Title
Modeling time-varying effects of multilevel risk factors of hospitalizations in patients on dialysis

Journal
Statistics in Medicine, 37(30)

ISSN
0277-6715

Authors
Li, Yihao
Nguyen, Danh V
Chen, Yanjun
et al.

Publication Date
2018-12-30

DOI
10.1002/sim.7950

Peer reviewed

HHS Public Access

Author manuscript

Stat Med. Author manuscript; available in PMC 2018 December 30.

Published in final edited form as:

Stat Med. 2018 December 30; 37(30): 4707–4720. doi:10.1002/sim.7950.

Modeling Time-Varying Effects of Multilevel Risk Factors of Hospitalizations in Patients on Dialysis

Yihao Li[a], **Danh V. Nguyen**[b], **Yanjun Chen**[c], **Connie M. Rhee**[b], **Kamyar Kalantar-Zadeh**[b], and **Damla Şentürk**[a,*]

[a]Department of Biostatistics, University of California, Los Angeles, California 90095, U.S.A.

[b]Department of Medicine, UC Irvine School of Medicine, Orange, CA 92868-3298, U.S.A.

[c]Institute for Clinical and Translational Science, University of California, Irvine, California 92687-1385, U.S.A.

Summary

For chronic dialysis patients, a unique population requiring continuous medical care, methodologies to monitor patient outcomes, such as hospitalizations, over time, after initiation of dialysis, are of particular interest. Contributing to patient hospitalizations are a number of multilevel covariates such as demographics and comorbidities at the patient-level, and staffing composition at the dialysis facility-level. We propose a varying coefficient model for multilevel risk factors (VCM-MR) to study the time-varying effects of covariates on patient hospitalization risk as a function of time on dialysis. The proposed VCM-MR also includes subject-specific random effects to account for within-subject correlation and dialysis facility-specific fixed effect varying coefficient functions to allow for modeling of flexible time-varying facility-specific risk trajectories. An approximate EM algorithm and an iterative Newton-Raphson approach are proposed to address the challenge of estimation of high-dimensional parameters (varying coefficient functions) for thousands of dialysis facilities in the United States. The proposed modeling allows for comparisons between time-varying effects of multilevel risk factors as well as testing of facility-specific fixed effects. The method is applied to model hospitalization risk using the rich hierarchical data available on dialysis patients initiating dialysis between January 1, 2006 and December 31, 2008 from United States Renal Data System, a large national database, where 331, 443 hospitalizations over time are nested within patients, and 89,889 patients are nested within 2,201 dialysis facilities. Patients are followed-up until December 31, 2013, where follow-up time is truncated five years after initiation of dialysis. Finite sample properties are studied through extensive simulations.

Keywords

end-stage renal disease; hospitalization risk; multilevel varying coefficient models; United States Renal Data System

*Correspondence to: Damla Şentürk, Department of Biostatistics, University of California, Los Angeles, California 90095.

Supporting Information

Additional Supporting Information may be found online, containing details of the estimation procedure, cohort description, the utilized bandwidth selection procedures for the varying coefficient functions and simulation design.

1 Introduction

The latest United States Renal Data System (USRDS) annual report[1] shows that there were over 678,000 individuals with end-stage-renal disease (ESRD) in the United States (US) as of December 31, 2014. About 70% of patients with ESRD were on dialysis, a life-sustaining treatment. Patients on dialysis are typically hospitalized twice a year, and hospitalization in this population remains a major mortality and morbidity burden. Modeling time-varying effects on patient outcomes, such as hospitalizations, is especially important in this unique cohort of patients because ESRD patients remain on dialysis for long periods of time (for the remainder of their lifetime or until receiving a kidney transplant). Hence studying leading risk factors of hospitalizations and characterizing their effects as a function of time on dialysis is essential in assessing health care improvement strategies, as the needs of dialysis patients may change the longer they remain on dialysis.

In addition to the need for time-dynamic modeling, another challenging aspect of studying hospitalization risk factors is the hierarchical structure of the data and the multilevel nature of the risk factors themselves. The rich hospitalization data available through USRDS is hierarchical where hospitalizations over time are nested within patients, and patients are nested within dialysis facilities across the US. There are a number of risk factors for hospitalizations at both the patient-level and facility-level in the hierarchy. The patients' baseline demographics and comorbidities at initiation of dialysis as well as facility staff level and composition are among the covariates that potentially affect hospitalization risk. Hence the desired modeling has to account for dependencies in the data within subjects and within facilities to facilitate efficient inference for the time-varying effects of multilevel risk factors.

Varying coefficient models are an effective tool in modeling time-varying regression effects[2–3] and there is a rich literature on their applications to longitudinal data.[4–8] However most of the literature is on a two-level hierarchy where observations over time are nested in subjects and is not applicable to data with higher levels of hierarchy such as observations over time nested in patients, and patients nested in dialysis facilities.[9–14] The few works that consider higher levels of hierarchy do not consider multilevel predictors or multilevel regression effects modeled through multilevel varying coefficient functions. You et al.[15] consider varying coefficient models with a three-level hierarchy, but only model time-varying effects of subject-level predictors and do not assume a particular structure for the dependencies within the hierarchy, utilizing an unstructured error covariance, which may not scale up well in large data applications.

As a novel departure from existing literature, we propose a varying coefficient model for multilevel risk factors (VCM-MR) with multilevel varying coefficient functions that are associated with them. We model the within-subject correlation via a subject-specific random effect. However, rather than a facility-level random effect, we model facility-specific deviations in hospitalization risk via facility-specific fixed varying coefficient functions. This is partly motivated by the fact that the correlation between predictors and a random effect higher in the hierarchy may lead to bias in estimation of regression effects in multilevel regression models.[16] While the inclusion of the facility-specific fixed varying

coefficient functions add flexibility in modeling time-varying facility-specific risk trajectories, they also pose a major computational challenge in estimation yielding a large number of varying coefficient functions, hence *high dimensionality in the parameter space.* Hence we propose a novel iterative estimation algorithm tailored to the specific computational challenges posed by our data application namely the high-dimensional parameter space and the large size of the data from USRDS, based on an approximate EM algorithm and Newton-Raphson maximization. We rely on bootstrap confidence intervals, constructed by resampling from facilities, for assessing variation in the estimated subject- and facility-level time-varying effects and develop a hypothesis testing procedure to assess whether facility-specific fixed effects are significant or time-varying. Finally, taking advantage of the multilevel structure of VCM-MR, we propose predicted multilevel (patient- and facility-level) hospitalization risk trajectories throughout dialysis treatment. Patient-level predictions can help guide patients in choosing a suitable dialysis facility at the initiation of dialysis and can further provide subject-specific predicted risk trajectories utilizing the patients' case-mix at baseline and facility-level covariates from facilities of interest. The goal of facility-level predictions is to provide feedback to a facility for improvement of patient care.

The proposed VCM-MR as well as the iterative estimation procedure based on the approximate EM algorithm and inference for the model components are developed in Section 2, with technical details included in the Supporting Information. The proposed estimation framework is an extension of the iterative estimation algorithm of Estes et al.,[17] recently proposed for a multilevel varying coefficient model used for time-dynamic facility profiling; see also He et al.[18] for time-static profiling. The goal of the previous work[17] was to develop a time-dynamic risk-standardized index to compare a facility's performance to a national norm. Hence the multilevel varying coefficient model of Estes et al.[17] does not model time-varying effects of subject- or facility-level covariates and only adjusts for non-time-varying effects of subject case-mix, suitable for the goal of facility profiling. Our modeling and overarching goals are quite different in this paper where the main focus is on studying the effects of multilevel (subject- and facility-level) covariates on dialysis patient outcomes as a function of the time indexing the change in patient needs (i.e. time on dialysis), rather than modeling facility quality of care. Applications to USRDS data to model hospitalization risk of patients over time on dialysis and simulation studies are presented in Sections 3 and 4, respectively.

2 Varying Coefficient Model for Multilevel Risk Factors

2.1 Model Specification

Let $i = 1, \ldots, I$ index dialysis facilities and $j = 1, \ldots, N_i$ index subjects belonging to the ith facility with N_i number of total subjects. Further let $Z_{i(j)} = \{Z_{1i(j)}, \ldots, Z_{pi(j)}\}^\top$ denote the vector of p facility-level predictors of facility i and $X_{ij} = (X_{1ij}, \ldots, X_{rij})^\top$ denote the vector of r subject-level predictors. Note that the facility-level predictors, such as total number of subjects or staff decomposition, are reported only once a year. Hence they are also indexed by the subject counter j, since their values are assigned using reported facility characteristics in the previous calendar year from the time the jth subject within facility i initiates dialysis.

