Statistics and Its Interface Volume 7 (2014) 87–99

Estimation of rank-tracking probabilities using nonparametric mixed-effects models for longitudinal data

Xin Tian* and Colin O. Wu

An important scientific objective of longitudinal studies involves tracking the probability of a subject having certain health status over the course of the study. Proper definitions and estimates of disease risk tracking have important implications in the design and analysis of long-term biomedical studies and in developing guidelines for disease prevention and intervention. We study in this paper a class of "rank-tracking probabilities" (RTP) to describe a subject's conditional probabilities of having certain health outcomes at two different time points. Structural nonparametric estimation and inferences for the RTPs and their functions are developed based on nonparametric mixed-effects models and B-spline smoothing methods. Statistical properties of our procedures are investigated through a simulation study. We apply our methods to an epidemiological study of childhood cardiovascular risk factors, and demonstrate that the RTPs and their nonparametric estimators provide useful tools to quantitatively evaluate whether the cardiovascular risks, such as obesity and hypertension, can be tracked from early childhood to adolescence.

AMS 2000 subject classifications: Primary 62H10, 62G08; secondary 62P10, 65D10.
Keywords and phrases: Basis approximation, Conditional distribution, Longitudinal study, Mixed model, Time-varying coefficient model, Rank-tracking probability.

1. INTRODUCTION

Because the subjects are repeatedly measured over time, longitudinal studies are commonly used in biomedical research for the evaluation of population-means or subject-specific temporal trends of the outcome variables. Most statistical methods in longitudinal analysis, such as the mixed-effects models or nonparametric regression models, are focused on evaluating the effects of time and covariates on the conditional-means of the outcome variables with the potential serial correlations taken into account. Recent summaries of longitudinal methods can be found, for example, in Verbeke and Molenberghs (2000), Diggle et al. (2002) and Fitzmaurice et al. (2009), among others. In addition to the conditional-mean based regression approaches, conditional-distribution or quantile based regression models have also been shown to be an effective tool for the analysis of repeated measurements data (e.g., Hall et al., 1999; Wei et al., 2006; Wu et al., 2010). These methods focus on evaluating the covariate effects on the distributions of the outcome variables over time, and may lead to better interpretations when the underlying scientific objectives are specified by the distribution functions.

In addition to the above regression analysis, many biomedical studies require the evaluation of subjects at multiple time points. An important scientific objective of longitudinal studies is to track the likelihood of a subject having certain health status at a later time point given the subject's health status at an earlier time point. Kavey et al. (2003) discussed the importance of tracking the cardiovascular risk factors over the years beginning in childhood with regard to primary prevention of the subsequent cardiovascular disease in adulthood. The existing statistical methods for longitudinal analysis mentioned above, although useful in various settings, do not provide a direct measure for this type of "tracking ability" of disease risk factors. Another class of statistical methods that is somewhat relevant to the concept of tracking ability is the estimation of serial correlations across different time points. Intuitively, if a subject's health conditions at different time points are positively correlated, then subjects with undesirable health status at an earlier time are expected to be more likely to have undesirable health status at a later time. Statistical evidence for the strength of correlation is then presented by the estimates of the covariance matrices. Some recent covariance estimation methods are discussed, for example, in Wu and Pourahmadi (2003) and Fan and Wu (2008). Serial correlations, however, may give some evidence of the tracking ability, but are insufficient to be used as a quantitative measure of the likelihood of risk factor tracking over time.

The National Growth and Health Study (NGHS) is a good example that illustrates the importance of developing a novel statistical quantity to directly measure the tracking probability under this context. This is a large epidemiological study of childhood growth and cardiovascular risks of 2,379 girls, who were 9 or 10 years old at enrollment,

*Corresponding author.

with up to 10 annual visits during 1986–1997 and detailed anthropometric and laboratory measurements obtained during each visit (NGHSRG, 1992). Previous publications, such as Daniels et al. (1998) and Obarzanek et al. (2010), investigated the effects of age, race and obesity on the cardiovascular risk factors, such as blood pressures (BP), using the conditional-mean based methods. One question that has not been investigated for the NGHS data is: What is the probability that a girl, who had obesity or hypertension at a younger age, will still have such high risk factors at an older age? Since obesity and hypertension of a child are defined by the conditional distributions and conditional quantiles of body mass index (BMI) and BP given the child's age, gender and height (NHBPEP, 2004; Obarzanek et al., 2010), an appropriate answer to this question may be used to justify longitudinal studies in young children and track those who are already overweight or hypertensive.

We study in this paper the estimation of a class of conditional probabilities, namely the "rank-tracking probabilities" (RTP), to quantitatively measure a subject's conditional probabilities of having certain health status at two different time points. The RTPs and their nonparametric estimators have been recently studied by Wu and Tian (2013a) with discrete and time-invariant covariates and Wu and Tian (2013b) under a two-step local smoothing method with a class of time-varying transformation models. We propose a class of structural nonparametric global smoothing methods for the estimation of the RTPs and the related RTP ratios (RTPR) with continuous and time-dependent covariates. Two classes of nonparametric mixed-effects models are used to predict the subject-specific outcome trajectories and estimate the RTPs of the outcome variables given certain time-invariant or time-dependent covariates. A resampling-bootstrap procedure is used to construct the pointwise confidence intervals. Applying our procedures to the NGHS data, we demonstrate that the RTPs and their estimates lead to useful interpretations for large longitudinal studies. For the statistical properties, we conduct a simulation study to evaluate the biases, variances and mean-squared errors of our smoothing estimation methods. Due to the limitation of space, we focus on the methodology development and potential applications of the RTPs and their functions, so that the asymptotic properties of our estimation methods are not developed here.

The rest of the paper is organized as follows. We introduce the definitions of the RTPs and RTPRs in Section 2, derive the estimation methods and their bootstrap confidence intervals in Section 3, and present in Section 4 and Section 5 the application of our procedures to the NGHS data and the results from our simulation study, respectively. In Section 6, we discuss some potential extensions and modeling approaches for the RTPs and their applications in longitudinal studies.

2. RANK-TRACKING PROBABILITIES

2.1 Data structure and assumptions

We consider a data structure that is both mathematically tractable and commonly used in longitudinal studies. Following the NGHS design (NGHSRG, 1992), we assume that our longitudinal sample contains n independent subjects and the ith subject has n_i number of visits at time points $t_{ij} \in \mathcal{T}$, where $\mathcal{T}$ is the time interval of the study. At any time $t \in \mathcal{T}$, $Y(t)$ is the real-valued outcome variable and $\mathbf{X}(t) = (X_1(t), \ldots, X_P(t))^T$ is the R^P-valued covariate vector including time-invariant baseline covariates or time-dependent covariates. The observed longitudinal sample for $\{\mathbf{X}(t), Y(t), t\}$ is $\{\mathbf{X}_i(t_{ij}), Y_i(t_{ij}), t_{ij};\ j = 1, \ldots, n_i,\ i = 1, \ldots, n\}$. While the study investigators usually prespecify the numbers of visits and each visit time at the design stage, in practice, the actual visit times and the number of visits may vary across individual subjects.

2.2 Rank-tracking probabilities

In many situations, a subject's health status is determined by its rank relative to the population of interest, such as quantiles or conditional quantiles, of certain outcome variables. This is particularly important in pediatric studies, since risk classifications established for adults may not be appropriate for children. To introduce the idea of "rank-tracking", we consider first the case of time-invariant covariate, i.e., $\mathbf{X}(t) \equiv \mathbf{X}$ for all $t \in \mathcal{T}$. For any $t \in \mathcal{T}$, we have a risk set $A(\mathbf{X}, t)$ such that the health status of a subject at time t is determined by whether $Y(t) \in A(\mathbf{X}, t)$. For example, $Y(t)$ is the BMI for young girls, and $A(\mathbf{X}, t)$ is the set of overweight and obese girls at age t defined by the 85th percentile of the CDC growth chart (Obarzanek et al. 2010).

Following the definition of Wu and Tian (2013a), for a given set $\mathcal{X}$ and a subject with covariate $\mathbf{X} \in \mathcal{X}$, the tracking ability of $Y(t)$ at any two time points $s_1 < s_2$ can be measured by the conditional probability of $Y(s_2) \in A(\mathbf{X}, s_2)$ given $Y(s_1) \in A(\mathbf{X}, s_1)$ and $\mathbf{X} \in \mathcal{X}$. A natural definition for the RTP based on $A(\cdot, \cdot)$ at $s_1 < s_2$ is

$$\begin{aligned}(1)\quad &\mathrm{RTP}_A\big(s_1, s_2; \mathcal{X}\big)\\ &= P\Big[Y(s_2) \in A\big(\mathbf{X}, s_2\big)\Big|Y(s_1) \in A\big(\mathbf{X}, s_1\big), \mathbf{X} \in \mathcal{X}\Big].\end{aligned}$$

In some applications, we may not need the condition of $\mathbf{X} \in \mathcal{X}$, so that $\mathcal{X}$ is chosen to be the entire space of the covariates and the RTP is

$$(2)\quad \mathrm{RTP}_A\big(s_1, s_2\big) = P\Big[Y(s_2) \in A\big(\mathbf{X}, s_2\big)\Big|Y(s_1) \in A\big(\mathbf{X}, s_1\big)\Big].$$

When $A(\cdot, \cdot)$ does not depend on $\mathbf{X}$, $\mathrm{RTP}_A(s_1, s_2; \mathcal{X}) = P[Y(s_2) \in A(s_2)|Y(s_1) \in A(s_1), \mathbf{X} \in \mathcal{X}]$ and $\mathrm{RTP}_A(s_1, s_2) = P[Y(s_2) \in A(s_2)|Y(s_1) \in A(s_1)]$.

