University of California

Los Angeles

Time Dynamic Modeling and Inference Approaches for Outcomes in Patients on Dialysis

A dissertation submitted in partial satisfaction

of the requirements for the degree

Doctor of Philosophy in Biostatistics

by

Jason Peter Estes

2015

c Copyright by

Jason Peter Estes

2015

Abstract of the Dissertation

Time Dynamic Modeling and Inference Approaches for Outcomes in Patients on Dialysis

by

Jason Peter Estes

Doctor of Philosophy in Biostatistics

University of California, Los Angeles, 2015

Professor Damla Şentürk, Chair

In the first chapter of this work, we characterize the dynamics of cardiovascular event risk trajectories for patients on dialysis while conditioning on survival status via multiple time indices: (1) time since the start of dialysis, (2) time since the pivotal initial infection-related hospitalization and (3) the patient's age at the start of dialysis. This is achieved by using a new class of generalized multiple-index varying coefficient (GM-IVC) models utilizing a multiplicative structure and one-dimensional varying coefficient functions along each time and age index. We develop a two-step estimation procedure for the GM-IVC models based on local maximum likelihood, and report new insights on the dynamics of cardiovascular events risk among the dynamic cohort of survivors using the United States Renal Data System database, which collects data on nearly all patients with end-stage renal disease in the U.S.

In the second chapter of 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. We compare and contrast partly and fully conditional PL-GVCMs in our aforementioned application and develop generalized likelihood ratio tests.

In the third chapter of this work, we introduce a time-varying standardized dynamic ratio (SDR) to aid in the evaluation of a dialysis facility's performance with respect to patient readmission rates as a function of time that patients are on dialysis. The estimation of SDR consists of two steps. First, we model the dependence of readmission events on facilities and patient-level characteristics using a multilevel varying coefficient model (MVCM) with fixed facility time-varying effects, with or without subject random effects. Second, using results from the models, standardization is achieved by computing the ratio of the sum of the predicted number of 30-day readmissions to the sum of the predicted number of 30-day readmissions assuming a reference standard and given the case-mix in that facility. A challenging aspect of our data application is that the number of model parameters is very large, and the estimation of high-dimensional parameters is troublesome. To overcome this problem, we propose a Newton Rhapson algorithm for the MVCM without the random effects, and an approximate EM algorithm for the MVCM with random effects. We propose a test statistic to facilitate in the identification of facilities whose outcomes are outside of normal expectations, and obtain p-values using re-sampling and simulation techniques. Finally, our method of identifying outlier facilities involves converting the observed p-values to Z-statistics and using the empirical null distribution, which accounts for over dispersion in the data.

The dissertation of Jason Peter Estes is approved.

Catherine Sugar

Gang Li

Danh Nyguen

Damla Şentürk, Committee Chair

University of California, Los Angeles

2015

Table of Contents

1 Cardiovascular Event Risk Dynamics Over Time in Older Patients on Dialysis: A Generalized Multiple-Index Varying Coefficient Model Approach . . . 1

1.1 Introduction . . . 1

1.2 Proposed Generalized Multiple-Index Varying Coefficient Model . . . 4

1.2.1 Model Specification . . . 4

1.2.2 Model Interpretation and Assumption . . . 6

1.3 Two-Step Estimation Via Local Maximum Likelihood . . . 8

1.3.1 Step I: Estimation of $_0(t_0)$ and $_1(t_1)$ . . . 9

1.3.2 Step II: Estimation of $_0(a_i)$, $_1(a_i)$ and $_r(a_i)$ . . . 12

1.4 Multiple-Index Cardiovascular Event Risk Trajectories in Older Patients on Dialysis . . . 14

1.4.1 Description of the Study Cohort . . . 14

1.4.2 Results: Cardiovascular Event Risk Trajectories . . . 15

1.5 Simulation Studies . . . 18

1.6 Discussion . . . 21

2 Time-Varying Effect Modeling with Longitudinal Data Truncated by Death: Conditional Models, Interpretations and Inference . . . 23

2.1 Introduction . . . 23

2.2 Partly and Fully Conditional Time-Varying Effect Modeling . . . 27

2.2.1 Model Specification: Conditional PL-GVCM . . . 27

2.2.2 Estimation . . . 29

2.2.3 Generalized Likelihood Ratio Test Under Follow-up Truncation by Death 33
2.3 Applications to Infection-Cardiovascular Risk Modeling 36
2.3.1 Description of the Study Cohort 36
2.3.2 Cardiovascular Outcome Risk Trajectories 37
2.4 Simulation Studies 41
2.4.1 Simulation Model and Design 41
2.4.2 Simulation Results 43
2.5 Discussion 45

3 Time-Dynamic Dialysis Facility Profiling 47
3.1 Introduction 47
3.2 Multilevel Varying Coefficient Modeling With Subject Specific Random Effects 50
3.2.1 Standardized Dynamic Ratio (SDR) 51
3.2.2 Estimation 52
3.2.3 Hypothesis Testing 56
3.3 Applications to USRDS Data 58
3.3.1 Description of the Study Cohort 58
3.3.2 Results 58
3.4 Simulation Studies 61
3.4.1 Simulation Model and Design 61
3.4.2 Hypothesis Test: Validity 63
3.5 Discussion 63

A 65

B 88

B.1 Supplementary Materials: Elaborations on Selection of Number of Bins and Robustness of Data Analysis Results to the Number of Bins Selected 88

C **92**

C.1 Supplementary Materials: Illustration of Different Targets of Inference for a Simple GVCM 92

C.2 Supplementary Materials: Simple Partly Conditional GVCM in Different Cohorts - Figure A.8 93

C.3 Supplementary Materials: Bandwidth Selection for Data Analysis Section 2.3.2.2 93

D **95**

D.1 Supplementary Materials: Details of the Proposed Approximate EM Algorithm of Estimation Section 3.2.2 95

List of Figures

A.1 Example of follow-up data for a subject (a) without and (b) with an infection-related hospitalization along with the proposed models for cardiovascular risk before (light gray) and after (dark gray) the infection-related hospitalization. Note that the model for cardiovascular risk after the initial infection-related hospitalization appropriately accounts for vintage until the infection-related hospitalization (term $_1(a_i)Z_i$). See Section 1.2 for details. 72

A.2 Baseline age-stratified varying coefficient function estimates after the (a) start of dialysis (vintage t_0) and (b) initial infection-related hospitalization (t_1). Final varying coefficient function estimates as a function of (c) vintage and time since the initial infection-related hospitalization and (d) baseline age at dialysis. 90% bootstrap confidence intervals are given as dashed lines in (c) and (d). 73

A.3 Estimated probabilities of cardiovascular (CV) events for white male patients with diabetes and with average levels of eGFR and BMI (with the vintage until the first infection-related hospitalization of $Z = 1{\cdot}4$ years) with baseline ages (a) 65, (b) 78 and (c) 90. Plot (d) overlays/combines the estimated probability trajectories from the three baseline ages. 90% bootstrap confidence intervals are given as dashed lines in (a), (b), and (c). 74

A.4 Estimated probabilities of cardiovascular (CV) events during the course of dialysis for patients experiencing the pivotal initial infection-related hospitalization at 3, 2 and 1 year after the start of dialysis with baseline ages of 65, 78 and 90 (columns left, middle and right, respectively). 90% bootstrap confidence intervals are given dashed. 75

A.5 Age-varying effects of baseline covariates. Given are estimated baseline age-varying coefficient functions for (a) vintage up to first infection, (b) sex - male, (c) race - black, (d) race - other, (e) congestive heart failure, (f) coronary heart disease, (g) peripheral vascular disease, (h) diabetes, (i) eGFR, and (j) BMI. 90% bootstrap confidence intervals are given as dashed lines. The reference horizontal line at zero indicates no effects. 76

A.6 Simulation results for $n = 3000$. The cross-sectional median curves of the proposed estimates are given along with 5% and 95% cross-sectional percentiles (dotted) overlaying the true varying coefficient functions (solid). 77

A.7 Illustration of partly conditional, fully conditional and unconditional model estimates of the varying coefficient function targets in a simple generalized varying coefficient model. 78

A.8 (a) Fits from a simple partly conditional GVCM, $g[E\,fY_i(t_i)jS_i > t_i g] = \;_P(t_i)$ along with 90% bootstrap CIs, 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, $g[E\,fY_i(t_i)jS_i\ 2\ D_j g] = \;_{j;F}(t_i)$, 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 time-invariant median varying coefficient function estimates over 200 Monte Carlo runs. 79

A.9 (a) Estimated varying coefficient functions from partly conditional PL-GVCM fits $b_{0;P}(t_0)$ (black), $b_{1;P}(t_1)$ (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. 80

A.10 (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. 81

A.11 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. The x-axis s denotes the midpoints of the death bins D_j. 82

A.12 (a)-(b) The null hypotheses of constancy (i.e., Test I: $H_0 : \ _0(t_0) = c_0$ and $H_0 : \ _1(t_1) = c_1$) and equality (i.e., Test II: $H_0 : \ _0(t_0) = \ _1(t_1)$) (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 ^{2}-distribution with 10 degrees of freedom (solid). (d)-(e) Estimated power for both tests at significance level $.05$ with 40-80% truncation by death. . . . 83

A.13 The median, 5th and 95th percentiles of the estimated (a) varying coefficient functions and (b) regression parameters in the partly conditional PL-GVCM from simulation studies of Section 2.4 at $n = 2000$ over 200 Monte Carlo runs. 84

A.14 L2-norm of the difference between the estimated facility $SDR_i(t)$ functions under Models 1 and 2 divided by the L2-norm of $SDR_i(t)$ under Model 1 stratified by facility size. 85

A.15 Histogram of Z scores based on Model 1 by facility size (a) small, (b) medium, and (c) large. The $N(0; 1)$ density is superimposed on the histograms (grey) along with a normal curve fitted to the center of the histograms using a robust M-estimation procedure. 86

A.16 Flowchart depicting inclusion/exclusion rules. 87

B.1 Varying coefficient function estimates corresponding to fits with different total number of bins. 90

B.2 Estimated age-varying coefficient functions corresponding to baseline covariates from fits with different total number of bins. 91

List of Tables

A.1 Baseline characteristics of $n = 294{,}511$ patients aged 59 to 96 used in Section 1.4.1. Data presented are mean standard deviation (SD) for continuous variables or count (percent) for categorical variables. 65

A.2 Relative mean squared deviation error of the estimated varying coefficient functions from simulation studies in Section 1.5. Median and 25th and 75th percentiles of the deviation measures are presented based on 200 Monte Carlo runs. 66

A.3 Baseline characteristics of $n = 243{,}730$ patients aged 65 to 90 used in Section 2.3.1. Data presented are mean standard deviation (SD) for continuous variables or count (percent) for categorical variables. 67

A.4 Summaries of the relative mean squared deviation error of the estimated varying coefficient functions and estimated mean squared error of the regression coefficients in Section 2.4. Median and 25th and 75th percentiles of the deviation measures are presented based on 200 Monte Carlo runs. 68

A.5 Estimated bias, standard error (SE), and mean squared error (MSE) of the estimated parameters from simulation studies of Section 3.4.1 based on 200 Monte Carlo runs. 68

A.6 Relative mean squared deviation error of the estimated time-varying facility effect functions and standard dynamic readmission ratios stratified by facility size and facility effect shape from simulation studies of Section 3.4.1 based on 200 Monte Carlo runs. 69

A.7 Performance of $\widehat{SDR}_i(t)$ in testing $H_0 : {}_1(t) = {}_0(t)$ from Section 3.4.2 where increasing values of indicate deviating further from the null. We provide the median MSDE of $\widehat{SDR}_i(t)$ and the estimated acceptance probability (AP) stratified by , facility size, and model. 70

A.8 Number and percent of significant outlier facilities in our data application in Section 3.3.2 determined by nominal p-value $< .025$. 71

A.9 Number and percent (with respect to facility size) of total facilities flagged as significantly worse or other than the national standard across time in our data application in Section 3.3.2. 71

Acknowledgments

There are many individuals who provided guidance, support, and encouragement during my matriculation at UCLA. Unquestionably, I first want to thank my adviser, Dr. Damla Şentürk, and Dr. Danh Nguyen for their guidance, support and motivating work ethic that significantly contributed to my development and growth as a researcher. I want to thank my committee members, Dr. Catherine Sugar and Dr. Gang Li, for their helpful suggestions and sincere interest in my progress. To my friends and family, I thank you for your contributions in making this experience memorable and positive. Finally, I would like to thank Dr. Weiss and Dr. Donatello Telesca for their invaluable time, guidance and many discussions.

Vita

2003	B.S. (Mathematics), California Polytechnic State University, San Luis Obispo.
2006–2007	Teaching Associate, Mathematics Department, California Polytechnic State University, San Luis Obispo.
2007	M.S. (Mathematics), California Polytechnic State University, San Luis Obispo.
2007–2010	Part-time faculty, Mathematics Department, California Polytechnic State University, San Luis Obispo.
2007–2011	Part-time faculty, Mathematics Department, Cuesta College.
2010–2011	Full-time faculty, Mathematics Department, Allan Hancock College.
2011–2015	Teaching Assistant, Biostatistics Department, UCLA.

Publications

Estes, J. P., Nguyen, D. V., Dalrymple, L. S., Mu, Y. and Şentürk, D. (2014), Cardiovascular event risk dynamics over time in older patients on dialysis: A generalized multiple-index varying coefficient model approach. Biometrics, 70: 751–761.

Estes, J. P., Nguyen, D.V., Dalrymple, L. S., Mu, Y. and Şentürk, D. (2015), Time-varying effect modeling with longitudinal data truncated by death: Conditional models, interpretations and inference. Statistics in Medicine, in press.

Bristow, C. C., Adu-Sarkodie Y., Ondondo, R., Bukusi, E. A., Dagnra, C. A., Yi K., Htoo E., Khamsay C., Houng, L., Campuzano, R., **Estes, J.**, Klausner, J. (2014). Multisite Laboratory Evaluation of a Dual Human Immunodeficiency Virus (HIV)/Syphilis Point-of-Care Rapid Test for Simultaneous Detection of HIV and Syphilis Infection. Open Forum Infectious Diseases.

CHAPTER 1

Cardiovascular Event Risk Dynamics Over Time in Older Patients on Dialysis: A Generalized Multiple-Index Varying Coefficient Model Approach

1.1 Introduction

As of 2010, end-stage renal disease affected more than 570,000 adults in the United States. Of these, more than 400,000 were on dialysis, a life-sustaining treatment (United States Renal Data System Annual Data Report [USRDS ADR], 2012). End-stage renal disease is associated with premature death, and cardiovascular disease is the leading cause of death in this population (USRDS ADR, 2012). An area of particular interest is whether infection contributes to the high risk of cardiovascular disease observed in this population, as infections are relatively common in patients on dialysis (Dalrymple et al., 2010; USRDS ADR, 2012). Previous studies have used an interval Poisson model, a Cox proportional hazards model and case-series analysis to support the notion that infection may contribute to a higher risk of cardiovascular disease in both the general population (Smeeth et al., 2004) and in the dialysis population (Foley et al., 2004; Ishani et al., 2005; Dalrymple et al., 2011; Mohammed et al., 2012). However, to date, studies have not fully elucidated how the risk (probability) of cardiovascular events changes over time for patients on dialysis and furthermore how the risk trajectory depends on individual characteristics.

Our primary objective is to understand how the risk of cardiovascular events dynamically evolves over time, and, in particular, how the changes depend simultaneously on multiple key time indices of: (a) time since the start of dialysis (vintage), (b) time since the initial infection-related hospitalization during dialysis and (c) baseline age at dialysis. While the dynamic cardiovascular risk trajectories as a function of the multiple time indices are of main interest, it is also important to characterize the effects of baseline covariates, which

may potentially depend on baseline age. Baseline covariates of interest include demographic characteristics (sex, race), comorbidities (diabetes, coronary heart disease, congestive heart failure, peripheral vascular disease), body mass index (BMI) and estimated glomerular filtration rate (eGFR).

In addition to modeling risk trajectories over multiple indices, an important methodological challenge in the analysis of longitudinal data from USRDS is follow-up truncated by death. This is particularly relevant to the dialysis population because nationally, the annual mortality in the dialysis population is 20-25% (USRDS ADR, 2012). For the analysis of infection and cardiovascular risk in the dialysis population (USRDS data), the predominant dropout is due to death and it is certainly related to cardiovascular events (outcome). A cardiovascular event is defined as myocardial infarction, unstable angina, stroke, or transient ischemic attack; for a more detailed description, see Section 1.4.1. When dropout is due to death, analysis demands careful consideration of the relevant target of inference. Kurland and Heagerty (2005) and Kurland et al. (2009) have considered truncation by death in longitudinal studies of geriatric populations, including studies examining disability or cognitive function outcomes and have proposed a 'partly conditional' target of inference where the analysis is conditional on being alive. Authors argue that an unconditional target of inference as is commonly used in drop-out or missing data literatures may not be a meaningful target when the missing data is due primarily to truncation by death, since it concerns a population where there are no deaths. Instead, more relevant scientific questions can be addressed through a partly conditional model for the dynamic cohort of survivors. Some overall questions of particular clinical relevance for our study include:

(a) What is the cardiovascular risk trajectory during the course of dialysis for the dynamic cohort of survivors and how does it depend on baseline age or other baseline covariates?

(b) What is the cardiovascular risk at, for instance, 2 years after dialysis for patients who survive at least 2 years on dialysis without an infection?

(c) What is the cardiovascular risk for patients who acquired an infection at 2 years after

dialysis and who survive 2 or more years?

To address the aforementioned modeling objectives, we propose generalized multiple-index varying coefficient (GM-IVC) models for generalized outcome data that (a) accommodate several time indices, (b) utilize one dimensional varying coefficient functions along each time index to facilitate ease of interpretability similar to standard varying coefficient models, (c) allow for multiple cross-sectional and longitudinal covariates and (d) target a partly conditional inference, conditional on survival status. It is known that modeling time-varying effects with multiple indices generally is unreasonably difficult because of the curse of dimensionality. To address the curse of dimensionality, we utilize a multiplicative structure for the multiple-index varying effects that is able to capture several time-dynamic cardiovascular risk trajectories. As detailed in Section 1.2.1, the proposed GM-IVC models are adaptive to the time period before and after the pivotal initial infection-related hospitalization. That is, the cardiovascular risk is modeled as a function of vintage for patients who never experience the pivotal infection and for patients who do experience the pivotal infection before their initial infection-related hospitalization. The GM-IVC models then shift to the time period after the initial infection-related hospitalization to estimate the cardiovascular risk as a function of time since the initial infection. Furthermore, since the cardiovascular risk as a function of these two time indices (vintage and time since infection) is associated with baseline age at dialysis, the models allow for baseline age as a third index.

We note that the literature on the standard varying coefficient models (Cleveland et al., 1991; Hastie and Tibshirani, 1993), generalized varying coefficient models (Cai et al., 2000; Zhang et al., 2004; Qu and Li, 2006), and their adaptations for analyzing longitudinal data (e.g., see Hoover et al., 1998; Wu et al., 2000; Chiang et al., 2001; Fan et al., 2000; 2003; Huang et al., 2002; Huang et al., 2004; and references therein) are limited to a single time index for the varying coefficient functions due to the curse of dimensionality. When understanding how the response trajectory changes with respect to each index is not of interest, a dimension reduction approach where a linear combination of several indices serve as a one-dimensional index of the varying coefficient model was proposed by Fan et al. (2003).

Although these approaches are very useful in their respective areas of applications, they are not directly applicable to our objective of modeling/understanding cardiovascular risk over multiple indices.

We also note that although the proposed GM-IVC models are motivated by our goal to better understand the dynamics of cardiovascular risk over several time and age indices for patients on dialysis with the initial infection-related hospitalization as the pivotal exposure, the models are sufficiently general for a variety of other potential applications. In many longitudinal investigations, a pivotal exposure of interest marks the shift to a new 'time' index for modeling the response trajectory. The remainder of this paper is organized as follows. We introduce our proposed generalized multiple-index varying coefficient model along with model interpretation and assumptions in Section 1.2. Section 1.3 outlines the proposed estimation algorithm based on local maximum likelihood. In Section 1.4, we examine the aforementioned cardiovascular risk trajectories in older patients on dialysis with data from the USRDS. Section 1.5 contains simulation studies to demonstrate the efficacy of the proposed estimation method, followed by concluding remarks in Section 1.6.