To study the effect of multilevel risk factors, $Z_{i(j)}$ and X_{ij}, on a patient's hospitalization risk, we model the binary outcome of having a hospitalization within a three month interval in the follow-up time after initiation of dialysis. The outcome $Y_{ijk} \equiv Y_{ij}(t_{ijk})$ equals one if the jth patient within facility i experienced a hospitalization during the kth three month follow-up interval and equals zero otherwise, where $k = 1, \ldots, N_{ij}$, with N_{ij} denoting the total number of three month intervals in the follow-up of subject j and t_{ijk} denoting the midpoint of the kth three month interval. A similar approach of modeling grouped response in intervals over follow-up time was considered by Liu et al.[19] in studying hospitalization days among dialysis patients using an event rate model. The goal is to model the expected outcome:

$$p_{ij}(t) = E\left\{Y_{ij}(t) \mid Z_{i(j)}, X_{ij}, b_{ij}, S_{ij} > t\right\},$$

where $b_{ij} \sim N\left(0, \sigma_b^2\right)$ denotes the subject-specific random effects to account for within-subject correlation and S_{ij} denotes the death time of subject j. The expected outcome $p_{ij}(t)$ defines a 'partly conditional' target conditional on the patients being alive $S_{ij} > t$. Note that since our target is only conditional on the patient being alive, we assume that the probability $p_{ij}(t)$ conditioning on $S_{ij} > t$ is the same as conditioning on $S_{ij} > t^*$ for any $t^* > t$, i.e., conditioning on different time points is exchangeable as long as the patient is alive at both time points (similar to most frailty models). For the outcome defined in three month intervals, the expected outcome is defined for follow-up intervals such that $S_{ij} > t_{ijk}$, i.e. for intervals where the subject survived at least half of the three month interval. The 45 day cutoff is used as a compromise between having too few days in the interval for hospitalization opportunities if the cutoff is lower and eliminating more last intervals in the follow-up of patients where the death occurred (leading to data loss) if the cutoff is higher. A sensitivity analysis with cut-offs of 30 and 60 days in our applications to USRDS data lead to similar inference as the chosen cutoff of 45 days. Partly conditional models study the dynamic cohort of survivors and have been considered in the context of generalized linear models for longitudinal data where missingness is primarily due to truncation by death.[20] Estes et al.[8,21] considered partly conditional target of inference for varying coefficient models.

In the proposed VCM-MR, the facility-level effects have two parts. The first part explains the time-varying effects of the facility-level covariates, denoted by $\theta(t)$ and the second part represents the facility-specific fixed effects, denoted by $\gamma_i(t)$. The logit link function, denoted by $g\{p_{ij}(t)\} = \log[p_{ij}(t)/\{1 - p_{ij}(t)\}]$, is used to connect the conditional expected outcome to the time-varying effects of the predictors via

$$g[E\{Y_{ij}(t,c) \mid X_{ij}, Z_{i(j)}, b_{ij}, S_{ij} > t\}] = g\{p_{ij}(t,c)\} = \eta(c) + \gamma_i(t) + Z_{i(j)}^{\top}\theta(t) + X_{ij}^{\top}\beta(t) + b_{ij}, \quad (1)$$

where $\beta(t) = \{\beta_1(t), \ldots, \beta_I(t)\}^{\top}$ and $\theta(t) = \{\theta_1(t), \ldots, \theta_p(t)\}^{\top}$ denote the time-varying effects of the subject- and facility-level covariates, respectively. Note that hospitalization risk is

assumed not to depend on the subject's actual survival time, but rather only on the length of time in the follow-up after initiation of dialysis. Even though this assumption may be a strong one depending on the the target of inference that is of interest, it is standard in 'partly conditional' models.[20] In addition, $\eta(c)$ denotes calendar time effects with c denoting calendar time at initiation of dialysis, for cases where the cohort includes patients initiating dialysis over multiple calendar years. Hence, while $\eta(c)$ adjusts for potential differences in the overall hospitalization risk of cohorts of patients initiating dialysis over different calendar years, $\gamma_i(t)$ captures facility effects over time on dialysis, a time period over which the needs of dialysis patients may change. Note that $\gamma_i(t)$ may still include some changes in facility performance over follow-up calendar time as well but such an effect is at best an average effect over the dialysis initiation years since it is estimated using cohorts initiating dialysis over multiple calendar years. Also, since both $\eta(c)$ and $\gamma_i(t)$ are playing the role of a y-intercept, they are not identifiable without restrictions. Therefore, we normalize $\eta(c)$ such that $\int_0^C \eta(c)dc = 0$ and allow $\gamma_i(t)$ to carry the magnitude of the y-intercept.

Finally, note that the proposed model in (1) is a partially pooled model for facility effects. In other words, it strikes a balance between complete pooling and no pooling, where complete pooling would pool data without keeping track of which facility they belong to and would not consider facility-specific effects $\gamma_i(t) + Z_{i(j)}^{\top}\theta(t)$ and in no pooling, estimation of the facility effects would only use facility-specific data, i.e., facility-level effects would be modeled only by $\gamma_i(t)$, which may cause over-fitting problems, especially in small facilities. Partial pooling stabilizes the estimation of facility-level effects through inclusion of facility-level covariates while still producing facility-specific predictions.

2.2 Estimation Procedure

We outline the proposed estimation procedure, based on an approximate EM algorithm, for the proposed VCM-MR. Let $L_{Yij}\{\gamma_i(t), \theta(t), \beta(t), \eta(c)\}$ denote the joint distribution of the outcome of the jth subject $\left(Y_{ij1}, \ldots, Y_{ijN_{ij}}\right)$ observed at the time points $t_{ij} = \left(t_{ij1}, \ldots, t_{ijN_{ij}}\right)$, conditional on b_{ij}, X_{ij}, $Z_{i(j)}$ and $S_{ij} > t_{ij}$. For mathematical convenience, we assume that the within-subject correlation among $\left(Y_{ij1}, \ldots, Y_{ijN_{ij}}\right)$ is explained by two independent sources: the subject-specific random effects b_{ij} and the dependency of Y_{ijk}, $k = 1, \ldots, N_{ij}$, on the patient's death time S_{ij}. A similar assumption is made in Liu et al.[19] who also consider a partly conditional model (referred to as a partial marginal model), where the death time is assumed to be independent of the frailty (represented by a within-subject random effect) capturing the within-subject correlation. Using the independence between b_{ij} and S_{ij}, the joint distribution of $\left(Y_{ij1}, \ldots, Y_{ijN_{ij}}, b_{ij}\right)$ conditional on X_{ij}, $Z_{i(j)}$ and $S_{ij} > t_{ij}$, denoted by $L_{ij}\{b_{ij}, \sigma_b, \gamma_i(t), \theta(t), \beta(t), \eta(c)\}$, can be given as

$$L_{ij}\{b_{ij}, \sigma_b, \gamma_i(t), \theta(t), \beta(t), \eta(c)\} = L_{Y_{ij}}\{\gamma_i(t), \theta(t), \beta(t), \eta(c)\} \times \frac{\exp\{-b_{ij}^2/(2\sigma_b^2)\}}{\sqrt{2\pi\sigma_b^2}},$$

for a normally distributed subject-specific random effect b_{ij}. Hence, the complete likelihood corresponding to the VCM-MR in (1) is

$$L\{b, \sigma_b, \gamma_1(t), \ldots, \gamma_I(t), \theta(t), \beta(t), \eta(c)\} = \prod_{i=1}^{I} \prod_{j=1}^{N_i} L_{ij}\{b_{ij}, \sigma_b, \gamma_i(t), \theta(t), \beta(t), \eta(c)\}.$$

In addition, viewing the subject-specific random effects as unobserved covariates, the incomplete likelihood available for estimation of $\{\sigma_b, \gamma_1(t), \ldots, \gamma_I(t), \theta(t), \beta(t), \eta(c)\}$ is

$$L\{\sigma_b, \gamma_1(t), \ldots, \gamma_I(t), \theta(t), \beta(t), \eta(c)\} = \prod_{i=1}^{I} \prod_{j=1}^{N_i} \left[\int_{-\infty}^{\infty} L_{ij}\{b_{ij}, \sigma_b, \gamma_i(t), \theta(t), \beta(t), \eta(c)\} db_{ij}\right].$$

We propose an approximate EM algorithm, where the expectation step targets the approximate conditional expectation of the complete likelihood by utilizing a Taylor's expansion and the first two moments of the subject-specific random effects. Then the maximization step optimizes the approximate expected likelihood with respect to model parameters. For the expectation step, the posterior distribution of b_{ij} can be given as

$$D_{ij}\{b_{ij} | Y_{ij}, \sigma_b, \gamma_i(t), \theta(t), \beta(t), \eta(c), S_{ij} > t_{ij}\} = \frac{L_{ij}\left\{b_{ij}, \sigma_b, \gamma_i(t), \theta(t), \beta(t), \eta(c)\right\}}{\int_{-\infty}^{\infty} L_{ij}\left\{b_{ij}, \sigma_b, \gamma_i(t), \theta(t), \beta(t), \eta(c)\right\} db_{ij}}.$$

Using this posterior distribution, we define the posterior mean and variance of b_{ij}, denoted by b_{ij0} and v_{ij0}, respectively, as

$$b_{ij0} = \int_{-\infty}^{\infty} b_{ij} D_{ij}\{b_{ij} | Y_{ij}, \sigma_b, \gamma_i(t), \theta(t), \beta(t), \eta(c), S_{ij} > t_{ij}\} db_{ij} \text{ and} \quad (2)$$

$$v_{ij0} = \int_{-\infty}^{\infty} (b_{ij} - b_{ij0})^2 D_{ij}\{b_{ij} | Y_{ij}, \sigma_b, \gamma_i(t), \theta(t), \beta(t), \eta(c), S_{ij} > t_{ij}\} db_{ij}. \quad (3)$$