In general, both the outcome $Y(t)$ and the risk set $A(\cdot, \cdot)$ may depend on the time-varying covariates $\mathbf{X}(t)$. The results of Thompson et al. (2007) and Obarzanek et al. (2010)

suggest that tracking the risk factors of cardiovascular disease, such as BP and lipids, depends on the subject's time-varying covariates, such as height and BMI. In this case, we have a time-varying subset $\mathcal{X}(t)$ for the covariates and a risk set $A[\mathbf{X}(t), t)]$. Substituting $\{\mathcal{X}, A[\mathbf{X}, t]\}$ of (1) with $\{\mathcal{X}(t), A[\mathbf{X}(t), t]\}$ and conditioning on $\mathbf{X}(s_1) \in \mathcal{X}(s_1)$, the RTP is

$$\begin{aligned}(3)\quad &\mathrm{RTP}_A\big[s_1, s_2; \mathcal{X}(s_1)\big] \\ &= P\Big\{Y(s_2) \in A\big[\mathbf{X}(s_2), s_2\big] \Big| Y(s_1) \in A\big[\mathbf{X}(s_1), s_1\big], \\ &\qquad\qquad \mathbf{X}(s_1) \in \mathcal{X}(s_1)\Big\}.\end{aligned}$$

Similarly, if $\mathcal{X}(s_1)$ is the entire space of the covariates at time s_1, the RTP, which generalizes (2) to time-varying covariates $\mathbf{X}(t)$, is

$$\begin{aligned}(4)\quad &\mathrm{RTP}_A\big(s_1, s_2\big) \\ &= P\Big\{Y(s_2) \in A\big[\mathbf{X}(s_2), s_2\big] \Big| Y(s_1) \in A\big[\mathbf{X}(s_1), s_1\big]\Big\}.\end{aligned}$$

Other special cases, such as $A[\mathbf{X}(t), t] \equiv A(t)$, may be derived from (3) and (4). Although the choices of $A(\cdot, \cdot)$ depend on the specific scientific objectives, it is common in biomedical studies to define the health status at time t based on the conditional quantiles of $Y(t)$ given $\mathbf{X}(t)$, so that $A[\mathbf{X}(t), t]$ may be specified as

$$(5)\qquad A_\alpha\big[\mathbf{X}(t), t\big] = \Big\{Y(t) : Y(t) > q_\alpha\big[t, \mathbf{X}(t)\big]\Big\},$$

where $q_\alpha[t, \mathbf{X}(t)]$ is the $(100 \times \alpha)$th quantile of $Y(t)$ given $\mathbf{X}(t)$.

2.3 Rank-tracking probability ratios

The strength of "rank-tracking ability" is measured by comparing $\mathrm{RTP}_A[s_1, s_2; \mathcal{X}(s_1)]$ with $P_A[s_2; \mathcal{X}(s_1)] = P\{Y(s_2) \in A[\mathbf{X}(s_2), s_2] | \mathbf{X}(s_1) \in \mathcal{X}(s_1)\}$, and $\mathrm{RTP}_A(s_1, s_2)$ with $P_A(s_2) = P\{Y(s_2) \in A[\mathbf{X}(s_2), s_2]\}$. Thus, it is convenient to measure the "rank-tracking abilities" of $Y(t)$ by the RTP-Ratios (RTPRs),

$$\begin{aligned}(6)\quad &\mathrm{RTPR}_A\big[s_1, s_2; \mathcal{X}(s_1)\big] \\ &= \mathrm{RTP}_A\big[s_1, s_2; \mathcal{X}(s_1)\big] / P_A\big[s_2; \mathcal{X}(s_1)\big]\end{aligned}$$

where $P_A[s_2; \mathcal{X}(s_1)] = P\{Y(s_2) \in A[\mathbf{X}(s_2), s_2] | \mathbf{X}(s_1) \in \mathcal{X}(s_1)\}$, and

$$(7)\qquad \mathrm{RTPR}_A(s_1, s_2) = \mathrm{RTP}_A\big(s_1, s_2\big) / P_A(s_2),$$

respectively. For the RTP of (3), the strength of "rank-tracking ability" is measured by comparing $\mathrm{RTPR}_A[s_1, s_2; \mathcal{X}(s_1)]$ with 1. If $\mathrm{RTPR}_A[s_1, s_2; \mathcal{X}(s_1)] = 1$, $Y(s_1)$ has no "tracking ability" for $Y(s_2)$. If $\mathrm{RTPR}_A[s_1, s_2; \mathcal{X}(s_1)] < 1$, $Y(s_1)$ has "negative tracking ability" for $Y(s_2)$. The strength of "positive tracking ability" is then determined by how much $\mathrm{RTPR}_A[s_1, s_2; \mathcal{X}(s_1)]$ is larger than 1. Both the RTPs and RTPRs are useful tools for identifying the risk factors with strong "rank-tracking abilities" in longitudinal studies, which may be used to develop new guidelines and intervention strategies for early disease prevention.

3. METHODS OF ESTIMATION AND INFERENCES

We establish a class of smoothing methods based on B-spline approximations for the estimation and inferences of the RTPs in (3) and (4) and the RTPRs in (6) and (7). When the covariates are discrete and time-invariant, nonparametric estimation based on kernel smoothing has been studied by Wu and Tian (2013a). For the more general situations that involve continuous and time-varying covariates, the kernel methods may be computationally infeasible because of the well-known "curse of dimensionality". Our proposed estimation methods have the advantage of incorporating multiple continuous and time-dependent covariates.

3.1 Nonparametric mixed-effects models

As a natural extension of the linear mixed-effects models, global smoothing through basis approximations is a popular approach in nonparametric longitudinal analysis. For the simple case of evaluating $\{Y(t), t; t \in \mathcal{T}\}$ without covariates, Shi et al. (1996) and Rice and Wu (2001) suggested to model $Y_i(t)$ at time t by the nonparametric mixed-effects model,

$$(8)\qquad Y_i(t) = \mu(t) + \zeta_i(t) + \epsilon_i(t),$$

where $\mu(t)$ is the mean curve of $Y_i(t)$, $\zeta_i(t)$ is the random departure from $\mu(t)$ for the ith subject with $E[\zeta_i(t)] = 0$, and $\epsilon_i(t)$ are the mean zero measurement errors.

When a set of covariates $\mathbf{X}(t)$ is incorporated, Liang et al. (2003) proposed a class of multivariate extensions of (8). For simplicity, we illustrate the case of a single covariate $X(t)$ with $P = 1$. The case of multivariate covariates can be extended analogously. Let $\{Y_i(t), X_i(t)\}$ be the outcome and covariate of the ith subject at time t, $\{\beta_0(t), \beta_1(t)\}$ be two smooth functions of t, $\beta_{0i}(t) = \beta_0(t) + \gamma_{0i}(t)$ and $\beta_{1i}(t) = \beta_1(t) + \gamma_{1i}(t)$, $\{\gamma_{0i}(t), \gamma_{1i}(t)\}$ be the mean zero stochastic processes that represent the individual random deviations. The mixed-effects varying-coefficient model of Liang et al. (2003) is

$$(9)\qquad Y_i(t_{ij}) = \beta_{0i}(t_{ij}) + X_i(t_{ij})\,\beta_{1i}(t_{ij}) + \epsilon_i(t_{ij}),$$

where $\epsilon_i(t)$ are mean zero measurement error processes with $\rho_\epsilon(s, t) = \mathrm{cov}\{\epsilon_i(s), \epsilon_i(t)\}$, $\epsilon_i(t)$ and $\{\gamma_{0i}(t), \gamma_{1i}(t)\}$ are mutually independent for given i and $\{\epsilon_i(t), \gamma_{0i}(t), \gamma_{1i}(t)\}$ and $\{\epsilon_k(t), \gamma_{0k}(t), \gamma_{1k}(t)\}$ are independent for $i \neq k$.

When the model structure of (9) does not hold, Zhou et al. (2008) suggested a joint model framework for predicting the multivariate subject-specific trajectories of correlated time-dependent curves. The approach of Zhou

et al. (2008) does not need the distinction between outcome and covariate variables, but requires the specification of the joint model correlation structures of the time-dependent curves.

3.2 B-spline estimation and prediction of trajectories

Under the data structure of Section 2.1, e.g., the NGHS design, the correlation structures of the data are completely unknown, and our objective is to estimate the RTPs and RTPRs nonparametrically. The first step is to estimate the coefficient curves and predict the outcome trajectories based on the mixed-effects models of (8) and (9) with B-spline basis approximations. Other smoothing methods, such as smoothing splines and penalized splines, may also be applied for curve estimation and outcome trajectory prediction. We focus on B-splines because of their good numerical properties and simplicity in practical implementation.

