, \kappa_{1,i1}, \ldots, \kappa_{1,iN_{1i}}\}$ and $\widetilde{Y}_i(\widehat{\mathbf{c}}_m) = (Y_{0,i1} - \widehat{p}_{0,i1}, \ldots, Y_{0,iN_{0i}} - \widehat{p}_{0,iN_{0i}}, Y_{1,i1} - \widehat{p}_{1,i1}, \ldots, Y_{1,iN_{1i}} - \widehat{p}_{1,iN_{1i}})^{\mathrm{T}}$, for a Bernoulli distributed response. For modeling a Poisson distributed response, $W_{1i}(\widehat{\mathbf{c}}_m) = \mathrm{diag}\{\kappa_{0,i1}\widehat{p}_{0,i1}, \ldots, \kappa_{0,iN_{0i}}\widehat{p}_{0,iN_{0i}}, \kappa_{1,i1}\widehat{p}_{1,i1}, \ldots, \kappa_{1,iN_{1i}}\widehat{p}_{1,iN_{1i}}\}$. For subjects who do not have any infection-related hospitalization, the predictor matrix reduces to size $N_{0i} \times (p+4)$ and sizes of the above quantities adjust accordingly.

2.2.2 Step 2: Estimation of $\beta_{r,P}$

In the second step, we target $\beta_{r,P}$, by using the VCF estimators $\widehat{\alpha}_{0,P}(t_{0ik})$ and $\widehat{\alpha}_{1,P}(t_{1ik'})$ obtained in step 1 in the global likelihood,

$$\ell_2(\mathbf{e}) = \frac{1}{\sum_{i=1}^{n} N_i} \sum_{i=1}^{n} \left(\sum_{k=1}^{N_{0i}} \ell\left[g^{-1}\left\{\widehat{\alpha}_{0,P}(t_{0ik}) + LC\right\}, Y_{0,ik}\right] + \sum_{k'=1}^{N_{1i}} \ell\left[g^{-1}\left\{\widehat{\alpha}_{1,P}(t_{1ik'}) + e_1 Z_i + LC\right\}, Y_{1,ik'}\right]\right),$$

resulting in the ML estimators $\widehat{\beta}_{r,P} = \widehat{e}_r$ for $r = 1, \ldots, p$, where LC denotes $\sum_{r=2}^{p} e_r X_{ri}$. The maximization can be done using the Newton-Raphson algorithm with the $m+1$ iteration update given by

$$\widehat{\mathbf{e}}_{m+1} = \widehat{\mathbf{e}}_m + \left\{\sum_{i=1}^{n} \mathcal{X}_{2i}^{T} W_i(\widehat{\mathbf{e}}_m)\mathcal{X}_{2i}\right\}^{-1} \left\{\sum_{i=1}^{n} \mathcal{X}_{2i}^{T} \widetilde{Y}_i(\widehat{\mathbf{e}}_m)\right\} \quad \text{where} \quad \mathcal{X}_{2i} = \left[\begin{array}{c|cccc} 0 & X_{2i} & X_{3i} & \ldots & X_{pi} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & X_{2i} & X_{3i} & \ldots & X_{pi} \\ \hline Z_i & X_{2i} & X_{3i} & \ldots & X_{pi} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ Z_i & X_{2i} & X_{3i} & \ldots & X_{pi} \end{array}\right]$$

is the predictor matrix of size $N_i \times p$, $\widehat{p}_{0,ik} = g^{-1}\{\widehat{\alpha}_{0,P}(t_{0ik}) + \sum_{r=2}^{p} \widehat{e}_{r,m} X_{ri}\}$, $\widehat{p}_{1,ik'} = g^{-1}\{\widehat{\alpha}_{1,P}(t_{1ik'}) + \widehat{e}_{1,m} Z_i + \sum_{r=2}^{p} \widehat{e}_{r,m} X_{ri}\}$, $W_i(\widehat{\mathbf{e}}_m) = \text{diag}\{\widetilde{p}_{0,i1}, \ldots, \widetilde{p}_{0,iN_{0i}}, \widetilde{p}_{1,i1}, \ldots, \widetilde{p}_{1,iN_{1i}}\}$, and $\widetilde{Y}_i(\widehat{\mathbf{e}}_m) = (Y_{0,i1} - \widehat{p}_{0,i1}, \ldots, Y_{0,iN_{0i}} - \widehat{p}_{0,iN_{0i}}, Y_{1,i1} - \widehat{p}_{1,i1}, \ldots, Y_{1,iN_{1i}} - \widehat{p}_{1,iN_{1i}})^{\mathrm{T}}$, for a Bernoulli distributed response. For modeling a Poisson distributed response, $W_i(\widehat{\mathbf{e}}_m) = \text{diag}(\widehat{p}_{0,i1}, \ldots, \widehat{p}_{0,iN_{0i}}, \widehat{p}_{1,i1}, \ldots, \widehat{p}_{1,iN_{1i}})$.

2.2.3 Step 3: Final Estimation of $\alpha_{0,P}(t_0)$ and $\alpha_{1,P}(t_1)$

In step 3, we use the final global estimates for $\beta_{r,P}$, to arrive at the final VCF estimators. For this, we maximize the local likelihood given in step 1, where b_r are replaced with $\widehat{\beta}_{r,P}$, $r = 1, \ldots, p$ from step 2. Hence, the $N_i \times 4$ design matrix $\mathcal{X}_{1i}$ uses the first 4 columns of the design matrix defined in step 1 and $\widehat{p}_{0,ik}$ and $\widehat{p}_{1,ik'}$ are redefined as $g^{-1}\{\widehat{c}_0 + \widehat{c}_1(t_{0ik} - s_0) + \sum_{r=2}^{p} \widehat{\beta}_{r,P} X_{ri}\}$ and $g^{-1}\{\widehat{d}_0 + \widehat{d}_1(t_{1ik'} - s_0) + \widehat{\beta}_{1,P} Z_i + \sum_{r=2}^{p} \widehat{\beta}_{r,P} X_{ri}\}$, respectively.

2.3 Generalized Likelihood Ratio Test Under Follow-up Truncation by Death

The proposed PL-GVCM aims to characterize the time-varying CV outcome trajectories from the start of dialysis and to compare patterns of CV outcome risk before and after an infection. These time-varying effects are described by the VCFs, $\alpha_0(t_0)$ and $\alpha_1(t_1)$, for the time periods before and after infection, respectively. Thus, we consider hypothesis tests on the VCFs. The first hypothesis of interest involves whether the VCFs are constant over time (before and after infection), i.e., $H_0 : \alpha_0(t_0) = c_0$ and $\alpha_1(t_1) = c_1$, as illustrated in Figure 6(a). This hypothesis encompasses the case where the infection event induces a constant change (shift) in the CV outcome risk (i.e., when $c_0 \neq c_1$). A second hypothesis of interest involves a comparison of time-varying effects before and after an initial infection-related hospitalization, specifically $H_0 : \alpha_0(t_0) = \alpha_1(t_1)$, as illustrated in Figure 6(b). This hypothesis examines whether the infection event leads to a transient change (e.g. an increase) in CV risk, but the CV risk pattern over time after infection parallels the CV risk trajectory before the infection event.

In the first hypothesis test for constancy of the varying coefficient functions, the null hy-

pothesis is parametric, while the alternative is nonparametric. In the second hypothesis test, both null and alternative hypotheses are nonparametric. Fan et al. (2001) extended GLRTs for nonparametric inferences in a variety of models. More specifically, they showed that the Wilks phenomenon that the asymptotic null distributions of the GLRTs are independent of nuisance parameters holds for a variety of nonparametric problems for i.i.d. data. Based on these ideas, we consider the GLRTs for the above two hypotheses in the PL-GVCM with longitudinal data substantially truncated by death. Because the within subject correlation for the response is quite weak in our data application (~ 0.02), we consider extensions of the Fan et al. (2001) i.i.d. framework to longitudinal data where the test statistic defined via log-likelihoods and the bootstrap data generation under the null hypotheses assume independence for repetitions within a subject. We study the validity and power of the proposed GLRTs using simulations in Section 4, where high follow-up truncation by death ranges from 40-80%.