1.2 Proposed Generalized Multiple-Index Varying Coefficient Model

1.2.1 Model Specification

Let a_i denote the age of the ith patient at the initiation of dialysis and S_i denote the survival time of the ith patient. While t_i will be used to denote overall follow-up times after initiation of dialysis, t_{0i} and t_{1i} will specifically track follow-up times before and after the potential infection-related hospitalization, respectively. Hence for patients who had a pivotal initial infection-related hospitalization at time Z_i, we note that $t_i = t_{0i} \mathbf{I}\{t_i < Z_i\} + (Z_i + t_{1i}) \mathbf{I}\{t_i \geq Z_i\}$, where $\mathbf{I}\{A\}$ denotes the indicator function for event A. For patients who do not experience a pivotal infection during follow up time, $Z_i = 0$ and $t_i = t_{0i}$. To

examine the changes in cardiovascular event probability (risk) while conditioning on survival status over these time indices, we model the binary indicator of having a cardiovascular event within a three month follow-up interval. We consider a binary outcome instead of a count outcome in our modeling, since having more than one cardiovascular event in a three month interval is very rare; it is less than 0.1% in our data. The goal is to model the expected outcome, denoted

$$_{i} \qquad _{i}(a_i; t_i; t_{0i}; t_{1i}) = E\, f Y_i(a_i; t_i; t_{0i}; t_{1i}) j Z_i; X_i; \mathsf{I}_{\mathsf{P}_i}(t_i); S_i > t_i g;$$

where $Y_i(a_i; t_i; t_{0i}; t_{1i})$ is the indicator of a cardiovascular event for subject i in a three month time interval centered around a fixed value of t_{0i} or t_{1i}; Z_i is the vintage till first infection-related hospitalization given that the ith patient has at least one infection-related hospitalization ($Z_i = 0$ for patients who do not experience an infection-related hospitalization); $\mathsf{I}_{\mathsf{P}_i}(t_i)$ denotes a subject-specific time-varying indicator of infection-related hospitalization *prior* to time t_i (i.e. equals 1 for $Z_i > t_i$ and zero otherwise); X_i is a vector of $p \quad 1$ additional baseline covariates. A link (transformation) function, denoted $g(\ _i)$, connects the conditional expected outcome (cardiovascular event risk) to the time-varying effects and age-varying effects corresponding to the multiple time indices and the covariates. More precisely, our proposed generalized multiple-index varying coefficient model has the form:

$$g(\ _i) = \ _0(a_i)\ _0(t_{0i}) f 1 \quad \mathsf{I}_{\mathsf{P}_i}(t_i) g + \ _1(a_i)\ _1(t_{1i}) \mathsf{I}_{\mathsf{P}_i}(t_i) + \ _1(a_i) Z_i \mathsf{I}_{\mathsf{P}_i}(t_i) + \sum_{r=2}^{p} \ _r(a_i) X_{ri}; \quad (1.1)$$

where the term $_0(a_i)\ _0(t_{0i})$ jointly captures vintage- and age-varying effects; the term $_1(a_i)\ _1(t_{1i})$ captures the age- and time-varying effects since the initial infection-related hospitalization; the age-varying coefficient functions $_r(a_i)$, $r = 1; : : : ; p$, correspond to vintage prior to the first infection-related hospitalization and baseline covariates. For formality, the supports for the varying coefficient functions in (1.1) are: $t_{0i}\ 2\ [0; T_{0i}]$, $t_{1i}\ 2\ [0; T_{1i}]$, $T_{0i} \quad T; T_{1i} \quad T$ and $a_i\ 2\ [A_0; A_1]$, where T is the maximum follow-up duration along each time axis. In our application, we model the cardiovascular event risk during a maximum follow-up period along each time axis with $T = 5$ years, both after the initiation of dialysis and after the initial infection-related hospitalization. We estimate the age-varying effects

for $a_i \in [65; 90] = [A_0; A_1]$. The target population is older patients on dialysis since the cardiovascular event probability is expected to be higher in this cohort.

Note that for our application, the outcome $Y_i(a_i; t_i; t_{0i}; t_{1i})$ is binary so that $\mu_i = \Pr\{Y_i(a_i; t_i; t_{0i}; t_{1i}) = 1 | Z_i; X_i; I_{P_i}(t_i); S_i > t_i\}$ and we use the logit link function, $\text{logit}(\mu_i) = \log\{\mu_i/(1-\mu_i)\}$. Finally, we note that a classical generalized varying coefficient model with a single time index is a special case of our proposed GM-IVC model (1.1). More specifically, when $\beta_0(t_{0i})$ and $\beta_1(t_{1i})$ are constant functions, model (1.1) reduces to a standard generalized baseline age-varying coefficient model with cross-sectional covariates (X_{ri}'s) and longitudinal covariates (namely, $I_{P_i}(t_i)$ and $Z_i I_{P_i}(t_i)$),

$$g(\mu_i) = \beta_0(a_i) + \{\beta_1(a_i) - \beta_0(a_i)\} I_{P_i}(t_i) + \beta_1(a_i) Z_i I_{P_i}(t_i) + \sum_{r=2}^{K} \beta_r(a_i) X_{ri}.$$

Other simplifications such as parametric forms can be considered for the varying coefficient functions. The proposed model is given for the most general setting for potentially complex features of the varying coefficient functions in diverse applications allowing for nonparametric forms along each time index.

1.2.2 Model Interpretation and Assumption

The proposed GM-IVC model (1.1) adapts to the follow-up time periods of patients before and after a potential infection-related hospitalization in order to model changes in the cardiovascular event risk over several time and age indices. This aspect is illustrated in Figure A.1. With respect to vintage, model (1.1) reduces to

$$g(\mu_i) = \beta_0(a_i) \cdot \beta_0(t_{0i}) + \sum_{r=2}^{K} \beta_r(a_i) X_{ri}. \tag{1.2}$$

Note that for modeling infection-free vintage, times from subjects who had no infection-related hospitalization during their entire follow-up and times prior to the initial infection-related hospitalization for patients with at least one infection-related hospitalization contribute to model (1.2). Hence, before a potential infection-related hospitalization, the risk of a cardiovascular event is modeled as a function of baseline age a_i (at the start of dialysis),

vintage t_{0i} and baseline covariates, whose effects are allowed to vary with baseline age. On the other hand, for those subjects with at least one infection-related hospitalization, after their initial infection-related hospitalization, model (1.1) shifts to

$$g(\ \ _i) = \ _1(a_i)\ \ _1(t_{1i}) + \ _1(a_i)Z_i + \sum_{r=2}^{X} \ _r(a_i)X_{ri}: \tag{1.3}$$

Therefore, after the initial infection-related hospitalization, we model the cardiovascular event risk primarily as a function of baseline age a_i and time since the initial infection-related hospitalization t_{1i} (along with baseline covariates). Thus, the infection-related hospitalization introduces an additional time index, namely time since the initial infection-related hospitalization. Note that model (1.3) also accounts for vintage till the initial infection-related hospitalization (Z_i). The time varying indicator, namely $I_{P_i}(t_i)$, in model (1.1) allows the switch between models (1.2) and (1.3) determined by the time of the initial infection-related hospitalization during a patient's course of dialysis. This flexibility will allow us to study the longitudinal effects of a pivotal initial infection-related hospitalization on the cardiovascular event risk and also compare these effects with the longitudinal effects along the time since dialysis index. In this respect the proposed model does have similarities with a change point varying coefficient model with subject specific change points at the potential initial infection-related hospitalizations. However the main innovation of the proposed GM-IVC model remains in that it can accommodate multiple time indices which is also novel in the change point models for survival or longitudinal data.

Model (1.1) addresses the curse of dimensionality from accommodating multiple time indices via the multiplicative forms $\ _0(a)\ \ _0(t_0)$ and $\ _1(a)\ \ _1(t_1)$. Hence, effects along the two time indices and age are modeled through one-dimensional coefficient functions, rather than bivariate varying coefficient functions (e.g., $h_0(a; t_0)$ and $h_1(a; t_1)$); this leads to easier interpretation and more straight forward comparisons along different time indices. The proposed multiplicative forms in model (1.1) are not identifiable without restrictions, hence we assume the following identifiability conditions:

$$\int_0^T \ ^2_0(t_0)dt_0 = 1; \quad \int_0^T \ ^2_1(t_1)dt_1 = 1; \quad \ _0(0) > 0 \ \text{ and } \ \ _1(0) > 0: \tag{1.4}$$

These identifiability conditions imply that the estimated effects along the time indices, t_0 and t_1, are normalized and that the estimated coefficients, $_0(\ )$ and $_1(\ )$, carry the shapes of the regression effects, while the magnitude and the sign of the effects are reflected through the coefficient functions, $_0(a)$ and $_1(a)$. We note that the assumed multiplicative forms along with the proposed identifiability conditions imply that the cardiovascular risk probabilities as a function of vintage and time since the initial infection-related hospitalization, for patients initiating dialysis at different ages, share a common shape captured by $_0(t_0)$ and $_1(t_1)$, respectively. Also, the magnitude of these trajectories are allowed to change as functions of baseline age at dialysis. We will illustrate in the analysis of the USRDS data in Section 1.4 that the plausibility of the assumed multiplicative forms can be easily assessed graphically during the implementation of the proposed estimation algorithm.

1.3 Two-Step Estimation Via Local Maximum Likelihood

For estimation in the GM-IVC model, we propose a two-step estimation algorithm that utilizes an extension of the local maximum likelihood estimator of Cai et al. (2000) to longitudinal data. In the first step, we target $_0(t_0)$ and $_1(t_1)$ based on the observation that for fixed a, the proposed model reduces to a generalized varying coefficient model in t_0 and t_1, both indexing the longitudinal follow-up of each subject. We bin patients according to baseline age a and obtain stratified estimates of the varying coefficient functions within each bin. This is equivalent to estimating slices of the two dimensional surfaces $h_0(a; t_0) = \ _0(a)\ _0(t_0)$ and $h_1(a; t_1) = \ _1(a)\ _1(t_1)$ at fixed a values. Hence we estimate features of the two-dimensional surfaces by estimation in one dimension. Since the stratified estimates share a common shape according to our identifiability conditions (1.4), we combine and normalize the stratified estimates to obtain our final estimators for $_0(t_0)$ and $_1(t_1)$. The second step of the proposed estimation algorithm utilizes the observation that for known $_0(t_0)$ and $_1(t_1)$, the proposed model reduces to a baseline age-varying coefficient model in a with longitudinal and cross-sectional covariates. Thus, using the estimated $_0(t_0)$ and

$_1(t_1)$ from the first step, we estimate $_0(a)$, $_1(a)$ and $f\ _r(a); r = 1;\cdots;pg$ of the baseline age-varying coefficient model in the second step of the estimation algorithm.

1.3.1 Step I: Estimation of $_0(t_0)$ and $_1(t_1)$

We begin by binning the subjects according to their baseline age a_i. In our application to the USRDS data, we use two year intervals. Denote by $fA_j;\ j = 1;\cdots;Jg$ the disjoint sets of patient indices that partition the cohort. Next, in each age bin A_j, we partition each patient's follow-up period into disjoint three month intervals both after the start of dialysis and after the initial infection-related hospitalization if the patient has at least one infection-related hospitalization. For time since dialysis, patients are followed up to their initial infection-related hospitalization or to the end of follow-up (for patients with no infection-related hospitalization). For time since the initial infection-related hospitalization, patients are followed to the end of their follow-up.

In our application we consider five year maximum follow-up periods, $T = 5$ in model (1.1), both after the start of dialysis and after the initial infection-related hospitalization, since the median follow-up in the entire cohort is approximately 2 years. Define t_{0ik} and t_{1ik^0} to be the midpoints of the kth and k^0th three month time intervals since dialysis start and time since the initial infection-related hospitalization intervals, respectively. Also, let $i(j)$ denote the ith patient in the age bin A_j. We define the binary response variable $Y_{0;ijk}$ $Y_{i(j)}(a_i; t_{i(j)} = t_{0i(j)} = t_{0i(j)k}) = 1$, if the ith patient in baseline age bin A_j had at least one cardiovascular event in the kth three month interval after the start of dialysis. Similarly, $Y_{1;ijk^0}$ $Y_{i(j)}(a_i; t_{i(j)} = Z_{i(j)} + t_{1i(j)}; t_{1i(j)} = t_{1i(j)k^0}) = 1$ if the ith patient in age bin A_j had at least one cardiovascular event in the k^0th three month interval after the initial infection-related hospitalization. In addition, we denote by $X_{ri(j)}$ and $Z_{i(j)}$ the value of the rth baseline covariate and the vintage until the initial infection-related hospitalization of patient i in age bin A_j, respectively.

We note that for a fixed age a, for patients within the age stratum/bin A_j, the proposed

GM-IVC model (1.1) reduces to the following varying coefficient model in the longitudinal follow-up time ($t_{i(j)}$, $t_{0i(j)}$ and $t_{1i(j)}$ all tracking longitudinal time),

$$g(\ _{i(j)}) = \ _{0j}(t_{0i(j)})f1 \quad \mathrm{I}_{\mathrm{P}_{i(j)}}(t_{i(j)})g + \ _{1j}(t_{1i(j)})\mathrm{I}_{\mathrm{P}_{i(j)}}(t_{i(j)}) + b_{1j}Z_{i(j)}\mathrm{I}_{\mathrm{P}_{i(j)}}(t_{i(j)}) + \sum_{r=2} b_{rj}X_{ri(j)}; \tag{1.5}$$

where $_{i(j)} \quad E\,fY_{i(j)}(a_{i(j)}; t_{i(j)}; t_{0i(j)}; t_{1i(j)})jZ_{i(j)}; X_{ri(j)}; \mathrm{I}_{\mathrm{P}_{i(j)}}(t_{i(j)}); S_{i(j)} > t_{i(j)}g$, $g(\ )$ is a known link function, $_{0j}(t_{0i(j)}) \quad _{0}(a_{i(j)})\ _{0}(t_{0i(j)})$, $_{1j}(t_{1i(j)}) \quad _{1}(a_{i(j)})\ _{1}(t_{1i(j)})$, $b_{1j} \quad _{1}(a_{i(j)})$ and $fb_{rj} \quad _{r}(a_{i(j)}); r = 2; \ldots; pg$. The data available for estimation is $f(t_{0i(j)k}; t_{1i(j)k'}; X_{ri(j)}; Z_{i(j)}; Y_{0;i(j)k}; Y_{1;i(j)k'})$; $i \ 2 \ A_j; j = 1; \ldots; J; k = 1; \ldots; N_{0i(j)}; k' = 1; \ldots; N_{1i(j)}g$, where $N_{0i(j)}$ and $N_{1i(j)}$ are the number of three month intervals since the start of dialysis and since the initial infection-related hospitalization, in the follow-up period of patient $i \ 2 \ A_j$, respectively. Using this data we fit model (1.5) by local maximum likelihood. Assuming $_{0j}(t_0)$ and $_{1j}(t_1)$ have continuous second derivatives, we approximate each function locally by $_{0j}(t_0) \quad c_{0j} + c_{1j}(t_0 \quad s_0)$ and $_{1j}(t_1) \quad d_{0j} + d_{1j}(t_1 \quad s_0)$ for t_0 and t_1 in a neighborhood of the fixed time point s_0. Note that higher degree approximations can be accommodated easily in the proposed framework, however local linear approximations are usually enough to capture complex local features. Maximizing the local log-likelihood $\grave{}_n(\mathbf{c}_j)$, defined by

$$\begin{aligned}\grave{}_n(\mathbf{c}_j) \quad = \quad & \frac{1}{\sum_{i\in A_j} N_{i(j)}} \sum_{i\in A_j} \sum_{k=1}^{N_{0i(j)}} \grave{} \ g^{-1} \ c_{0j} + c_{1j}(t_{0i(j)k} \quad s_0) + \sum_{r=2} b_{rj}X_{ri(j)} \ ; Y_{0;ijk} \ K_h(t_{0i(j)k} \quad s_0) \\ & + \sum_{k'=1}^{N_{1i(j)}} \grave{} \ g^{-1} \ d_{0j} + d_{1j}(t_{1i(j)k'} \quad s_0) + b_{1j}Z_{i(j)} + \sum_{r=2} b_{rj}X_{ri(j)} \ ; Y_{1;ijk'} \ K_h(t_{1i(j)k'} \quad s_0) \ ;\end{aligned}$$

results in the local maximum likelihood estimators for the varying coefficient functions, namely $\mathrm{b}_{0j}(t_0) = \mathrm{b}_{0j}$, $\mathrm{b}_{1j}(t_1) = \mathrm{b}_{0j}$ and b_{rj}. The terms in the above local log-likelihood, $\grave{}_n(\mathbf{c}_j)$, are: $K_h(\ ) = K(\ =h)=h$, where $K(\ )$ denotes a kernel function and h is the bandwidth; $\mathbf{c}_j \quad (c_{0j}; c_{1j}; d_{0j}; d_{1j}; b_{1j}; \ldots; b_{pj})^{\mathsf{T}}$; $N_{i(j)} = N_{0i(j)} + N_{1i(j)}$ and $\grave{}(\ ;\ )$ denotes the log-likelihood function. Note that the above likelihood only includes data from subjects who are still alive at t_0 and t_1. This equivalent to including a survival status indicator in the likelihood in the estimation of the partly conditional models as was proposed by Kurland and Heagerty (2005). The maximization can be implemented using the Newton-Raphson algorithm, with the $m+1$ iteration update given by

$$\mathrm{b}_{j;m+1} = \mathrm{b}_{j;m} + \sum_{i2A_j} X_{ij}^{\mathsf{T}} W_{1ij}(\mathrm{b}_{j;m}) X_{ij} \quad {}^{1} \sum_{i2A_j} X_{ij}^{\mathsf{T}} W_{2ij} \hat{r}_{ij}(\mathrm{b}_{j;m});$$

where

$$X_{ij} = \left[\begin{array}{cc|cc|cccc} 1 & (t_{0i(j)1} \quad s_0) & 0 & 0 & 0 & X_{2i(j)} & \cdots & X_{pi(j)} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & & \vdots \\ 1 & (t_{0i(j)N_{0i(j)}} \quad s_0) & 0 & 0 & 0 & X_{2i(j)} & \cdots & X_{pi(j)} \\ \hline 0 & 0 & 1 & (t_{1i(j)1} \quad s_0) & Z_{i(j)} & X_{2i(j)} & \cdots & X_{pi(j)} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 1 & (t_{1i(j)N_{1i(j)}} \quad s_0) & Z_{i(j)} & X_{2i(j)} & \cdots & X_{pi(j)} \end{array}\right]$$

is the predictor matrix of size $N_{i(j)} \quad (p+4)$,

$$\begin{aligned} \boldsymbol{\beta}_{0;ijk} &= g^{-1}\{\boldsymbol{\alpha}_{0j;m} + \boldsymbol{\alpha}_{1j;m}(t_{0i(j)k} \quad s_0) + \sum_{r=2}^{p} \boldsymbol{\theta}_{rj;m} X_{ri(j)}\}; \\ \boldsymbol{\beta}_{1;ijk'} &= g^{-1}\{\boldsymbol{\alpha}_{0j;m} + \boldsymbol{\alpha}_{1j;m}(t_{1i(j)k'} \quad s_0) + \boldsymbol{\theta}_{1j;m} Z_{i(j)} + \sum_{r=2}^{p} \boldsymbol{\theta}_{rj;m} X_{ri(j)}\}; \end{aligned}$$

$W_{1ij}(\boldsymbol{\alpha}_{j;m}) = \mathrm{diag}\{K_h(t_{0i(j)1} \quad s_0)\boldsymbol{\beta}_{0;ij1}(1 \quad \boldsymbol{\beta}_{0;ij1}), \cdots, K_h(t_{0i(j)N_{0i(j)}} \quad s_0)\boldsymbol{\beta}_{0;ijN_{0i(j)}}(1 \quad \boldsymbol{\beta}_{0;ijN_{0i(j)}}), K_h(t_{1i(j)1} \quad s_0)\boldsymbol{\beta}_{1;ij1}(1 \quad \boldsymbol{\beta}_{1;ij1}), \cdots, K_h(t_{1i(j)N_{1i(j)}} \quad s_0)\boldsymbol{\beta}_{1;ijN_{1i(j)}}(1 \quad \boldsymbol{\beta}_{1;ijN_{1i(j)}})\}$, $W_{2ij} = \mathrm{diag}\{K_h(t_{0i(j)1} \quad s_0), \cdots, K_h(t_{0i(j)N_{0i(j)}} \quad s_0); K_h(t_{1i(j)1} \quad s_0), \cdots, K_h(t_{1i(j)N_{1i(j)}} \quad s_0)\}$ and $\tilde{Y}_{ij}(\boldsymbol{\alpha}_{j;m}) = (Y_{0;ij1} \quad \boldsymbol{\beta}_{0;ij1}, \cdots, Y_{0;ijN_{0i(j)}} \quad \boldsymbol{\beta}_{0;ijN_{0i(j)}}; Y_{1;ij1} \quad \boldsymbol{\beta}_{1;ij1}, \cdots, Y_{1;ijN_{1i(j)}} \quad \boldsymbol{\beta}_{1;ijN_{1i(j)}})^{\mathsf{T}}$, for a Bernoulli distributed response. For modeling a Poisson distributed response, $W_{1ij}(\boldsymbol{\alpha}_{j;m}) = \mathrm{diag}\{K_h(t_{0i(j)1} \quad s_0)\boldsymbol{\beta}_{0;ij1}, \cdots, K_h(t_{0i(j)N_{0i(j)}} \quad s_0)\boldsymbol{\beta}_{0;ijN_{0i(j)}}, K_h(t_{1i(j)1} \quad s_0)\boldsymbol{\beta}_{1;ij1}, \cdots, K_h(t_{1i(j)N_{1i(j)}} \quad s_0)\boldsymbol{\beta}_{1;ijN_{1i(j)}}\}$. For subjects who do not have any infection-related hospitalization, the predictor matrix reduces to size $N_{0i(j)} \quad (p+4)$ and sizes of the above quantities adjust accordingly.