When B-splines are used for (9), we have the approximations $\beta_0(t) \approx b_0(t)^T\xi_0$, $\beta_1(t) \approx b_1(t)^T\xi_1$, $\gamma_{0i}(t) \approx b_0(t)^T\eta_i$, and $\gamma_{1i}(t) \approx b_1(t)^T\phi_i$ based on the B-spline basis functions $b_0(t) = (b_{01}(t), \ldots, b_{0m}(t))^T$ and $b_1(t) = (b_{11}(t), \ldots, b_{1q}(t))^T$ for some integers $m, q > 0$. Here $\xi_0 = (\xi_{01}, \ldots, \xi_{0m})^T$ and $\xi_1 = (\xi_{11}, \ldots, \xi_{1q})^T$ are the vectors of coefficients for the fixed-effects components, $\eta_i = (\eta_{i1}, \ldots, \eta_{im})^T$ and $\phi_i = (\phi_{i1}, \ldots, \phi_{iq})^T$ are the vectors of coefficients for the subject-specific normal random components with mean zero and covariance matrices Γ and Φ. If we denote by Y_i and X_i the column vectors consisting of the observed $Y_i(\mathbf{t_i})$ and $X_i(\mathbf{t_i})$ values at the time points $\mathbf{t_i} = (t_{i1}, \ldots, t_{in_i})$, B_{0i} and B_{1i} the corresponding $n_i \times m$ and $n_i \times q$ spline basis matrices, and ϵ_i the measurement errors evaluated at these time points, respectively, the B-spline approximation for (9) is

$$Y_i = B_{0i}(\xi_0 + \eta_i) + B_{1i}(\xi_1 + \phi_i) * X_i + \epsilon_i, \tag{10}$$

where "$*$" denotes a component-wise product. The B-spline approximation for (8) is simply $Y_i = B_{0i}(\xi_0 + \eta_i) + \epsilon_i$.

For a given distribution function of ϵ_i, such as $\epsilon_i \sim N(0, \Sigma)$, the maximum likelihood estimators (MLEs) or the restricted MLEs of $\{\xi_0, \xi_1, \Sigma, \Gamma, \Phi\}$ are $\{\widehat{\xi_0}, \widehat{\xi_1}, \widehat{\Sigma}, \widehat{\Gamma}, \widehat{\Phi}\}$, and the best linear unbiased predictors (BLUPs) of the random effects $\widehat{\eta_i}$ and $\widehat{\phi_i}$ can be computed by the EM algorithm as described in Liang et al. (2003). By plugging in the coefficient estimates, the B-spline predicted outcome trajectory curve for the ith subject at any time point t can be computed by

$$\widehat{Y}_i(t) = b_0(t)^T\big(\widehat{\xi_0} + \widehat{\eta_i}\big) + b_1(t)^T\big(\widehat{\xi_1} + \widehat{\phi_i}\big)X_i(t). \tag{11}$$

For the estimation and prediction of (8), only ξ_0 and η_i need to be estimated, so that the predicted subject-specific outcome trajectory curve is

$$\widehat{Y}_i^*(t) = b_0(t)^T\big(\widehat{\xi_0} + \widehat{\eta_i}\big). \tag{12}$$

When the time-varying covariate $X_i(t)$ are measured with errors, we can approximate $X_i(t)$ by a smoothing function with a random component as in model (8),

$$X_i(t) = \mu_x(t) + \zeta_{xi}(t) + u_i(t), \tag{13}$$

and use similar B-spline basis approximations to compute $\widehat{\mu}_x(t)$, the estimator of the mean curve $\mu_x(t)$, and $\widehat{\zeta}_{xi}(t)$, the predicted subject-specific curve $\zeta_{xi}(t)$. The predicted subject-specific trajectory curve $\widehat{X}_i(t)$ for $X_i(t)$ can be obtained by setting $u_i(t) = 0$ and substituting $\{\mu_x(t), \zeta_{xi}(t)\}$ of (13) with $\{\widehat{\mu}_x(t), \widehat{\zeta}_{xi}(t)\}$.

Remark 3.1. The predicted subject-specific trajectory curves $\widehat{X}_i(t)$ can be used in the RTP and RTPR estimation when $X_i(t)$ are expected to have measurement errors. When the mixed-effects varying-coefficient model (9) is satisfied and $X_i(t)$ are measured with errors, Liang et al. (2003) suggests replacing the observed $X_i(t)$ in model (9) with the predicted curves $\widehat{X}_i(t)$ to correct the measurement errors in $X_i(t)$. Thus in the first approach, we use $\{\widehat{Y}_i(t), \widehat{X}_i(t);\ i = 1, \ldots, n\}$ from models (9) and (13) to estimate the RTPs and RTPRs. Note that the predicted trajectory curves $\widehat{Y}_i(t)$ from (9) depend on the linear relationship between $Y_i(t)$ and $X_i(t)$ at each time point t, so that it may not be appropriate to estimate the RTPs and RTPRs based on $\widehat{Y}_i(t)$ when the time-varying model (9) does not necessarily hold. Thus as an alternative approach, for the more general case that (9) is not satisfied and the actual relationship between $Y_i(t)$ and $X_i(t)$ is unknown, a plausible unstructured nonparametric approach is to first compute the trajectories $\widehat{Y}_i^*(t)$ and $\widehat{X}_i(t)$ from the mixed-effects models (8) and (13), respectively, and then estimate the RTPs and RTPRs based on the available $\{\widehat{Y}_i^*(t), \widehat{X}_i(t);\ i = 1, \ldots, n\}$. For the cases with more than one covariate $X_i(t)$, i.e., $P > 1$ in $\mathbf{X}_i(t)$, each continuous time-varying component of the $\mathbf{X}_i(t)$ may be predicted by model (13) separately. For multivariate covariates in models (9) and (10), the same estimation methods can be applied by a simple modification of the design matrix.

3.3 Estimation with predicted outcome trajectories

We present the estimation and inferences of the RTPs and RTPRs based on the predicted trajectory curves obtained under three scenarios: (a) the mixed-effects model of (8) without covariates, (b) the mixed-effects varying-coefficients model of (9), and (c) the combined unstructured mixed-effects models of (8) and (13). Each scenario is a special case of the later one. Because different predicted trajectory curves are used under each of these scenarios, the estimation methods for the RTPs and RTPRs are different.

3.3.1 Estimation without covariates

We first consider the estimation of the RTPs and RTPRs based on the observations $\{Y_i(t_{ij});\ j = 1, \ldots, n_i,\ i =$

$1, \ldots, n\}$ without covariates. By (1), the RTP is

$$\text{RTP}_A(s_1, s_2) = \frac{E\{1_{[Y(s_2)\in A(s_2), Y(s_1)\in A(s_1)]}\}}{E\{1_{[Y(s_1)\in A(s_1)]}\}}, \tag{14}$$

where $1_{[\cdot]}$ is the indicator function and $A(t)$ is a prespecified and known risk set at time t, such as $A_\alpha(t) = \{Y(t) : Y(t) > q_\alpha(t)\}$ with $q_\alpha(t)$ being the known $(100\times\alpha)$th quantile of $Y(t)$. The $\text{RTPR}_A(s_1, s_2)$ of (6) is given by dividing $\text{RTP}_A(s_1, s_2)$ of (14) by $E\{1_{[Y(s_2)\in A(s_2)]}\}$. The estimator of $\text{RTP}_A(s_1, s_2)$ based on the predicted curves of (12) is

$$\widehat{\text{RTP}}_A(s_1, s_2) = \frac{\sum_{i=1}^n 1_{[\widetilde{Y}_i^*(s_2)\in A(s_2), \widetilde{Y}_i^*(s_1)\in A(s_1)]}}{\sum_{i=1}^n 1_{[\widetilde{Y}_i^*(s_1)\in A(s_1)]}}, \tag{15}$$

and the estimator of $\text{RTPR}_A(s_1, s_2)$ is

$$\widehat{\text{RTPR}}_A(s_1, s_2) = \frac{\widehat{\text{RTP}}_A(s_1, s_2)}{(1/n)\sum_{i=1}^n 1_{[\widetilde{Y}_i^*(s_2)\in A(s_2)]}}. \tag{16}$$

where $\widetilde{Y}_i^*(t) = \widehat{Y}_i^*(t) + \widehat{\epsilon}_i(t)$, $\widehat{Y}_i^*(t)$ is the B-spline predicted trajectory curves given in (12) and $\widehat{\epsilon}_i(t)$ is the estimated errors.

In some cases, $A(\cdot)$ is not known and has to be estimated from the same sample that is used to estimate the RTPs and RTPRs. For example, the $(100 \times \alpha)$th quantile $q_\alpha(t)$ used in $A_\alpha(t)$ may not be known for a given study population and has to be estimated from the predicted trajectories. We describe in Section 3.3.4 a split sample approach for dealing with such situations.

Remark 3.2. Because $\widehat{Y}_i^*(t)$ is the subject-specific mean curve for the ith subject at t, we need to use $\widetilde{Y}_i^*(t)$, which includes the random measurement errors of the subject, in the estimators (15) and (16), instead of using $\widehat{Y}_i^*(t)$ alone. Here, the estimated errors $\widehat{\epsilon}_i(t)$ may be computed from the fitted model residuals or from the maximum likelihood estimators of $\epsilon_i(t)$ in model (8). If the estimated measurement errors $\widehat{\epsilon}_i(t)$ were ignored and $\widetilde{Y}_i^*(t)$ in (15) and (16) were replaced by $\widehat{Y}_i^*(t)$, the estimators of $\text{RTP}_A(s_1, s_2)$ and $\text{RTPR}_A(s_1, s_2)$ could be biased. The potential biases of replacing $\widetilde{Y}_i^*(t)$ with $\widehat{Y}_i^*(t)$ in (15) and (16) can be seen from a simulation. The asymptotic results of the estimators warrant further research.