The GLRT statistic, denoted T, is of the form $T = r_k\{\ell(H_1) - \ell(H_0)\}$ where $r_k = \{K(0) - 0.5\int K^2(u)du\}/[\int\{K(u) - 0.5(K*K)(u)\}^2du]$, $K*K$ denoting the convolution of K with itself and $\ell(H_0)$ and $\ell(H_1)$ denoting the log-likelihoods under the null and alternative hypothesis, respectively. The form of log-likelihoods, $\ell(\cdot)$, is given by

$$\sum_{i=1}^{n}\left[\sum_{k=1}^{N_{0i}}\{Y_{0,ik}\log(p_{0,ik}) + (1-Y_{0,ik})\log(1-p_{0,ik})\} + \sum_{k'=1}^{N_{1i}}\{Y_{1,ik'}\log(p_{1,ik'}) + (1-Y_{1,ik'})\log(1-p_{1,ik'})\}\right]. \tag{8}$$

In our application, we use the Epanechnikov kernel where $r_k = 2.1153$. Fan et al. (2001) showed that the GLRT statistic follow a χ^2-distribution asymptotically; however, the level of the test may not be achieved consistently. To alleviate this issue, we will adopt the approach by Cai et al. (2000) by using a conditional bootstrap procedure which provides an improved estimate of the null distribution with moderate sample size for GVCMs. More precisely, we will use a nonparametric bootstrap method to estimate the null distribution. The main steps of the GLRT algorithm are:

(i) Estimate the PL-GVCM parameters under the null and alternative hypothesis, providing $\ell(H_0)$ and $\ell(H_1)$.

(ii) Compute the GLRT statistic $T = 2.1153\{\ell(H_1) - \ell(H_0)\}$.

(iii) Generate a bootstrap sample of response values conditional on the estimates of the model parameters under the null hypothesis.

(iv) Compute the test statistic T based on the bootstrap sample by repeating steps (i)-(ii); denote this bootstrap statistic by T^*.

(v) Use the distribution of the bootstrap test statistic, T^*, to approximate the distribution of T under the null.

For the first test $H_0 : \alpha_0(t_0) = c_0$ and $\alpha_1(t_1) = c_1$, the proposed model in (3) reduces to the generalized linear model $g(\mu_{i,P}) = c_0\{1 - \mathbb{I}_{\mathrm{G}_i}(t_i)\} + c_1\mathbb{I}_{\mathrm{G}_i}(t_i) + \beta_{1,P}Z_i\mathbb{I}_{\mathrm{G}_i}(t_i) + \sum_{r=2}^{p}\beta_{r,P}X_{ri}$ under the null hypothesis. Parameters $\beta_{r,P}$ can be estimated by maximizing the global likelihood in step 2 of the proposed estimation algorithm where $\widehat{\alpha}_{0,P}(t_{0i})$ and $\widehat{\alpha}_{1,P}(t_{1i})$ would be replaced by c_0 and c_1 specified in the null, respectively. To obtain the parameter estimates of the partly conditional PL-GVCM under the alternative, we utilize the proposed 3-step fitting algorithm described in Section 2.2. Next, the test statistic is computed using the log-likelihoods given in (8) under the null and alternative hypotheses; where under the null $\widehat{p}_{0,ik} = g^{-1}(c_0 + \widehat{\tau}_i)$ and $\widehat{p}_{1,ik} = g^{-1}(c_1 + \widehat{\beta}_{1,P}Z_i + \widehat{\tau}_i)$ with $\widehat{\tau}_i = \sum_{r=2}^{p}\widehat{\beta}_{r,P}X_{ri}$. Similarly, under the alternative $\widehat{p}_{0,ik} = g^{-1}\{\widehat{\alpha}_0(t_{0i}) + \widehat{\tau}_i\}$ and $\widehat{p}_{1,ik} = g^{-1}\{\widehat{\alpha}_1(t_{1i}) + \widehat{\beta}_{1,P}Z_i + \widehat{\tau}_i)$, all evaluated using parameter estimates under respective hypotheses. The response values $(Y^*_{0,i1}, Y^*_{0,i2}, \ldots, Y^*_{0,iN_{0i}}, Y^*_{1,i1}, Y^*_{1,i2}, \ldots, Y^*_{1,iN_{1i}})^{\mathrm{T}}$ in the bootstrap sample are generated using parameter estimates under the null according to $Y^*_{0,ik} \sim Bernoulli\{g^{-1}(c_0 + \widehat{\tau}_i)\}$ and $Y^*_{1,ik} \sim Bernoulli\{g^{-1}(c_1 + \widehat{\beta}_{1,P}Z_i + \widehat{\tau}_i)\}$. After B bootstrap test statistics are obtained based on B

bootstrap samples, the χ^2-distribution of T under the null is approximated via estimating the degrees of freedom of the distribution based on the distribution of the bootstrap test statistics. B is taken to be 500 in our applications.

For the second test $H_0 : \alpha_0(t_0) = \alpha_1(t_1)$, model (3) reduces to $g(\mu_{i,P}) = \alpha(t_{0i})\{1-\mathbb{I}_{\mathrm{G}_i}(t_i)\}+ \alpha(t_{i1})\mathbb{I}_{\mathrm{G}_i}(t_i) + \beta_{1,P}Z_i\mathbb{I}_{\mathrm{G}_i}(t_i) + \sum_{r=2}^{p}\beta_{r,P}X_{ri}$ under the null, where $\alpha(\cdot)$ denotes the common varying coefficient function under the equality $\alpha_0(t_0) = \alpha_1(t_1)$. An adaptation of the proposed estimation algorithm is used to estimate the parameters under the null hypothesis, where d_0 and d_1 is replaced with c_0 and c_1, respectively in (6), $\mathcal{X}_{1i}$ in (7) reduces down to a $N_i \times (p+2)$ matrix with second to fourth columns replaced with $(t_{0i1} - s_0, \ldots, t_{0iN_{0i}} - s_0, t_{1i1} - s_0, \ldots, t_{1iN_{1i}} - s_0)^{\mathrm{T}}$ and similar adjustment are made in step 3. For the unrestricted partly conditional PL-GVCM under the alternative, parameters are targeted with the proposed 3-step estimation algorithm of Section 2.2. Similar to the the first hypothesis test, the test statistic is computed using the likelihood in (8) under the null and alternative hypotheses, where under the null $\widehat{p}_{0,ik} = g^{-1}\{\widehat{\alpha}(t_{0i})+\widehat{\tau}_i\}$ and $\widehat{p}_{1,ik} = g^{-1}\{\widehat{\alpha}(t_{1i})+\widehat{\beta}_{1,P}Z_i+\widehat{\tau}_i\}$; under the alternative $\widehat{p}_{0,ik} = g^{-1}\{\widehat{\alpha}_0(t_{0i})+ \widehat{\tau}_i\}$, $\widehat{p}_{1,ik} = g^{-1}\{\widehat{\alpha}_1(t_{1i}) + \widehat{\beta}_{1,P}Z_i + \widehat{\tau}_i\}$ using parameter estimates under respective hypotheses. The bootstrap response $(Y^*_{0,i1}, Y^*_{0,i2}, \ldots, Y^*_{0,iN_{0i}}, Y^*_{1,i1}, Y^*_{1,i2}, \ldots, Y^*_{1,iN_{1i}})^{\mathrm{T}}$ is generated under the null according to $Y^*_{0,ik} \sim Bernoulli[g^{-1}\{\widehat{\alpha}(t_{0i})+\widehat{\tau}_i\}]$ and $Y^*_{1,ik} \sim Bernoulli[g^{-1}\{\widehat{\alpha}(t_{1i})+\widehat{\beta}_{1,P}Z_i+ \widehat{\tau}_i\}]$. The bootstrap test statistics are used to approximate the distribution of T under the null similar to the first test.

3 Applications to Infection-Cardiovascular Risk Modeling

3.1 Description of the Study Cohort

We use data from the USRDS, a national data system that collects information on nearly all patients with end-stage renal disease in the US, including data on inpatient care patient demographics and baseline patient factors prior to the start of dialysis. The population of

inference are adults aged 65 to 90 who newly initiated dialysis between January 1, 2000 and December 31, 2007 without a prior history of renal transplant. Eligibility criterion included (a) having survived the first 90 days of dialysis and did not recover renal function or receive a kidney transplant during this interval, (b) having Medicare as the primary payer on day 91 of dialysis, and (c) receiving hemodialysis or peritoneal dialysis on day 91. Thus, the observation period began on day 91 and subjects were followed-up until death (80%), study end on December 31, 2009 or after 5 years of observation (from the initiation of dialysis or the initial infection-related hospitalization). We exclude 1.3% of the cohort that recovered renal function and 2.1% of the cohort that received a kidney transplant, since the evaluation of candidates for transplant relates to overall health.

The outcome, CV events were defined as a myocardial infarction, unstable angina, stroke, or transient ischemic attack, determined from primary discharge diagnosis and based on the International Classification of Disease, 9th Revision, Clinical Modification (ICD-9-CM) codes. An infection-related hospitalization was determined from discharge diagnosis, also based on ICD-9-CM codes, and included the following types of infection: blood stream infections and sepsis; central nervous system; cardiovascular; peritoneal; gastrointestinal and hepatobiliary; genitourinary; pulmonary; skin and soft tissue; bone and joint; dialysis access and central venous catheters; device, procedure and surgery-related. Table 1 summarizes the baseline covariates included in the study.