Note that the stratified estimators from different A_j's target $\Lambda_{0j}(t_0) \quad \Lambda_0(a_j) \quad \Lambda_0(t_0)$ and $\Lambda_{1j}(t_1) \quad \Lambda_1(a_j) \quad \Lambda_1(t_1)$, and that they share the same shape as $\Lambda_0(t_0)$ and $\Lambda_1(t_1)$, respectively. Hence, to arrive at our final estimators for $\Lambda_0(t_0)$ and $\Lambda_1(t_1)$, we aggregate the stratified estimators coming from different age strata using the identifiability conditions via

$$\Lambda_0(t_0) = \frac{\int_0^5 \{\sum_j \Lambda_{0j}(t_0)\}}{[\int_0^5 \{\sum_j \Lambda_{0j}(t_0)\}^2 dt_0]^{1/2}}(\Lambda \quad 1)^{\prime\prime}; \qquad \Lambda_1(t_1) = \frac{\int_0^5 \{\sum_j \Lambda_{1j}(t_1)\}}{[\int_0^5 \{\sum_j \Lambda_{1j}(t_1)\}^2 dt_1]^{1/2}}(\Lambda \quad 1)^{\prime\prime};$$

where I_0 and I_1 denote the indicator functions for $\sum_j b_{0j}(0) < 0$ and $\sum_j b_{1j}(0) < 0$.

Note that the number of bins selected does not need to be large as long as there is enough sample size to obtain stable estimates from the fitted generalized varying coefficient models in each age bin. There are a couple of factors that play a role in determining adequate sample size for fitting a generalized varying coefficient model: 1) nature of the response (e.g. continuous or binary), 2) number of predictors in the model, 3) amount of truncation by death. For a more detailed discussion of selection of number of bins, readers are referred to Appendix B.1.

1.3.2 Step II: Estimation of $_0(a_i)$, $_1(a_i)$ and $_r(a_i)$

For estimation of the 's and 's, we observe that for known $_0(t_0)$ and $_1(t_1)$, the proposed GM-IVC model reduces to a varying coefficient model in the single age index a. Therefore, we use the estimators of $_0(t_0)$ and $_1(t_1)$ from step I to target the 's and the 's in the baseline age-varying coefficient model

$$g(\ _i) = \ _0(a_i)b_0(t_{0i})f1 \quad I_{P_i}(t_i)g + \ _1(a_i)b_1(t_{1i})I_{P_i}(t_i) + \ _1(a_i)Z_i I_{P_i}(t_i) + \sum_{r=2} \ _r(a_i)X_{ri}; \quad (1.6)$$

in the second step of the estimation algorithm.

We aggregate the indicator responses proposed in step I of the estimation algorithm across the age groups, to obtain the binary indicators $Y_{0;ik} \quad Y_i(a_i; t_i = t_{0i} = t_{0ik})$ and $Y_{1;ik^0} \quad Y_i(a_i; t_i = Z_i + t_{1i}; t_{1i} = t_{1ik^0})$ that will equal one if the ith patient has at least one cardiovascular event during the kth three month interval for time since dialysis or the k^0th three month interval for time since the initial infection-related hospitalization, respectively, and equal zero otherwise. We estimate the 's and and 's in model (1.6) using the aggregate data $f(a_i; Y_{0;ik}; Y_{1;ik^0}; b_{0ik}; b_{1ik^0}; Z_i; X_{ri}); k = 1; \cdots; N_{0i}; k^0 = 1; \cdots; N_{1i}g$ on all subjects where $b_{0ik} \quad b_0(t_{0ik})$ is the value of b_0 at the midpoint of the kth three month interval for time since dialysis, and $b_{1ik^0} \quad b_1(t_{1ik^0})$ is the value of b_1 at the midpoint of the k^0th three month interval for the potential time since the initial infection-related hospitalization. Assuming $_0(a)$, $_1(a)$ and $_r(a)$ have continuous second derivatives, we approximate each function

locally by $_0(a) \quad e_0 + e_1(a \quad a_0)$, $_1(a) \quad f_0 + f_1(a \quad a_0)$ and $_r(a) \quad b_{r0} + b_{r1}(a \quad a_0)$ for a in a neighborhood of the fixed time point a_0.

Similar to the local maximization in step I, maximizing the local log-likelihood, $\grave{}_n(\mathbf{e})$, defined by

$$\grave{}_n(\mathbf{e}) = \frac{1}{\sum_{i=1}^{n} N_i} \sum_{i=1}^{n} \sum_{k=1}^{N_{0i}} \grave{} \; g^{\;1} \;\; (e_0 + e_1\bar{a}_i)\mathrm{b}_{0ik} + \sum_{r=2}^{p} (b_{r0} + b_{r1}\bar{a}_i)X_{ri} \;\; ; Y_{0;ik} \;\; K_h(\bar{a}_i)$$

$$+ \sum_{k^0=1}^{N_{1i}} \grave{} \; g^{\;1} \;\; (f_0 + f_1\bar{a}_i)\mathrm{b}_{1ik^0} + (b_{10} + b_{21}\bar{a}_i)Z_{ri} + \sum_{2=1}^{p} (b_{r0} + b_{r1}\bar{a}_i)X_{ri} \;\; ; Y_{1;ik^0} \;\; K_h(\bar{a}_i) \;\; ;$$

results in the local maximum likelihood estimators for the varying coefficient functions $\mathrm{b}_0(a) = \hat{\mathrm{b}}_0$, $\mathrm{b}_1(a) = \hat{\mathrm{b}}_0$ and $\mathrm{b}_r(a) = \hat{\mathrm{b}}_{r0}$. In the above local likelihood, $K_h(\;) = K(\;=h)=h$, $K(\;)$ denotes a kernel function, h is the bandwidth, $\mathbf{e} \quad (e_0; e_1; f_0; f_1; b_{10}; b_{11}; : : : ; b_{p0}; b_{p1})^{\mathsf{T}}$, $N_i = N_{0i} + N_{1i}$, $\bar{a}_i = (a_i \quad a_0)$ and $\grave{}(\;;\;)$ denotes the log-likelihood function. Similar to step I, the maximization can be implemented using the Newton-Raphson algorithm, with the $m+1$ iteration update given by

$$\hat{\mathrm{b}}_{m+1} = \hat{\mathrm{b}}_m + \sum_{i=1}^{n} X_i^{\mathsf{T}} W_{1i}(\hat{\mathrm{b}}_m) X_i \quad {}^{1} \sum_{i=1}^{n} X_i^{\mathsf{T}} W_{2i} \hat{r}_i(\hat{\mathrm{b}}_m);$$

but with the following changes. The $N_i \quad (2p+4)$ predictor matrix X_i is equal to

$$\begin{bmatrix}
\mathrm{b}_{0i1} & \bar{a}_i\mathrm{b}_{0i1} & 0 & 0 & 0 & 0 & X_{2i} & \bar{a}_iX_{2i} & : : : & X_{pi} & \bar{a}_iX_{pi} \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & & \vdots & \vdots \\
\mathrm{b}_{0iN_{0i}} & \bar{a}_i\mathrm{b}_{0iN_{0i}} & 0 & 0 & 0 & 0 & X_{2i} & \bar{a}_iX_{2i} & : : : & X_{pi} & \bar{a}_iX_{pi} \\
\hline
0 & 0 & \mathrm{b}_{1i1} & \bar{a}_i\mathrm{b}_{1i1} & Z_i & \bar{a}_iZ_i & X_{2i} & \bar{a}_iX_{2i} & : : : & X_{pi} & \bar{a}_iX_{pi} \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & & \vdots & \vdots \\
0 & 0 & \mathrm{b}_{1iN_{1i}} & \bar{a}_i\mathrm{b}_{1iN_{1i}} & Z_i & \bar{a}_iZ_i & X_{2i} & \bar{a}_iX_{2i} & : : : & X_{pi} & \bar{a}_iX_{pi}
\end{bmatrix};$$

$$\hat{\mathrm{b}}_{0;ik} = g^{\;1} \;\; (\hat{\mathrm{b}}_{0;m} + \hat{\mathrm{b}}_{1;m}\bar{a}_i)\mathrm{b}_{0ik} + \sum_{r=2}^{p} (\hat{\mathrm{b}}_{r0;m} + \hat{\mathrm{b}}_{r1;m}\bar{a}_i)X_{ri} \;\; ;$$

$$\hat{\mathrm{b}}_{1;ik^0} = g^{\;1} \;\; (\hat{f}_{0;m} + \hat{f}_{1;m}\bar{a}_i)\mathrm{b}_{1ik^0} + (\hat{\mathrm{b}}_{10;m} + \hat{\mathrm{b}}_{11;m}\bar{a}_i)Z_i + \sum_{r=2}^{p} (\hat{\mathrm{b}}_{r0;m} + \hat{\mathrm{b}}_{r1;m}\bar{a}_i)X_{ri} \;\; ;$$

$W_{1i}(\boldsymbol{\theta}_m) = \text{diag}\{K_h(\bar{a}_i)\boldsymbol{b}_{0;i1}(1 - \boldsymbol{b}_{0;i1}), \ldots, K_h(\bar{a}_i)\boldsymbol{b}_{0;iN_{0i}}(1 - \boldsymbol{b}_{0;iN_{0i}}), K_h(\bar{a}_i)\boldsymbol{b}_{1;i1}(1 - \boldsymbol{b}_{1;i1}), \ldots, K_h(\bar{a}_i)\boldsymbol{b}_{1;iN_{1i}}(1 - \boldsymbol{b}_{1;iN_{1i}})\}$, $W_{2ij} = \text{diag}\{K_h(\bar{a}_i), \ldots, K_h(\bar{a}_i)\}$ and $\hat{\boldsymbol{Y}}_i(\boldsymbol{\theta}_m) = (Y_{0;i1} - \boldsymbol{b}_{0;i1}, \ldots, Y_{0;iN_{0i}} - \boldsymbol{b}_{0;iN_{0i}}; Y_{1;i1} - \boldsymbol{b}_{1;i1}, \ldots, Y_{1;iN_{1i}} - \boldsymbol{b}_{1;iN_{1i}})^{\mathsf{T}}$, for binary response. For modeling a Poisson distributed count response, $W_{1i}(\boldsymbol{\theta}_m) = \text{diag}\{K_h(\bar{a}_i)\boldsymbol{b}_{0;i1}, \ldots, K_h(\bar{a}_i)\boldsymbol{b}_{0;iN_{0i}}, K_h(\bar{a}_i)\boldsymbol{b}_{1;i1}, \ldots, K_h(\bar{a}_i)\boldsymbol{b}_{1;iN_{1i}}\}$. For subjects who do not have any infection-related hospitalization, the predictor matrix reduces to size $N_{0i} \times (2p + 4)$ and sizes of above quantities adjust accordingly.

1.4 Multiple-Index Cardiovascular Event Risk Trajectories in Older Patients on Dialysis

1.4.1 Description of the Study Cohort

We use data from the United States Renal Data System (USRDS), which collects data on nearly all patients with end-stage renal disease in the U.S. The USRDS is a national database that collects and maintains standard analytic files, including data on inpatient hospitalizations submitted to Medicare, patient demographics, dialysis modality, comorbidities and laboratory measures at the start of dialysis (USRDS, 2011). The defined population of inference in our study are patients aged 65 and older who newly initiated dialysis between January 1, 2000 and December 31, 2007 without a prior history of renal transplant. Patients were eligible for inclusion if (a) they survived the first 90 days of dialysis and did not recover renal function or receive a kidney transplant during this interval, (b) had Medicare as the primary payer on day 91 of dialysis, and (c) were receiving hemodialysis or peritoneal dialysis on day 91 of dialysis. Thus, the observation period began on day 91 of dialysis. The follow-up period for an individual ended at death (78.5%) or study end on December 31, 2009 or after 5 years of observation (from the start of dialysis or after the initial infection-related hospitalization, 21.5%). In our analysis, we exclude 1.3% of the cohort that had renal function recovery and 3.0% of the cohort that received kidney transplantation, since the evaluation of candidates for transplantation relates to overall health and benefit. We include patients

whose follow-up was administratively censored, since there are no differences expected in cohorts initiating dialysis at different times. A follow-up sensitivity analysis run on the entire cohort that included patients with renal function recovery and kidney transplantation lead to very similar findings and hence results are not reported here.

A cardiovascular event was defined as 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 ascertained from discharge diagnoses, also based on ICD-9-CM classification (categories include bacteremia, candidemia and viremia; central nervous system; cardiovascular; peritonitis and peritoneal abscess; gastrointestinal and hepatobiliary; genitourinary; pulmonary; skin and soft tissue; bone and joint; dialysis access and central venous catheters; device, procedure and surgery-related.) Baseline covariates of interest include demographic characteristics (sex, race), comorbidities (diabetes, coronary heart disease, congestive heart failure, peripheral vascular disease), body mass index (BMI) and MDRD eGFR (estimated glomerular filtration rate based on the 4-variable Modification of Diet in Renal Disease (MDRD) equation from the National Kidney Foundation (2002)). Our assessment of the cardiovascular risk trajectories will focus on older patients on dialysis with baseline age $a_i \; 2 \; (65; 90)$; however, we conservatively include in our estimation procedure patients aged 59 to 96 years at the start of dialysis to reduce boundary effects. Table A.1 summarizes the baseline covariates for $n = 294; 511$ patients used in fitting the GM-IVC model, of which $245; 874$ patients were aged 65 to 90.

1.4.2 Results: Cardiovascular Event Risk Trajectories

We begin our proposed estimation procedure for the GM-IVC model (1.1) by obtaining the age-stratified $_{0j}$ and $_{1j}$ estimates. For this, we bin patients into 2 year baseline age strata, where bins are a little wider at 3 years for strata above age 84, to obtain stable estimates at very high ages, yielding a total of 11 bins. A sensitivity analysis has been run where the total number of bins were selected as 8 and 14; data analysis results were very similar and readers

are referred to Appendix B.1 for details. The age-stratified estimates (b_{0j}'s and b_{1j}'s) are plotted as a function of vintage and time (years) since the pivotal initial infection-related hospitalization in Figure A.2(a) and (b), respectively. The plotted age-stratified estimates roughly share a similar increasing pattern, indicating that the multiplicative assumption of model (1.1) is reasonable in our application.

The final estimated time-varying coefficient functions over the two time indices, namely $b_0(t_0)$ and $b_1(t_1)$, along with the age-varying coefficient functions $b_0(a)$ and $b_1(a)$ are displayed in Figure A.2(c) and (d), respectively. Also provided along with the cardiovascular event risk trajectories are percentile bootstrap confidence intervals based on 200 bootstrap replications formed by resampling from subject trajectories with replacement. The bandwidths used ($h = 1.5$ years for $\{b_0(t_0); b_1(t_1)\}$ in the first step and $h = 4$ years for $\{b_0(a); b_1(a); b_r(a)\}$ in the second step of the local maximum likelihood estimation procedure) were chosen by the 20-fold cross-validation, similar to Cai et al. (2000). To reduce boundary effects, a bandwidth of $h = 2.5$ years was used at grid points close to 5 years in estimation of $\hat{}_1(t_1)$. Recall (from Section 1.2.2) that in assessing the estimated varying coefficient functions, b_0 and b_1 do not carry the sign or the magnitude of the regression effects, and they should be compared based on their shapes. Because b_0 and b_1 are negative (Figure A.2(d)), the general increasing patterns of b_0 and b_1 (in Figure A.2(c)), with respect to both time indices (t_0 and t_1), imply decreasing cardiovascular event probabilities after the start of dialysis and after the initial infection-related hospitalization for the dynamic cohort of survivors. Consistent with the slight convex pattern of b_1, around 0, we estimate a faster decrease in cardiovascular event probabilities after the start of dialysis compared to follow-up after the initial infection-related hospitalization. From Figure A.2(d), it can be seen that the estimated age-varying coefficient function $b_1(a)$ is greater than $b_0(a)$; this implies that cardiovascular event risk is nominally higher across all ages conditional on survival status after the initial infection-related hospitalization compared to after the start of dialysis. Furthermore, both $b_1(a)$ and $b_0(a)$ are increasing with age a and converging together as age a approaches 90; this suggests, not surprisingly, that cardiovascular event risk increases with

age generally.

To compare the cardiovascular event risk trajectories directly, as a function of vintage, time since the initial infection-related hospitalization and patient age at dialysis, Figure A.3 provides the cardiovascular event probabilities and their respective bootstrap confidence intervals over both time indices for baseline ages of 65, 78 and 90. The following observations can be made about the cardiovascular risk trajectories from the results in Figure A.3.

1. The risk is significantly greater across the five year follow-up time after the pivotal initial infection-related hospitalization compared to the time after the start of dialysis conditional on survival of the patient.

2. The highest risk is near the time of dialysis start and the time of the initial infection-related hospitalization and declines with both time indices.

3. The risks over both time indices increase with increasing baseline age.

4. The effect of baseline age on the risk of cardiovascular event is much more pronounced for time after the start of dialysis compared to the time after the initial infection-related hospitalization. Furthermore, from Figure A.2(d), the increasing cardiovascular event probability among the dynamic cohort of survivors after the initial infection-related hospitalization plateaus after baseline age 72.

5. However, the *di erence* in risks for time since dialysis and time since the initial infection-related hospitalization declines with increasing baseline age at dialysis.

The later two points are made transparent by Figure A.3(d), which overlays the estimated cardiovascular event probabilities across baseline ages. To illustrate the pattern of cardiovascular risk dynamics above, we selected the estimates for white male patients with diabetes and average levels of eGFR and BMI to display in Figure A.3.

Figure A.4 displays the estimated cardiovascular event risk trajectories for both time indices simultaneously, with the initial infection-related hospitalization occurring at 3, 2

and 1 year(s) after the start of dialysis; similar to Figure A.3, the risk trajectories are provided with bootstrap confidence intervals for individuals with baseline ages of 65, 78 and 90 (Figure A.4: left, middle and right column, respectively). The increased cardiovascular event probabilities remain elevated and do not decrease to their original levels; for example, even after one year from the initial infection-related hospitalization. This appears to hold independent of when the initial infection related-hospitalization occurred (1, 2 or 3 years after the start of dialysis). Also, consistent with impact of baseline age at dialysis described above, the sustained elevated risks are particularly pronounced for relatively younger patients at the start of dialysis (e.g., age 65 and 78 compared to age 90).

The effects of baseline covariates (potentially modified by baseline age) and vintage on the cardiovascular event probability are presented in Figure A.5. We find significant negative associations between vintage until the initiating infection, male gender, eGFR and BMI. Comorbidities, namely congestive heart failure, coronary heart disease, peripheral vascular disease and diabetes, were significantly associated with increased cardiovascular event risk. The effects of coronary heart disease and diabetes on cardiovascular risk decline with increasing baseline age.

