Research Article

Statistics in Medicine

Received 18 July 2014, Accepted 31 August 2015 Published online 24 September 2015 in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.6743

A flexible model for correlated medical costs, with application to medical expenditure panel survey data

Jinsong Chen,[a] Lei Liu,[b*†] Ya-Chen T. Shih,[c‡] Daowen Zhang[d] and Thomas A. Severini[e]

We propose a flexible model for correlated medical cost data with several appealing features. First, the mean function is partially linear. Second, the distributional form for the response is not specified. Third, the covariance structure of correlated medical costs has a semiparametric form. We use extended generalized estimating equations to simultaneously estimate all parameters of interest. B-splines are used to estimate unknown functions, and a modification to Akaike information criterion is proposed for selecting knots in spline bases. We apply the model to correlated medical costs in the Medical Expenditure Panel Survey dataset. Simulation studies are conducted to assess the performance of our method.

Keywords: model selection; generalized linear model; health econometrics; semiparametric regression; generalized estimating equation

1. Introduction

Escalating medical costs are a central fiscal challenge for the United States. National spending on health care accounted for 18% of gross domestic product (GDP) in 2011 and is projected to reach 25% of GDP by 2037 [1]. In a recent New York Times interview, world-renowned health economist Victor Fuchs emphasized that, to solve the nation's fiscal problems, attention must be paid to health care spending [2]. Thus, the analysis of medical cost data plays an important role in health economics and health policy.

While documenting the growth in medical costs actuarially serves the purpose of alerting policy makers and the public the scope and urgency of the health care spending problem, without identifying key cost drivers, these actuary estimates offer little guidance on strategies to contain the growth of medical costs. Because of several characteristics of medical cost data, understanding factors associated with medical costs requires innovative statistical methods. Specifically, the challenging features of medical cost data include heteroscedasticity, severe skewness, and non-normality. Until recently, most of the published literatures concern the analysis of medical costs in the cross-sectional setting. A review of earlier methods to analyze medical cost data can be found in Diehr *et al*. [3]. A comprehensive overview was published in a special issue devoted to this topic in Medical Care (vol. 47, no. 7 supplement 1, July 2009). In brief, to address the challenge in analyzing medical cost data, researchers [4–6] used generalized linear models (GLM) describing the link and variance structure by pre-specified functions, for example, log link with gamma error distribution. The extensions of GLM to address the aforementioned issues have been developed. Chiou and Müller [7] proposed a nonparametric quasi-likelihood method to model the

[a]*Department of Medicine, University of Illinois at Chicago, Chicago, IL, U.S.A.*
[b]*Department of Preventive Medicine, Northwestern University, Chicago, IL, U.S.A.*
[c]*Department of Medicine, The University of Chicago, Chicago, IL, U.S.A.*
[d]*Department of Statistics, North Carolina State University, Raleigh, NC, U.S.A.*
[e]*Department of Statistics, Northwestern University, Evanston, IL, U.S.A.*
[*]*Correspondence to: Lei Liu, Department of Preventive Medicine, Northwestern University, 680 North Lakeshore Drive, Suite 1400, Chicago, IL 60611, U.S.A.*
[†]*E-mail: Lei.liu@northwestern.edu*
[‡]*Current address: Department of Health Services Research, M. D. Anderson Cancer Center, TX, U.S.A.*

variance as an unspecified function of the mean using kernel smoothing methodology. Basu and Rathous [8] developed a model where the variance structure is modeled via a flexible parametric function of the mean. Chen *et al*. [9] suggested a generalized partial linear model with a nonlinear variance function for cross-sectional data and adopted penalized nonparametric quasi-likelihood for estimation.

More recently, there has been increasing interest in analyzing correlated medical cost data. The correlation might stem from the structure of clustered medical cost data, where medical costs of subjects from the same cluster are correlated because of similarities in their health status, socioeconomic characteristics, and shared genetic traits [10]. Another source of correlation arises in longitudinal medical cost data where repeated measures of costs in different time intervals (e.g., monthly medical costs) are correlated for the same subject [11].

Our motivating example is the clustered medical costs from elderly households in the medical expenditure panel survey (MEPS) dataset. We include 2139 individuals aged between 65 and 84 years within 1556 households from the MEPS 2010 full year consolidated data file of the household survey. The histogram of medical costs is shown in Figure 1. It can be seen that medical costs are skewed to the right and possibly heteroscedastic. Furthermore, medical costs of individuals in elderly households are likely to be correlated for several reasons. For example, health behaviors, attitudes, and believes tend to be shared among individuals in the same household; these factors can translate into similar medical care-seeking behaviors, which then affect medical costs. Finally, based on our previous study, there may be nonlinear covariate effects for some covariates of interest, for example, age [9], which could be accounted for by spline-based methods.

For such correlated data, marginal models with generalized estimating equations (GEE) [12] (termed GEE1 in [13]) are a useful extension of GLMs. In GEE1, the goal is to estimate regression parameters (first moment), while the correlation parameters (second moment) are treated as nuisance. Prentice and Zhao [14] and Zhao and Prentice [15] suggested a second-order extension of GEE1 whose goal is to obtain more efficient estimates of the correlation parameters, assuming the response follows a quadratic exponential model (GEE2). In GEE2, the third and fourth moments are treated as nuisance, while the incorrect specification of which will not affect the consistent estimates of the first and second moments. However, GEE2 pays the price that the correct specification of the mean and covariance models is necessary to ensure consistent estimates in the first and second moments. Hall and Severini [16] improved GEE2 through extended generalized estimating equations (EGEE). It retains the strength of GEE2, which can efficiently estimate correlation structure. On the other hand, EGEE does not require a correct covariance specification for the consistent estimation of regression parameters. Further, no third or fourth moments are needed in the EGEE approach, simplifying the computation.

However, both the mean and covariance structure of EGEE [16] contain only parametric components (termed 'EGEE-P' hereafter). In this article, we combine the features of our previous nonparametric model for medical costs [9] and the EGEE method. The result is an EGEE with a partially linear mean

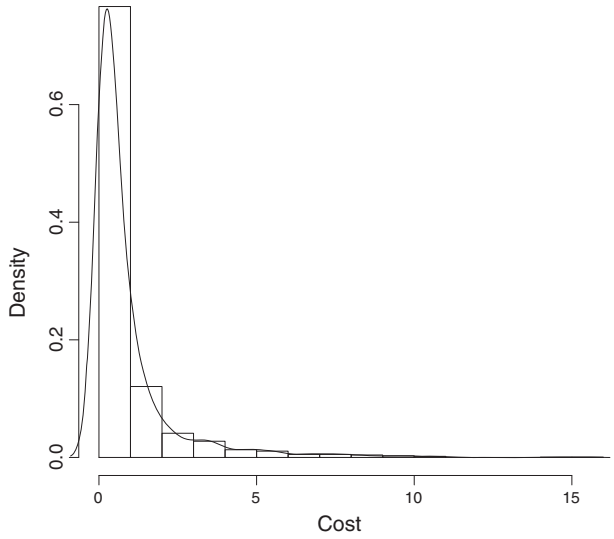

Figure 1. Histogram of annual medical costs (in US$1000) for the medical expenditure panel survey data. The solid line is the estimated density.

function and semiparametric covariance structure for correlated (e.g., clustered or longitudinal) medical cost data. We would call our new method as 'EGEE-N'. This method is more flexible in capturing possible nonlinear covariate effects in the mean function, as well as the unspecified relationship between the mean and variance. At the same time, we retain efficiency in estimating these parameters as in EGEE.