3.2 Cardiovascular Outcome Risk Trajectories

3.2.1 Partly and Fully Conditional Time-Varying Models without Covariates

To explore partly and fully conditional time-varying effects, we first consider the CV outcome risk trajectories over time from the initiation of dialysis without covariates. For this, the partly conditional GVCM is $g[E\{Y_i(t_i)|S_i > t_i\}] = \alpha_P(t_i)$, where the model fits to 3 cohorts are shown in Figure 2(a): (i) patients who die, (ii) patients followed to the end of study (EOS), and (iii)

all patients combined. The VCF estimates (the CV risk trajectories) have generally increasing trends over time after dialysis, both in the cohort of patients (i) whose death is observed and (ii) followed to the EOS. As expected, the CV risk over time is lower for the cohort of patients alive at the EOS compared to the cohort of patients who die during follow-up; and due to the high mortality of patients on dialysis, the ratio of sample sizes of cohort (i) over cohort (ii) is sharply decreasing with follow-up time (Figure 2(a), solid gray line). Even though CV risk trajectories are increasing in cohorts (i) and (ii), for the combined cohort of all patients (iii) the CV risk trajectory has an overall decreasing trend, especially within the first 2 years after starting dialysis. This is related to the fact that the partly conditional model describes different (dynamic) cohorts at each time point in the follow-up. That is, while the CV risk is high at the initiation of dialysis because the dynamic cohort of survivors consists mostly of patients with observed death and higher CV risk, this CV risk decreases over time as the ratio of the number of patient who die relative to patients alive at the end of follow-up declines in the dynamic cohort of survivors. This is illustrated in Figure 2(a), where the combined cohort VCF estimate (dashed line) represents a weighted average of the estimates for cohorts (i) and (ii), which depends on the changing sample size ratios (gray line) over time.

Figure 2(b) displays 4 fully conditional model fits to data from 4 death bins (strata) with midpoints 1.125, 2.125, 3.125, and 4.12 years (time of death). These fully conditional analyses can be interpreted simply as stratified analyses. As expected, we also see an overall global increasing CV trajectory for each death bin and CV risk is substantially higher for early death stratum. We emphasize that while the partly conditional model is fitted to the entire cohort, the fully conditional model can only be fitted in a subset of the cohort for patients whose death is observed since it conditions on death time. Thus, the estimated VCFs should be interpreted accordingly. We note that the phenomenon of opposing trends observed in the partly and fully conditional models was replicated in a simulation study (details in supplemental Appendix B).

A key aspect in replicating this phenomenon is the inclusion of a high proportion of subjects with observed mortality early on near the start of dialysis where this proportion gradually decreases with follow-up time (Figure 2(c)-(d)).

3.2.2 Time-Varying CV Risks Before and After Infection, and Baseline Factors

We next turn to the main study objectives, which are to examine the CV risk trajectories during the time periods before and after an initial infection-related hospitalization and to assess the association of vintage and patient baseline characteristics, including comorbidities, on CV outcome. For this, we fit the partly and fully conditional PL-GVCMs described in Section 2.1 with covariates demographic characteristics (age, sex, race), comorbidities (diabetes, coronary heart disease, congestive heart failure, peripheral vascular disease), body mass index (BMI) and estimated glomerular filtration rate (eGFR). Bandwidth selection are given in supplemental Appendix C.

The estimated partly conditional VCFs before and after infection, namely $\widehat{\alpha}_P(t_0)$ and $\widehat{\alpha}_P(t_1)$, and the corresponding CV risk trajectories are given in Figures 3(a) and 3(b), respectively. Also, given are 90% bootstrap percentile confidence intervals (CIs) based on 200 bootstrap samples where entire subject trajectories are sampled with replacement. We formally tested whether the partly conditional VCFs characterizing CV risks are constant over time (Test I) and whether they are equal to each other (Test II) using the GLRTs described in Section 2.3. There is strong evidence indicating that there is differential time-varying effects before and after infection (both null hypotheses rejected with p-value $< .0001$). As evident from Figures 3(a)-(b), both VCFs (and corresponding CV outcome risk trajectories) are decreasing in time for the dynamic cohort of survivors. Furthermore, the initial infection-related hospitalization marks a significant increase in CV risk with non-overlapping CIs for $\alpha_{0,P}(t_0)$ and $\alpha_{1,P}(t_1)$. Figures 3(c)-(f) show the estimated CV risk trajectories where the initial infection-related hospitalization occurs at 1-4 years after starting dialysis. This indicates a

sustained increase in CV risk across the duration of follow-up after infection, in the sense that the CV risk levels after infection do not return to the levels observed at initiation of dialysis. In addition, the CV risk declines at a faster rate within the first year after initiation of dialysis compared to the linear decrease after the initial infection-related hospitalization.

Results for the fully conditional model fits, stratified by death bins, show that CV risk has a general decreasing trend as survival of the patients in the bins increase (with bin midpoints or time of death at 1.125, 2.125, 3.125, and 4.125 years); this pattern of results (omitted) is similar to Figure 2(b). Figures 4(a)-(c) show the typical pattern of increased CV risk after the initial infection in the fully conditional model fits, consistently across death bins/strata. While the partly conditional model provides information about the dynamic cohort of survivors, the fully conditional model provides an opportunity to compare estimated effects across cohorts with differential death strata directly.

The estimated effects of baseline covariates, $\{\widehat{\beta}_{r,F}\}$, on CV risk for a sequence of fully conditional models are summarized in Figure 5. Being male is associated with lower CV risk in both the fully and partly conditional model ($\widehat{\beta}_{3,P} = -.125$, 95% bootstrap CI [95% bCI]: $(-.143, -.110)$). Baseline comorbidities, including coronary heart disease ($\widehat{\beta}_{7,P} = .201$, 95% bCI: $(.185, .218)$) and diabetes ($\widehat{\beta}_{9,P} = .179$, 95% bCI: $(.164, .198)$) are associated with higher CV outcome risk in both the partly and fully conditional models. Several comorbidities, specifically congestive heart failure ($\widehat{\beta}_{6,P} = .045$, 95% bCI: $(.026, .065)$) and peripheral vascular disease ($\widehat{\beta}_{8,P} = .093$, 95% bCI: $(.070, .115)$), in addition to baseline age ($\widehat{\beta}_{2,P} = .009$, 95% bCI: $(.007, .010)$), are found to be associated with increased CV risk in the partly conditional model. But once conditioned on death time, are not found significant in most of the death bins (Figure 5). This may be related to some comorbidities and age being related to CV risk via their effect on survival in the entire cohort, where once conditioned on death time may no longer be associated with CV risk, while others such as coronary heart disease and diabetes having a

more direct effect on CV across differential survival. Among those who survive longer, higher BMI is associated with lower CV outcome risk, and among those who die within 3 years of dialysis, higher eGFR is associated with lower CV risk, consistent with the general trends observed in the partly conditional model ($\widehat{\beta}_{10,P} = -.008$, 95% bCI: $(-.010, -.006)$ for BMI; $\widehat{\beta}_{11,P} = -.009$, 95% bCI: $(-.011, -.008)$ for eGFR). Finally, the particular infection time does not seem to have a strong association with CV risk in either the partly ($\widehat{\beta}_{11,P} = -.021$, 95% bCI: $(-.030, -.010)$) or the fully conditional models (Figure 5(a)).

4 Simulation Studies

As described in Section 2.2, the fully conditional estimation involves fitting the PL-GVCM within each death bin, where subjects with similar death times are grouped together. Thus, the issue of truncation by death is handled by stratification by death time (death bins) and the model fits within each death bin follow a standard estimation algorithm for PL-GVCM. In contrast, the partly conditional PL-GVCM is fitted based on subjects who have differential follow-up, where many individuals' follow-up times are truncated by death. Thus, our simulation studies here will focus on the finite sample properties of the proposed estimation method for the partly conditional PL-GVCM; similarly we will examine the validity and power of the proposed GLRTs in Section 2.3.

