Statistica Sinica Preprint No: SS-2019-0127	
Title	Risk-predictive probabilities and dynamic nonparametric conditional quantile models for longitudinal analysis
Manuscript ID	SS-2019-0127
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202019.0127
Complete List of Authors	Seonjin Kim Hyunkeun Ryan Cho and Colin Wu
Corresponding Author	Hyunkeun Ryan Cho
E-mail	hyunkeun-cho@uiowa.edu
Notice: Accepted version subject to English editing.	

Statistica Sinica

Risk-predictive probabilities and dynamic nonparametric conditional quantile models for longitudinal analysis

Seonjin Kim, Hyunkeun Ryan Cho, and Colin Wu

Miami University, University of Iowa, and National Heart, Lung and Blood Institute

Abstract: Tracking subjects with disease risks at multiple time points is an important objective for disease prevention and preventive medicine (Wilsgaard et al., 2001; Obarzanek et al., 2010). Appropriate statistical tracking models are essential for identifying the risk factors that remain persistent over time and early detection of subjects with high disease risks. Since disease risks are often defined by multivariate response variables, we propose a class of bivariate risk-predictive probability models to quantify the likelihood of future disease risk for an individual. These models describe the relationships between bivariate risk outcomes at a later time point and covariates at an early time point through a class of conditional quantile-based joint distribution functions. We develop a simulation based procedure under the stratified bivariate time-varying quantile regression framework to estimate the conditional joint distributions and risk-predictive probabilities, demonstrate through theoretical and simulation studies that the estimation procedure yields consistent estimates, and propose a statistical quantity that measures the relative risk to identify high-risk individuals. Through an application to the National Growth and Health Study, we illustrate how the proposed models and procedures can be used to identify early adolescent girls who are more likely to be diagnosed with hypertension at late adolescence.

Key words and phrases: Bivariate longitudinal outcome; Conditional joint distributions; Nonparametric regression; Quantile regression; Time-varying coefficients.

1. Introduction

Longitudinal tracking of disease risk factors over time is important for guiding early preventive interventions in public health (e.g., Wilsgaard et al., 2001; Obarzanek et al., 2010). Given that a main objective of preventive medicine is to reduce the incidence of future disease risks through early intervention to individuals with high risk factors relative to the population, an appropriate statistical model would play a critical role in the identification of persistent disease risk factors and individuals who would develop high disease risks in the future. Like the proverb that the past is a mirror of the future, the past and present health status of the subject is likely an important indicator of the development of a disease. In this article, we develop a class of models for the risk-predictive probability (RPP) that measures how likely a disease would occur in the future given an individual's current condition. The RPP and its models can serve as an effective tool for early identification of persistent disease risk factors by identifying high-risk groups relative to the population.

This work is motivated by an epidemiological study of pediatric cardiac risk factors for children and adolescents, the National Growth and Health Study (NGHS) conducted from 1986 to 1997. This prospective cohort study was designed to explore the trend of cardiovascular risk factors in girls over an adolescent period. Various characteristics, such as systolic and diastolic blood pressure (SBP, DBP), race, height, and body mass index (BMI), were measured annually from 2379 African American and Caucasian girls up to ten visits. Kavey et al. (2003), Thompson et al. (2007), and Obarzanek et al. (2010) have studied the NGHS and raised the following important question: What types of features in early adolescent girls have an influence on the presence of hypertension at late adolescence? Since normal and abnormal levels of blood pressure (BP) for children and adolescents are defined jointly by the SBP and DBP percentiles (Flynn et al.,

2017), a major obstacle for addressing this question is the lack of an appropriate statistical model which describes the joint distributions of the bivariate longitudinal outcomes, SBP and DBP at late adolescence, conditioning on their values and other covariates at early adolescent.

Disease risk factors defined by bivariate (or more generally, multivariate) longitudinal outcomes are common in biomedical studies. For example, in biomarker studies of human immunodeficiency viruses (HIV), the bivariate outcome formed by CD4 cells and HIV viral load (HIV-RNA) in blood is often used as a prognostic measure on HIV progression (Thiébaut et al., 2002; Thiébaut et al., 2005; Ghosh et al., 2007); in cardiovascular studies, it is demonstrated by Barter et al. (2007) that risks for cardiovascular events may be affected by the levels of high-density lipoprotein (HDL) cholesterol and low-density lipoprotein (LDL) cholesterol jointly.

In a conditional distribution-based attempt for longitudinal analysis, Wu and Tian (2013a, 2013b) and Tian and Wu (2014) considered a statistical quantity, the "rank-tracking probability" (RTP), to measure the tracking abilities of disease risk factors over time in longitudinal studies. Their statistical framework, however, is limited to univariate longitudinal outcomes and not applicable to joint distributions with bivariate outcomes because of the complexity of time-varying nonparametric modeling structures that are both clinically meaningful and mathematically flexible (Wu and Tian, 2018, Sections 12.2 and 12.6). Statistical methods for multivariate longitudinal data have been mostly studied under the framework of conditional means and variance-covariance structures. Examples of multivariate longitudinal analysis in the literature include Rochon (1996), Chaganty and Naik (2002), Fieuws and Verbeke (2006), Kim and Zimmerman (2012), Xu and Mackenzie (2012), Xiang et al. (2013), Verbeke et al. (2014), Cho (2016), and Kohlia et al. (2016), among others. These results are focused on modeling conditional means and variance-covariance structures based on the concurrently observed multivariate outcomes and covariates.

Kwak (2017a, 2017b) and Kürüm et al. (2018) considered a number of copula-based models for evaluating the conditional distribution functions with multivariate longitudinal data. But, again, these models do not describe the dynamic relationships involving both past and future variables formed by the multivariate outcomes and covariates.

In contrast to the existing multivariate longitudinal methods, we propose in this paper a class of conditional distribution-based models to evaluate the "tracking" relationship between the bivariate response vector at a later time point and the response and covariate values at an earlier time point. Let $(Y_1(t), Y_2(t))$ be the bivariate vector of real-valued responses $Y_1(t)$ and $Y_2(t)$ and $Z(t) = (X(t)^T, Y_1(t), Y_2(t))^T$ be the vector of covariates and responses at any time point $t \in \mathcal{T}$, where $X(t)$ is the $p \geq 1$ dimensional vector of covariates and the time range $\mathcal{T}$ is a bounded subset of $[0, \infty)$. For any two time points $u < v$ in $\mathcal{T}$, our goal is to model and estimate the conditional distribution function and their functional of $(Y_1(v), Y_2(v))$ given $Z(u) = z(u)$. Here, $z(u) = (x(u)^T, y_1(u), y_2(u))^T$ represents the known "health status" for a subject at time u which, in general, includes both the covariate and response variables. As useful special cases of $z(u)$, we may consider the situations "without covariates," i.e. $z(u) = (y_1(u), y_2(u))^T$, and "without outcomes," i.e. $z(u) = x(u)$. We note that u represents an early time point of interest so it is often used to denote an individual's current age or the most recent time point when the health status $z(u)$ is measured and v represents a later time point of interest, so it is used to denote a specific future time such as 10 years later from u, i.e., $v = u + 10$.

Since a completely unstructured nonparametric model of the conditional distribution functions of $(Y_1(v), Y_2(v))$ given $Z(u)$ is generally not mathematically tractable and biologically interpretable, we propose a class of structured nonparametric regression models for the RPPs (Section 2.1) based on conditional quantiles. In order to focus on the main objective of tracking the mul-

tivariate outcomes across the time range $\mathcal{T}$, our nonparametric quantile regression models rely on linking the outcomes and covariates at time points (u, v) through linear structures with bivariate functional parameters. This is different from the longitudinal quantile regression in the literature, such as Kim and Yang (2011) and Cho, Hong and Kim (2016). There are two main advantages for this conditional quantile-based modeling approach for evaluating the RPPs for tracking multivariate longitudinal outcomes. First, the RPPs and the related conditional distributions can be simply estimated based on the nonparametric estimators of the functional quantile regression parameters through a simulation-based procedure. Second, the nonparametric conditional quantile models have natural interpretations for applications where health status is classified based on conditional quantiles, such as abnormal levels of blood pressure defined in Flynn et al. (2017).

In the main results, we first demonstrate that the joint condition distribution functions can be estimated by a simulation-based procedure constructed using Wei (2008, Lemma 1) and a class of quantile regression models with bivariate time-varying coefficients. Then we show that the RPP estimators obtained from the simulation-based procedure are consistent under the bivariate time-varying coefficient models. For the practically interesting objective of determining whether an "unhealthy" individual in the past is more likely to have higher future disease risk, we propose a statistical inference procedure based on resampling-subject bootstrapping. The inference procedure compares the RPP with the unconditional joint probability of response variables at any v without knowing their values at any $u < v$. In our application to the NGHS data, we estimate the RPPs that a preadolescent girl with various BP and BMI levels develops abnormal levels of BP at later adolescent years. Furthermore, we demonstrate in a simulation study the consistency of the RPP estimators by comparing their values with the RPPs obtained without imposing any modeling structures.

2. Methodology

2.1 Risk-predictive probability models

We introduce in this section the RPP and present a simulation-based procedure for estimating the RPP. For any sets of events $A_1(v) \subset \mathbb{R}$ and $A_2(v) \subset \mathbb{R}$ on the real line at time v, we define the RPP as

$$\text{RPP}\{A_1(v), A_2(v)|z(u)\} = P\{Y_1(v) \in A_1(v), Y_2(v) \in A_2(v)|Z(u) = z(u)\}, \tag{2.1}$$

which is the conditional joint probability of $Y_1(v) \in A_1(v)$ and $Y_2(v) \in A_2(v)$ given an individual's health status, i.e., outcomes and covariates, at time u, $Z(u) = z(u)$ where $u < v$. For any given covariate values, the RPP defined in (2.1) is a function on the bivariate time scale (u, v). Consequently, a estimator of (2.1) is a bivariate curve on (u, v), which allows the investigator to evaluate the risk-predictive ability at any time point pairs within the range of interest. The statistical objective is to estimate the RPP based on a flexible and clinically meaningful structured nonparametric model. Since the RPP measures how likely a subject with health status $z(u)$ at a earlier time u belongs to the event $\{Y_1(v) \in A_1(v), Y_2(v) \in A_2(v)\}$ at a later time v, it provides a direct statistical index to track subjects who are likely to have the event in the future.