The rest of the paper is organized as follows. The model is introduced in Section 2. Section 3 describes estimation and inference. In Section 4, we apply our method to correlated medical costs in MEPS. The results of simulation studies are presented in Section 5. A summary of our conclusions and implications of this work are given in the final section.

2. Model

For the ith subject/cluster ($i = 1, \cdots, m$), let $\mathbf{Y}_i = (Y_{i1}, \cdots, Y_{ij}, \cdots, Y_{in_i})^T$ be the $n_i \times 1$ vector of correlated outcomes, $\mathbf{X}_i = (\mathbf{x}_{i1}, \cdots, \mathbf{x}_{ij}, \cdots, \mathbf{x}_{in_i})^T$ be the $n_i \times p$ matrix of linear covariates, and $\mathbf{z}_i = (z_{i1}, \cdots, z_{ij}, \cdots, z_{in_i})^T$ be the $n_i \times 1$ vector of nonlinear covariates. Define $E(Y_{ij}|\mathbf{x}_{ij}, z_{ij}) = \mu_{ij}$. We propose a generalized partially linear marginal model with a semiparametric covariance structure for correlated data as follows:

$$\begin{aligned} g(\mu_{ij}) &= \mathbf{x}_{ij}^T\boldsymbol{\beta} + f(z_{ij}), \\ \log\{\mathrm{Var}(Y_{ij})\} &= \eta(\mu_{ij}), \end{aligned} \tag{1}$$

where $g(\cdot)$ is a known monotone and differentiable link function, for example, log link for medical costs, $\boldsymbol{\beta}$ is a vector of linear regression coefficients, and $f(\cdot)$ is an unknown function. Variance $\mathrm{Var}(Y)$ is modeled as an unknown but smooth function of μ, that is, $\exp\{\eta(\mu)\}$. Of note, model (1) is an adaptation of Chen *et al*. [9] for cross-sectional data to the correlated data setting. Without loss of generality, we consider one-dimensional nonlinear predictor z in model (1). It can be extended for multiple nonlinear predictors.

For correlated data, it is important to account for the correlation among observations within the same subject/cluster. We model the covariance matrix of response with the incorporation of a working correlation matrix. For the ith cluster, the $n_i \times n_i$ covariance matrix $\mathrm{Var}(\mathbf{Y}_i) = \mathbf{V}_i$ is assumed to have the structure

$$\mathbf{V}_i = \mathbf{A}_i^{\frac{1}{2}}\mathbf{R}(\boldsymbol{\alpha})\mathbf{A}_i^{\frac{1}{2}},$$

where $\mathbf{A}_i = \mathrm{diag}\{\mathrm{Var}(Y_{i1}), \ldots, \mathrm{Var}(Y_{in_i})\} = \mathrm{diag}[\exp\{\eta(\mu_{i1})\}, \ldots, \exp\{\eta(\mu_{in_i})\}]$, and $\mathbf{R}(\boldsymbol{\alpha})$ is a symmetric working correlation matrix (e.g., compound symmetry or AR(1)) of $\mathbf{Y}_i$. $\mathbf{R}(\boldsymbol{\alpha})$ is specified by the $s \times 1$ vector $\boldsymbol{\alpha}$, which we assume is not a function of the mean in this paper. Thus, the covariance matrix has a semiparametric structure.

3. Estimation and inference

We use B-splines to estimate the nonlinear functions because they are numerically stable and attractive for theoretical development [17, 18]. The unknown function $f(\cdot)$ can be approximated by $f = \mathbf{B}\boldsymbol{\tau}$, where $\mathbf{B}$ and $\boldsymbol{\tau}$ are the B-splines basis with degree q and coefficient vector, respectively. The mean function is then $g(\mu) = \mathbf{x}^T\boldsymbol{\beta} + \mathbf{B}\boldsymbol{\tau} = \mathbf{X}\boldsymbol{\theta}$, where $\mathbf{X} = (\mathbf{x}^T, \mathbf{B})$ and $\boldsymbol{\theta} = (\boldsymbol{\beta}^T, \boldsymbol{\tau}^T)^T$. Similarly, we can approximate the variance function as $\eta(\mu) = \mathbf{B}_\mu\boldsymbol{\gamma}$.

Let $\boldsymbol{\omega} = (\boldsymbol{\theta}^T, \boldsymbol{\gamma}^T, \boldsymbol{\alpha}^T)^T$, where the elements of $\boldsymbol{\theta}$, $\boldsymbol{\gamma}$, and $\boldsymbol{\alpha}$ are denoted as $(\theta_1, \ldots, \theta_l)$, $(\gamma_1, \ldots, \gamma_r)$, and $(\alpha_1, \ldots, \alpha_s)$, respectively. Following Hall and Severini [16], we use extended generalized estimating functions with a nonparametric component for the ith cluster

$$\mathbf{D}_i(\boldsymbol{\theta}) = \left(\frac{\partial\boldsymbol{\mu}_i^T}{\partial\boldsymbol{\theta}}\right)^T \mathbf{V}_i^{-1}(\mathbf{y}_i - \boldsymbol{\mu}_i);$$

$$\mathrm{D}_i(\gamma_b) = -(\mathbf{y}_i - \boldsymbol{\mu}_i)^T\frac{\partial\mathbf{V}_i^{-1}}{\partial\gamma_b}(\mathbf{y}_i - \boldsymbol{\mu}_i) + \mathrm{Tr}\left(\mathbf{V}_i\frac{\partial\mathbf{V}_i^{-1}}{\partial\gamma_b}\right), b = 1, \ldots, r;$$

$$\mathrm{D}_i(\alpha_c) = -(\mathbf{y}_i - \boldsymbol{\mu}_i)^T\frac{\partial\mathbf{V}_i^{-1}}{\partial\alpha_c}(\mathbf{y}_i - \boldsymbol{\mu}_i) + \mathrm{Tr}\left(\mathbf{V}_i\frac{\partial\mathbf{V}_i^{-1}}{\partial\alpha_c}\right), c = 1, \ldots, s$$

to simultaneously estimate $\boldsymbol{\theta}$, γ_b, and α_c, where $\partial \boldsymbol{\mu}_i^T/\partial\boldsymbol{\theta} = (\partial\mu_{i1}/\partial\boldsymbol{\theta}, \ldots, \partial\mu_{in_i}/\partial\boldsymbol{\theta})$ and Tr is the matrix trace.