1.5 Simulation Studies

We carry out simulation studies to examine the efficacy of the proposed estimation procedure to target the true time- and age-varying coefficient functions when there is truncation by death. Similar to the model used for the data analysis of Section 1.4, we consider the following GM-IVC model

$$g(\ \ _i) = \ _0(a_i)\ \ _0(t_{0i}) f 1 \quad \mathsf{I}_{\mathsf{P}_i}(t_i) g + \ _1(a_i)\ \ _1(t_{1i}) \mathsf{I}_{\mathsf{P}_i}(t_i) + \ _1(a_i) Z_i \mathsf{I}_{\mathsf{P}_i}(t_i) + \ _2(a_i) X_{2i} + \ _3(a_i) X_{3i};$$

for $t_{0i}\ 2\ [0; T_{0i}]$, $t_{1i}\ 2\ [0; T_{1i}]$, $T_{0i} \quad T; \mathrm{T}_{1i} \quad T$, $T = 5$, $a_i\ 2\ [65; 90]$ and $g(\ )$ the logistic link function. For the age index, the ith subject's age, a_i, is generated from a uniform random variable between 59 and 96. The longitudinal binary indicator of whether the ith subject experiences a pivotal exposure (that effects the time-varying indicator variable $\mathsf{I}_{\mathsf{P}_i}(t_i)$) in their

total follow-up time is determined according to a Bernoulli random variable with probability 0.67 to mimic the rate of infections in our data application. X_{2i} is a Bernoulli random variable with mean .52 and X_{3i} is a Gamma random variable with shape 4 and rate 6.

The response and the survival time are generated jointly where binary $Y_{0;ik}$ $Y_i(a_i; t_i = t_{0i} = t_{0ik})$ and $Y_{1;ik^0}$ $Y_i(a_i; t_i = Z_i + t_{1i}; t_{1i} = t_{1ik^0})$, defined in Section 1.3.1, are generated as indicators for $(Y_{0;ik} > 0)$ and $(Y_{1;ik^0} > 0)$, respectively. For subjects who did not experience a pivotal exposure, we generate $(Y_{0;i1}; \cdots; Y_{0;i21}; S_i)^\mathsf{T}$ according to a 22 dimensional multivariate normal distribution with mean vector $[(\ _{0;i1}; \cdots;\ _{0;i21})^\mathsf{T}; E(S_i) = 4]^\mathsf{T}$ where $_{0;i} = (\ _{0;i1}; \cdots;\ _{0;i21})^\mathsf{T}$ represents the mean vector $E\{Y_i(a_i; t_i; t_{0i}; t_{1i}) | Z_i; X_i; \mathrm{I}_{\mathrm{P}_i}(t_i)\}$ 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 three months for a maximum of 5 year follow-up. The covariance matrix of the 22 dimensional multivariate normal distribution is $\Sigma = [\Sigma_{11}; \Sigma_{12}; \Sigma_{21}; \Sigma_{22}]$, Σ_{11} has dimension 21 21 with ones down the diagonal and .25 off the diagonal, $\Sigma_{12} = \Sigma_{21}^\mathsf{T}$ has dimension 21 1 with .25 in every position, and Σ_{22} has dimension 1 1 with element .50. Unconditional means $_{0;i} = (\ _{0;i1}; \cdots;\ _{0;i21})$ are found through the following correspondence:

$$_{0;ik} = E[Y_{0;ik} | S_i > t_{0ik}] = P(Y_{0;ik} > 0 | S_i > t_{0ik}) = \frac{P(Y_{0;ik} > 0; S_i > t_{0ik})}{P(S_i > t_{0ik})}. \tag{1.7}$$

In other words, we have that $P(Y_{0;ik} > 0; S_i > t_{0ik}) =\ _{0;ik}\ P(S_i > t_{0ik})$ and we find $_{0;i} = (\ _{0;i1}; \cdots;\ _{0;i21})^\mathsf{T}$ using the bisection method (similar to Kurland and Heagerty, 2005), where $_{0;ik} = \{\ ^{1}f\ _{0}(a_i)\ _{0}(t_{0ik}) +\ _{2}(a_i)X_{1i} +\ _{3}(a_i)X_{2i}\}$. The generated $(Y_{0;i1}; \cdots; Y_{0;i21})^\mathsf{T}$ vector is truncated such that $t_{0;ik}$ S_i to create the observed outcomes for $k = 1; \cdots; N_{0i}$. For subjects who experienced a pivotal exposure during follow up, we generate $Z_i = \frac{1}{4}\{4W_i\}$ where W_i $N(3; .5)$ and $\{\ \}$ denotes the floor function. We consider Z_i observed and generate $(Y_{0;i1}; \cdots; Y_{0;iN_{0i}}; Y_{1;i1}; \cdots; Y_{1;i20}; S_i)^\mathsf{T}$ $N_{N_{0i}+20}([\ _{0;i1}; \cdots;\ _{0;iN_{0i}};\ _{1;i1}; \cdots\ _{1;i20}; E(S_i) = 4 + Z_i]^\mathsf{T}; \Sigma)$ where $N_{0i} = 4Z_i + 1$, $\Sigma = [\Sigma_{11}; \Sigma_{12}; \Sigma_{21}; \Sigma_{22}]$, Σ_{11} has dimension $(N_{0i}+20)$ $(N_{0i}+20)$ with ones down the diagonal and .25 off the diagonal, $\Sigma_{12} = \Sigma_{21}^\mathsf{T}$ has dimension $(N_{0i}+20)$ 1 with .25 in every position, and Σ_{22} has dimension 1 1 with element .50. The unconditional means $(\ _{0;i1}; \cdots;\ _{0;iN_{0i}})$ are found as described in equation (1.7) and $(\ _{1;i1}; \cdots;\ _{1;i20})$ are

found similarly by the bisection method using $P(Y_{1;ik^0} > 0; S_i > Z_i + t_{1ik^0}) = \ _{1;ik^0} \quad P(S_i > Z_i + t_{1ik^0})$ where $\ _{1;ik^0} = g \ ^1 f \ _1(a_i) \ \ _1(t_{1ik^0}) + \ _1(a_i) Z_i + \ _2(a_i) X_{1i} + \ _3(a_i) X_{2i} g$. The generated $(Y_{1;i1}; \cdots; Y_{1;i21})^{\mathsf{T}}$ vector is truncated such that $Z_i + t_{1;ik^0} \quad S_i$ to create the observed outcomes after the pivotal exposure for $k^0 = 1; \cdots; N_{1i}$.

The shape time-varying coefficient functions are taken to be $\ _0(t_0) = 0.0148 t_0^2 \quad 0.0083 t_0 + 0.3333$ and $\ _1(t_1) = 0.0250 t_1^2 \quad 0.0167 t_1 + 0.2500$; both functions satisfy the identifiability conditions (4). The age-varying effects before and after the pivotal exposure are $\ _0(a) = \ 0.003 a^2 + 0.58 a \quad 29$ and $\ _1(a) = 0.002 a^2 \quad 0.22 a + 2$. The baseline age-varying covariate effects are $\ _1(a) = 0.025 a \quad 1.6$, $\ _2(a) = \ \sin(0.02 \ a \ \ 4 \ )$ and $\ _3(a) = \ 0.0012 a^2 + 0.14 a \quad 5$.

To study the performance of the proposed estimation procedure for the GM-IVC, we utilize a relative mean squared deviation error (MSDE) defined as

$$\text{MSDE}_{\ 0} = \int_0^T f \ _0(t_0) \quad \mathrm{b}_0(t_0) g^2 dt_0 \ \Big/ \int_0^T \ _0^2(t_0) dt_0$$

for the time-vary function $\ _0(t_0)$. The MSDEs for the other time- and age-varying coefficient functions, namely MSDE $_1$, MSDE $_0$, MSDE $_1$, MSDE $_1$, MSDE $_2$ and MSDE $_3$, are defined similarly. Table A.2 gives the median, first quartile and third quartile of the estimated MSDE values in percentage for the varying coefficient functions over 200 Monte Carlo runs. Results are presented at two sample sizes $n = 3000$ and 5000 with bandwidth chosen by 20-fold cross-validation as described in Cai et al. (2000). Bandwidths used were chosen in a preliminary simulation study yielding $h = (1.5; 1.5)$ for $\hat{\ }_0(t_0); \hat{\ }_1(t_1)$ and $h = (3.75; 2.5)$ for $\hat{\ }_0(a); \hat{\ }_1(a); \hat{\ }_2(a)$ for $n = 3000$ and $n = 5000$, respectively. Overall, the MSDE values reported are small and get smaller with increased sample size. In addition, Figure A.6 displays the estimated median, 5th and 95th percentile varying coefficient functions overlaying the true curves at the sample size $n = 3000$. The estimated functions track the true varying coefficient functions. Overall, the simulation studies illustrate that the proposed estimation procedure for the GM-IVC models is effective in capturing the true time- and age-varying dynamics for data truncated by death.

1.6 Discussion

Infection and cardiovascular disease remain the leading causes of hospitalization and death in patients on dialysis in the United States (USRDS 2011). Understanding the complex cardiovascular risk trajectories is important for potential strategies to target cardiovascular risk reduction, including implementation of overall infection control or prevention strategies. The results highlight the significant impact of the first infection-related hospitalization on cardiovascular event risk and the dependence of this effect on age at the start of dialysis. An important finding is that the infection-related hospitalization results in *sustained* increases in cardiovascular event risk among the dynamic cohort of survivors; for instance, even one year after the infection-related hospitalization, the cardiovascular event probability is still higher than the cardiovascular risk at the start of dialysis, a time of high cardiovascular risk with respect to vintage. This pattern of cardiovascular risk dynamics, with respect to vintage (t_0) and time since the pivotal initial infection-related hospitalization (t_1), holds for most older patients starting dialysis, although the difference in cardiovascular event probabilities before and after the infection declines with increasing baseline age at dialysis. The difference in cardiovascular risk converges (equalizes) only for very elderly patients starting dialysis, near 90 years of age.

From a technical perspective, the proposed GM-IVC models add important and necessary flexibility to the current varying coefficient modeling toolkit by the introduction of additional time (and age) indices. We achieved this by employing a sensible multiplicative structure to capture the multiple time- and age-varying effects, and at the same time, avoiding the curse of dimensionality, a known limiting factor in modeling dynamic, varying effects. The multiplicative structure assumption can be assessed in the first step of the proposed estimation algorithm via assessing whether the binned estimators of $_0(t_0)$ and $_1(t_1)$ share a common shape. If this does not seem to be a plausible assumption, the more general two dimensional regression surfaces $h_0(a; t_0)$ and $h_1(a; t_1)$ need to be targeted. Also, as illustrated with our USRDS data application, the GM-IVC models provide natural graphical displays of time-

and age-varying dynamics that are fairly easy to interpret; thus, retaining a popular feature of standard/classical varying coefficient models. We believe the proposed GM-IVC models are widely applicable since characterizing the outcome trajectories over multiple indices, including time since a pivotal exposure event is often of interest in longitudinal analysis. Finally due to a large percent of truncation by death, we developed a model targeting a partly conditional inference target, conditional on the survival status of the patients. Partly conditional models was originally proposed for generalized linear models (Kurland and Heagerty, 2005); the current work extends them to varying coefficient models incorporating multiple indices. Investigation of the theoretical properties of the proposed estimators is an open problem. We provide R codes for our GM-IVC model at http://dsenturk.bol.ucla.edu/.

CHAPTER 2

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

2.1 Introduction

As of 2011, more than 430,000 adults in the United States were on dialysis, a life-sustaining treatment (United States Renal Data System Annual Data Report [USRDS ADR], 2013). Cardiovascular (CV) disease and infection remain the leading causes of mortality and hospitalization in patients on dialysis (USRDS ADR, 2013). 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 in older patients on dialysis. Understanding the time-dynamic changes in patients' CV outcome trajectories over time to allow for identification of time frames of increased CV risk (probability) is an important step towards exploring effective approaches to disease management. Hence, the main goal of the manuscript is to develop *conditional* modeling approaches to model the time-varying effects of risk factors including the initial infection-related hospitalization and baseline co-existing illnesses on patients' CV outcome trajectories over time, from the start of dialysis. We consider a time-variant binary indicator outcome of having a CV event within a 3 month follow-up interval where a CV event is defined as a myocardial infarction, unstable angina, stroke, or transient ischemic attack, determined from hospitalization primary discharge diagnosis.

Our study 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. Hence a major challenge in the time-varying effects modeling is the high mortality

in the dialysis population. 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* time-varying effects models for handling truncation by death. More precisely, 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 et al., 2004; Qu and Li, 2006; Senturk and Mueller, 2009; Senturk et al., 2013), have been adapted for analyzing longitudinal data (e.g., see Hoover et al., 1998; Wu et al., 2000; Chiang et al., 2001; Fan et al., 2000; 2003; 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 time-invariant. In our current work, we consider the following PL-GVCM, containing both time-invariant and time-variant predictors, necessary for our application:

$$g[E\,fY(t)jX;U(t)g] = \sum_{r=1}^{X} {}_{r}X_r + \sum_{s=1}^{X} {}_{s}(t)U_s(t);$$

where $Y(t)$ is the outcome trajectory, $g(\ )$ is a known link function, $X = (X_1;\cdots;X_p)^{\mathsf{T}}$ is the vector of time-invariant covariates, and $U(t) = fU_1(t);\cdots;U_q(t)g^{\mathsf{T}}$ is the vector of time-variant covariates. The coefficients, $\ = (\ {}_1;\cdots;\ {}_p)^{\mathsf{T}}$, describe constant effects corresponding to time-invariant factors and the time-varying regression coefficients, $(\ ) = f\ {}_1(\ );\cdots;\ {}_q(\ )g^{\mathsf{T}}$, capture the dynamic effects of the time-variant 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 and missing data literatures targeting an unconditional mean trajectory model, specifically $E\,fY(t)jX;U(t)g$, 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 ${}_P\ \ E\,fY(t)jX;U(t);S > tg$ (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, ${}_F\ \ E\,fY(t)jX;U(t);S = tg$, which conditions on the actual death time (i.e., $S = t$). Note that the fully conditional model is not intended for the purpose of prediction since it conditions on the time of death. However, it is a useful modeling tool to explore association of risk factors and characterizing trends in response trajectories in retrospective population-based studies.

We note that the ideas of conditioning on the survival and on actual death time, namely

partly and fully conditionals, were introduced in Kurland and Heagerty (2005) and Kurland et al. (2009) for standard generalized linear models and that Estes et al. (2014) is the only work extending the partly conditional model to the varying coefficient models. However, Estes et al. (2014) does not fully explore partly conditional model interpretations or comparisons with other conditional approaches. The current paper is the first in literature to contrasts the targets/goals of inference (including related interpretations) for partly and fully conditional time-varying effects models in the context of PL-GVCMs and to provide guidance on their applications in the context of longitudinal data with substantial truncation by death. In addition, this work is also the first in literature to study efficacy of inference procedures (generalized likelihood ratio tests) for the conditional time-varying effects models.

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 (simulation study detailed in Appendix C.1). Figure A.7 displays the partly conditional and fully conditional (along with the unconditional) estimates of the varying coefficient function targets. We could imagine a scenario where t denotes years on treatment which reduces relapse probability of a disease condition in patients. 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 relapse risk trajectory during the first two years of treatment among patients who survive at least two years on dialysis?"* In Figure A.7, 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; this reflects the fact that $_P(t) \neq (t)$, particularly for high level of mortality during follow-up. In contrast, a fully conditional model conditions on a specific time of death t and, thus, the inferential interest focuses on the time-varying trajectory for the stratum 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 A.7 for death times $t = 3; 4$, and 5 years) are estimated to compare trends in the expected

outcome trajectories for the death strata. The fully conditional model approach would allow one *to compare the relapse risk trajectories for a series of patient cohorts who die around for example 1, 2, and 3 years.*

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 2.2. Section 2.3 provides modeling results and interpretations, followed by simulation studies in Section 2.4 and a discussion in Section 2.5.

2.2 Partly and Fully Conditional Time-Varying Effect Modeling

2.2.1 Model Specification: Conditional PL-GVCM

As introduced in Section 2.1, 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 initial 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 initial 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-variant CV event probability (risk), we model the time-variant 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:

$$\mu_{i;P} \quad \mu_{i;P}(t_i; t_{0i}; t_{1i}) = E\{Y_i(t_i; t_{0i}; t_{1i}) | Z_i; X_i; I_{G_i}(t_i); S_i > t_i\};$$

where $I_{G_i}(t_i)$ denotes a time-variant indicator of infection-related hospitalization *prior* to time t_i; Z_i is the vintage till the initial infection-related hospitalization if patient i has at least one infection-related hospitalization; $X_i = (X_{2i}; \ldots; X_{pi})^T$ are time-invariant 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:

$$g(\mu_{i;P}) = \alpha_{0;P}(t_{0i})\{1 - I_{G_i}(t_i)\} + \alpha_{1;P}(t_{1i}) I_{G_i}(t_i) + \gamma_{1;P} Z_i I_{G_i}(t_i) + \sum_{r=2}^{p} \gamma_{r;P} X_{ri}; \qquad (2.1)$$

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 $\{\gamma_{r;P}\}_{r=1}^{p}$, correspond to the effects of vintage prior to the initial infection and time-invariant covariates. The supports for the varying coefficient functions in (2.1) 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-variant indicator, $I_{G_i}(t_i)$, in the PL-GVCM (2.1) 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 $\mu_{i;P} = g^{-1}\{\alpha_{0;P}(t_{0i}) + \sum_{r=2}^{p} \gamma_{r;P} X_{ri}\}$. For patients with at least one infection-related hospitalization, we see from (2.1) that the CV risk model *after* the initial infection-related hospitalization transitions to $\mu_{i;P} = g^{-1}\{\alpha_{1;P}(t_{1i}) + \gamma_{1;P} Z_i + \sum_{r=2}^{p} \gamma_{r;P} X_{ri}\}$. Note that this model appropriately accounts for vintage till the initial infection-related hospitalization, namely Z_i.

The proposed model can be characterized as a segmented nonparametric regression model where the time of the initial infection-related hospitalization acts as an observed change

point and the CV risk trajectory before and after this observed change point is allowed be a nonparametric function of time. For simplicity we concentrate on the initial infection-related hospitalization, since the initial hospitalization leads to the largest increase in CV risk in our preliminary findings. However the model can be extended to incorporate multiple infection-related hospitalizations with multiple change points or a change point with multiple states than only two. The latter extension would require the same number of varying coefficient functions as the levels of the states of the change point variable.

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

$$_{ij;F} \quad _{ij;F}(t_i; t_{0i}; t_{1i}) = E\, f Y_i(t_i; t_{0i}; t_{1i}) j Z_i; X_i; \mathrm{I}_{G_i}(t_i); S_i \;\; 2 \;\; D_j g;$$

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

$$g(\;_{ij;F}) = \;_{0j;F}(t_{0i}) f 1 \quad \mathrm{I}_{G_i}(t_i) g + \;_{1j;F}(t_{1i}) \mathrm{I}_{G_i}(t_i) + \;_{1j;F} Z_i \mathrm{I}_{G_i}(t_i) + \sum_{r=2} \;_{rj;F} X_{ri}; \qquad (2.2)$$

The parameters and varying coefficient functions in (2.2) above are analogously defined as in the partly conditional model (2.1).

2.2.2 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 $_s$'s first, then targeting the $_r$'s via maximum likelihood using the estimated VCFs, followed by re-estimation of the VCFs using the estimated $_r$'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.

2.2.2.1 Step 1: Initial Estimation of $_{0;P}(t_0)$ and $_{1;P}(t_1)$

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^0} to be the midpoints of the kth and k^0th 3-month time intervals since initiation of dialysis and since the initial infection-related hospitalization, respectively. We define the binary response variable $Y_{0;ik}$ $Y_i(t_i = t_{0i} = t_{0ik}) = 1$, if the ith patient had at least one CV event in the kth 3-month interval after initiation of dialysis. Similarly, $Y_{1;ik^0}$ $Y_i(t_i = Z_i + t_{1i}; t_{1i} = t_{1ik^0}) = 1$ if the ith patient had at least one CV event in the k^0th 3-month interval after the initial infection-related hospitalization. Hence, the available data is $f(t_{0ik}; t_{1ik^0}; X_{ri}; Z_i; Y_{0;ik}; Y_{1;ik^0}) : i = 1; \cdots; n;\ k = 1; \cdots; N_{0i};\ k^0 = 1; \cdots; N_{1i}g$, where n is the total number of subjects.

We utilize three month intervals to facilitate obtaining stable fits for both the partly and fully conditional models, allowing comparisons. Note that there are multiple factors that effect the stability of the PL-GVCM fit in general: 1) nature of the response (continuous, binary, etc.), 2) number of predictors in the model, and 3) amount of truncation by death. A binary response with a small probability of success, increasing number of predictors and larger amount of truncation by death would require a larger sample size for a stable fit, especially towards the end of follow-up, which would require larger partitions.

