 DOI: 10.1111/biom.12356

Merging Multiple Longitudinal Studies with Study-Specific Missing Covariates: A Joint Estimating Function Approach

Fei Wang,[1,*] Peter X.-K. Song,[2,] and Lu Wang[2,***]**

[1]Global Analytics, Ford Motor Credit, Dearborn, Michigan 48126, U.S.A.
[2]Department of Biostatistics, University of Michigan, Ann Arbor, Michigan 48109, U.S.A.
*email: fwang55@ford.com
**email: pxsong@umich.edu
***email: luwang@umich.edu

Summary. Merging multiple datasets collected from studies with identical or similar scientific objectives is often undertaken in practice to increase statistical power. This article concerns the development of an effective statistical method that enables to merge multiple longitudinal datasets subject to various heterogeneous characteristics, such as different follow-up schedules and study-specific missing covariates (e.g., covariates observed in some studies but missing in other studies). The presence of study-specific missing covariates presents great statistical methodology challenge in data merging and analysis. We propose a joint estimating function approach to addressing this challenge, in which a novel nonparametric estimating function constructed via splines-based sieve approximation is utilized to bridge estimating equations from studies with missing covariates to those with fully observed covariates. Under mild regularity conditions, we show that the proposed estimator is consistent and asymptotically normal. We evaluate finite-sample performances of the proposed method through simulation studies. In comparison to the conventional multiple imputation approach, our method exhibits smaller estimation bias. We provide an illustrative data analysis using longitudinal cohorts collected in Mexico City to assess the effect of lead exposures on children's somatic growth.

Key words: Data merging; Imputation; Meta analysis; Quadratic inference function; Sieve estimation.

1. Introduction

Analyzing combined datasets collected from multiple similar studies has been popular in practice in order to achieve greater power in statistical analysis. When parameters across multiple study populations are common and thus can be estimated using more observations from the merged datasets, performances in both statistical estimation and inference can be improved. In addition, combined data potentially provide richer information to answer some questions that otherwise may not be answered using data from each individual study.

Potentially increased power gained from data merging is subject to additional complexities, one of which is missing covariates considered in this article. Our work is motivated by a cohort study involving multiple longitudinal cohorts gathered in Mexico City (Afeiche et al., 2011). Our analysis concerns two birth cohorts established by the same study team from two hospitals in Mexico City, termed as cohort B and cohort C throughout the article. Two types of lead exposure recorded in the study include mother's blood lead concentration (PBL) and child's cord blood lead concentration (CBL), where the former is fully recorded in both cohorts but the latter is only fully measured in cohort C. One of the primary objectives was to assess the association between CBL and child's somatic growth adjusting for other covariates available. A key challenge in the analysis of merged data from both cohorts pertains to the fact that CBL measurements in cohort B are nearly completely missing and regression coefficients (e.g., the effect of CBL) are different between two cohorts.

Besides the study-specific missing covariates mentioned above, inter-study heterogeneity is another issue often complicating or even impairing the modeling strategy for multiple longitudinal data. For instance, data collected from hospitals located in urban areas might be more volatile than those collected from hospitals located in rural areas because hospitals in cities tend to have more diversified patient populations. Similarly, multi-center clinical trials, even administrated by a common protocol, may still vary in actual operations for data collection, due, for example, to study coordinator's personal effort on retaining patient's follow-up visits. Joint modeling of mean and covariance (e.g., Leng et al., 2010) has been discussed to account for covariance heterogeneity. However, diagnostic tools for covariance models have been little considered in the literature and a mis-specified covariance model can lead to incorrect statistical inference and misleading data analysis. All of these, as a result, may offset the benefit of estimation efficiency from merged data.

Wang et al. (2012) proposed a joint estimating approach to assessing the validity of data merging and to analyzing the merged longitudinal dataset. It is shown that their approach is flexible to handle covariance heterogeneity (e.g., different within-subject correlations across cohorts) and provides proper control of type I error in hypothesis testing. However, their method is limited only to the case of fully

observed data and is not applicable to the aforementioned study where measurements of CBL in cohort B are substantially missing. Multiple imputation (Little and Rubin, 2002) technique is a popular approach to handling missing data. Kim (2011) proposed parametric fractional imputation, which uses fractional weights to approximate the observed likelihood. While the imputation approach may be a simple and direct solution to the problem, such a strategy may fail to work properly when the parameter of interest in missing data is different from the one in observed data. Robins et al. (1994), among others, developed various versions of inverse probability weighted (IPW) estimators and augmented IPW (AIPW) estimators to analyze incomplete longitudinal data. So far, IPW, AIPW, and multiple imputation approaches have been mainly developed for a single study. Applying them to the analysis of combined data requires nontrivial statistical work, especially when the merged dataset involves study-specific missing covariates. Chen and Ibrahim (2006), Shi et al. (2009), among others, proposed different approaches for missing covariates in parametric regression, but these methods are all constructed under selection model (Little, 1993) and thus are not able to be applied directly in our situation.

We propose a new estimating function approach to analyzing merged data from multiple studies with study-specific missing covariates. The novelty of our method lies in the idea of joining study-specific estimating functions, instead of directly joining multiple datasets. In this way, we allow great flexibility to accommodate different covariance structures across studies. Given that it is not feasible to evaluate estimating functions of studies with missing covariates, integrating these estimating functions with respect to missing covariates is inevitable. The resulting estimating functions are then evaluated nonparametrically without assuming specific distributions on covariates. We show that if the study-specific mean models are correctly specified in all individual studies, then under Assumptions 1 and 2 in Section 2 our proposed joint estimating functions are asymptotically unbiased, leading to valid estimation and inference.

Section 2 presents notation and models, followed by estimating procedures in Section 3. In Section 4, we derive the relevant asymptotic properties, and in Section 5 we discuss implementations. After presenting simulation results in Section 6, we illustrate our method by analyzing the motivating data in Section 7. Section 8 contains some discussions, and all technical details and extra simulations are included in the Web Appendix.

2. Model

We consider datasets collected from $K \geq 2$ longitudinal studies with n_k number of subjects in study k, $k = 1, \ldots, K$, and the total number of subjects is $n = \sum_{k=1}^{K} n_k$. Let $D_i \in \{1, \ldots, K\}$ be the study indicator of subject i, and Y_{ij} be the outcome measured for subject i at visit time j, $j = 1, \ldots, m_{D_i}$, and m_{D_i} denotes the number of visits in study D_i for subject $i = 1, \ldots, n$. For the ease of exposition, we assume that subjects in the same study have the same number of repeated measurements in the rest of the article. Let $\boldsymbol{X}_{ij}$ denote a p-dimensional vector of covariates fully observed in all K studies and let $\boldsymbol{Z}_{ij}$ denote a q-dimensional vector of covariates completely observed only in study $k \in \mathcal{S}_o \subset \{1, \ldots, K\}$ and missing in study $k \in \mathcal{S}_m \subset \{1, \ldots, K\}$, where $\mathcal{S}_m \cup \mathcal{S}_o = \{1, \ldots, K\}$. Correspondingly, we let $n_o = \sum_{l \in \mathcal{S}_o} n_l$ denote the number of subjects in studies belonging to $\mathcal{S}_o$. In our motivating example mentioned in the previous section, $\boldsymbol{Z}_{ij}$ represents child's cord blood lead exposure, which is missing in cohort B, and $\boldsymbol{X}_{ij}$ represents mother's blood lead exposure and other covariates which are fully observed for both cohorts. Suppose that the mean of Y_{ij}, given all covariates $\boldsymbol{X}_{ij}$ and $\boldsymbol{Z}_{ij}$ in study k, satisfies the following model:

$$\begin{aligned} \mu_{k,ij} &= E(Y_{ij} \mid \boldsymbol{X}_{ij}, \boldsymbol{Z}_{ij}, D_i = k) \\ &= h(\boldsymbol{X}_{ij}^T \boldsymbol{\beta}_{0,k} + \boldsymbol{Z}_{ij}^T \boldsymbol{\lambda}_{0,k}), \quad k = 1, \ldots, K, \end{aligned} \tag{1}$$

where $h(\cdot)$ is a known link function and $\boldsymbol{\theta}_{0,k} = (\boldsymbol{\beta}_{0,k}^T, \boldsymbol{\lambda}_{0,k}^T)^T$ is the true regression parameter defined in a compact set $\mathcal{B} \subseteq R^{p+q}$. Here, we assume that the true parameters are fully or partially shared across studies. The conditional variance of Y_{ij} in study k takes the form: $\text{var}(Y_{ij} \mid \boldsymbol{X}_{ij}, \boldsymbol{Z}_{ij}, D_i = k) = \phi_k v(\mu_{k,ij})$, where $v(\cdot)$ is a known variance function and ϕ_k is the dispersion parameter.

The missing pattern in this article is similar to one scenario studied by Little (1992) for merging studies with missing covariates, but our case is more complex and more challenging. Here, we consider different regression models for longitudinal studies with missing covariates. Let $\boldsymbol{Y}_i = (Y_{i1}, \ldots, Y_{im_{D_i}})^T$, $\boldsymbol{X}_i = (\boldsymbol{X}_{i1}, \ldots, \boldsymbol{X}_{im_{D_i}})^T$, and similarly $\boldsymbol{Z}_i$, $i = 1, \ldots, n$. To borrow information from studies with fully observed data, we rely on the following missing data mechanism (Rubin, 1976; Little and Zhang, 2011):

ASSUMPTION 1. $D_i \perp \boldsymbol{Z}_i \mid \boldsymbol{X}_i$, for all i.

Assumption 1 implies that given fully observed $\boldsymbol{X}_i$, the study indicator D_i is conditionally independent of missing covariates $\boldsymbol{Z}_i$. This differs from the usual MAR assumption, $D_i \perp \boldsymbol{Z}_i \mid (\boldsymbol{X}_i, \boldsymbol{Y}_i)$. More detailed explanation about this discrepancy can be found in the Web Appendix. Assumption 1 enables us to develop feasible parameter estimation using the following integral as a bridge for studies with missing $\boldsymbol{Z}_i$:

$$\begin{aligned} &f(\boldsymbol{Y}_i \mid \boldsymbol{X}_i, D_i = k \in \mathcal{S}_m) \\ &= \int f(\boldsymbol{Y}_i, \boldsymbol{Z}_i \mid \boldsymbol{X}_i, D_i = k \in \mathcal{S}_m) d\boldsymbol{Z}_i \\ &= \int f_{\boldsymbol{\theta}_{0,k}}(\boldsymbol{Y}_i \mid \boldsymbol{Z}_i, \boldsymbol{X}_i, D_i = k \in \mathcal{S}_m) f(\boldsymbol{Z}_i \mid \boldsymbol{X}_i, D_i \in \mathcal{S}_o) d\boldsymbol{Z}_i. \end{aligned}$$

Assumption 1 and the usual MAR assumption do not typically hold simultaneously unless $D_i \perp (\boldsymbol{Y}_i, \boldsymbol{Z}_i) \mid \boldsymbol{X}_i$ or $(D_i, \boldsymbol{Y}_i) \perp \boldsymbol{Z}_i \mid \boldsymbol{X}_i$. But with the mean model specification of $\boldsymbol{Y}_i \mid (\boldsymbol{X}_i, \boldsymbol{Z}_i, D_i)$ considered in (1), where the regression coefficients are allowed to be different across different studies, both $D_i \perp (\boldsymbol{Y}_i, \boldsymbol{Z}_i) \mid \boldsymbol{X}_i$ and $(D_i, \boldsymbol{Y}_i) \perp \boldsymbol{Z}_i \mid \boldsymbol{X}_i$ are incompatible. Thus, Assumption 1, $D_i \perp \boldsymbol{Z}_i \mid \boldsymbol{X}_i$, is required in this article.