Let $\mathbf{D}_i(\boldsymbol{\omega}) = \{\mathbf{D}_i(\boldsymbol{\theta})^T, \mathbf{D}_i(\boldsymbol{\gamma})^T, \mathbf{D}_i(\boldsymbol{\alpha})^T\}^T$, where $\mathbf{D}_i(\boldsymbol{\gamma}) = \{\mathrm{D}_i(\gamma_1), \ldots, \mathrm{D}_i(\gamma_r)\}^T$ and $\mathbf{D}_i(\boldsymbol{\alpha}) = \{\mathrm{D}_i(\alpha_1), \ldots, \mathrm{D}_i(\alpha_s)\}^T$. The estimates $\hat{\boldsymbol{\omega}}$ for model (1) are the solutions to the extended generalized estimating equation (EGEE-N):

$$\mathbf{D}(\boldsymbol{\omega}) = \sum_{i=1}^{m} \mathbf{D}_i(\boldsymbol{\omega}) = 0.$$

EGEE-N can be solved by the Fisher scoring algorithm:

$$\boldsymbol{\omega}^{(\varkappa+1)} = \boldsymbol{\omega}^{(\varkappa)} - \left[\sum_{i=1}^{m} E\left\{\frac{\partial \boldsymbol{D}_i(\boldsymbol{\omega}^{(\varkappa)})}{\partial \boldsymbol{\omega}^{(\varkappa)}}\right\}\right]^{-1} \mathbf{D}(\boldsymbol{\omega}^{(\varkappa)}). \tag{2}$$

The ith expected Jacobian can be shown to be

$$E\left\{\frac{\partial \boldsymbol{D}_i(\boldsymbol{\omega})}{\partial \boldsymbol{\omega}}\right\} = \begin{pmatrix} -\left(\frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\theta}}\right)^T \mathbf{V}^{-1} \left(\frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\theta}}\right) & \mathbf{0} & \mathbf{0} \\ \left[\mathrm{Tr}\left(\frac{\partial \mathbf{V}_i}{\partial \boldsymbol{\theta}_a}\frac{\partial \mathbf{V}_i^{-1}}{\partial \boldsymbol{\gamma}_b}\right)\right]_{r\times l} & \left[\mathrm{Tr}\left(\frac{\partial \mathbf{V}_i}{\partial \boldsymbol{\gamma}_{b_1}}\frac{\partial \mathbf{V}_i^{-1}}{\partial \boldsymbol{\gamma}_{b_2}}\right)\right]_{r\times r} & \left[\mathrm{Tr}\left(\frac{\partial \mathbf{V}_i}{\partial \boldsymbol{\alpha}_c}\frac{\partial \mathbf{V}_i^{-1}}{\partial \boldsymbol{\gamma}_b}\right)\right]_{r\times s} \\ \left[\mathrm{Tr}\left(\frac{\partial \mathbf{V}_i}{\partial \boldsymbol{\theta}_a}\frac{\partial \mathbf{V}_i^{-1}}{\partial \boldsymbol{\alpha}_c}\right)\right]_{s\times l} & \left[\mathrm{Tr}\left(\frac{\partial \mathbf{V}_i}{\partial \boldsymbol{\gamma}_b}\frac{\partial \mathbf{V}_i^{-1}}{\partial \boldsymbol{\alpha}_c}\right)\right]_{s\times r} & \left[\mathrm{Tr}\left(\frac{\partial \mathbf{V}_i}{\partial \boldsymbol{\alpha}_{c_1}}\frac{\partial \mathbf{V}_i^{-1}}{\partial \boldsymbol{\alpha}_{c_2}}\right)\right]_{s\times s} \end{pmatrix}, \tag{3}$$

where the $r \times l$ matrix is

$$\left[\mathrm{Tr}\left(\frac{\partial \mathbf{V}_i}{\partial \boldsymbol{\theta}_a}\frac{\partial \mathbf{V}_i^{-1}}{\partial \boldsymbol{\gamma}_b}\right)\right]_{r\times l} \text{ with the } (a,b)th \text{ element being } \mathrm{Tr}\left(\frac{\partial \mathbf{V}_i}{\partial \boldsymbol{\theta}_a}\frac{\partial \mathbf{V}_i^{-1}}{\partial \boldsymbol{\gamma}_b}\right).$$

The estimate of $E\{\partial \boldsymbol{D}_i(\boldsymbol{\omega})/\partial\boldsymbol{\omega}\}$ can be obtained by substituting $\hat{\boldsymbol{\omega}}$ into expression (3).

EGEE-N simultaneously estimates the linear parameters, nonlinear function, variance function, and correlation parameters. Chiou and Müller [7] and Chen *et al.* [9] used an alternative strategy to estimate variance function $V(\cdot)$ based on equation $\varepsilon^2 = V(\mu) + \zeta$, where $\varepsilon^2 = (Y-\mu)^2$ and ζ is a mean zero error term. In practice, $\hat{\varepsilon}^2 = V(\hat{\mu}) + \zeta$, where $\hat{\varepsilon}^2 = (Y-\hat{\mu})^2$, is used for nonparametric regression because ε^2 and μ are unknown. We refer this approach as the residual-based method because the residual $Y - \hat{\mu}$ is inserted into the estimation process. A drawback of the residual-based method is that the nonparametric regression suffers from the problem of errors in variable, for example, $(\hat{\varepsilon}^2 = \varepsilon^2 + e_{\varepsilon^2}, \hat{\mu} = \mu + e_\mu)$ contain unobservable errors $(e_{\varepsilon^2}, e_\mu)$. Consequently, this approach results in biased estimation of the variance function [19]. Furthermore, the computation of the residual-based method is demanding. It often relies on a two-stage iterative procedure: (i) estimate parameters in the mean function given the estimated variance function and (ii) obtain the estimate of the variance function given the mean function. In contrast, the EGEE-N method avoids these problems by estimating all parameters simultaneously, having computational and inferential advantages.

The consistency of regression parameter estimates in both GEE and EGEE-P does not require the correct specification of the covariance structure. However, we have less-efficient estimates of regression parameters when the variance function in GEE or EGEE-P is misspecified. On the other hand, a nonparametric alternative has been considered by Chiou and Müller [7] and Li [20], modeling the conditional covariance of response variable, $\mathrm{cov}(Y_{i1}, Y_{i2}|\mathbf{x}_1, \mathbf{x}_2)$, through a bivariate smoothing method. Lin and Pan [21] proposed separate nonparametric estimates for the correlation structure and the error variance for longitudinal data. The covariance estimate in EGEE-N can be viewed as a semiparametric estimate in that the variance function estimate is nonparametric while the working correlation matrix is given a parametric form. Our EGEE-N approach, in between a fully parametric approach and a fully nonparametric approach, provides a trade-off between model complexity and robustness to misspecification of the variance function.

3.1. Selection of initial values

Selecting good starting values for algorithm (2) is crucial for convergence and efficiency of the algorithm. Denote $\boldsymbol{\omega}^{(0)} = [\{\boldsymbol{\theta}^{(0)}\}^T, \{\boldsymbol{\gamma}^{(0)}\}^T, \{\boldsymbol{\alpha}^{(0)}\}^T]^T$ as the starting values. They are chosen based on the GEE method. At first, we obtain $\boldsymbol{\theta}^{(0)}$ via GEE using an independent constant covariance structure. That is, we have $\boldsymbol{\theta}^{(0)}$ via the GEE iterative procedure (expression (8) in [22]) using identity covariance matrix. Then, $\boldsymbol{\gamma}^{(0)}$ is acquired from the model $\log(\hat{\varepsilon}_i^2) = \eta(\hat{\mu}_i) + \zeta_i$. Finally, we compute correlation parameters $\boldsymbol{\alpha}^{(0)}$ using method-of-moments estimators based on Pearson residuals [22].