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

$$\begin{aligned}\ell_1(\mathbf{c}) &= \frac{1}{\sum_{i=1}^{n} N_i} \sum_{i=1}^{n} \sum_{k=1}^{N_i} \ell\left(g^{-1}\left(c_0 + c_1(t_{0ik} - s_0) + \sum_{r=2}^{p} b_r X_{ri}\right); Y_{0;ik}\right) K_h(t_{0ik} - s_0) \\ &\quad + \sum_{k'=1}^{N_{1i}} \ell\left(g^{-1}\left(d_0 + d_1(t_{1ik'} - s_0) + b_1 Z_i + \sum_{r=2}^{p} b_r X_{ri}\right); Y_{1;ik'}\right) K_h(t_{1ik'} - s_0), \quad (2.3)\end{aligned}$$

provides the initial local ML estimators for the VCFs, namely $\hat{\alpha}_{0;P}(t_0) = \hat{c}_0$ and $\hat{\alpha}_{1;P}(t_1) = \hat{d}_0$. In the above local log-likelihood, $K_h(\cdot) = K(\cdot/h)/h$, $K(\cdot)$ denotes a kernel function and h is the bandwidth; $\mathbf{c} = (c_0; c_1; d_0; d_1; b_1; \ldots; b_p)^{\mathsf{T}}$; $N_i = N_{0i} + N_{1i}$ and $\ell(\cdot;\cdot)$ denotes the log-likelihood function.

Let $\hat{\mu}_{0;ik} = g^{-1}\{\hat{c}_0 + \hat{c}_1(t_{0ik} - s_0) + \hat{\eta}_i\}$ and $\hat{\mu}_{1;ik'} = g^{-1}\{\hat{d}_0 + \hat{d}_1(t_{1ik'} - s_0) + \hat{b}_1 Z_i + \hat{\eta}_i\}$, where $\hat{\eta}_i = \sum_{r=2}^{p} \hat{b}_r X_{ri}$. Also, define $\dot{\mu}_{v;ij} = K_h(t_{vij} - s_0)\{g'_{j=1}^{N_{vi}}\}$, $\dot{\mu}_{v;ij}(1 - \hat{\mu}_{v;ij})\{g'_{j=1}^{N_{vi}}\}$, and $\dot{\mu}_{v;ij}\hat{\mu}_{v;ij}\{g'_{j=1}^{N_{vi}}\}$, for $v = 0; 1$. Then the Newton-Raphson update at iteration $m+1$ is given by

$$\hat{\mathbf{c}}_{m+1} = \hat{\mathbf{c}}_m + \left(\sum_{i=1}^{n} X_{1i}^{\mathsf{T}} W_{1i}(\hat{\mathbf{c}}_m) X_{1i}\right)^{-1} \left(\sum_{i=1}^{n} X_{1i}^{\mathsf{T}} W_{2i} \hat{\varepsilon}_i(\hat{\mathbf{c}}_m)\right);$$

where

$$X_{1i} = \left[\begin{array}{ccc|ccc|cccc} 1 & (t_{0i1} - s_0) & & 0 & & 0 & 0 & X_{2i} & \cdots & X_{pi} \\ \vdots & \vdots & & \vdots & & \vdots & \vdots & \vdots & & \vdots \\ 1 & (t_{0iN_{0i}} - s_0) & & 0 & & 0 & 0 & X_{2i} & \cdots & X_{pi} \\ \hline 0 & 0 & & 1 & (t_{1i1} - s_0) & & Z_i & X_{2i} & \cdots & X_{pi} \\ \vdots & \vdots & & \vdots & \vdots & & \vdots & \vdots & & \vdots \\ 0 & 0 & & 1 & (t_{1iN_{1i}} - s_0) & & Z_i & X_{2i} & \cdots & X_{pi} \end{array}\right]$$

is the predictor matrix of size $N_i \times (p+4)$, $W_{1i}(\hat{\mathbf{c}}_m) = \mathrm{diag}\{\hat{\mu}_{0;i1}, \ldots, \hat{\mu}_{0;iN_{0i}}, \hat{\mu}_{1;i1}, \ldots, \hat{\mu}_{1;iN_{1i}}\}$, $W_{2i} = \mathrm{diag}\{\dot{\mu}_{0;i1}, \ldots, \dot{\mu}_{0;iN_{0i}}, \dot{\mu}_{1;i1}, \ldots, \dot{\mu}_{1;iN_{1i}}\}$ and $\hat{\varepsilon}_i(\hat{\mathbf{c}}_m) = (Y_{0;i1} - \hat{\mu}_{0;i1}, \ldots, Y_{0;iN_{0i}} - \hat{\mu}_{0;iN_{0i}}, Y_{1;i1} - \hat{\mu}_{1;i1}, \ldots, Y_{1;iN_{1i}} - \hat{\mu}_{1;iN_{1i}})^{\mathsf{T}}$, for a Bernoulli distributed response. For modeling a Poisson

distributed response, $W_{1i}(\mathbf{b}_m) = \text{diag}\{ \ _{0;i1}\mathbf{b}_{0;i1}; \cdots; \ _{0;iN_{0i}}\mathbf{b}_{0;iN_{0i}}; \ _{1;i1}\mathbf{b}_{1;i1}; \cdots; \ _{1;iN_{1i}}\mathbf{b}_{1;iN_{1i}}\}$. For subjects who do not have any infection-related hospitalization, the predictor matrix reduces to size $N_{0i} \quad (p+4)$ and sizes of the above quantities adjust accordingly.

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 (2.3) tacitly utilizes local working independence for repeated values within a subject. This is motivated by two points. First, the within-subject correlation for the response is quite weak in our data application (0.02). Second, and more generally, Kurland and Heagerty (2005) 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 authors report that when the working correlation is non-diagonal, the variance inverse terms depend on the specific value of the death time, which leads to biased estimation in targeting the partly conditional target.

2.2.2.2 Step 2: Estimation of $_{r;P}$

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

$$\begin{aligned}
\ell_2(\mathbf{e}) &= \frac{1}{\sum_{i=1}^{n} N_i} \sum_{i=1}^{n} \sum_{k=1}^{N_{0i}} \ell \ g^{-1}\left(\mathbf{b}_{0;P}(t_{0ik}) + LC_i \ ; Y_{0;ik}\right) \\
&+ \sum_{k'=1}^{N_{1i}} \ell \ g^{-1}\left(\mathbf{b}_{1;P}(t_{1ik'}) + e_1 Z_i + LC_i \ ; Y_{1;ik'} \ ;\right.
\end{aligned}$$

resulting in the ML estimators $\mathbf{b}_{r;P} = \mathbf{e}_r$ for $r = 1; \cdots; p$, where LC_i denotes $\sum_{r=2}^{p} e_r X_{ri}$. The maximization is carried out using the Newton-Raphson algorithm with the $m+1$ iteration

update given by

$$\boldsymbol{\theta}_{m+1} = \boldsymbol{\theta}_m + \left(\sum_{i=1} X_{2i}^T W_i(\boldsymbol{\theta}_m) X_{2i} \right)^{-1} \left(\sum_{i=1} X_{2i}^T \hat{\boldsymbol{r}}_i(\boldsymbol{\theta}_m) \right) \quad \text{where} \quad X_{2i} = \left[\begin{array}{c|cccc} 0 & X_{2i} & X_{3i} & \cdots & X_{pi} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & X_{2i} & X_{3i} & \cdots & X_{pi} \\ \hline Z_i & X_{2i} & X_{3i} & \cdots & X_{pi} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ Z_i & X_{2i} & X_{3i} & \cdots & X_{pi} \end{array}\right]$$

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

2.2.2.3 Step 3: Final Estimation of $\beta_{0;P}(t_0)$ and $\beta_{1;P}(t_1)$

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 $\hat{\mathbf{b}}_{r;P}$, $r = 1, \ldots, p$ from step 2. Hence, the $N_i \times 4$ design matrix X_{1i} uses the first 4 columns of the design matrix defined in step 1 and $\boldsymbol{\mu}_{0;ik}$ and $\boldsymbol{\mu}_{1;ik'}$ are redefined as $g^{-1}\{\mathbf{b}_0 + \mathbf{b}_1(t_{0ik} - s_0) + \sum_{r=2}^{p} \mathbf{b}_{r;P} X_{ri}\}$ and $g^{-1}\{\mathbf{b}_0 + \mathbf{b}_1(t_{1ik'} - s_0) + \mathbf{b}_{1;P} Z_i + \sum_{r=2}^{p} \mathbf{b}_{r;P} X_{ri}\}$, respectively.

2.2.3 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, $\beta_0(t_0)$ and $\beta_1(t_1)$, 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., $H_0 : \beta_0(t_0) = c_0$ and $\beta_1(t_1) = c_1$, as illustrated in Figure A.12(a). This hypothesis encompasses the case where the infection event induces a

constant change (shift) in the CV outcome risk (i.e., where $c_0 \neq c_1$). A second hypothesis of interest involves a comparison of time-varying e ects before and after an initial infection-related hospitalization, speci cally $H_0 : {}_0(t_0) = {}_1(t_1)$, as illustrated in Figure A.12(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 rst hypothesis test for constancy of the varying coe cient functions, the null hypothesis 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 speci cally, 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 de ned 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 2.4, where high follow-up truncation by death ranges from 40-80%.

The GLRT statistic, denoted T, is of the form $T = r_k f \ell(H_1) \quad \ell(H_0) g$ where $r_k = f K(0) \quad 0{:}5 \int K^2(u)du g = [\int f K(u) \quad 0{:}5(K \quad K)(u) g^2 du]$, $K \quad K$ 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 the log-likelihoods $\ell(\ )$, are given by

$$\sum_{i=1}^{n} \sum_{k=1}^{N_{0i}} f Y_{0;ik} \log(p_{0;ik}) + (1 \quad Y_{0;ik}) \log(1 \quad p_{0;ik}) g + \sum_{k^0=1}^{N_{1i}} f Y_{1;ik^0} \log(p_{1;ik^0}) + (1 \quad Y_{1;ik^0}) \log(1 \quad p_{1;ik^0}) g : \tag{2.4}$$

In our application, we use the Epanechnikov kernel where $r_k = 2{:}1153$. Fan et al. (2001) showed that the GLRT statistic follow a ^{2}-distribution asymptotically; however, the level

of the test may not be achieved consistently. To alleviate this issue, we will adopt the approach in 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, using $\ell(H_0)$ and $\ell(H_1)$; (ii) compute the GLRT statistic $T = 2 \cdot 1153 \{\ell(H_1) - \ell(H_0)\}$; (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 (2.1) reduces to the generalized linear model $g(\mu_{i;P}) = c_0\{1 - I_{G_i}(t_i)\} + c_1 I_{G_i}(t_i) + \gamma_{1;P} Z_i I_{G_i}(t_i) + \sum_{r=2}^{p} \gamma_{r;P} X_{ri}$ under the null hypothesis. Parameters $\gamma_{r;P}$ can be estimated by maximizing the global likelihood in step 2 of the proposed estimation algorithm (Section 2.2.2) where $b_{0;P}(t_{0i})$ and $b_{1;P}(t_{1i})$ 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 detailed in Section 2.2.

Next, the test statistic is computed using the log-likelihoods given in (2.4) under the null and alternative hypotheses; where under the null $\hat{\mu}_{0;ik} = g^{-1}(c_0 + \hat{\eta}_i)$ and $\hat{\mu}_{1;ik} = g^{-1}(c_1 + b_{1;P} Z_i + \hat{\eta}_i)$ with $\hat{\eta}_i = \sum_{r=2}^{p} \hat{\gamma}_{r;P} X_{ri}$. Similarly, under the alternative $\hat{\mu}_{0;ik} = g^{-1}\{b_0(t_{0i}) + \hat{\eta}_i\}$ and $\hat{\mu}_{1;ik} = g^{-1}\{b_1(t_{1i}) + \hat{\gamma}_{1;P} Z_i + \hat{\eta}_i\}$, all evaluated using parameter estimates under respective hypotheses. The response values $Y^*_{0;i} = (Y_{0;i1}, Y_{0;i2}, \ldots, Y_{0;iN_{0i}}, Y_{1;i1}, Y_{1;i2}, \ldots, Y_{1;iN_{1i}})^T$ in the bootstrap sample are generated using parameter estimates under the null according to $Y_{0;ik} \sim \text{Bernoulli}\{g^{-1}(c_0 + \hat{\eta}_i)\}$ and $Y_{1;ik} \sim \text{Bernoulli}\{g^{-1}(c_1 + \hat{\gamma}_{1;P} Z_i + \hat{\eta}_i)\}$. 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 $H_0 : {}_0(t_0) = {}_1(t_1)$, model (2.1) reduces to $g(\ {}_{i;P}) = \ (t_{0i})f\,1 - I_{G_i}(t_i)g + \ (t_{i1})I_{G_i}(t_i) + \ {}_{1;P}Z_i I_{G_i}(t_i) + \sum_{r=2}^{p} \ {}_{r;P}X_{ri}$ under the null, where $(\ )$ denotes the common varying coe cient function under the equality ${}_0(t_0) = \ {}_1(t_1)$. An adaptation of the proposed estimation algorithm is used to estimate the parameters under the null hypothesis, where d_0 and d_1 are replaced with c_0 and c_1, respectively in (2.3), X_{1i} in step 1 of Section 2.2.2 reduces down to $N_i \ (p+2)$ matrix with second to fourth columns replaced with $(t_{0i1} \ s_0; : : : ; t_{0iN_{0i}} \ s_0; t_{1i1} \ s_0; : : : ; t_{1iN_{1i}} \ s_0)^T$ 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 Section 2.2. Similar to the the rst hypothesis test, the test statistic is computed using the likelihood in (2.4) under the null and alternative hypotheses, where under the null $p_{0;ik} = g^{\ 1}f\,b(t_{0i}) + b_i g$ and $p_{1;ik} = g^{\ 1}f\,b(t_{1i}) + \ {}_{1;P}Z_i + b_i g$; under the alternative $p_{0;ik} = g^{\ 1}f\,b_0(t_{0i}) + b_i g$, $p_{1;ik} = g^{\ 1}f\,b_1(t_{1i}) + \ {}_{1;P}Z_i + b_i g$ using parameter estimates under respective hypotheses. The bootstrap response $Y^*_{0;i1}; Y_{0;i2}; : : : ; Y_{0;iN_{0i}}; Y_{1;i1}; Y_{1;i2}; : : : ; Y_{1;iN_{1i}})^T$ is generated under the null according to $Y_{0;ik} \ \ \text{Bernoulli}\,[g^{\ 1}f\,b(t_{0i}) + b_i g]$ and $Y_{1;ik} \ \ \text{Bernoulli}\,[g^{\ 1}f\,b(t_{1i}) + \ {}_{1;P}Z_i + b_i g]$. The bootstrap test statistics are used to approximate the distribution of T under the null similar to the rst test.

2.3 Applications to Infection-Cardiovascular Risk Modeling

2.3.1 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 rst 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 de ned as a myocardial infarction, unstable angina, stroke, or transient ischemic attack, determined from primary discharge diagnosis and based on the International Classi cation of Disease, 9th Revision, Clinical Modi cation (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 A.3 summarizes the baseline covariates included in the study.

2.3.2 Cardiovascular Outcome Risk Trajectories

2.3.2.1 Partly and Fully Conditional Time-Varying Models without Covariates

To explore partly and fully conditional time-varying e ects, we rst consider the CV outcome risk trajectories over time from the initiation of dialysis without covariates. For this, the partly conditional GVCM is $g[E f Y_i(t_i) j S_i > t_i g] = {}_P(t_i)$, where the model ts to 3 cohorts are shown in Figure A.8(a): (i) patients who die, (ii) patients followed to the end of study (EOS), and (iii) all patients combined. Also, given are 90% bootstrap percentile con dence intervals (CIs) based on 200 bootstrap samples where entire subject trajectories are sampled with replacement. 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 A.8(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 rst 2 years after starting dialysis. This is related to the fact that the partly conditional model describes di erent (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 patients who die relative to patients alive at the end of follow-up declines in the dynamic cohort of survivors. This is illustrated in Figure A.8(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 A.8(b) displays 4 fully conditional model ts 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 strati ed 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 tted to the entire cohort, the fully conditional model can only be tted 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 deferred to Appendix (C.2)). 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 A.8(c)-(d)).

2.3.2.2 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 t the partly and fully conditional PL-GVCMs described in Section 2.2.1 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 ltration rate (eGFR). Selected kernel and bandwidth values are given in Appendix (C.3).

The estimated partly conditional VCFs before and after infection, namely $b_P(t_0)$ and $b_P(t_1)$, and the corresponding CV risk trajectories, along with 90% bootstrap CIs are given in Figures A.9(a) and A.9(b), respectively. 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 2.2.3. There is strong evidence indicating that there is di erential time-varying e ects before and after infection (both null hypotheses rejected with p-value $< .0001$). As evident from Figures A.9(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 signi cant increase in CV risk with non-overlapping CIs for $_{0;P}(t_0)$ and $_{1;P}(t_1)$. Figures A.9(c)-(f) show the estimated CV risk trajectories where the initial infection-related hospitalization occurs at 1-4 years after starting dialysis. Note that modeling e ects of vintage prior to initial infection-related hospitalization (Z_i) allows the trajectories in (c)-(f) to be di erent; nevertheless the estimated ts do not di er too much due to the small estimated value of $_{1;P}$. Results indicate 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 rst year after initiation of dialysis compared to the linear decrease after the

initial infection-related hospitalization.

Results for the fully conditional model ts, strati ed 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 A.8(b). Figures A.10(a)-(c) show the typical pattern of increased CV risk after the initial infection in the fully conditional model ts, consistently across death bins/strata. Note that the CV risk decreases for subjects in larger death bins as expected and that due to the additional strati cation of the data into death bins, the sample size used in the ts is smaller leading to larger CIs, especially after the initial infection-related hospitalization. While the partly conditional model provides information about the dynamic cohort of survivors, the fully conditional model provides an opportunity to compare estimated e ects across cohorts with di erential death strata directly.

The estimated e ects of baseline covariates $\{b_{F;P}\}$, on CV risk for a sequence of fully conditional models are summarized in Figure (A.11). Being male is associated with lower CV risk in both the fully and partly conditional model ($b_{3;P}$ = :125, 95% bootstrap CI [95% bCI]: (:143; :110)). Baseline comorbidities, including coronary heart disease ($b_{7;P}$ = :201, 95% bCI: (185; :218)) and diabetes ($b_{9;P}$ = :179, 95% bCI: (164; :198)) are associated with higher CV outcome risk in both the partly and fully conditional models. Several comorbidities, speci cally congestive heart failure ($b_{6;P}$ = :045, 95% bCI: (026; :065)) and peripheral vascular disease ($b_{8;P}$ = :093, 95% bCI: (070; :115)), in addition to baseline age ($b_{2;P}$ = :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 signi cant in most of the death bins (A.11). This may be related to some comorbidities and age being related to CV risk via their e ect 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 e ect on CV across di erential 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 ($\hat{b}_{10;P} = \ :008$, 95% bCI: (:010; :006) for BMI; $\hat{b}_{11;P} = \ :009$, 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 ($\hat{b}_{1;P} = \ :021$, 95% bCI: (:030; :010)) or the fully conditional models (Figure A.11(a)).

2.4 Simulation Studies

As described in Section 2.2.2, the fully conditional estimation involves tting 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 strati cation by death time (death bins) and the model ts within each death bin follow a standard estimation algorithm for PL-GVCM. In contrast, the partly conditional PL-GVCM is tted based on subjects who have di erential follow-up, where many individuals' follow-up times are truncated by death. Thus, our simulation studies, here will focus on the nite 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 2.2.3.

