A Semiparametric Single-Index Risk Score Across Populations

Shujie Ma
Department of Statistics, University of California at Riverside, Riverside, CA 92521,
shujie.ma@ucr.edu

Yanyuan Ma
Department of Statistics, University of South Carolina, Columbia, SC 29208,
yanyuanma@stat.sc.edu

Yanqing Wang
Fred Hutchinson Cancer Research Center, Seattle, WA 98109, ywang237@fredhutch.org

Eli S. Kravitz
Department of Statistics, Texas A&M University, 3143 TAMU, College Station, TX
77843-3143, kravitze@tamu.edu

Raymond J. Carroll
Department of Statistics, Texas A&M University, 3143 TAMU, College Station, TX
77843-3143, and School of Mathematical Sciences, University of Technology Sydney,
Broadway NSW 2007, carroll@stat.tamu.edu

Abstract

We consider a problem motivated by issues in nutritional epidemiology, across diseases and populations. In this area, it is becoming increasingly common for diseases to be modeled by a single diet score, such as the Healthy Eating Index, the Mediterranean Diet Score, etc. For each disease and for each population, a partially linear single-index model is fit. The partially linear aspect of the problem is allowed to differ in each population and disease. However, and crucially, the single-index itself, having to do with the diet score, is common to all diseases and populations, and the nonparametrically estimated functions of the single-index are the same up to a scale parameter. Using B-splines with an increasing number of knots, we develop a method to solve the problem, and display its asymptotic theory. An application to the NIH-AARP Study of Diet and Health is described, where we show the advantages of using multiple diseases and populations simultaneously rather than one at a time in understanding the effect of increased Milk consumption. Simulations illustrate the properties of the methods.

Some Key Words: Asymptotic theory; B-splines; Combining data sets; Healthy Eating Index; Logistic regression; Partially linear single-index models; Semiparametric models; Single-index models.

Short title: Single Index Models Across Populations

1 Introduction

We describe a novel partially linear logistic single index model in which there are multiple populations, and multiple diseases within each population, but where the single index part of the model is shared across the populations and diseases. In the case of a single disease across independent populations, we derive B-spline based semiparametric efficient methodology. In other cases, such as multiple populations with multiple diseases, our B-spline based methods are consistent and we derive their asymptotic theory.

The problem arises from common practice in nutritional epidemiology, where the goal is to relate nutritional intakes to disease. In this area, it is increasingly common to relate the patterns of multiple dietary components, rather than an individual dietary component, to a disease. One popular way to summarize dietary intake patterns is through a dietary pattern score. While there are many flavors of dietary pattern scores, in our example we use the U.S. Department of Agriculture's (USDA's) Healthy Eating Index-2005 (HEI-2005, http://www.cnpp.usda.gov/HealthyEatingIndex.htm). It is based on the key recommendations of the 2005 Dietary Guidelines for Americans available at http://www.health.gov/dietaryguidelines/dga2005/document/default.htm. The HEI-2005 comprises 12 distinct component scores. Intakes of each food or nutrient, represented by one of the 12 components, are expressed as a ratio to energy (caloric) intake, assessed, and given a score. See Table 1 for a list of these components and the standards for scoring, and see Guenther et al. (2008) and Guenther et al. (2008) for details. The 12 different component scores are then summed to get a total score, ranging from 0 for a terrible diet to 100 for the best possible diet.

The key concept here is that the total score is developed before any health outcome data are considered. *Once a total score is developed, it is then used, across multiple populations, in risk models to relate any disease to the total score.* As an example, Panagiotakos et al. (2006) show that for colorectal cancer in the NIH-AARP Study of Diet and Health (Schatzkin et al., 2001), with diet assessed by a food frequency questionnaire, higher HEI-2005 total scores are statistically significantly associated with lower disease risks. They also consider three other dietary pattern scores. George et al. (2010) show that among breast cancer survivors, higher

HEI-2005 total scores are associated with lower levels of chronic inflammation. Chiuve et al. (2012) show that the HEI-2005 total score and the Alternative Healthy Eating Index (AHEI) (McCullough et al., 2002) are significant predictors of chronic diseases such as coronary heart disease, diabetes, stroke and cancer, and that closer adherence to the 2005 Dietary Guidelines may lower the risk of major chronic diseases. The AHEI is also associated with all cause mortality (Akbaraly et al., 2011).

In its most general form, there are $k = 1, ..., K$ populations. Within population k, there are $\ell = 1, ..., L_k$ diseases. As in the HEI-2005, there are $j = 0, ..., J$ dietary components. Let the $J + 1$ individual component scores in population k be $(X_{0k}, ..., X_{Jk})$: in our case, $J + 1 = 12$. Then the current practice in nutritional epidemiology would be to form a total score $\sum_{j=0}^{J} X_{jk}$ for population k and use it as the risk predictor *for all populations/diseases.* Thus, for example, in a logistic regression with a binary outcome $Y_{k\ell}$, and with $H(\cdot)$ being the logistic distribution function, the model for population k and disease ℓ is

$$\text{pr}(Y_{k\ell} = 1 | X_{0k}, ..., X_{Jk}) = H(\beta_{0k\ell} + \beta_{1k\ell} \textstyle\sum_{j=0}^{J} X_{jk}). \tag{1}$$

This is important and convenient from a public health perspective, because it enables nutritional epidemiologists to use the same predictor, namely $\sum_{j=0}^{J} X_{jk}$, for all diseases, and to describe the effect of that predictor through a single quantity, $\beta_{1k\ell}$.

Crucially, it is undesirable to try to fit different parameters for each component score. That is, instead of fitting (1), one might fit

$$\text{pr}(Y_{k\ell} = 1 | X_{0k}, ..., X_{Jk}) = H(\beta_{0k\ell} + \textstyle\sum_{j=0}^{J} \beta_{jk\ell} X_{jk}). \tag{2}$$

The reason why (1) is preferred to (2) for public health purposes is that it is much more interpretable. Model (1) describes how a *single, interpretable score*, $\sum_{j=0}^{J} X_{jk}$, affects disease risk. Model (2) is chaotic because it requires policy makers to say things such as "if you are in population $k = 1$ and are worried about disease $\ell = 1$ then your diet improves your risk if you eat this kind of food more and that kind of food less, but for disease $\ell = 2$ you need to consider your dietary composition in another way". Interpretability is even more complicated because the component scores have a reasonably complex pattern of correlations, see Table S.1 of the **Supplementary Material**. Since there are so many diseases and populations, this is not helpful practically and would not be used. Indeed, as seen above, the single HEI-2005 score is associated with colon cancer (Reedy et al., 2008), chronic inflammation, (George et al., 2010) and many chronic diseases (Chiuve et al., 2012), and its ease of interpretation is apparent.

Our goal is to develop a single interpretable score that, unlike the HEI-2005 score or other scores, is calibrated to different populations or diseases. We do this not by summing the component scores, but by weighting them and allowing a more flexible shape. Thus, for an unknown function $m(\cdot)$ and weights $(\alpha_0, ..., \alpha_J)$, we propose the single-index score $m(\sum_{j=0}^{J} X_{jk}\alpha_j)$, and model the risk as

$$\mathrm{pr}(Y_{k\ell} = 1|X_{0k}, ..., X_{Jk}) = H\{\beta_{0k\ell} + \beta_{1k\ell} m(\textstyle\sum_{j=0}^{J} X_{jk}\alpha_j)\}. \quad (3)$$

Model (3), like model (1), is based upon a single interpretable score, $m(\sum_{j=0}^{J} X_{jk}\alpha_j)$, that can be used across populations and/or diseases.

In Section 2, we present our model more formally. Section 3 describes how to fit the model for a single disease across independent populations, and we show that our method is semiparametric efficient in this case. In Section 4 we describe generalizations. Section 4.1 considers a single population with multiple diseases, while Section 4.2 describes the real goal of our data analysis, where there are multiple populations and multiple diseases. Section 5 gives results of the data example, while Section 6 describes simulation results. Computational and technical details are in an Appendix and in **Supplementary Material**.

2 Multiple Population Single-Index Model

2.1 Model and Splines

In this section, we consider a single disease across multiple different populations. There are $k = 1, ..., K$ populations. For the k^{th} population, there are $i = 1, ..., n_k$ individuals with binary responses Y_{ik}. Define $X_{ik} = (X_{ik1}, \ldots, X_{ikJ})^{\mathrm{T}}$. In addition to the responses, for individual i in population k we observe $(G_{ik}, X_{ik0}, X_{ik}, Z_{ik}, G_{ik}W_{ik})$, defined as follows. The $J + 1$ dietary component scores are $(X_{ik0}, X_{ik}^{\mathrm{T}})^{\mathrm{T}}$. Covariates that are observed for all individuals are $Z_{ik} = (Z_{ik1}, \ldots, Z_{ikd})^{\mathrm{T}}$. Further, we allow for a subset of individuals to have additional covariates $W_{ik} = (W_{ik1}, \ldots, W_{ika})^{\mathrm{T}}$, and define G_{ik} to be the binary indicator that these individuals have such additional covariates. For example, Reedy et al. (2008) fit models to men and women with many common covariates such as age and levels of educational status, but for women they in addition include indicators of types of hormone replacement therapy.

For the i^{th} individual and the k^{th} outcome, we posit the marginal model

$$\begin{aligned} &\mathrm{pr}(Y_{ik} = 1 \mid G_{ik}, X_{ik0}, X_{ik}, Z_{ik}, G_{ik}W_{ik}) \\ &\quad = H_{ik} = H\{(\beta_{k1} + \beta_{k2}G_{ik})m(X_{ik0} + X_{ik}^{\mathrm{T}}\alpha) + Z_{ik}^{\mathrm{T}}(\theta_{k1} + \theta_{k2}G_{ik}) + G_{ik}W_{ik}^{\mathrm{T}}\theta_{k3}\}. \end{aligned} \tag{4}$$

Here $\alpha \in \mathbb{R}^J$, $\theta_{k1} \in \mathbb{R}^d$, $\theta_{k2} \in \mathbb{R}^d$, $\theta_{k3} \in \mathbb{R}^a$, $\beta_{k1} \in \mathbb{R}$ and $\beta_{k2} \in \mathbb{R}$. Crucially, for use in practice, the function $m(\cdot)$ and the parameter α do not depend on k.

Remark 1 In model (4), the most general form of the single-index is $m(X_{ik0}\alpha_0 + X_{ik}^{\mathrm{T}}\alpha)$. However, because $m(\cdot)$ is modeled nonparametrically, such a formulation is not identifiable. There are three equivalent ways to obtain identifiability. The first common method, what we have done, is to select one variable that is known to be related to the response, which we label as X_{ik0}, and to set its parameter $\alpha_0 = 1$. A second common method is to make the restriction that $\alpha_0^2 + \alpha^{\mathrm{T}}\alpha = 1$. In the context of our problem, there is a third way. Since common nutritional epidemiology practice is to weight each variable the same, namely $= 1$, the sum of the weights $= J + 1$. For comparison purposes, we can achieve identifiability via the restriction $\alpha_0 + (1, ..., 1)\alpha = J + 1$. We use the first method in our computations, but report results for the third.

Model (4) generalizes the now-classical generalized partially linear single-index model (Carroll et al., 1997), with the novelty being both the context and that the same single-index $m(X_{ik0} + X_{ik}^{\mathrm{T}}\alpha)$ is used across multiple outcomes.

Single-index models have been widely used as a popular tool in multivariate nonparametric regression to alleviate the "curse of dimensionality" (Bellman, 1961). For example, recently Yu and Ruppert (2002) used penalized spline least squares estimation for single-index models with independent and identically distributed observations: their number of knots was fixed, unlike in our development. Wang and Yang (2009a) proposed polynomial spline estimation and extended the results to weakly dependent response variables. Cui et al. (2011) developed a kernel estimating function method for generalized single-index models, while Ma and Zhu (2013) constructed robust and efficient estimation with high dimensional covariates. These papers are restricted to the case of one population, $K = 1$, and one outcome, $L = 1$. Here we consider multiple populations and multiple outcomes. We propose a regression spline based profile estimation procedure and establish the asymptotic properties of the estimators in model (4).