3.3.2 Estimation with mixed-effects varying-coefficient models

When the model (9) is satisfied with a time-varying covariate vector $\mathbf{X}(t)$, we can compute the B-spline predicted subject-specific trajectory curves $\widehat{Y}_i(t)$ and $\widehat{\mathbf{X}}_i(t)$ from (11) and (13), respectively. For a given time-varying subset $\mathcal{X}(s_1)$ at time s_1, we estimate the RTPs of (3) and (4) by

$$\widehat{\text{RTP}}_A\big[s_1, s_2; \mathcal{X}(s_1)\big] = \frac{\sum_{i=1}^n 1_{\{\widetilde{Y}_i(s_2)\in A[\widehat{\mathbf{X}}_i(s_2), s_2], \widetilde{Y}_i(s_1)\in A[\widehat{\mathbf{X}}_i(s_1), s_1], \widehat{\mathbf{X}}_i(s_1)\in\mathcal{X}(s_1)\}}}{\sum_{i=1}^n 1_{\{\widetilde{Y}_i(s_1)\in A[\widehat{\mathbf{X}}_i(s_1), s_1], \widehat{\mathbf{X}}_i(s_1)\in\mathcal{X}(s_1)\}}} \tag{17}$$

and

$$\widehat{\text{RTP}}_A(s_1, s_2) = \frac{\sum_{i=1}^n 1_{\{\widetilde{Y}_i(s_2)\in A[\widehat{\mathbf{X}}_i(s_2), s_2], \widetilde{Y}_i(s_1)\in A[\widehat{\mathbf{X}}_i(s_1), s_1]\}}}{\sum_{i=1}^n 1_{\{\widetilde{Y}_i(s_1)\in A[\widehat{\mathbf{X}}_i(s_1), s_1]\}}}, \tag{18}$$

where $\widetilde{Y}_i(t) = \widehat{Y}_i(t) + \widehat{\epsilon}_i(t)$ and $\widehat{\epsilon}_i(t)$ is the maximum likelihood estimator of $\epsilon_i(t)$ or the estimated error computed from the fitted model residuals. Based on (17) and (18), the estimators of the corresponding RTPRs are given by

$$\widehat{\text{RTPR}}_A\big[s_1, s_2; \mathcal{X}(s_1)\big] = \frac{\widehat{\text{RTP}}_A\big\{s_1, s_2; \mathcal{X}(s_1)\big\}}{(1/n_{\mathcal{X}})\sum_{i=1}^n 1_{\{\widetilde{Y}_i(s_2)\in A[\widehat{\mathbf{X}}_i(s_2), s_2], \widehat{\mathbf{X}}_i(s_1)\in\mathcal{X}(s_1)\}}}, \tag{19}$$

where $n_{\mathcal{X}} = \sum_{i=1}^n 1_{[\widehat{\mathbf{X}}_i(s_1)\in\mathcal{X}(s_1)]}$, and

$$\widehat{\text{RTPR}}_A(s_1, s_2) = \frac{\widehat{\text{RTP}}_A(s_1, s_2)}{(1/n)\sum_{i=1}^n 1_{\{\widetilde{Y}_i(s_2)\in A[\widehat{\mathbf{X}}_i(s_2), s_2]\}}}. \tag{20}$$

Remark 3.3. The predicted covariate trajectories $\widehat{\mathbf{X}}_i(t)$ are used in (17) through (20) for the cases that $\mathbf{X}_i(t_{ij})$ are expected to have measurement errors. When $\mathbf{X}_i(t_{ij})$ are not expected to have measurement errors or $\mathbf{X}_i(t_{ij})$ are time – invariant baseline covariates, the observed $\mathbf{X}_i(t_{ij})$ are used. In addition to estimators in (17) through (20), which rely on the predicted trajectory curves $\{\widetilde{Y}_i(s_1), \widetilde{Y}_i(s_2)\}$, an alternative method for the estimation of the RTPs and RTPRs based on the model (9) is to first estimate the mean, variance and covariance curves of (9) using the procedures described in Liang et al. (2003), and then estimate the RTPs and RTPRs based on the estimated mean, variance and covariance curves and the assumptions that the random effects and the measurement errors of (9) and (13) have normal distributions. This estimation approach gives similar results to our estimators of (17) through (20) in our simulation study when the normal distribution assumptions on (9) hold.

3.3.3 Estimation with unstructured mixed-effects models

For the general situations that the time-varying linear structure of (9) may not hold, we approximate the subject-specific curves of $Y_i(t)$ and $\mathbf{X}_i(t)$ using B-splines on (8) and (13), respectively. When $\mathbf{X}_i(t)$ is multivariate with $P > 1$, (13) is used to approximate each component of $\mathbf{X}_i(t)$. Based on the predicted trajectory curves $\{\widehat{Y}_i^*(t), \widehat{\mathbf{X}}_i(t);\ i = 1, \ldots, n\}$, we obtain $\{\widetilde{Y}_i^*(t);\ i = 1, \ldots, n\}$ as in (16), and

compute the estimators $\widehat{\text{RTP}}^*_A[s_1, s_2; \mathcal{X}(s_1)]$, $\widehat{\text{RTP}}^*_A(s_1, s_2)$, $\widehat{\text{RTPR}}^*_A[s_1, s_2; \mathcal{X}(s_1)]$ and $\widehat{\text{RTPR}}^*_A(s_1, s_2)$ by substituting $\{\widetilde{Y}_i(s_1), \widetilde{Y}_i(s_2)\}$ with $\{\widetilde{Y}^*_i(s_1), \widetilde{Y}^*_i(s_2)\}$ in (17), (18), (19) and (20), respectively. As discussed in Remark 3.1, the advantage of using the trajectory curves $\{\widetilde{Y}^*_i(t), \widehat{X}_i(t);\ t \in \mathcal{T}\}$ is that the RTP and RTPR estimators can still be computed as long as the B-spline approximations for (8) and (13) hold, while the relationship between $Y_i(t)$ and $\mathbf{X}_i(t)$ is unknown.

3.3.4 A split sample approach for the estimation of $A(\cdot, \cdot)$

When $A[\mathbf{X}(t), t]$ is unknown but can be estimated from the available sample, a practical approach is to randomly split the data into sub-samples, so that one sub-sample can be used to estimate $A[\mathbf{X}(t), t]$, while the other sub-sample can be use to estimate the RTPs and RTPRs using the estimator $\widehat{A}[\mathbf{X}(t), t]$ of $A[\mathbf{X}(t), t]$. Specifically, we randomly split the sample into sub-samples I_1 and I_2 with corresponding sample sizes n_{I_1} and n_{I_2}, such that $n_{I_1} + n_{I_2} = n$. The first sub-sample I_1 is used to estimate $A[\mathbf{X}(t), t]$, and then the RTPs and RTPRs can be estimated using the second sub-sample with the methods of Section 3.3.1 through Section 3.3.3 and $\widehat{A}[\mathbf{X}(t), t]$ in place of $A[\mathbf{X}(t), t]$. The estimators $\widehat{A}[\mathbf{X}(t), t]$ depend on the specific definitions of $A[\mathbf{X}(t), t]$ and have to be constructed on a case-by-case basis. For example, for $A_\alpha[\mathbf{X}(t), t]$ defined in (5), $\widehat{A}_\alpha[\mathbf{X}(t), t]$ can be obtained by using the estimated conditional quantile $\widehat{q}_\alpha[t, \mathbf{X}(t)]$ from the first sub-sample.

3.4 Bootstrap pointwise confidence intervals

Asymptotically approximated inferences for the estimators developed in Sections 3.3 are still unavailable because the asymptotic distributions of these estimators have not yet been explicitly derived. As a practical approach, we can use the "resampling-subject" bootstrap which has been commonly used in longitudinal analysis (e.g., Hoover et al., 1998). In this approach, we obtain B bootstrap samples by resampling the subjects with replacement one at a time, and compute the corresponding estimates within each bootstrap sample. The average and the lower and upper $[100 \times (\alpha/2)]$th percentiles of the B bootstrap estimates are obtained as our bootstrap estimate and the $[100 \times (1-\alpha)]$th bootstrap confidence intervals. Alternatively, we can also compute the sample standard deviations (SD) of the estimates from the bootstrap samples, and approximate the $[100 \times (1 - \alpha)]\%$ confidence intervals by the "estimate $\pm$ $z_{\alpha/2} \times SD$" error bands.

4. APPLICATION TO NGHS DATA

4.1 Brief background

The NGHS is a multi-center population-based cohort study, which enrolled Caucasian and African-American girls at 9 or 10 years of age and measured their height, weight, BP and other cardiovascular risk factors annually for up to 10 visits. Among the 2,379 girls enrolled in the NGHS, there were 1,166 Caucasians and 1,213 African-Americans. The number of observed follow-up visits varied from 1 to 10, and had median 10, mean 8.8 and standard deviation 2.0. The study design was described in NGHSRG (1992), and the main findings were reported in Thompson et al. (2007).

Based on NHBPEP (2004) and Obarzanek et al. (2010), a lower bound of the age- and sex-adjusted 85th percentile of the Centers for Disease Control and Prevention (CDC) BMI growth chart is used to define overweight and obese girls at a given age. The age-, sex-, and height-specific conditional percentiles for the systolic blood pressure (SBP) and diastolic blood pressure (DBP) are used to define prehypertension, and stage 1 and stage 2 hypertension in children. Due to the limitation of the statistical methodology, existing publications on the temporal trends of BMI and BP among children and adolescents, such as Thompson et al. (2007) and Obarzanek et al. (2010), have not systematically investigated the question of how to quantitatively measure the "rank-tracking abilities" for BMI and BP over age.