The integrals in (2) and (3) are approximated numerically via a Gauss-Hermite quadrature calculation with 20 sample points. Assuming that the within-subject correlation introduced by the dependence on death time S_{ij} is weak compared to that introduced by the random effect b_{ij}, we can approximate the joint likelihood $L_{Y_{ij}}\{\gamma_i(t), \theta(t), \beta(t), \eta(c)\}$ conditional on b_{ij} (needed in (2) and (3)) utilizing the working independence assumption:

$$L_{Y_{ij}}\{\gamma_i(t), \theta(t), \beta(t), \eta(c)\} \approx \prod_{k=1}^{N_{ij}} \frac{\exp[\{\gamma_i(t_{ijk}) + b_{ij} + Z_{i(j)}^{\top}\theta(t_{ijk}) + X_{ij}^{\top}\beta(t_{ijk}) + \eta(c_{ij})\}Y_{ijk}]}{1 + \exp\{\gamma_i(t_{ijk}) + b_{ij} + Z_{i(j)}^{\top}\theta(t_{ijk}) + X_{ij}^{\top}\beta(t_{ijk}) + \eta(c_{ij})\}}.$$

The working independence assumption will also be used in the M-step of the proposed EM algorithm to estimate the partly conditional target. Kurland and Heagerty[20] point out that a standard likelihood-based method or an estimating equation approach without a working independence structure will not lead to valid inference for a partly conditional target.

For approximating the conditional expectation of the complete likelihood in the E-step, let $\{\sigma_b^*, \gamma_1^*(t), \ldots, \gamma_I^*(t), \theta^*(t), \beta^*(t), \eta^*(c)\}$ be the current parameter estimates, b_{ij0}^*, v_{ij0}^* denote the estimated posterior mean and variance of b_{ij} based on the current parameter estimates and $\ell_{ij}\{b_{ij}, \sigma_b, \gamma_i(t), \theta(t), \beta(t), \eta(c)\}$ denote the log of $L_{ij}\{b_{ij}, \sigma_b, \gamma_i(t), \theta(t), \beta(t), \eta(c)\}$. Because the closed form for $\sum_{i=1}^{I}\sum_{j=1}^{N_i} E[\ell_{ij}\{b_{ij}, \sigma_b, \gamma_i(t), \theta(t), \beta(t), \eta(c)\} | Y_{ij}, \sigma_b^*, \gamma_i^*(t), \theta^*(t), \beta^*(t), \eta^*(c), S_{ij} > t_{ij}]$ is not available, we use the second order Taylor series expansion to approximate the expected log-likelihood around b_{ij0}^* by

$$\sum_{i=1}^{I}\sum_{j=1}^{N_i}\left(\sum_{k=1}^{N_{ij}}\left[Y_{ijk}\{g(p_{0,ijk}^*)\} + \log(q_{0,ijk}^*) - \frac{v_{ij0}^*}{2}p_{0,ijk}^* q_{0,ijk}^*\right] - \frac{(b_{ij0}^*)^2 + v_{ij0}^*}{2(\sigma_b^*)^2} - \frac{1}{2}\log\{2\pi(\sigma_b^*)^2\}\right) \equiv \sum_{i=1}^{I} L_i\{\sigma_b^*, \gamma_i^*(t), \theta^*(t), \beta^*(t), \eta^*(c)\}, \quad (4)$$

where $p_{0,ijk}^* = g^{-1}\left\{\gamma_i^*\left(t_{ijk}\right) + b_{ij0}^* + Z_{i(j)}^{\top}\theta^*\left(t_{ijk}\right) + X_{ij}^{\top}\beta^*\left(t_{ijk}\right) + \eta^*\left(c_{ij}\right)\right\}$, $q_{0,ijk}^* = 1 - p_{0,ijk}^*$ and $L_i\{\sigma_b^*, \gamma_i^*(t), \theta^*(t), \beta^*(t), \eta^*(c)\}$ is defined implicitly. (For details, see Appendix A of the Supporting Information.)

The main challenge in maximizing the approximate expected log-likelihood in (4) (M-step) is the high dimensionality of the parameter space when the number of facilities is large. Nonetheless, the approximate expected log-likelihood is separable into I components, denoted by $L_i\{\sigma_b^*, \gamma_i^*(t), \theta^*(t), \beta^*(t), \eta^*(c)\}$. Hence, maximizing the approximate expected log-likelihood with respect to $\gamma_i(t)$ is equivalent to maximizing $L_i\{\sigma_b^*, \gamma_i^*(t), \theta^*(t), \beta^*(t), \eta^*(c)\}$, with respect to $\gamma_i(t)$ and therefore $\gamma_i(t)$ can be estimated utilizing data from the ith facility. Hence for a fixed set of $\{\sigma_b, \theta(t), \beta(t), \eta(c)\}$, $\gamma_i(t)$ is updated based on data from the ith facility. To estimate $\{\sigma_b, \theta(t), \beta(t), \eta(c)\}$, for fixed $\gamma_i(t)$, $i = 1, \ldots, I$, we maximize the entire approximate expected log-likelihood. Therefore, we propose an iterative Newton-Raphson algorithm which iterates between estimation of $\gamma_i(t)$ and $\{\sigma_b, \theta(t), \beta(t), \eta(c)\}$. This iterative nature, separating the maximization of $\gamma_i(t)$ and the rest of the model parameters is the key to the computational feasibility of the proposed algorithm, since updating of $\gamma_i(t)$ estimates only uses data within facilities. The computational time for fitting VCM-MR is 2.0 and 9.7 minutes in our simulation set-up introduced in Section 4 for $I = 100$ and $I = 500$ facilities, respectively, and 1.3 hours for the application to USRDS data in Section 3 on a modest DELL XPS 8910 desktop with 6th generation Intel® Core™ i7–6700 processor.

For estimation of the varying coefficient functions $\gamma_i(t)$, $\beta(t)$, $\theta(t)$ and $\eta(c)$, we utilize their local linear expansions in time (follow-up time or calendar year at initiation of dialysis) and target the parameters in these expansions by maximizing the approximate local log-likelihood defined in a local neighborhood in time, by a one-step Newton-Raphson iteration. We begin by estimation of σ_b via maximizing the approximate expected log-likelihood, followed by the estimation of $\gamma_i(t)$ with fixed $\{\sigma_b, \theta(t), \beta(t), \eta(c)\}$. Then $\gamma_i(t)$, $\eta(c)$, σ_b are fixed and $\beta(t)$ and $\theta(t)$ are updated. Finally, we update $\eta(c)$ with fixed $\{\sigma_b, \gamma_i(t)\ \theta(t), \beta(t)\}$. The estimation steps of the proposed algorithm are provided in Appendix A of the Supporting Information.

2.3 Inference for Model Parameters

For inference on the effects of the multilevel risk factors ($\theta(t)$, $\beta(t)$) and the calendar year at initiation of dialysis ($\eta(c)$), we utilize a bootstrap procedure which samples from facilities, the highest level of the hierarchy in the data. For each bootstrap dataset, we sample the same number of total facilities as the original data and repeat the proposed estimation procedure to obtain the bootstrap estimates of the varying coefficient functions of interest. Bootstrap confidence intervals are constructed using the pointwise percentiles of the set of bootstrap estimates.

While our main inferential focus is on the effects of multilevel risk factors for hospitalizations, we also propose a computationally efficient hypothesis testing algorithm for the significance of the facility-specific effects captured by $\gamma_i(t)$. Testing for $H_0 : \gamma_i(t) = 0$ can be interpreted as testing whether the facility performance deviates from what would be explained or expected based on its facility- and subject-level covariates. Since the proposed bootstrap procedure for inference on the multilevel risk factors samples from facilities to preserve the correlation structure in the data, it cannot be used for inference on the facility-specific effects. The proposed hypothesis testing procedure utilizes the fact that estimation of $\beta(t)$, $\theta(t)$, $\eta(c)$ and σ_b^2 is quite precise, based on the entire data which is large, made up of hospitalizations of patients from all dialysis facilities across the US, and fixes these estimates once they are estimated throughout the algorithm. Hence the proposed hypothesis testing procedure only resamples data for subjects within the ith facility and is therefore computationally efficient. The proposed testing procedure measures departures of $\hat{\gamma}_i(t)$ from 0, under the null $H_0 : \gamma_i(t) = 0$ via the test statistic $r_i = \sqrt{\int \{\hat{\gamma}_i(t) - 0\}^2 dt}$. The specific steps are as follows.

a. Estimate all model parameters from the initial data fit and compute the test statistic r_i based on the observed data. Denote this observed test statistic by r_i^O. Fix $\theta(t)$, $\beta(t)$, $\eta(c)$ and σ_b at their estimated values $\hat{\theta}(t)$, $\hat{\beta}(t)$, $\hat{\eta}(c)$ and $\hat{\sigma}_b$.