In practice, proper choices of $A_1(v)$ and $A_2(v)$ are determined by the study objectives. For example, for the study of adolescent BP levels, the probability of having an "abnormal BP level defined by the 95th percentiles" (Flynn et al., 2017) at time v given the subject's BP and other

covariates at time u is

$$\text{RPP}_{\text{abnormal BP}}(v|u) = 1 - \text{RPP}\{(-\infty, Q_{.95}\{Y_1(v)\}), (-\infty, Q_{.95}\{Y_2(v)\})|z(u)\}, \quad (2.2)$$

while $\text{RPP}\{(-\infty, Q_{.95}\{Y_1(v)\}), (-\infty, Q_{.95}\{Y_2(v)\})|z(u)\}$ is the probability of not having an "abnormal BP level" at time v given $z(u)$, where $Y_1(t)$ and $Y_2(t)$ are SBP and DBP, respectively, at time t and $Q_{\tau_k}\{Y_k(t)\}$ is the $\tau_k \times 100$th quantile of $Y_k(t)$ for $k = 1, 2$. Note that, in this example, $A_1(v)$ and $A_2(v)$ are defined using percentiles of response variables, but, in general they could be defined with predetermined values. For adult BP studies (Chobanian et al., 2013), the RPP of having hypertension can be defined as $\text{RPP}_{\text{abnormal BP}}(v|u) = 1 - \text{RPP}\{(-\infty, 140), (-\infty, 90)|z(u)\}$.

Comparing the RPP with the unconditional joint probability

$$P\{A_1(v), A_2(v)\} = P\{Y_1(v) \in A_1(v), Y_2(v) \in A_2(v)\},$$

we can examine how the occurrence of the event at time v is influenced by the observed health status at time u. Suppose that a subject is classified to have a "high disease risk" at time v if the bivariate outcomes are in the event $\{Y_1(v) \in A_1(v), Y_2(v) \in A_2(v)\}$. If the subject's $\text{RPP}\{A_1(v), A_2(v)|z(u)\}$ is greater than $P\{A_1(v), A_2(v)\}$, the subject is more likely to have "high disease risk" than the population of interest because of the subject's health status $z(u)$ at time u. The magnitude of the increased disease risk can be quantified by the ratio of RPP relative to the

benchmark $P\{A_1(v), A_2(v)\}$

$$\text{RR}\{A_1(v), A_2(v)|z(u)\} = \frac{\text{RPP}\{A_1(v), A_2(v)|z(u)\}}{P\{A_1(v), A_2(v)\}}, \tag{2.3}$$

which we refer to as the *relative risk* (RR). Note that a subject having RR > 1 is more likely to have "higher disease risk" then the population of interest.

2.2 Nonparametric dynamic conditional quantile models

When the sample size is very large, the estimation of the RPP might be achieved using a smoothing method without imposing any modeling structures. However, an unstructured smoothing for the RPP is usually infeasible in practice because of the well-known "curse of dimensionality," e.g., Wu and Tian (2018, Section 1.3.3). A useful alternative is to consider a modeling structure for the RPP that is sufficiently flexible. Using a structured nonparametric approach, we consider a class of time-varying coefficient quantile regression models with some structural assumptions between $(Y_1(v), Y_2(v))$ and $Z(u)$.

The following lemma of Wei (2008) describes a useful relationship between the marginal, conditional and joint distributions of multivariate random variables. This lemma suggests that, in order to estimate the conditional distributions of $(Y_1(v), Y_2(v))$ given $Z(u)$, we need to model the conditional distributions of $Y_1(v)$ given $Z(u)$ and $Y_2(v)$ given $(Z(u), Y_1(v))$, respectively.

Lemma 1. *Suppose that (Y_1, Y_2) is a pair of random variables with absolute continuous joint distribution F_{Y_1,Y_2}, and let U_1 and U_2 be two independent random variables uniformly distributed on (0,1), then*

$$(F^{-1}_{Y_1}(U_1), F^{-1}_{U_2|U_1}(U_2|U_1)) \sim F_{Y_1,Y_2},$$

where $F_{Y_1}(A_1)$ *is the marginal distribution of* Y_1 *and* $F_{Y_2|Y_1}$ *is the conditional distribution of* Y_2 *given* Y_1.

Since the inverse function of a cumulative distribution function (CDF) is a quantile function, Lemma 1 ensures that a bivariate random sample generated sequentially from $Q_\tau\{Y_1(v)|Z(u)\}$ and $Q_\tau\{Y_2(v)|Z(u), Y_1(v)\}$ follows the conditional distribution of $(Y_1(v), Y_2(v))$ given $Z(u)$. By imposing a linear modeling structure with coefficients to be time-varying curves, we propose the following dynamic models for $Q_\tau\{Y_1(v)|Z(u)\}$ and $Q_\tau\{Y_2(v)|Z(u), Y_1(v)\}$, such that

$$Q_\tau\{Y_1(v)|Z(u)\} = \alpha_{\tau,1}(v|u) + Z^T(u)\alpha_{\tau,2}(v|u) \quad \text{and} \tag{2.4}$$

$$Q_\tau\{Y_2(v)|Z(u), Y_1(v)\} = \beta_{\tau,1}(v|u) + Z^T(u)\beta_{\tau,2}(v|u) + \beta_{\tau,3}(v|u)Y_1(v), \tag{2.5}$$

where $\alpha_{\tau,1}(v|u), \alpha_{\tau,2}(v|u), \beta_{\tau,1}(v|u), \beta_{\tau,2}(v|u)$, and $\beta_{\tau,3}(v|u)$ are unknown coefficient functions of both u and v. Intuitively, (2.4) shows that, for any $0 < \tau < 1$ and a pair of time points (u, v), the τth quantile of $Y_1(v)$ depends on $Z(u)$ through a linear relationship with the time-varying regression quantiles $\alpha_{\tau,1}(v|u)$ and $\alpha_{\tau,2}(v|u)$. Since the functional coefficients can vary with two distinct time points, the above models can be used to explore the dynamic relationship between the bivariate response variables and covariates measured at different time points across the quantiles. In addition, these functional parameters determine the conditional quantiles, hence, the conditional distribution functions. Consequently the estimates of these functional parameters can be used to estimate the conditional distribution functions.

2.3 Estimation of the dynamic conditional distributions

If the dynamic conditional quantiles of models (2.4) and (2.5) are available, Lemma 1 suggests that RPP$\{A_1(v), A_2(v)|z(u)\}$ can be estimated by the following simulation-based procedure:

1. Draw q_1 from a uniform distribution on $(0, 1)$ and obtain the conditional q_1th quantile of $Y_1(v)$ given $z(u)$, denoted by $Y_1^*(v)$, from an estimated model of (2.4).

2. Draw q_2 from a uniform random variable on $(0, 1)$ and obtain the conditional q_2th quantile of $Y_2(v)$ given $z(u)$ and $Y_1^*(v)$, denoted by $Y_2^*(v)$, from an estimated model of (2.5).

3. Generate a sufficiently large number of $(Y_1^*(v), Y_2^*(v))$ by repeating 1–2 many times.

4. Estimate the RPP by computing the proportion of the simulated pairs within $Y_1^*(v) \in A_1(v)$ and $Y_2^*(v) \in A_2(v)$.

Note that the bivariate random sample of $Y_1^*(v)$ and $Y_2^*(v)$ can also be obtained by switching the order of $Y_1(t)$ and $Y_2(t)$. Unless there is a natural ordering between $Y_1(t)$ and $Y_2(t)$, we can in practice simulate the data in both orders and take the combined data to estimate the conditional joint distribution.

Statistical inferences for the RPP and RR can be constructed using the common approach of resampling-subject bootstrap for longitudinal data (e.g., Hoover et al., 1998). In addition to the RPP, statistical inferences for the RR have clinical implications for identifying individuals who are more likely to have "high disease risks" in future times relative to others in the population. In particular, we would like to determine if RR > 1 for the time range of interest. Using the resampling-subject bootstrap procedure of Hoover et al. (1998), we construct the one-sided pointwise confidence interval for the RR. This procedure relies on three steps: (a) generating B

bootstrap samples by resampling the subjects with replacement, (b) estimating the corresponding RRs from each of the B bootstrap samples, and (c) computing the $(100 \times \alpha)$th empirical quantile of the estimated RRs from the B bootstrap samples as the lower bound of the one-sided α-level confidence interval.

2.4 Estimation of the time-varying regression quantiles

In this section, we propose a novel estimation procedure for the time-varying regression quantiles in (2.4) and (2.5) based on the following longitudinal sample that consists of n randomly selected subjects. The ith subject, $1 \leq i \leq n$, has $m_i \geq 1$ measurements at time points t_{ij}, $j = 1, \ldots, m_i$, such that $(Y_{1,ij}, Y_{2,ij})$ and X_{ij} are the bivariate outcome and a vector of p covariates, respectively, at time t_{ij}.