To set the main ideas clearly, it is convenient to first assume that the Y_{ik} are independent, see Section 4 for the more general cases of interest in the HEI-2005 problem. To ensure identifiability, we set $\beta_{11} = -1$, and we set the first component of $\theta_{11} = (\theta_{111}, ..., \theta_{11d})$ equal to zero, i.e., $\theta_{111} = 0$. Hence, our parameters are (ν, m), where m is an unspecified function that is sufficiently smooth, while

$$\nu = (\alpha^{\mathrm{T}}, \beta_{12}, \beta_{21}, \beta_{22}, \ldots, \beta_{K1}, \beta_{K2}, \theta_{112}, \ldots, \theta_{11d}, \theta_{12}^{\mathrm{T}}, \theta_{13}^{\mathrm{T}}, \theta_{21}^{\mathrm{T}}, \theta_{22}^{\mathrm{T}}, \theta_{23}^{\mathrm{T}}, \ldots, \theta_{K1}^{\mathrm{T}}, \theta_{K2}^{\mathrm{T}}, \theta_{K3}^{\mathrm{T}})^{\mathrm{T}}.$$

Thus, ν has total dimension $d_\nu = J + 2K + 2Kd + Ka - 2$.

Let $U_{ik}(\alpha) = X_{ik0} + X_{ik}^{\mathrm{T}}\alpha$ be the realizations of $U_k(\alpha) = X_{k0} + X_k^{\mathrm{T}}\alpha$. The unknown function $m(\cdot)$ is estimated by polynomial splines described as follows. Without loss of generality, assume $u \in [a, b]$. Let $N = N_n$ be the number of interior knots. Divide $[a, b]$ into $(N + 1)$ subintervals $I_p = \{(\xi_p, \xi_{p+1}), p = r, r + 1, \ldots, N + r - 1\}$, $I_N = (\xi_{N+r}, 1)$, where $(\xi_p)_{p=r+1}^{N+r}$ is a sequence of interior knots, given as

$$\xi_1 = \cdots = a = \xi_r < \xi_{(r+1)} < \cdots \xi_{(r+N)} < b = \xi_{N+r+1} = \cdots = \xi_{N+2r}.$$

Define the distance between neighboring knots as $h_p = \xi_{p+1} - \xi_p, r \leq p \leq N + r$, and $h = \max_{r \leq p \leq N+r} h_p$. Let G_n be the space of B-splines of order r, so that $P_n = N + r$ is the number of functions in G_n. For $u \in [a, b]$, let G_n be the linear space spanned by the B-spline functions $B_r(u) = \{B_{r,p}(u), 1 \leq p \leq P_n\}^{\mathrm{T}}$. Then $m(u)$ can be approximated by $\widetilde{m}(u) = \sum_{p=1}^{P_n} B_{r,p}(u)\lambda_p = B_r^{\mathrm{T}}(u)\lambda$, where $\lambda = (\lambda_1, \ldots, \lambda_{P_n})^{\mathrm{T}}$. B-splines have been used frequently to estimate the nonparametric functions in nonparametric and semiparametric models because they are easy to compute with derivable asymptotic theory. See Huang

(2003) and Wang and Yang (2009b) for their utility in nonparametric models, Stone (1985) and Huang and Yang (2004) in additive models, and Huang et al. (2002), Liu et al. (2011) and Wang et al. (2011) in semiparametric models.

3 Profile Estimating Procedure

Our estimation is performed through a conceptually simple profiling procedure, as described below.

Step 1. Define

$$\widetilde{H}_{ik} = H[(\beta_{k1} + \beta_{k2} G_{ik}) B_r^{\mathrm{T}}\{U_{ik}(\alpha)\}\lambda + Z_{ik}^{\mathrm{T}}(\theta_{k1} + \theta_{k2} G_{ik}) + G_{ik} W_{ik}^{\mathrm{T}} \theta_{k3}].$$

Treating ν as a fixed parameter, estimate $m(u)$ by spline functions $\widehat{m}(u,\nu) = \sum_{p=1}^{P_n} B_{r,p}(u)\widehat{\lambda}_p(\nu)$ with $\widehat{\lambda}(\nu) = \{\widehat{\lambda}_1(\nu), \cdots, \widehat{\lambda}_{P_n}(\nu)\}^{\mathrm{T}}$ through maximizing

$$L_n(\lambda,\nu) = \textstyle\sum_{k=1}^{K}\sum_{i=1}^{n_k}\{Y_{ik}\mathrm{log}(\widetilde{H}_{ik}) + (1-Y_{ik})\mathrm{log}(1-\widetilde{H}_{ik})\}. \qquad (5)$$

To prepare for the second step, we perform the following additional calculations. Let $B_{r-1}(u) = \{B_{r-1,p}(u) : 2 \le p \le P_n\}^{\mathrm{T}}$ be the B-spline functions of order $r-1$. We estimate $m'(u,\nu)$, the first derivative of m, through $\widehat{m}'(u,\nu) = \sum_{p=2}^{P_n} B_{r-1,p}(u)\widehat{\lambda}_p^{(1)}(\nu)$, where $\widehat{\lambda}_p^{(1)}(\nu) = (r-1)\{\widehat{\lambda}_p(\nu) - \widehat{\lambda}_{p-1}(\nu)\}/(\xi_{p+r-1} - \xi_p)$, for $2 \le p \le P_n$. This is because the first derivative of a spline function can be expressed in terms of a spline of one order lower see page 116 of de Boor (2001). Let $D = (d_{jj'})_{1\le j,j' \le P_n - 1}$ be a $(P_n - 1)\times(P_n - 1)$ diagonal matrix with $d_{jj} = 1/(\xi_{j+r} - \xi_{j+1})$ and $d_{jj'} = 0$ for $j \neq j'$, and let $D_{11} = (-D, 0_{P_n-1})_{(P_n-1)\times P_n}$ and $D_{12} = (0_{P_n-1}, D)_{(P_n-1)\times P_n}$, where 0_{P_n-1} is the (P_n-1)-dimensional vector with 0's as its elements. Then $\widehat{m}'(u,\nu) = B_{r-1}^{\mathrm{T}}(u) D_1 \widehat{\lambda}(\nu)$, where $D_1 = (r-1)(D_{11} + D_{12})$. For $u \in [a,b]$, define

$$\widehat{\sigma}^2(u,\nu) = B_r^{\mathrm{T}}(u)[\textstyle\sum_{k=1}^{K}\sum_{i=1}^{n_k} V_{ik}(\nu)(\beta_{k1} + \beta_{k2}G_{ik})^2 B_r\{U_{ik}(\alpha)\} B_r\{U_{ik}(\alpha)\}^{\mathrm{T}}]^{-1} B_r(u), \qquad (6)$$

where $V_{ik}(\nu) = H_{ik}(\nu)(1 - H_{ik}(\nu))$, and

$$H_{ik}(\nu) = H\{(\beta_{k1} + \beta_{k2}G_{ik}) m(X_{ik0} + X_{ik}^{\mathrm{T}}\alpha) + Z_{ik}^{\mathrm{T}}(\theta_{k1} + \theta_{k2}G_{ik}) + G_{ik}W_{ik}^{\mathrm{T}}\theta_{k3}\}.$$

Step 2. Define

$$\widehat{H}_{ik}(\nu) = H\{(\beta_{k1} + \beta_{k2}G_{ik})B_r^{\mathrm{T}}(U_{ik}(\alpha))\widehat{\lambda}(\nu) + Z_{ik}^{\mathrm{T}}(\theta_{k1} + \theta_{k2}G_{ik}) + G_{ik}W_{ik}^{\mathrm{T}}\theta_{k3}\}. \quad (7)$$

Estimate ν by $\widehat{\nu}$ through maximizing

$$L_n(\nu) = \textstyle\sum_{k=1}^{K}\sum_{i=1}^{n_k}[Y_{ik}\log\{\widehat{H}_{ik}(\nu)\} + (1 - Y_{ik})\log\{1 - \widehat{H}_{ik}(\nu)\}].$$

Once we obtain $\widehat{\nu}$, we can plug it into $\widehat{m}$ to obtain $\widehat{\nu}, \widehat{m}(u, \widehat{\nu})$ as the final estimator. To prepare the description of the asymptotic properties of our procedure, we define

$$\begin{aligned} Q_{i1}(\nu) \;=\; & \big\{(-1 + \beta_{12}G_{i1})m'(X_{i10} + X_{i1}^{\mathrm{T}}\alpha)X_{i1}^{\mathrm{T}}, G_{i1}m(X_{i10} + X_{i1}^{\mathrm{T}}\alpha), 0_{1,2K-2}, Z_{1i2}, \cdots, \\ & Z_{i1d}, G_{i1}Z_{i1}^{\mathrm{T}}, G_{i1}W_{i1}^{\mathrm{T}}, 0_{1,(K-1)(2d+a)}\big\}^{\mathrm{T}}, \end{aligned}$$

and for $k = 2, ..., K$ define

$$\begin{aligned} Q_{ik}(\nu) \;=\; & \big\{(\beta_{k1} + \beta_{k2}G_{ik})m'(X_{ik0} + X_{ik}^{\mathrm{T}}\alpha)X_{ik}^{\mathrm{T}}, 0_{1,2k-3}, m(X_{ik0} + X_{ik}^{\mathrm{T}}\alpha), G_{ik}m(X_{ik0} + X_{ik}^{\mathrm{T}}\alpha), \\ & 0_{1,2K-2k+(k-1)(2d+a)-1}, Z_{ik}^{\mathrm{T}}, G_{ik}Z_{ik}^{\mathrm{T}}, G_{ik}W_{ik}^{\mathrm{T}}, 0_{1,(K-k)(2d+a)}\big\}^{\mathrm{T}}. \end{aligned}$$

Denote the elements of $Q_{ik}(\nu) = (Q_{ik,\ell}(\nu))_{l=1}^{d_\nu}$. Let

$$\nu^0 = (\alpha^{0\mathrm{T}}, \beta_{12}^0, \beta_{21}^0, \beta_{22}^0, \ldots, \beta_{K1}^0, \beta_{K2}^0, \theta_{112}^0, \ldots, \theta_{11d}^0, \theta_{12}^{0\mathrm{T}}, \theta_{13}^{0\mathrm{T}}, \theta_{21}^{0\mathrm{T}}, \theta_{22}^{0\mathrm{T}}, \theta_{23}^{0\mathrm{T}}, \ldots, \theta_{K1}^{0\mathrm{T}}, \theta_{K2}^{0\mathrm{T}}, \theta_{K3}^{0\mathrm{T}})^{\mathrm{T}}$$

be the collection of the true parameters. Let $[a_0, b_0]$ be the support of $X_{k0} + X_k^{\mathrm{T}}\alpha^0$, where α^0 is the true population parameter. Denote $\|\cdot\|$ as the L_2 norm of any square integrable function on $[a_0, b_0]$. For $1 \leq \ell \leq d_\nu$, let $\eta_\ell^0(\cdot)$ be the function $\eta_\ell(\cdot) \in \mathrm{L}_2([a_0, b_0])$ that minimizes $E[\sum_{k=1}^{K} n_k\{Q_{ik,\ell}(\nu^0) - (\beta_{k1}^0 + \beta_{k2}^0G_{ik})\eta_\ell(U_{ik}\,(\alpha^0))\}^2V_{ik}]$. Also, define $\eta^0\{U_{ik}(\alpha)\} = \{\eta_1^0\{U_{ik}(\alpha)\}, \ldots, \eta_{d_\nu}^0\{U_{ik}(\alpha)\}]^{\mathrm{T}}$. For simplicity of notations, we let $V_{ik} = V_{ik}(\nu^0)$, $Q_{ik} = Q_{ik}(\nu^0)$ and $U_{ik} = U_{ik}(\alpha^0)$. Let $n = \sum_{k=1}^{K} n_k$.

In the following three theorems, we establish the consistency, asymptotic normality and efficiency of our procedure.