4.1 Simulation Model and Design

To study the efficacy of the estimation method under truncation by death, we consider a model for the partly conditional outcome mean, $\mu_{i,P} = E\{Y(t_i, t_{0i}, t_{1i})|Z_i, X_{1i}, X_{2i}, \mathbb{I}_{\mathrm{G}_i}(t_i), S_i > t_i\}$, through the following PL-GVCM:

$$\log\{\mu_{i,P}/(1-\mu_{i,P})\} = \alpha_{0,P}(t_{0i})\{1-\mathbb{I}_{\mathrm{G}_i}(t_i)\}+\alpha_{1,P}(t_{1i})\mathbb{I}_{\mathrm{G}_i}(t_i)+\beta_{1,P}Z_i\mathbb{I}_{\mathrm{G}_i}(t_i)+\beta_{2,P}X_{1i}+\beta_{3,P}X_{2i},$$

where $\alpha_{0,P}(t) = -.05t^2 + .025t - 1.25$, $\alpha_{1,P}(t) = -.03t^2 - .05t$, $(\beta_{1,P}, \beta_{2,P}, \beta_{3,P}) = (-.5, .5, 1)$, and the time support is $t_{vi} \in [0, T_{vi}]$, $T_{vi} \leq T = 5$ (for $v = 0, 1$). X_{1i} is generated from a Gamma distribution with {shape, rate} parameters $\{4, 6\}$ and $X_{2i} \sim Bernoulli(.52)$. In order to generate the time-varying indicator variable, $\mathbb{I}_{\mathrm{G}_i}(t_i)$, we first generate a binary indicator of whether or not a subject experiences an infection-related hospitalization according to a Bernoulli distribution with probability .68 to mimic the infection rate in our data application. For those subjects who experience an infection-related hospitalization, we generate $Z_i = \frac{1}{4}\lfloor 4W_i \rfloor$ where $W_i \sim N(1.25, .25)$ and $\lfloor \cdot \rfloor$ denotes the floor function.

The response vector and survival time are generated jointly using the bisection algorithm, similar to Estes et al. (2014) and Kurland and Heagerty (2005). This data simulation design mimics the real data in that within-subject correlation of the response is low ($\sim$ 0.04) and truncation by death is high during the 5-year follow-up, ranging from 40-80%. The binary response $Y_{0,ik}$ and $Y_{1,ik'}$, are generated as indicators for $(Y^*_{0,ik} > 0)$ and $(Y^*_{1,ik'} > 0)$, respectively. For subjects who do not experience an infection-related hospitalization, we generate $(Y^*_{0,i1}, \ldots, Y^*_{0,i21}, S_i)^{\mathrm{T}}$ according to a 22-dimensional normal distribution with mean vector $[\mu^{*\mathrm{T}}_{0,i}, E(S_i) = 3.38]^{\mathrm{T}}$ where $\mu^*_{0,i} = (\mu^*_{0,i1}, \ldots, \mu^*_{0,i21})^{\mathrm{T}}$ is the mean vector $E\{Y_i(t_i, t_{0i}, t_{1i})|Z_i, X_i, \mathbb{I}_{\mathrm{G}_i}(t_i)\}$ of the ith subject, *unconditional* on survival status. We include a maximum of 21 repeated measures per subject on the outcome similar to the outcome in USRDS data measured every 3 months for a maximum of 5 years of follow-up. The covariance matrix of the 22-dimensional normal distribution is $\Sigma = \left[I_{21}, -.05\eta_{21}; -.05\eta^{\mathrm{T}}_{21}, .5\right]$, where I_a is an identity matrix of size a and η_a is a vector of ones of size a. Elements of the unconditional mean vector, $\mu^*_{0,i}$, are computed through the correspondence:

$$\mu_{0,ik} = E[Y_{0,ik}|S_i > t_{0ik}] = P(Y^*_{0,ik} > 0|S_i > t_{0ik}) = P(Y^*_{0,ik} > 0, S_i > t_{0ik})/P(S_i > t_{0ik}), \quad (9)$$

where $\mu_{0,ik} = g^{-1}\{\alpha_{0,P}(t_{0ik}) + \beta_{2,P}X_{1i} + \beta_{3,P}X_{2i}\}$. Through (9) $P(Y^*_{0,ik} > 0, S_i > t_{0ik})$ is computed via $\mu_{0,ik} \times P(S_i > t_{0ik})$ and we find $\{\mu^*_{0,il}\}_{l=1}^{21}$ using the bisection method. The

generated $(Y_{0,i1},\ldots,Y_{0,i21})^{\mathrm{T}}$ vector is truncated such that $t_{0,ik} \leq S_i$ to create the observed outcomes for $k = 1,\ldots,N_{0i}$.

For subjects who experience an infection-related hospitalization, based on the previously generated Z_i, we generate $(Y^*_{0,i1},\ldots,Y^*_{0,iN_{0i}},Y^*_{1,i1},\ldots,Y^*_{1,i20},S_i)^{\mathrm{T}} \sim N_{N_{0i}+20}([\mu^*_{0,i1},\ldots,\mu^*_{0,iN_{0i}},$ $\mu^*_{1,i1},\ldots\mu^*_{1,i20},E(S_i)=3.38+Z_i]^{\mathrm{T}},\Sigma)$ where $N_{0i}=4Z_i+1$, $\Sigma=\left[I_a,-.05\eta_a;-.05\eta_a^{\mathrm{T}},.5\right]$ with $a=N_{0i}+20$. The unconditional means $(\mu^*_{0,i1},\ldots,\mu^*_{0,iN_{0i}})$ are calculated as described in (9) and $(\mu^*_{1,i1},\ldots,\mu^*_{1,i20})$ are calculated similarly by the bisection method using $P(Y^*_{1,ik'}>0,S_i>$ $Z_i+t_{1ik'})=\mu_{1,ik'}\times P(S_i>Z_i+t_{1ik'})$ where $\mu_{1,ik'}=g^{-1}\{\alpha_{1,P}(t_{1ik'})+\beta_{1,P}Z_i+\beta_{2,P}X_{1i}+\beta_{3,P}X_{2i}\}$. The generated $(Y_{1,i1},\ldots,Y_{1,i21})^{\mathrm{T}}$ vector is truncated such that $Z_i+t_{1,ik'}\leq S_i$ to create the observed outcomes after the pivotal exposure for $k'=1,\ldots,N_{1i}$.

4.2 Simulation Results

4.2.1 Estimation

We generated 200 datasets at sample sizes of $n=500$ and 2000. For the estimation, bandwidths were chosen by 20-fold cross-validation as described in Cai et al. (2000). Bandwidths utilized were chosen in a preliminary simulation study yielding $h=(1.5,1.5)$ for $\hat{\alpha}_{0,P}(t_0)$, $\hat{\alpha}_{1,P}(t_1)$ at $n=(500,2000)$, respectively. To study the performance of the proposed estimation procedure, we utilize a relative mean squared deviation error (MSDE) defined as

$$\mathrm{MSDE}_{\alpha_v}=\left[\int_0^T\{\alpha_{v,P}(t_v)-\widehat{\alpha}_{v,P}(t_v)\}^2dt_v\right]\Big/\int_0^T\alpha^2_{v,P}(t_v)dt_v$$

for the VCFs, $v=0,1$, and mean squared error MSE_{β_r} for the constant coefficients $\{\beta_{r,P}\}_{r=1}^3$. The median and first and third quartiles of the estimated MSDE and MSE measures over 200 Monte Carlo runs are presented in Table S1 of the supplemental Appendix. The MSDE and MSE values are relatively small and decrease with increasing sample size, indicating the overall effectiveness of the estimation in targeting partly conditional PL-GVCMs using longitudinal data truncated by death (at 80%; results are similar for other levels of truncation by death). In

addition, Figure S1 of the supplemental Appendix displays the estimated median and 5th and 95th percentiles of the VCF estimates along with the true curves for $n = 2000$. The estimated functions track the true VCFs.

4.2.2 Hypothesis Tests

We also examine the validity and power of the two proposed GLRTs, namely Test I: H_0 : $\alpha_{0,P}(t_0) = c_0$ and $\alpha_{1,P}(t_1) = c_1$ and Test II: H_0 : $\alpha_{0,P}(t_0) = \alpha_{1,P}(t_1)$ (illustrated in Figure 6(a)-(b), respectively) for longitudinal data under high levels of truncation by death, similar to our data application (ranging from 40-80%).

We first study the Wilks phenomenon under the high level of truncation by death (at 80%), that the null distribution of the test statistic approximately follows a χ^2-distribution and does not depend on the specific null values considered. For Test I, we consider 5 different sets of null values: $(c_0, c_1) \in \{(-1, 1), (-1, 0), (0, -1), (0, 1), (1, 0)\}$. The parametric bootstrap procedure (Section 2.3) is used for $n = 500$ to estimate the null distribution of the test statistic under these 5 settings. The estimated densities of the GLRT statistic, T, based on $B = 500$ bootstrap samples are given in Figure 6(c) along with the density of the χ^2-distribution. The degrees of freedom of the χ^2-distribution is chosen to be close to the sample mean of the bootstrap test statistic values across null configurations. The plotted densities of T are close to the χ^2 density, indicating that the Wilks phenomenon holds for the partly conditional PL-GVCMs under substantial truncation by death.