To clarify the relationship between the response and covariates at different time points, it is convenient to denote the longitudinal observations as follows. Within each subject i, for any $j < j'$, $Y_{ij'} = (Y_{1,ij'}, Y_{2,ij'})$ is a pair of future response variable relative to $Z_{ij} = (Y_{ij}, X_{ij}^T)$. The longitudinal data are then expressed as $(Y_{ij'}, Z_{ij}, t_{ij'}, t_{ij})$ for $i = 1, \ldots, n$, $j = 1, \ldots, m_i - 1$ and $j' = j + 1, \ldots, m_i$, so that the future response variables are paired with the past outcomes and covariates. For example, if the first subject, i.e. $i = 1$, is measured four times, then the subject's data used to estimate the coefficients in (2.4) and (2.5) are

$$(Y_{12}, Z_{11}),\ (Y_{13}, Z_{11}),\ (Y_{13}, Z_{12}),\ (Y_{14}, Z_{11}),\ (Y_{14}, Z_{12}),\ (Y_{14}, Z_{13}),$$

where $Y_{ij'} = (Y_{1,1j'}, Y_{2,1j'})$ and $Z_{ij} = (Y_{1,ij}, Y_{2,ij}, X_{ij}^T)$. Since Z_{ij} is the available observation for the ith subject at time t_{ij}, it could include both the covariates X_{ij} and the bivariate outcomes Y_{ij}.

However, due to practical reasons, some longitudinal studies may not have observed outcomes at every visit. For instance, if Y_{12} and Y_{14} are measured during four visits while all the covariates are measured at every visit of the subject $i = 1$, then the subject's observations are

$$(Y_{12}, Z_{11}),\ (Y_{14}, Z_{11}),\ (Y_{14}, Z_{12}),\ (Y_{14}, Z_{13}),$$

where $Z_{ij} = X_{ij}$. Similarly, when the covariates are not available at time t_{ij}, we have $Z_{ij} = Y_{ij}$.

We note that model (2.4) is a special case of model (2.5), which depends on the ordering between $Y_1(t)$ and $Y_2(t)$. On the other hand, $Y_1(v)$ and $Y_2(v)$ in (2.5) are exchangeable. Thus, it suffices to only present the estimation and asymptotic properties for the quantile regression model (2.5). Without loss of generality, we suppose that the response and covariates are measured on each visit. For any $j < j'$, $Y_{2,ij'}$ is a future response variable in view of a predictor $Z_{ijj'} = (1, Y_{1,ij}, Y_{2,ij}, X_{1,ij}^T, Y_{1,ij'})$ so that the longitudinal data are expressed as $(Y_{2,ij'}, Z_{ijj'}, t_{ij'}, t_{ij})$ for $i = 1, \ldots, n$, $j = 1, \ldots, m_i - 1$ and $j' = j + 1, \ldots, m_i$. Restructured notation can be made similarly to other cases.

Let $\theta_\tau(v|u) = (\beta_{\tau,1}(v|u), \beta_{\tau,2}(v|u)^T, \beta_{\tau,3}(v|u))^T$ be the vector of functional parameters which depend on two distinct time points $u < v$. A local estimator of $\theta_\tau(v|u)$, denoted by $\hat{\theta}_\tau(v|u)$, is obtained by minimizing the local linear quantile regression criterion

$$\begin{aligned}
&\Big(\hat{\theta}_\tau(v|u), \hat{\theta}^*_\tau(v|u), \hat{\theta}^\#_\tau(v|u)\Big) \\
&= \underset{\theta,\theta^*,\theta^\#}{\operatorname{argmin}} \sum_{i=1}^{n} \sum_{j=1}^{m_i-1} \sum_{j'=j+1}^{m_i} \Big[\rho_\tau\Big(Y_{2,ij'} - Z_{ijj'}^T\theta - Z_{ijj'}^T\{\theta^*(t_{ij} - u) + \theta^\#(t_{ij'} - v)\}\Big) \\
&\qquad \times K\Big(\frac{t_{ij} - u}{b_1}\Big) K\Big(\frac{t_{ij'} - v}{b_2}\Big)\Big],
\end{aligned}$$

where $\rho_\tau(u) = u\{\tau - \mathbf{1}(u < 0)\}$ is the check function with an indicator function $\mathbf{1}(\cdot)$, $K(\cdot)$ is a kernel function with bandwidths b_1 and b_2 and $\theta^*(v|u)$ and $\theta^\#(v|u)$ are the first partial derivatives of $\theta(v|u)$ with respect to u and v, respectively. Here, the kernel function assigns more weight to the longitudinal observations whose time points $(t_{ij}, t_{ij'})$ are closer to target time points (u, v). If $(t_{ij}, t_{ij'})$ moves away from (u, v), the contribution from this observation to the quantile estimator diminishes, which leads to the reduction of potential estimation bias.

2.5 Asymptotic properties of kernel estimators

For simplicity, we only develop in this article the asymptotic properties of $\hat{\theta}_\tau(v|u)$ for the case that the response variables are observed at all the measurement times. Similar derivations with more tedious calculations can be extended to the cases with outcome variables not completely observed at all the measurement times for all subjects. The asymptotic properties of $\hat{\theta}_\tau(v|u)$ are established under the following regularity assumptions:

1. For any $u, v \in \mathcal{T}$, $\Gamma_Z(u, v) = E\{Z(u, v)Z(u, v)^T\}$ is positive-definite and differentiable, where $Z(u, v) = (1, Y_1(u), Y_2(u), X(u)^T, Y_1(v))^T$.

2. Let $N = \sum_{i=1}^n m_i(m_i - 1)/2$. As $n \to \infty$, $Nb_1b_2 \to \infty$, $Nb_1b_2(b_1^6 + b_2^6) \to 0$, and $\sum_{i=1}^n m_i^4\Big(1/\sqrt{N^3b_1b_2} + (b_1^2 + b_2^2)/N\Big) \to 0$.

3. The time-varying coefficient function $\theta(v|u)$ and the bivariate density function of (u, v), denoted by $p(\cdot, \cdot)$, are twice continuously differentiable.

4. The kernel function $K(\cdot)$ is symmetric with bounded support and bounded derivative. Write $\mu_K = \int u^2K(u^2)du$ and $\varphi_K = \int K^2(u)du$.

5. Let $\xi_{ijj'} = Y_{2,ij'} - Q_\tau(Y_{2,ij'}|Z_{ijj'})$. Denote by $f_\xi(\cdot)$ and $F_\xi(\cdot)$ the density and distribution functions of $\xi_{ijj'}$, respectively. Here, $f_\xi(\cdot)$ is bounded, positive, and twice continuous differentiable on $\{v : 0 < F_\xi(v) < 1\}$.

The above assumptions are comparable to the ones used in kernel estimation with longitudinal data (e.g., Hoover et al., 1998; Wu and Tian, 2013b). In particular, Assumption 2 specifies the necessary conditions with respect to the number of within subject measurements and bandwidths. For ease of presentation, we consider the special case of $m_i = m$ for all i, so that Assumption 2 reduces to $nm^2b_1b_2 \to \infty$, $nm^2b_1b_2(b_1^6 + b_2^6) \to 0$ and

$$nm^4\Big(1/\sqrt{n^3m^6b_1b_2} + (b_1^2 + b_2^2)/nm^2\Big) = m/\sqrt{nb_1b_2} + m(b_1 + b_2) \to 0.$$

In particular, if $b_1 = O(N^{-1/6})$ and $b_2 = O(N^{-1/6})$ are used, we have that $nm^2b_1b_2 \to \infty$ and $nm^2b_1b_2(b_1^6+b_2^6) \to 0$ always hold, but, in addition, $m = o(n^{1/4})$ is needed to ensure that $mb_1 \to 0$, $mb_2 \to 0$, and $m/\sqrt{nb_1b_2} \to 0$. Therefore, the data types specified by Assumption 2 include both sparse (i.e., m is bounded) and some dense ($m = n^\gamma$ for $\gamma < 1/4$) longitudinal data.

Theorem 1. *Let $u < v$ be two fixed time points in the interior of $\mathcal{T}$. If Assumptions 1 to 5 hold, then, for any given $\tau \in (0, 1)$, we have the following asymptotic normality result for $\hat{\theta}_\tau(v|u)$*

$$\begin{aligned}&\sqrt{Nb_1b_2}\left\{\hat{\theta}_\tau(v|u) - \theta_\tau(v|u) + \frac{\mu_K}{2}\Big(\frac{\partial^2\theta_\tau(v|u)}{\partial u^2}b_1^2 + \frac{\partial^2\theta_\tau(v|u)}{\partial v^2}b_2^2\Big)\right\}\\ &\xrightarrow{d} N\left(0, \frac{\tau(1-\tau)\varphi_K^2}{p(u,v)f_\xi^2(0)}\Gamma_Z^{-1}(u,v)\right),\end{aligned} \tag{2.6}$$

as $n \to \infty$, where "$\xrightarrow{d}$" denotes convergence in distribution.

A direct conclusion of Theorem 1 is that, under the mild regularity conditions, the local linear

quantile regression method leads to a consistent estimator of $\theta_\tau(v|u)$. If the linearity assumptions on the conditional quantile functions in (2.4) and (2.5) are satisfied, $z^T\hat{\theta}_\tau(v|u)$ is a consistent estimator of the τth conditional quantile of $Y_2(v)$ given $z = (1, z(u)^T, y_1(y_2))^T$. This consistency result suggests that the model-based simulation procedure described in Section 2.1 provides a consistent estimate of the RPP. Furthermore, if the outcomes are not observed at every visit, Theorem 1 still holds, but the convergence rate in Theorem 1 is affected, because N, the total number of observations used in the estimation, decreases.

2.6 Smoothing Estimators and Cross-Validation Bandwidths

Similar to kernel-type local smoothing in the literature, the choice of bandwidths plays a crucial role in the appropriateness of the smoothing estimators. We present here a "leave-one-subject-out cross-validation" (LSCV) method for selecting the data-driven bandwidths b_1 and b_2 for the local smoothing estimators of the time-varying regression quantiles and RR$\{A_1(v), A_2(v)|z(u)\}$.