b. Resample subject-specific random effects from the posterior distribution $D_{ij}\{b_{ij} \mid Y, \hat{\sigma}_b, \hat{\theta}(t), \hat{\beta}(t), \hat{\eta}(c), \gamma_i(t) = 0, S_{ij} > t_{ij}\}$ defined under the null. Compute the posterior mean and variance, b_{ij0} and v_{ij0} using the fixed $\hat{\theta}(t)$,$\hat{\beta}(t)$, $\hat{\eta}(c)$ and $\hat{\sigma}_b$ values from step (a). Approximate the posterior distribution by a normal density

with mean b_{ij0} and variance v_{ij0} and draw an independent sample of size F for each subject $j = 1, \ldots, N_i$ within facility i: $b_{ij}^{(f)} \sim N\left(b_{ij0}, v_{ij0}\right)$, $f = 1, \ldots, F$.

c. Draw F samples of the outcome $\{Y_{ijk}^{(f)} : j = 1, \ldots, N_i, k = 1, \ldots, N_{ij}, f = 1, \ldots, F\}$ where each observation, conditional on the resampled subject-specific random effects, is generated from a Bernoulli distribution under the null ($H_0 : \gamma_i(t) = 0$):

$$Y_{ijk}^{(f)} \mid b_{ij}^{(f)} \sim Ber\left[\frac{\exp\{\hat{\eta}(c_{ij}) + b_{ij}^{(f)} + Z_{i(j)}^{\top}\hat{\theta}(t_{ijk}) + X_{ij}^{\top}\hat{\beta}(t_{ijk})\}}{1 + \exp\{\hat{\eta}(c_{ij}) + b_{ij}^{(f)} + Z_{i(j)}^{\top}\hat{\theta}(t_{ijk}) + X_{ij}^{\top}\hat{\beta}(t_{ijk})\}}\right].$$

a. stimate b_{ij0}, v_{ij0} and $\gamma_i(t)$ and the test statistic, $r_i^{(f)}$, based on each resampled dataset $f = 1, \ldots, F$. Note that since $\theta(t)$, $\beta(t)$, $\eta(c)$ and σ_b are fixed in step (a), we only need to iterate between estimation steps 2, 4 and 7 to obtain the parameter estimates.

b. Calculate the nominal p-value $Pr\left(r_i > r_i^O \mid H_0\right)$ by $(1/F)\sum_{f=1}^{F} \mathbb{I}\left\{r_i^{(f)} > r_i^O\right\}$, where $\mathbb{I}\{\cdot\}$ denotes the indicator function.

Note that the hypothesis testing procedure can be extended for also testing whether the facility-specific fixed effect is time-varying, i.e., $H_0 : \gamma_i(t) = c$, by substituting c for 0 above, using the test statistic $r_i = \sqrt{\int \{\hat{\gamma}_i(t) - c\}^2 dt}$.

3 Multilevel Risk Factors of Hospitalization Among Patients on Dialysis

3.1 Description of the USRDS Study Cohort

We utilize hospitalization data from the United States Renal Data System (USRDS), which collects information on nearly all patients with end-stage renal disease (ESRD), including patient demographics and comorbidities prior to the initiation of dialysis. The study cohort includes dialysis patients 18 years of age or older who initiated dialysis between January 1, 2006 and December 31, 2008. The follow-up is until December 31, 2013, where the follow-up time is truncated five years after initiation of dialysis. The detailed descriptions of the study cohort and the exclusion rules are provided in Supporting Information Appendix B. Our final study cohort includes 89, 889 patients receiving dialysis at a total of 2, 201 facilities. The number of patients per facility varies between 20 and 162 where we refer to facilities with 20–31, 31–44 and > 44 patients as small, medium and large facilities, respectively (the cutoff values are taken to be the tertiles of the distribution).

3.2 Time-Varying Effects of Multilevel Risk Factors

To study the effects of multilevel risk factors, 27 patient-level and three facility-level covariates are considered for the proposed VCM-MR. The patient-level covariates include age, gender, body mass index (BMI), whether diabetes is the cause of ESRD and 23 comorbidities, ranging from chronic obstructive pulmonary disease (COPD), seizure disorder, ulcers, drug and alcohol disorders, end-stage liver disease, severe cancer to psychiatric comorbidities and transplants. Each of the 23 comorbidities (indicator variables)

are determined based on the presence of the condition from the previous 12 months prior to the initiation of dialysis treatment for each person based on Medicare claims. Total number of patients, nurse-to-staff ratio and non-nurse-to-patient ratio (i.e., patient care technician (PCT)-to-patient ratio) are included for the facility-level covariates. For convenience of interpretation, we consider nurse-to-patient ratio and non-nurse-to-patient ratio as percentages, by multiplying the relevant ratios by 100. The standard deviation of the subject-specific random effects is estimated to be $\widehat{\sigma_b} = 1.12$, larger than the magnitude of most estimated $\gamma_l(t)$ values, signaling that the variation of the overall hospitalization risk across patients is generally larger than the variation across dialysis facilities. For details on the selection of the bandwidths of the varying coefficient functions, corresponding to the multilevel risk factors, see Appendix C of the Supporting Information.

All of the patient-level risk factors are found to have significant effects on patient hospitalization risk except for the two comorbidities considered, fibrosis of the lung or other chronic lung disorders and respirator dependence. In addition, all of the comorbidities that are found significant are associated with an increase in hospitalization risk, as expected. Figure 1 displays the estimated varying coefficient functions for a sample of eight patient-level risk factors which are found significant, (a) age, (b) BMI, (c) whether diabetes is the cause of ESRD, (d) gender, COPD, (f) ulcers, (g) transplants and (h) seizure disorders and convulsions. The point-wise 95% bootstrap confidence intervals based on 200 bootstrap replications are also provided (dashed lines). For easier comparison, we plot effect sizes corresponding to the changes in age and BMI in 10-year and 5-unit increments, respectively (close to their respective unit standard deviations). Older age at initiation of dialysis is associated with higher hospitalization risk except for the first few years on dialysis. Although time-varying effects of age on the risk of hospitalization has not been examined in this population, this finding may be partly attributed to the cumulative burden of dialysis treatment which typically leads to the deteriorating conditions of end stage renal disease. Diabetes being the cause of ESRD is associated with higher hospitalization risk with the effect getting stronger as patients stay longer on dialysis. Females have more hospitalizations than males, but this difference in hospitalization risk gets smaller in the later years of dialysis treatment. As observed also in other chronic conditions, higher BMI is associated with lower hospitalization risk, with a protective effect. All four comorbidities displayed are associated with higher hospitalization risk, with some time-varying effects throughout the course of dialysis. For example, the association between seizure disorders and convulsions and increased hospitalization risk gets weaker as patients stay longer on dialysis.

At the facility-level, Figure 2(a)-(c) display the time-varying effects of the three risk factors considered: (a) nurse-to-patient ratio (in percent), (b) non-nurse-to-patient ratio (in percent) and (c) total number of patients. The effect sizes plotted correspond to changes of 5-percentage point increments in both nurse-to-patient and non-nurse-to-patient ratios and a change by 50 patients in the total number of patients. Higher number of total patients (hence larger facility size) is significantly correlated with lower hospitalization risk. Both nurse-to-patient ratio and PCT-to-patient ratio have a significant effect on the patients' hospitalization risk, where the higher ratio of nurse-to-patient and PCT-to-patient are both correlated with a

lower risk of hospitalization, as expected. Note that the effect size of PCT-to-patient ratio is found to be larger than that for the nurse-to-patient ratio on hospitalization risk. For example, at the end of the first year of dialysis, a five percentage points increase in the PCT-to-patient ratio is associated with a seven percent decrease in the odds of hospitalizations, whereas the same amount increase in the nurse-to-patient ratio is only associated with a 4 percent decrease in the odds. The effects of all significant risk factors at the facility-level increase with time on dialysis, i.e. for patients who have been on dialysis longer, an increase in facility size or nurse-to-patient ratio is associated with larger reductions in hospitalization risk. Also plotted in Figure 2 is the estimated varying coefficient function for the calendar year effect at initiation of dialysis. The bootstrap confidence interval for $\eta(c)$ contains a constant function around zero except at the boundaries, providing evidence for non-significant calendar year effects in the study cohort.

In addition, as explained in Section 2.3, $\gamma_i(t)$ captures facility-specific fixed effects. We want to caution the reader that in the presence of facility-level covariates in the model, $\gamma_i(t)$ alone does not reflect facility performance and should not be used for facility comparisons; instead it should be interpreted as a residual facility-specific deviation captured beyond what is explained by baseline case-mix and facility-level covariates. Hence, positive $\hat{\gamma}_i(t)$ correspond to higher risks of hospitalization after adjusting for patient case-mix and facility-level risk factors, while negative $\hat{\gamma}_i(t)$ correspond to lower risks of hospitalization. All facilities have been tested for significant effects ($H_0 : \gamma_i(t) = 0$). Overall 15% of facilities have significant effects, where the significant facility effects represent 16.2% of small facilities (0.1% always positive, 15.7% always negative and 0.4% mixed), 14.8% of medium facilities (0.1% always positive, 14.4% always negative and 0.3% mixed) and 13.9% of large facilities (13.8% always negative and 0.1% mixed). For illustration of the different trends, we plot a sample of facilities whose time-varying effect $\gamma_i(t)$ estimates are found significantly different than zero and are always negative, always positive or mixed in the Supporting Information Figure S1.