Next, we study the power and validity of the two proposed hypothesis tests. For Test I, the power is evaluated at a sequence of alternatives indexed by δ: $H_1 : \alpha_{0,P}(t_0) = c_0(1-\delta)+\delta\alpha_0^0(t_0)$ and $\alpha_{1,P}(t_1) = c_1(1 - \delta) + \delta\alpha_1^0(t_1)$ where $\alpha_0^0(t) = -.05t^2 + .025t - 1.25$, $\alpha_1^0(t) = -.03t^2 - .05t$, $\delta \in [0, 1]$, $c_0 = E[\alpha_0(t_0)]$ and $c_1 = E[\alpha_1(t_1)]$. Similarly, for Test II, we consider the alternative $H_1 : \alpha_{0,P}(t_0) = (1 - \delta)\alpha_0^0(t_0) + \delta\alpha_0^0(t_0)$ and $\alpha_{1,P}(t_1) = (1 - \delta)\alpha_0^0(t_1) + \delta\alpha_1^0(t_1)$ where $\alpha_0^0(t) = -.05t^2+.025t-1.25$, $\alpha_1^0(t) = -.03t^2-.05t$ and $\delta \in [0, 1]$. Note that in both cases, larger values

of δ correspond to further deviations from the null. Figure 6(d) gives the 3 power curves, at level .05, for 80%, 60% and 40% truncation by death for Test I. Results are presented based on 200 replications at $n = 500$. Similarly Figure 6(e) gives the 3 power curves for Test II. As expected, the power increases with effect size ($\delta \uparrow$) and the power degrades with increasing level of truncation by death. The validity of a test is indicated by the empirical power under the null ($\delta = 0$, Type I error), which should coincide approximately with the level of the test. For Test I at significance levels (.05, .1, .2, .5), the corresponding empirical Type I errors are (.06, .13, .23, .56) for 80% truncation by death. The results are similar for 60% and 40% truncation by death: (.04, .12, .24, .51) and (.06, .11, .19, .53). Similarly, validity of Test II holds, indicated by the following empirical Type I error rates: (.04, .1, .19, .48), (.04, .11, .19, .47) and (.04, .09, .17, .48) for 80%, 60% and 40% truncation by death.

5 Discussion

In this work, we proposed partly and fully conditional approaches to modeling time-varying effects for longitudinal data with substantial truncation by death. We provided an in-depth comparative study of these conditional modeling approaches with applications to further understand the time-varying effect of infection on patients' CV outcome trajectories over time, from the start of dialysis. While the partly conditional approach provides information on an evolving/dynamic cohort of survivors, the fully conditional approach conditions on the actual death time where the analysis involves fitting a sequence of stratum-specific time-varying effect models (within death bins). Thus, the later approach enables direct comparison of time-varying effects for each death bin/stratum as well as variation in baseline covariate effects on outcome across death bins. We note that another approach for longitudinal data truncated by death is joint modeling of CV risk and survival. The joint modeling typically focuses on the survival outcome (and the longitudinal outcome) and is not suitable to our application of modeling

the infection-CV outcome relationship. For inference via hypothesis testing, we proposed an extension of the GLRT statistic to longitudinal data with substantial truncation by death, like the dialysis population. Empirical estimate of power and validity via simulation shows the efficacy of the proposed tests. We provide R codes for the proposed partly conditional and fully conditional PL-GVCM at http://dsenturk.bol.ucla.edu/PLVCM_algorithm_JASA.pdf.

References

Ahmad, I., Leelahanon, S., and Li, Q. (2005). "Efficient Estimation of a Semiparametric Partially LinearVarying Coefficient Model." *Annals of Statistics*, 33, 258-283.

Cai, Z., Fan, J. and Li, R. Z. (2000). "Efficient estimation and inferences for varying-coefficient models." *Journal of the American Statistical Association*, 95, 888-902.

Chiang, C. T., Rice, J. A. and Wu, C. O. (2001). "Smoothing spline estimation for varying coefficient models with repeatedly measured dependent variables." *JASA*, 96, 605-619.

Cleveland, W. S., Grosse, E. and Shyu, W. M. (1991). "Local regression models." *In Statistical Models in S (Chambers, J. M. and Hastie, T. J., eds)*, 309-376. Wadsworth & Brooks, Pacific Grove.

Dalrymple, L. S., Mohammed, S. M., Mu, Y., Johansen, K. L., Chertow, G. M., Grimes, B., Kaysen, G. A. and Nguyen, D. V. (2011). "The risk of cardiovascular-related events following infection-related hospitalizations in older patients on dialysis." *Clinical Journal of the American Society of Nephrology*, 6, 1708-1713.

Estes, J. P., Nguyen, D. V., Dalrymple, L. S., Mu, Y. and Senturk, D. (2014). "Cardiovascular Event Risk Dynamics Over Time in Older Patients on Dialysis: A Generalized Multiple-Index Varying Coefficient Model Approach." *Biometrics* in press.

Fan, J. and Huang, T. (2005). "Profile likelihood inferences on semiparametric varying-coefficient partially linear models." *Bernoulli*, 11, no. 6, 1031-1057.

Fan, J. and Zhang, J. T. (2000). "Two-step estimation of functional linear model with application to longitudinal data." *JRSSB*, 62, 303-322.

Fan, J., Zhang, C. and Zhang, J. (2001). "Generalized likelihood ratio statistics and wilks phenomenon." *Annals of Statistics*, 29, 153-193.

Hastie, T. and Tibshirani, R. (1993). "Varying coefficient models." *JRSSB* 55, 757-796.

Hoover, D. R., Rice, J. A., Wu, C. O. and Yang, L. P. (1998). "Nonparametric smoothing estimates of time-varying coefficient models with longitudinal data." *Biometrika*, 85, 809-822.

Huang, J. Z., Wu, C. O. and Zhou, L. (2002). "Varying-coefficient models and basis function approximations for the analysis of repeated measurements." *Biometrika*, 89, 111-128.

Huang, J. Z., Wu, C. O. and Zhou, L. (2004). "Polynomial spline estimation and inference for varying coefficient models with longitudinal data." *Statistica Sinica*, 14, 763-788.

Kurland, B. F. and Heagerty, P. J. (2005). "Directly parameterized regression conditioning on being alive: Analysis of longitudinal data truncated by deaths." *Biostatistics*, 6, 241-258.

Kurland, B. F., Johnson, L. L., Egleston, B. L. and Diehr, P. H. (2009). "Longitudinal data with follow-up truncated by death: Match the analysis method to research aims." *Statistical Science*, 24, 211-222.

Lu, Y. (2008) "Generalized partially linear varying-coefficient models." *Journal of Statistical Planning and Inference*, 138(4), 901-914.

Mohammed S. M., Dalrymple, L. S., Senturk, D. and Nguyen, D. V. (2013). "Naive hypothesis testing for case series models with time-varying exposure onset measurement error: Inference for infection-cardiovascular risk in patients on dialysis." *Biometrics*, 69, 520-529.

Mohammed, S. M., Senturk, D., Dalrymple, L. S. and Nguyen, D. V. (2012). "Measurement error case series models with application to infection-cardiovascular risk in older patients on dialysis." *JASA*, 107, 1310-1323.

Qu, A. and Li, R. (2006). "Quadratic inference functions for varying coefficient models with longitudinal data." *Biometrics*, 62, 379-391.

Senturk, D., Dalrymple, L. S., Mohammed, S. M., Kaysen, G. A. and Nguyen, D. V. (2013). "Modeling time varying effects with generalized and unsynchronized longitudinal data." *Statistics in Medicine*, 32, 2971-2987.

Senturk, D. and Mueller, H.G. (2009). "Covariate adjusted generalized linear models." *Biometrika*, 96, 357-370.

Senturk, D. and Mueller, H.G. (2010). "Functional varying coefficient models for longitudinal data." *J. Am. Statist. Assoc.*, 105, 1256-1264.

Senturk, D. and Nguyen, D. V. (2011). "Varying coefficient models for sparse noise-contaminated longitudinal data." *Statistica Sinica*, 21, 1831-1856.

United States Renal Data System (2013). *USRDS 2013 Annual Data Report: Atlas of Chronic Kidney Disease and End-Stage Renal Disease in the United States*, National Institutes of Health, National Institute of Diabetes and Digestive and Kidney Diseases, Bethesda, MD.

Wu, C. O. and Chiang, C. T. (2000). "Kernel smoothing on varying coefficient models with longitudinal dependent variable." *Statistica Sinica*, 10, 433-456.

Xia, Y., Zhang, W. and Tong, H. (2004). "Efficient estimation for semivarying coefficient models." *Biometrika*, 91, 661-681.