Suppose that the bandwidths b_1 and b_2 have the same order of magnitude. It follows directly from the asymptotic distribution in Theorem 1 that the optimal bandwidths, which minimize the mean square error of $\hat{\theta}_\tau(v|u)$, are of order $O(N^{-1/6})$. Following the bandwidth selection strategy described in Yu and Jones (1998), the bandwidths are selected as

$$b_i = h_i[\tau(1-\tau)/\phi^2\{\Phi^{-1}(\tau)\}]^{1/6}, \quad i = 1, 2, \tag{2.7}$$

where ϕ and Φ are standard normal density and distribution functions, respectively, and h_1 and h_2

are bandwidths selected for the corresponding regression mean estimation, which minimizes

$$\sum_{i=1}^{n}\sum_{j=1}^{m_i-1}\sum_{j'=j+1}^{m_i}\Big[Y_{2,ij'}-Z_{ijj'}^T\theta-Z_{ijj'}^T\{\theta^*(t_{ij}-u)+\theta^{\#}(t_{ij'}-v)\}\Big]^2 \times K\Big(\frac{t_{ij}-u}{h_1}\Big)K\Big(\frac{t_{ij'}-v}{h_2}\Big). \tag{2.8}$$

To select the data-driven bandwidths h_1 and h_2, we use the LSCV bandwidths (Rice and Silverman, 1991) given by

$$(h_1,h_2)=\underset{h_1^*,h_2^*}{\operatorname{argmin}}\sum_{i=1}^{n}\sum_{j=1}^{m_i-1}\sum_{j'=j+1}^{m_i}\Big\{Y_{2,ij}-Z_{ijj'}^T\hat{\theta}^{-i}(t_{ij'}|t_{ij};h_1^*,h_2^*)\Big\}^2,$$

where $\hat{\theta}^{-i}(\cdot|\cdot;h_1^*,h_2^*)$ is the estimator of the mean regression coefficients based on remaining data with all the observations of the ith subject deleted.

Since, by (2.3), $RR\{A_1(v),A_2(v)|z(u)\}$ depends on $P\{A_1(v),A_2(v)\}$ as its denominator, we would like to estimate $P\{A_1(v),A_2(v)\}$ using the kernel estimator

$$\widehat{P}\{A_1(v),A_2(v)\}=\frac{\sum_{i=1}^{n}\sum_{j=1}^{m_i}\mathbf{1}\{Y_{1,ij}\in A_1(v),Y_{2,ij}\in A_2(v)\}K\Big(\frac{t_{ij}-v}{h}\Big)}{\sum_{i=1}^{n}\sum_{j=1}^{m_i}K\Big(\frac{t_{ij}-v}{h}\Big)}$$

with the same kernel function $K(\cdot)$ as in (2.8) and a bandwidth $h>0$. The data-driven bandwidth of h can be selected using the LSCV procedure, which is given by

$$h=\underset{h^*}{\operatorname{argmin}}\sum_{i=1}^{n}\sum_{j=1}^{m_i}\Big[\mathbf{1}\{Y_{1,ij}\in A_1(v),Y_{2,ij}\in A_2(v)\}-\widehat{P}^{-i}\{A_1(v),A_2(v);h^*\}\Big]^2,$$

where $\widehat{P}^{-i}\{A_1(v),A_2(v);h^*\}$ is the kernel estimator of $P\{A_1(v),A_2(v)\}$ based on the remaining

data with all the observations of the ith subject deleted. The resulting estimator of $RR\{A_1(v), A_2(v)|z(u)\}$ is

$$\widehat{\mathrm{RR}}\{A_1(v), A_2(v)|z(u)\} = \frac{\widehat{\mathrm{RPP}}\{A_1(v), A_2(v)|z(u)\}}{\widehat{P}\{A_1(v), A_2(v)\}},$$

which is obtained by plugging the corresponding estimators in the expression of (2.3).

3. Application to the NGHS BP Data

We apply our estimation and inference procedures to the NGHS to evaluate the predictive probabilities of the bivariate BP outcomes, SBP and DBP, during adolescent years with race, BMI percentile, and height percentile as covariates. As discussed in the introduction, this is a prospective cohort study of the cardiovascular risk factors of 1166 Caucasian and 1213 African American girls who were enrolled in the study at either 9 or 10 years of age and had up to an annual physical examination until 18 or 19 years old. Some further details and statistical analyses of this study are discussed in Wu and Tian (2018, Sections 1.2 and 13.4). Since some study participants have missing measurements due to reasons completely unrelated to the study or their health status, it is reasonable to assume that these data are missing completely at random, so after deleting the missing data, our analysis uses the longitudinal observations from 1164 Caucasian and 1212 African American girls. The number of repeated measurements has a range of 1 to 10 with a median of 9.0, mean of 8.3, and a standard deviation of 2.0. All covariates and bivariate outcomes are measured at each visit. The girls' BMI and height percentiles are computed based on the Centers for Disease Control and Prevention (CDC) growth chart as in Wu and Tian (2018, Section 13.4). Among two attempts in the literature to investigate the conditional distributions of a univariate longitudinal outcome with the NGHS data, Wu, Tian, and Yu (2010) studies the time-varying effects of race, BMI

percentile, and height percentile on the SBP and Wu and Tian (2013a) estimates the time-trends of the conditional distributions of the SBP.

Let $Y_1(t)$, $Y_2(t)$, X_1, $X_2(t)$, and $X_3(t)$ be the SBP, DBP, race, and BMI percentile, and height percentile, respectively, at age t, where $X_1 = 1$ if the girl is African American and 0 if Caucasian. We recall the probability of having an "abnormal BP level" defined as

$$\text{RPP}_{\text{abnormal BP}}(v|u) = 1 - \text{RPP}\{(-\infty, Q_{.95}\{Y_1(v)\}), (-\infty, Q_{.95}\{Y_2(v)\})|Z(u) = z(u)\}, \quad (3.1)$$

where $Z(t) = (X_1, X_2(t), X_3(t), Y_1(t), Y_2(t))^T$ and $Q_\tau\{Y_1(t)\}$ and $Q_\tau\{Y_2(t)\}$ are the $(\tau \times 100)$th quantiles of SBP and DBP at age t. Since all subjects enrolled in the study at the age of 9 or 10 and were followed for 9 years, we can estimate $\text{RPP}_{\text{abnormal BP}}(v|u)$ for $9 \le u < v \le 19$ yet for the purpose of illustration, $\text{RPP}_{\text{abnormal BP}}(18|10)$ is analyzed.

We first use the NGHS data of girls whose age is within the interval $[17.5, 18.5)$ and estimate $(Q_{0.95}\{Y_1(18)\}, Q_{0.95}\{Y_2(18)\})$, which is $(\hat{Q}_{0.95}\{Y_1(18)\}, \hat{Q}_{0.95}\{Y_2(18)\}) = (123, 80)$. Using the kernel smoothing estimators of Section 2.6 with Gaussian kernel, the joint probability of having SBP or DBP above their corresponding 95th percentile at the age of 18 is estimated to be 7.7%, i.e.,

$$1 - \widehat{P}\{(-\infty, \hat{Q}_{0.95}\{Y_1(18)\}), (-\infty, \hat{Q}_{0.95}\{Y_2(18)\})\} = 0.077.$$

To illustrate how the observed health status at an earlier adolescent period influences the probability of future abnormal BP levels, we estimate the RPP of having SBP or DBP over the 95th percentile at the age of 18 under various combinations of height percentile and BMI percentile, and three BP groups at the age of 10: the "medium-BP" $(Q_{0.5}\{Y_1(10)\}, Q_{0.5}\{Y_2(10)\})$, the "above median-BP" $(Q_{0.75}\{Y_1(10)\}, Q_{0.75}\{Y_2(10)\})$, and the "elevated-BP" $(Q_{0.9}\{Y_1(10)\}, Q_{0.9}\{Y_2(10)\})$.

These quantiles are estimated using data from girls within the age interval $[9.5, 10.5)$. We compute the kernel smoothing estimators with the Gaussian kernel using the bandwidths $(h_1, h_2) = (1.0, 1.5)$ and $(h_1, h_2) = (2.3, 1.5)$, respectively, for the following two sets of quantile regression models at age v: (a) a marginal quantile model of $Y_1(v)$ and a quantile regression model of $Y_2(v)$ conditioning on $Y_1(v)$; (b) a marginal quantile model of $Y_2(v)$ and a quantile regression model of $Y_1(v)$ conditioning on $Y_2(v)$. The models in (a) and (b) differ in their orders of $Y_1(v)$ and $Y_2(v)$. These bandwidths are selected by the LSCV procedure of Section 2.3 and the quantile adjustment given in (2.7). In each order, a bivariate random sample of 1000 is generated from the proposed simulation based procedure.

Figure 1 shows the heat maps of the estimated $\text{RPP}_{\text{abnormal BP}}(18|10)$. The color on the heat maps gradually changes from red to green representing the gradually decreasing estimated $\text{RPP}_{\text{abnormal BP}}(18|10)$. The colors of the estimated probability become lighter when their values are closer to the estimated probability $1 - \widehat{P}\{(-\infty, \hat{Q}_{0.95}\{Y_1(v)\}), (-\infty, \hat{Q}_{0.95}\{Y_2(v)\})\}$, which is 0.077 and represented by white on the heat maps. For the effects of the covariates, we observe in Figure 1 that a ten-year old girl with larger BMI and height percentiles is more likely to have her SBP or DBP over their corresponding 95th percentiles at age 18. For the dynamic effects of BP over time, it is seen in Figure 1 that there is a positive dependence between the BP levels at earlier and later adolescent periods in the sense that higher SBP and DBP levels at age 10 are associated with higher probability of having SBP or DBP over their 95th percentiles at age 18. In particular, for any covariate values (i.e., race, BMI and height), the estimated RPP of SBP or DBP over their 95th percentiles at age 18 for the girls with high SBP and DBP levels at age 10 is always higher than the estimated probability of SBP or DBP over their 95th percentiles at age 18 without conditioning on the SBP and DBP levels at age 10. The effects of race can be seen from the observation that the

African American girls always have higher estimated $\text{RPP}_{\text{abnormal BP}}(18|10)$ than the Caucasian girls under the same BP levels, and height and BMI percentiles at age 10. This suggested race effect is worthwhile to be further investigated in other pediatric studies.