3.2. Knot selection

Akaike information criterion (AIC) is a useful model-selection criterion for semiparametric models [23, 24]. However, the likelihood-based AIC cannot be directly applied to GEE where there is no well-defined likelihood. Pan [25] proposed a modification to AIC, quasi-likelihood under the independence model criterion (QIC), for GEE. In this paper, we suggest a nonparametric version of QIC for selecting spline bases in EGEE-N and name it nonparametric QIC (NQIC). Our focus is to estimate linear coefficients $\boldsymbol{\beta}$ and the nonlinear function $f(\cdot)$, and treat the variance $\mathrm{Var}(Y)$ as nuisance. Therefore, NQIC is used to choose spline bases in the mean function while a fixed set of dense knots is given to the spline for the variance function to reduce the computational burden.

The nonparametric quasi-likelihood for data under independence is defined in [7, 9]:

$$\tilde{Q}(\boldsymbol{\theta}, \mathbf{I}) = \sum_{i=1}^{m} \sum_{j=1}^{n_i} \int_{y_{ij}}^{\mu_{ij}} \frac{y_{ij} - t}{\hat{\mathcal{V}}(t)} dt,$$

where $\hat{\mathcal{V}}(\cdot)$ is the nonparametrically estimated variance function, which is $\exp\{\hat{\eta}(\cdot)\}$ in our case. The quasi-likelihood assumes a working independence correlation structure. It is known that a more general working correlation structure does not guarantee the existence of a quasi-likelihood, and it is generally difficult to construct even if it exists [26]. In this paper, we propose a criterion based on quasi-likelihood under the working independence model with an estimated $\boldsymbol{\theta}$ from EGEE-N.

Denote by $\hat{\boldsymbol{\theta}}_{EGEE-N}$ the estimate of $\boldsymbol{\theta}$ by solving EGEE-N. Following Pan [25], we define NQIC for estimated $\hat{\boldsymbol{\theta}}_{EGEE-N}$ as follows:

$$NQIC(\hat{\boldsymbol{\theta}}_{EGEE-N}) = -2\tilde{Q}(\hat{\boldsymbol{\theta}}_{EGEE-N}, \mathbf{I}) + 2\mathrm{Tr}(\hat{\mathbf{H}}_I \hat{V}_{\boldsymbol{\theta}}),$$

where $\hat{\mathbf{H}}_I = -\partial^2 \tilde{Q}(\boldsymbol{\theta}, \mathbf{I}) / \partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_{EGEE-N}}$ and $\hat{V}_{\boldsymbol{\theta}}$ is the consistent estimate of $\mathrm{cov}(\hat{\boldsymbol{\theta}}_{EGEE-N})$, which can be obtained by the sandwich estimate described in Section 3.3.

Knots of spline basis are selected by minimizing NQIC. That is, different sets of knots result in different spline bases, and selecting knots is absorbed in the process of model-selection with various spline bases. Knot selection includes the choice of the location and number of knots. A convenient option is to place knots in the sample quantiles of the continuous covariate z [27]. Eilers and Marx [28] suggested equally spaced knots in practice to simplify the computation, which also facilitates interpolation. The cubic B-splines with equidistant knots are used in our study. We select the number of knots via a grid search. The optimal number of knots is chosen by minimizing NQIC over 11 different numbers of knots from 2 to 12 in simulation studies and the application of MEPS data.

3.3. Inferential results

Define $\boldsymbol{\psi} = \left(\boldsymbol{\beta}^T, f^T, \eta^T, \boldsymbol{\alpha}^T\right)^T$ to be the true parameter value. The EGEE-N estimator is $\hat{\boldsymbol{\psi}} = \left(\hat{\boldsymbol{\beta}}^T, \hat{f}^T, \hat{\eta}^T, \hat{\boldsymbol{\alpha}}^T\right)^T$, where $\hat{f} = \mathbf{B}\hat{\boldsymbol{\tau}}$ and $\hat{\eta} = \mathbf{B}_\mu \hat{\boldsymbol{\gamma}}$. The covariance matrix of $\hat{\boldsymbol{\psi}}$ is

$$V(\hat{\boldsymbol{\psi}}) = \mathbf{G}\left[E\left\{\frac{\partial \mathbf{D}(\boldsymbol{\omega})}{\partial \boldsymbol{\omega}}\right\}\right]^{-1}\left[\sum_{i=1}^{m} E\left\{\mathbf{D}_i(\boldsymbol{\omega})\mathbf{D}_i(\boldsymbol{\omega})^T\right\}\right]\left[E\left\{\frac{\partial \mathbf{D}(\boldsymbol{\omega})}{\partial \boldsymbol{\omega}}\right\}\right]^{-1}\mathbf{G}^T, \tag{4}$$

where $\mathbf{G} = \mathrm{blockdiag}(\mathbf{I}_{\boldsymbol{\beta}}, \mathbf{B}, \mathbf{B}_\mu, \mathbf{I}_{\boldsymbol{\alpha}})$ with identity matrices $\mathbf{I}_{\boldsymbol{\beta}}$ and $\mathbf{I}_{\boldsymbol{\alpha}}$ compatible in dimension with $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$.

Following Hall and Severini [16], the consistency of $\hat{\boldsymbol{\psi}}$ require a condition: the solution to $E\{\mathbf{D}(\boldsymbol{\omega})\} = 0$ is unique. Whether this condition is satisfied for $\left(\hat{\boldsymbol{\beta}}^T, \hat{f}^T\right)$ does not depend on the correct specification of

the covariance matrix. Therefore, the consistency of $\left(\hat{\boldsymbol{\beta}}^T, \hat{f}^T\right)$ is satisfied regardless of whether the covariance matrix is correctly specified. The consistency of $\hat{\alpha}$ requires the correct specification of covariance. The correct specification of $\mathbf{R}(\alpha)$ is a condition for the consistency of $\hat{\eta}$.

In covariance matrix (4), the center part is as follows:

$$E\left\{\mathbf{D}_i(\boldsymbol{\omega})\mathbf{D}_i(\boldsymbol{\omega})^T\right\} = \begin{pmatrix} \mathbf{D}_i(\boldsymbol{\theta})\mathbf{D}_i(\boldsymbol{\theta})^T & \mathbf{D}_{i_{\theta\gamma}} & \mathbf{D}_{i_{\theta\alpha}} \\ \mathbf{D}_{i_{\theta\gamma}}^T & \mathbf{D}_i(\boldsymbol{\gamma})\mathbf{D}_i(\boldsymbol{\gamma})^T & \mathbf{D}_i(\boldsymbol{\gamma})\mathbf{D}_i(\boldsymbol{\alpha})^T \\ \mathbf{D}_{i_{\theta\alpha}}^T & \mathbf{D}_i(\boldsymbol{\alpha})\mathbf{D}_i(\boldsymbol{\gamma})^T & \mathbf{D}_i(\boldsymbol{\alpha})\mathbf{D}_i(\boldsymbol{\alpha})^T \end{pmatrix},$$

where $\mathbf{D}_{i_{\theta\gamma}} = -(\partial\boldsymbol{\mu}_i/\partial\boldsymbol{\theta})^T\mathbf{V}_i^{-1}(\mathbf{y}_i-\boldsymbol{\mu}_i)[(\mathbf{y}_i-\boldsymbol{\mu}_i)^T(\partial\mathbf{V}_i^{-1}/\partial\gamma_b)(\mathbf{y}_i-\boldsymbol{\mu}_i)]_{r\times 1}^T$ and $\mathbf{D}_{i_{\theta\alpha}} = -(\partial\boldsymbol{\mu}_i/\partial\boldsymbol{\theta})^T\mathbf{V}_i^{-1}(\mathbf{y}_i-\boldsymbol{\mu}_i)[(\mathbf{y}_i-\boldsymbol{\mu}_i)^T(\partial\mathbf{V}_i^{-1}/\partial\alpha_c)(\mathbf{y}_i-\boldsymbol{\mu}_i)]_{s\times 1}^T$. The covariance estimate $\hat{V}(\hat{\boldsymbol{\psi}})$ can be obtained via substituting $\hat{\boldsymbol{\omega}}$ for $\boldsymbol{\omega}$ in expression (4).