3.3 Predicted Multilevel Hospitalization Trajectories

Using the estimated model components, the proposed VCM-MR can also be used for obtaining predictions of hospitalization risk trajectories. Similar to the hierarchical nature of the proposed modeling, prediction obtained from VCM-MR is also multilevel, at the patient- and facility-levels. The goal of the patient-level prediction considered is to provide information to patients in selecting facilities at initiation of dialysis and for patients to predict their specific risk trajectories after initiation of dialysis based on their baseline covariates. While for selecting facilities, multiple risk trajectories can be obtained using facility characteristics from multiple facilities and 'average' case-mix values, for the second goal of creating subject-specific predicted hospitalization risk trajectories, subject-specific case-mix would be utilized in obtaining risk predictions. For illustration, we plot the patient-level hospitalization risk predictions for a single patient from the USRDS cohort using three facilities including the patient's current facility (Figure 3(a)). Because calendar year effect $\eta(c)$ is not found significant, it does not contribute to patient-level predictions given by $\hat{p}'_{ij}(t) = g^{-1}\{\hat{\gamma}_i(t) + Z^{\top}_{i(j)}\hat{\theta}(t) + X^{\top}_{ij}\hat{\beta}(t)\}$ for patient j at facility i, targeting $p'_{ij}(t) = g^{-1}\{\gamma_i(t) + Z^{\top}_{i(j)}\theta(t) + X^{\top}_{ij}\beta(t)\}$. Note that the dependence on calendar year c is

suppressed due to the nonsignificance of $\eta(c)$ and that patient-specific predictions only use information at baseline (initiation of dialysis) and therefore do not include estimates of subject-specific random effects.

Figure 3(a) displays the three predicted risk trajectories for the patient from the USRDS study cohort along with the observed risk trajectory. As expected, the predicted risk trajectory using the current case-mix and facility-level risk factors is found to be the closest to the smooth of the observed outcome. Also, the two predicted risk trajectories corresponding to a medium (with 32–44 patients) and a small (with 20–31 patients) facility are above the predictions from the current large facility (with > 44 patients), showing that smaller facilities have higher risk of hospitalization, which agrees with data analysis results outlined in Section 3.2. For a new patient who is outside the study cohort used to build the model, patient-level prediction would use the patient's current case-mix at the initiation of dialysis and facility-level covariates from facilities which the patient is considering for receiving dialysis. While we assume these candidate facilities would exist in the original data used to build the model (with their estimated $\gamma_i(t)$), their $Z_{i(j)}$ would have to be obtained to reflect facility characteristics from the time of predictions. Note that the predicted hospitalization risk based on the partly conditional VCM-MR, conditions on the patient being alive, with the predicted patient-level trajectory representing the patient's hospitalization risk t years after initiation of dialysis if the patient were alive at that time.

Prediction at the facility-level provides information to facilities for improving patient care. More specifically, it provides guidance on how much decrease in patient hospitalization risk is associated with the change in a modifiable facility-level risk factor while keeping the decomposition of the patients whom the facility is serving fixed. In facility-level prediction, the patient case-mix of the facility, including patient comorbidities, are assumed to be known and fixed, while the differing predictions correspond to the differing choices of the facility-level characteristics considered. Define the mean hospitalization risk as $p_i''(t) = \left(1/N_{it}\right)\sum_{j=1}^{N_{it}} p_{ij}''(t)$ where N_{it} is the number of patients who are alive and receiving dialysis treatment at facility i at time t and $p_{ij}''(t) = g^{-1}\left\{\gamma_i(t) + Z_{i(j)}^{\top}\theta(t) + X_{ij}^{\top}\beta(t) + b_{ij}\right\}$. Note that different from $p_{ij}'(t)$, $p_{ij}''(t)$ includes patient-specific random effects b_{ij} since the decomposition of the patient-level characteristics are assumed known and fixed in facility-level prediction. The mean hospitalization risk can be interpreted as the average hospitalization risk of all patients who are receiving treatment at facility i at time t. Using the mean risk trajectory defined above, we obtain the facility-level predicted hospitalization risk trajectory as $\hat{p}_i''(t) = \left(1/N_{it}\right)\sum_{j=1}^{N_{it}} \hat{p}_{ij}''(t)$ with $\hat{p}_{ij}''(t) = g^{-1}\left\{\hat{\gamma}_i(t) + Z_{i(j)}^{\top}\hat{\theta}(t) + X_{ij}^{\top}\hat{\beta}(t) + \hat{b}_{ij}\right\}$. Note that different from $\hat{p}_{ij}'(t)$ in patient-level prediction obtained throughout the entire follow-up, conditional on the patient being alive, the $\hat{p}_{ij}''(t)$ is obtained only over the time period for which the patient is alive, based on the assumption that the patients' survival time (S_{ij}) is known.

Figure 3(b) shows three facility-level predicted mean risk trajectories for a large facility with current PCT-to-patient (ptp) ratio of 4.5%, along with two predictions that correspond to an

increase in the facility's ptp ratio to 9.5% and 14.5%. Also plotted is the observed risk, which is the smooth of observations from all patients within the facility. The predicted mean risk trajectory using the current data (in gray) is the closest to the observed risk, as expected. Increasing the ptp ratio within the facility is associated with a decrease in its predicted average patient hospitalization risk, consistent with the data analysis findings of Section 3.2. Facility- and patient-level prediction can be assessed using relative mean squared deviation error (MSDE), $\left[\int\left\{\hat{p}_i^{\prime\prime}(t)-p_i^{\prime\prime}(t)\right\}^2dt\right]/\int\left\{p_i^{\prime\prime}(t)\right\}^2dt$, and mean squared error (MSE), $\left[\sum_{k=1}^{N_{ij}}\left\{\hat{p}_{ij}^{\prime}\left(t_{ijk}\right)-Y_{ijk}\right\}^2\right]/N_{ij}$, respectively. For the facility-level prediction, $p_i^{\prime\prime}(t)$ is set to the observed hospitalization risk in facility i. MSE is used instead of MSDE for patient-level prediction since it is well-defined even for subjects with no hospitalizations in their follow-up (where the denominator of MSDE would be zero) and for subjects with a single observation (where the integral in the denominator of MSDE cannot be computed). The (25%, 50%, 75%) percentiles of the MSDE for facility-level prediction are (.021, .044, .077), (.018, .034, .061) and (.014, .025, .043) for small, medium and large facilities, respectively. Note that the facility-level prediction gets smaller with increasing facility size, mainly due to more precise estimation of $\gamma_i(t)$ for larger facilities. The (25%, 50%, 75%) percentiles of the MSE for patient-level prediction are (.100, .200, .291) in the data application. The mean of the patient-level prediction error (.218) is slightly below the benchmark approach of using the overall mean hospitalization rate 0.27 in place of all $\hat{p}_{ij}^{\prime}\left(t_{ijk}\right)$, yielding an average prediction error of .232.

4 Simulation Studies

We carry out simulation studies to examine the finite sample properties of the proposed estimation and inference procedures. Studies include assessment of the validity of the hypothesis testing procedure, performance of the bootstrap confidence intervals and multilevel predictions. Similar to our modeling in applications to USRDS data, we consider the following partly conditional model:
$g[E\{Y_{ij}(t,c)\mid Z_{i(j)},X_{ij},b_{ij},S_{ij}>t\}]=\eta(c)+\gamma_i(t)+Z_{i(j)}^{\top}\theta(t)+X_{ij}^{\top}\beta(t)+b_{ij}$ for $t\in[0,5]$ and $c\in[0,3]$. The details of the simulation design are deferred to Appendix D of the Supporting Information.

4.1 Estimation

A preliminary simulation using the sequential 10-fold cross-validation described in Appendix B of the Supporting Information is conducted for choosing the bandwidths of the varying coefficient functions in the full simulations. The most commonly selected bandwidths for small, medium and large facilities are fixed at 1.85, 1.65 and 1.35, respectively, for estimation of $\gamma_i(t)$. In addition, the selected bandwidths for $\{\theta(t),\beta(t)\}$ and $\eta(c)$ are 1.8 and 2.4, respectively. Mean squared error (MSE) is used to assess estimation of the time-invariant model parameter σ_b^2, and relative mean squared deviation error (MSDE), $\mathrm{MSDE}_{\hat{\xi}}=\left[\int\left\{\hat{\xi}(t)-\xi(t)\right\}^2dt\right]/\int\xi^2(t)dt$ (for a generic function $\xi(t)$), is used to assess estimation of the time-varying functions, $\gamma_i(t)$, $\theta(t)$, $\beta(t)$ and $\eta(c)$. In addition, multilevel

prediction is assessed by MSE and MSDE for patient- and facility-level prediction as defined in Section 3.3, with the difference that true $p'_i(t,c)$ and $p''_{ij}\left(t_{ijk}, c_{ij}\right)$ available in the simulation setting replace the observed values utilized in data analysis. More specifically, for patient-level prediction, $p'_{ij}(t,c) = g^{-1}\{\gamma_i(t) + Z^{\top}_{i(j)}\theta(t) + X^{\top}_{ij}\beta(t) + \eta(c)\}$ and $\hat{p}'_{ij}(t,c) = g^{-1}\{\hat{\gamma}_i(t) + Z^{\top}_{i(j)}\hat{\theta}(t) + X^{\top}_{ij}\hat{\beta}(t) + \hat{\eta}(c)\}$ and for facility-level prediction $p''_{ij}(t,c) = g^{-1}\{\gamma_i(t) + Z^{\top}_{i(j)}\theta(t) + X^{\top}_{ij}\beta(t) + b_{ij} + \eta(c)\}$ and $p''_{ij}(t,c) = g^{-1}\{\hat{\gamma}_i(t) + Z^{\top}_{i(j)}\hat{\theta}(t) + X^{\top}_{ij}\hat{\beta}(t) + \hat{b}_{ij} + \hat{\eta}(c)\}$.