2.4.1 Simulation Model and Design

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

$$\log\{ _{i;P} =(1 \quad _{i;P})\} = \ _{0;P}(t_{0i})\{1 \quad I_{G_i}(t_i)\} + \ _{1;P}(t_{1i})I_{G_i}(t_i) + \ _{1;P}Z_i I_{G_i}(t_i) + \ _{2;P}X_{1i} + \ _{3;P}X_{2i};$$

where $_{0;P}(t) = \ :05t^2 + :025t \quad 1:25$, $_{1;P}(t) = \ :03t^2 \quad :05t$, $(_{1;P}; \ _{2;P}; \ _{3;P}) = (\ :5; :5; 1)$, and the time support is $t_{vi} \ 2 \ [0; T_{vi}]$, $T_{vi} \quad T = 5$ (for $v = 0; 1$). X_{1i} is generated from a Gamma distribution with rate parameters $\{4; 6\}$ and $X_{2i} \quad$ Bernoulli (:52). In order to generate the time-varying indicator variable, $I_{G_i}(t_i)$, we rst 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 $W_i \quad N(1:25; :25)$ 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). The simulation design mimics the real data in that within-subject correlation of the response is low ($0:04$) and truncation by death is high during the 5-year follow-up, ranging from 40-80%. The response vector and survival time are generated jointly using the bisection algorithm. The binary response $Y_{0;ik}$ and $Y_{1;ik0}$, are generated as indicators for $(Y_{0;ik} > 0)$ and $(Y_{1;ik0} > 0)$, respectively. For subjects who do not experience an infection-related hospitalization, we generate $(Y_{0;i1}; \ldots; Y_{0;i21}; S_i)^T$ according to a 22-dimensional normal distribution with mean vector $[\,_{0;i}^T; E(S_i) = 3:38]^T$ where $_{0;i} = (\,_{0;i1}; \ldots; \,_{0;i21})^T$ is the mean vector $E\{Y_i(t_i; t_{0i}; t_{1i}) | Z_i; X_i; I_{G_i}(t_i)\}$ 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 $= I_{21}; \;\; :05\,_{21}; \;\; :05\,_{21}^T; :5$, where I_a is an identity matrix of dimension a by a and $_a$ is a vector of ones of size a. Elements of the unconditional mean vector, $_{0;i}$, are computed through the correspondence:

$$_{0;ik} = E[Y_{0;ik} | S_i > t_{0ik}] = P(Y_{0;ik} > 0 | S_i > t_{0ik}) = P(Y_{0;ik} > 0; S_i > t_{0ik}) / P(S_i > t_{0ik});$$

where $_{0;ik} = g^{-1}\{\,_{0;P}(t_{0ik}) + \,_{2;P}X_{1i} + \,_{3;P}X_{2i}\}$. Through (2.4.1) $P(Y_{0;ik} > 0; S_i > t_{0ik})$ is computed via $_{0;ik} \quad P(S_i > t_{0ik})$ and we find $\{\,_{0;il}\}_{l=1}^{21}$ using the bisection method. The generated $(Y_{0;i1}; \ldots; Y_{0;i21})^T$ vector is truncated such that $t_{0;ik} \quad S_i$ 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 $(Y_{0;i1}; \ldots; Y_{0;iN_{0i}}; Y_{1;i1}; \ldots; Y_{1;i20}; S_i)^T \quad N_{N_{0i}+20}([\,_{0;i1}; \ldots; \,_{0;iN_{0i}}; \,_{1;i1}; \ldots \,_{1;i20}; E(S_i) = 3:38 + Z_i]^T; \;)$ where $N_{0i} = 4Z_i + 1$, $= \quad I_a; \;\; :05\,_a; \;\; :05\,_a^T; :5$ with

$a = N_{0i} + 20$. The unconditional means $(\ _{0;i1}; \ldots;\ _{0;iN_{0i}})$ are calculated as described above and $(\ _{1;i1}; \ldots;\ _{1;i20})$ are calculated similarly by the bisection method using $P(Y_{1;ik^0} > 0; S_i > Z_i + t_{1ik^0}) =\ _{1;ik^0}\ \ P(S_i > Z_i + t_{1ik^0})$ where $_{1;ik^0} = g^{-1}f\ _{1;P}(t_{1ik^0}) +\ _{1;P}Z_i +\ _{2;P}X_{1i} +\ _{3;P}X_{2i}g$. The generated $(Y_{1;i1}; \ldots; Y_{1;i21})^T$ vector is truncated such that $Z_i + t_{1;ik^0}\ \ S_i$ to create the observed outcomes after the pivotal exposure for $k^0 = 1; \ldots; N_{1i}$.

2.4.2 Simulation Results

2.4.2.1 Estimation

We generated 200 datasets at sample sizes of $n = 500$ and 2000. For the estimation, we utilized a Epanechnikov kernel and chose bandwidths 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 $\hat{\ }_{0;P}(t_0)$, $\hat{\ }_{1;P}(t_1)$ at $n = (500; 2000)$, respectively. To study the performance of the proposed estimation procedure, we utilize a relative mean squared deviation error (MSDE) de ned as

$$\text{MSDE}_v = \int_0^T f\ _{v;P}(t_v)\ \ b_{v;P}(t_v)g^2 dt_v \ \Big/ \int_0^T\ _{v;P}^2(t_v)dt_v$$

for the VCFs, $v = 0; 1$, and mean squared error MSE_r for the constant coe cients $f\ _{r;P}g_{r=1}^3$. The median and rst and third quartiles of the estimated MSDE and MSE measures over 200 Monte Carlo runs are presented in Table (A.4). The MSDE and MSE values are relatively small and decrease with increasing sample size, indicating the overall e ectiveness 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 (A.13) 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.

2.4.2.2 Hypothesis Tests

We also examine the validity and power of the two proposed GLRTs, namely Test I: H_0 : $_{0;P}(t_0) = c_0$ and $_{1;P}(t_1) = c_1$ and Test II: H_0 : $_{0;P}(t_0) = {}_{1;P}(t_1)$ (illustrated in Figure A.12(a)-(b), respectively) for longitudinal data under high levels of truncation by death, similar to our data application (ranging from 40-80%).

We rst 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 ^{2}-distribution and does not depend on the speci c null values considered. For Test I, we consider 5 different sets of null values: $(c_0; c_1)$ 2 f (1; 1); (1; 0); (0; 1); (0; 1); (1; 0)g. The parametric bootstrap procedure (Section 2.2.3) is used for $r = 500$ 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 A.12(c) along with the density of the ^{2}-distribution. The degrees of freedom of the ^{2}-distribution is chosen to be close to the sample mean of the bootstrap test statistic values across null con gurations. The plotted densities of T are close to the ^{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 H_1 : $_{0;P}(t_0) = c_0(1 \quad) + {}^0_0(t_0)$ and $_{1;P}(t_1) = c_1(1 \quad) + {}^0_1(t_1)$ where ${}^0_0(t) = \quad :05t^2 + :025t \quad 1:25$, ${}^0_1(t) = \quad :03t^2 \quad :05t$, 2 [0; 1], $c_0 = E[\ _0(t_0)]$ and $c_1 = E[\ _1(t_1)]$. Similarly, for Test II, we consider the alternative H_1 : $_{0;P}(t_0) = (1 \quad) \ {}^0_0(t_0) + {}^0_0(t_0)$ and $_{1;P}(t_1) = (1 \quad) \ {}^0_0(t_1) + {}^0_1(t_1)$ where ${}^0_0(t) = \quad :05t^2 + :025t \quad 1:25$, ${}^0_1(t) = \quad :03t^2 \quad :05t$ and 2 [0; 1]. Note that in both cases, larger values of correspond to further deviations from the null. Figure A.12(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 $r = 500$. Similarly Figure A.12(e) gives the 3 power curves for Test II. As expected, the power increases with e ect size" () 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 (= 0, Type I error), which should coincide approximately with the level of the test. For Test I at signi cance 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.

2.5 Discussion

In this work, we proposed partly and fully conditional approaches to modeling time-varying e ects 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 e ect 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 tting a sequence of stratum-speci c time-varying e ect models (within death bins). Thus, the later approach enables direct comparison of time-varying e ects for each death bin/stratum as well as variation in baseline covariate e ects on outcome across death bins.

We note that other modeling approaches to the current problem of addressing truncation by death include modeling the times to CV events (multiple events per subject) using the counting process formulation of a time-varying e ect Anderson-Gill survival model. Another alternative approach is the joint modeling of survival and the longitudinal outcome of CV risk. Nevertheless, implementing these approaches for a data set of the size of the USRDS data is a major computational challenge since survival modeling is commonly built on likelihood based approaches that use the entire data. Time-varying extensions of survival modeling with time-varying coe cients modeled through basis expansions require that the

data be rearranged into a repeated measures format. Hence a data set with 105,000 CV events can lead to an expanded dataset with more than 12 million records. The joint modeling approaches similarly do not scale up to the current USRDS data size. The proposed estimation procedure for PL-GVCM modeling addresses this computational challenge by using a local likelihood approach where only data from a small neighborhood in time is used in maximizing the likelihood. This local log-likelihood approach is key in handling the big data size in our application. In addition, while for most joint modeling approaches the interest is on modeling both the survival and the longitudinal outcome, our focus is primarily on studying the e ects of infections on CV risk, hence the proposed approach is a more direct one in addressing this goal without the need to explicitly model survival.

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. Motivated by our data application with low within-subject correlation, we have shown the e cacy of the proposed test through studying the empirical estimates of power and validity via simulations in cases with low within-subject correlation. Preliminary works of Li et al. (2015) show that Wilks phenomenon hold for longitudinal data if a working independence correlation structure is assumed in the test statistic and that it does not hold when the working variance function is misspeci ed. In addition they point to the fact that incorporating the correlation into the test statistic does not necessarily improve the power of the test. Further research is needed for incorporation of the within-subject correlation in estimation of partly conditional models as well as extensions of generalized likelihood ratio tests for semiparametric modeling of longitudinal data.

We provide R codes for the proposed partly conditional and fully conditional PL-GVCM at http://dsenturk.bol.ucla.edu/PLVCM_algorithm.pdf.

CHAPTER 3

Time-Dynamic Dialysis Facility Pro ling

3.1 Introduction

As of 2011, end-stage renal disease (ESRD) a ected more than 615,000 adults in the U.S. Of these, more than 430,000 were on dialysis, a life-sustaining treatment (United States Renal Data System Annual Data Report [USRDS ADR], 2013). On average, dialysis patients are admitted to a hospital nearly twice a year and hospitalizations account for approximately 38% of total Medicare expenditures for dialysis patients [USRDS ADR, 2013]. An unplanned hospital readmission is de ned as any unplanned hospital admission that occurs within 30 days of discharge from a previous admission. Approximately 30% of patients with end stage renal disease (ESRD) discharged from a hospital have an unplanned readmission, and previous studies have shown that a good portion of unplanned readmissions are preventable [USRDS ADR, 2013]. In this paper our objective is to propose a time-dynamic measure of unplanned hospital readmissions at dialysis facilities to identify potential facility speci c problems that contribute to rising costs in health care of dialysis patients.

This objective relates to a broader national e ort under the new health care legislation (A ordable Care Act), care provider monitoring (`pro ling'), which aims to evaluate a provider's quality of care to a reference standard (e.g., overall `national' average standard) with the main goal of identifying providers that deviate in important ways in the delivery of patient care (Ash et al. 2011; Normand et al. 1997; He et al. 2013; Horwitz et al. 2011; Kalb eisch and Wolfe 2013). Care provider performance indices are usually tied to regulation and payment, and have largely been time-static, summarized by a single numeric measure. Examples include medical care provider pro ling using standardized mortality ratios (e.g., among U.S. hospitals or coronary artery bypass graft surgery (Ash et al. 2011; Normand et al. 1997; Krumholz et al. 2011; Normand and Shahian 2007; Lin et al. 2009; Liu et al.

2003; Paddock et al. 2006; Yang et al. 2014)) and standardized hospital readmission ratios (e.g., hospitals, dialysis facilities (He et al. 2013, Horwitz et al. 2011; Kalb eisch and Wolfe 2013; CMS 2014)). The only time-varying performance metrics in the literature are the risk-adjusted CUmulative SUMmation (CUSUM and observed-expected CUMSUM) techniques (Biswas and Kalb eisch 2008; Sun and Kalb eisch 2013), although these are tailored to survival time outcomes.

As a novel departure from prior literature, we introduce time-dynamic pro ling to provide a continuous metric for monitoring performance over time for generalized patient outcomes (hospital readmission rates). Our objective is to evaluate a facility's performance as a function of time that patients are on dialysis, relative to a national standard, accounting for di erences in patient-level characteristics prior to starting dialysis (case-mix). The proposed approach is particularly relevant for chronic conditions, such as ESRD, where it is critical to understand the outcome trajectories throughout the time period that patients are on dialysis. In addition it provides more informative feedback to care providers for quality improvement since it can identify speci c time periods of under- or over-performance during dialysis relative to a reference standard.

Our proposal of time-dynamic pro ling is based on the proposed multilevel varying co-e cient model (MVCM) with xed facility time-varying e ects and subject speci c random e ects to accommodate the multilevel data structure of the data with patients nested within dialysis facilities, and observations over time nested within patients. Inclusion of xed versus random facility e ects have been studied in the time-static pro ling literature. While the average absolute error in estimation is typically smaller overall for random e ects models, this average gain is likely achieved mostly in the center of the distribution of the outcomes. When the interest is on inference for the more extreme facilities deviating from a standard norm, as in pro ling, xed e ects models have been reported to be e ective in identifying outliers in dialysis facility pro ling (Kalb eisch and Wolf 2013). An additional advantage of xed e ects models is that they do not have the inherent problem of confounding of patient and facility-level e ects in multilevel modeling for dialysis facility pro ling. To accommo-

date time dependence among the repeated measures obtained on the same individuals, we introduce subject speci c random e ects. He et al. (2013) also evaluates the impact of discharging hospitals on static pro ling of dialysis facilities based on hospital readmissions. Authors propose an additional hospital level random e ect, but report minimal changes in dialysis facility pro ling results in applications to USRDS data; hence we opt out of this additional level of complexity in the proposed extension to the time-dynamic pro ling setting.

Standard hierarchical methods, designed to provide inference and prediction in the setting of multilevel longitudinal data, have proven to be indispensible for modeling time-static e ects (xed coe cients) (Gelman 2006, refs. therein). However, for modeling multilevel time-dynamic/time-varying coe cients the available methods are limited in scope and applicability. An important tool in exploring time-dynamic patterns is varying coe cient models (VCMs) (Cleveland et al. 1991; Hastie and Tibshirani 1993) which are natural extensions of parametric models with a diverse array of applications in biomedical science, epidemiology, ecology, and economics (Fan and Zhang 2008). The literature on the standard VCMs, generalized VCMs (GVCMs) (Cai et al. 2000; Qu and Li 2006; Senturk et al. 2013), and their adaptations for analyzing longitudinal data (Hoover et al. 1998; Estes et al. 2014) have largely been for `single-level' data. Limited works have considered mixed VCMs (Zhang 2004; Li et al. 2012; Liang et al. 2003; Rice and Wu 2001; Wu and Liang 2004; Wu and Zhang 2002), mostly for the analysis of regular longitudinal data (i.e., without the higher-level unit). Recently there is growing literature on multilevel functional linear models and Crainiceanu et al. (2009) considered functional predictors but in modeling a scalar response. A multilevel VCM for a space and time varying response and space and time varying predictors was proposed in (Serban 2011). The model was proposed for modeling service accessibility outcomes identi ed by location and time and was not proposed for analysis of patient-level longitudinal data, which is a key necessity in modeling multilevel patient outcome data. Another work examined marginal regression models in a randomized clinical trial (Chen and Wang 2010) for multilevel functional data via penalized spline estimating equations.

The proposed MVCM and time-dynamic facility performance index, standardized dy-

namic ratio (SDR), are developed in Section 3.2. We also develop a hypothesis testing procedure in Section (3.2.3) suitable for pro ling goals. Novel studies based on USRDS data to determine the time-dynamic facility performance index for dialysis facilities in the U.S. and simulation studies are given in Sections 3 and 4, respectively. We also include comparisons of the performance of the proposed model to one ignoring within subject correlation and conclude with a brief discussion section.

3.2 Multilevel Varying Coe cient Modeling With Subject Spe-ci c Random E ects

Let $i = 1; \ldots; I$ index dialysis facilities, $j = 1; \ldots; N_i$ index subjects belonging to the ith facility having N_i total subjects, and $k = 1; \ldots; N_{ij}$ index hospitalizations for the jth subject belonging to the ith facility having N_{ij} total hospitalizations. Let the outcome $Y_{ijk} \quad Y_{ij}(t_{ijk})$ equal 1 if the kth index hospitalization of the jth patient within facility i results in a readmission within 30 days, and equal 0 otherwise. There are several important aspects in building e ective models for the outcome of hospital readmissions for the goal of time-dynamic dialysis facility pro ling. First, the model needs to respect the multilevel structure of the data and needs to allow for time-varying e ects of facilities. Second, longitudinal subject-level predictors or post-dialysis cross-sectional covariates (such as adverse events during dialysis or patient health attributes after dialysis which might be the result of care) needs to be excluded to avoid confounding with the time-dynamic facility-level e ects, the main quantity of interest. Third, the extent to which the proposed modeling will be useful will depend on having a set of rich baseline covariates $Z_{ij}(= (Z_{1ij}; \ldots; Z_{rij})^T$ for the jth patient within facility i) that together captures patient heath characteristics prior to dialysis (case-mix).

Hence, to achieve the above goals we propose the logistic MVCM

$$g\ E\ Y_{ij}(t)\ j\ Z_{ij}; b_{ij} \quad = g f p_{ij}(t) g = \ _i(t) + b_{ij} + Z_{ij}^T; \tag{3.1}$$

where g is the logit link function, the parameters $_i(t)$ correspond to the xed time-varying facility level e ects, $b_i = (b_{i1}; : : : ; b_{iN_i})^T$ correspond to subject speci c random e ects within the ith facility, $= (_1; : : : ; _r)^T$ is a vector of regression parameters, and $p_{ij}(t) \quad E \; Y_{ij}(t) \; j \; Z_{ij} ; b_j \quad = g^{1} \quad _i(t) + b_j + Z_{ij}^T \quad$. The MVCM in (3.1), to which we will refer to as Model 1, extends the standard xed e ects logistic regression model of He et al. (2013) for time-static dialysis facility pro ling, to model time-dynamic e ects via facility-level varying coe cient functions.

3.2.1 Standardized Dynamic Ratio (SDR)

To capture the performance of the ith facility relative to a reference standard accounting for patient case-mix, we introduce the standardized dynamic ratio

$$SDR_i(t) = \frac{\sum_{j=1}^{n_i} p_{ij}(t)}{\sum_{j=1}^{n_i} p_{ij;M}(t)}; \tag{3.2}$$

where $p_{ij}(t) = g^{1} f \; _i(t) + b_j + Z_{ij}^T \; g$ and $p_{ij;M} = g^{1} f \; _M(t) + b_j + Z_{ij}^T \; g$ with $_M(t)$ denoting the cross-sectional median of $f \; _1(t); : : : ; \; _I(t) g$. In (3.2), the facility performance (numerator) is obtained from the patient-level risk model de ned in (3.1), referred to as the facility-speci c model. The national standard (denominator) for the types of patients treated at the dialysis facility (its case-mix) is obtained from the national-level model with $_M(t)$. The SDR is the ratio of the sum of the expected number of 30-day readmissions at facility i at time t over the subjects treated at the facility to the sum of the expected readmissions of a counterfactual median facility at time t again over the case-mix of facility i. To produce a fair assessment, both the facility-speci c and national models will account for case-mix prior to dialysis. These adjustments for di erences in patients' health at baseline are required to ensure that variations in reported performance apply to facility's contributions to their patients outcomes rather than to the intrinsic di culty of the patients they treat. This is also the reason why adjustments are not made for longitudinal subject level predictors post-initiation of dialysis, since they are on the pathway to the outcome and adjusting for

them would reduce the magnitude of the facility e ects.

In (3.2), $SDR_i(t) = 1$ would indicate typical performance for a given case-mix, where each facility performance is compared to a population having the same case-mix. This feature protects against extrapolation outside of the facility's treated case-mix; importantly, pair-wise comparisons of $SDR_i(t)$ with $SDR_{i^0}(t)$ would only be meaningful to the extent that the distributions of Z_{ij} and Z_{i^0j} overlap and are balanced.

A natural estimator of the proposed $SDR_i(t)$ is $\hat{SDR}_i(t) = \sum_{j=1}^{n_i} \hat{p}_{ij}(t) = \sum_{j=1}^{n_i} \hat{p}_{ij;M}(t)$ where $\hat{p}_{ij}(t) = g^{-1}\{b_i(t) + \hat{\theta}_{ij} + Z_{ij}^T b\}$, $\hat{p}_{ij;M} = g^{-1}\{b_M(t) + \hat{\theta}_{ij} + Z_{ij}^T b\}$ and $\hat{\theta}_{ij}$ are predicted subject speci c random e ects obtained from the means of the posterior distributions of θ_{ij}. Estimators of the model parameters , $_1(t); : : : ; _I(t)$ and predicted random e ects, will be obtained via an approximate EM algorithm due to the large number of facilities, alternating between the estimation of $_i(t)$, , and the predicted random e ects until convergence as outlined in the next section.