Of note, our confidence interval formulas treat selected knots as fixed. That is, we do not take into account the variability introduced during the knot selection in the model estimation and inference. The results in statistical literature [28] and our experience indicate that ignoring this variability has a minimum effect on the performance of the estimated parameters and nonparametric functions. One may use the bootstrap approach to account for this variability. However, this will be computationally very expensive.

4. Application

We construct a subset of elderly households from the MEPS 2010 Full Year Consolidated Data File of the Household Survey, consisting of households in which every member was 65 years of age or over in the survey year. Using the elderly households as our study sample allows us to isolate the statistical issue to clustered medical costs without having to address additional analytical issues, such as massive zeros in medical costs (because very few elderly individuals would have zero medical expenditures) or endogeneity associated with the choice of insurance (because Medicare eligibility starts at age 65, except for those with end-stage renal disease or disabled). The final study sample includes 2139 individuals aged between 65 and 84 years within 1556 households. Because age in MEPS is top-coded at age 85 years, that is, age over 85 years is coded at 85, we choose the upper range as 84 to avoid erroneously attributing age effects of individuals over age 85 years to those at 85.

The response variable in our model is annual medical costs in US dollars. Medical costs are highly skewed to the right (Figure 1), as is evident by a much larger mean (US$9235) than median (US$3955). Of the individuals in our study sample, the proportions for male, white race, having any hospitalization in the survey year, and death are 41.1%, 79.5%, 15.0%, and 1.4%, respectively. To account for the impact of an individual's underlying disease(s) on his/her medical costs, we consolidate 11 highly prevalent health conditions collected in MEPS into five disease categories: heart and blood vessel disease (HBV, 82.7%), respiratory disease (16.9%), body movement disorder (76.6%), cancer (28.9%), and diabetes (21.9%). Each disease category is quantified as a dichotomous variable and included as a covariate in our analysis. There are a small portion (3.6%) of subjects with no cost, which is changed to $1 to accommodate the log link function.

We apply the semiparametric model (1) with exchangeable correlation to capture the possible correlation in medical costs for members within the same household. Specifically, the mean function of the model has the form $\log(\mu_{ij}) = \mathbf{x}_{ij}^T\boldsymbol{\beta} + f(z_{ij})$. The covariate vector in the linear term includes 10 binary (1 for yes and 0 for no) predictors: male, white, death, hospitalization, each of the five disease categories, and a dichotomous variable indicating family composition (1 for family size greater than one, 0 for living alone). The continuous predictor in the nonlinear term z_{ij} is age.

The estimated coefficients for linear predictors are shown in Table I. Gender and race are not significantly associated with annual medical costs, whereas significantly positive associations are observed between annual medical costs and death, hospitalization, and all five disease categories. Elderly living alone tended to have higher medical costs, although the effect is only marginally significant (P-value = 0.082). Such association can be explained by the fact that compared with those who lived alone, elderly couples tended to take care of each other and thus reduce the need for home-based medical service or extended hospitalization because of the concern of unsafe home environment if patients were to be discharged home alone. The estimated correlation parameter is 0.041 and marginally

significant (P-value = 0.062), suggesting elderly members in the same family shared similar medical spending patterns.

The knot selection method described in Section 3.2 is applied to select six knots for the B-splines basis for estimating the nonlinear function of age, while we use eight knots for variance function estimation. The estimated curve for age is shown in the left panel of Figure 2. We observe a nonlinear age effect on medical costs. There is a small dip at first, then the age effect increases from 67 years, reaches a maximum value around 75 years, and fluctuates after age 75 years. The small dip right after age 65 years may reflect a tempered effect in health care utilization after a sharp increase in the use of medical services among those who were previously uninsured or underinsured before they became Medicare eligible at age 65 years. Medical costs then increase with age up to 75 years as the health status deteriorates with age. After reaching 75 years of age, patients might receive less aggressive treatments because of concerns of multiple chronic conditions so their costs fluctuate and do not exhibit a clearly increasing pattern. The estimated log variance function $\eta(\cdot)$ is presented in the right panel of Figure 2. There is a clear nonlinear curvature, which indicates evidence of heteroscedasticity, as the variance increases with the mean.

Table I. Estimates of linear coefficients for medical expenditure panel survey data.

Covariates	Estimate	s.e.	P-value
White	−0.048	0.067	0.468
Male	0.026	0.058	0.655
Death	0.706	0.140	< 0.001
Hospitalization	1.489	0.066	< 0.001
HBV disease	0.409	0.100	0.001
Respiratory disease	0.235	0.065	< 0.001
Body movement disorder	0.222	0.077	0.004
Cancer	0.252	0.062	0.001
Diabetes	0.386	0.062	< 0.001
Family	−0.099	0.057	0.082

Note: HBV, heart and blood vessel disease.

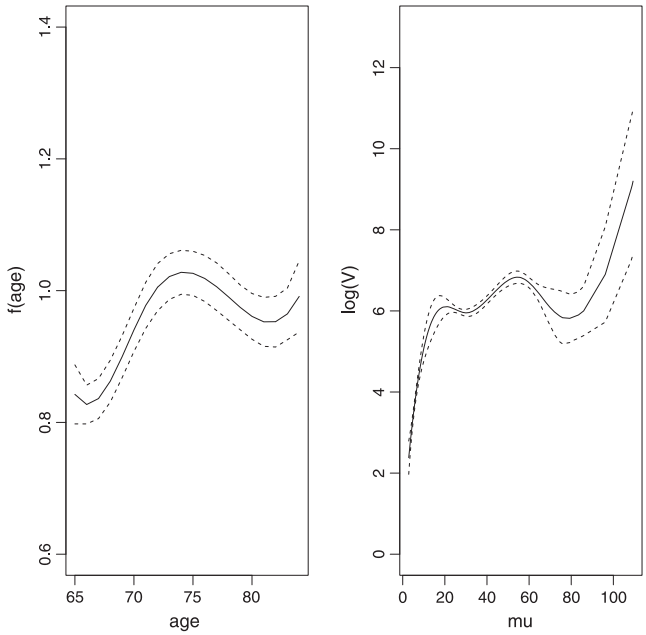

Figure 2. Curve estimation for medical expenditure panel survey data. Left: estimated $f(age)$ with 95% pointwise confidence interval; right: estimated variance function ('mu' represents the mean medical cost in the unit of US$1000.)