3. Estimation

We propose a quadratic inference function (QIF) approach (Qu et al., 2000; Wang et al., 2012) to estimating all

study-specific regression parameters $\boldsymbol{\theta}_0 = (\boldsymbol{\theta}_{0,1}^T, \ldots, \boldsymbol{\theta}_{0,K}^T)^T$, where subject-level data from multiple studies are accessible. A nonparametric sieve estimation is applied to estimate the unknown function resulted from integration of inference functions with respect to missing covariates. In the following, we focus on an important scenario of $\boldsymbol{Z}_i$ being fully missing in study $k \in \mathcal{S}_m$. The proposed method is also applicable when $\boldsymbol{Z}_i$ is partially missing (i.e., $\boldsymbol{Z}_i$ being observed on some subjects) in study $k \in \mathcal{S}_m$ as long as the missing data mechanism Assumption 1 holds.

3.1. *Conditional Moments*

Firstly note that in model (1), it is not feasible to estimate $\boldsymbol{\theta}_{0,k}$ for $k \in \mathcal{S}_m$ only using data of study k due to the missingness of $\boldsymbol{Z}_i$. So we consider an induced model by integrating the full conditional model (1) with respect to missing covariates $\boldsymbol{Z}_i$. Precisely, let $\eta_k(\boldsymbol{X}_{ij}, \boldsymbol{\theta}_k)$ denote the conditional expectation of $h_k(\boldsymbol{X}_{ij}, \boldsymbol{Z}_{ij}, \boldsymbol{\theta}_k) \equiv h(\boldsymbol{X}_{ij}^T\boldsymbol{\beta}_k + \boldsymbol{Z}_{ij}^T\boldsymbol{\lambda}_k)$ with respect to $\boldsymbol{Z}_{ij}|\boldsymbol{X}_{ij}$ in study $k \in \mathcal{S}_m$. Suppose that the resulting mean model $\eta_k(\boldsymbol{X}_{ij}, \boldsymbol{\theta}_k)$ is a smooth function such that there exists a unique $\boldsymbol{\theta}_{0,k}$ satisfying $\eta_k(\boldsymbol{X}_{ij}, \boldsymbol{\theta}_{0,k}) = E(Y_{ij} \mid \boldsymbol{X}_{ij}, D_i = k)$. In this case, Assumptions 1 and model (1) imply that

$$\begin{aligned}\eta_k(\boldsymbol{X}_{ij}, \boldsymbol{\theta}_{0,k}) &= E\{h_k(\boldsymbol{X}_{ij}, \boldsymbol{Z}_{ij}, \boldsymbol{\theta}_{0,k}) \mid \boldsymbol{X}_{ij}, D_i = k \in \mathcal{S}_m\} \\ &= E\{h_k(\boldsymbol{X}_{ij}, \boldsymbol{Z}_{ij}, \boldsymbol{\theta}_{0,k}) \mid \boldsymbol{X}_{ij}, D_i \in \mathcal{S}_o\}.\end{aligned}$$

This means that the mean model function $\eta_k(\cdot, \boldsymbol{\theta}_{0,k})$ can be estimated by using data from studies in $\mathcal{S}_o$. Similarly, the conditional variance $\nu_k(\boldsymbol{X}_{ij}, \boldsymbol{\theta}_{0,k})$ of Y_{ij} conditioning on $\boldsymbol{X}_{ij}$ in study $k \in \mathcal{S}_m$ can be rewritten as follows:

$$\begin{aligned}\nu_k(\boldsymbol{X}_{ij}, \boldsymbol{\theta}_{0,k}) &= \mathrm{Var}(Y_{ij} \mid \boldsymbol{X}_{ij}, D_i = k \in \mathcal{S}_m) \\ &= \phi_k E[v\{h_k(\boldsymbol{X}_{ij}, \boldsymbol{Z}_{ij}, \boldsymbol{\theta}_{0,k})\} \mid \boldsymbol{X}_{ij}, D_i = k \in \mathcal{S}_m] \\ &\quad + E\{h_k(\boldsymbol{X}_{ij}, \boldsymbol{Z}_{ij}, \boldsymbol{\theta}_{0,k})^2 | \boldsymbol{X}_{ij}, D_i \in \mathcal{S}_o\} - \{\eta_k(\boldsymbol{X}_{ij}, \boldsymbol{\theta}_{0,k})\}^2.\end{aligned}$$

When Y_{ij} follows a normal distribution, $\nu_k(\boldsymbol{X}_{ij}, \boldsymbol{\theta}_{0,k})$ is $\phi_k + Var(\boldsymbol{Z}_{ij} \mid \boldsymbol{X}_{ij}, D_i \in \mathcal{S}_o)\boldsymbol{\lambda}_{0,k}^2$. When Y_{ij} is binary, $\nu_k(\boldsymbol{X}_{ij}, \boldsymbol{\theta}_{0,k})$ is $\eta_k(\boldsymbol{X}_{ij}, \boldsymbol{\theta}_{0,k})\{1 - \eta_k(\boldsymbol{X}_{ij}, \boldsymbol{\theta}_{0,k})\}$. In these two popular cases in practice, $\nu_k(\boldsymbol{X}_{ij}, \boldsymbol{\theta}_{0,k})$, $k \in \mathcal{S}_m$, can be estimated using data from studies in $\mathcal{S}_o$.

Since both $\eta_k(\cdot, \cdot)$ and $\boldsymbol{\theta}_{0,k}$ are unknown in $\eta_k(\boldsymbol{X}_{ij}, \boldsymbol{\theta}_{0,k})$, we cannot uniquely identify $\eta_k(\cdot, \cdot)$ and $\boldsymbol{\theta}_{0,k}$ unless extra restrictions are imposed. Ichimura (1993) refers to $E(Y_{ij} \mid \boldsymbol{X}_{ij}, D_i = k) = \eta_k(\boldsymbol{X}_{ij}, \boldsymbol{\theta}_{0,k})$ as a single-index model. Thus, we postulate Assumption 2 similar to the identification condition in single-index model (Ichimura, 1993; Carroll et al., 1998):

ASSUMPTION 2. *(a) $\theta_l = 1$ for some l, $1 \le l \le K(p+q)$; and (b) $\eta_k(\boldsymbol{X}_{ij}, \boldsymbol{\theta}_k)$ is differentiable and not constant on the support of $\boldsymbol{\theta}_k$ for $k \in \mathcal{S}_m$.*

Assumption 2 (a) restricts $\boldsymbol{\theta}_k$ to be identified uniquely. Assumption 2 (b) eliminates the case where $\eta_k(\boldsymbol{X}_{ij}, \boldsymbol{\theta}_k)$ does not have enough variability to estimate $\boldsymbol{\theta}_{0,k}$. For example, suppose $E(Y_{i,j} \mid X_i, Z_i, D_i = k) = \beta_{k,0} + \beta_{k,1}X_i + \lambda_{k,1}Z_i$, for studies $k = 1, 2$, where Z_i is completely missing in study two. If $E(Z_i \mid X_i) = 0$, then $\eta_2(X_i, \boldsymbol{\theta}_2) = \beta_{2,0} + \beta_{2,1}X_i$ is a constant on the support of $\lambda_{2,1}$. Thus, $\lambda_{2,1}$ cannot be identified. But Assumption 2 precludes this scenario.

3.2. *Estimation with Missing Covariates*

For the ease of exposition, sometimes we suppress covariates in the short-handed notation, for example, we denote $\eta_{k,ij}(\boldsymbol{\theta}_k) \equiv \eta_k(\boldsymbol{X}_{ij}, \boldsymbol{\theta}_k)$, $h_{k,ij}(\boldsymbol{\theta}_k) \equiv h_k(\boldsymbol{X}_{ij}, \boldsymbol{Z}_{ij}, \boldsymbol{\theta}_k)$, and $\nu_{k,ij}(\boldsymbol{\theta}_k) \equiv \nu_k(\boldsymbol{X}_{ij}, \boldsymbol{\theta}_k)$ and so forth. The corresponding vectors for subject i are denoted by $\boldsymbol{\eta}_{k,i}(\boldsymbol{\theta}_k)$, $\boldsymbol{h}_{k,i}(\boldsymbol{\theta}_k)$, and $\boldsymbol{\nu}_{k,i}(\boldsymbol{\theta}_k)$, respectively. Following Newey (1994), we can show that our proposed estimators of the regression coefficients are consistent and asymptotically normal, as long as the plug-in estimator of $\eta_{k,ij}(\boldsymbol{\theta}_k)$ satisfies a convergence rate faster than $n^{-1/4}$. This rate is achievable when $\eta_{k,ij}(\boldsymbol{\theta}_k)$ is sufficiently smooth in $\boldsymbol{x}$, which can be accomplished by the sieve least square method (Newey, 1997). A sieve estimator of $\eta_{k,ij}(\boldsymbol{\theta}_k)$, $k \in \mathcal{S}_m$, takes the form:

$$\hat{\eta}_{k,ij}(\boldsymbol{\theta}_k) \equiv \hat{\eta}_k(\boldsymbol{x}_{ij}, \boldsymbol{\theta}_k) = \sum_{l=1}^{t_{n_k}} a_{k,l}(\boldsymbol{\theta}_k) b_l(\boldsymbol{x}_{ij}) = \boldsymbol{b}(\boldsymbol{x}_{ij})^T \boldsymbol{a}_k(\boldsymbol{\theta}_k),$$

where $\boldsymbol{a}_k(\boldsymbol{\theta}_k) = \{a_{k,1}(\boldsymbol{\theta}_k), \ldots, a_{k,t_{n_k}}(\boldsymbol{\theta}_k)\}^T$ is a vector of unknown coefficients to be estimated, $\boldsymbol{b}(\boldsymbol{x}_{ij}) = \{b_1(\boldsymbol{x}_{ij}), \ldots, b_{t_{n_k}}(\boldsymbol{x}_{ij})\}^T$ are basis functions, and the number of basis functions, t_{n_k}, increases along the increase of sample size n. Estimation of $\boldsymbol{a}_k(\boldsymbol{\theta}_k)$ is carried out by minimizing the following objective function using all studies from $\mathcal{S}_o$:

$$\begin{aligned}\hat{\boldsymbol{a}}_k(\boldsymbol{\theta}_k) = \underset{\boldsymbol{a}_k(\boldsymbol{\theta}_k)}{\arg\min} \sum_{i=1}^{n} \sum_{j=1}^{m_{D_i}} I[D_i \in \mathcal{S}_o] \Big\{ & h_k(\boldsymbol{X}_{ij}, \boldsymbol{Z}_{ij}, \boldsymbol{\theta}_k) \\ & - \boldsymbol{b}(\boldsymbol{X}_{ij})^T \boldsymbol{a}_k(\boldsymbol{\theta}_k) \Big\}^2, \; k \in \mathcal{S}_m,\end{aligned}$$

where $I[A]$ is the indicator function for set A. For subject i in study $D_i = l$, we define the following notation: an $m_l \times t_{n_k}$ matrix $\boldsymbol{W}_i = \{\boldsymbol{b}(\boldsymbol{X}_{i1}), \ldots, \boldsymbol{b}(\boldsymbol{X}_{im_l})\}^T$, a $t_{n_k} \times \sum_{l \in \mathcal{S}_o} n_l m_l$ matrix $\boldsymbol{U}^T = (\boldsymbol{W}_i^T)_{D_i = l \in \mathcal{S}_o}$, and $(\sum_{l \in \mathcal{S}_o} n_l m_l)$-dimensional vector $\boldsymbol{H}_k(\boldsymbol{\theta}_k) = \{\boldsymbol{h}_{k,i}(\boldsymbol{\theta}_k)^T\}_{D_i = l \in \mathcal{S}_o}^T$. Thus, we have $\hat{\boldsymbol{a}}_k(\boldsymbol{\theta}_k) = (\boldsymbol{U}^T\boldsymbol{U})^{-1}\boldsymbol{U}^T\boldsymbol{H}_k(\boldsymbol{\theta}_k)$, and moreover $\hat{\boldsymbol{\eta}}_{k,i}(\boldsymbol{\theta}_k) = \boldsymbol{W}_i\hat{\boldsymbol{a}}_k(\boldsymbol{\theta}_k)$. Correspondingly, the estimated $\partial\hat{\boldsymbol{\eta}}_{k,i}(\boldsymbol{\theta}_k)/\partial\boldsymbol{\theta}_k$ is $\nabla_{\boldsymbol{\theta}_k}\hat{\boldsymbol{\eta}}_{k,i}(\boldsymbol{\theta}_k) = \boldsymbol{W}_i\nabla_{\boldsymbol{\theta}_k}\hat{\boldsymbol{a}}_k(\boldsymbol{\theta}_k)$, where $\nabla_a f(a)$ denotes the gradient vector of function f with respect to a.