Simulations are conducted for two cases with $I = 100$ and 500 total number of facilities and results are presented based on 200 Monte Carlo runs. Supporting Information Figure S2 displays the estimated time-varying coefficient functions of the multilevel risk factors and $\eta(c)$, along with their 95% bootstrap CIs from the simulation run with the median MSDE based on $I = 100$ total facilities. The estimates track the true functions which lie within the CIs for most of the time points. The (25th, 50th, 75th) percentiles of the MSDEs obtained for the varying coefficient functions from both simulation cases are summarized in Table 1. Also given in Table 1 are the MSDE and MSE for facility- and patient-level prediction, respectively. The increase in the total number of facilities leads to smaller MSDE values in estimation of $\beta(t), \theta(t)$ and $\eta(c)$, as expected, but does not affect MSDEs for $\gamma_i(t)$, since their estimation is based only on within facility data. The results for $\gamma_i(t)$ are categorized by facility size where the precision in estimation of $\gamma_i(t)$ improves for larger facilities. This is also the reason for the decrease in the MSE and MSDEs of patient- and facility-level predictions with increasing facility size. Since $\gamma_i(t)$ is the component estimated with the least precision (based on facility-specific data) among the varying coefficient functions, the improvement in their estimation has the largest effect on the improvement observed in the multilevel prediction. Multilevel prediction is less affected by the increase in the total number of facilities. Note that MSDE values from estimation of $\eta(c)$ are higher compared to other varying coefficient functions due to the smaller norm of the varying coefficient function which is centered around zero for identifiability. Overall, the estimation is on target as illustrated in Supporting Information Figure S2.

4.2 Inference: Bootstrap Confidence Intervals and Hypothesis Testing

We also examine the performance of the bootstrap confidence intervals proposed for the varying coefficient functions $\beta(t), \theta(t)$, $\eta(c)$ and the validity of the proposed hypothesis testing procedure to identify significant facility-specific effects.

To study the coverage and length of the proposed bootstrap CIs for $\beta(t)$, $\theta(t)$, $\eta(c)$, results are reported from 200 Monte Carlo runs for $I = 100$ and $I = 500$ total number of facilities, at three time points in Table 2. As expected, the length of the CIs decrease with increasing number of facilities. Observed coverage probabilities (CPs) typically range between 85% −95% with a low of 77% (Table 2). This is to be expected since the proposed CIs are pointwise CIs.

To assess the validity of the proposed hypothesis testing procedure for $H_0 : \gamma_i(t) = \gamma^0(t)$, data for all facilities were generated from the simulation design described in Appendix D of

Supporting Information except for the first facility with facility-specific effects $\gamma_1(t) = \gamma_{0\delta}(t)$, where $\gamma_{0\delta}(t) = (1-\delta)\gamma^0(t) + \delta(-\sqrt{t/5}-1)$ for $\delta = 0, .25, .50, .75$, and 1, and $\gamma^0(t) = 0$. Under this setup, the first facility effect deviates more from H_0 with increasing δ, where $\gamma_{0\delta}(t) = \gamma^0(t)$ for $\delta = 0$, and $\gamma_{0\delta}(t) = -\sqrt{t/5}-1$ for $\delta = 1$. We tested $H_0 : \gamma_1(t) = \gamma^0(t)$ for $\delta = 0, .25, .50, .75$ and 1, to assess the level and power of the proposed hypothesis test. In addition to varying δ, we also generated the first facility at three different facility sizes ($N_1 = 27$, 43 and 69 subjects similar to real data, referred to as the small, medium and large facility, respectively). We considered two simulation cases with $I = 100$ and 500 facilities and calculated the test statistics r_1 and the associated p-values from 500 Monte-Carlo runs. Supporting Information Figure S3 displays three power curves from varying δ based on $I = 100$ total facilities corresponding to small, medium and large facilities that are tested (results from $I = 500$ facilities are similar). The level of the test is on target at .028, .030, .052 for small, medium, and large facilities, respectively; and the power increases with increasing δ, as expected. More specifically, note that the power at $\delta = 1$ are .452, .682 and .924 for small, medium and large facilities, respectively, and that the power increases more rapidly with increasing facility size.

5 Discussion

Studying leading risk factors of hospitalizations for patients on dialysis is important for identifying strategies that can improve their health. Due to the multilevel (subject- and facility-level) nature of the risk factors, we propose a novel varying coefficient model for multilevel risk factors to characterize effects as a function of the time patients are on dialysis. In addition to capturing multilevel effects, the proposed model allows for comparison of the effects of significant subject- and facility-level factors. To handle the computational challenges due to the high-dimensional parameter space and the large size of the data from USRDS, we develop a novel iterative estimation algorithm and an efficient hypothesis testing procedure. In the USRDS application, VCM-MR identifies significant multilevel risk factors for patient hospitalizations and leads to insights on modifiable facility-level risk factors (e.g. nurse-to-patient and PCT-to-patient ratios) which are associated with reductions in patient hospitalization risk.

We note that the proposed partly conditional modeling targets the hospitalization risk directly as the patient outcome. This can be extended through joint modeling to also handle patient survival. However, this extension requires further research as the multilevel varying coefficient models for joint modeling have not been considered to date. Finally, the goal of the proposed multilevel predictions, especially at the subject-level, is to make an entire risk trajectory prediction over the course of dialysis treatment using patient characteristics from initiation of dialysis. Time-dynamic predictions using time-varying multilevel covariates during dialysis, while also of interest, require further research. Additional issues to be considered for time-dynamic predictions would be the nature of the dependency of the time-varying response on the time-varying covariates, whether it be contemporaneous, or involving delayed covariate values or dependent on the entire covariate history.

We provide R codes for running our VCM-MR algorithm on simulated datasets on Github (https://github.com/dsenturk/VCM-MR).

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

We are grateful to two referees for insightful suggestions which improved the paper. This study was supported by research grants from the National Institute of Diabetes and Digestive and Kidney Diseases (R01 DK092232 - DS, DVN, KK, CMR, YC; K23 DK102903 - CMR, KK, DVN). The interpretation and reporting of the data presented here are the responsibility of the authors and in no way should be seen as an official policy or interpretation of the United States government.

References

1. United States Renal Data System. USRDS 2016 Annual Data Report: Epidemiology of Kidney Disease in the United States, National Institutes of Health, National Institute of Diabetes and Digestive and Kidney Diseases, Bethesda, MD, 2016.
2. Cleveland WS, Grosse E, Shyu WM. Local regression models In: Chambers JM, Hastie TJ, eds. Statistical Models in S. Pacific Grove, CA: Wadsworth & Brooks; 1991:309–376.
3. Hastie T, Tibshirani R. Varying Coefficient Models. Journal of the Royal Statistical Society: Series B. 1993;55:757–796.
4. Fan J, Zhang W. Statistical Methods With Varying Coefficient Models. Statistics and its Interface. 2008;1:179–195. [PubMed: 18978950]
5. Hoover DR, Rice JA, Wu CO, Yang LP. Nonparametric smoothing estimates of time-varying coefficient models with longitudinal data. Biometrika. 1998;85:809–822.
6. Qu A, Li R. Quadratic inference functions for varying coefficient models with longitudinal data. Biometrics. 2006;62:379–391. [PubMed: 16918902]
7. Senturk D, Dalrymple LS, Mohammed SM, Kaysen GA, Nguyen DV. Modeling time varying effects with generalized and unsynchronized longitudinal data. Statistics in Medicine. 2013;32:2971–2987. [PubMed: 23335196]
8. Estes JP, Nguyen DV, Dalrymple LS, Mu Y, Senturk D. Cardiovascular Event Risk Dynamics Over Time in Older Patients on Dialysis: A Generalized Multiple-Index Varying Coefficient Model Approach. Biometrics. 2014;70:754–764. [PubMed: 24766178]
9. Zhang D Generalized linear mixed models with varying coefficients for longitudinal data. Biometrics. 2004;60:8–15. [PubMed: 15032768]
10. Liang H, Wu H, Carroll RJ. The Relationship Between Virologic and Immunologic Responses in AIDS Clinical Research Using Mixed-effects Varying-coefficient Models With Measurement Error. Biostatistics. 2003;4:297–312. [PubMed: 12925523]
11. Wu H, Zhang JT. Local Polynomial Mixed-effects Models for Longitudinal Data. Journal of the American Statistical Association. 2002;97:883–897.
12. Crainiceanu CM, Staicu AM, Di CZ. Generalized Multilevel Functional Regression. Journal of the American Statistical Association. 2009;1104:1550–1561.
13. Serban N A Space-time Varying Coefficient Model: The Equity of Service Accessibility. Annals of Applied Statistics. 2011;5(3):2024–2051.
14. Chen H, Wang YA. Penalized Spline Approach to Functional Mixed Effects Model Analysis. Biometrics. 2010;67:861–870. [PubMed: 21155747]
15. You J, Wan AT, Liu S, Zhou Y. A varying-coefficient approach to estimating multi-level clustered data models. Test. 2015;24:417–440.
16. Multilevel Gelman A. (Hierarchical) Modeling: What It Can and Cannot Do. Technometrics. 2006;48:432–435.