5. Simulation

Simulation studies are conducted based on two different settings for longitudinal data. There are 1000 datasets generated in each setting for 100 subjects and five observations within each subject.

Setting 1: The simulated data are generated from a multivariate normal distribution. The response is generated from the following model:

$$Y_{ij} = x_{ij}\beta + f(t_{ij}) + e_{ij}$$

where $\beta = 1$ and $f(t) = \sin(2\pi t)/2$. The covariate x is uniformly distributed over $[-1, 1]$. We let time t be a covariate varying within each subject with 100 equally spaced points in $[0, 1]$, that is

$$t_{ij} = \frac{\text{trun}\{(i+4)/5\}}{100} + 0.2(j-1) - 0.01.$$

Table II. Estimates of linear coefficients and correlation parameter in the simulation studies.

Method	$\eta(\mu)$	Bias	SD	SE	CP (%)	$\eta(\mu)$	Bias	SD	SE	CP (%)
Data generated from exchangeable correlation										
Estimation of β										
	Normal data					Gamma data				
EGEE-P	0	0.0064	0.062	0.060	95.0	$\log(\mu^2/2)$	−0.0016	0.042	0.041	94.8
EGEE-N	$\hat{\eta}$	0.0073	0.063	0.059	94.4	$\hat{\eta}$	−0.0011	0.042	0.041	94.2
Estimation of α										
	Normal data					Gamma data				
EGEE-P	0	0.0065	0.034	0.032	91.6	$\log(\mu^2/2)$	0.0179	0.072	0.068	93.2
EGEE-N	$\hat{\eta}$	0.0029	0.048	0.046	93.9	$\hat{\eta}$	0.0039	0.070	0.069	92.1
Data generated from AR(1) correlation										
Estimation of β										
	Normal data					Gamma data				
EGEE-P	0	−0.0020	0.067	0.061	94.3	$\log(\mu^2)$	0.0009	0.064	0.062	94.9
EGEE-N	$\hat{\eta}$	−0.0012	0.067	0.062	94.3	$\hat{\eta}$	0.0014	0.062	0.061	94.8
Estimation of α										
	Normal data					Gamma data				
EGEE-P	0	0.0044	0.034	0.033	90.8	$\log(\mu^2)$	0.0113	0.065	0.064	93.3
EGEE-N	$\hat{\eta}$	0.0025	0.042	0.040	92.9	$\hat{\eta}$	−0.0067	0.066	0.063	92.5

Note: SD, sampling standard deviation of estimates; SE, average of the estimated standard errors; CP, empirical coverage probability; $\hat{\eta}$, nonparametric estimate of $\eta(\mu)$; EGEE, extended generalized estimating equations.

Table III. Mean coverage probabilities of estimates of $f(t)$ in the simulation studies.

Method	$\eta(\mu)$	CP (%)	$\eta(\mu)$	CP (%)
Data generated from exchangeable correlation				
	Normal data		Gamma data	
EGEE-P	0	94.8	$\log(\mu^2/2)$	94.4
EGEE-N	$\hat{\eta}$	94.3	$\hat{\eta}$	94.1
Data generated from AR(1) correlation				
	Normal data		Gamma data	
EGEE-P	0	94.8	$\log(\mu^2)$	94.6
EGEE-N	$\hat{\eta}$	94.4	$\hat{\eta}$	94.6

Note: EGEE, extended generalized estimating equations.

The random error e_{ij} associated with observation j for subject i has variance 1 with two correlation structures: exchangeable correlation $\alpha = 0.5$ and AR(1) correlation with $\alpha = 0.5$.

Setting 2: The response variable follows a gamma distribution G(a, b) with density $f(y) = 1/\{\Gamma(a)b^a)\}y^{a-1}\exp(-y/b)$. The link function is

$$\log(\mu_{ij}) = x_{ij}\beta + f(t_{ij})$$

where $\beta = 0.3$ and $f(t) = \sin(2\pi t)/2$. The linear covariate x and nonlinear covariate t are generated in the same way as in Setting 1. The correlation structures under consideration are exchangeable with $\alpha = 0.5$ and AR(1) with $\alpha = 0.5$. We take $y_{ij} \sim G(2, \mu_{ij}/2)$ with exchangeable correlation and $y_{ij} \sim G(\mu_{ij}, 1)$ with AR(1) correlation. The variance functions are $V(y) = \mu^2/2$ and $V(y) = \mu^2$ for exchangeable and AR(1), respectively. The data are simulated based on the algorithm of Park and Shin [29] using a gamma distribution with a given mean structure and variance function.

We fit data generated from Settings 1 and 2 using the EGEE-P method with a correct parametric variance function and EGEE-N with an unspecified variance function. To fit each dataset, the NQIC criterion and grid search method are used to select spline knots to estimate $f(\cdot)$ in the mean function, while eight fixed knots are used to estimate the unknown variance function of the mean. EGEE-P serves as the gold standard to assess the performance of EGEE-N. It can be seen from Table II that our proposed EGEE-N procedure performs well in estimating both the linear coefficient and correlation parameter in terms of biases, sample standard deviations (SD), estimated standard errors (SE), and empirical coverage probabilities (CP). The mean coverage probabilities of estimates of $f(t)$ (averaged across all 500 observations within each dataset, then over 1000 simulation datasets) are satisfactory as shown in Table III. The point-wise biases, mean squared errors (MSEs), and empirical CP of $f(t)$ are shown in Figures 3 and 4. The estimates of nonlinear functions of EGEE-N are very close to those of EGEE-P. Of note, the