Zhang D. (2004). "Generalized linear mixed models with varying coefficients for longitudinal data." *Biometrics*, 60, 8-15.

Zhang,W., Lee, S.Y., Song, X., (2002). "Local polynomial fitting in semivarying coefficient models." *J. Multivariate Annal*, 82, 168-188.

Table 1: Baseline characteristics of $n = 243,730$ patients aged 65 to 90. Data presented are mean ± standard deviation (SD) for continuous variables or count (percent) for categorical variables.

Variable	Mean ± SD/ Count (%)
Baseline age	75.78 ± 6.25
Male	125,875 (52)
Race	
Black	53,704 (22)
White	176,780 (73)
Other	13,246 (5)
Congestive heart failure	100,896 (41)
Coronary heart disease	87,532 (36)
Peripheral vascular disease	46,357 (19)
Diabetes	138,682 (57)
Estimated glomerular filtration rate	10.923 ± 5.445
Body mass index	26.973 ± 6.783

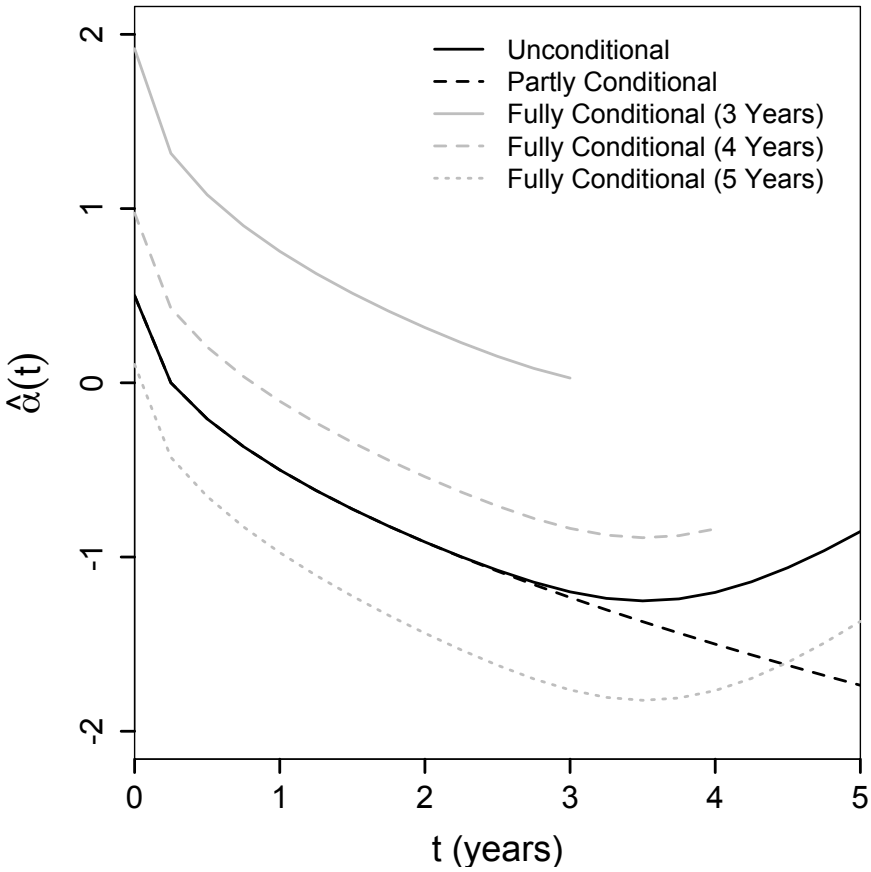

Figure 1: Illustration of partly conditional, fully conditional and unconditional model estimates of the varying coefficient function targets in a simple generalized varying coefficient model.

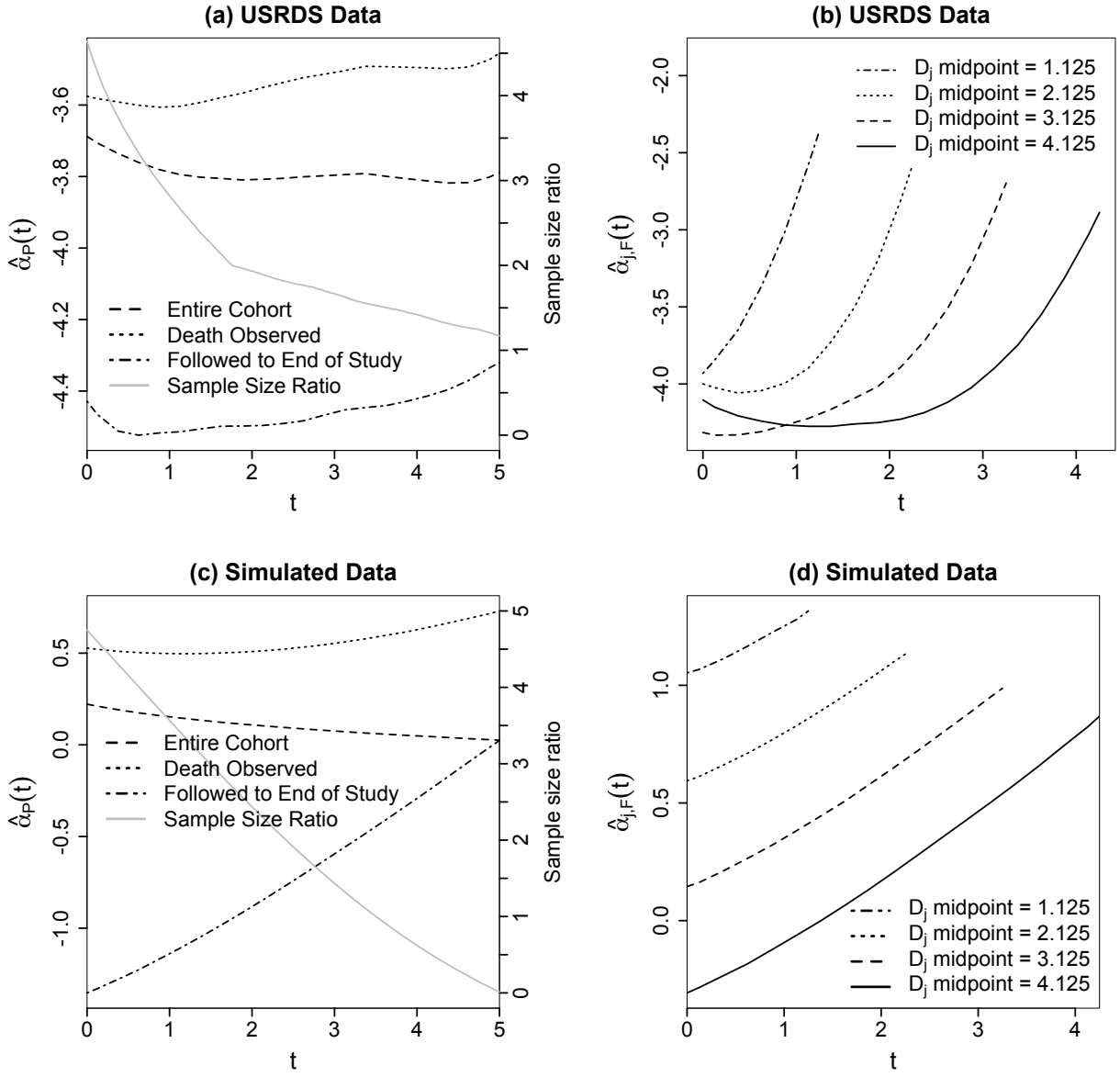

Figure 2: (a) Fits from a simple partly conditional GVCM, $g[E\{Y_i(t_i)|S_i > t_i\}] = \alpha_P(t_i)$ using 3 USRDS cohorts. Also displayed (gray line) is the sample size ratio for the cohort whose death is observed over the cohort who were followed to the end of the study. (b) Fits from a fully conditional GVCM, $g[E\{Y_i(t_i)|S_i \in D_j\}] = \alpha_{j,F}(t_i)$, for subjects in 3-month death bins with midpoints $1.125, 2.125, 3.125$, and 4.125 years. (c and d) Fits from simulated data under the simple partly and fully conditional GVCMs. Presented are the cross-sectional median varying coefficient function estimates over 200 Monte Carlo runs.