We further estimate the RR at ages $(u, v) = (10, 18)$ to quantify the relative strength of the RPP at these ages over the probability of having abnormal SBP or DBP levels at 18 years of age. The girls with high RR values, e.g, significantly greater than 1, can be identified as the ones having high risk of developing abnormal BP levels at young adulthood. Since the BMI is a well-known risk factor for pediatric hypertension (Obarzanek et al., 2010), we estimate the RR values at $(u, v) = (10, 18)$ over a sequence of BMI percentiles $\{0.05, 0.1, \ldots, 0.95\}$ given a fixed height percentile.

Figure 2 shows the lower bounds of the one-sided 95% confidence intervals (CI) for the RRs of African American and Caucasian girls conditioning on the medium height and the 75th SBP and DBP quantiles at age 10. For both African American and Caucasian girls, the lower bounds of the CIs increase linearly as the BMI percentile increases. Except for the Caucasian girls with BMI percentiles below 25, the lower CI bounds of the RRs are all greater than one for both races. This suggests that the majority of the girls within the given hight and BP range have higher probabilities of developing abnormal SBP or DBP levels at age 18. Similar phenomena are also observed for the RRs and their corresponding one-sided CIs under various other scenarios of covariate values and BP levels at age 10, e.g., the girls with medium height, and SBP and DBP values at their medians and 90th quantiles.

4. Simulation Study

In order to validate that the proposed method performs properly for analyzing the NGHS data, the simulation setups reflect the NGHS design. We generate longitudinal data of 1000 subjects from the following bivariate models for $j = 1, \ldots, 10$:

$$Y_{ij1}(t_{ij}) = \alpha_0(t_{ij}) + \alpha_1(t_{ij})X_{ij1} + \alpha_2(t_{ij})X_{ij2}(t_{ij}) + \alpha_3(t_{ij})X_{ij3}(t_{ij}) + e_i + \epsilon_{ij}, \quad (4.1)$$

$$Y_{ij2}(t_{ij}) = \beta_0(t_{ij}) + \beta_1(t_{ij})X_{ij1} + \beta_2(t_{ij})X_{ij2}(t_{ij}) + \beta_3(t_{ij})X_{ij3}(t_{ij}) + e_i + \varepsilon_{ij}. \quad (4.2)$$

We independently generate X_{ij1}, $X_{ij2}(t_{ij})$, $X_{ij3}(t_{ij})$, e_i, and $(\epsilon_{ij}, \varepsilon_{ij})^T$ each other as follows: $X_{ij1} \sim$ Bernoulli(0.5), $100X_{ij2}(t_{ij}) = a_i + \xi_{ij}$, $a_i \sim U(5, 95)$, $\xi_{ij} \sim U(-5, 5)$, $100X_{ij3}(t_{ij}) = b_i + \varphi_{ij}$, $b_i \sim U(5, 95)$, $\varphi_{ij} \sim U(-5, 5)$, $t_{ij} \sim U(j + 8, j + 9)$, $e_i \sim N(0, 6^2)$, and $(\epsilon_{ij}, \varepsilon_{ij})^T$ are generated from the bivariate normal distribution with zero means, $Var(\epsilon_{ij}) = 36$, $Var(\varepsilon_{ij}) = 64$ and $Cov(\epsilon_{ij}, \varepsilon_{ij}) = 14$. In addition, the coefficients are set as: $\alpha_0(t) = 72 + 3t - 0.07t^2$, $\alpha_1(t) = -0.1 + 0.06t$, $\alpha_2(t) = -3 + 1.3t - 0.03t^2$, $\alpha_3(t) = 4 + 1.1\cos(\pi t/6) - 0.3\sin(\pi t/6)$, $\beta_0(t) = 15 + 5.27t - 0.15t^2$, $\beta_1(t) = 1 - 0.1t + 0.007t^2$, $\beta_2(t) = 23 - 3t + 0.11t^2$, $\beta_3(t) = 3 + 0.85\cos(\pi t/6) - 0.42\sin(\pi t/6)$.

Note that Y_{ij1}, Y_{ij2} X_{ij1}, X_{ij2}, and X_{ij3} approximate SBP, DBP, race, BMI, and height percentiles in the NGHS data, respectively, and the coefficients are set based on the estimates obtained from fitting the models (4.1) and (4.2) to the NGHS data. The within-subject correlation is imposed by using subject errors e_i and the correlation between the bivariate response variables is also considered by using bivariate normal errors $(\epsilon_{ij}, \varepsilon_{ij})$. We note that the conditional distribution $(Y_1(v), Y_2(v)|Z(u))$ is not appropriate to use for generating ordinary longitudinal data $(X(t), Y_1(t), Y_2(t))$, that are measured concurrently.

For 1000 simulation replicates, we estimate the the same RPP in (3.1) considered in NGHS data analysis:

$$1 - \text{RPP}\{(-\infty, Q_{.95}\{Y_1(18)\}), (-\infty, Q_{.95}\{Y_2(18)\})|Z(10) = z(10)\},$$

where $z(10) = (x_1, x_2(10), x_3(10), y_1(10), y_2(10))$. We evaluate the performance of the proposed method by computing the RPP for $x_1 = 0, 1$, $x_2(10) = 0.05, \ldots, 0.95$, and $x_3(10) = 0.5$ with $(y_1(10), y_2(10)) = (Q_{0.5}\{Y_1(10)\}, Q_{0.5}\{Y_2(10)\})$. Since the quantiles of $(y_1(t), y_2(t))$ are unknown, they are estimated in the same manner as the NGHS data analysis using subjects whose age is in $[t - 0.5, t + 0.5)$. For estimation of the RPP given $z(10)$, stratified quantile regression models with both orders of the bivariate response variables with a random sample of 1000 are used.

We remark that the true value of RPP is infeasible to be obtained because data are generated based on models (4.1)–(4.2) while the stratified quantile regression models (2.4)-(2.5) are considered to obtain the RPP. Alternatively we generate a sufficiently large number of subjects (e.g. $n = 1,000,000$) and evaluate the true RPP without imposing any structure between the bivariate response variable at time 18 and predictors at time 10.

Figure 2 displays the unstructured RPP curves and the average, 2.5% percentiles and 97.5% percentiles of the estimated RPP curves. It can be shown that the average of the estimated RPP curves by the proposed method is quite close to the unstructured RPP curves. Moreover, the variation of estimated RPP curves is reasonably small even though relatively smaller longitudinal data is used compared to the NGHS data. Therefore, these simulation results validate that the proposed estimator of the RPP is consistent and that the estimated RPPs for analyzing NGHS data are reliable.

5. Discussion

The statistical methodology and its application to the NGHS data studied in this paper provide a useful exploratory tool for evaluating the dynamic relationship with multivariate longitudinal data. We propose the RPP and its functional as a natural and direct means to quantify the past information on the likelihood of future health status and disease risks. The dynamic quantile regressions presented in Section 2 lead to a class of novel and flexible structured nonparametric models for computing the RPP and its functional. This conditional quantile based approach can be applied to a wide range of biomedical studies where the scientific objectives are to discover the factors that have a significant influence on the future disease risks. In practice, statistical estimates and inferences of the RPP and its functional can be used to identify the individuals who are more likely than the general population to develop unfavorable disease risks.

In some situations, it is possible to consider a "risk score" which combines several risk factors into a single univariate outcome variable. For example, one of the cardiovascular disease risk factors evaluated in Redheuil et al. (2014) is the mean brachial blood pressure (MBP), which is defined as $(2\text{DBP}+\text{SBP})/3$. However, as discussed in Redheuil et al. (2014), the MBP is only one of the many risk factors to be investigated in cardiovascular studies, and it is by no means a unique substitute for the bivariate (SBP, DBP). For the NGHS analysis of hypertension in adolescent girls, clinical implications of abnormal levels of blood pressure is discussed in Obarzanek et a. (2010) using joint distributions of (SBP, DBP) conditional on age, height, BMI and other covariates.

The RPP and the conditional quantile models proposed in this paper differ from the conditional distribution based "rank-tracking probabilities" (RTP) in the literature (e.g., Wu and Tian, 2013a, 2013b) in three important aspects. First, our RPP and conditional quantile models allow

for longitudinal data with bivariate outcome variables while the RTP based methods can only be applied to univariate outcomes. Second, unlike the RTP, our RPP defined at any time points $u < v$ allows for any given outcome and covariate values at the previous time u while the RTP requires the outcome at u to belong to some prespecified set of "events." Third, our conditional quantile regression models allow for dynamic dependence of outcomes and covariates at both time points $u < v$ while the currently available conditional distribution based models for RTP do not allow the outcomes and covariates simultaneously appear at both time points. These three unique features enable our conditional quantile based RPP and its functional to be more generally applied than the RTP based methods in the literature.

Compared with the nonparametric estimation methods for conditional based models (Wu and Tian, 2013a, 2013b), the simulation and kernel smoothing estimation procedure proposed in this paper appear to be very unique because of the inclusion of the simulation step. This simulation step is appropriate and essential because the current modeling framework is based on the conditional quantiles. The application to the NGHS data suggests that our proposed models and estimation methods lead to clinical conclusions that are consistent with the findings observed in the literature. The statistical properties established by the simulation study and the asymptotic development suggest that our simulation and kernel smoothing based estimation methods lead to consistent results.

Since the proposed estimation of the conditional distribution of bivariate outcomes involves two-dimensional kernel estimation, a sufficient number of observations are required to obtain reliable estimation results. In order to check how sensitive the estimation of RPP is in terms of a sample size, we perform the simulation studies with the number of subjects $n = 1000$, which is substantially smaller than NHGS data $(n = 2376)$, and the simulation results suggest the proposed estimator is reliable and consistent. We further note that restructuring longitudinal data helps to

overcome the bidimensional problem by increasing the number of observations. When both the response variables and covariates are concurrently measured, the number of observations in the restructured longitudinal data is $nm(m-1)/2$ where $m = m_1 = \cdots = m_n$. This becomes much larger than the number of observations nm in the original longitudinal data as m increases.