<u>Theorem</u> 1 *Under the conditions in Appendix A.2, when ν is the collection of the true parameters or a $\sqrt{n}$-consistent estimator of ν^0, (a) $|\widehat{m}(u, \nu) - m(u)| = O_p\{(nh)^{-1/2} + h^q\}$ uniformly in $u \in [a_0, b_0]$; (b) $|\widehat{m}'(u, \nu) - m'(u)| = O_p(n^{-1/2}h^{-3/2} + h^{q-1})$ uniformly in $u \in [a_0, b_0]$; and (c) as $n \to \infty$, $\widehat{\sigma}^{-1}(u, \nu^0)\{\widehat{m}(u, \nu) - m(u)\} \to$ Normal$(0, 1)$.*

Theorem 2 *Define $n = n_1$ and for $k = 2, ..., K$, define $n_k = nc_{kn}$, where there are constants $c_* > 0$ and $c_{**} < \infty$ such that $c_* \le c_k = \lim_{n\to\infty} c_{kn} \le c_{**}$. Under the conditions in Appendix A.2, $\|\widehat{\nu} - \nu^0\|_2 = O_p\left(n^{-1/2}\right)$, and*

$$\begin{aligned} n^{1/2}(\widehat{\nu} - \nu^0) \;\; &= \;\; [n^{-1}\textstyle\sum_{k=1}^{K}\sum_{i=1}^{n_k} H_{ik}(1 - H_{ik})(Q_{ik} - (\beta_{k1}^0 + \beta_{k2}^0 G_{ik})\eta^0(U_{ik}))^{\otimes 2}]^{-1} \\ &\quad \times [n^{-1/2}\textstyle\sum_{k=1}^{K}\sum_{i=1}^{n_k}(Y_{ik} - H_{ik})\{Q_{ik} - (\beta_{k1}^0 + \beta_{k2}^0 G_{ik})\eta^0(U_{ik})\}] + o_p(1), \end{aligned} \tag{8}$$

for $\widehat{\nu}$ in a neighborhood of ν^0. Then as $n \to \infty$, $n^{1/2}(\widehat{\nu} - \nu^0) \to Normal(0_{d_\nu}, \mathbf{\Sigma})$, where

$$\mathbf{\Sigma} = (\textstyle\sum_{k=1}^{K} c_k E[V_{ik}\{Q_{ik} - (\beta_{k1}^0 + \beta_{k2}^0 G_{ik})\eta^0(U_{ik})\}^{\otimes 2}])^{-1},$$

and 0_{d_ν} is a d_ν-dimensional vector with "0" as its elements. Here and throughout the text, $a^{\otimes 2} \equiv aa^{\mathrm{T}}$ for any matrix or vector a.

In practice, $\mathbf{\Sigma}$ is estimated by

$$\widehat{\mathbf{\Sigma}} = n(\textstyle\sum_{k=1}^{K}\sum_{i=1}^{n_k}[\widehat{V}_{ik}(Q_{ik}(\widehat{\nu}) - \widehat{\Pi}_n Q_{ik}(\widehat{\nu}))^{\otimes 2}])^{-1},$$

where $\widehat{V}_{ik} = \widehat{H}_{ik}(\widehat{\nu})(1 - \widehat{H}_{ik}(\widehat{\nu}))$, $\widehat{\Pi}_n Q_{ik}(\widehat{\nu}) = (\widehat{\Pi}_n Q_{ik,\ell}(\widehat{\nu}), 1 \le \ell \le d_\nu)^{\mathrm{T}}$, and for $1 \le l \le d_\nu$, $\widehat{\Pi}_n Q_{ik,\ell}(\widehat{\nu}) = (\widehat{\beta}_{k1} + \widehat{\beta}_{k2} G_{ik}) B_r^{\mathrm{T}}(U_{ik}(\widehat{\alpha}))\widehat{\delta}_\ell$, where

$$\widehat{\delta}_\ell = \arg\min_{\delta_\ell \in R^{P_n}} \textstyle\sum_{k=1}^{K}\sum_{i=1}^{n_k} \left\{Q_{ik,\ell}(\widehat{\nu}) - ((\widehat{\beta}_{k1} + \widehat{\beta}_{k2} G_i) B_r^{\mathrm{T}}(U_{ik}(\widehat{\alpha}))\delta\right\}^2 \widehat{V}_{ik}.$$

In addition, under the assumption of independence, or conditional independence of the Y_{ik} given the covariates, our estimation method is semiparametric efficient. We state this as

Theorem 3 *Under the conditions in Appendix A.2, profile likelihood estimation of the parameter ν reaches the semiparametric efficiency bound. The minimum variance bound for estimating ν can be further simplified to*

$$cov_{\mathrm{opt}}\{n^{1/2}(\widehat{\nu} - \nu^0)\} = \left\{E\left(\textstyle\sum_{k=1}^{K} c_k V_{ik}\left[Q_{ik}Q_{ik}^{\mathrm{T}} - (\beta_{k1}^0 + \beta_{k2}^0 G_{ik})^2\{\eta^0(U_{ik})\}^{\otimes 2}\right]\right)\right\}^{-1}.$$

The proofs of the theorems are given in the Appendix.

4 Generalizations

4.1 Single Population, Multiple Diseases

In this section, we relax the assumption of independence of the Y_{ik} given the covariates, and consider the case of a single population with K outcomes, with a common sample size n. The response indicators remain as $(Y_{i1}, ..., Y_{iK})$, but now the covariates are the same for each response, and are written as $\mathbb{C}_i = (G_i, X_{i0}, X_i, Z_i, G_i W_i)$, and now we use $U_i(\alpha) = X_{i0} + X_i^{\mathrm{T}}\alpha$. Ignoring this correlation and invoking a "working independence" principle, the profile likelihood procedure described in Section 3 will still provide consistent estimation. However, more efficient estimation can be generally obtained through taking into account the correlation structure.

Specifically, the derivative of $Y_{ik}\mathrm{log}(\widetilde{H}_{ik}) + (1 - Y_{ik})\mathrm{log}(1 - \widetilde{H}_{ik})$ with respect to λ is $(Y_{ik} - \widetilde{H}_{ik})(\beta_{k1} + \beta_{k2}G_i)B_r(U_i(\alpha))$. Translated to the setting of this section, Step 1 in Section 3 is equivalent to solving

$$\textstyle\sum_{k=1}^{K} \sum_{i=1}^{n_k}(Y_{ik} - \widetilde{H}_{ik})(\beta_{k1} + \beta_{k2}G_i)B_r(U_i(\alpha)) = 0.$$

Here, we modify this step to

Step 1d. Let $\Omega_i = (\Omega_{i,k,k'})_{k,k'=1}^{K}$ represent a working covariance matrix of $(Y_{i1}, ..., Y_{iK})$ conditional on $\mathbb{C}_i$. Let $\mathbf{B}_i(\nu)$ be a $K \times K$ matrix with the (k, k') entry $\mathbf{B}_{i,k,k'}(\nu) = \Omega_{i,k,k'}(\beta_{k1} + \beta_{k2}G_i)(\beta_{k'1} + \beta_{k'2}G_i)$. Obtain $\widehat{\lambda}_w(\nu)$ by solving $\sum_{i=1}^{n} B_r\{U_i(\alpha)\}\widetilde{\mathbf{A}}_i(\nu)\mathbf{B}_i(\nu)^{-1}\mathbf{\Phi}_i(\nu) = 0$, where

$$\begin{aligned}
\mathbf{\Phi}_i(\nu) &= \{(Y_{i1} - \widetilde{H}_{i1})(\beta_{11} + \beta_{12}G_i), \ldots, (Y_{iK} - \widetilde{H}_{iK})(\beta_{K1} + \beta_{K2}G_i)\}^{\mathrm{T}}, \\
\widetilde{\mathbf{A}}_i(\nu) &= \{\widetilde{V}_{i1}(\beta_{11} + \beta_{12}G_i)^2, \ldots, \widetilde{V}_{ik}(\beta_{K1} + \beta_{K2}G_i)^2\},
\end{aligned}$$

where $\widetilde{V}_{ik} = \widetilde{H}_{ik}(1 - \widetilde{H}_{ik})$.

Using $\widehat{\lambda}_w(\nu)$, we form the corresponding estimators of $m(u)$ and $m'(u)$, which are $\widehat{m}_w(u, \nu) = B_r^{\mathrm{T}}(u)\widehat{\lambda}_w(\nu)$ and $\widehat{m}_w'(u, \nu) = \{B_r'(u)\}^{\mathrm{T}}\widehat{\lambda}_w(\nu)$. Define $\mathbf{A}_i(\nu) = \{V_{i1}(\beta_{11} + \beta_{12}G_i)^2, \ldots, V_{ik}(\beta_{K1} + \beta_{K2}G_i)^2\}$. Let $\mathbf{A}_i = \mathbf{A}_i(\nu^0)$ and $\mathbf{B}_i = \mathbf{B}_i(\nu^0)$. For $u \in [a, b]$, define $\widehat{\sigma}_w^2(u,\nu) = B_r^{\mathrm{T}}(u)\Pi_n^{-1}\Xi_n\Pi_n^{-1}B_r(u)$, where

$$\begin{aligned}
\Pi_n &= \textstyle\sum_{i=1}^n B_r(U_i)\mathbf{A}_i\mathbf{B}_i^{-1}\mathbf{A}_i^{\mathrm{T}}B_r(U_i)^{\mathrm{T}},\\
\Xi_n &= \textstyle\sum_{i=1}^n B_r(U_i)\mathbf{A}_i\mathbf{B}_i^{-1}\mathbf{Q}_i\mathbf{B}_i^{-1}\mathbf{A}_i^{\mathrm{T}}B_r(U_i)^{\mathrm{T}},
\end{aligned}$$

and where $\mathbf{Q}_i$ is a $K\times K$ matrix with the (k,k') entry

$$\mathbf{Q}_{i,k,k'}(\nu) = \{E(Y_{ik}Y_{ik'}\mid \mathbb{C}_i) - H_{ik}H_{ik'}\}(\beta_{k1}^0 + \beta_{k2}^0 G_i)(\beta_{k'1}^0 + \beta_{k'2}^0 G_i).$$

In the description above, $\mathbf{\Omega}_i$ is a generic working covariance matrix. Here is how we implemented it. Let $\mathbf{\Omega}_i$ be the conditional covariance matrix of $\mathbf{Y}_i = (Y_{i1},\ldots,Y_{iK})^{\mathrm{T}}$ given $\mathbb{C}_i$. Then the (k,k') entry of $\mathbf{\Omega}_i$ is $\mathbf{\Omega}_{i,k,k'} = E(Y_{ik}Y_{ik'}\mid\mathbb{C}_i) - H_{ik}H_{ik'}$. In practice, we estimate $\mathbf{\Omega}_i$ by $\widehat{\mathbf{\Omega}}_i = \widehat{\mathbf{V}}_i^{1/2}\widehat{\mathbf{R}}\widehat{\mathbf{V}}_i^{1/2}$, where $\widehat{\mathbf{V}}_i$ is a $K\times K$ diagonal matrix with the k^{th} diagonal as $\widehat{H}_{ik}(1-\widehat{H}_{ik})$ and $\widehat{\mathbf{R}}$=$n^{-1}\sum_{i=1}^n \widehat{\mathbf{V}}_i^{-1/2}\left(\mathbf{Y}_i - \widehat{\mathbf{H}}_i\right)\left(\mathbf{Y}_i - \widehat{\mathbf{H}}_i\right)^{\mathrm{T}}\widehat{\mathbf{V}}_i^{-1/2}$, where $\widehat{\mathbf{H}}_i = \left(\widehat{H}_{i1},\ldots,\widehat{H}_{iK}\right)^{\mathrm{T}}$ and $\widehat{H}_{ik} = \widehat{H}_{ik}(\widehat{\nu})$.