4.2 Rank-tracking for BMI

We first estimate the RTPs and RTPRs of BMI based on the NGHS subjects with race taken as a binary covariate, so that $X_i = 0$ if the ith girl is Caucasian, and $X_i = 1$ if she is African-American. We consider the set $A_{0.85}(t)$ to be the set of subjects whose BMI values at t years old are greater than $q_{0.85}(t)$, the 85th percentile of BMI for girls at age t years as defined in the CDC BMI growth chart. After fitting the nonparametric mixed-effect model (8) separately for the Caucasian and African-American girls using the BMI observations over age based on cubic B-spline approximations with four and three equally spaced interior knots selected from BIC, respectively, we computed the subject-specific BMI trajectory curves over age, and estimated $\text{RTP}_A(s_1, s_1 + \delta; x)$ and $\text{RTPR}_A(s_1, s_1 + \delta; x)$ of BMI for Caucasian ($x = 0$) and African-American ($x = 1$) girls over the age range $9 \le s_1 \le 17$, $\delta = 2$ and $\delta = 4$.

Figure 1 shows the estimated 2-year and 4-year RTPs and RTPRs curves for both Caucasian and African-American girls with their 95% pointwise bootstrap percentile confidence intervals based on $B = 500$ bootstrap replications. Figure 1 (a–b) and (e–f) show that the conditional probabilities of being overweight or obese are 78% to 94% and 75% to 91% for those girls who were already overweight or obese 2 years and 4 years earlier, respectively. Since the NGHS participants are slightly more overweight than the general population of the CDC growth chart, we examined the relative strength of the BMI tracking ability through the RTPR curves displayed in Figure 1 (c–d) and (g–h). Our spline smoothing estimation results are consistent with the results based on the kernel estimation procedures of Wu and Tian (2013a), suggesting that BMI has high rank-tracking ability for adolescent girls in both ethnic groups.

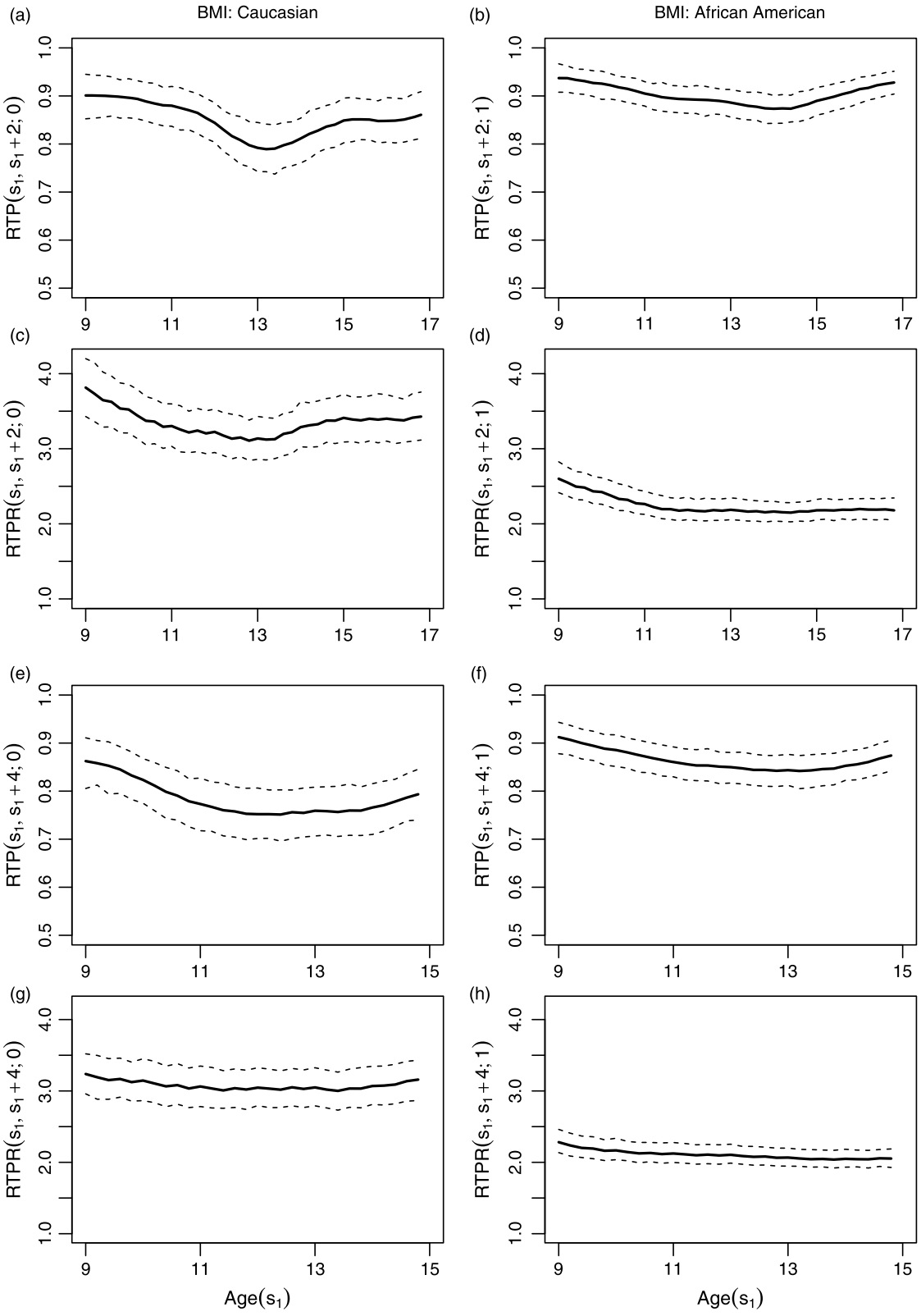

Figure 1. The estimated BMI $RTP_{A_{0.85}}(s_1, s_1+2; x)$, $RTPR_{A_{0.85}}(s_1, s_1+2; x)$, $RTP_{A_{0.85}}(s_1, s_1+4; x)$, $RTPR_{A_{0.85}}(s_1, s_1+4; x)$ *and their 95% bootstrap percentile confidence intervals for Caucasian* $(x = 0)$ *and African-American* $(x = 1)$ *girls.*

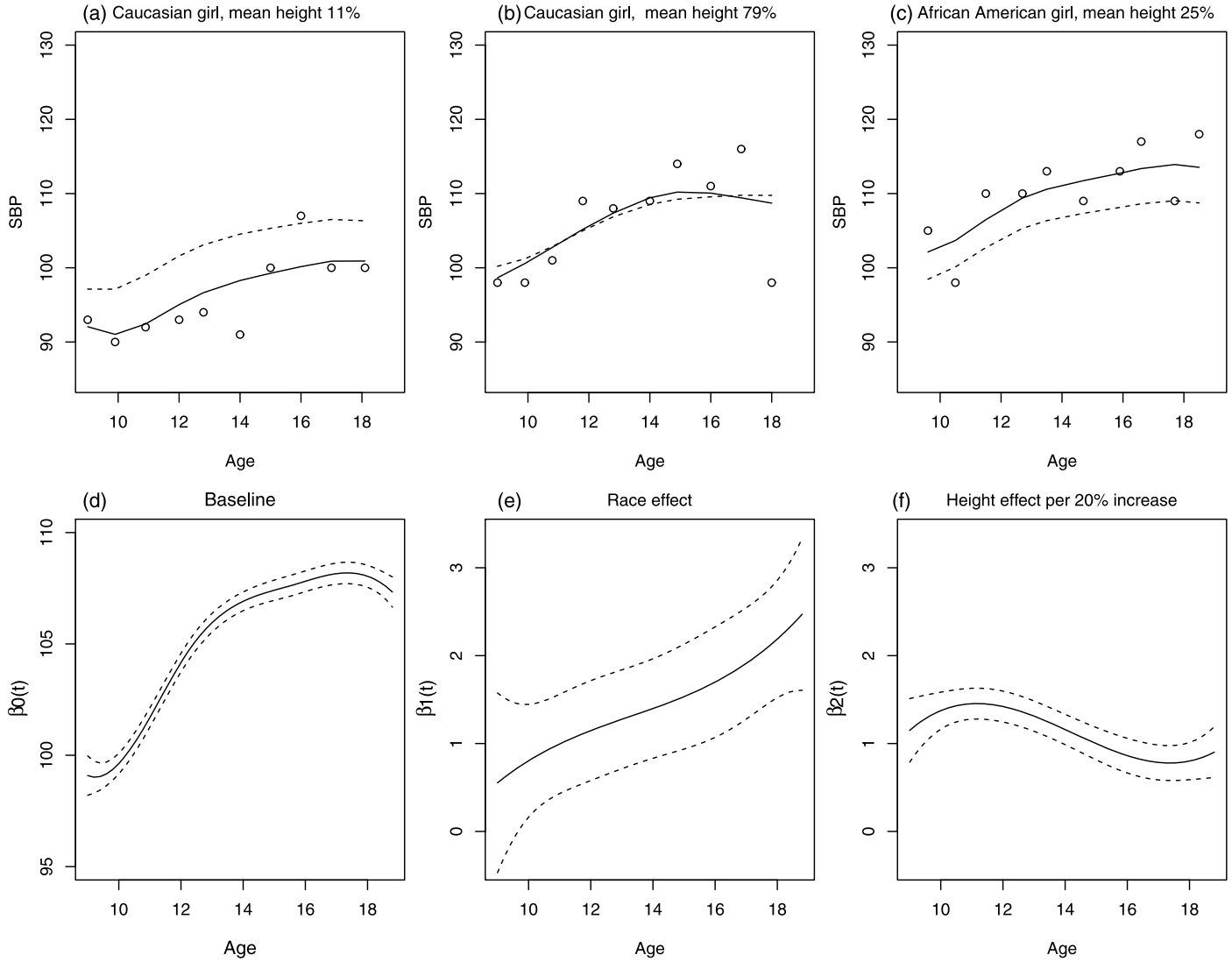

Figure 2. Upper panels (a)–(c): The longitudinal SBP measurements for three girls from NGHS with predicted subject-specific curves and mean population curves plotted in solid and dash lines. Lower panels (d)–(f): The mean baseline curve $\beta_0(t)$, and two coefficient curves for race and height percentile $\beta_1(t)$ and $\beta_2(t)$ with 95% confidence interval based on model (9) with the cubic B-spline basis approximations.