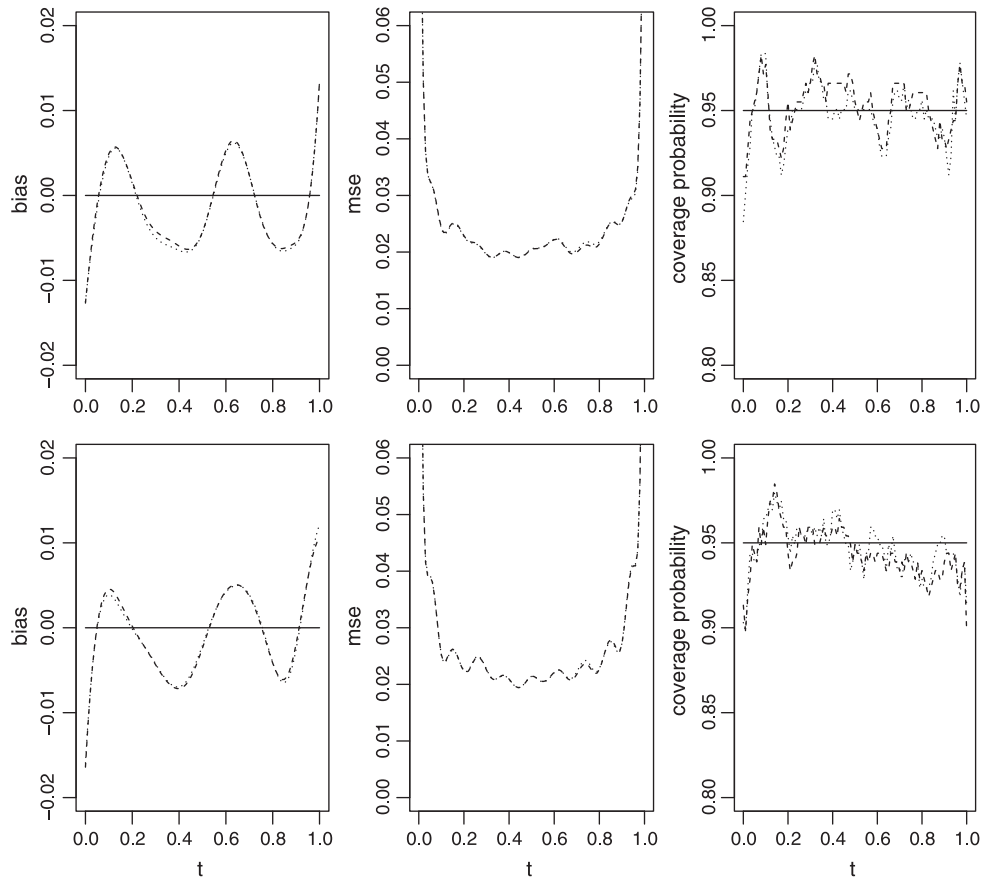

Figure 3. Simulation results of estimated $f(t)$ for normal data. Upper: exchangeable correlation; lower: AR(1) correlation. Left: plot of point-wise biases; middle: plot of point-wise mean squared errors; right: plot of point-wise empirical coverage probabilities. Dashed line: extended generalized estimating equations (EGEE)-N; dotted line: EGEE-P.

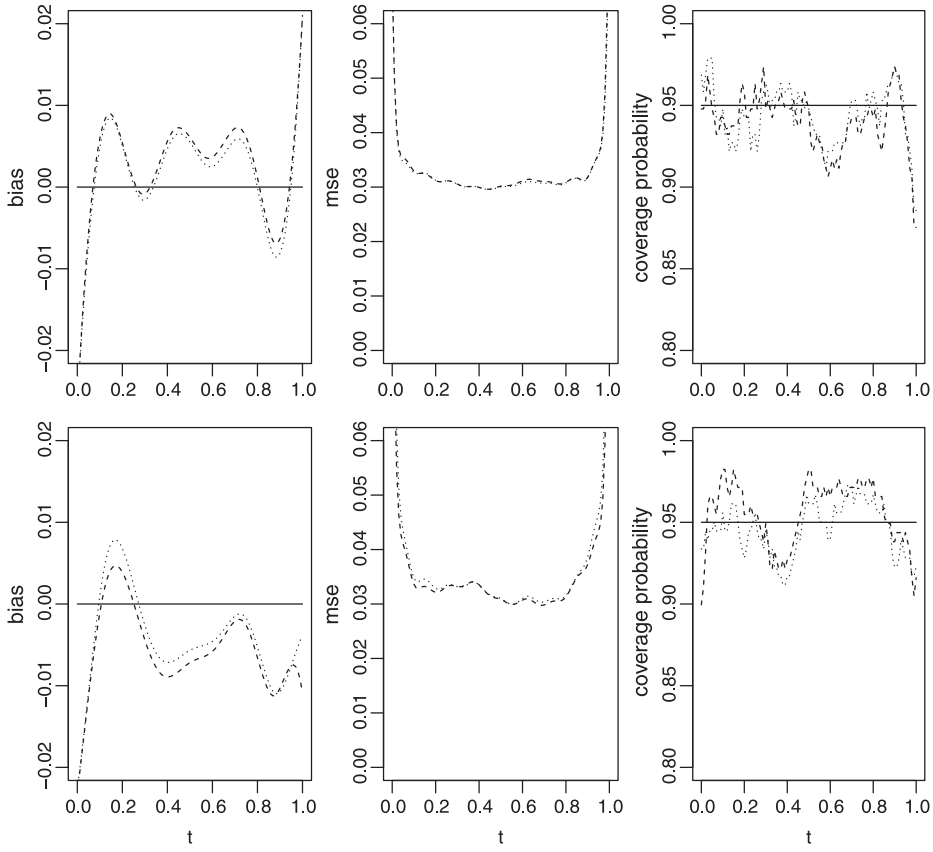

Figure 4. Simulation results of estimated $f(t)$ for gamma data. Upper: exchangeable correlation; lower: AR(1) correlation. Left: plot of point-wise biases; middle: plot of point-wise mean squared errors; right: plot of point-wise empirical coverage probabilities. Dashed line: extended generalized estimating equations (EGEE)-N; dotted line: EGEE-P.

Table IV. Results of predicted μ in the simulation studies.

Method	Data generated from exchangeable correlation					
	Normal data			Gamma data		
	Bias	MSE	CP (%)	Bias	MSE	CP (%)
EGEE-P	−0.006	0.026	94.0	0.003	0.012	93.8
EGEE-N	−0.005	0.027	93.5	−0.005	0.012	93.6
Method	Data generated from AR(1) correlation					
	Normal data			Gamma data		
	Bias	MSE	CP (%)	Bias	MSE	CP (%)
EGEE-P	−0.006	0.024	94.0	−0.004	0.017	93.2
EGEE-N	−0.005	0.025	93.7	−0.015	0.017	93.1

Note: MSE, average mean squared errors; CP, empirical coverage probabilities; EGEE, extended generalized estimating equations.

slightly larger biases in the EGEE-P estimation of α for gamma data setting in Table II might be due to random variation caused by the relatively small sample size. Additional simulation with a cluster size of 200 indicates the unbiased estimates of α of the EGEE-P method (results available upon request). Finally, per the suggestion of a reviewer, we show the estimates of μ, the mean of bias, MSE, and empirical CP (averaged across all 500 observations, then over 1000 simulation datasets) in Table IV, which indicate the appropriateness of our method in predicting the mean response.

In Chen *et al*. [9], in the cross-sectional setting, we compared our model (1) with the generalized additive model (GAM) (e.g., Wood [30]), which assumed a parametric variance structure and a parametric distribution. Through simulation studies, we found that when the parametric assumptions do not hold,

the GAM, implemented by the gam() function in the R mgcv package, had relatively larger biases and SDs compared with the results of our methods. We expect similar results for the correlated medical cost data. Similarly, the results from extensive simulation in [9] showed that the penalized quasi-likelihood method with the misspecified variance function did not have substantial impact on the consistency of the estimates, but resulted in a loss of efficiency in estimating both linear coefficients and smoothing functions. For our proposed EGEE-N method for correlated data, preliminary simulation studies show a similar scenario.

6. Discussion

Motivated by the analysis of clustered medical cost data, we proposed a flexible generalized partially linear model with a semiparametric covariance structure. We developed the extended generalized estimating equation approach using B-splines and a modification to AIC for selecting the spline basis. The method performed reasonably well in simultaneously estimating regression coefficients, nonlinear functions, correlation parameters, and variance function. The simulation studies confirmed the validity of our method and the application of MEPS demonstrated the utility of our method.