Similarly, the derivative of the (i,k) term in (7) with respect to ν is

$$(Y_{ik} - \widehat{H}_{ik}(\nu))\{\widehat{Q}_{ik}(\nu) + (\beta_{k1} + \beta_{k2}G_i)\{\widehat{\lambda}_w'(\nu)\}^{\mathrm{T}}B_r(X_{i0} + X_i^{\mathrm{T}}\alpha)\},$$

where $\widehat{H}_{ik}(\nu)$ is the same as $\widetilde{H}_{ik}$ except that λ in $\widetilde{H}_{ik}$ is replaced by $\widehat{\lambda}_w(\nu)$ in $\widehat{H}_{ik}(\nu)$, and $\widehat{Q}_{ik}(\nu)$ is the same as $Q_{ik}(\nu)$ except that $m(\cdot), m'(\cdot)$ in $Q_{ik}(\nu)$ are replaced by $\widehat{m}_w(\cdot,\nu), \widehat{m}_w'(\cdot,\nu)$ in $\widehat{Q}_{ik}$, and $\widehat{\lambda}_w'(\nu) = \partial\widehat{\lambda}_w(\nu)/\partial\nu^{\mathrm{T}}$ is the $P_n\times d_\nu$ derivative matrix of $\widehat{\lambda}_w(\nu)$ with respect to ν. We thus modify Step 2 to

Step 2d. Let $\mathbf{\Psi}_i(\nu)$ be the $d_\nu K\times 1$ vector formed by K vectors, each of length d_ν, with the $k^{th}, k=1,\ldots,K$, vector being $(Y_{ik} - \widehat{H}_{ik}(\nu))\{\widehat{Q}_{ik}(\nu) + (\beta_{k1}+\beta_{k2}G_i)\{\widehat{\lambda}_w'(\nu)\}^{\mathrm{T}}B_r(X_{i0}+X_i^{\mathrm{T}}\alpha)\}$. Obtain $\widehat{\nu}_w$ from solving $\sum_{i=1}^n \widehat{\mathbf{C}}_i(\nu)\widehat{\mathbf{D}}_i(\nu)^{-1}\mathbf{\Psi}_i(\nu) = 0$, where $\widehat{\mathbf{C}}_i(\nu)$ is a $d_\nu\times d_\nu K$ matrix, with k^{th} block

$$\widehat{\mathbf{C}}_{i,k}(\nu) = \widehat{V}_{ik}(\nu)\{\widehat{Q}_{ik}(\nu) + (\beta_{k1} + \beta_{k2}G_i)\widehat{\lambda}_w'(\nu)^{\mathrm{T}}B_r(X_{i0} + X_i^{\mathrm{T}}\alpha)\}^{\otimes 2},$$

where $\widehat{V}_{ik}(\nu) = \widehat{H}_{ik}(\nu)(1-\widehat{H}_{ik}(\nu))$ and $\widehat{\mathbf{D}}_i(\nu)$ is a $d_\nu K\times d_\nu K$ matrix, with (k,k') block

$$\begin{aligned}
\widehat{\mathbf{D}}_{i,k,k'}(\nu) &= \Omega_{i,k,k'}\{\widehat{Q}_{ik}(\nu) + (\beta_{k1} + \beta_{k2}G_i)\widehat{\lambda}_w'(\nu)^{\mathrm{T}}B_r(X_{i0} + X_i^{\mathrm{T}}\alpha)\},\\
&\quad\times\{\widehat{Q}_{ik'}(\nu) + (\beta_{k'1} + \beta_{k'2}G_i)\widehat{\lambda}_w'(\nu)^{\mathrm{T}}B_r(X_{i0} + X_i^{\mathrm{T}}\alpha)\}^{\mathrm{T}},
\end{aligned}$$

and $\widehat{\lambda}_w'(\nu)$ can be obtained via numerical differentiation. Let $\beta_1+\beta_2 G_i = \{(\beta_{k1}+\beta_{k2}G_i), 1 \le k \le K\}^{\mathrm{T}}$, and $\Theta_i(\nu)$ =diag($\{V_{ik}(\nu)(\beta_{k1}+\beta_k G_i)\}_{k=1}^{K}$. Then $\mathbf{A}_i(\nu) = (\beta_1+\beta_2 G_i)^{\mathrm{T}}\Theta_i(\nu)$. Denote $\mathbf{1}_{d_\nu}$ as the d_ν-dimensional vector with 1's as its elements. Let $\Theta_i = \Theta_i(\nu^0)$. Let $Q_i = (Q_{i1}, ..., Q_{iK})^{\mathrm{T}}$ and let η be a vector of functions $\eta(u) = \{\eta_1(u), \ldots, \eta_{d_\nu}(u)\}^{\mathrm{T}}$ with $\eta_\ell(\cdot) \in \mathrm{L}_2([a,b])$ that minimizes

$$\mathbf{1}_{d_\nu}^{\mathrm{T}} E\left[\left\{Q_i - (\beta_1^0+\beta_2^0 G_i)\eta^{\mathrm{T}}(U_i)\right\}^{\mathrm{T}} \Theta_i \mathbf{B}_i^{-1}\Theta_i \left\{Q_i - (\beta_1^0+\beta_2^0 G_i)\eta^{\mathrm{T}}(U_i)\right\}\right]\mathbf{1}_{d_\nu}.$$

Define $\mathbf{C}_i$ as a $d_\nu \times d_\nu K$ matrix, with k^{th} block $\mathbf{C}_{i,k} = V_{ik}\{Q_{ik} - (\beta_{k1}^0 + \beta_{k2}^0 G_i)\eta(U_i)\}^{\otimes 2}$. Define $\mathbf{D}_i$ as a $d_\nu K \times d_\nu K$ matrix, with (k,k') block

$$\mathbf{D}_{i,k,k'} = \Omega_{i,k,k'}\{Q_{ik} - (\beta_{k1}^0 + \beta_{k2}^0 G_i)\eta(U_i)\}\{Q_{ik'} - (\beta_{k'1}^0 + \beta_{k'2}^0 G_i)\eta(U_i)\}^{\mathrm{T}},$$

and define $\mathbf{D}_i^*$ as a $d_\nu K \times d_\nu K$ matrix, with (k,k') block

$$\begin{aligned}\mathbf{D}_{i,k,k'}^* &= \{E(Y_{ik}Y_{ik'} \mid \mathbb{C}_i) - H_{ik}H_{ik'}\}\{Q_{ik} - (\beta_{k1}^0 + \beta_{k2}^0 G_i)\eta(U_i)\} \\ &\quad \times\{Q_{ik'} - (\beta_{k'1}^0 + \beta_{k'2}^0 G_i)\eta(U_i)\}^{\mathrm{T}}.\end{aligned}$$

In the following two theorems, we establish the consistency and asymptotic normality of our procedure. Different from the independent disease case, without a correct specification of the correlation structure of the occurrences of different diseases, we can no longer achieve semiparametric efficiency.

Theorem 4 *Under the conditions in Appendix A.2, when ν is the collection of the true parameters or a $\sqrt{n}$-consistent estimator of ν, (a) $\widehat{\sigma}_w^{-1}(u,\nu^0)\{\widehat{m}_w(u,\nu) - m(u)\} \to Normal(0,1)$; (b) $|\widehat{m}_w(u,\nu) - m(u)| = O_p\left\{(nh)^{-1/2} + h^q\right\}$ uniformly in $u \in [a_0,b_0]$; and (c) $|\widehat{m}_w'(u,\nu) - m'(u)| = O_p(n^{-1/2}h^{-3/2} + h^{q-1})$ uniformly in $u \in [a_0,b_0]$.*

Let $\widehat{\mathbf{D}}_i^*$ be a $d_\nu K \times d_\nu K$ matrix, with (k,k') block

$$\widehat{\mathbf{D}}_{i,k,k'}^* = \widehat{\Omega}_{i,k,k'}\{Q_{ik}(\widehat{\nu}) - (\widehat{\beta}_{k1} + \widehat{\beta}_{k2}G_i)\widehat{\eta}(U_i(\widehat{\nu}))\}\{Q_{ik'}(\widehat{\nu}) - (\widehat{\beta}_{k'1} + \widehat{\beta}_{k'2}G_i)\widehat{\eta}(U_i(\widehat{\nu}))\}^{\mathrm{T}},$$

where $\widehat{\eta}(U_i(\widehat{\nu})) = \{\widehat{\eta}_1(U_i(\widehat{\nu})), \ldots, \widehat{\eta}_{d_\nu}(U_i(\widehat{\nu}))\}^{\mathrm{T}}$ and $\widehat{\eta}_\ell(U_i(\widehat{\nu})) = B_r^{\mathrm{T}}(U_{ik}(\widehat{\nu}))\widehat{\tau}_\ell$ with $\{\widehat{\tau}_\ell\}$ minimizing

$$\begin{aligned}&\mathbf{1}_{d_\nu}^{\mathrm{T}}\textstyle\sum_{i=1}^n \left[\left\{Q_i(\widehat{\nu}) - (\widehat{\beta}_1 + \widehat{\beta}_2 G_i)\widehat{\eta}^{\mathrm{T}}(U_i(\widehat{\nu}))\right\}^{\mathrm{T}} \Theta_i(\widehat{\nu})\mathbf{B}_i(\widehat{\nu})^{-1}\Theta_i(\widehat{\nu})\right. \\ &\left.\times \left\{Q_i(\widehat{\nu}) - (\widehat{\beta}_1 + \widehat{\beta}_2 G_i)\widehat{\eta}^{\mathrm{T}}(U_i(\widehat{\nu}))\right\}\right]\mathbf{1}_{d_\nu}.\end{aligned}$$

Theorem 5 *Let* $(\mathbf{C}, \mathbf{D}, \mathbf{D}^{*})$ *be generic notation for random variables with the same distribution as* $(\mathbf{C}_i, \mathbf{D}_i, \mathbf{D}_i^{*})$*. Under the conditions in Appendix A.2,* $\sqrt{n}(\widehat{\nu} - \nu^0) \rightarrow \mathit{Normal}(0_{d_\nu}, \mathbf{\Sigma})$ *for* $\widehat{\nu}$ *in a neighborhood of* ν^0*, where*

$$\mathbf{\Sigma} = \left\{E(\mathbf{C}\mathbf{D}^{-1}\mathbf{C}^{\mathrm{T}})\right\}^{-1} E(\mathbf{C}\mathbf{D}^{-1}\mathbf{D}^{*}\mathbf{D}^{-1}\mathbf{C}^{\mathrm{T}}) \left\{E(\mathbf{C}\mathbf{D}^{-1}\mathbf{C}^{\mathrm{T}})\right\}^{-1}.$$

Here, $\mathbf{\Sigma}$ *is consistently estimated by the sandwich estimator*

$$\widehat{\mathbf{\Sigma}} = (n^{-1}\textstyle\sum_{i=1}^{n}\widehat{\mathbf{C}}_i\widehat{\mathbf{D}}_i^{-1}\widehat{\mathbf{C}}_i^{\mathrm{T}})^{-1}(n^{-1}\sum_{i=1}^{n}\widehat{\mathbf{C}}_i\widehat{\mathbf{D}}_i^{-1}\widehat{\mathbf{D}}_i^{*}\widehat{\mathbf{D}}_i^{-1}\widehat{\mathbf{C}}_i^{\mathrm{T}})(n^{-1}\sum_{i=1}^{n}\widehat{\mathbf{C}}_i\widehat{\mathbf{D}}_i^{-1}\widehat{\mathbf{C}}_i^{\mathrm{T}})^{-1}. \qquad (9)$$

4.2 Multiple Populations and Multiple Diseases

Finally, we consider the general case that there are $k = 1, ..., K$ independent populations, and within the k^{th} population, there are $\ell = 1, ..., L_k$ diseases and $i = 1, ..., n_k$ observations. The outcomes are $Y_{ik\ell}$ and the covariates are $\mathbb{C}_{ik} = (G_{ik}, X_{ik0}, X_{ik}, Z_{ik}, G_{ik}W_{ik})$. The model is

$$\begin{aligned} \mathrm{pr}(Y_{ik\ell} = 1 \mid \mathbb{C}_{ik}) &= H_{ik\ell} \qquad (10) \\ &= H\{(\beta_{k\ell 1} + \beta_{k\ell 2}G_{ik})m(X_{ik0} + X_{ik}^{\mathrm{T}}\alpha) + Z_{ik}^{\mathrm{T}}(\theta_{k\ell 1} + \theta_{k\ell 2}G_{ik}) + G_{ik}W_{ik}^{\mathrm{T}}\theta_{k\ell 3}\}. \end{aligned}$$

We make the same assumptions as in Section A.2. As in Theorem 2, we write $n = n_1$ and for $k = 2, .., K$, define $n_k = nc_{kn}$, where there are constants $c_* > 0$ and $c_{**} < \infty$ such that $c_* \le c_k = \lim_{n\rightarrow\infty} c_{kn} \le c_{**}$.