17. Estes J, Nguyen DV, Chen Y, et al. Time-Dynamic Profiling with Application to Hospital Readmission Among Patients on Dialysis [published online ahead of print June 5 2018]. Biometrics. 2018. doi: 10.1111/biom.12905.

18. He K, Kalbfleisch JD, Li Y, Li Y. Evaluating Hospital Readmission Rates in Dialysis Facilities; Adjusting for Hospital Effects. Lifetime Data Analysis. 2017;19:490–512.

19. Liu D, Schaubel DE, Kalbfleisch JD. Computationally efficient marginal models for clustered recurrent event data. Biometrics. 2012;68:637–647. [PubMed: 21957989]

20. Kurland BF, Heagerty PJ. Directly parameterized regression conditioning on being alive: Analysis of longitudinal data truncated by deaths. Biostatistics. 2005;6:241–258. [PubMed: 15772103]

21. Estes JP, Nguyen DV, Dalrymple LS, Mu Y, Senturk D. Time-varying Effect Modeling With Longitudinal Data Truncated by Death: Conditional Models, Interpretations, and Inference. Statistics in Medicine. 2015;35:1834–1847. [PubMed: 26646582]

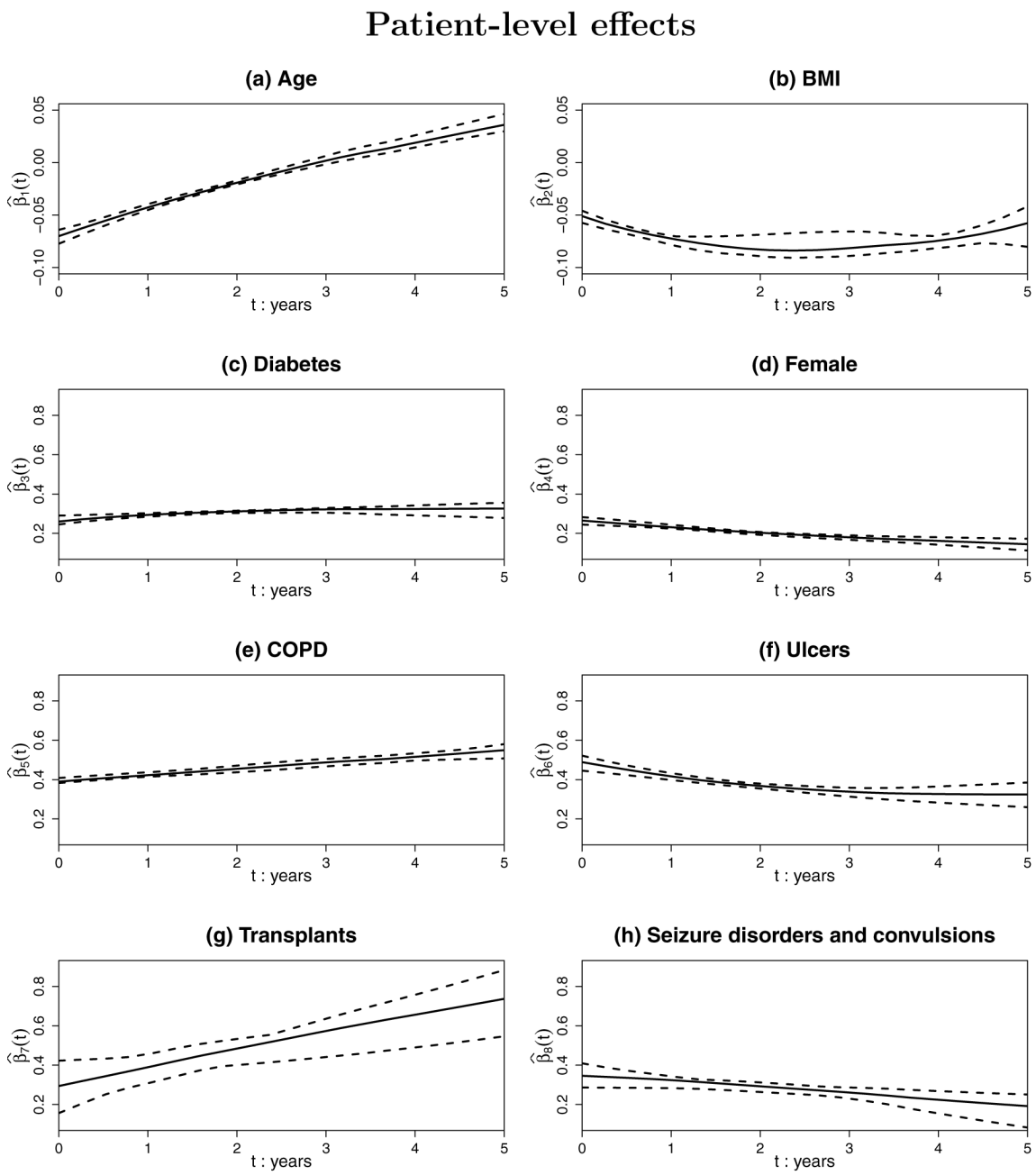

Figure 1:

Estimated patient-level effects $\hat{\beta}(t)$ (solid) on hospitalization risk corresponding to (a) a change of 10-years in age at initiation of dialysis, (b) a change of 5-units in BMI, (c) whether diabetes is the cause of ESRD, (d) gender, (e) COPD, (f) ulcers, (g) transplants, (h) seizure disorders and convulsions along with their 95% confidence intervals (dashed).

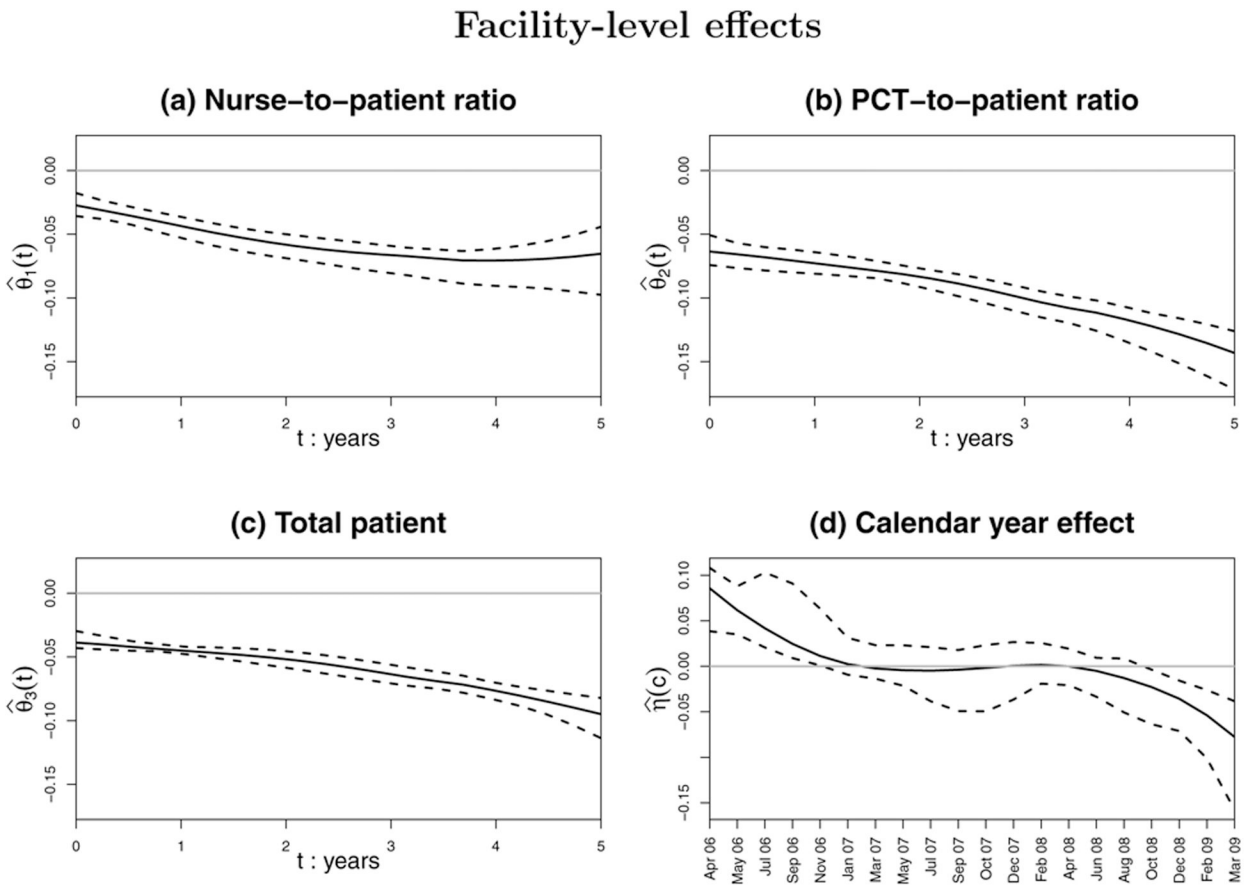

Figure 2:

Estimated facility-level effects $\hat{\theta}(t)$ (solid) on hospitalization risk corresponding to a change of (a) 5-percentage points in nurse-to-patient ratio, (b) 5-percentage points in patient care technician (PCT)-to-patient ratio and (c) 50 patients in total number of patients, along with their 95% confidence intervals (dashed). Horizontal lines at zero are plotted in gray for reference and positive numbers on the y-axis correspond to increased risk of hospitalization. All three facility-level covariates are associated with protective effects on risk of hospitalization where PCT-to-patient ratio has a stronger association than nurse-to-patient ratio. Estimated calendar year effect $\hat{\eta}(c)$ and its corresponding 95% confidence interval are plotted in (d).