Given estimated $\boldsymbol{\eta}_{k,i}(\boldsymbol{\theta}_k)$, we use the quadratic inference function (QIF) to estimate the regression parameter $\boldsymbol{\theta}_k$, which provides great flexibility to account for inter-study heterogeneities. Briefly, QIF begins with an expansion on the inverse of a working correlation matrix for study k of the form: $\boldsymbol{R}_k^{-1}(\alpha_k) \approx \sum_{s=1}^{s_k} \rho_{k,s}\boldsymbol{M}_{k,s}$, where $\rho_{k,1}, \ldots, \rho_{k,s_k}$ are constants possibly dependent on nuisance correlation parameter α_k, and $\boldsymbol{M}_{k,1}, \ldots, \boldsymbol{M}_{k,s_k}$ are known basis matrices with elements 0 and 1 determined by the given working correlation matrix $\boldsymbol{R}_k(\alpha_k)$. Refer to Wang et al. (2012) for more details.

Now denote the estimating function for subject i in study $k \in \mathcal{S}_m$ by $\boldsymbol{g}_{k,i}(\boldsymbol{\theta}_k, \hat{\boldsymbol{\eta}}_{k,i})$, which is expressed with an explicit involvement of $\hat{\boldsymbol{\eta}}_{k,i}$. The same treatment is given to other notation whenever applicable. The extended score vector $\bar{\boldsymbol{g}}_k(\boldsymbol{\theta}_k, \hat{\boldsymbol{\eta}}_k)$

takes the form:

$$\bar{\boldsymbol{g}}_k(\boldsymbol{\theta}_k, \hat{\boldsymbol{\eta}}_k) = \frac{1}{n_k}\sum_{i=1}^{n_k} \boldsymbol{g}_{k,i}(\boldsymbol{\theta}_k, \hat{\boldsymbol{\eta}}_{k,i}) \overset{def.}{=} \frac{1}{n_k}\sum_{i=1}^{n_k} \left\{ \boldsymbol{g}_{k,i,1}^T(\boldsymbol{\theta}_k, \hat{\boldsymbol{\eta}}_{k,i}), \cdots, \boldsymbol{g}_{k,i,s_k}^T(\boldsymbol{\theta}_k, \hat{\boldsymbol{\eta}}_{k,i}) \right\}^T,$$

where for s=1, …, s_k, $\boldsymbol{g}_{k,i,s}(\boldsymbol{\theta}_k, \hat{\boldsymbol{\eta}}_{k,i})$=$\nabla_{\boldsymbol{\theta}_k}\hat{\boldsymbol{\eta}}_{k,i}(\boldsymbol{\theta}_k)^T \boldsymbol{V}_{k,i,s}(\boldsymbol{\theta}_k)\{\boldsymbol{Y}_i - \hat{\boldsymbol{\eta}}_{k,i}(\boldsymbol{\theta}_k)\}$, with $\boldsymbol{V}_{k,i,s}(\boldsymbol{\theta}_k) = \boldsymbol{A}_{k,i}{}^{-1/2}\boldsymbol{M}_{k,s}\boldsymbol{A}_{k,i}{}^{-1/2}$ and $\boldsymbol{A}_{k,i} = \text{diag}\{\nu_{k,i1}(\boldsymbol{\theta}_k), \ldots, \nu_{k,im_k}(\boldsymbol{\theta}_k)\}$. Minimizing a quadratic function $Q_k(\boldsymbol{\theta}_k, \hat{\boldsymbol{\eta}}_k) = n_k \bar{\boldsymbol{g}}_k(\boldsymbol{\theta}_k, \hat{\boldsymbol{\eta}}_k)^T \boldsymbol{C}_k^-(\boldsymbol{\theta}_k, \hat{\boldsymbol{\eta}}_k)\bar{\boldsymbol{g}}_k(\boldsymbol{\theta}_k, \hat{\boldsymbol{\eta}}_k)$, $k \in \mathcal{S}_m$, we have

$$\hat{\boldsymbol{\theta}}_k = \arg\min_{\boldsymbol{\theta}_k} Q_k(\boldsymbol{\theta}_k, \hat{\boldsymbol{\eta}}_k), \tag{2}$$

where $\boldsymbol{C}_k(\boldsymbol{\theta}_k, \hat{\boldsymbol{\eta}}_k)$ is given by $\boldsymbol{C}_k(\boldsymbol{\theta}_k, \hat{\boldsymbol{\eta}}_k) = \frac{1}{n_k}\sum_{i=1}^{n_k} \boldsymbol{g}_{k,i}(\boldsymbol{\theta}_k, \hat{\boldsymbol{\eta}}_{k,i})\boldsymbol{g}_{k,i}(\boldsymbol{\theta}_k, \hat{\boldsymbol{\eta}}_{k,i})^T$. Note that as discussed earlier, estimate for $\nu_{k,ij}(\boldsymbol{\theta}_k)$ is not needed in a linear model or in a logistic model. Even if an estimate of $\nu_{k,ij}(\boldsymbol{\theta}_k)$ is needed (e.g., in a log-linear model), the large sample properties in Section 4 for $\hat{\boldsymbol{\theta}}_k$ still hold, as long as it is a root-n consistent estimator.

3.3. *Joint Estimation with Complete and Incomplete Datasets*

An advantage of performing joint analysis of merged data is to improve estimation efficiency on the regression coefficients across studies (Wang et al., 2012). This property is expected to prevail for our proposed method when some covariates are not observed in some studies. Let $\mathcal{M}_l \subset \{1, \ldots, K\}$, $l = 1, \ldots, p+q$, be the subset of studies within which the l^{th} covariate has a common effect size. The parameter space constrained by all $\mathcal{M}_l$, $l = 1, \ldots, p+q$, is denoted by Ω with $\Omega = \{(\boldsymbol{\theta}_1^T, \ldots, \boldsymbol{\theta}_K^T)^T : \theta_{kl} = \theta_{k'l} \text{ for } \forall\ k \neq k' \in \mathcal{M}_l, l = 1, \ldots, p+q\}$ representing the subspace of parameters restricted under all conditions of common regression coefficients. In study $k \in \mathcal{S}_o$, we define the extended score vector $\bar{\boldsymbol{g}}_k(\boldsymbol{\theta}_k, \boldsymbol{h}_k)$ as

$$\bar{\boldsymbol{g}}_k(\boldsymbol{\theta}_k, \boldsymbol{h}_k) = \frac{1}{n_k}\sum_{i=1}^{n_k} \boldsymbol{g}_{k,i}(\boldsymbol{\theta}_k, \boldsymbol{h}_{k,i}) \overset{def.}{=} \frac{1}{n_k}\sum_{i=1}^{n_k} \begin{pmatrix} \nabla_{\boldsymbol{\theta}_k}\boldsymbol{h}_{k,i}(\boldsymbol{\theta}_k)^T \boldsymbol{V}_{k,i,1}(\boldsymbol{\theta}_k)\{\boldsymbol{Y}_i - \boldsymbol{h}_{k,i}(\boldsymbol{\theta}_k)\} \\ \vdots \\ \nabla_{\boldsymbol{\theta}_k}\boldsymbol{h}_{k,i}(\boldsymbol{\theta}_k)^T \boldsymbol{V}_{k,i,s_k}(\boldsymbol{\theta}_k)\{\boldsymbol{Y}_i - \boldsymbol{h}_{k,i}(\boldsymbol{\theta}_k)\} \end{pmatrix},$$

where $\boldsymbol{h}_{k,i}(\boldsymbol{\theta}_k)$ is defined in (1) and $\nabla_{\boldsymbol{\theta}_k}\boldsymbol{h}_{k,i}(\boldsymbol{\theta}_k) = \partial \boldsymbol{h}_{k,i}(\boldsymbol{\theta}_k)/\partial \boldsymbol{\theta}_k^T$. Now, we are ready to form a joint quadratic inference function to simultaneously estimate all regression coefficients using all K studies. This objective function is

$$Q(\boldsymbol{\theta}, \hat{\boldsymbol{\eta}}) = n\bar{\boldsymbol{g}}(\boldsymbol{\theta}, \hat{\boldsymbol{\eta}})^T \boldsymbol{C}^-(\boldsymbol{\theta}, \hat{\boldsymbol{\eta}})\bar{\boldsymbol{g}}(\boldsymbol{\theta}, \hat{\boldsymbol{\eta}}),$$

where $\bar{\boldsymbol{g}}(\boldsymbol{\theta}, \hat{\boldsymbol{\eta}}) = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{g}_i(\boldsymbol{\theta}, \hat{\boldsymbol{\eta}}_i) \overset{def.}{=} \frac{1}{n}\sum_{i=1}^{n} \left(I[D_i = 1]\boldsymbol{g}_{1,i}^T, \ldots, I[D_i = K]\boldsymbol{g}_{K,i}^T \right)^T$, with $\boldsymbol{g}_{k,i} = I[D_i = k \in \mathcal{S}_o]\boldsymbol{g}_{k,i}(\boldsymbol{\theta}_k, \boldsymbol{h}_{k,i}) + I[D_i = k \in \mathcal{S}_m]\boldsymbol{g}_{k,i}(\boldsymbol{\theta}_k, \hat{\boldsymbol{\eta}}_{k,i})$ for $k = 1, \ldots, K$, and $\boldsymbol{C}(\boldsymbol{\theta}, \hat{\boldsymbol{\eta}})$ is a block-diagonal matrix, $\boldsymbol{C}(\boldsymbol{\theta}, \hat{\boldsymbol{\eta}}) = \frac{1}{n}\sum_{i=1}^{n} \text{diag}\{\boldsymbol{g}_{1,\text{i}}\boldsymbol{g}_{1,\text{i}}^\top, \ldots, \boldsymbol{g}_{\text{K,i}}\boldsymbol{g}_{\text{K,i}}^\top\}$. Parameter vector $\boldsymbol{\theta} = (\boldsymbol{\theta}_1^T, \ldots, \boldsymbol{\theta}_K^T)^T$ is then estimated by minimizing $Q(\boldsymbol{\theta}, \hat{\boldsymbol{\eta}})$ over Ω, that is

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta} \in \Omega} Q(\boldsymbol{\theta}, \hat{\boldsymbol{\eta}}). \tag{3}$$

We show that this joint QIF estimator $\hat{\boldsymbol{\theta}}$ acquires efficiency gains compared to study-specific QIF estimator $\hat{\boldsymbol{\theta}}_k$ in our setting of missing covariates, with details presented in Section 4.

4. Asymptotic Properties and Efficiency Gain

For convenience, the study-specific expectation under the distribution generating data of study k is denoted by $E_k(\cdot) = E(\cdot \mid D_i = k)$, $k = 1, \ldots, K$. Likewise $E_o(\cdot) = E(\cdot \mid D_i \in \mathcal{S}_o)$ denotes the expectation for all studies with fully observed data. Denote the Euclidean norm of a vector $\boldsymbol{b}$ by $\|\boldsymbol{b}\|$, the induced norm of a matrix $\boldsymbol{A}$ by $\|\boldsymbol{A}\| = \sup_{\|\boldsymbol{b}\|=1} \|\boldsymbol{A}\boldsymbol{b}\|$, the sup-norm of a function $f(\boldsymbol{x})$ by $\|f\|_\infty = \sup_{\boldsymbol{x}} \|f(\boldsymbol{x})\|$, and the L_2 norm of a random vector $\boldsymbol{X}$ by $\|\boldsymbol{X}\|_2$. We further impose some regularity conditions listed in the Web Appendix.