References

Barter, P., Gotto, A. M., LaRosa, J. C., Maroni, J., Szarek, M., Grundy, S. M, John, J. P., Kastelein, Bittner, V. and Fruchart, J. C. (2007) "HDL cholesterol, very low levels of LDL cholesterol, and cardiovascular events," *New England Journal of Medicine*, 357, 1301–1310.

Cho, H. (2016), "The analysis of multivariate longitudinal data using multivariate marginal models," *Journal of Multivariate Analysis*, 143, 481–491.

Cho, H., Hong H. G. and, Kim, M. O. (2016). "Efficient quantile marginal regression for longitudinal data with dropouts," *Biostatistics*, 17, 561–575.

Chobanian, A. A. V., Bakris, G. L., Black, H. R., Cushman, W. C., Green, L. A., Izzo, J. L., Jones, D. W., Materson, B. J., Oparil, S., Wright, J. T. and others (2003). "Seventh report of the joint national committee on prevention, detection, evaluation, and treatment of high blood pressure," *Hypertension*, 42, 1206–1252.

Chaganty, N. R. and Naik, D. N. (2002), "Analysis of multivariate longitudinal data using quasi-least squares," *Journal of Statistical Planning and Inference*, 103, 421–436.

Falkner, B. Daniels, S. R., Flynn, J. T. and Gidding, S. and Green, L. A., Ingelfinger, J. R., Lauer, R. M. and Morgenstern, B. Z., Portman, R. J., Prineas, R. J., Rocchini, A. P., Rosner B.,

Sinaiko, A. R., Stettler, N., Urbina E., Roccella, E. J., Hoke T., Hunt, C. E., Pearson G., Karimbakas, J. and Horton, A. (2004), "The fourth report on the diagnosis, evaluation, and treatment of high blood pressure in children and adolescents," *Pediatrics*, 114, 555–576.

Fieuws, S. and Verbeke, G. (2006), "Pairwise fitting of mixed models for the joint modeling of multivariate longitudinal profiles," *Biometrics*, 62, 424–431.

Flynn, J. T., Kaelber, D. C., Baker-Smith, c. M., et al. (2017), "Clinical practice guideline for screening and management of high blood pressure in children and adolescents," *Pediatrics*, 140(3):e20171904.

Ghosh, P., Branco, M. D. and Chakraborty, H. (2007), "Bivariate random effect model using skew-normal distribution with application to HIV-RNA," *Statistics in Medicine*, 26, 1255–1267.

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.

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.

Kim, C. and Zimmerman, D. L. (2012), "Unconstrained models for the covariance structure of multivariate longitudinal data," *Journal of Multivariate Analysis*, 107, 104–118.

Kim, M. O. and Yang, Y. (2011). "Semiparametric approach to a random effects quantile regression model," *Journal of the American Statistical Association*, 106, 1405–1417.

Kohlia, P., Garcia, T. P. and Pourahmadi, M. (2016), "Modeling the Cholesky factors of covariance matrices of multivariate longitudinal data," *Journal of Multivariate Analysis*, 145, 87–100.

Kürüm, E., Hugues, J., Li, R. and Shiffman, S. (2018), "Time-varying copula models for longitudinal data," *Statistics and Its Interface*, 11, 203–221.

Kwak, M. (2017), "Estimation and inference on the joint conditional distribution for bivariate longitudinal data using Gaussian copula," *Journal of the Korean Statistical Society*, 46, 349–364.

Kwak, M. (2017), "Estimation and inference of the joint conditional distribution for multivariate longitudinal data using nonparametric copulas," *Journal of Nonparametric Statistics*, 29, 491–514.

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," Journal of Pediatrics, 157, 461–467.

Redheuil A., Wu, C. O., Kachenoura, N., Ohyama, Y., Yan, R. T., Bertoni, A. G., Hundley, G. W., Duprez, D. A., Jacobs, D. R. and Daniels, L. B., Darwin, C., Sibley, C., Bluemke, D. A. and Lima, J. (2014), "Proximal aortic distensibility is an independent predictor of all-cause mortality and incident cardiovascular events in the Multi-Ethnic Study of Atherosclerosis," *Journal of American College of Cardiology*, 64, 261–2629.

Rice, J.A. and Silverman, B.W. (1991), "Estimating the mean and covariance structure nonparametrically when the data are curves," *Journal of the Royal Statistical Society, Series B*, 53,

233–243.

Rochon, J. (1996), "Analyzing bivariate repeated measures for discrete and continuous outcome variables," *Biometrics*, 52, 740–750.

Thiébaut, R., Jacqmin-Gadda, H., Chêne, G., Leport, C. and Commenges, D. (2002), "Bivariate linear mixed models using SAS proc MIXED," *Computer Methods and Programs in Biomedicine*, 69, 249–256.

Thiébaut, R., Jacqmin-Gadda, H., Babiker, A., Commenges, D. and the Cascade Collaboration (2005), "Joint modelling of bivariate longitudinal data with informative dropout and left-censoring, with application to the evolution of CD4+ cell count and HIV RNA viral load in response to treatment of HIV infection," *Statistics in Medicine*, 24, 65–82.

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.

Tian, X., and Wu, C. O. (2014), "Estimation of rank-tracking probabilities using nonparametric mixed-effects models for longitudinal data," *Statistics and Its Interface*, 7, 87–99.

Verbeke, G., Fieuws, S., Molenberghs, G. and Davidian M. (2014), "The analysis of multivariate longitudinal data: A review," *Statistical Method in Medical Research*, 23, 42–59.

Wei, Y. (2008), "An approach to multivariate covariate-dependent quantile contours with application to bivariate conditional growth charts," *Journal of the American Statistical Association*, 103, 397–409.

Wilsgaard, T. and Jacobsen, B. K. and Schirmer, H. and Thune, I. and Løchen, M.-L. and Njølstad, I. and Arnesen, E. (2001), "Tracking of cardiovascular risk factors: the Tromsø study, 1979–1995," *American journal of epidemiology*, 154, 418–426.

Wu, C. O. and Tian, X. (2013a), "Nonparametric estimation of conditional distributions and rank-tracking probabilities with time-varying transformation models in longitudinal studies," *Journal of the American Statistical Association*, 108, 971–982.

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

Wu, C. O. and Tian, X. (2018), *Nonparametric Models for Longitudinal Data: With Implementation in R*. Chapman & Hall/CRC Press, Monographs on Statistics & Applied Probability 159, Boca Raton.

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.

Xiang, D., Qiu, P. and Pu, X. (2013), "Nonparametric regression analysis of multivariate longitudinal data," *Statistica Sinica*, 23, 769–89.

Xu, J. and Mackenzie, G. (2012), "Modeling covariance structure in bivariate marginal models for longitudinal data," *Biometrika*, 99, 649–662.

Yu, K. and Jones, M.C. (1998), "Local linear quantile regression," *Journal of the American Statistical Association*, 93, 228–237.

Appendix

Proof of Theorem 1. Let $u_{ij} = (t_{ij} - t_1)/b_1$, $v_{ij'} = (t_{ij'} - t_2)/b_2$, $K_{ijj'} = K(u_{ij})K(v_{ij'})$,

$$\Delta = \begin{bmatrix} \Delta_\theta \\ \Delta_\theta^* \\ \Delta_\theta^\# \end{bmatrix} = \sqrt{Nb_1b_2} \begin{bmatrix} \theta - \theta_\tau(t_2|t_1) \\ b_1\{\theta^* - \frac{\partial\theta_\tau(t_2|t_1)}{\partial t_1}\} \\ b_2\{\theta^\# - \frac{\partial\theta_\tau(t_2|t_1)}{\partial t_2}\} \end{bmatrix} \quad \text{and} \quad J_{ijj'} = \begin{bmatrix} Z_{ijj'} \\ Z_{ijj'}u_{ij} \\ Z_{ijj'}v_{ij'} \end{bmatrix}.$$

Then we can write

$$Y_{2,ij'} - Z_{ijj'}^T\theta - Z_{ijj'}^T\{\theta^*(t_{ij} - t_1) + \theta^\#(t_{ij'} - t_2)\} = \xi_{ijj'} + d_{ijj'} - J_{ijj'}^T\Delta/\sqrt{Nb_1b_2}$$

where $d_{ijj'} = Z_{ijj'}^T\{\theta_\tau(t_{ij'}|t_{ij}) - \theta_\tau(t_2|t_1) - (t_{ij} - t_1)\partial\theta_\tau(t_2|t_1)/\partial t_1 - (t_{ij'} - t_2)\partial\theta_\tau(t_2|t_1)/\partial t_2\}$. Since $\left(\hat\theta_\tau(t_2|t_1), \hat\theta_\tau^*(t_2|t_1), \hat\theta_\tau^\#(t_2|t_1)\right)$ minimizes (2.8), the re-scaled vector $\hat\Delta$ minimizes the re-parameterized function of Δ:

$$L(\Delta) = \sum_{i=1}^{n}\sum_{j=1}^{m_i-1}\sum_{j'=j+1}^{m_i}\left\{\rho_\tau\left(\xi_{ijj'} + d_{ijj'} - J_{ijj'}^T\Delta/\sqrt{Nb_1b_2}\right) - \rho_\tau\left(\xi_{ijj'} + J_{ijj'}^T d_{ijj'}\right)\right\}K_{ijj'}.$$