3.2.2 Estimation

To develop the intuition behind the proposed estimation algorithm, rst consider a simpler logistic MVCM without subject speci c random e ects,

$$g\{E(Y_{ij}(t) \mid Z_{ij})\} = g\{p_{ij}(t)\} = {}_i(t) + Z_{ij}^T ; \tag{3.3}$$

with the likelihood function

$$L\{ {}_1(t); : : : ; {}_I(t); \} = \prod_{i=1}^{I} \prod_{j=1}^{n_i} \prod_{k=1}^{N_{ij}} \frac{\exp\{ {}_i(t_{ijk}) + Z_{ij}^T \} Y_{ijk}}{1 + \exp\{ {}_i(t_{ijk}) + Z_{ij}^T \}} : \tag{3.4}$$

We will utilize this model in comparisons to the proposed Model 1 with subject speci c random e ects, and will refer to this model as Model 2 throughout the paper. Facility speci c xed e ect functions $_i(t)$ can be approximated locally and derived parameters can be estimated via maximization of the local likelihoods. However, when the number of facilities are high (nearly 4000 dialysis facilities in our data application), maximizing local likelihoods poses a serious computational challenge. Nevertheless, the likelihood in (3.4) is separable

through the outmost product over facilities into I components where the ith component depends only on $\{\beta_i(t), \gamma\}$. Hence, given γ, $\beta_i(t)$ can be estimated by maximizing the local likelihood based on data only from the ith facility; and given $\{\beta_1(t), \ldots, \beta_I(t), \gamma\}$, can be estimated based on a global likelihood without the need for localization. Therefore a sequential approach can easily be implemented.

Now consider the proposed logistic MVCM with subject specific random effects given in (3.1). Let $b = (b_{ij} : i = 1, \ldots, I; j = 1, \ldots, N_i)^T$ denote the vector of independent and identically distributed subject specific random effects with mean zero and variance σ_b^2. Viewing the subject specific random effects as missing data, we propose an approximate EM algorithm. The complete likelihood corresponding to Model 1 is

$$L\{b, \sigma_b, \beta_1(t), \ldots, \beta_I(t), \gamma\} = \prod_{i=1}^{I} \prod_{j=1}^{N_i} \prod_{k=1}^{N_{ij}} \frac{\exp\{\beta_i(t_{ijk}) + b_{ij} + Z_{ij}^T \gamma\} Y_{ijk}}{1 + \exp\{\beta_i(t_{ijk}) + b_{ij} + Z_{ij}^T \gamma\}} \frac{\exp\{-b_{ij}^2/(2\sigma_b^2)\}}{\sqrt{2\pi\sigma_b^2}}. \tag{3.5}$$

The incomplete (or observed) likelihood that is available for the estimation of $\{\sigma_b, \beta_1(t), \ldots, \beta_I(t), \gamma\}$ is

$$L\{\sigma_b, \beta_1(t), \ldots, \beta_I(t), \gamma\} = \prod_{i=1}^{I} \prod_{j=1}^{N_i} \int_{-\infty}^{\infty} L_{ij}\{b_{ij}, \sigma_b, \beta_i(t), \gamma\}\, db_{ij} \tag{3.6}$$

where

$$L_{ij}\{b_{ij}, \sigma_b, \beta_i(t), \gamma\} = \prod_{k=1}^{N_{ij}} \frac{\exp\{\beta_i(t_{ijk}) + b_{ij} + Z_{ij}^T \gamma\} Y_{ijk}}{1 + \exp\{\beta_i(t_{ijk}) + b_{ij} + Z_{ij}^T \gamma\}} \frac{\exp\{-b_{ij}^2/(2\sigma_b^2)\}}{\sqrt{2\pi\sigma_b^2}}.$$

Further, let Y be the vector of all outcomes Y_{ijk}, $i = 1, \ldots, I$, $j = 1, \ldots, N_i$, $k = 1, \ldots, N_{ij}$. The posterior distribution of b_{ij} given the data and $\{\beta_i(t), \gamma, \sigma_b\}$ is

$$D_{ij}\{b_{ij} \mid Y, \sigma_b, \beta_i(t), \gamma\} = \frac{L_{ij}\{b_{ij}, \sigma_b, \beta_i(t), \gamma\}}{\int_{-\infty}^{\infty} L_{ij}\{b_{ij}, \sigma_b, \beta_i(t), \gamma\}\, db_{ij}}.$$

Hence the posterior mean and variance of b_{ij} are $\hat{b}_{ij0} = C_{ij}^{-1} \int_{-\infty}^{\infty} b_{ij} L_{ij}\{b_{ij}, \sigma_b, \beta_i(t), \gamma\}\, db_{ij}$ and $\hat{v}_{ij0} = C_{ij}^{-1} \int_{-\infty}^{\infty} (b_{ij} - \hat{b}_{ij0})^2 L_{ij}\{b_{ij}, \sigma_b, \beta_i(t), \gamma\}\, db_{ij}$, respectively, where $C_{ij} = \int_{-\infty}^{\infty} L_{ij}\{b_{ij}, \sigma_b, \beta_i(t), \gamma\}\, db_{ij}$. We use a Gauss-Hermite quadrature calculation with 20 quadrature points to numerically approximate $\hat{b}_{ij0}$ and $\hat{v}_{ij0}$ (Lange, 1999).

The E-step of the proposed approximate EM algorithm pertains to the calculation of the conditional expectation of the complete log-likelihood. Because the closed form for $E[\log L(\{b_j\}; \beta_b; \beta_1(t); \ldots; \beta_I(t); \gamma \mid Y; \beta_b; \beta_1(t); \ldots; \beta_I(t); \gamma)] = \sum_{i=1}^{I}\sum_{j=1}^{J} E[\ell_{ij}(\{b_j\}; \beta_b; \beta_i(t); \gamma \mid Y; \beta_b; \beta_i(t); \gamma)]$ is not available where $\{\beta_b; \beta_1(t); \ldots; \beta_I(t); \gamma\}$ are the current estimates of the parameters and $\ell_{ij}(\{b_j\}; \beta_b; \beta_i(t); \gamma) = \log L_{ij}(\{b_j\}; \beta_b; \beta_i(t); \gamma)$, we approximate $\ell_{ij}(\{b_j\}; \beta_b; \beta_i(t); \gamma)$ using a second order Taylor's expansion about $\{b_{j0}\}$ to obtain the approximate expected log-likelihood

$$\sum_{i=1}^{I}\sum_{j=1}^{J} E\left[\ell_{ij}(\{b_j\}; \beta_b; \beta_i(t); \gamma) \mid Y; \beta_b; \beta_i(t); \gamma\right] \approx \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{n_{ij}} Y_{ijk}\left(\beta_i(t_{ijk}) + b_{j0} + Z_{ij}^T\right.$$

$$\left. + \log\left(\frac{\gamma}{2}\right) p_{0;ijk}\left(\gamma\right) q_{0;ijk}\left(\frac{v_{ij0}}{2}\right) - \frac{(b_{j0})^2 + v_{ij0}}{2(\sigma_b)^2} - \frac{1}{2}\log\left(2\pi(\sigma_b)^2\right)\right. \tag{3.7}$$

where $p_{0;ijk} = \left(\gamma\right)^{-1} f(\beta_i(t_{ijk}) + b_{j0} + Z_{ij}^T\gamma)$, $q_{0;ijk} = 1 - p_{0;ijk}$, b_{j0} and v_{ij0} are the posterior mean and variance of $\{b_j\}$ given the current parameter estimates. Detailed derivations are deferred to Appendix D.1.

The M-step maximizes the expectation of the complete log-likelihood utilizing the approximation in (3.7). A key observation is that the approximation to the expected value of the complete log-likelihood is again separable through the outmost sum over facilities into I components where the ith component depends only on $\{\beta_i(t); \gamma; \beta_b\}$ and the posterior mean b_{j0} and variance v_{ij0}, similar to the likelihood in (3.4). Hence while joint maximization with respect to all model parameters is a big computational challenge for a large number of facilities, a sequential approach similar to the one discussed for Model 2 can be easily implemented. We begin with the estimation of β_b by maximizing the approximation of the expected global log-likelihood with respect to β_b. In the next step, we estimate $\beta_i(t)$ by maximizing the approximation of the expected local log-likelihood using data from only the ith facility, given the current estimates γ; β_b, and $\{b_{j0}; v_{ij0}\}$ via a one-step Newton-Raphson iteration. Finally, γ is estimated by maximizing the approximation of the expected global log-likelihood using all estimated quantities again utilizing a one-step Newton-Raphson iteration. The proposed estimation algorithm is summarized by the following steps.

(1) Set initial values for $^{(0)}$, $_i^{(0)}(t)$, $_b^{(0)}$.

(2) Estimate the posterior means $b_{ij\,0}^{(m)}$ and variances $v_{ij\,0}^{(m)}$ of the subject speci c random e ects in the mth iteration step using $_i(t) = {}_i^{(m\ \ 1)}(t)$, $\ = {}^{(m\ \ 1)}$, and $_b = {}_b^{(m\ \ 1)}$ from the previous iteration step. Use the estimated posterior means and variances of the random e ects to approximate the expected value of the complete log-likelihood.

(3) Maximize the approximation of the expected log-likelihood with respect to $_b$ to derive at the estimator

$$_b^{(m)} = \left(\sum_{i=1}^{I} N_i \right)^{1} \sum_{i=1}^{I} \sum_{j=1}^{N_i} \left\{ (b_{ij\,0}^{(m)})^2 + v_{ij\,0}^{(m)} \right\}^{1=2} :$$

(4) Approximate $_i(t)$ locally via $_i(t) \quad {}_{0i} + {}_{1i}(t \quad t_0)$ for t in the neighborhood of a xed t_0 leading to the part of the approximate local expected complete log-likelihood which depends on only data from the ith facility given by

$$\sum_{j=1}^{p_i} \sum_{k=1}^{p_j} Y_{ijk} \quad {}_{0i} + {}_{1i}(t_{ijk} \quad t_0) + b_{ij\,0} + Z_{ij}^T \quad + \log(\bar{q}_{0;ijk}) \quad \frac{v_{ij\,0}}{2} \bar{p}_{0;ijk}\, \bar{q}_{0;ijk}$$

$$\frac{(b_{ij\,0})^2 + v_{ij\,0}}{2N_{ij}\,(\ _b)^2} \quad \frac{1}{2N_{ij}} \log \ 2\ (\ _b)^2 \quad K_h(t_{ijk} \quad t_0)$$

where $\bar{p}_{0;ijk} = g^{\ 1}f\ {}_{0i} + {}_{1i}(t_{ijk} \quad t_0) + b_{ij\,0} + Z_{ij}^T \quad g$, $\bar{q}_{0;ijk} = 1 \quad \bar{p}_{0;ijk}$ and $K_h(\ ) = K(\ =h)=h$ with $K(\ )$ denoting a kernel function and h the bandwidth. Estimate $_i^{(m)}(t)$ by a one-step Newton Raphson algorithm maximizing the above approximate local likelihood with respect to $(\ _{0i};\ _{1i})$ based on data from only the ith facility and current estimates of the other parameters.

(5) Estimate $^{(m)}$ by a one-step Newton Raphson algorithm maximizing the approximate expected global log-likelihood using estimated quantities from the above steps.

(6) If $\max_{i;j;k} |p_{ijk}^{(m)} \quad p_{ijk}^{(m\ \ 1)}| > \ $, where $\ $ is some tolerance level and $p_{ijk}^{(m)} = g^{\ 1}f\ {}_i^{(m)}(t_{ijk}) + b_{ij\,0}^{(m)} + Z_{ij}^T\ {}^{(m)}g$, set $m = m+1$ and go back to Step 2. The stopping rule is motivated by

our de nition of SDR which directly uses predicted probabilities that can be sensitive to change (compounding e ect) had tolerance levels been speci ed for each parameter estimate individually.

Explicit expressions of the estimators and derivations are deferred to Appendix D.1. More information on the selection of the bandwidths in Step 4 are deferred to Sections 3.3.2 and 3.4.1.

3.2.3 Hypothesis Testing

Making statistical inference about our proposed time-dynamic facility performance index, $SDR_i(t)$, is challenging given its de nition and the high-dimensional nature of the xed parameters. A particular clinical interest, motivated by rising costs and e ective health care, is to identify facilities that deviate from the national reference facility, a hypothetical facility whose e ect $_M(t)$ is de ned by taking cross-sectional median facility e ects across time. For facilities whose performance across time coincides with the national reference facility, $SDR_i(t)$ will be a constant function equal to 1 across time. Time intervals at which $SDR_i(t) < 1$ would indicate that the facility's predicted readmission rates are less than expected based on national rates. Time intervals at which $SDR_i(t) > 1$ would indicate that the facility's predicted readmission rates are greater than expected based on national rates. Thus, $H_0 : SDR_i(t) = 1$ for all t is a hypothesis test of interest. We note that $SDR_i(t) = 1$ implies $\sum_{j=1}^{N_i} p_{ij}(t) = \sum_{j=1}^{N_i} p_{ij;M}(t)$ which also implies $_i(t) = _M(t)$. Thus, we de ne our test statistic to be

$$R_i = \int \left(\sum_{j=1}^{N_i} \hat{p}_{ij}(t) - \sum_{j=1}^{N_i} \hat{p}_{ij;M}(t) \right)^2 dt \Big)^{1/2};$$

which quanti es the departure of $\sum_{j=1}^{N_i} \hat{p}_{ij}(t)$ from $\sum_{j=1}^{N_i} \hat{p}_{ij;M}(t)$ using the L^2 norm. For facilities whose performance across time coincides with the national reference facility, R_i will be equal to 0 and otherwise it will be positive. We use a resampling procedure to estimate the distribution of the test statistic, and hence a nominal p-value described as follows.

(i) We x , $_M(t)$ and $_b$ at their estimated values $\hat{b}$, $\hat{b}_M(t)$, and $\hat{b}_b$ (using the proposed estimation algorithm given in Section 3.2.2) in the steps below since their estimation uses the entire data set which will be large in our data application.

(ii) Approximate $\hat{b}_{ij0}$ and $\hat{v}_{ij0}$ under the null $H_0 : {}_i(t) = {}_M(t)$ using the posterior distribution $D_{ij}\{b_{ij} \mid Y; \hat{b}_b; \hat{b}_M(t); \hat{b}\}$ de ned in Section 3.2.2. Then generate subject random e ects from the subject-speci c posterior distribution approximated by a normal distribution with mean and variance $\hat{b}_{ij0}$ and $\hat{v}_{ij0}$ respectively. Thus, we draw an independent sample of size S,

$$b_{ij}^s \quad N(\hat{b}_{ij0}; \hat{v}_{ij0}); \; s = 1; \ldots; S;$$

for each subject $j = 1; \ldots; N_{ij}$ within facility i.

(iii) Draw S samples $\{Y_{ijk}^s : j = 1; \ldots; N_i; k = 1; \ldots; N_{ij}\}$ where each observation is generated under the null, independently from a Bernoulli distribution

$$Y_{ijk}^s \mid b_{ij}^s \quad \text{Ber} \; \frac{\exp\{\hat{b}_M(t_{ijk}) + b_{ij}^s + Z_{ij}^T \hat{b}\}}{1 + \exp\{\hat{b}_M(t_{ijk}) + b_{ij}^s + Z_{ij}^T \hat{b}\}} \; :$$

(iv) Calculate the value of the test statistic R_i for each sample in (iii) by iterating between steps 2, 4, and 6 of the proposed estimation algorithm given in Section 3.2.2 with ${}_i(t) = \hat{b}_M(t)$, $= \hat{b}$, and ${}_b = \hat{b}_b$. We skip steps 1, 3, and 5 since we are xing, $_M(t)$ and $_b$ at the large sample estimates $\hat{b}$, $\hat{b}_M(t)$, and $\hat{b}_b$ (and therefore do not require these parameters to be estimated) and denote the observed test statistics for each resampled data by r_i^s. We note that this reduced resampled estimation procedure only relies on data from the ith facility, and does not require the entire data set. This leads to major computational savings in implementation.

(v) We use $(1/S)\sum_{s=1}^{S} I\{r_i^s > r_i\}$ to approximate the nominal p-value $P(R_i \quad r_i \mid H_0)$, where $I\{A\}$ is the indicator function for event A.

3.3 Applications to USRDS Data

3.3.1 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 de ned population of inference in our study are dialysis facilities who service patients aged 18 and older who newly initiated dialysis between January 1, 2006 and December 31, 2009. Patients were eligible for inclusion if they survived the rst 90 days of dialysis. Thus, the observation period began on day 91 of dialysis. Our study cohort consisted of 615,251 discharges nested within 189,943 patients nested within 3,901 dialysis facilities. Of these discharges, 189,309 resulted in an unplanned hospital readmission within 30 days of discharge yielding an overall readmission rate of 30.8%. The number of subjects per facility varied from 20 to 192, with a mean of 49 and a median of 43 patients. The inclusion/exclusion rules included reducing the cohort based on facility size, availability of baseline covariates and sparsity (less than 10 discharges per year per facility) of follow-up data to facilitate the time-varying e ects estimation. We also truncated follow-up after a patient switches dialysis facilities. For a complete ow chart of inclusion/exclusion rules, we refer the reader to Figure A.16. The maximum of follow-up was three years where 57% of the initial cohort was observed throughout the full three years.

3.3.2 Results

We used Models 1 and 2 to estimate the time-dynamic facility performance index $SDR_i(t)$ for each facility using the patient-level adjustments for age, sex, body mass index at incidence of ESRD, diabetes as cause of ESRD, and an additional thirty-two past-year comorbidities (from day 91 on dialysis) and discharge diagnoses that are rare but have a high rate of readmission. We selected a separate bandwidth for each facility using 10-fold cross-validation with candidate bandwidths ranging from .8 to 1.3 (in years). The cross-validation compared

observed outcomes to the predicted probabilities $g^{-1}\{b_i(t_{ijk}) + Z_{ij}^T b\}$ estimated without the random e ects using the entire data set excluding data from the left out subjects as similarly de ned in Wu and Zhang 2002. In Model 1, the variance of the random e ects was estimated to be $b_b^2 = :68$. Thus, the estimated time-dynamic facility performance indices $\hat{SDR}_i(t)$ were di erent between the two models. These di erences, strati ed by facility size, are displayed in Figure A.14 using the L2-norm ratio $[\int_0^{R_3}\{\hat{SDR}_{1i}(t) - \hat{SDR}_{2i}(t)\}^2 dt]^{1=2} = [\int_0^{R_3}\hat{SDR}_{1i}^2(t)dt]^{1=2}$ as a di erence measure where $\hat{SDR}_{1i}(t)$ is the estimate of $SDR_i(t)$ under Model 1 and $\hat{SDR}_{2i}(t)$ is the estimate of $SDR_i(t)$ under Model 2. The gure shows that the di erences decrease with increasing facility size which is due to the fact that larger facilities contain more subjects.

Next we perform the test $SRD_i(t) = 1$ for each facility $i = 1; : : : ; 3901$ using both Models 1 and 2 to identify a subset of facilities as deviating from a national standard across time. There are three cases to consider: (a) the agged facility consistently overperforms ($\hat{SDR}_i(t) < 1$ for all t), (b) the agged facility consistently underperforms ($\hat{SDR}_i(t) > 1$ for all t), and (c) the agged facility under or over performs during a subset of the observation period. In case (c), further investigation will need to be conducted. One can easily extend the proposed hypothesis testing procedure to provide insights into facility performance over speci c time regions of follow up instead of the follow up. Table A.8 presents the pairwise comparison of the number and percentage of outlier facilities identi ed by the nominal p-values calculated under Models 1 and 2. For example, 118 out of 532 facilities agged signi cantly worse (or other) than the national standard under Model 1 were not agged under Model 2. On the other hand, 172 out of 583 facilities agged signi cantly worse (or other) than the national standard under Model 2 were not agged under Model 1. This shows that adjusting for subject speci c random e ects has some in uence on the classi cation of facility outlier status, and this in uence is directly attributable to the estimation of the time-dynamic SDRs (as displayed in Figure A.14). We also note that further investigations (not included in this paper) revealed a higher proportion of agreement between Models 1 and 2 in agging outlier facilities among large facilities due to the fact that larger facilities

contain more subjects.