THEOREM 1. *Let $n_o = \sum_{l \in \mathcal{S}_o} n_l$. Suppose that (i) the mean model* (1) *is correctly specified, and that (ii) missing mechanism assumption 1 holds. Under Assumption A in the Web Appendix, if $t_{n_k} \to \infty$ and $t_{n_k} = o(n_o)$, estimator $\hat{\boldsymbol{\theta}}_k$ for $k \in \mathcal{S}_m$ given in* (2) *is consistent, namely, $\hat{\boldsymbol{\theta}}_k \xrightarrow{p} \boldsymbol{\theta}_{0,k}$ as $n_o \to \infty$.*

The proof of Theorem 1 is relatively straightforward. We first establish $\|\hat{\boldsymbol{\eta}}_k - \boldsymbol{\eta}_k\|_\infty = o_p(1)$, which is the focus of Lemma 1 in the Web Appendix, by applying similar arguments to those given in Chen et al. (2005). Consequently, we can show the uniform consistency for the score functions, $\sup_{\boldsymbol{\theta}_k \in \mathcal{B}} \|\bar{\boldsymbol{g}}_k(\boldsymbol{\theta}_k, \hat{\boldsymbol{\eta}}_k) - \bar{\boldsymbol{g}}_k(\boldsymbol{\theta}_k, \boldsymbol{\eta}_k)\| = o_p(1)$, and achieve the consistency of $\hat{\boldsymbol{\theta}}_k$ according to Glivenko–Cantelli Theorem and Lemma 5.2 of Newey (1994). The following theorem concerns asymptotic normality for $\hat{\boldsymbol{\theta}}_k$.

THEOREM 2. *Under Assumptions A and B in the Web Appendix, if $t_{n_k} \to \infty$ and $t_{n_k} = o(n_o)$ for $k \in \mathcal{S}_m$, the estimated score function $\bar{\boldsymbol{g}}_k(\boldsymbol{\theta}_{0,k}, \hat{\boldsymbol{\eta}}_k)$ can be represented by*

$$n_k^{1/2}\bar{\boldsymbol{g}}_k(\boldsymbol{\theta}_{0,k}, \hat{\boldsymbol{\eta}}_k) = n_k^{-1/2}\sum_{D_i=k} \boldsymbol{g}_{k,i}(\boldsymbol{\theta}_{0,k}, \boldsymbol{\eta}_{k,i}) + \tau_k^{1/2} n_o^{-1/2}\sum_{D_i \in \mathcal{S}_o} \boldsymbol{q}_{k,i}(\boldsymbol{\theta}_{0,k}, \boldsymbol{h}_{k,i}) + o_p(1),$$

where $\frac{n_k}{n_o} \to \tau_k$ as $n_k \to \infty, n_o \to \infty$ and $\boldsymbol{q}_{k,i}(\boldsymbol{\theta}_{0,k}, \boldsymbol{h}_{k,i}) = \left\{ \boldsymbol{q}_{k,i,1}(\boldsymbol{\theta}_{0,k}, \boldsymbol{h}_{k,i})^T, \ldots, \boldsymbol{q}_{k,i,s_k}(\boldsymbol{\theta}_{0,k}, \boldsymbol{h}_{k,i})^T \right\}^T$ consists of elements $\boldsymbol{q}_{k,i,s}(\boldsymbol{\theta}_{0,k}, \boldsymbol{h}_{k,i}) = \frac{f(\boldsymbol{X}_i \mid D_i = k)}{f(\boldsymbol{X}_i \mid D_i \in \mathcal{S}_o)} \nabla \boldsymbol{\eta}_{k,i}(\boldsymbol{\theta}_{0,k})^T \boldsymbol{V}_{k,i,s}\{\boldsymbol{\eta}_{k,i}(\boldsymbol{\theta}_{0,k}) - \boldsymbol{h}_{k,i}(\boldsymbol{\theta}_{0,k})\}$, $s = 1, \ldots, s_k$. Moreover, the asymptotic distribution of $\hat{\boldsymbol{\theta}}_k$ is given by

$$\sqrt{n_k}(\hat{\boldsymbol{\theta}}_k - \boldsymbol{\theta}_{0,k}) \xrightarrow{d} N\{\boldsymbol{0}, (\boldsymbol{G}_k^T \boldsymbol{\Sigma}_k^{-1} \boldsymbol{G}_k)^{-1}\},$$

where $\boldsymbol{G}_k = E_k\{\nabla \boldsymbol{g}_{k,i}(\boldsymbol{\theta}_{0,k}, \boldsymbol{\eta}_{k,i})\}$, *and* $\boldsymbol{\Sigma}_k = \boldsymbol{\Sigma}_{k,1} + \tau_k \boldsymbol{\Sigma}_{k,2}$ *with*

$$\boldsymbol{\Sigma}_{k,1} = E_k\{\boldsymbol{g}_{k,i}(\boldsymbol{\theta}_{0,k}, \boldsymbol{\eta}_{k,i})\boldsymbol{g}_{k,i}(\boldsymbol{\theta}_{0,k}, \boldsymbol{\eta}_{k,i})^T\}, \text{ and}$$
$$\boldsymbol{\Sigma}_{k,2} = E_o\{\boldsymbol{q}_{k,i}(\boldsymbol{\theta}_{0,k}, \boldsymbol{h}_{k,i})\boldsymbol{q}_{k,i}(\boldsymbol{\theta}_{0,k}, \boldsymbol{h}_{k,i})^T\}.$$

From Theorem 2, the representation of $n_k^{1/2}\bar{\boldsymbol{g}}_k(\boldsymbol{\theta}_{0,k}, \hat{\boldsymbol{\eta}}_{k,i})$ constitutes two components: $n_k^{-1/2} \times \sum_{D_i=k} \boldsymbol{g}_{k,i}(\boldsymbol{\theta}_{0,k}, \boldsymbol{\eta}_{k,i})$ and $\tau_k^{1/2} n_o^{-1/2} \sum_{D_i \in \mathcal{S}_o} \boldsymbol{q}_{k,i}(\boldsymbol{\theta}_{0,k}, \boldsymbol{h}_{k,i})$. It is interesting to note that the latter component is related to the weighted likelihood (e.g., Hu and Zidek, 2002; Wang and Zidek, 2005). Since covariate $\boldsymbol{Z}_i$ is not collected in study $k \in \mathcal{S}_m$, $\tau_k^{1/2} n_o^{-1/2} \sum_{D_i \in \mathcal{S}_o} \boldsymbol{q}_{k,i}(\boldsymbol{\theta}_{0,k}, \boldsymbol{h}_{k,i})$ presents an inference function using the observed data on $\boldsymbol{Z}_i$ from other studies in $\mathcal{S}_o$ weighted by the measure of relevance via a factor $f(\boldsymbol{X}_i \mid D_i = k)/f(\boldsymbol{X}_i \mid D_i \in \mathcal{S}_o)$. Thus, naturally our method yields the asymptotic variance of $\hat{\boldsymbol{\theta}}_k$ that consists of two pieces, $\boldsymbol{\Sigma}_{k,1}$ and $\boldsymbol{\Sigma}_{k,2}$, where $\boldsymbol{\Sigma}_{k,1}$ gives the asymptotic variance of $\hat{\boldsymbol{\theta}}_k$ when $\boldsymbol{\eta}_{k,i}$ were known, while $\boldsymbol{\Sigma}_{k,2}$ characterizes the additional variance incurred by the nonparametric sieve estimation of $\boldsymbol{\eta}_{k,i}$. The extra contribution by $\boldsymbol{\Sigma}_{k,2}$ toward the total variance of $\boldsymbol{\Sigma}_k$ is weighted according to a rate τ_k; when n_o exceeds n_k in the sense of $\frac{n_k}{n_o} \to 0$, the contribution of $\boldsymbol{\Sigma}_{k,2}$ will vanish and be ignored asymptotically. To evaluate $(\boldsymbol{G}_k^T \boldsymbol{\Sigma}_k^{-1} \boldsymbol{G}_k)^{-1}$, we need to replace $\boldsymbol{G}_k$ and $\boldsymbol{\Sigma}_k$ by their consistent estimates, respectively. This step involves estimating an unknown density ratio between $f(\boldsymbol{X}_i \mid D_i = k)$ and $f(\boldsymbol{X}_i \mid D_i \in \mathcal{S}_o)$. Note that we may rewrite this ratio as $\frac{f(D_i=k|\boldsymbol{X}_i) f(D_i \in \mathcal{S}_o)}{f(D_i \in \mathcal{S}_o|\boldsymbol{X}_i) f(D_i=k)}$, where $\frac{f(D_i=k|\boldsymbol{X}_i)}{f(D_i \in \mathcal{S}_o|\boldsymbol{X}_i)}$ may be estimated by a multinomial logistic model and $\frac{f(D_i \in \mathcal{S}_o)}{f(D_i=k)}$ by $\frac{n_o}{n_k}$. Given an estimated density ratio, the linearization variance estimation (Demnati and Rao, 2004, 2010) can also be applied. But obviously those approaches need some additional model assumptions which may not be easily checked in practice. An alternative way is to perform a bootstrap variance estimation, which avoids making extra model assumptions in the above density ratio estimation and hence is recommended in Section 5.

Now, we turn to the joint QIF estimator $\hat{\boldsymbol{\theta}}$ given in (3). Using similar arguments, we obtain the following two representations for the extended scores $n^{1/2}\bar{\boldsymbol{g}}(\boldsymbol{\theta}_0, \hat{\boldsymbol{\eta}})$: for $k \in \mathcal{S}_o$,

$$n^{-1/2} \sum_{D_i=k} g_{k,i}(\boldsymbol{\theta}_{0,k}, \boldsymbol{h}_{k,i}) = \left(\frac{\tau_k}{1+\tau_{\mathcal{S}_m}}\right)^{1/2} n_k^{-1/2} \sum_{D_i=k} \boldsymbol{g}_{k,i}(\boldsymbol{\theta}_{0,k}, \boldsymbol{h}_{k,i}) + o_p(1),$$

and for $k \in \mathcal{S}_m$, $n^{-1/2} \sum_{D_i=k} g_{k,i}(\boldsymbol{\theta}_{0,k}, \hat{\boldsymbol{\eta}}_{k,i})$ is given by

$$\left(\frac{\tau_k}{1+\tau_{\mathcal{S}_m}}\right)^{1/2} \left\{ n_k^{-1/2} \sum_{D_i=k} \boldsymbol{g}_{k,i}(\boldsymbol{\theta}_{0,k}, \boldsymbol{\eta}_{k,i}) + \tau_k^{1/2} n_o^{-1/2} \sum_{D_i \in \mathcal{S}_o} \boldsymbol{q}_{k,i}(\boldsymbol{\theta}_{0,k}, \boldsymbol{h}_{k,i}) \right\} + o_p(1),$$

where $\tau_{\mathcal{S}_m} = \sum_{k \in \mathcal{S}_m} \tau_k$. Thus, the asymptotic variance of $n^{1/2}\bar{\boldsymbol{g}}(\boldsymbol{\theta}_0, \hat{\boldsymbol{\eta}})$, $\boldsymbol{\Sigma}$, is a block-diagonal matrix, whose k-th block-element is given as follows:

$$\frac{\tau_k}{1+\tau_{\mathcal{S}_m}} \boldsymbol{\Sigma}_k I[k \in \mathcal{S}_o] + \frac{\tau_k}{1+\tau_{\mathcal{S}_m}} \boldsymbol{\Sigma}_{k,1} I[k \in \mathcal{S}_m] + \frac{\tau_k^2}{1+\tau_{\mathcal{S}_m}} \boldsymbol{\Sigma}_{k,2} I[k \in \mathcal{S}_m]. \quad (4)$$

Here $\boldsymbol{\Sigma}_k = E_k\{\boldsymbol{g}_{k,i}(\boldsymbol{\theta}_{0,k}, \boldsymbol{h}_{k,i})\boldsymbol{g}_{k,i}(\boldsymbol{\theta}_{0,k}, \boldsymbol{h}_{k,i})^T\}$, and the other two covariances, $\boldsymbol{\Sigma}_{k,1}$ and $\boldsymbol{\Sigma}_{k,2}$, are given in Theorem 2. The block-diagonal structure for $\boldsymbol{\Sigma}$ is due to the fact that $\boldsymbol{g}_{k,i}(\boldsymbol{\theta}_{0,k}, \boldsymbol{h}_{k,i})$ and $\boldsymbol{q}_{l,i}(\boldsymbol{\theta}_{0,k}, \boldsymbol{h}_{k,i})$ for study k and study l, $k \neq l$, are uncorrelated. When there exist shared parameters, namely $dim(\Omega) < (p+q)K$, the joint QIF estimation can improve efficiency for all regression coefficients by applying similar arguments in Wang et al. (2012). When the shared parameters contain part of parameters in $\boldsymbol{\lambda}_{0,k}$ for $k \in \mathcal{S}_m$, the proposed joint estimation approach can achieve higher efficiency than that from (2) using individual datasets.