4.3 Rank-tracking for SBP

The BP levels have been shown to depend on the girl's race and height percentile (Daniels et al. 1998, Wu et al. 2010). We can estimate the RTPs and RTPRs of SBP based on the predicted SBP trajectory curves obtained either from the model (9) or the separate univariate mixed-effects models (8) and (13) with the NGHS SBP data $\mathcal{D} = \{Y_i(t_{ij}), X_{1i}, X_{2i}(t_{ij}); j = 1, \ldots, n_i, i = 1, \ldots, n\}$. Here $Y_i(t)$, X_{1i} and $X_{2i}(t)$ are ith girl's SBP (in mmHg), race and age-adjusted height percentile at t years of age, with $X_{1i} = 0$ if the girl is Caucasian and $X_{1i} = 1$ if she is African-American. The random variables corresponding to $\mathcal{D}$ at time t are $\{Y(t), X_1, X_2(t)\}$. Let $q_{0.9}[t, X_2(t)]$ be the conditional 90th percentile of SBP for girls given that they are t years old and have height percentile $X_2(t)$. Our objective is to estimate the 2-year "rank-tracking ability" of SBP, $\mathrm{RTP}_A(s_1, s_1 + 2; x_1)$ and $\mathrm{RTPR}_A(s_1, s_1 + 2; x_1)$, for Caucasian ($x_1 = 0$) and African-American ($x_1 = 1$) girls based on $A_{0.9}[X_2(t), t]$ defined in (5). Here $A_{0.9}[X_2(t), t]$ represents the girls whose SBP values are above the conditional 90th percentile for the given age and height percentile.

Using the framework of (9), the mixed-effects varying-coefficient model for $\mathcal{D}$ is

(21)
$$Y_i(t_{ij}) = \beta_{0i}(t_{ij}) + X_{1i}\beta_{1i}(t_{ij}) + X_{2i}(t_{ij})\beta_{2i}(t_{ij}) + \epsilon_i(t_{ij}),$$

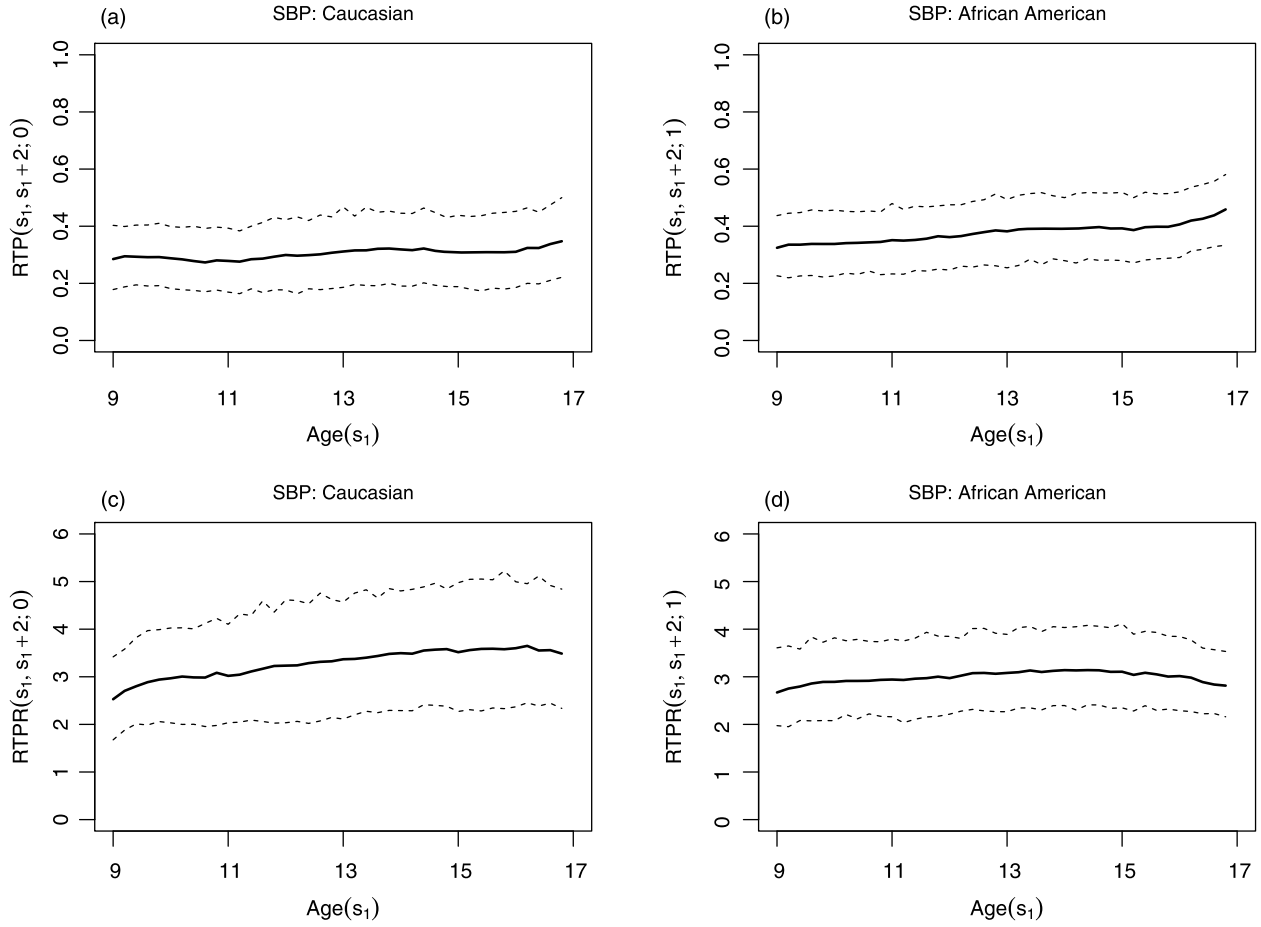

Figure 3. The estimated SBP $RTP_{A_{0.9}}(s_1, s_1+2; x_1)$ *and* $RTPR_{A_{0.9}}(s_1, s_1+2; x_1)$*, and their 95% bootstrap percentile confidence intervals for Caucasian* ($x_1 = 0$) *and African-American* ($x_1 = 1$) *girls.*

where, for $l = 0, 1, 2$, $\beta_{li}(t) = \beta_l(t) + \gamma_{li}(t)$ with mean curves $\beta_l(t)$ and mean zero stochastic processes $\gamma_{li}(t)$, and $\epsilon_i(t)$ are independent normal random variables with mean zero and variance σ^2. Figure 2 (a–c) show the raw data of longitudinal SBP measurements for three randomly selected subjects, and the mean curves and predicted subject-specific SBP trajectories based on (21). These scatter plots and trajectory curves suggest that the model (21) provides a reasonable fit to the data. Figure 2 (d–f) show the mean baseline curve $\beta_0(t)$, and the two mean coefficient curves for race and height percentile $\beta_1(t)$ and $\beta_2(t)$, respectively, estimated by cubic B-spline basis approximation with the numbers of knots chosen from BIC. These curves suggest that the mean SBP increases with age, and the SBP measurements also depend on race and height percentile with both coefficient curves being positive and varying with age. The African-American girls tend to have higher SBP values than the Caucasian girls, and the mean SBP differences in race increase with age. The effect of height percentiles on SBP tapers off at older age.

Since the conditional 90th percentile $q_{0.9}[t, X_2(t)]$ is not known, we randomly split the subjects into two sub-samples with approximately equal sample sizes, and computed the estimates of $q_{0.9}[t, X_2(t)]$ using the first sub-sample. We then used the cubic B-spline predicted SBP trajectories values to compute the estimators $\widehat{\mathrm{RTP}}_A(s_1, s_1+2; x_1)$ and $\widehat{\mathrm{RTPR}}_A(s_1, s_1+2; x_1)$ of the 2-year RTPs and RTPRs for SBP at ages s_1 and s_1+2 for both Caucasian ($x_1 = 0$) and African-American ($x_1 = 1$) girls. Figure 3 shows the corresponding RTPs and RTPRs estimates and their 95% pointwise bootstrap percentile confidence intervals obtained with $B = 500$ bootstrap replications. The RTPs in Figure 3 (a–b) are between approximately 30% and 40% for both Caucasian and African-American girls across $s_1 \in [9, 17]$ years, and these values are significantly larger than 10%, which is approximately the conditional probability of SBP greater than $q_{0.9}[t, X_2(t)]$ at $t = s_1+2$ given race only. To evaluate the relative strength of the "rank-tracking ability" of SBP for these girls, we compare the estimates of $\mathrm{RTPR}_A(s_1, s_1+2; x_1)$ in Figure 3 (c–d) for Caucasian and African-American girls, and the estimated RTPRs are approximately between 2.6 and 3.6 for Caucasian girls and between 2.9 and 3.1 for African-American girls. The 95% confidence intervals for the RTPRs are greater than 1, suggesting that SBP has high