Write $\delta_{ijj'} = J_{ijj'}^T\Delta/\sqrt{Nb_1b_2}$. Applying Knight's identity $\rho_\tau(u - \theta) - \rho_\tau(u) = -\theta(\tau - \mathbf{1}_{u<0}) + \int_0^\theta(\mathbf{1}_{u\le s} - \mathbf{1}_{u\le 0})ds$, we can write $\mathcal{L}(\Delta) = -A_n\Delta + I_n$, where

$$\begin{aligned} A_n &= \frac{1}{\sqrt{Nb_1b_2}}\sum_{i=1}^{n}\sum_{j=1}^{m_i-1}\sum_{j'=j+1}^{m_i}(\tau - \mathbf{1}_{d_{ijj'}+\xi_{ijj'}<0})K_{ijj'}J_{ijj'}^T, \\ I_n &= \sum_{i=1}^{n}\sum_{j=1}^{m_i-1}\sum_{j'=j+1}^{m_i}K_{ijj'}\eta_{ijj'}, \quad \eta_{ijj'} = \int_0^{\delta_{ijj'}}(\mathbf{1}_{d_{ijj'}+\xi_{ijj'}\le s} - \mathbf{1}_{d_{ijj'}+\xi_{ijj'}\le 0})ds. \end{aligned}$$

Consider I_n. Since K has bounded support, it suffices to consider $|t_{ij} - t_1| = O(b_1)$ and $|t_{ij'} -$

$t_2| = O(b_2)$. By Condition 1, $|\delta_{ijj'}| = O_p\{(Nb_1b_2)^{-1/2}\}$ and $|d_{ijj'}| = O_p(b_1^2 + b_2^2)$. Since $|\eta_{ijj'}| \leq |\delta_{ijj'}|\mathbf{1}_{-|\delta_{ijj'}|\leq\xi_{ijj'}+d_{ijj'}\leq|\delta_{ijj'}|}$, we have $\mathbb{E}(K_{ijj'}^2\eta_{ijj'}^2) = O(\rho_n/N)$, where $\rho_n = 1/\sqrt{Nb_1b_2} + b_1^2 + b_2^2$. By the Cauchy-Schwarz inequality, we have

$$\begin{aligned}\mathrm{var}(I_n) &= \sum_{i=1}^{n}\mathrm{var}\left(\sum_{j=1}^{m_i-1}\sum_{j'=j+1}^{m_i}K_{ijj'}\eta_{ijj'}\right) \leq \sum_{i=1}^{n}\left[\frac{m_i(m_i-1)}{2}\sum_{j=1}^{m_i-1}\sum_{j'=j+1}^{m_i}\mathbb{E}(K_{ijj'}^2\eta_{ijj'}^2)\right] \\ &= O\left(\sum_{i=1}^{n}m_i^4\rho_n/N\right) = O\left\{\sum_{i=1}^{n}m_i^4\Big(\frac{1}{\sqrt{N^3b_1b_2}} + \frac{b_1^2+b_2^2}{N}\Big)\right\} \to 0, \qquad (5.3)\end{aligned}$$

in view of Condition 2. By $d_{ijj'} = O_p(b_1^2 + b_2^2)$ and simple Taylor's expansion,

$$\mathbb{E}(\eta_{ijj'}|Z_{ij}, t_{ij}, t_{ij'}) = \int_0^{\delta_{ijj'}}\Big[F_\xi(s - d_{ijj'}) - F_\xi(-d_{ijj'})\Big]ds \asymp \delta_{ijj'}^2\frac{f_\xi(0)}{2}, \qquad (5.4)$$

uniformly for all (i, j, j'). Recall $\Gamma_Z(t_1, t_2) = \mathbb{E}[Z(t_1, t_2)Z(t_1, t_2)^T]$. Note that

$$\sum_{i=1}^{n}\sum_{j=1}^{m_i-1}\sum_{j'=j+1}^{m_i}\mathbb{E}(K_{ijj'}\delta_{ijj'}^2) \to p(t_1, t_2)\Delta^T\Omega(t_1, t_2)\Delta, \qquad (5.5)$$

where $\Omega(t_1, t_2) = \mathrm{diag}\{\Gamma_Z(t_1, t_2), \mu_K\Gamma_Z(t_1, t_2), \mu_K\Gamma_Z(t_1, t_2)\}$ is a block diagonal matrix. By (5.3)–(5.5), we have the convergence in probability:

$$\begin{aligned}I_n = \mathbb{E}(I_n) + o_p(1) &= \sum_{i=1}^{n}\sum_{j=1}^{m_i-1}\sum_{j'=j+1}^{m_i}\mathbb{E}[K_{ijj'}\mathbb{E}(\eta_{ijj'}|Z_{ij}, t_{ij}, t_{ij'})] + o_p(1) \\ &\to \frac{f_\xi(0)}{2}p(t_1, t_2)\Delta^T\Omega(t_1, t_2)\Delta.\end{aligned}$$

Recall $\hat{\Delta} = \text{argmin}_{\Delta}\, \mathcal{L}(\Delta)$. By the quadratic approximation and convexity lemma,

$$\begin{aligned}\hat{\Delta} &= \underset{\Delta}{\text{argmin}} \Big\{ - A_n\Delta + \frac{f_\xi(0)}{2} p(t_1,t_2)\Delta^T\Omega(t_1,t_2)\Delta \Big\} + o_p(1) \\ &= \frac{\Omega(t_1,t_2)^{-1}A_n^T}{p(t_1,t_2)f_\xi(0)} + o_p(1).\end{aligned}$$

For the $\hat{\theta}$ components of $\hat{\Delta}$, we have

$$\begin{aligned}&\hat{\theta}_\tau(t_2|t_1) - \theta_\tau(t_2|t_1) \\ &= \frac{\Gamma_Z^{-1}(t_1,t_2)}{p(t_1,t_2)f_\xi(0)} \frac{1}{Nb_1b_2} \sum_{i=1}^{n} \sum_{j=1}^{m_i-1} \sum_{j'=j+1}^{m_i} (\tau - \mathbf{1}_{\xi_{ijj'}<0} + \zeta_{ijj'}) K_{ijj'} Z_{ij} + o_p[(Nb_1b_2)^{-1/2}], \quad (5.6)\end{aligned}$$

where $\zeta_{ijj'} = \mathbf{1}_{\xi_{ijj'}<0} - \mathbf{1}_{d_{ijj'}+\xi_{ijj'}<0}$. Let $R_n = (Nb_1b_2)^{-1} \sum_{i=1}^{n} \sum_{j=1}^{m_i-1} \sum_{j'=j+1}^{m_i} \zeta_{ijj'} K_{ijj'} Z_{ij}$. By the arguments in (5.3)–(5.4) and Taylor's expansion $d_{ijj'} = \{b_1^2 u_{ij}^2 \partial^2\theta_\tau(t_2|t_1)/\partial t_1^2 + b_2^2 v_{ij'}^2 \partial^2\theta_\tau(t_2|t_1)/\partial t_2^2 + b_1b_2u_{ij}v_{ij'}\partial^2\theta_\tau(t_2|t_1)/(\partial t_1\partial t_2)\}/2 + O(b_1^3 + b_2^3)$,

$$\begin{aligned}\mathbb{E}(R_n) &= \frac{1}{Nb_1b_2} \sum_{i=1}^{n} \sum_{j=1}^{m_i-1} \sum_{j'=j+1}^{m_i} \mathbb{E}\{K_{ijj'}Z_{ij}\mathbb{E}(\zeta_{ijj'}|Z_{ij}, t_{ij}, t_{ij'})\} \\ &= \frac{p(t_1,t_2)f_\xi(0)\mu_K}{2}\Gamma_Z(t_1,t_2)\Big(\frac{\partial^2\theta_\tau(t_2|t_1)}{\partial t_1^2}b_1^2 + \frac{\partial^2\theta_\tau(t_2|t_1)}{\partial t_2^2}b_2^2\Big) + o(b_1^3 + b_2^3), \quad (5.7)\end{aligned}$$

and $\text{var}(R_n) = o\{(Nb_1b_2)^{-1/2}\}$. Note that $b_1^3 + b_2^3 = o\{(Nb_1b_2)^{-1/2}\}$. Thus, by (5.6)–(5.7),

$$\begin{aligned}&\sqrt{Nb}\Big\{\hat{\theta}_\tau(t_2|t_1) - \theta_\tau(t_2|t_1) - \frac{\mu_K}{2}\Big(\frac{\partial^2\theta_\tau(t_2|t_1)}{\partial t_1^2}b_1^2 + \frac{\partial^2\theta_\tau(t_2|t_1)}{\partial t_2^2}b_2^2\Big)\Big\} \\ &= \frac{\Gamma_X^{-1}(t_1,t_2)}{p(t_1,t_2)f_\xi(0)} \frac{1}{\sqrt{Nb_1b_2}} \sum_{i=1}^{n} \varrho_i + o_p(1), \quad (5.8)\end{aligned}$$

where $\varrho_i = \sum_{j=1}^{m_i-1} \sum_{j'=j+1}^{m_i} \varrho_{ijj'}$ with $\varrho_{ijj'} = [\tau - \mathbf{1}_{\xi_{ijj'}<0}]K_{ijj'}Z_{ij}$. Note that $\mathbb{E}(\varrho_{ijj'}\varrho_{i\ell\ell'}^T) =$

$O(b_1^2 b_2^2)$ for $j \neq \ell$ and $j' \neq \ell'$, $\mathbb{E}(\varrho_{ijj'}\varrho_{i\ell\ell'}^T) = O(b_1 b_2^2)$ for $j = \ell$ and $j' \neq \ell'$, and $\mathbb{E}(\varrho_{ijj'}\varrho_{i\ell\ell'}^T) = O(b_1^2 b_2)$ for $j \neq \ell$ and $j' = \ell'$. Thus,

$$\begin{aligned}
\mathrm{var}\Big(\frac{1}{\sqrt{Nb_1b_2}}\sum_{i=1}^{n}\varrho_i\Big) &= \frac{1}{Nb_1b_2}\sum_{i=1}^{n}\sum_{j=1}^{m_i-1}\sum_{j'=j+1}^{m_i}\mathbb{E}\{[\tau-\mathbf{1}_{\xi_{ijj'}<0}]^2K_{ijj'}^2Z_{ij}Z_{ij}^T\} \\
&+ \frac{1}{Nb_1b_2}\sum_{i=1}^{n}O(m_i^3b_1^2b_2)+\frac{1}{Nb_1b_2}\sum_{i=1}^{n}O(m_i^3b_1b_2^2)+\frac{1}{Nb_1b_2}\sum_{i=1}^{n}O(m_i^4b_1^2b_2^2) \\
&\to \tau(1-\tau)p(t_1,t_2)\varphi_K^2\Gamma_Z(t_1,t_2). \qquad (5.9)
\end{aligned}$$