We summarize the above discussion in the following Theorem.

THEOREM 3. *Under Assumptions A and B given in the Web Appendix, the joint estimator* $\hat{\boldsymbol{\theta}}$ *given in* (3) *is asymptotically normally distributed with mean* $\mathbf{0}$ *and asymptotic variance* $(\boldsymbol{G}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{G})^{-1}$, *namely*

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \xrightarrow{d} N\{\mathbf{0}, (\boldsymbol{G}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{G})^{-1}\}, \text{ as } n \to \infty$$

where $\boldsymbol{\Sigma}$ *is a block-diagonal matrix whose kth block-element is given in* (4) *and* $\boldsymbol{G} = (\boldsymbol{G}_1^T, \ldots, \boldsymbol{G}_K^T)^T$ *with the k-th matrix* $\boldsymbol{G}_k$ *given by* $\boldsymbol{G}_k = \begin{cases} E_k\{\nabla_{\boldsymbol{\theta}_k} \boldsymbol{g}_{k,i}(\boldsymbol{\theta}_{0,k}, \boldsymbol{h}_{k,i})\}, & k \in \mathcal{S}_o \\ E_k\{\nabla_{\boldsymbol{\theta}_k} \boldsymbol{g}_{k,i}(\boldsymbol{\theta}_{0,k}, \boldsymbol{\eta}_{k,i})\}, & k \in \mathcal{S}_m \end{cases}$. *When there exist shared parameters across studies,* $\hat{\boldsymbol{\theta}}$ *has a smaller asymptotic variance than any* $\hat{\boldsymbol{\theta}}_k$, $k = 1, \ldots, K$, *obtained by* (2) *using data from individual studies.*

5. Implementation

This section focuses on two key elements in the implementation of our method: (i) bootstrap variance estimation and (ii) selection of the number of basis functions in the nonparametric estimation of $\boldsymbol{\eta}(\boldsymbol{x}, \boldsymbol{\theta}_k)$.

Following Chen et al. (2003) and Hall and Horowitz (1996), we establish a procedure to estimate the asymptotic variance by using bootstrap resampling techniques. Let $\{\boldsymbol{Y}_i^*, \boldsymbol{X}_i^*, \boldsymbol{Z}_i^*, D_i^*\}_{i=1}^n$ be a bootstrap sample, which is generated by the scheme of stratified sampling with individual studies as strata, so that the resulting bootstrap sample constitutes the same proportions of subjects from K studies and preserves the same within-subject correlation as that of the original sample. According to Hall and Horowitz (1996), a bootstrap version of extended score $\bar{\boldsymbol{g}}_k^*(\boldsymbol{\theta}_k, \hat{\boldsymbol{\eta}}_{k,i})$ needs to be centered, given by

$$\bar{\boldsymbol{g}}_k^c(\boldsymbol{\theta}_k, \hat{\boldsymbol{\eta}}_k^*) = \bar{\boldsymbol{g}}_k^*(\boldsymbol{\theta}_k, \hat{\boldsymbol{\eta}}_k^*) - \bar{\boldsymbol{g}}_k(\hat{\boldsymbol{\theta}}_k, \hat{\boldsymbol{\eta}}_k),$$

where $\hat{\boldsymbol{\theta}}_k$ and $\hat{\boldsymbol{\eta}}_k$ are estimated from the original sample and $\hat{\boldsymbol{\eta}}_k^*$ is estimated from the bootstrap sample. The rea-

son for the need of centering is that the QIF estimator is obtained as a minimizer of an objective function, and the resulting estimated moments of the extended scores are not necessarily equal to $\mathbf{0}$. It is imperative to subtract $\bar{\boldsymbol{g}}_k(\hat{\boldsymbol{\theta}}_k, \hat{\boldsymbol{\eta}}_k)$ from $\bar{\boldsymbol{g}}_k^*(\boldsymbol{\theta}_k, \hat{\boldsymbol{\eta}}_k^*)$ to obtain asymptotically unbiased estimating functions, which is critical to ensure consistent estimation. Consequently, the bootstrap estimator $\hat{\boldsymbol{\theta}}_k^*$ is defined as the minimizer of $Q_k(\boldsymbol{\theta}_k, \hat{\boldsymbol{\eta}}_k^*)$ given in (2), where $\bar{\boldsymbol{g}}_k(\boldsymbol{\theta}_k, \hat{\boldsymbol{\eta}}_k)$ is replaced by its bootstrap version $\bar{\boldsymbol{g}}_k^c(\boldsymbol{\theta}_k, \hat{\boldsymbol{\eta}}_k^*)$. Repeating this bootstrap procedure a certain number of times, we yield a set of bootstrap estimates of $\boldsymbol{\theta}_k$, which are then used to calculate the bootstrap variances. The same procedure can be established for the joint estimation of $\boldsymbol{\theta}$.

Another critical issue for implementing the proposed method is to determine the number of basis functions when estimating $\boldsymbol{\eta}_{k,i}(\boldsymbol{\theta}_k)$. Since a nonparametric regression is used to estimate the conditional mean model instead of estimating regression coefficients selecting the number of basis functions is more relevant to estimation of $\boldsymbol{\eta}_{k,i}(\boldsymbol{\theta}_k)$ than estimation of $\boldsymbol{\theta}_k$. There are several criteria potentially useful to serve for such a selection purpose, including Schwarz's (1978) Bayesian information criterion (BIC), Craven and Wahba's (1979) generalized cross-validation (GCV), and Wang and Qu's (2009) QIF-based BIC (BIQIF). Because BIQIF tends to select underfitting models, we follow He et al. (2002) and propose a new BIC-type model selection criterion:

$$\mathrm{BIC}(t_{n_k}) = Q(\hat{\boldsymbol{\theta}}_k^{(t_{n_k})}, \hat{\boldsymbol{\eta}}_k) + \frac{\log n}{2n}(p + t_{n_k}),\ k = 1, \cdots, K,$$

where p is the number of regression parameters, t_{n_k} is the number of basis functions, and $\hat{\boldsymbol{\theta}}_k^{(t_{n_k})}$ is the estimate of $\boldsymbol{\theta}_k$ when t_{n_k} basis functions are used. Within a sufficiently wide range of candidate values, the best t_{n_k} is the one with the smallest $\mathrm{BIC}(\mathrm{t_{n_k}})$.

6. Simulation Studies

We run a simulation study to compare our proposed method with two existing methods, GEE and QIF, using full data, imputed data by either parametric multiple imputation or nonparametric hot-deck multiple imputation (Little and Rubin, 2002). The focus of this comparison is to illustrate that the single and multiple imputation methods are not applicable for the study-specific missing data, which certifies a need of the proposed methodology for this special missing structure. We draw 4000 datasets from the following model:

$$Y_{ij} = \begin{cases} \beta_0 + \beta_1 X_{ij} + \beta_2 Z_{ij} + \epsilon_{ij},\ D_i = 1 \\ \beta_0 + \beta_3 X_{ij} + \beta_4 Z_{ij} + \epsilon_{ij},\ D_i = 2 \end{cases}, \quad j = 1, \ldots, m, i = 1, \ldots, n,$$

where the true regression coefficients $\boldsymbol{\theta}_0 = (\beta_0, \beta_1, \beta_2, \beta_3, \beta_4)^T = (1, 1, -0.5, 2, 0.5)^T$, $n = 200$ subjects and $m = 4$ repeated measurements. Covariate X_{ij} is generated from $\mathrm{Unif}(0, 1)$, and covariate Z_{ij} is generated from a conditional model given X_{ij} of the form: $Z_{ij} = sin(4\pi X_{ij}) + \zeta_{ij}$, where $\zeta_{ij} \overset{iid}{\sim} N(0, 0.5)$. Here $\boldsymbol{Z}_i$ is treated as a study-specific missing covariate whose state of missingness, D_i, is determined by a logistic model on X_{i1}, $\mathrm{logit}\{\mathrm{P(D_i = 2 \mid X_i)}\} = 0.5 + 0.4\mathrm{X_{i1}}$. As a result, 39% of subjects are sampled from study 2 to have missing $\boldsymbol{Z}_i$. The above specification implies that $E(Y_{ij} \mid X_{ij}, D_i = 2) = \beta_0 + \beta_3 X_{ij} + \beta_4 sin(4\pi X_{ij})$. Error terms, $\epsilon_i = (\epsilon_{i1}, \ldots, \epsilon_{im})^T$, are independently generated from $N_m\{\mathbf{0}, \phi_k \boldsymbol{R}_k(\alpha_k)\}$, $k = 1, 2$, where the covariance matrix $\phi_k \boldsymbol{R}_k(\alpha_k)$ is specified in the following two cases:

- Case I. correlation matrices $\boldsymbol{R}_1(\cdot)$ and $\boldsymbol{R}_2(\cdot)$ in two studies are both AR-1 correlation with $(\alpha_1, \alpha_2) = (0.4, 0.4)$, and variance parameters are $(\phi_1, \phi_2) = (1, 1)$;
- Case II. correlation matrix $\boldsymbol{R}_1(\cdot)$ in study 1 is AR-1 with $\alpha_1 = 0.7$ while correlation matrix $\boldsymbol{R}_2(\cdot)$ in study 2 is compound symmetry with $\alpha_2 = 0.2$; variance parameters are different, $(\phi_1, \phi_2) = (10, 1)$.

The imputed datasets for study 2 are created according to the true conditional distribution of $\boldsymbol{Z}_i$ given $\boldsymbol{X}_i$ to avoid potential uncertainty in the estimation of this conditional distribution. Here, we use $f(\boldsymbol{Z}_i \mid \boldsymbol{X}_i)$ for imputation instead of $f(\boldsymbol{Z}_i \mid \boldsymbol{X}_i, \boldsymbol{Y}_i)$ because two studies are governed by two different regression models, and therefore $f(\boldsymbol{Z}_i \mid \boldsymbol{X}_i, \boldsymbol{Y}_i)$ in study 2 is not estimable using observed data in study 1 (refer to a detailed explanation provided in a paragraph below). Likewise, in the implementation of hot-deck imputation, we select a set of observed data that are similar to those who have missing $\boldsymbol{Z}_i$ in the sense of smaller Euclidean distances in their $\boldsymbol{X}_i$ values and randomly generate 10 imputed datasets.

The conditional mean function, $E(Y_{ij} \mid X_{ij}, D_i = 2) = \beta_0 + \beta_3 X_{ij} + \beta_4 sin(4\pi X_{ij})$, is estimated using six B-spline basis functions. Here, we compare our method with the imputation methods. Simulation results for cases I and II above are reported in Tables 1 and 2 under two working correlation structures. In the ideal case where the hypothetical full data are used, both QIF and GEE have shown little biases and reached desirable 95% nominal coverage for both correlation scenarios. When covariate $\boldsymbol{Z}_i$ is missing in study 2, both parametric and hot-deck multiple imputation methods produce noticeable estimation biases in GEE and QIF, particularly for those parameters exclusively belonging to study 2, where severe undercoverage exists for β_3 and β_4 (substantially lower than 95% nominal level).

The failure of both parametric multiple imputation and hot-deck imputation may be attributed to the validity of the imputation methods, which have been justified only under the selection model in the literature. Note that in a selection model regression parameters are present in the distribution $f(\boldsymbol{Y}_i \mid \boldsymbol{X}_i, \boldsymbol{Z}_i)$, which however is not the case in our problem where regression parameters are different across two studies. Thus, the imputation is in general not applicable to multiple studies governed by models with different parameters.