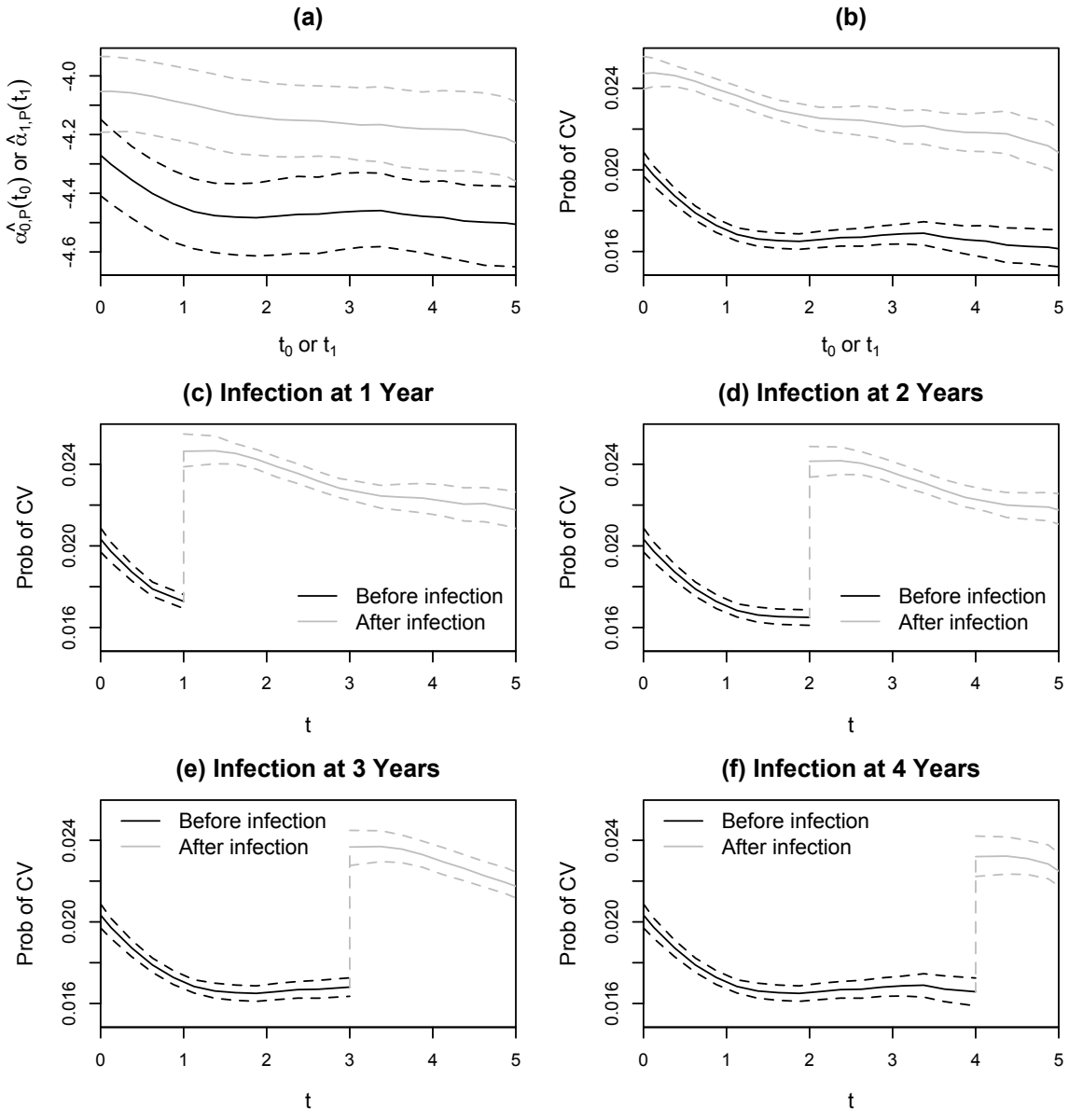

Figure 3: (a) Estimated varying coefficient functions from partly conditional PL-GVCM fits $\widehat{\alpha}_{0,P}(t_0)$ (black), $\widehat{\alpha}_{1,P}(t_1)$ (gray). (b) Estimated CV risk since initiation of dialysis (black) and since the initial infection-related hospitalization (gray) for a white diabetic male who initiated dialysis at age 75.5 with a median levels of eGFR and BMI (9.83 and 25.81, respectively). (c)-(f) Estimated CV risk trajectories for an adult described above where the patient experiences the initial infection-related hospitalization at $1-4$ years after initiation of dialysis. 90% bootstrap confidence intervals given as dashed lines.

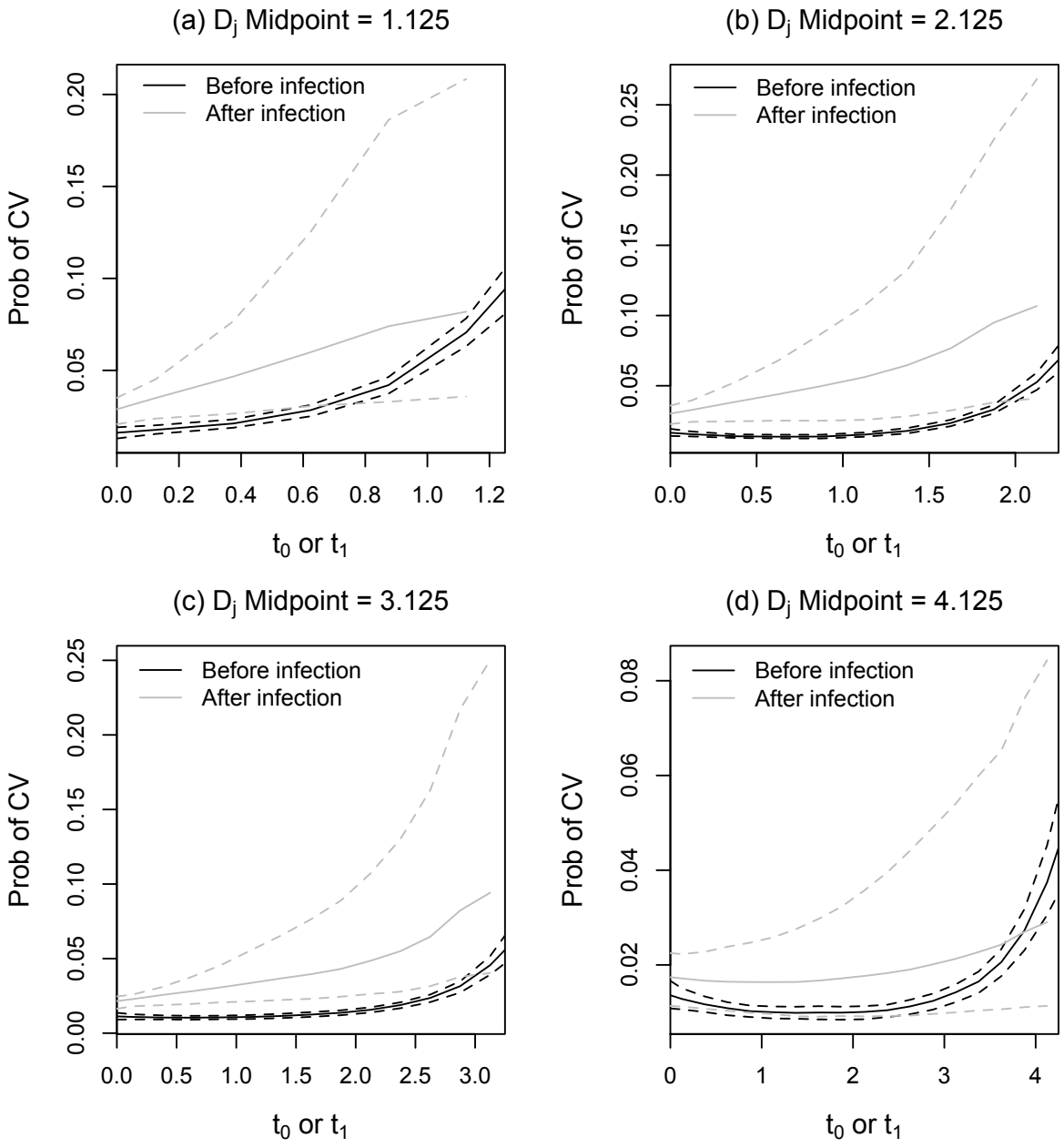

Figure 4: (a)-(h) Estimated CV risk based on the fully conditional PL-GVCM fits from 3-month death bins with midpoints 1.125, 2.125, 3.125 and 4.125, respectively, for a white male diabetic initiating dialysis at age 75.25 with a median levels of eGFR and BMI (9.79 and 26.05, respectively). Time of the initial infection-related hospitalization (vintage) was selected as the median value within each death bin at 0.90., 1.62, 2.06 and 2.43, respectively. 90% bootstrap confidence intervals given as dashed lines.