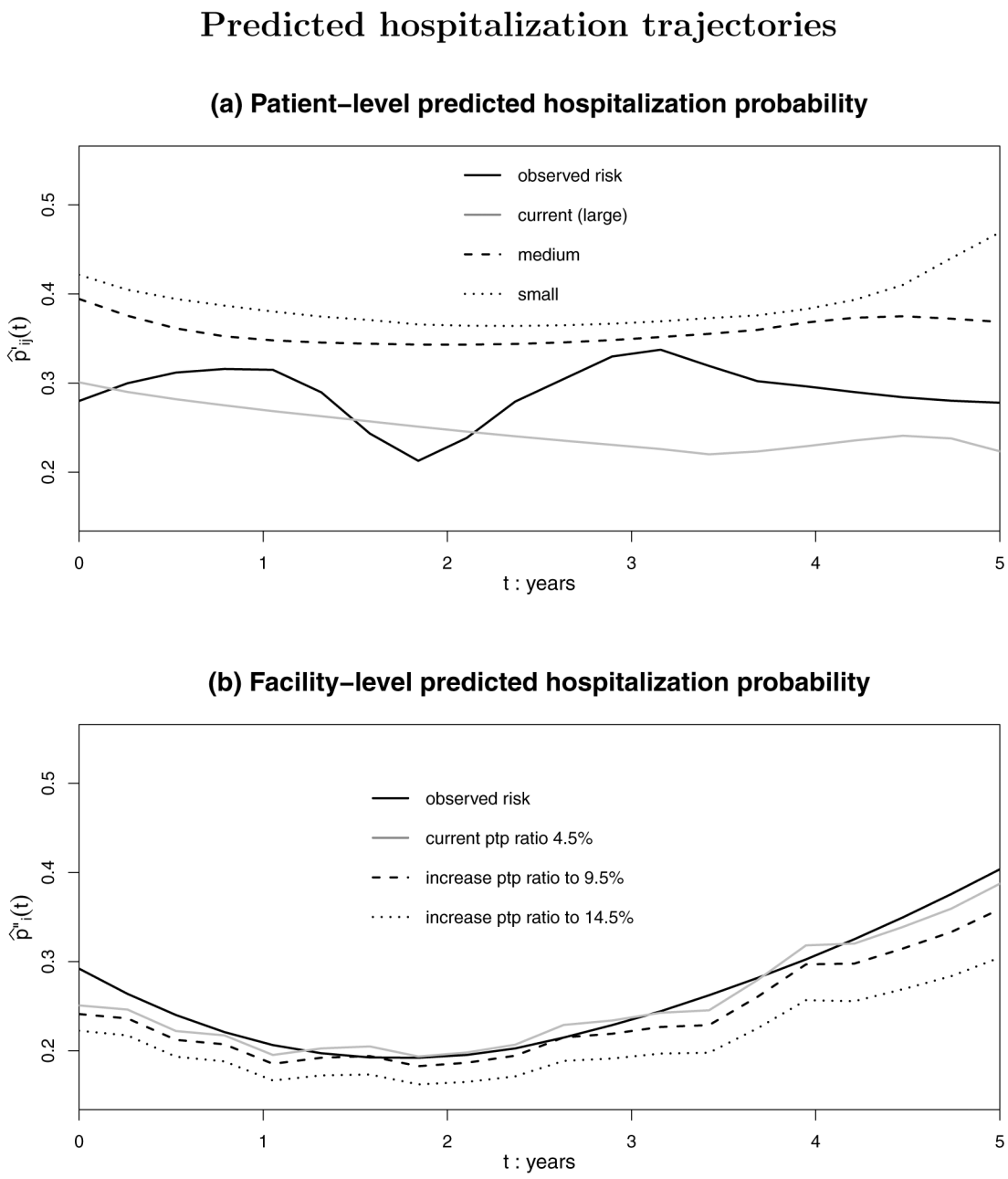

Figure 3:

(a) Patient-level predicted hospitalization probability. The smooth of a patient's observed outcomes is given in solid black. Predicted hospitalization risk trajectories from the patient's current large facility (> 44 patients) and two new facilities of sizes medium (32–44 patients) and small (20–31 patients) are given in gray, dashed and dotted, respectively. (b) Facility-level predicted hospitalization probability for a large facility. The observed risk given in solid black represents the smooth of observations from all patients within a facility. Predicted mean hospitalization risk trajectories with the current PCT-to-patient (ptp) ratio

(4.5%) and higher ptp ratios of 9.5% and 14.5%, are given in gray, dashed and dotted, respectively.

Table 1:

Percentiles of the mean squared deviation error (MSDE) for the time-varying coefficient estimates of facility-specific fixed effects $\gamma_I(t)$, effects of facility-level covariates $\theta(t)$, effects of subject-level covariates $\beta(t)$, calendar time effect at initiation of dialysis $\eta(c)$ and for facility-level prediction, based on 200 Monte Carlo runs. Percentiles of mean squared error (MSE) are reported for patient-level prediction error. MSDE for $\gamma_I(t)$ and facility-level prediction and MSE for patient-level prediction are stratified by small (20–34 patients), medium (35–54 patients) and large (> 54 patients) facility sizes.

Part I		**I=100**			**I=500**	
MSDE	**25%**	**50%**	**75%**	**25%**	**50%**	**75%**
$\widehat{\theta_1}(t)$	0.012	0.023	0.04	0.003	0.005	0.009
$\widehat{\theta_2}(t)$	0.014	0.027	0.041	0.003	0.006	0.008
$\hat{\beta}_1(t)$	0.011	0.02	0.032	0.002	0.005	0.008
$\hat{\beta}_2(t)$	0.01	0.018	0.032	0.003	0.005	0.008
$\hat{\eta}(c)$	0.382	0.583	0.917	0.29	0.425	0.53

Part II	MSDE $\hat{\gamma}_i(t)$			**Facility-level prediction (MSDE)**			**Patient-level prediction (MSE)**		
I=100	**25%**	**50%**	**75%**	**25%**	**50%**	**75%**	**25%**	**50%**	**75%**
All	.022	.045	.099	.008	.015	.028	<.001	.001	.003
Small	.035	.077	.158	.013	.024	.044	.001	.002	.006
Medium	.024	.045	.095	.008	.015	.027	<.001	.002	.004
Large	.012	.028	.056	.005	.009	.016	<.001	<.001	.002
I=500									
All	.020	.044	.096	.007	.014	.028	<.001	.001	.003
Small	.033	.073	.154	.012	.023	.043	<.001	.002	.005
Medium	.022	.046	.093	.008	.015	.027	<.001	.002	.003
Large	.013	.027	.053	.005	.009	.016	<.001	<.001	.002

Table 2:

Coverage probability (CP, in %) and length (LEN) of the 95% bootstrap confidence intervals at three time points (beginning, middle and end of follow-up) for varying coefficient functions (VCFs) associated with facility-level covariates ($\theta(t)$), subject-level covariates ($\beta(t)$) and calendar time effects at initiation of dialysis ($\eta(c)$).

	$I = 100$						$I = 500$					
Year	$t = 0$		$t = 2.5$		$t = 5$		$t = 0$		$t = 2.5$		$t = 5$	
VCF	**CP**	**LEN**	**CP**	**LEN**	**CP**	**LEN**	**CP**	**LEN**	**CP**	**LEN**	**CP**	**LEN**
$\theta_1(t)$	87.0	.464	76.5	.233	89.5	1.04	92.0	.221	80.0	.119	93.5	.496
$\theta_2(t)$	91.0	.468	79.0	.235	93.0	1.05	86.0	.222	83.5	.117	92.0	.492
$\beta_1(t)$	85.0	.426	77.5	.234	94.0	.974	83.5	.202	77.0	.115	95.0	.472
$\beta_2(t)$	89.0	.426	82.5	.233	94.0	.976	87.0	.202	78.0	.115	92.0	.467
$\eta(c)$	94.0	.272	90.0	.126	90.0	.265	91.5	.182	92.5	.091	93.5	.181