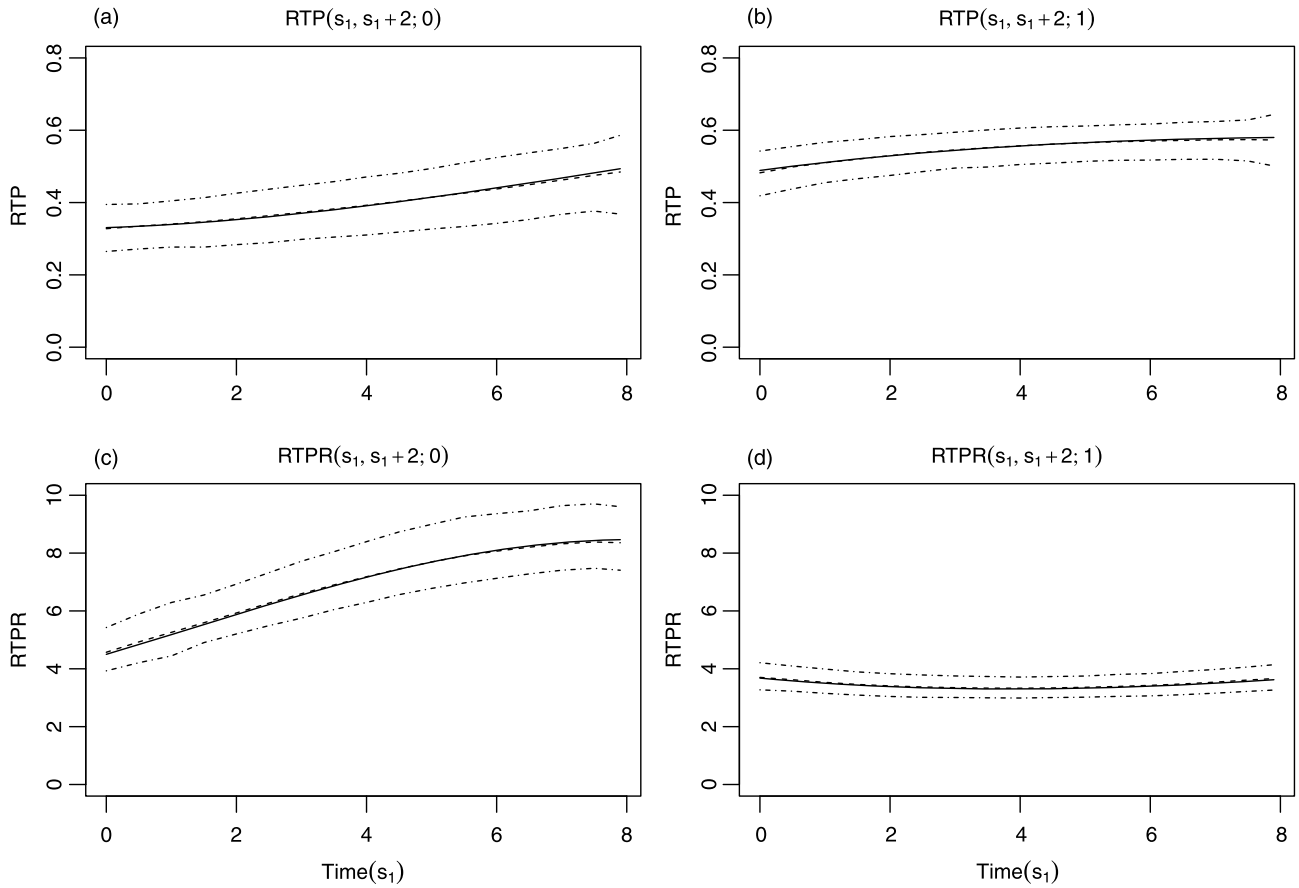

Figure 4. The solid lines are true curves of $RTP_{A_{0.9}}(s_1, s_1+2; x_1)$ *in (a) and (b) and* $RTPR_{A_{0.9}}(s_1, s_1+2; x_1)$ *in (c) and (d) for* $x_1 = 0$ *and* $x_1 = 1$. *The middle dashed lines are the averages of the estimated curves, and the dotted dash lines are the pointwise lower and upper 2.5% percentiles of the estimated curves at* $s_1 \in [0, 8]$ *from 1,000 simulations.*

"positive tracking ability" for both Caucasian and African-American girls within this age range. We also estimate RTPs and RTPRs based on the separate univariate mixed-effects models (8) and (13) and the procedure in Section 3.3.3, the estimate curves are very similar to the results presented in Figure 3.

5. SIMULATION

We consider a simulation study that mimic the NGHS data structure. The simulated samples are generated based on the model $Y_i(t) = \beta_0(t) + \gamma_{0i}(t) + X_{1i}(t)\beta_1(t) + X_{2i}(t)\beta_2(t) + \epsilon_i(t)$ for $t \in [0, 10]$. Each simulated sample has $n = 1,000$ subjects, and each subject has 10 visits. Within each sample, we generate the ith subject's visiting time of the jth visit t_{ij} from the uniform distribution $U[(j-1), j]$ for $j = 1, \ldots, 10$, so that, $t_{i1} \sim U[0, 1], \ldots, t_{i10} \sim U[9, 10]$. Given each t_{ij}, $X_{1i}(t_{ij}) = X_{1i}$ is a binary time-invariant covariate with the Bernoulli(0.5) distribution, and $X_{2i}(t_{ij})$ is a time-varying covariate generated from the discrete uniform distribution on $\{1, 2, \ldots, 100\}$. Similar to the patterns of the coefficient curves in Figure 2, we choose $\beta_0(t) = 100 - 0.1t^2 + 1.9t$, $\beta_1(t) = 0.65t - 0.04t^2$, and $\beta_2(t) = 0.02 + 0.02\cos(0.11\pi t)$. The random coefficient is $\gamma_{0i}(t) = \gamma_{01i} + \gamma_{02i}t$, where $(\gamma_{01i}, \gamma_{02i})^T$ follows the bivariate normal distribution with zero-mean and covariance $\Gamma = (\Gamma_1, \Gamma_2)$ such that $\Gamma_1 = (6.25, 2.5)^T$ and $\Gamma_2 = (2.5, 4)^T$. The random error $\epsilon(t_{ij})$ is uncorrelated with $\gamma_0(t_{ij})$ and has the $N(0, 4)$ distribution.

For each simulated sample, our objective is to estimate $\text{RTP}_{A_\alpha}(s_1, s_2; x_1)$ and $\text{RTPR}_{A_\alpha}(s_1, s_2; x_1)$ defined in (3) and (6) for $x_1 = 0$ and $x_1 = 1$, where $A_\alpha(\cdot)$ is defined in (5) with $q_\alpha[t, X_2(t)]$ computed from the normal distribution based on the simulation model. Using the procedures in Section 3, we fit a mixed-effects varying coefficient model (9) with the cubic B-spline basis and equally spaced knots selected from BIC to each sample, and compute the estimators $\widehat{\text{RTP}}_{A_\alpha}(s_1, s_2; x_1)$ and $\widehat{\text{RTPR}}_{A_\alpha}(s_1, s_2; x_1)$ using the predicted subject-specific curves $\widehat{Y}_i(t)$ for $t \in [0, 10]$. The bootstrap confidence intervals for the RTP and RTPR estimators are computed with B = 500 bootstrap samples.

We repeated the simulation $M = 1,000$ times. Figure 4 shows the true curves computed based on the simulation model with known coefficient curves, the averages of their

Table 1. The true values of $RTP_{A_\alpha}(s_1, s_1+\delta; x_1)$ under the simulation model, the averages and the square root of the MSEs for the estimates and the empirical coverage probabilities of the bootstrap 95% pointwise confidence intervals from 1,000 simulations

			$\alpha = 90\%$				$\alpha = 75\%$			
		s_1	True Value	Ave. Est.	Ave. Root MSE	Cov. Prob.	True Value	Ave. Est.	Ave. Root MSE	Cov. Prob.
$\delta = 2$	$x_1 = 0$	0	0.331	0.327	0.032	0.955	0.472	0.471	0.024	0.945
		2	0.353	0.354	0.035	0.976	0.488	0.491	0.024	0.978
		4	0.391	0.392	0.040	0.986	0.521	0.524	0.025	0.982
		6	0.440	0.437	0.045	0.985	0.564	0.565	0.027	0.983
		8	0.496	0.487	0.053	0.978	0.611	0.610	0.033	0.973
	$x_1 = 1$	0	0.489	0.481	0.033	0.982	0.631	0.627	0.022	0.952
		2	0.529	0.529	0.027	0.969	0.666	0.668	0.017	0.986
		4	0.557	0.556	0.025	0.968	0.687	0.690	0.016	0.980
		6	0.573	0.569	0.026	0.960	0.698	0.698	0.016	0.974
		8	0.580	0.573	0.035	0.947	0.701	0.699	0.023	0.948
$\delta = 4$	$x_1 = 0$	0	0.290	0.292	0.032	0.929	0.427	0.430	0.025	0.946
		2	0.337	0.339	0.037	0.952	0.470	0.474	0.026	0.973
		4	0.399	0.396	0.044	0.976	0.527	0.528	0.027	0.976
		6	0.471	0.461	0.053	0.982	0.592	0.588	0.034	0.972
	$x_1 = 1$	0	0.545	0.549	0.032	0.965	0.682	0.681	0.021	0.986
		2	0.561	0.564	0.027	0.983	0.693	0.692	0.017	0.982
		4	0.565	0.565	0.026	0.964	0.694	0.691	0.017	0.962
		6	0.561	0.555	0.034	0.934	0.687	0.681	0.023	0.927

estimates and the lower and upper 2.5% pointwise percentiles of the estimated curves $\widehat{\mathrm{RTP}}_{A_\alpha}(s_1, s_1+\delta; x_1)$ and $\widehat{\mathrm{RTPR}}_{A_\alpha}(s_1, s_1+\delta; x_1)$ for $x_1 = 0$ and 1 computed at $\alpha = 90\%$, $\delta = 2$ and $s_1 \in [0, 8]$. These plots indicate that these average curves are very close to the true curves, and the widths of the lower and upper 2.5% pointwise percentiles are reasonably small. To evaluate the overall performance of our estimators, we computed the empirical mean squared errors

$$\mathrm{MSE}\Big[\widehat{\mathrm{RTP}}_{A_\alpha}(t, t+\delta; x_1)\Big] = \frac{1}{M}\sum_{m=1}^{M}\Big\{\widehat{\mathrm{RTP}}^{(m)}_{A_\alpha}(t, t+\delta; x_1) - \mathrm{RTP}_{A_\alpha}(t, t+\delta; x_1)\Big\}^2,$$

where t can be chosen from a grid of time points in $[0, 10-\delta]$ and $\widehat{\mathrm{RTP}}^{(m)}_{A_\alpha}(\cdot)$ is the estimate computed from the mth simulated sample.