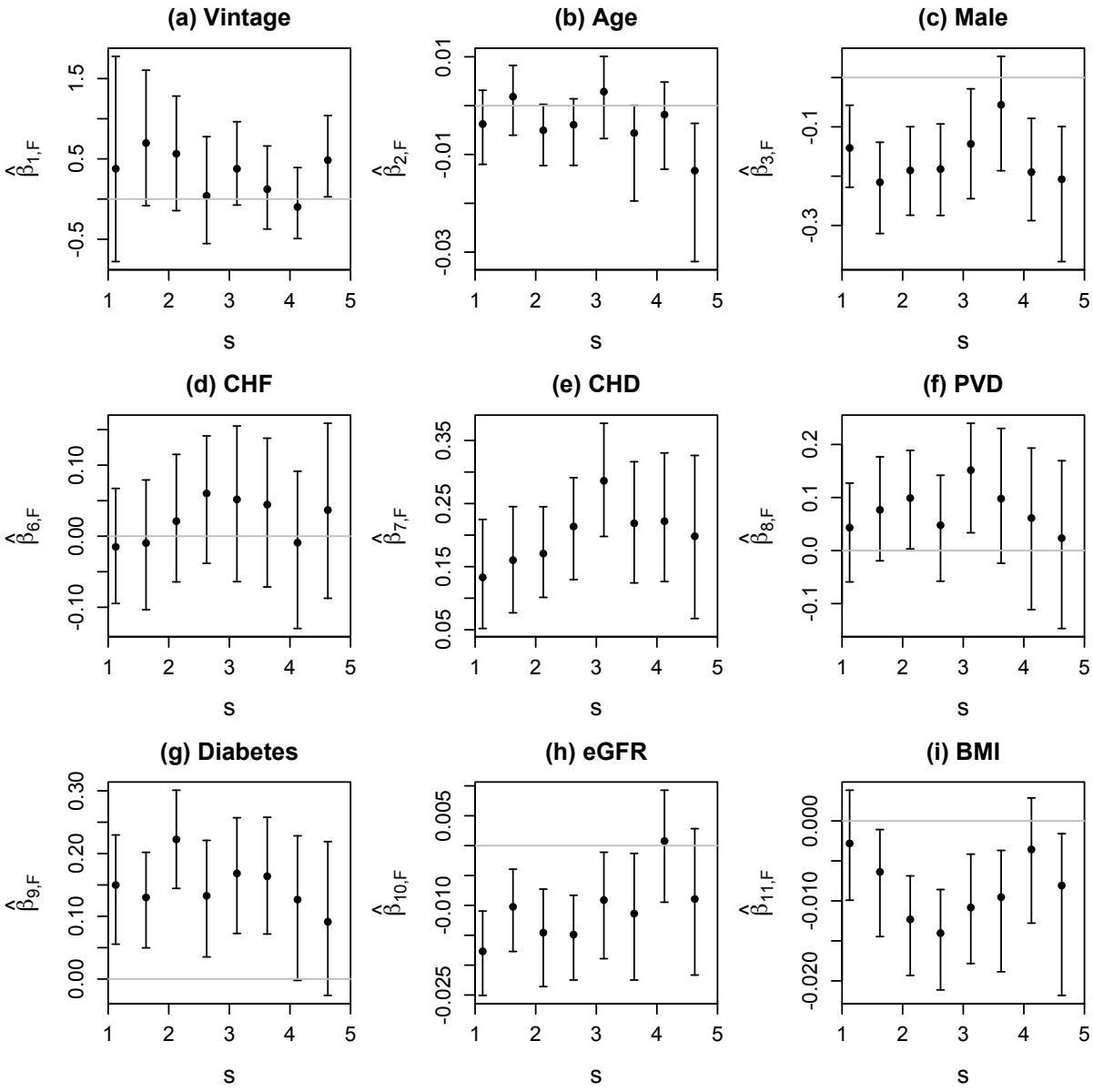

Figure 5: Estimated coefficients for baseline covariates (a) vintage, (b) age, (c) gender-male, (d) congestive heart failure, (e) coronary heart disease, (f) peripheral vascular disease, (g) diabetes, (h) eGFR, and (i) BMI for a sequence of fully conditional PL-GVCMs from death bins with midpoints $D_j = 1.125, 1.625, 2.125, 2.625, 3.125, 3.625, 4.125, 4.625$ years, respectively from left to right. 90% bootstrap confidence intervals are displayed as whiskers. The gray horizontal line at zero (no effect) is included for reference.

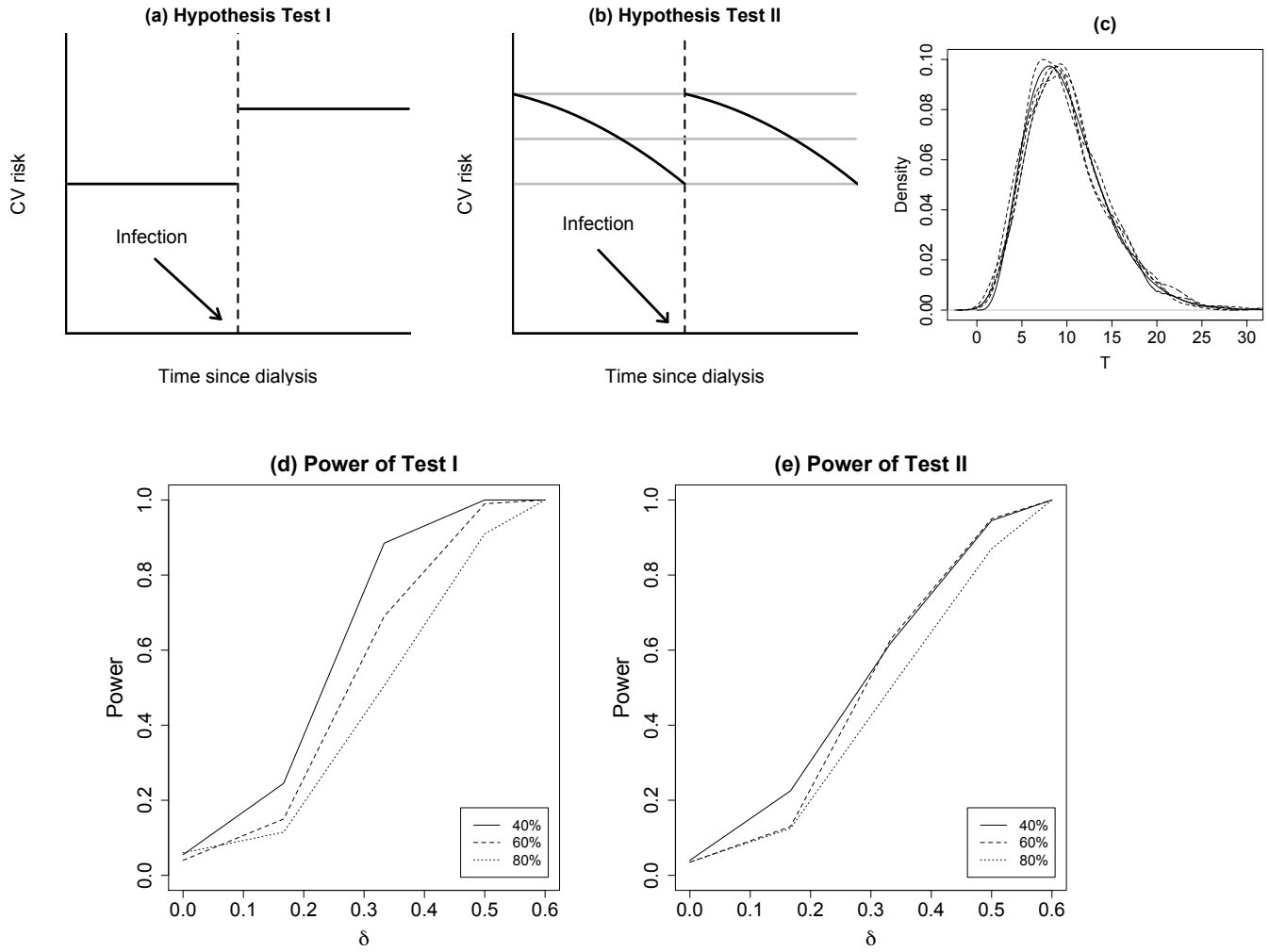

Figure 6: (a)-(b) The null hypotheses of constancy (i.e., Test I: $H_0 : \alpha_0(t_0) = c_0$ and $H_0 : \alpha_1(t_1) = c_1$) and equality (i.e., Test II: $H_0 : \alpha_0(t_0) = \alpha_1(t_1)$) (c) Estimated densities of the generalized likelihood ratio test statistic, T, from 5 different sets of (c_0, c_1) values (dashed) along with the density function of a χ^2-distribution with 10 degrees of freedom (solid). Empirical power estimated at significance level .05 for Test I and II with 40-80% truncation by death.