Make the definitions of the terms in Section 4.1 appropriate to population $k = 1, .., K$, e.g., $\widetilde{\mathbf{A}}_{ik}(\nu)$, $\mathbf{B}_{ik}(\nu)$, $\mathbf{\Phi}_{ik}(\nu)$, $\widehat{\mathbf{C}}_{ik}(\nu)$, $\widehat{\mathbf{D}}_{ik}(\nu)$, $\mathbf{\Psi}_{ik}(\nu)$, $\mathbf{C}_{ik}$, $\mathbf{D}_{ik}$, Π_{nk}, Ξ_{nk}, etc. Obtain $\widehat{\lambda}_w(\nu)$ by solving $\sum_{k=1}^{K}\sum_{i=1}^{n_k} B_r\{U_{ik}(\alpha)\}\widetilde{\mathbf{A}}_{ik}(\nu)\mathbf{B}_{ik}(\nu)^{-1}\mathbf{\Phi}_{ik}(\nu) = 0$, and obtain $\widehat{\nu}_w$ by solving $\sum_{k=1}^{K}\sum_{i=1}^{n_k}\widehat{\mathbf{C}}_{ik}(\nu)\widehat{\mathbf{D}}_{ik}(\nu)^{-1}\mathbf{\Psi}_{ik}(\nu) = 0$. Define

$$\begin{aligned} \widehat{\sigma}_w^2(u, \nu) &= B_r^{\mathrm{T}}(u)\{\textstyle\sum_{k=1}^{K}(n_k/n)\Pi_{nk}\}^{-1}\{\sum_{k=1}^{K}(n_k/n)\Xi_{nk}\}\{\sum_{k=1}^{K}(n_k/n)\Pi_{nk}\}^{-1}B_r(u); \\ \mathbf{\Sigma} &= \left\{\textstyle\sum_{k=1}^{K}c_k E(\mathbf{C}_{ik}\mathbf{D}_{ik}^{-1}\mathbf{C}_{ik}^{\mathrm{T}})\right\}^{-1}\left\{\sum_{k=1}^{K}c_k E(\mathbf{C}_{ik}\mathbf{D}_{ik}^{-1}\mathbf{D}_{ik}^{*}\mathbf{D}_{ik}^{-1}\mathbf{C}_{ik}^{\mathrm{T}})\right\} \\ &\qquad \times\left\{\textstyle\sum_{k=1}^{K}c_k E(\mathbf{C}_{ik}\mathbf{D}_{ik}^{-1}\mathbf{C}_{ik}^{\mathrm{T}})\right\}^{-1}; \\ \widehat{\mathbf{\Sigma}} &= (n^{-1}\textstyle\sum_{k=1}^{K}\sum_{i=1}^{n_k}\widehat{\mathbf{C}}_{ik}\widehat{\mathbf{D}}_{ik}^{-1}\widehat{\mathbf{C}}_{ik}^{\mathrm{T}})^{-1}(n^{-1}\sum_{k=1}^{K}\sum_{i=1}^{n_k}\widehat{\mathbf{C}}_{ik}\widehat{\mathbf{D}}_{ik}^{-1}\widehat{\mathbf{D}}_{ik}^{*}\widehat{\mathbf{D}}_{ik}^{-1}\widehat{\mathbf{C}}_{ik}^{\mathrm{T}}) \\ &\qquad \times(n^{-1}\textstyle\sum_{k=1}^{K}\sum_{i=1}^{n_k}\widehat{\mathbf{C}}_{ik}\widehat{\mathbf{D}}_{ik}^{-1}\widehat{\mathbf{C}}_{ik}^{\mathrm{T}})^{-1}. \end{aligned}$$

Then Theorems 4-5 hold with these definitions, see Appendix A.8, and $\widehat{\Sigma}$ remains a sandwich estimator.

As in other problems involving correlated binary data and generalized estimating equations, the semiparametric efficiency established in Theorem 3 does not hold for the multiple populations and multiple diseases case, mainly due to the fact that the responses are correlated among different diseases and the correlation structure is unknown. Discussion of the working correlation matrix in parametric generalized estimating equation problems can be found in many papers, see for example Chaganty and Joe (2004).

Instead of embedding the problem in the generalized estimating equation framework, as we have done, there is some literature on developing a likelihood function that allows correlation among the binary responses while having the marginal probabilities be of logistic form, see for example Zhao and Prentice (1990) and Le Cessie and Van Houwelingen (1994). Our methods can be extended to this approach, but the ease of computation associated with a generalized estimating equation approach is a considerable advantage. This computational advantage is one of the reasons that generalized estimating equation methods are so widely employed in practice.

5 Data Analysis

5.1 Spline Setup

In all our implementations, we used cubic splines ($r = 4$) with equally spaced knots to approximate the nonparametric function $m(\cdot)$. We selected the number of interior knots N by minimizing a BIC criterion, where $\text{BIC}(N) = -2L_n(\widehat{\lambda}, \widehat{\nu}) + (N + p)\log(n)$. See Xue and Yang (2006) and Ma and Yang (2011) for the properties of the BIC criterion.

5.2 Dietary Score Example

We applied our methods to the NIH-AARP Study of Diet and Health (Schatzkin et al., 2001). The method used for assessing dietary component intakes is the National Cancer Institute's Dietary History Questionnaire (DHQ) (Subar et al., 2001). There were 294,673 men and 199,285 women in the data set. There were also dummy variables for various groups of age, body mass index, education, ethnicity, physical activity and smoking, making up the variables Z. In addition, for women, there were two dummy variables for hormone replacement therapy, making up the variables W. The HEI-2005 score for whole grains were taken as X_{ik0} and X_{ik}. The sum of the weights was normalized to equal $J+1=12$ for ease of comparison with the HEI-2005 total score, all of whose weights $=1$: the standard errors of these weights were obtained by the delta-method after fitting the data as described in Sections 3 and 4.

For women, the data set contains four diseases, breast cancer, ovarian cancer, colorectal cancer and lung cancer, while for men there are prostate cancer, colorectal cancer and lung cancer. See Table 2 for the numbers and percentages of cancer cases. The minimum HEI-2005 total score in the data set was $x_{\min}=19.67$, while the maximum was $x_{\max}=96.61$.

We used $F\left\{U\left(\widehat{\alpha}\right)\right\} = \Phi\left(\left[U\left(\widehat{\alpha}\right)-E\left\{U\left(\widehat{\alpha}\right)\right\}\right]/\sqrt{\mathrm{Var}\left\{U\left(\widehat{\alpha}\right)\right\}}\right)$ to construct B-spline functions, where $\Phi\left(\cdot\right)$ is the distribution function of the standard normal distribution and $U\left(\widehat{\alpha}\right) = X_0 + X^{\mathrm{T}}\widehat{\alpha}$. Thus the nonparametric function m is estimated by $\widehat{m}\{u\left(\widehat{\alpha}\right),\nu\} = \sum_{p=1}^{P_n} B_{r,p}\left[F\left\{u\left(\widehat{\alpha}\right)\right\}\right]\widehat{\lambda}_p(\nu)$.

We performed two analyses. In the first, we took a single disease and the two independent populations of men and women, using the method in Section 3, and applied to colorectal cancer and lung cancer separately. In the second, we analyzed all the cancer outcomes, using the method in Section 4.2. The point of doing the former is to illustrate that analyzing single diseases at a time can lead to very different results than those from analyzing multiple diseases simultaneously, a point we made in Section 1.

5.3 Independent Populations, Single Disease

Our first analysis uses the setup in Section 3, where there are $K=2$ independent populations, men and women, and $L=1$ disease. We performed analyses separately for colorectal cancer and lung cancer, and display here the results for both. Because hormone replacement therapy occurs only for women, the right had side of model (4) is identifiable when the parameter subscripts do not involve k, e.g., $\beta_1+\beta_2 G_{ik}$.

Table 3 shows the estimates of the weights of the component scores, their standard errors and their p-values for testing whether the weights = 1, i.e., whether the weight equals the HEI-2005 weight.

The main conclusion of Table 3 is that the weights for the HEI-2005 component scores are strikingly different for total fruit, whole grains and dairy, depending on whether one is interested in colorectal cancer or lung cancer. This is a point we made in the discussion after equation (2) about having a single score and not one for every disease. Table 3 suggests that if one is worried about colorectal cancer, one should increase consumption of whole grains and dairy products, but if one is worried about lung cancer, such consumption would have only a minor effect, but total fruit intake should be increased.

One can see in Table 3 that the weights of many of the individual component scores differ from HEI-2005's weight of 1.0. We also tested whether the HEI-2005 weights fit the data as well, by testing $H_0 : \alpha_1 = \alpha_2 = \cdots = \alpha_J = 1$. To this end, we constructed the Wald chi-square statistic $\chi^2_W = (\widehat{\alpha} - \mathbf{1}_J)^{\mathrm{T}} \{\widehat{\mathrm{V}}\,(\widehat{\alpha})\}^{-1} (\widehat{\alpha} - \mathbf{1}_J)$, which has an asymptotic chi-square distribution with J degrees of freedom under H_0. Here, $\widehat{\mathrm{V}}\,(\widehat{\alpha})$ is the estimated asymptotic variance-covariance matrix of $\widehat{\alpha}$, and is calculated following Theorem 5. The p-value for this hypothesis is < 0.0001.

5.4 Multiple Populations and Multiple Diseases Analysis

Our second analysis uses the setup in Section 4.2, with all the cancers available in our data set: lung, colorectal, breast and ovarian cancers for women, and lung, colorectal and prostate cancers for men. We found that the working correlations among men and women were all < 0.03 in absolute value, so we report results for the working independence estimate.

Table 4 shows the estimated weights of the component scores and their standard errors. Because we are using multiple diseases and populations, and not just colorectal or lung cancer separately, but all the cancers simultaneously, we can expect differences between Table 4 and either analysis in Table 3. One of the striking difference is the vast down-weighting of increased Milk consumption compared to the results for colorectal cancer only. In the HEI-2005 score, increase of Milk consumption results in a monotonically increasing score for

Milk. In the colorectal cancer case, a person who gets the top score of 10 on Milk contributes 24.4 to the single index. After accounting for the other diseases, however, the contribution is 0.61, a vast decrease. To us, this makes perfect sense, because the value of increased Milk consumption in adults is hardly universally accepted. For example, The Alternative Healthy Eating Index (McCullough et al., 2002) does not even include dairy as part of its index, i.e., increased Milk consumption gets zero weight. The Modified Mediterranean Diet Score (Trichopoulou et al., 2005) and the MedDietScore (Panagiotakos et al., 2006), have been shown to be related to overall survival and coronary heart disease respectively, but for these scores, increases in Milk consumption lead to *decreases* in the score for Milk, i.e., negative weight.

Table 5 shows the estimates, standard errors (se) and p-values for the coefficients β. The β coefficient for men associated with lung cancer was $= -1$ for identifiability. However, when we instead set the coefficient for lung cancer for women to be $= -1$, the estimated coefficient for lung cancer for men was -1.06, and the p-value was very small. It is clear from the table that the real practical impact of diet here is its contribution to decreases in risk for lung and colorectal cancers, and for both men and women, and that the impact is greater for lung cancers. See below for a discussion of the relative risks, displayed in Figure 2, which supports our conclusion. The estimated values for all other groups are also negative except for the two groups: (a) men and prostate cancer; and (b) women and ovarian cancer, where the coefficients are very small: they have both been set $= 0$ under the constraint that a better diet is not a risk factor for either cancer. In the figures that we discuss, the index (x-axis) plotted is from the 3^{rd} to the 97^{th} percentiles of the actual index.