In effect, $f(\boldsymbol{Z}_i \mid \boldsymbol{Y}_i, \boldsymbol{X}_i, D_i = 2) \propto f(\boldsymbol{Z}_i \mid \boldsymbol{X}_i, D_i = 2) f(\boldsymbol{Y}_i \mid \boldsymbol{X}_i, \boldsymbol{Z}_i, D_i = 2)$. By Assumption 1, $f(\boldsymbol{Z}_i \mid \boldsymbol{X}_i, D_i = 2) = f(\boldsymbol{Z}_i \mid \boldsymbol{X}_i)$ can be estimated from study 1. However, $f(\boldsymbol{Y}_i \mid \boldsymbol{X}_i, \boldsymbol{Z}_i, D_i = 2)$ cannot be estimated from study 1 because $f(\boldsymbol{Y}_i \mid \boldsymbol{X}_i, \boldsymbol{Z}_i, D_i = 2) \neq f(\boldsymbol{Y}_i \mid \boldsymbol{X}_i, \boldsymbol{Z}_i, D_i = 1)$ and $\boldsymbol{\theta}_{0,1} \neq \boldsymbol{\theta}_{0,2}$ in model (1). Therefore, $f(\boldsymbol{Z}_i \mid \boldsymbol{Y}_i, \boldsymbol{X}_i, D_i = 2)$ is not estimable and cannot be used to impute missing $\boldsymbol{Z}_i$ in study 2. Even if here the true conditional distribution $f(\boldsymbol{Z}_i \mid \boldsymbol{X}_i)$ is used in the imputation, imputed values for missing $\boldsymbol{Z}_i$ may still violate unbiasedness of $E\{\boldsymbol{Y}_{ij} - h(\boldsymbol{X}_{ij}^T \boldsymbol{\beta}_{0,2} + \boldsymbol{Z}_{ij}^T \boldsymbol{\lambda}_{0,2}) \mid$

Table 1

Summary of regression parameter estimates for data generated in Case I under AR-1 working correlation (or under compound symmetry working correlation, presented in the parenthesis). Full, Par-MI, and Hot-deck represent full data, data imputed by parametric multiple imputation, and data imputed by hot-deck multiple imputation, respectively. E.S.E. is the empirical standard error computed from 4000 simulated datasets. A.S.E. is the asymptotic standard error. For our method, A.S.E. is the bootstrap standard error computed using 400 bootstrap samples. The coverage probability, C.P., is computed by using the asymptotic standard error.

Method	$\boldsymbol{\theta}$	$\hat{\boldsymbol{\theta}}$	Bias	E.S.E	A.S.E.	MSE	C.P.
Full QIF	β_0	1.000 (1.001)	0.000 (0.001)	0.075 (0.078)	0.075 (0.077)	0.006 (0.006)	0.952 (0.946)
	β_1	1.001 (1.000)	0.001 (0.000)	0.124 (0.131)	0.126 (0.131)	0.015 (0.017)	0.948 (0.951)
	β_2	−0.501 (−0.501)	−0.001 (−0.001)	0.040 (0.043)	0.040 (0.042)	0.002 (0.002)	0.946 (0.944)
	β_3	2.006 (2.002)	0.006 (0.002)	0.158 (0.165)	0.160 (0.164)	0.025 (0.027)	0.956 (0.952)
	β_4	0.499 (0.498)	−0.001 (−0.002)	0.060 (0.062)	0.058 (0.060)	0.004 (0.004)	0.941 (0.946)
Full GEE	β_0	1.000 (1.001)	0.000 (0.001)	0.072 (0.076)	0.072 (0.075)	0.005 (0.006)	0.945 (0.947)
	β_1	1.002 (1.000)	0.002 (0.000)	0.122 (0.130)	0.122 (0.128)	0.015 (0.017)	0.948 (0.948)
	β_2	−0.501 (−0.501)	−0.001 (−0.001)	0.040 (0.042)	0.039 (0.041)	0.002 (0.002)	0.939 (0.940)
	β_3	2.003 (2.002)	0.003 (0.002)	0.151 (0.160)	0.152 (0.158)	0.023 (0.026)	0.949 (0.952)
	β_4	0.499 (0.498)	−0.001 (−0.002)	0.057 (0.060)	0.055 (0.058)	0.003 (0.004)	0.951 (0.940)
Par-MI QIF	β_0	1.035 (1.036)	0.035 (0.036)	0.077 (0.079)	0.079 (0.081)	0.007 (0.008)	0.927 (0.931)
	β_1	0.961 (0.958)	−0.039 (−0.042)	0.126 (0.133)	0.129 (0.134)	0.017 (0.019)	0.939 (0.929)
	β_2	−0.504 (−0.504)	−0.004 (−0.004)	0.040 (0.043)	0.040 (0.042)	0.002 (0.002)	0.944 (0.942)
	β_3	1.865 (1.863)	−0.135 (−0.137)	0.169 (0.172)	0.180 (0.183)	0.047 (0.048)	0.894 (0.894)
	β_4	0.241 (0.240)	−0.259 (−0.260)	0.049 (0.050)	0.083 (0.084)	0.069 (0.070)	0.040 (0.047)
Par-MI GEE	β_0	1.042 (1.042)	0.042 (0.042)	0.074 (0.077)	0.076 (0.078)	0.007 (0.008)	0.917 (0.914)
	β_1	0.951 (0.949)	−0.049 (−0.051)	0.123 (0.130)	0.125 (0.131)	0.018 (0.020)	0.923 (0.926)
	β_2	−0.505 (−0.505)	−0.005 (−0.005)	0.040 (0.042)	0.039 (0.041)	0.002 (0.002)	0.939 (0.934)
	β_3	1.851 (1.853)	−0.149 (−0.147)	0.162 (0.169)	0.169 (0.174)	0.049 (0.050)	0.879 (0.872)
	β_4	0.239 (0.239)	−0.261 (−0.261)	0.047 (0.049)	0.078 (0.080)	0.070 (0.071)	0.017 (0.031)
Hot-deck QIF	β_0	1.034 (1.035)	0.034 (0.035)	0.076 (0.079)	0.079 (0.081)	0.007 (0.007)	0.931 (0.935)
	β_1	0.961 (0.959)	−0.039 (−0.041)	0.126 (0.132)	0.129 (0.134)	0.017 (0.019)	0.937 (0.933)
	β_2	−0.504 (−0.504)	−0.004 (−0.004)	0.040 (0.043)	0.040 (0.042)	0.002 (0.002)	0.944 (0.944)
	β_3	1.866 (1.864)	−0.134 (−0.136)	0.170 (0.173)	0.181 (0.183)	0.047 (0.048)	0.893 (0.893)
	β_4	0.237 (0.236)	−0.263 (−0.264)	0.049 (0.050)	0.083 (0.084)	0.072 (0.072)	0.039 (0.038)
Hot-deck GEE	β_0	1.041 (1.042)	0.041 (0.042)	0.073 (0.076)	0.076 (0.078)	0.007 (0.008)	0.923 (0.917)
	β_1	0.952 (0.949)	−0.048 (−0.051)	0.123 (0.130)	0.125 (0.131)	0.018 (0.019)	0.924 (0.927)
	β_2	−0.504 (−0.505)	−0.004 (−0.005)	0.040 (0.042)	0.039 (0.041)	0.002 (0.002)	0.937 (0.936)
	β_3	1.852 (1.855)	−0.148 (−0.145)	0.162 (0.169)	0.170 (0.174)	0.048 (0.050)	0.879 (0.871)
	β_4	0.235 (0.235)	−0.265 (−0.265)	0.047 (0.049)	0.078 (0.080)	0.073 (0.073)	0.015 (0.022)
Our Method	β_0	1.002 (1.002)	0.002 (0.002)	0.080 (0.083)	0.081 (0.084)	0.006 (0.007)	0.944 (0.942)
	β_1	1.000 (0.999)	0.000 (−0.001)	0.129 (0.136)	0.130 (0.136)	0.017 (0.018)	0.943 (0.943)
	β_2	−0.501 (−0.501)	−0.001 (−0.001)	0.040 (0.043)	0.040 (0.042)	0.002 (0.002)	0.949 (0.940)
	β_3	1.992 (1.992)	−0.008 (−0.008)	0.206 (0.210)	0.207 (0.211)	0.042 (0.044)	0.948 (0.960)
	β_4	0.483 (0.483)	−0.017 (−0.017)	0.211 (0.210)	0.210 (0.214)	0.045 (0.044)	0.936 (0.943)

$\boldsymbol{X}_{ij}, \boldsymbol{Z}_{ij}, D_i = 2\} = 0$. Therefore, both GEE and QIF with the imputed data are impaired and yield significant estimation biases. Molenberghs and Kenward (2007,Chapter 2) examine the performance of GEE with multiple imputation for missing responses, where by comparing IPW GEE with imputation-based GEE under the selection model, they show that imputation-based GEE produces significantly larger bias as well as mean squared error (MSE) than IPW GEE in various longitudinal data settings. Our findings are in agreement with theirs.

In contrast to the imputation methods, our proposed method demonstrates satisfactory performances in terms of bias and coverage probability. For example, the coverage of β_4 is close to the nominal 95% level in various settings. This is because our method uses asymptotically unbiased estimating functions derived by plugging in a consistent nonparametric estimation of $E(\boldsymbol{Y}_i \mid \boldsymbol{X}_i, D_i = 2)$. As shown in Table 1 for case I and Table 2 for case II, the price paid to gain the benefit of desirable coverage is the larger standard deviations compared to the ideal QIF and GEE using the hypothetical full data. This is not surprising because $E(\boldsymbol{Y}_i \mid \boldsymbol{X}_i, D_i = 2)$ is estimated nonparametrically in our method. This further confirms the theoretical results in Theorems 2 and 3 regarding the asymptotic covariances, where, as explained already, the uncertainty from the nonparametric estimation is to be accounted for.

We evaluate the performance of our proposed method with various missing data percentages in the Web Appendix, and the results are stable. We also examine how the proposed BIC

Table 2

Summary of regression parameter estimates for data generated in Case II under AR-1 working correlation (or under compound symmetry working correlation, presented in the parenthesis). Full, Par-MI, and Hot-deck represent full data, data imputed by parametric multiple imputation, and data imputed by hot-deck multiple imputation, respectively. E.S.E. is the empirical standard error computed from 4000 simulated datasets. A.S.E. is the asymptotic standard error. For our method, A.S.E. is the bootstrap standard error computed using 400 bootstrap samples. The coverage probability, C.P., is computed by using the asymptotic standard error.