The desired result then easily follows from (5.8) and the independence of $\varrho_1, \ldots, \varrho_n$. ♢

Department of Statistics, Miami University, Oxford, OH 45056

E-mail: (kims20@miamioh.edu)

Department of Biostatistics, University of Iowa, Iowa City, IA 52242

E-mail: (hyunkeun-cho@uiowa.edu)

Office of Biostatistics Research, National Heart, Lung and Blood Institute, Bethesda, MD 20892

E-mail: (wuc@nhlbi.nih.gov)

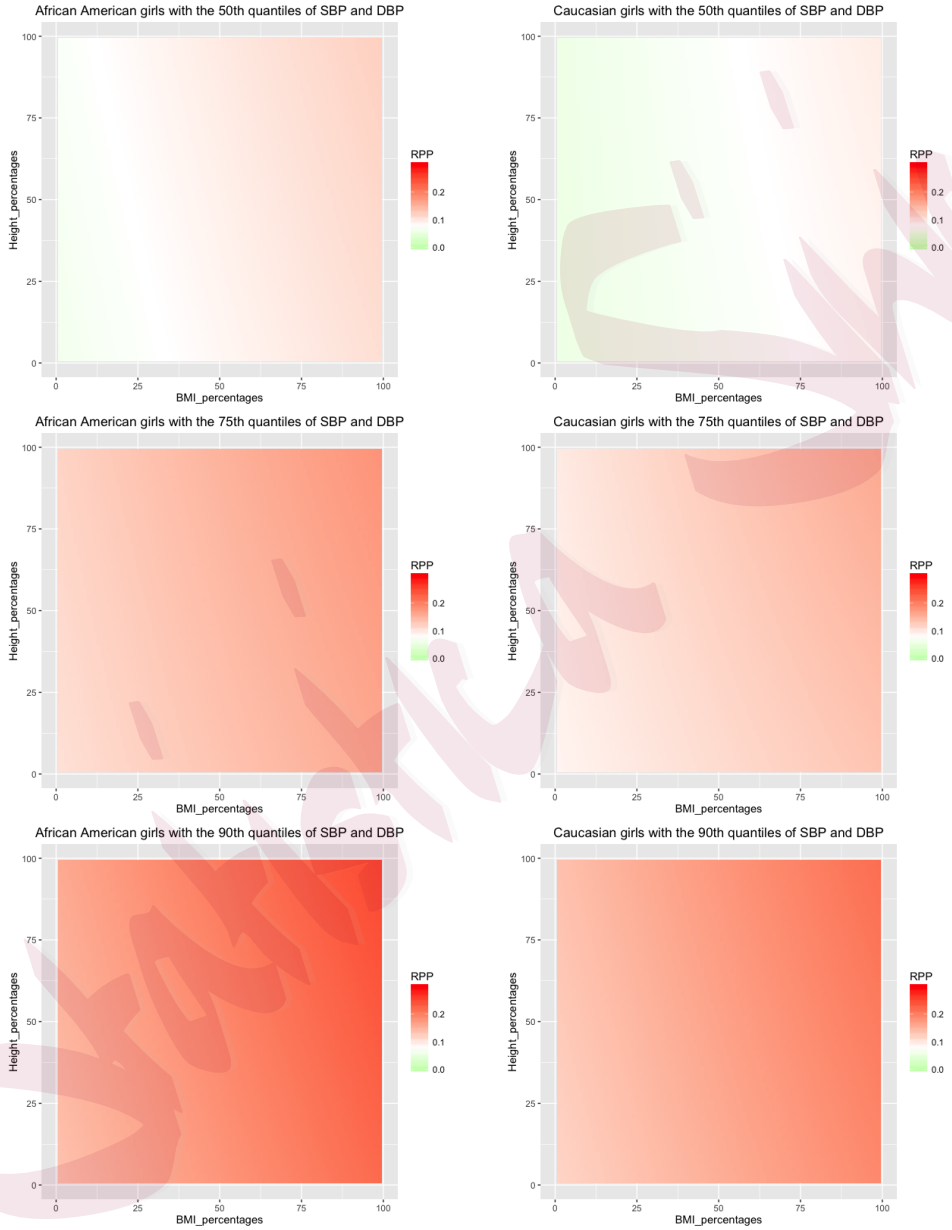

Figure 1: Heat maps of the estimated RPP of having hypertension at age 18.

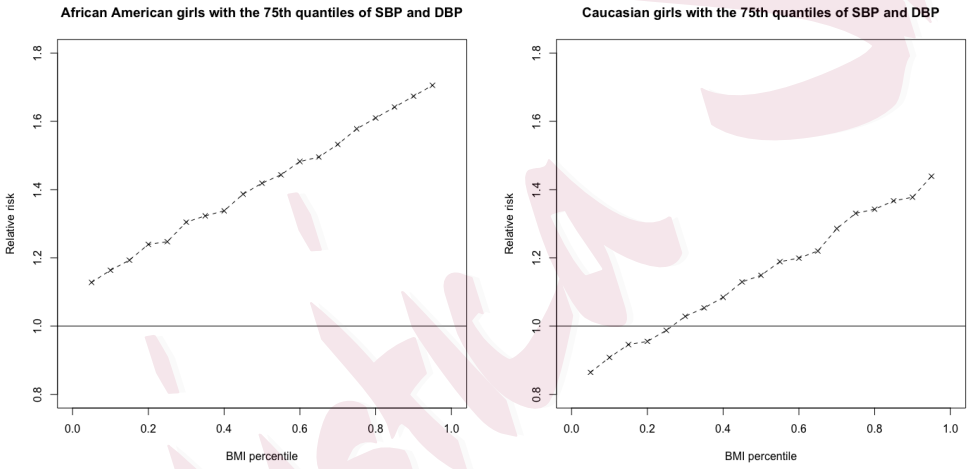

Figure 2: Lower bounds, marked with x, of one-sided 95% confidence intervals for the relative risk of RPP with the 50th height percentile and the 75th quantile of SBP and DBP at age 10.

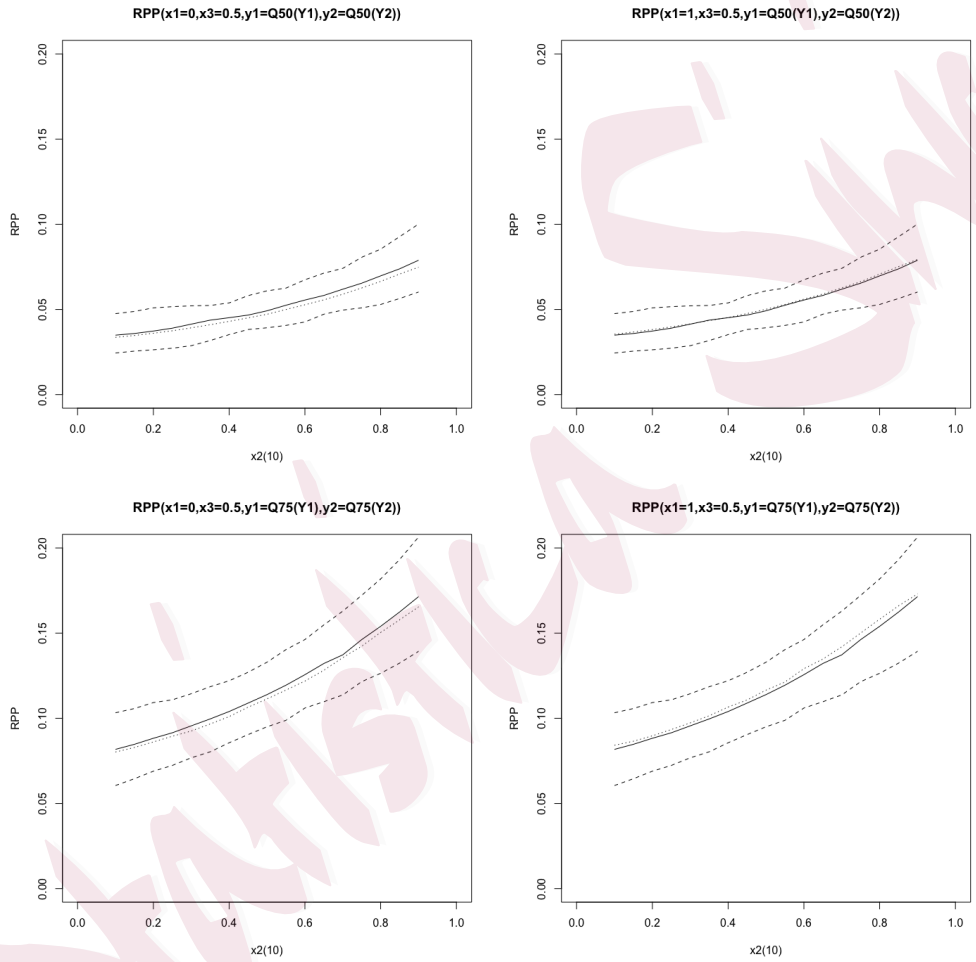

Figure 3: The solid lines are the average of the estimated RPP curves; the dashed lines are the pointwise 2.5% and 97.5% percentile of the estimated RPP curves; the dotted lines are the structured RPP curves.