The link function $g(\cdot)$ in model (1) is not restricted to identity or log link functions for continuous data. The incorporation of other flexible and appropriate link functions in our method may be helpful for the analysis of medical cost data [8, 31]. Our semiparametric modeling for the covariance structure satisfies positive definiteness. In our model, we do not investigate the model-selection for choosing the appropriate working correlation structure because our goal is to estimate the mean function. Moreover, it will be interesting to model the conditional covariance $\mathrm{cov}(Y_{i1}, Y_{i2}|\mathbf{x}_1, \mathbf{x}_2)$ or correlation $\mathrm{corr}(Y_{i1}, Y_{i2}|\mathbf{x}_1, \mathbf{x}_2)$ as a smoothing surface using two-dimensional B-splines approximation, for example, bivariate tensor-product B-splines [32]. We can also extend our model by assuming that the correlation parameter depends on the mean.

In our study sample constructed from the MEPS data, there is a small portion of individuals with zero medical costs, and we simplify the analysis by changing the zeros to \$1 to accommodate the log link function. When the portion of zero values is substantial, we can use the two-part model [10, 33, 34] to study the odds of costs being zero (no medical services) and the amount of medical costs among those with positive costs, accounting for the cross-part correlation between these two quantities. Researchers who are interested in incorporating the two-part model to our proposed EGEE-N method can use the full household sample from the MEPS, which includes individuals at all ages and should have a large proportion of zero costs because of the inclusion of young, healthy individuals. Another interesting methodological research is to expand the EGEE-N method to address the issue of endogeneity. Here, the non-elderly households in the MEPS offer an opportunity to explore this research because insurance variable is often endogeneous.

Acknowledgements

The authors thank the referees and associate editor for their careful reading and valuable comments. This research is partly supported by AHRQ R01 HS 020263, NIH/NCI R01 CA 85848 and NSF DMS-1308009. Dr Liu is a consultant to Celladon, Zensun, and Outcome Research Solutions, Inc. The R code is available from the corresponding author.

References

1. CBO (congressional budget office). The 2012 long-term budget outlook, 2012. https://www.cbo.gov/publication/43288. [Accessed on 15 September 2015].
2. Kolata G. Knotty challenges in health care costs. *The New York Times*. March 6, 2012; http://www.nytimes.com/2012/03/06/health/policy/an-interview-with-victor-fuchs-on-health-care-costs.html. [Accessed on 15 September 2015].
3. Diehr P, Yanez D, Ash A, Hornbrook M, Lin DY. Methods for analyzing health care utilization and costs. *Annual Review of Public Health* 1999; **20**:125–144.
4. Blough DK, Madden CW, Hornbrook MC. Modeling risk using generalized linear models. *Journal of Health Economics* 1999; **18**:153–171.
5. Manning WG, Mullahy J. Estimating log models: to transform or not to transform?. *Journal of Health Economics* 2001; **20**:461–494.
6. Manning WG, Basu A, Mullahy J. Generalized modeling approaches to risk adjustment of skewed outcomes data. *Journal of Health Economics* 2005; **20**:465–488.
7. Chiou JM, Muller HG. Nonparametric quasi-likelihood. *The Annals of Statistics* 1999; **27**:36–64.

8. Basu A, Rathous PJ. Estimating marginal and incremental effects on health outcomes using flexible link and variance function models. *Biostatistics* 2005; **6**:93–109.
9. Chen J, Liu L, Zhang D, Shih YT. A flexible model for the mean and variance functions, with application to medical cost data. *Statistics in Medicine* 2013; **32**:4306–4318.
10. Liu L, Strawderman RL, Cowen ME, Shih YCT. A flexible two-part random effects model for correlated medical costs. *Journal of Health Economics* 2010; **29**:110–123.
11. Liu L, Huang XL, O'Quigley J. Analysis of longitudinal data in the presence of informative observational times and a dependent terminal event, with application to medical cost data. *Biometrics* 2008; **64**:950–958.
12. Liang KY, Zeger SL. Longitudinal data analysis using generalized linear models. *Biometrika* 1986; **73**:13–22.
13. Liang KY, Zeger SL, Qaqish B. Multivariate regression analyses for categorical data (with discussion). *Journal of the Royal Statistical Society, Ser. B* 1992; **45**:3–40.
14. Prentice RL, Shao LP. Estimating equations for parameters in means and covariance of multivariate discrete and continuous responses. *Biometrics* 1991; **47**:825–839.
15. Zhao LP, Prentice RL. Correlated binary regression using a quadratic exponential model. *Biometrika* 1990; **77**:642–648.
16. Hall DB, Severini TA. Extended generalized estimating equations for clustered data. *Journal of the American Statistical Association* 1993; **93**:1365–1375.
17. Schumaker LL. *Spline Functions*. Wiley: New York, 1981.
18. Zhou S, Shen X, Wolfe DA. Local asymptotics for regression aplines and confidence regions. *The Annals of Statistics* 1998; **26**:1760–1782.
19. Fan J, Truong YK. Nonparametric regression with errors in variables. *The Annals of Statistics* 1993; **21**:1900–1925.
20. Li Y. Efficient semiparametric regression for longitudinal data with nonparametric covariance estimation. *Biometrika* 2011; **98**:355–370.
21. Liu H, Pan J. Nonparametric estimation for mean and covariance structures for longitudinal data. *The Canadian Journal of Statistics* 2013; **41**:557–574.
22. Liang KY, Zeger SL. Longitudinal data analysis using generalized linear models. *Biometrika* 1986; **73**:13–22.
23. Simonoff JS, Tsai CL. Semiparametric and additive model selection using an improved akaike information criterion. *Journal of Computational and Graphical Statistics* 1990; **8**:22–40.
24. Wood SN. Fast stable direct fitting and smoothness selection for generalized additive models. *Journal of the Royal Statistical Society, Ser. B* 2008; **70**:495–518.
25. Pan W. Akaike's information criterion in generalized estimating equations. *Biometrics* 2006; **57**:120–165.
26. McCullagh P, Nelder JA. *Generalized Linear Models*. Chapman Hall: New York, 1989.
27. He X, Zhu Z, Fung WK. Robust estimation in generalized partial linear mdoels for clustered data. *Journal of the American Statistical Association* 2005; **472**:1176–1184.
28. Eilers PHC, Marx BD. Splines, knots, and penalties. *Wiley Interdisciplinary Reviews: Computational Statistics* 2010; **2**:637–653.
29. Park CG, Shin DW. An algorithm for generating correlated random variables in a class of infinitely divisible distributions. *Journal of Statistical Computation and Simulation* 1998; **61**:127–139.
30. Wood SN. *Generalized Linear Models: An Introduction with R*. Chapman & Hall: New York, 2006.
31. Huang JZ, Zhang L, Zhou L. Efficient estimation in marginal partially linear models for longitudinal/clustered data using splines. *Scandinavian Journal of Statistics* 2007; **34**:451–477.
32. He X, Xi P. Bivariate tensor-product B-splines in a partly linear model. *Journal of Multivariate Analysis* 1996; **58**:162–181.
33. Tooze JA, Grunwald GK, Jones RH. Analysis of repeated measures data with clumping at zero. *Statistical Methods in Medical Research* 2002; **11**:341–355.
34. Chen J, Liu L, Johnson BA, O'Quigley J. Penalized likelihood estimation for nonparametric mixed models, with application to alcohol treatment research. *Statistics in Medicine* 2013; **32**:335–346.