Figure 1 shows the plot of the estimates of $m(\cdot)$ against the index $u\,(\widehat{\alpha})$ along with pointwise 95% confidence intervals, without any additional monotonicity constraints on $m(\cdot)$. The estimated function itself is monotone as expected. Observe that the estimated function is not an exact linear function, especially when considering the pointwise confidence intervals. Indeed, from the index value of 50 to 72, the estimated function has an increasing acceleration, then it becomes flat, and it increases quickly again when the index value is greater than 82. When we refit the data with a linear link, the results, while different, are in good agreement, both in the estimated functions, the tables, and the analyses that are described next.

Figure 2 displays the estimated relative risks of the various cancers and separately for men and women. Clearly, we see that the index predicts stronger decreases for lung cancer relative risks for better diet index score as compared to the other diseases. In Figure 2, the effect of better diet on prostate cancer in men and ovarian cancer in women are is nearly null, and the effect of better diet is quite modest on breast cancer. When analyzing the HEI-2005 total score, the p-values for prostate cancer, breast cancer and ovarian cancer were 0.15, 0.09, and 0.44, respectively, roughly what is seen in Figure 2.

Figure S.1 of the **Supplementary Material** shows how the estimated relative risks differ between men and women for lung and colorectal cancer. In both cases, women have the lower risk in general, with the largest difference being in colorectal cancer, but even there, the differences are not great. This agrees with the marginal rate of lung cancer for men and women being 2.08% and 1.82%, respectively, while the marginal rate of colorectal cancer for men and women 1.59% and 1.15%, respectively, see Table 2.

The hypothesis for testing that the weights all equal 1.0 is rejected with a p-value numerically very close to 0, as expected.

6 Simulation

In this section, we describe a simulation study to assess the finite-sample performance of our method in the case of two populations and multiple diseases. Section S.1 of the **Supplementary Material** has results for two independent populations and one disease. Here simulated data from the logistic model with multiple populations and diseases, so that

$$\mathrm{pr}(Y_{ik\ell} = 1 \mid \mathbb{C}_{ik}) = H\{\beta_{k\ell} m(X_{ik}^{\mathrm{T}}\alpha) + Z_{ik}^{\mathrm{T}}\theta_{k\ell 1} + G_{ik} W_{ik}^{\mathrm{T}}\theta_{k\ell 3}\},$$

for $i = 1, \ldots, n$ and $k = 1, 2$. We consider two independent populations, and within the k^{th} population, there are $\ell = 1, ..., L_k$ diseases and $i = 1, ..., n$ observations. We let $L_1 = 3$ and $L_2 = 4$, so that, as in the example of Section 5, the first and second populations have four and three diseases, respectively. There were $1,000$ simulated data sets.

For each simulated data set, we let $n = 3,000$, and set the covariates to be randomly selected, without replacement, from the real data in Section 5. We set each component of $\alpha =$ 1, and made the convention that the estimates should sum to 12. The true values of $\beta_{k\ell}$ are listed in Table 7. We simulated the components in $\theta_{k\ell 1}$ and $\theta_{k\ell 3}$ from the Uniform$[-0.5, 0.5]$ distribution, except that, for identifiability, the first component in θ_{k1} is taken to be zero. The true function is $m(u) = \exp(u/3)$.

To generate the correlated binary data, we use the following algorithm. Suppose we want to generate M binary variables with probabilities $(\pi_1, ..., \pi_M)$. Let $(U_1, ..., U_M)$ be equicorrelated standard normal random variables with correlation ρ, and for $m = 1, ..., M$, define $T_m = log\{\pi_m/(1-\pi_m)\} + \log[\Phi(U_m)/\{1-\Phi(U_m)\}]$. Then setting $Y_m = I(T_m > 0)$ creates correlated binary random variables with the desired probabilities. In our setting, the correlations of the binary variables were approximately $\rho/2$, so for the simulation in this section we set $\rho = 0.10$. This resulted in correlations somewhat higher than in the real data in Section 5.

We also conducted simulation experiments for independent binary outcomes and the sample size but with correlation nearly 0.10, ($\rho = 0.20$), and both correlations with $n = 2000$. These results similar to the results in this section, are in Section S.2 of the **Supplementary Material**.

To give some idea of how this simulation compares with the real data, Table 7 also lists the mean number of cases by disease and by population: the mean across both is $7,975$. These are many fewer than the number of cases seen in the actual data, see Table 2. Hence, since the effective sample size might be thought of as the number of cases, the simulation approximates a smaller study than the NIH-AARP data analyzed in Section 5.

Table 6 gives the results for the estimates of α, while Table 7 gives the results for the estimates of the $\beta_{k\ell}$. In both cases, the estimates are very nearly unbiased, the estimated standard errors very nearly equal the actual standard errors, and the coverage probabilities are close the nominal 95%.

Figure S.2 of the **Supplementary Material** shows that the mean estimated function across the simulated data sets is also very nearly unbiased. Overall, the simulation suggests that our methodology leads to nearly unbiased estimates and inferences that achieve their nominal levels.

7 Discussion

Based on motivation from the current practice in nutritional epidemiology, we have developed generalized partially linear logistic single index models in the case that there are several populations and/or diseases. The novelty of the modeling is that the single-index function itself is the same across populations and diseases. In the case that the populations/diseases are independent given covariates, we developed a computable B-spline based semiparametric efficient methodology. In the case that the populations/diseases are correlated given the covariates, our method makes no assumptions about how the diseases are related.

The importance of developing a score that is constructed for health risk prediction across multiple diseases and populations, versus a different score for each population and disease, were illustrated in our work. When we analyzed men for colorectal cancer solely, increasing Milk consumption was given a very high weight in the single index. However, when we fit the model simultaneously for multiple diseases, the weight for increasing Milk consumption became much smaller. The importance of Milk consumption is a source of controversy in the nutrition literature, and our results agree with the Alternative Healthy Eating Index (AHEI), which assigns zero weight to increased Milk consumption, and the Mediterranean diets scores assigns negative weight to increased Milk consumption.

Our results are focused on logistic regression, so that they are readily transparent, but they are easily adapted to apply to any generalized linear model, by replacing H and its related quantities with a more general link function and its corresponding related quantities in the derivation.

We have strived to use a single overall score, which we have argued (a) avoids different diet for different disease; and (b) is important for interpretability. It is important to use as many diseases and populations as possible, and to draw inferences and projections only to those populations and diseases. One can think of what we have done as to come up with a framework for a type of "average" version of the individual model fits across multiple diseases and populations. We do not average directly, instead, we average across the estimating functions.

Supplementary Material

The **Supplementary Material** contains results of addition simulations, and R and Matlab programs to run the analysis. The NIH-AARP data used in the data analysis are available from the NIH via a data transfer agreement (www.http://dietandhealth.cancer.gov/) but we are not allowed to distribute it. The program files include simulated data as described in Section 6.

Acknowledgments

S. Ma's research was supported by NSF grant DMS-1306972. Y. Ma's research was supported by NSF grant DMS-1206693 and NIH grant R01-NS073671. Carroll's research was supported by a grant from the National Cancer Institute (U01-CA057030).

References

Ai, C. and Chen, X. (2003). Efficient estimation of models with conditional moment restrictions containing unknown functions. *Econometrica*, 71, 1795–1843.

Akbaraly, T. N., Ferrie, J. E., Berr, C., Brunner, E. J., Head, J., Marmot, M. G., Singh-Manoux, A., Ritchie, K., Shipley, M. J., and Kivimaki, M. (2011). Alternative Healthy Eating Index and mortality over 18 y of follow-up: results from the Whitehall II cohort. *American Journal of Clinical Nutrition*, 194, 247–253.

Bellman, R. E. (1961). *Adaptive Control Processes.* Princeton University Press, Princeton.

Bosq, D. (1961). *Nonparametric Statistics for Stochastic Processes.* Springer, New York.

Carroll, R. J., Fan, J., Gijbels, I., and Wand, M. P. (1997). Generalized partially linear single-index models. *Journal of the American Statistical Association*, 92, 477–489.

Chaganty, N. R. and Joe, H. (2004). Efficiency of generalized estimating equations for binary responses. *Journal of the Royal Statistical Society: Series B*, 66, 851–860.

Chiuve, S. E., Fung, T. T., Rimmand, E. B., Hu, F. B., McCullough, M. L., Wang, M., Stampfer, M. J., and Willett, W. C. (2012). Alternative dietary indices both strongly predict risk of chronic disease. *Journal of Nutrition*, 142, 1009–1018.

Cui, X., Härdle, W. K., Zhu, L., et al. (2011). The efm approach for single-index models. *Annals of Statistics*, 39, 1658–1688.

de Boor, C. (2001). *A Practical Guide to Splines.* Springer, New York.

DeVore, R. A. and Lorentz, G. G. (1993). *Constructive Approximation.* Springer, Berlin.

Fan, J. and Lv, J. (2011). Nonconcave penalized likelihood with NP-dimensionality. *IEEE Transactions on Information Theory*, 57, 5467–5484.

George, S. M., Neuhouser, M. L., Mayne, S. T., Irwin, M. L., Albanes, D., Gail, M. H., Alfano, C. M., Bernstein, L., McTiernan, A., Reedy, J., Smith, A. W., Ulrich, C. M., and Ballard-Barbash, R. (2010). Postdiagnosis diet quality is inversely related to a biomarker of inflammation among breast cancer survivors. *Cancer Epidemiology, Biomarkers & Prevention*, 19, 2220–2228.

Guenther, P. M., Reedy, J., and Krebs-Smith, S. M. (2008). Development of the Healthy Eating Index-2005. *Journal of the American Dietetic Association*, 108, 1896–1901.

Guenther, P. M., Reedy, J., Krebs-Smith, S. M., and Reeve, B. B. (2008). Evaluation of the Healthy Eating Index-2005. *Journal of the American Dietetic Association*, 108, 1854–1864.

Huang, J. Z. (2003). Local asymptotics for polynomial spline regression. *Annals of Statistics*, 31, 1600–1635.

Huang, J. Z., Wu, C. O., and Zhou, L. (2002). Varying-coefficient models and basis function approximations for the analysis of repeated measurements. *Biometrika*, 89, 111–128.

Huang, J. Z. and Yang, L. (2004). Identification of non-linear additive autoregressive models. *Journal of the Royal Statistical Society: Series B*, 66, 463–477.

Le Cessie, S. and Van Houwelingen, J. (1994). Logistic regression for correlated binary data. *Applied Statistics*, 43, 95–108.

Liu, X., Wang, L., and Liang, H. (2011). Estimation and variable selection for semiparametric additive partial linear models. *Statistica Sinica*, 21, 1225.

Ma, S. and Yang, L. (2011). A jump-detecting procedure based on spline estimation. *Journal of Nonparametric Statistics*, 23, 67–81.

Ma, Y. and Zhu, L. (2013). Doubly robust and efficient estimators for heteroscedastic partially linear single-index models allowing high dimensional covariates. *Journal of the Royal Statistical Society: Series B*, 75, 305–322.

McCullough, M. L., Feskanich, D., Stampfer, M. J., Giovannucci, E. L., Rimm, E. B., Hu, F. B., Spiegelman, D., Hunter, D. J., Colditz, G. A., and Willett, W. C. (2002). Diet quality and major chronic disease risk in men and women: moving toward improved dietary guidance. *American Journal of Clinical Nutrition*, 76, 1261–1271.

Panagiotakos, D. B., Pitsavos, C., and Stefanadis, C. (2006). Dietary patterns: a mediterranean diet score and its relation to clinical and biological markers of cardiovascular disease risk. *Nutrition, Metabolism and Cardiovascular Diseases*, 16, 559–568.

Reedy, J. R., Mitrou, P. N., Krebs-Smith, S. M., Wirfält, E., Flood, A. V., Kipnis, V., Leitzmann, M., Mouwand, T., Hollenbeck, A., Schatzkin, A., and Subar, A. F. (2008). Index-based dietary patterns and risk of colorectal cancer: the nih-aarp diet and health study. *American Journal of Epidemiology*, 168, 38–48.