Method	$\boldsymbol{\theta}$	$\hat{\boldsymbol{\theta}}$	Bias	E.S.E	A.S.E.	MSE	C.P.
Full QIF	β_0	1.002 (1.001)	0.002 (0.001)	0.132 (0.129)	0.124 (0.123)	0.017 (0.017)	0.917 (0.929)
	β_1	0.992 (0.988)	−0.008 (−0.012)	0.287 (0.306)	0.282 (0.301)	0.082 (0.094)	0.938 (0.947)
	β_2	−0.497 (−0.497)	0.003 (0.003)	0.093 (0.104)	0.096 (0.105)	0.009 (0.011)	0.952 (0.958)
	β_3	2.002 (2.004)	0.002 (0.004)	0.219 (0.212)	0.210 (0.205)	0.048 (0.045)	0.936 (0.940)
	β_4	0.503 (0.503)	0.003 (0.003)	0.063 (0.061)	0.064 (0.062)	0.004 (0.004)	0.947 (0.941)
Full GEE	β_0	1.009 (1.011)	0.009 (0.011)	0.201 (0.211)	0.186 (0.197)	0.041 (0.045)	0.923 (0.935)
	β_1	0.984 (0.979)	−0.016 (−0.021)	0.290 (0.330)	0.283 (0.319)	0.084 (0.109)	0.944 (0.934)
	β_2	−0.498 (−0.498)	0.002 (0.002)	0.089 (0.105)	0.090 (0.104)	0.008 (0.011)	0.957 (0.952)
	β_3	1.995 (1.998)	−0.005 (−0.002)	0.243 (0.241)	0.229 (0.228)	0.059 (0.058)	0.929 (0.928)
	β_4	0.502 (0.502)	0.002 (0.002)	0.069 (0.062)	0.069 (0.063)	0.005 (0.004)	0.943 (0.947)
Par-MI QIF	β_0	1.102 (1.105)	0.102 (0.105)	0.140 (0.137)	0.136 (0.135)	0.030 (0.030)	0.884 (0.881)
	β_1	0.921 (0.906)	−0.079 (−0.094)	0.289 (0.308)	0.285 (0.305)	0.090 (0.103)	0.928 (0.935)
	β_2	−0.502 (−0.504)	−0.002 (−0.004)	0.093 (0.104)	0.096 (0.105)	0.009 (0.011)	0.954 (0.954)
	β_3	1.783 (1.779)	−0.217 (−0.221)	0.231 (0.222)	0.237 (0.232)	0.100 (0.098)	0.866 (0.866)
	β_4	0.235 (0.234)	−0.265 (−0.266)	0.049 (0.048)	0.089 (0.086)	0.073 (0.073)	0.049 (0.034)
Par-MI GEE	β_0	1.052 (1.054)	0.052 (0.054)	0.203 (0.212)	0.188 (0.198)	0.044 (0.048)	0.916 (0.918)
	β_1	0.952 (0.943)	−0.048 (−0.057)	0.290 (0.330)	0.283 (0.319)	0.086 (0.112)	0.941 (0.935)
	β_2	−0.501 (−0.501)	−0.001 (−0.001)	0.089 (0.105)	0.090 (0.104)	0.008 (0.011)	0.957 (0.952)
	β_3	1.800 (1.810)	−0.200 (−0.190)	0.251 (0.246)	0.251 (0.246)	0.103 (0.097)	0.882 (0.892)
	β_4	0.237 (0.236)	−0.263 (−0.264)	0.055 (0.049)	0.096 (0.087)	0.072 (0.072)	0.093 (0.036)
Hot-deck QIF	β_0	1.101 (1.104)	0.101 (0.104)	0.141 (0.137)	0.137 (0.136)	0.030 (0.030)	0.886 (0.882)
	β_1	0.922 (0.908)	−0.078 (−0.092)	0.289 (0.308)	0.285 (0.305)	0.090 (0.103)	0.935 (0.941)
	β_2	−0.502 (−0.504)	−0.002 (−0.004)	0.093 (0.104)	0.096 (0.105)	0.009 (0.011)	0.952 (0.952)
	β_3	1.786 (1.782)	−0.214 (−0.218)	0.232 (0.223)	0.238 (0.233)	0.099 (0.098)	0.856 (0.864)
	β_4	0.232 (0.231)	−0.268 (−0.269)	0.050 (0.049)	0.089 (0.087)	0.074 (0.075)	0.056 (0.044)
Hot-deck GEE	β_0	1.052 (1.054)	0.052 (0.054)	0.203 (0.212)	0.188 (0.198)	0.044 (0.048)	0.916 (0.920)
	β_1	0.952 (0.944)	−0.048 (−0.056)	0.290 (0.330)	0.283 (0.319)	0.086 (0.112)	0.942 (0.936)
	β_2	−0.500 (−0.500)	0.000 (0.000)	0.090 (0.105)	0.090 (0.104)	0.008 (0.011)	0.957 (0.953)
	β_3	1.801 (1.811)	−0.199 (−0.189)	0.251 (0.247)	0.252 (0.247)	0.102 (0.096)	0.881 (0.891)
	β_4	0.233 (0.232)	−0.267 (−0.268)	0.057 (0.052)	0.096 (0.087)	0.074 (0.074)	0.101 (0.045)
Our method	β_0	1.007 (1.008)	0.007 (0.008)	0.179 (0.176)	0.163 (0.164)	0.032 (0.031)	0.931 (0.929)
	β_1	0.989 (0.982)	−0.011 (−0.018)	0.298 (0.319)	0.288 (0.311)	0.089 (0.102)	0.937 (0.944)
	β_2	−0.497 (−0.498)	0.003 (0.002)	0.093 (0.104)	0.095 (0.104)	0.009 (0.011)	0.953 (0.955)
	β_3	1.990 (1.988)	−0.010 (−0.012)	0.326 (0.316)	0.309 (0.309)	0.106 (0.100)	0.941 (0.952)
	β_4	0.487 (0.485)	−0.013 (−0.015)	0.259 (0.246)	0.255 (0.252)	0.067 (0.061)	0.956 (0.956)

criterion behaves in the selection of basis functions. Under the same settings of the previous simulation study, we increase the number of basis functions from 4 to 12 in the estimation of $E(Y_{ij} \mid \boldsymbol{X}_{ij}, D_i = 2)$, and summarize the results in Figure 1. This figure indicates that BIC criterion is minimized at six, after which the MSE cannot be improved significantly with more basis functions being used. This evidence implies that our criterion tends to chose a parsimonious nonparametric model with small MSE.

To illustrate the efficiency gain in the joint analysis, we further run a simulation study to compare the standard errors obtained from the joint analysis and those obtained from the individual analysis. This is to confirm the theoretical result given in Theorem 3. The data are generated in the same way as in case I of the previous simulation. The joint analysis utilizes the fact that two studies have a common intercept parameter, while the individual analysis ignores this fact and includes different intercepts in the respective models. The standard errors are calculated by the bootstrap method discussed in Section 5. Summarized results over 100 replications in Table 3 clearly show that the joint analysis has given smaller standard errors for all regression coefficients. This efficiency improvement appears very substantial for study 2 where missing covariates are present. The individual analysis only uses 61% of the sample size to obtain parameter estimation. In conclusion, it is clearly beneficial to borrow data information from study 1 to improve inference for the parameters in study 2.

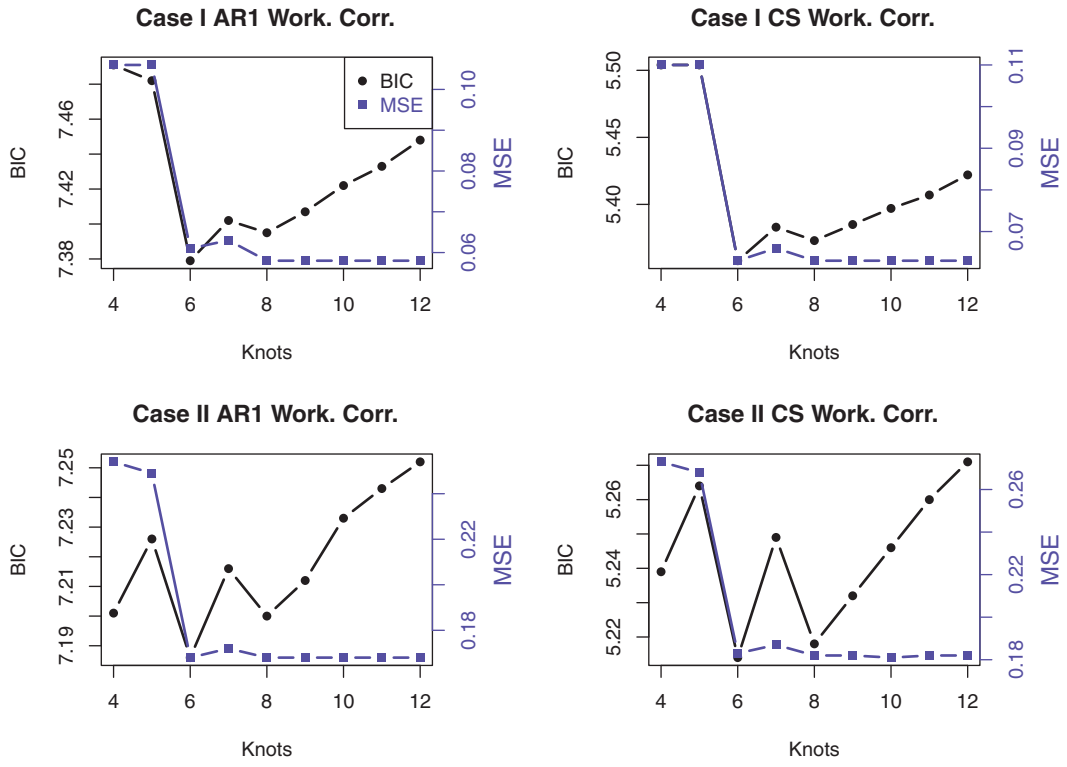

Figure 1. BIC and MSE profile curves for selecting the number of basis functions.

We have also conducted simulation experiments on binary longitudinal outcomes, and similar findings were yielded. Refer to the details in the Web Appendix.

7. Application

We apply our method to analyze the lead exposure data collected from two longitudinal birth cohorts in Mexico City. Between 1994 and 2005, the study recruited 89 mother–infant pairs in cohort B and 492 mother–infant pairs in cohort C at two maternity hospitals serving low-to-moderate income populations (Afeiche et al., 2011). We are interested in studying the effect of cord blood lead exposure on child's weight growth. Child's weight was measured repeatedly at 0, 3, 6, 12, 18, 24, 30, 36, 48, and 60 months after birth in cohort B, while at 0, 1, 4, 7, 12, 18, 24, 30, 36, 42, and 48 months in cohort C. Two lead exposure measures, mother's blood lead (PBL), and child's cord blood lead (CBL), are recorded at baseline visit (time 0). PBL was measured for all mothers in both cohorts while CBL was collected for all infants in cohort C and approximately 46% of infants in cohort B due to child's or maternal refusal, inability to give blood or because a blood lead measure was not scheduled.

Table 3

Comparison of standard errors from joint estimation and individual estimation in Case I under AR-1 and compound symmetry (CS) working correlations. For our method, standard error is the bootstrap standard error computed using 200 bootstrap samples.

		Standard error			
		AR-1		CS	
Study	$\hat{\beta}$	Joint	Individual	Joint	Individual
I	β_0	0.076	0.089	0.079	0.092
	β_1	0.125	0.136	0.131	0.142
	β_2	0.039	0.040	0.041	0.042
II	β_0	0.076	0.258	0.079	0.255
	β_3	0.174	0.448	0.178	0.442
	β_4	0.097	0.228	0.100	0.225

The upper panel in Figure 2 displays trajectories of child's weights versus child's ages across cohorts B and C, and its lower panel includes the scatter-plots of child's weights versus child's CBL in log scale across the two cohorts. Adjusting for child's gender and age, we estimate the effect of CBL on weight growth via the following model:

$$\begin{aligned} E\left(Y_{k,ij} \mid X_{k,i}, Z_{k,i}, G_{k,i}, t_{k,ij}\right) &= \beta_1^k + \beta_2^k X_{k,i} + \beta_3^k B_1(Z_{k,i}) \\ &\quad + \beta_4^k B_2(Z_{k,i}) + \beta_5^k G_{k,i} \\ &\quad + \beta_6^k B_1(t_{k,ij}) + \beta_7^k B_2(t_{k,ij}) \\ &\quad + \beta_8^k B_3(t_{k,ij}), \quad k = 1, 2, \end{aligned} \tag{5}$$

where cohorts C and B are denoted by $k = 1$ and $k = 2$, respectively. For subject i at the jth visit, variable $Y_{k,ij}$, $t_{k,ij}$, $X_{k,i}$, $Z_{k,i}$, and $G_{k,i}$ are log (weight), child's ages (year), log (PBL), log (CBL), and child's gender (1 for male and 0 for female),