We carried out the simulation for a range of (α, δ) values. For $\alpha = 90\%$ or 75% and $\delta = 2$ or 4, the averages and MSEs of the RTP estimates and the empirical coverage probabilities for the bootstrap percentile confidence intervals are summarized in Table 1. These results show that the RTP estimates based on our proposed models and procedures are closed to the true RTP curves and the coverage probabilities of the bootstrap confidence intervals are also close to the nominal level of 95%. We omit the simulation results for the RTPRs, because the RTPR estimators are also close to the true RTPR curves and their bootstrap confidence intervals have satisfactory empirical coverage probabilities. Note, we have applied the procedure in Section 3.3.3 with unstructured mixed-effects models (8) and (13) to the same simulation samples, the estimates are similar to Figure 4 and Table 1 and thus due to the limit of space those results are not presented here.

6. DISCUSSION

We developed in this paper a class of global smoothing methods for the estimation of the RTPs and RTPRs, which quantitatively measure the "rank-tracking abilities" of a time-dependent outcome variable in a longitudinal study. The RTPs can be intuitively interpreted as the conditional probabilities of a subject's health status at a later time point given that the same subject has certain health status at an earlier time point. Compared with serial correlations, which have been commonly used in the literature to evaluate the "tracking abilities" of time-dependent variables in longitudinal studies, the RTPs have two distinct advantages. First, as an intuitive quantitative measure that can incorporate subject's baseline or time-varying covariates, the RTPs and RTPRs have simple and straightforward interpretations in practice. Second, the RTPs and RTPRs do not depend on the restrictive assumptions of serial correlations, which may not be satisfied in many practical settings. In contrast, serial correlations may not have adequate interpretations if the relationship between the outcome variable evaluated at two time points are not linear or their joint distribution is significantly different from normality. The estimation and

inferences of an RTP may be constructed entirely nonparametrically or under certain structural nonparametric models which are sufficiently flexible in many longitudinal settings. The inherent model flexibility and scientific interpretability enable the RTPs and RTPRs to be used as a convenient statistical tool to identify certain disease risk factors that track over time.

We have attempted to minimize the data structural assumptions in developing our estimation methods, although the nonparametric mixed-effects models have been used to reduce the model and computational complexity. However, our application to the NGHS BMI and SBP data shows that the RTPs and RTPRs may be approximated by certain parametric or nonparametric models in practice. For example, if $\mathrm{RTP}_A(s_1, s_1+\delta)$ does not change with s_1 for any given risk set $A(\cdot)$, then it is appropriate to consider this RTP as a parametric or nonparametric function of δ only, so that more efficient estimators than the basis approximation estimators or other smoothing approaches may be constructed. Clearly, depending on the scientific questions and the nature of the data, various statistical models for the RTPs and RTPRs may be considered in a given situation. Thus, goodness-of-fit tests or model selection methods are needed to evaluate the adequacy of the RTP and RTPR models. Further research is warranted for the estimation and inferences of the RTPs and RTPRs under appropriate models.

ACKNOWLEDGMENTS

The authors would like to thank Drs. Eric Leifer and Dihua Xu for reviewing the initial draft, and Drs. Eva Obarzanek, Stephen R. Daniels, Rae-Ellen W. Kavey, Gail D. Pearson, and Jeffrey A. Cutler for explaining the importance of "rank-tracking" in pediatric studies and the objectives of the National Growth and Health Study. The National Growth and Health Study was supported by contracts #NO1-HC-55023-26 and grants #U01-HL48941-44 from the National Heart, Lung and Blood Institute.

Received 29 March 2013

REFERENCES

Daniels, S. R., McMahon, R. P., Obarzanek, E., Waclawiw, M. A., Similo, S. L., Biro, F. M, Schreiber, G. B., Kimm, S. Y. S., Morrison, J. A. and Barton, B. A. (1998). Longitudinal correlates of change in blood pressure in adolescent girls. *Hypertension* **31**, 97–103.

Diggle, P. J., Heagerty, P., Liang, K.-Y. and Zeger, S. L. (2002). *Analysis of Longitudinal Data*, 2nd ed. Oxford: Oxford University Press. MR2049007

Fan, J. and Wu, Y. (2008). Semiparametric estimation of covariance matrixes for longitudinal data. *Journal of the American Statistical Association* **103**, 1520–1533. MR2504201

Fitzmaurice, G., Davidian, M., Verbeke, G. and Molenberghs, G. (Editors) (2009). *Longitudinal Data Analysis*. Boca Raton, FL: Chapman & Hall/CRC. MR1500110

Hall, P., Wolff, R. C. L. and Yao, Q. (1999). Methods for estimating a conditional distribution function. *J. Amer. Statist. Assoc.* **94**, 154–163. MR1689221

Hoover, D. R., Rice, J. A., Wu, C. O. and Yang, L. P. (1998). Nonparametric smoothing estimates of time-varying coefficient models with longitudinal data. *Biometrika* **85**, 809–822. MR1666699

Kavey, R. E. W., Daniels, S. R., Lauer, R. M., Atkins, D. L., Hayman, L. L. and Taubert, K. (2003). American Heart Association guidelines for primary prevention of atherosclerotic cardiovascular disease beginning in childhood. *Circulation* **107**, 1562–1566.

Liang, H., Wu, H. and Carroll, R. J. (2003). The relationship between virologic and immunologic responses in AIDS clinical research using mixed-effects varying-coefficient models with measurement error. *Biostatistics* **4**, 297–312.

National High Blood Pressure Education Program Working Group on High Blood Pressure in Children and Adolescents (NHBPEP Working Group) (2004). The fourth report on the diagnosis, evaluation, and treatment of high blood pressure in children and adolescents. *Pediatrics* **114**, 555–576.

National Heart, Lung, and Blood Institute Growth and Health Research Group (NGHSRG) (1992). Obesity and cardiovascular disease risk factors in black and white girls: The NHLBI Growth and Health Study. *Am J. Public Health* **82** 1613–1620.

Obarzanek, E., Wu, C. O., Cutler, J. A., Kavey, R. W., Pearson, G. D. and Daniels, S. R. (2010). Prevalence and incidence of hypertension in adolescent girls. *The Journal of Pediatrics* **157**(3), 461–467.

Rice, J. A., and Wu, C. O. (2001). Nonparametric mixed effects models for unequally sampled noisy curves. *Biometrics* **57**, 253–259. MR1833314

Shi, M., Weiss, R. E. and Taylor, J. M. G. (1996). An analysis of paediatric CD4 counts for acquired immune defciency syndrome using flexible random curves. *Appl. Statist.* **45**, 151–163.

Thompson, D. R., Obarzanek, E., Franko, D. L., Barton, B. A., Morrison, J., Biro, F. M., Daniels, S. R. and Striegel-Moore, R. H. (2007). Childhood overweight and cardiovascular disease risk factors: The National Heart, Lung, and Blood Institute Growth and Health Study. *Journal of Pediatrics* **150**, 18–25.

Verbeke, G., and Molenberghs, G. (2000). *Linear Mixed Models for Longitudinal Data*. New York: Springer. MR1880596

Wei, Y., Pere, A., Koenker, R., and He, X. (2006). Quantile regression methods for reference growth charts. *Statistics in Medicine* **25**, 1369–1382. MR2226792

Wu, B. and Pourahmadi, M. (2003). Nonparametric estimation of large covariance matrices of longitudinal data. *Biometrika* **90**, 831–844. MR2024760

Wu, C. O., Tian, X. and Yu, J. (2010). Nonparametric estimation for time-varying transformation models with longitudinal data. *Journal of Nonparametric Statistics* **22**, 133–147. MR2598958

Wu, C. O. and Tian, X. (2013a). Nonparametric estimation of conditional distribution functions and rank-tracking probabilities with longitudinal data. *Journal of Statistical Theory and Practice* **7**, 259–284.

Wu, C. O. and Tian, X. (2013b). Nonparametric estimation of conditional distributions and rank-tracking probabilities with time-varying transformation models in longitudinal studies. *Journal of the American Statistical Association*, in press.

Zhou, L., Huang, J. Z. and Carroll, R. J. (2008). Joint modelling of paired sparse functional data using principal components. *Biometrika* **95**, 601–619. MR2443178

Xin Tian
Office of Biostatistics Research
National Heart, Lung and Blood Institute
Bethesda, MD 20892
USA
E-mail address: tianx@nhlbi.nih.gov

Colin O. Wu
Office of Biostatistics Research
National Heart, Lung and Blood Institute
Bethesda, MD 20892
USA
E-mail address: wuc@nhlbi.nih.gov