Schatzkin, A., Subar, A. F., Thompson, F. E., Harlan, L. C., Tangrea, J., Hollenbeck, A. R., Hurwitz, P. E., Coyle, L., Schussler, N., Michaud, D. S., Freedman, L. S., Brown, C. C., Midthune, D., and Kipnis, V. (2001). Design and serendipity in establishing a large cohort with wide dietary intake distributions: the national institutes of health-aarp diet and health study. *American Journal of Epidemiology*, 154, 1119–1125.

Stone, C. J. (1985). Additive regression and other nonparametric models. *Annals of Statistics*, pages 689–705.

Subar, A. F., Thompson, F. E., Kipnis, V., Mithune, D., Hurwitz, P., McNutt, S., McIntosh, A., and Rosenfeld, S. (2001). Comparative validation of the block, willett, and national cancer institute food frequency questionnaires: The Eating at America's Table Study. *American Journal of Epidemiology*, 154, 1089–1099.

Trichopoulou, A., Orfanos, P., Norat, T., Bueno-de Mesquita, B., Ocké, M. C., Peeters, P. H., van der Schouw, Y. T., Boeing, H., Hoffmann, K., Boffetta, P., et al. (2005). Modified mediterranean diet and survival: Epic-elderly prospective cohort study. *British Medical Journal*, 330, 991.

Wang, J. and Yang, L. (2009a). Polynomial spline confidence bands for regression curves. *Statistica Sinica*, 19, 325.

Wang, L., Liu, X., Liang, H., and Carroll, R. J. (2011). Estimation and variable selection for generalized additive partial linear models. *Annals of Statistics*, 39, 1827.

Wang, L. and Yang, L. (2009b). Spline estimation of single-index models. *Statistica Sinica*, 19, 765.

Xue, L. and Yang, L. (2006). Additive coefficient modeling via polynomial spline. *Statistica Sinica*, 16, 1423.

Yu, Y. and Ruppert, D. (2002). Penalized spline estimation for partially linear single-index models. *Journal of the American Statistical Association*, 97, 1042–1054.

Zhao, L. P. and Prentice, R. L. (1990). Correlated binary regression using a quadratic exponential model. *Biometrika*, 77, 642–648.

Zhou, S., Shen, X., and Wolfe, D. A. (1998). Local asymptotics for regression splines and confidence regions. *The Annals of Statistics*, pages 1760–1782.

Zhou, S. and Wolfe, D. A. (2000). On derivative estimation in spline regression. *Statistica Sinica*, pages 93–108.

Appendix

A.1 Some Simplifications and Definitions

For simplicity of notation, we work through the asymptotics in the case that there is one population with a common sample size n, and thus the covariates are the same across the populations/diseases. The statements of Theorems 1-3 are readily verified when the responses are independent with different sample sizes and different covariates across $k = 1, ..., K$.

For any vector $\zeta = (\zeta_1, \ldots, \zeta_s)^{\mathrm{T}} \in R^s$, denote the norm $\|\zeta\|_r = (|\zeta_1|^r + \cdots + |\zeta_s|^r)^{1/r}$, $1 \le r \le \infty$. For positive numbers a_n and b_n, $n > 1$, let $a_n \asymp b_n$ denote that $\lim_{n\to\infty} a_n/b_n = c$, where c is some nonzero constant. Denote the space of the q^{th} order smooth functions as $C^{(q)}([a,b]) = \left\{\phi \,\middle|\, \phi^{(q)} \in C\,[a,b]\right\}$. For any $s \times s$ symmetric matrix $\mathbf{A}$, denote its L$_r$ norm as $\|\mathbf{A}\|_q = \max_{\varsigma \in R^s, \varsigma \neq 0} \|\mathbf{A}\varsigma\|_q \|\varsigma\|_q^{-1}$. Let $\|\mathbf{A}\|_\infty = \max_{1\le i\le s} \sum_{j=1}^{s} |a_{ij}|$. For a vector a, let $\|a\|_\infty = \max_{1\le i\le s} |a_i|$.

A.2 Regularity Conditions

(C1) The density function $f_{X_0+X^T\alpha}(x_0 + x^T\alpha)$ of random variable $X_0 + X^T\alpha$ is bounded away from 0 on S_α and satisfies the Lipschitz condition of order 1 on S_α, where $S_\alpha = \left\{X_0 + X^T\alpha, (X_0, X^{\mathrm{T}})^{\mathrm{T}} \in S\right\}$ and S is a compact support set of $(X_0, X^{\mathrm{T}})^{\mathrm{T}}$, for α in a neighborhood of its true values α^0.

(C2) $m(\cdot) \in C^{(q)}([a_0, b_0])$ for $q \ge 2$, $\eta_\ell^0 \in C^{(1)}([a_0, b_0])$, $1 \le l \le d_\nu$, and the spline order satisfies $r \ge q$.

(C3) There exists $0 < c < \infty$, such that the distances between neighboring knots satisfies

$$\max_{r\le p\le N+r} |h_{p+1} - h_p| = o(N^{-1}) \text{ and } h/\min_{r\le p\le N+r} h_J \le c.$$

Furthermore, the number of knots satisfies $N \to \infty$, as $n \to \infty$, $N^{-4}n \to \infty$ and $Nn^{-1/(2q+2)} \to \infty$.

(C4) $\sup_{i,k} |G_{ik}| \le M < \infty$. The eigenvalues of $\sum_{k=1}^{K} c_k E[V_{ik}\{Q_{ik} - (\beta_{k1}^0 + \beta_{k2}^0 G_{ik})\eta^0(U_{ik})\}^{\otimes 2}]$ are bounded below from zero. The eigenvalues of $E(\mathbf{C}\mathbf{D}^{-1}\mathbf{C}^{\mathrm{T}})$ given in Theorem 5 are bounded below from zero.

Conditions (C1)-(C3) are commonly used in the nonparametric smoothing literature; see, for example, Zhou et al. (1998) and Cui et al. (2011). Condition (C4) is needed for asymptotic normality of the parametric estimator.

A.3 Proof of Theorem 1

We first introduce two lemmas which will be used in the following proofs.

Lemma 1 *For any $a = (a_p : 1 \le p \le P_n)$, there exist constants $0 < c_B \le C_B < \infty$, such that for n large enough,*

$$c_B a^{\mathrm{T}} a h \le a^{\mathrm{T}} E\left\{B_r(U_{ik}(\alpha^0)) B_r^{\mathrm{T}}(U_{ik}(\alpha^0))\right\} a \le C_B a^{\mathrm{T}} a h. \tag{A.1}$$

$$\begin{aligned} &\max_{1 \le p,p' \le P_n} \left| n^{-1} \textstyle\sum_{k=1}^{K} \sum_{i=1}^{n_k} [B_{r,p}(U_{ik}(\alpha^0)) B_{r,p'}(U_{ik}(\alpha^0)) - E\left\{B_{r,p}(U_{ik}(\alpha^0)) B_{r,p'}(U_{ik}(\alpha^0))\right\}] \right| \\ = \;& O_p\{\sqrt{hn^{-1} log(n)}\}. \end{aligned} \tag{A.2}$$

Proof of Lemma 1: Result (A.1) follows from Theorem 5.4.2 of DeVore and Lorentz (1993), and (A.2) can be proved by Bernstein's inequality in Bosq (1961).

Define

$$\mathbf{V}_n^0(\nu) = n^{-1} \textstyle\sum_{k=1}^{K} \sum_{i=1}^{n_k} E\left\{V_{ik}(\beta_{k1} + \beta_{k2} G_i)^2 B_r(U_{ik}(\alpha)) B_r^{\mathrm{T}}(U_{ik}(\alpha))\right\}. \tag{A.3}$$

Lemma 2 *There are constants $0 < c_v < C_v < \infty$, and $0 < C_S < \infty$, such that for n large enough,*

$$\|\mathbf{V}_n^0(\nu^0)^{-1}\|_\infty \le C_S h^{-1}. \tag{A.4}$$

and

$$c_v h \le \|\mathbf{V}_n^0(\nu^0)\|_2 \le C_v h, C_v^{-1} h^{-1} \le \|\mathbf{V}_n^0(\nu^0)^{-1}\|_2 \le c_v^{-1} h^{-1}, \tag{A.5}$$

Proof of Lemma 2: Result (A.5) follows from (A.1). The result that $\|\mathbf{V}_n^0(\nu^0)^{-1}\|_\infty \le C_S h^{-1}$ follows from (A.5) and Theorem 13.4.3 in DeVore and Lorentz (1993).

If $m \in C^q[a_0, b_0]$, there exists $\lambda^0 \in R^{P_n}$, such that

$$\sup_{u \in [a_0, b_0]} |m(u) - \widetilde{m}(u)| = O(h^q), \tag{A.6}$$

where $\widetilde{m}(u) = B_r^{\mathrm{T}}(u)\lambda^0$ (de Boor, 2001). In the following, we prove the results for the nonparametric estimator $\widehat{m}(u, \nu)$ in Theorem 1 when $\nu = \nu^0$. Then the results also hold when ν is a $\sqrt{n}$ consistent estimator of ν^0, since the nonparametric convergence rate in Theorem 1 is slower than $n^{-1/2}$. Let $\alpha_n = n^{-1/2} P_n + P_n^{-q+1/2}$. We will show that for any given $\epsilon > 0$, for n sufficiently large, there exists a large constant $C > 0$ such that

$$\mathrm{pr}\{\sup_{\|\tau\|_2 = C} L_n(\lambda^0 + \alpha_n \tau, \nu^0) < L_n(\lambda^0, \nu^0)\} \ge 1 - \epsilon. \tag{A.7}$$

This implies that for n sufficiently large, with probability at least $1 - \epsilon$, there exists a local maximum for (5) in the ball $\{\lambda^0 + \alpha_n \tau : \|\tau\|_2 \le C\}$. Hence, there exists a local maximizer such that $\|\widehat{\lambda}(\nu^0) - \lambda^0\|_2 = O_p(\alpha_n)$. Since $L_n(\lambda, \nu^0)$ is a concave function of λ, the local maximizer is the global maximizer of (5).

Define

$$\widetilde{H}_{ik}(\lambda, \nu) = H[(\beta_{k1} + \beta_{k2} G_{ik}) B_r^{\mathrm{T}}\{U_{ik}(\alpha)\}\lambda + Z_{ik}^{\mathrm{T}}(\theta_{k1} + \theta_{k2} G_{ik}) + G_{ik} W_{ik}^{\mathrm{T}} \theta_{k3}],$$

then

$$\begin{aligned} & \partial L_n(\lambda, \nu)/\partial\lambda \\ = \;& \textstyle\sum_{k=1}^{K}\sum_{i=1}^{n_k}\{Y_{ik} - \widetilde{H}_{ik}(\lambda, \nu)\}(\beta_{k1} + \beta_{k2} G_{ik}) B_r\{U_{ik}(\alpha)\}, \\ & \partial^2 L_n(\lambda, \nu)/\partial\lambda\partial\lambda^{\mathrm{T}} \\ = \;& -\textstyle\sum_{k=1}^{K}\sum_{i=1}^{n_k}\widetilde{H}_{ik}(\lambda, \nu)\{1 - \widetilde{H}_{ik}(\lambda, \nu)\}(\beta_{k1} + \beta_{k2} G_i)^2 B_r\{U_{ik}(\alpha)\} B_r^{\mathrm{T}}\{U_{ik}(\alpha)\}. \end{aligned}$$