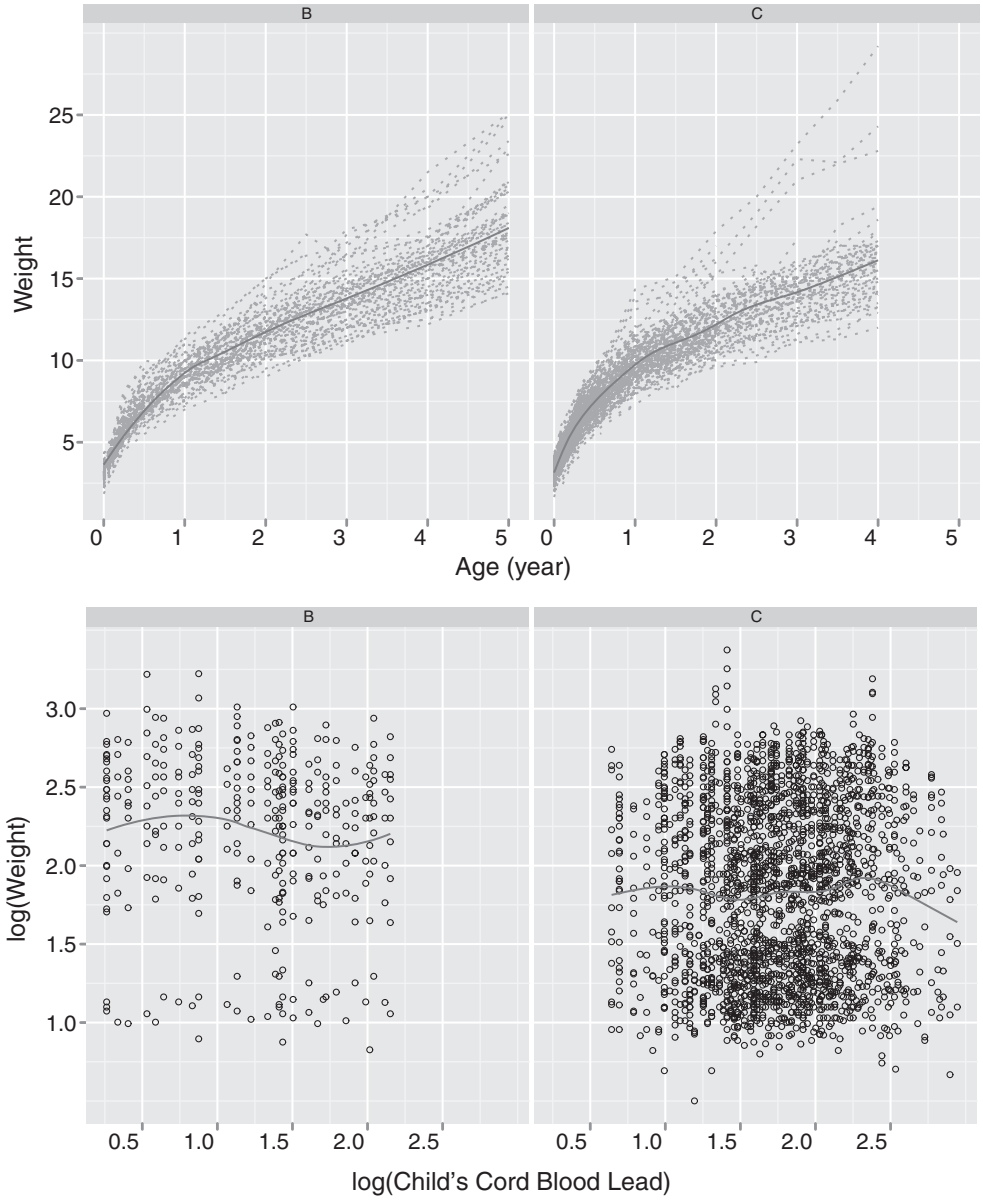

Figure 2. Trajectories of children's weights versus children's ages across the two cohorts (upper panel), and scatterplots of log-transformed children's weights versus log-transformed children's cord blood lead exposure across the two cohorts (lower panel).

respectively. We apply log-transformation on weight, PBL, and CBL to reduce skewness. Effects on time $t_{k,ij}$ and $Z_{k,i}$ are captured by linear splines with three basis functions, $B_1(t_{k,ij})$, $B_2(t_{k,ij})$, and $B_3(t_{k,ij})$, for covariate time $t_{k,ij}$ at knots 0.5 and 2, and two basis functions, $B_1(Z_{k,i})$ and $B_2(Z_{k,i})$, for $Z_{k,i}$ at knot 2.3 in log scale. The piecewise linear trend of child's weight versus child's age can be observed in Figure 2.

Given that 46% of CBL measurements are missing in cohort B, we estimate the effect of covariate CBL by merging the two cohorts. Through a routine model screening process using interactions between covariates and cohort dummy variables, we finally reach a model with common coefficients for $X_{k,i}$, $G_{k,i}$, $B_1(t_{k,ij})$, and $B_2(t_{k,ij})$ across two cohorts.

Results in Table 4 indicate that gender and age both are strongly associated with weight growth of children. For children age 2 or younger in two cohorts, they have similar weight growth on average. Children older than 2 years in cohort B grow faster than their peers in cohort C. As to the effect of lead exposure in child's cord blood, the effect of log(CBL) on weight growth in cohort C appears to be nearly significant when log(CBL) is greater than 2.3, or equivalently CBL concentration larger than 10μg/L.

8. Concluding Remarks

We have developed a novel estimating function approach to assessing covariate effects through merging datasets from mul-

Table 4
Estimates of regression parameters and p-values for children's lead exposure analysis. Intercept, log(PBL)*, Gender,* B_1(age)*, and* B_2(age) *have common coefficients across the two cohorts.*

	Cohort C		Cohort B	
Covariates	Estimates	p-values	Estimates	p-values
Intercept	1.156	<0.05	1.156	<0.05
log(CBL)	0.006	0.730	0.006	0.730
B_1(log(CBL))	0.035	0.330	1.561	0.980
B_2(log(CBL))	−0.111	0.052	−0.017	0.993
Gender	0.036	<0.05	0.036	<0.05
B_1(age)	0.910	<0.05	0.910	<0.05
B_2(age)	1.303	<0.05	1.303	<0.05
B_3(age)	1.561	<0.05	1.726	<0.05

tiple longitudinal studies. The proposed method accounts for various aspects of heterogeneity across studies so the resulting estimation and inference are not only synthesized with integrated data, but also adaptive to individual study features.

The innovation of our method lies in the strategies of handling multiple datasets with study-specific missing covariates, which often occur in data merging. When datasets of multiple studies are collected respectively from different subpopulations, it is problematic to use studies with fully observed data either to impute study-specific missing covariates or to adjust the chance of missingness by the method of inverse probability weighting. The failure of inverse probability weighting or doubly robust approaches lies in the fact that the regression coefficients for the completely observed data are not the same as in the missing data (happens in a different study in our setting), so that no appropriate weights can be allocated in this study-specific missing structure. Our approach features a sieve nonparametric estimation of a marginalized mean model which is resulted from integrating the set of missing covariates out of the original mean model in (1). Under Assumption 1, the marginalized mean model can be estimated properly by using studies with fully observed covariates and hence the resulting estimation for regression coefficients is consistent and asymptotically normal.

In addition, the implementation of our method is numerically straightforward. Both theoretical and numerical evidences are provided to show the large-sample properties and finite-sample performances of the proposed method. Although our method is developed using balanced longitudinal data, it can be applied to unbalanced longitudinal data with no additional burden. Please refer to Song et al. (2009) for details. Since our method relies on the nonparametric estimation of the marginalized estimating functions, it could be challenged when the number of observed covariates is large. Also when the number of studies is large, it would be computationally demanding to use traditional hypothesis testing method to determine commonly shared parameters across studies in the joint analysis. Providing a flexible and efficient way to detect common parameters in multiple studies in the presence of missing covariates is worth future exploration.

9. Supplementary Materials

Web Appendices referenced in Sections 2 to 6 are available with this paper at the *Biometrics* website on Wiley Online Library. The computer program in the R language implementing the proposed method and an illustration data example are also available at this website.

Acknowledgements

The authors would like to thank the editor, the associate editor, and two reviewers for their very helpful comments. Song's research is supported by an NSF Grant (DMS #1208939).

References

Afeiche, M., Peterson, K. E., Snchez, B. N., Cantonwine, D., Lamadrid-Figueroa, H., Schnaas, L., et al. (2011). Prenatal lead exposure and weight of 0- to 5-year-old children in mexico city. *Environmental Health Perspective* **119**, 1436–1441.

Carroll, R. J., Fan, J., Gijbels, I., and Wand, M. (1998). Generalized partially linear single-index models. *Journal of the American Statistical Association* **92**, 477–489.

Chen, Q. and Ibrahim, J. G. (2006). Semiparametric models for missing covariate and response data in regression models. *Biometrics* **62**, 177–184.

Chen, X., Hong, H., and Tamer, E. (2005). Measurement error models with auxiliary data. *Review of Economic Studies* **72**, 343–366.

Chen, X., Linton, O., and Keilegom, I. V. (2003). Estimation of semiparametric models when the criterion function is not smooth. *Econometrica* **71**, 1591–1608.

Craven, P. and Wahba, G. (1979). Smoothing noisy data with spline functions. *Numerische Mathematik* **31**, 377–403.

Demnati, A. and Rao, J. N. K. (2004). Linearization variance estimators for survey data. *Survey Methodology* **30**, 17–26.

Demnati, A. and Rao, J. N. K. (2010). Linearization variance estimators for model parameters from complex survey data. *Survey Methodology* **36**, 193–202.

Hall, P. and Horowitz, J. L. (1996). Bootstrap critical values for tests based on generalized method of moments estimators. *Econometrica* **64**, 891–916.

He, X., Zhu, Z.-Y., and Fung, W.-K. (2002). Estimation in a semiparametric model for longitudinal data with unspecified dependence structure. *Biometrika* **89**, 579–590.

Hu, F. and Zidek, J. V. (2002). The weighted likelihood. *The Canadian Journal of Statistics* **30**, 347–371.

Ichimura, H. (1993). Semiparametric least squares (SLS) and weighted SLS estimation of single-index models. *Journal of Econometrics* **58**, 71–120.

Kim, J. K. (2011). Parametric fractional imputation for missing data analysis. *Biometrika* **98**, 119–132.

Leng, C., Zhang, W., and Pan, J. (2010). Semiparametric mean covariance regression analysis for longitudinal data. *Journal of the American Statistical Association* **105**, 181–193.

Little, R. J. and Zhang, N. (2011). Subsample ignorable likelihood for regression analysis with missing data. *Journal of the Royal Statistical Society: Series C* **60**, 591–605.

Little, R. J. A. (1992). Regression With Missing X's: A Review. *Journal of the American Statistical Association* **87**, 1227–1237.

Little, R. J. A. (1993). Pattern-mixture models for multivariate incomplete data. *Journal of the American Statistical Association* **88**, 125–134.

Little, R. J. A. and Rubin, D. B. (2002). Wiley Series in Probability and Statistics. *Statistical Analysis with Missing Data*, New York, NY: Wiley.

Molenberghs, G. and Kenward, M. G. (2007). Missing Data in Clinical Studies. New York, NY: Wiley.

Newey, W. K. (1994). The asymptotic variance of semiparametric estimators. *Econometrica* **62**, 1349–1382.

Newey, W. K. (1997). Convergence rates and asymptotic normality for series estimators. *Journal of Econometrics* **79**, 147–168.

Qu, A., Lindsay, B. G., and Li, B. (2000). Improving generalised estimating equations using quadratic inference functions. *Biometrika* **87**, 823–836.

Robins, J. M., Rotnitzky, A., and Zhao, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association* **89**, 846–866.

Rubin, D. B. (1976). Inference and missing data. *Biometrika* **63**, 581–592.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics* **6**, 461–464.

Shi, X., Zhu, H., and Ibrahim, J. G. (2009). Local influence for generalized linear models with missing covariates. *Biometrics* **65**, 1164–1174.

Song, P. X.-K. X., Jiang, Z., Park, E., and Qu, A. (2009). Quadratic inference functions in marginal models for longitudinal data. *Statistics in Medicine* **28**, 3683–3696.

Wang, F., Wang, L., and Song, P. X. K. (2012). Quadratic inference function approach to merging longitudinal studies: Validation and joint estimation. *Biometrika* **99**, 755–762.

Wang, L. and Qu, A. (2009). Consistent model selection and data-driven smooth tests for longitudinal data in the estimating equations approach. *Journal of The Royal Statistical Society: Series B.* **71**, 177–190.

Wang, X. and Zidek, J. V. (2005). Selecting likelihood weights by cross-validation. *The Annals of Statistics* **33**, 463–500.

Received April 2014. *Revised April* 2015. *Accepted May* 2015.