To adjust for simultaneously testing a large number of facilities, we employ the method discussed in Kalb eish and Wolfe (2013) which accounts for unexpected over dispersion in the data based on the empirical null (Efron 2004, 2007). We start by converting the nominal p-value for each facility to a Z-score using the inverse cumulative distribution function $^{1}(\ )$ of a standard normal random variable. Then, the z-scores are strati ed into three groups (small: 20 - 34 patients; medium: 35 - 54 patients; large: 55 - 192 patients) based on facility size were cut-o s are determined by the tertiles of the distribution of the number of patients. To rank facility test statistics with an estimated p-value of 0, we utilize the z-scores of their test statistics from the hypothesis tes $H_0 : SDR_i(t) = 1$. The probability density function of a standard normal random variable is then superimposed onto each of the three histograms of the strati ed z-scores. We then t a normal curve to the center of the histograms using a robust M-estimation method implemented using the rlm function of the MASS library belonging to the R statistical package. These histograms, displayed in Figure A.15, show that the over dispersion of the Z-scores is substantial in facilities with a larger number of patients. This is consistent with the ndings in Kalb eisch and Wolfe (2013) who considered a time-invariant modeling approach, and motivates the strati cation used in the empirical null method to identify a small percentage of outlying facilities. We also see in Figure A.15 that the central location of the normal curve resulting from the robust M-estimation method lies to the left of zero, the central location of the superimposed standard normal curve, resulting in the empirical method to ag less facilities than when using the nominal p-value. This is consistent with Table A.9 which shows that 13.6% of facilities were agged using the nominal p-value and 4.5% using the empirical null method. In Table A.9, we see that 9.9%, 12.3%, and 18.9% of small, medium, and large facilities were agged using the nominal p-value respectively. This imbalance is due to the over dispersion mentioned above. However, the empirical null method reduced these proportions considerably to 3.0%, 4.1%, and 6.5% respectively, a more manageable proportion of facilities to further investigate.

3.4 Simulation Studies

As described in Section 3.2.2, we use an approximate EM algorithm to estimate the parameters in Model 3.1. Thus, our simulation studies here will focus on the nite sample properties of the proposed estimation algorithm for the proposed multilevel varying coe cient model and our standard dynamic readmission ratio. Similarly we will examine the validity of the proposed hypothesis tests in Section 3.2.3. Finally, we will investigate consequences of ignoring within subject correlation via comparisons to Model 2.

3.4.1 Simulation Model and Design

To study the e cacy of the proposed estimation algorithm and our standard dynamic readmission ratio, we simulate data under Model 1

$$g\ E\ \ Y_{ij}(t)\ j\ Z_{ij}\ ;b_j\ \ \ =\ \ _i(t) + b_j\ + Z_{ij}^T$$

where $Z_{ij} = (Z_{1ij}\ ; Z_{2ij}\ )^T$ and $\ \ = (\ \ _1;\ \ _2)^T = (0:5;\ \ 0:5)^T$. We generate Z_{ij} from a bivariate normal distribution with mean $(0; 0)^T$ and covariance matrix $[125; :0625;:0625; :125]$; b_j are independent and identically distributed normal random variables with mean 0 and variance .70 (similar to our data application); N_{ij} is generated from a discrete random variable with support $f\ 1; 2; 3; 4; 5; 6g$ and corresponding probabilities $f\ :02; :03; :05; :20; :30; :40g$. One third of the facilities are xed to be small, one third to be medium, and one third to be large, and their sizes N_i are generated from one of three discrete uniform random variables putting equal weights over their supports $f\ 20; 21; : : : ; 34g$, $f\ 35; 36; : : : ; 54g$, and $f\ 55; 56; : : : ; 120g$ respectively where the cuto s mimic values in from our data application. We generate $f\ t_{ij\ 1}; : : : ; t_{ijN\ _{ij}}\ g$ by taking a random sample from the set $f\ 0; 1=35; : : : ; 1g$, and de ne one third of the facility e ects $\ _i(t)$ to be $\log(3=7)$, one third to be $\sqrt{\bar{t}}\ \ \ 0:3$, and one third to be $(\ \ t\ \ \ 0:5)^2\ \ \ 0:75$. The response variables Y_{ijk} are generated from a Bernoulli distribution with mean $g^{\ 1}f\ \ _i(t_{ijk}\ ) + b_j\ + Z_{ij}^T\ \ g$. We generated 200 datasets using this set up for both $I\ = 100$ and $I\ = 1000$ to study the e cacy of the proposed estimation algorithm for varying facility size, and then compare the results when the within subject correlations are

ignored. We selected a separate bandwidth for each facility using 10-fold cross-validation as previously described in Section 3.3.2. The candidate bandwidths utilized were chosen in a preliminary simulation study ranging from .32 to .44 (in years). To study the performance of the estimators of $_i(t)$ and $SDR_i(t)$, we utilize a relative mean squared deviation error (MSDE)

$$MSDE_b = \int \hat{b}(t) \quad (t)^2 dt \cdot \int \quad ^2(t)dt;$$

and to study the performance of the time-invariant parameters we utilize mean squared error (MSE). The estimated bias, standard error, MSE, and the quartiles of the MSDE (strati ed by facility size and function shape) are given in Tables A.5 and A.6.

The estimated bias, standard error, and MSE values are relatively small and decrease when the number of facilities increase indicating the e ectiveness in the estimation of the time-invariant parameters in the model. The MSDEs of the time-varying function estimates also decrease when facility size increases (small to large), which is expected since larger facilities will have more data available in the estimation of $_i(t)$. While increasing facility number reduces the estimated MSE of the time-invariant model parameters, it is not expected to change the MSDEs of $_i(t)$ and $SDR_i(t)$ since the estimation of these quantities depend largely on data within facilities (hence facility size). Ignoring the within subject correlation in the estimation severely biases $\hat{b}$, where this bias increases with increasing the within subject variation in our simulation. Furthermore, the MSDEs of $\hat{b}_i(t)$ and $\hat{SDR}_i(t)$ are similar but slightly reduced when the within subject correlation is ignored. This is expected since ignoring the random e ects targets a marginal model drawing inference on population averages leading to attenuation. This is consistent with what is observed in the estimated mean square errors of the facility e ect estimates in a reduced setting utilizing a mixed e ects logistic regression model (under the same simulation set up) where the facility e ects are time-invariant constants. Finally, we note that we do not expect di erences among the error in estimations of the three di erent time-varying functions (at, square root, and quadratic) used in this simulation. Thus, the trend in the MSDEs for $\hat{b}_i(t)$ and $\hat{SDR}_i(t)$ with respect to facility shape follows from the numerical order of the integrals of the squared true underlying

functions.

3.4.2 Hypothesis Test: Validity

We study the level of the proposed hypothesis testing procedure by focusing on one outlier facility and performing the test for ve di erent null hypothesis indexed by . We consider H_0 : $_1(t) = \ _0(t)$ for $= 0; :25; :50; :75; 1$ with three di erent facility sizes (small, medium and large) using the median number of subjects (with respect to facility size) in our data application: small (27 subjects), medium (43 subjects), and large (69 subjects) where $_0(t) = (1 \quad) \ (t) + \ ^0(t)$, $(t) = \sqrt{\bar{x}} \quad 0:3$, and $^0(t) = (\ x \quad 0:5)^2 \quad 0:75$. We note that when $= 0$, $_0(t) = \ (t)$, and as increases to 1, $_0(t)$ becomes closer to $^0(t)$. The data is generated similar to the simulation set up in Section 3.4.1 for $I = 100; 1000$ where we x the rst facility to be either small, medium, or large and de ne $_1(t) = \ _0(t)$ and $_i(t) = \ (t) + \ _i$ with $_i \quad$ i.i.d $N(0; :2^2)$, $i = 2; : : : ; I$. We then calculate the observed test statistic r_1, de ned in Section 3.2.3, replacing $b_M(t_{1jk})$ with $_0(t_{1jk})$. To obtain a p-value, we perform steps (ii) - (v) from Section 3.2.3 replacing $b_M(t_{ijk})$ with $_0(t_{ijk})$ for $i = 1; : : : ; I$. This process is repeated 500 times yielding an estimate of the level of the test by looking at the proportion of the 500 runs resulting in a rejection of the null hypothesis.

The results of the hypothesis testing simulation are summarized in Table A.7 where it is seen that the acceptance probability is consistently slightly above the nominal value .95 under Model 1 in both cases when $I = 100$ and when $I = 1000$. When the within subject correlation is ignored, the acceptance probability is consistently below the nominal value of .95, and is worse with larger facilities.

3.5 Discussion

In this work, we developed a method for time-dynamic pro ling of dialysis facilities in the U.S. Our proposed multilevel varying coe cient model with xed facility time-varying e ects and subject speci c random e ects accommodates the multilevel data structure, and our in-

clusion of a rich set of patient baseline covariates adjusts for the patient heath characteristics prior to dialysis. The use of xed facility e ects (rather than random facility e ects) was to provide more precise estimation of the true time-varying e ects for those facilities with extreme outcomes. We developed an approximate EM algorithm to estimate the parameters in our MVCM, and proposed a time-dynamic facility performance index suitable for pro ling goals. Adjustments for multiple testing to control the overall type I error rate and the overall agging rate were made through the use of the empirical null method. In our data application, the nominal p-value agged roughly 15% of facilities compared to roughly 5% when using the empirical null method. One purpose of instituting a SDR measure for dialysis facilities is to identify potential facility speci c problems over regions of time that contribute to rising costs in health care of dialysis patients.

APPENDIX A

Table A.1: Baseline characteristics of $n = 294;511$ patients aged 59 to 96 used in Section 1.4.1. Data presented are mean standard deviation (SD) for continuous variables or count (percent) for categorical variables.

Variable	Mean	SD/ Count (%)
Baseline age	73.90	7.845
Male		151,847 (52)
Race		
Black		70,020 (24)
White		207,515 (70)
Other		16,976 (6)
Congestive heart failure		120,201 (41)
Coronary heart disease		102,009 (35)
Peripheral vascular disease		55,165 (19)
Diabetes		174,223 (59)
Estimated glomerular ltration rate	10.820	5.436
Body mass index	27.331	7.064

Table A.2: Relative mean squared deviation error of the estimated varying coe cient functions from simulation studies in Section 1.5. Median and 25th and 75th percentiles of the deviation measures are presented based on 200 Monte Carlo runs.

	n = 3000			n = 5000		
	Median	25th Percent	75th Percent	Median	25th Percent	75th Percent
$MSDE_0$	.0037	.0016	.0068	.0017	.0006	.0030
$MSDE_1$	.0078	.0038	.0157	.0039	.0022	.0076
$MSDE_0$	.0127	.0080	.0190	.0109	.0082	.0152
$MSDE_1$	.0242	.0118	.0545	.0160	.0093	.0338
$MSDE_1$	.0348	.0165	.0735	.0251	.0130	.0481
$MSDE_2$	.0085	.0051	.0140	.0083	.0055	.0111
$MSDE_3$	.0090	.0060	.0134	.0076	.0057	.0105

Table A.3: Baseline characteristics of $n = 243,730$ patients aged 65 to 90 used in Section 2.3.1. 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 ltration rate	10.923 5.445
Body mass index	26.973 6.783

Table A.4: Summaries of the relative mean squared deviation error of the estimated varying coe cient functions and estimated mean squared error of the regression coe cients in Section 2.4. Median and 25th and 75th percentiles of the deviation measures are presented based on 200 Monte Carlo runs.

	n = 500			n = 2000		
	Median	25%	75%	Median	25%	75%
$MSDE_0$	.0123	.0057	.0282	.0033	.0014	.0059
$MSDE_1$	.1280	.0598	.2438	.0309	.0161	.0588
MSE_1	.0082	.0017	.0182	.0013	.0003	.0037
MSE_2	.0011	.0003	.0030	.0002	< : 0001	.0006
MSE_3	.0022	.0005	.0062	.0007	.0001	.0015

Table A.5: Estimated bias, standard error (SE), and mean squared error (MSE) of the estimated parameters from simulation studies of Section 3.4.1 based on 200 Monte Carlo runs.

	I = 100						I = 1000					
	Model 1			Model 2			Model 1			Model 2		
Estimate	Bias	SE	MSE	Bias	SE	MSE	Bias	SE	MSE	Bias	SE	MSE
b_1	.009	.063	.004	-.057	.055	.006	.007	.021	< : 001	-.052	.018	.003
b_2	-.008	.064	.004	.059	.057	.007	-.006	.020	< : 001	.053	.018	.003
b_0^2	-.009	.042	.002				-.008	.013	< : 001			

Table A.6: Relative mean squared deviation error of the estimated time-varying facility e ect functions and standard dynamic readmission ratios strati ed by facility size and facility e ect shape from simulation studies of Section 3.4.1 based on 200 Monte Carlo runs.

	I = 100						I = 1000					
b(t)	Model 1			Model 2			Model 1			Model 2		
MSDE	25%	50%	75%	25%	50%	75%	25%	50%	75%	25%	50%	75%
All	.051	.106	.214	.049	.097	.184	.050	.105	.214	.049	.097	.185
Small	.093	.185	.356	.081	.160	.289	.092	.185	.353	.082	.159	.290
Medium	.056	.109	.202	.053	.100	.174	.056	.111	.206	.053	.101	.179
Large	.032	.061	.112	.033	.060	.106	.031	.060	.110	.033	.060	.105
Flat	.050	.100	.196	.047	.093	.172	.048	.098	.195	.046	.091	.170
Sqrt	.037	.075	.148	.038	.074	.140	.037	.075	.146	.038	.075	.137
Quadratic	.081	.158	.315	.068	.136	.250	.079	.160	.313	.070	.136	.256

$DR(t)	Model 1			Model 2			Model 1			Model 2		
MSDE	25%	50%	75%	25%	50%	75%	25%	50%	75%	25%	50%	75%
All	.011	.023	.045	.009	.020	.038	.011	.022	.045	.009	.019	.038
Small	.018	.038	.071	.016	.033	.060	.018	.037	.070	.016	.032	.060
Medium	.011	.023	.043	.010	.020	.037	.012	.023	.044	.010	.020	.037
Large	.007	.014	.026	.006	.013	.023	.007	.014	.025	.006	.012	.022
Flat	.015	.029	.053	.013	.025	.045	.015	.029	.053	.013	.025	.046
Sqrt	.016	.031	.057	.014	.027	.048	.016	.031	.057	.014	.026	.049
Quadratic	.007	.013	.024	.006	.011	.021	.006	.012	.023	.005	.011	.020

Table A.7: Performance of $\hat{SDR}_i(t)$ in testing $H_0 : {}_1(t) = {}_0(t)$ from Section 3.4.2 where increasing values of indicate deviating further from the null. We provide the median MSDE of $\hat{SDR}_i(t)$ and the estimated acceptance probability (AP) strati ed by , facility size, and model.

I = 100

	Small Facility				Medium Facility				Large Facility			
	Model 1		Model 2		Model 1		Model 2		Model 1		Model 2	
	MSDE	AP	MSDE	AP	MSDE	AP	MSDE	AP	MSDE	AP	MSDE	AP
0	.054	.972	.052	.856	.031	.966	.039	.812	.022	.958	.034	.766
.25	.043	.970	.035	.868	.024	.972	.020	.880	.017	.970	.016	.816
.50	.031	.974	.029	.910	.018	.974	.017	.894	.013	.978	.012	.882
.75	.025	.964	.022	.904	.015	.976	.015	.902	.010	.960	.010	.884
1	.018	.972	.017	.876	.012	.972	.011	.878	.008	.966	.007	.864

I = 1000

	Small Facility				Medium Facility				Large Facility			
	Model 1		Model 2		Model 1		Model 2		Model 1		Model 2	
	MSDE	AP	MSDE	AP	MSDE	AP	MSDE	AP	MSDE	AP	MSDE	AP
0	.051	.974	.055	.894	.034	.966	.039	.812	.023	.958	.023	.766
.25	.043	.960	.035	.862	.024	.972	.025	.880	.017	.970	.017	.816
.50	.031	.986	.029	.906	.018	.974	.017	.894	.013	.978	.013	.882
.75	.025	.972	.022	.896	.015	.976	.015	.902	.010	.960	.010	.884
1	.018	.978	.017	.928	.011	.972	.010	.878	.008	.966	.008	.864

Table A.8: Number and percent of significant outlier facilities in our data application in Section 3.3.2 determined by nominal p-values $< .025$.

Model 2	Model 1				
	Non-sig	Sig-better	Sig-worse	Other	Row-sum
Non-sig	3112 (79.8%)	26 (0.7%)	36 (0.9%)	82 (2.1%)	3256 (83.5%)
Sig-better	9 (0.2%)	46 (1.2%)	0 (0%)	7 (0.2 %)	62 (1.6%)
Sig-worse	49 (1.3%)	0 (0%)	153 (3.9%)	7 (0.2%)	209 (5.4%)
Other	123 (3.2%)	4 (0.1%)	18 (0.5%)	229 (5.9%)	374 (9.6%)
Column-sum	3293 (84.4%)	76 (1.9%)	207 (5.3%)	325 (8.3%)	3901 (100%)

Table A.9: Number and percent (with respect to facility size) of total facilities flagged as significantly worse or other than the national standard across time in our data application in Section 3.3.2.

	Model 1		Model 2	
Number of Subjects	Nominal p-value	Empirical null	Nominal p-value	Empirical null
$[20, 34]$	128 (9.9%)	39 (3.0%)	130 (10.1%)	44 (3.4%)
$[35, 54]$	166 (12.3%)	56 (4.1%)	185 (13.7%)	63 (4.7%)
$[55, 192]$	238 (18.9%)	82 (6.5%)	268 (21.3%)	119 (9.5%)
Overall	532 (13.6%)	177 (4.5%)	583 (14.9%)	226 (5.8%)

Figure A.1: Example of follow-up data for a subject (a) without and (b) with an infection-related hospitalization along with the proposed models for cardiovascular risk before (light gray) and after (dark gray) the infection-related hospitalization. Note that the model for cardiovascular risk after the initial infection-related hospitalization appropriately accounts for vintage until the infection-related hospitalization (term $_1(a_i)Z_i$). See Section 1.2 for details.

Figure A.2: Baseline age-strati ed varying coe cient function estimates after the (a) start of dialysis (vintage t_0) and (b) initial infection-related hospitalization (t_1). Final varying coe - cient function estimates as a function of (c) vintage and time since the initial infection-related hospitalization and (d) baseline age at dialysis. 90% bootstrap con dence intervals are given as dashed lines in (c) and (d).

Figure A.3: Estimated probabilities of cardiovascular (CV) events for white male patients with diabetes and with average levels of eGFR and BMI (with the vintage until the rst infection-related hospitalization of $Z = 1:4$ years) with baseline ages (a) 65, (b) 78 and (c) 90. Plot (d) overlays/combines the estimated probability trajectories from the three baseline ages. 90% bootstrap con dence intervals are given as dashed lines in (a), (b), and (c).

Figure A.4: Estimated probabilities of cardiovascular (CV) events during the course of dialysis for patients experiencing the pivotal initial infection-related hospitalization at 3, 2 and 1 year after the start of dialysis with baseline ages of 65, 78 and 90 (columns left, middle and right, respectively). 90% bootstrap con dence intervals are given dashed.

Figure A.5: Age-varying e ects of baseline covariates. Given are estimated baseline age{ varying coe cient functions for (a) vintage up to rst infection, (b) sex - male, (c) race - black, (d) race - other, (e) congestive heart failure, (f) coronary heart disease, (g) peripheral vascular disease, (h) diabetes, (i) eGFR, and (j) BMI. 90% bootstrap con dence intervals are given as dashed lines. The reference horizontal line at zero indicates no e ects.

Figure A.6: Simulation results for $n = 3000$. The cross-sectional median curves of the proposed estimates are given along with 5% and 95% cross-sectional percentiles (dotted) overlaying the true varying coe cient functions (solid).

Figure A.7: Illustration of partly conditional, fully conditional and unconditional model estimates of the varying coe cient function targets in a simple generalized varying coe cient model.

Figure A.8: (a) Fits from a simple partly conditional GVCM, $g[E\{Y_i(t_i) | S_i > t_i\}] = \ _P(t_i)$ along with 90% bootstrap CIs, 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, $g[E\{Y_i(t_i) | S_i \ 2 \ D_j\}] = \ _{j;F}(t_i)$, 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 time-invariant median varying coe cient function estimates over 200 Monte Carlo runs.

Figure A.9: (a) Estimated varying coe cient functions from partly conditional PL-GVCM ts $b_{0;P}(t_0)$ (black), $b_{1;P}(t_1)$ (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 con dence intervals given as dashed lines.

Figure A.10: (a)-(h) Estimated CV risk based on the fully conditional PL-GVCM ts 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 con dence intervals given as dashed lines.

Figure A.11: Estimated coe cients 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 con dence intervals are displayed as whiskers. The gray horizontal line at zero (no e ect) is included for reference. The x-axis denotes the midpoints of the death bins D_j .