By Taylor's expansion, we have

$$\begin{aligned} & L_n(\lambda^0 + \alpha_n\tau, \nu^0) - L_n(\lambda^0, \nu^0) \\ = \;& \{\partial L_n(\lambda^0, \nu^0)/\partial\lambda\}^{\mathrm{T}}\alpha_n\tau - [-2^{-1}(\alpha_n\tau)^{\mathrm{T}}\{\partial^2 L_n(\lambda^*, \nu^0)/\partial\lambda\partial\lambda^{\mathrm{T}}\}\alpha_n\tau], \end{aligned} \tag{A.8}$$

where $\lambda^* = \varrho\lambda + (1-\varrho)\lambda^0$ for some $\varrho \in (0, 1)$. Moreover,

$$|\{\partial L_n(\lambda^0, \nu^0)/\partial\lambda\}^{\mathrm{T}}\alpha_n\tau| \leq \alpha_n ||\partial L_n(\lambda^0, \nu^0)/\partial\lambda||_2 ||\tau||_2 = C\alpha_n ||\partial L_n(\lambda^0, \nu^0)/\partial\lambda||_2,$$

and $\partial L_n(\lambda^0, \nu^0)/\partial\lambda = \Delta_{n1} + \Delta_{n2}$, where

$$\begin{aligned} \Delta_{n1} &= \textstyle\sum_{k=1}^{K}\sum_{i=1}^{n_k}(Y_{ik} - H_{ik}(\nu^0))(\beta_{k1}^0 + \beta_{k2}^0 G_{ik}) B_r(U_{ik}(\alpha^0)), \\ \Delta_{n2} &= \textstyle\sum_{k=1}^{K}\sum_{i=1}^{n_k}(H_{ik}(\nu^0) - \widetilde{H}_{ik}(\lambda^0, \nu^0))(\beta_{k1}^0 + \beta_{k2}^0 G_{ik}) B_r(U_{ik}(\alpha^0)). \end{aligned}$$

Since $E(\Delta_{n1}) = \mathbf{0}$, and $E[\{Y_{ik} - H_{ik}(\nu^0)\}(\beta_{k1}^0 + \beta_{k2}^0 G_{ik}) B_{r,p}\{U_{ik}(\alpha^0)\}]^2 \leq C_1 h$ for some constant $0 < C_1 < \infty$, then $E(||n^{-1}\Delta_{n1}||_2^2) \leq P_n K n^{-1} C_1 h$. By Condition (C3), we have $h \leq c P_n^{-1}$. Then $E(||n^{-1}\Delta_{n1}||_2^2) \leq P_n K n^{-1} C_1 c P_n^{-1} = K C_1 c n^{-1}$. Then for any $\epsilon > 0$, by Chebyshev's inequality, we have $\mathrm{pr}(||n^{-1}\Delta_{n1}||_2 \geq \sqrt{n^{-1} K C_1 c \epsilon^{-1}}) \leq \epsilon$. Hence, there exists an event A_{n1} with $\mathrm{pr}(A_{n1}^C) \leq \epsilon$, such that on A_{n1} we have $||\Delta_{n1}||_2 < \sqrt{K C_1 c \epsilon^{-1}} n^{1/2}$. Moreover, by (A.6), we have $\sup_{i,k}|H_{ik}(\nu^0) - \widetilde{H}_{ik}(\lambda^0, \nu^0)| = O(h^q)$. Denote

$$\Delta_{ikp} = (H_{ik}(\nu^0) - \widetilde{H}_{ik}(\lambda^0, \nu^0))(\beta_{k1}^0 + \beta_{k2}^0 G_{ik}) B_{r,p}(U_{ik}(\alpha^0)).$$

Then, there exist constants $0 < C_2, C_2^{'} < \infty$ such that

$$\begin{aligned} & E(||\Delta_{n2}||_2) \\ \leq \;& P_n^{1/2}\{\sup_{1\leq p\leq P_n} E(\textstyle\sum_{k=1}^{K}\sum_{i=1}^{n_k}\Delta_{ikp})^2\}^{1/2} \\ \leq \;& P_n^{1/2}[\{\sup_{1\leq p\leq P_n}\textstyle\sum_{k=1}^{K}\sum_{i=1}^{n_k} E(\Delta_{ikp}^2)\}^{1/2} + \{\sup_{1\leq p\leq P_n}\displaystyle\sum_{(k,i)\neq(k',i')} E(\Delta_{ikp}\Delta_{i'k'p})\}^{1/2}] \\ \leq \;& P_n^{1/2}\{(C_2 n h^{2q} h)^{1/2} + (C_2^{'} n^2 h^{2q} h^2)^{1/2}\} \\ \leq \;& P_n^{1/2}(\sqrt{C_2} + \sqrt{C_2^{'}}) n h^{q+1} \leq P_n^{1/2} C_3 n c^{q+1} P_n^{-(q+1)} = C_3 c^{q+1} n P_n^{-q-1/2}, \end{aligned}$$

where $C_3 = (\sqrt{C_2}+\sqrt{C_2'})$, for n sufficiently large given that $nh \to \infty$. Again by Chebyshev's inequality, for any $\epsilon > 0$, we have $\text{pr}(||\Delta_{n2}||_2 \geq \epsilon^{-1/2} C_3 c^{q+1} n P_n^{-q-1/2}) \leq \epsilon$. Hence, there exists an event A_{n2} with $\text{pr}(A_{n2}^C) \leq \epsilon$, such that on A_{n2} we have $||\Delta_{n2}||_2 < \epsilon^{-1/2} C_3 c^{q+1} n P_n^{-q-1/2}$. Therefore, by the above results, we have for n sufficiently large, on the event $A_{n1} \cap A_{n2}$ with $\text{pr}(A_{n1} \cap A_{n2}) \geq 1 - 2\epsilon$, such that

$$\begin{aligned} |\{\partial L_n(\lambda^0, \nu^0)/\partial\lambda\}^{\mathrm{T}} \alpha_n \tau| &\leq C\alpha_n ||\partial L_n(\lambda^0, \nu^0)/\partial\lambda||_2 \leq C\alpha_n(||\Delta_{n1}||_2 + ||\Delta_{n2}||_2) \\ &\leq C\alpha_n(\sqrt{KC_1 c\epsilon^{-1}} n^{1/2} + \epsilon^{-1/2} C_3 c^{q+1} n P_n^{-q-1/2}). \qquad \text{(A.9)} \end{aligned}$$

Moreover, by (A.1) and (A.2), we have for n sufficiently large, with probability approaching 1,

$$-2^{-1}\tau^{\mathrm{T}}\{\partial^2 L_n(\lambda^*, \nu^0)/\partial\lambda\partial\lambda^{\mathrm{T}}\}\tau \geq nC_3 \tau^{\mathrm{T}}\tau h \geq C_4 C^2 n P_n^{-1}.$$

Thus, there exists an event A_{n3} with $\text{pr}(A_{n3}^C) \leq \epsilon$ for any $\epsilon > 0$, such that on A_{n3},

$$-2^{-1}(\alpha_n\tau)^{\mathrm{T}}\{\partial^2 L_n(\lambda^*, \nu^0)/\partial\lambda\partial\lambda^{\mathrm{T}}\}(\alpha_n\tau) \geq \alpha_n^2 C_4 C^2 n P_n^{-1}. \qquad \text{(A.10)}$$

Therefore, by (A.8), (A.9) and (A.10), for n sufficiently large, on the event $A_{n1} \cap A_{n2} \cap A_{n3}$ with $\text{pr}(A_{n1} \cap A_{n2} \cap A_{n3}) \geq 1 - 3\epsilon$, we have

$$\begin{aligned} &L_n(\lambda^0 + \alpha_n\tau, \nu^0) - L_n(\lambda^0, \nu^0) \\ \leq\ & C\alpha_n(\sqrt{KC_1 c\epsilon^{-1}} n^{1/2} + \epsilon^{-1/2} C_3 c^{q+1} n P_n^{-q-1/2}) - \alpha_n^2 C_4 C^2 n P_n^{-1} \\ =\ & C\alpha_n P_n^{-1}\{\sqrt{KC_1 c\epsilon^{-1}} n^{1/2} P_n + \epsilon^{-1/2} C_3 c^{q+1} n P_n^{-q+1/2} - CC_4 n\alpha_n\} \\ =\ & C\alpha_n P_n^{-1}\{\sqrt{KC_1 c\epsilon^{-1}} n^{1/2} P_n + \epsilon^{-1/2} C_3 c^{q+1} n P_n^{-q+1/2} - CC_4 n^{1/2} P_n - CC_4 n P_n^{-q+1/2}\} \\ <\ & 0, \end{aligned}$$

when $C > \max(C_4^{-1}\sqrt{KC_1 c\epsilon^{-1}}, \epsilon^{-1/2} C_4^{-1} C_3 c^{q+1})$. This shows (A.7). Hence, we have $\|\widehat{\lambda}(\nu^0) - \lambda^0\|_2 = O_p(\alpha_n) = O_p(n^{-1/2}P_n + P_n^{-q+1/2})$. A similar strategy for proving consistency has been used in the literature when the dimension of the parameter is diverging, see for example the proof of Theorem 3 in Fan and Lv (2011).

Next, let

$$V_{ik} = \text{var}(Y_{ik} \,|G_{ik}, X_{ik0}, X_{ik}, Z_{ik}, G_{ik}W_{ik}) = H_{ik}(1 - H_{ik}),$$

and

$$\mathbf{V}_n(\nu) = n^{-1}\textstyle\sum_{k=1}^K \sum_{i=1}^{n_k} V_{ik}(\beta_{k1} + \beta_{k2}G_{ik})^2 B_r(U_{ik}(\alpha)) B_r^{\mathrm{T}}(U_{ik}(\alpha)). \qquad \text{(A.11)}$$

By (A.2), (A.6) and the assumption that $P_n^4 n^{-1} = o(1)$,

$$\begin{aligned} &\left\|-n^{-1}\partial^2 L_n(\lambda^0, \nu^0)/\partial\lambda\partial\lambda^{\mathrm{T}} - \mathbf{V}_n(\nu^0)\right\|_\infty \\ =\ & O(h^q)\left\|n^{-1}\textstyle\sum_{k=1}^K \sum_{i=1}^{n_k} B_r(U_{ik}(\alpha^0)) B_r^{\mathrm{T}}(U_{ik}(\alpha^0))\right\|_\infty \\ =\ & O(h^q)\left\|E\left\{B_r(U_{ik}(\alpha^0)) B_r^{\mathrm{T}}(U_{ik}(\alpha^0))\right\}\right\|_\infty + O(h^q) P_n O_p\{\sqrt{hn^{-1}\log(n)}\} \\ =\ & O(h^q)O(h) + O(h^q) P_n O_p\{\sqrt{hn^{-1}\log(n)}\} = O_p(h^{q+1}). \end{aligned}$$

By (A.2) and (A.4), we have $\|\mathbf{V}_n(\nu^0)^{-1}\|_\infty = O_p(h^{-1})$. Thus by the above results, one has

$$\left\| \left\{ -n^{-1}\partial^2 L_n(\lambda^0,\nu^0)/\partial\lambda\partial\lambda^{\mathrm{T}} \right\}^{-1} - \mathbf{V}_n(\nu^0)^{-1} \right\|_\infty = O_p(h^{q-1}).$$

Let

$$\mathbf{D}_n(\nu) = n^{-1}{\textstyle\sum_{k=1}^{K}\sum_{i=1}^{n_k}}(Y_{ik} - H_{ik})(\beta_{k1} + \beta_{k2}G_{ik})B_r(U_{ik}(\alpha)).$$

Since $E\{(\beta_{k1}^0 + \beta_{k2}^0 G_{ik})B_{r,p}(U_{ik}(\alpha^0))\} = O(h)$, by Bernstein's inequality, we have

$$\left\| n^{-1}{\textstyle\sum_{i=1}^{n}\sum_{k=1}^{K}}(\beta_{k1}^0 + \beta_{k2}^0 G_{ik})B_r(U_{ik}(\alpha^0)) \right\|_\infty = O_p\left(h\right).$$

By the above result and (A.6),

$$\begin{aligned} & \left\| n^{-1}\partial L_n(\lambda^0,\nu^0)/\partial\lambda - \mathbf{D}_n(\nu^0) \right\|_\infty \\ = \; & O(h^q)\left\| n^{-1}{\textstyle\sum_{i=1}^{n}\sum_{k=1}^{K}}(\beta_{k1}^0 + \beta_{k2}^0 G_{ik})B_r(U_{ik}(\alpha^0)) \right\|_\infty = O_p\left(h^{q+1}\right). \end{aligned}$$

Let $\mathbb{C}_{ik} = (G_i, X_{ik0}, X_{ik}^{\mathrm{T}}, Z_{ik}^{\mathrm{T}}, G_i W_{ik}^{\mathrm{T}})^{\mathrm{T}}$, $\mathbb{C}_i = (\mathbb{C}_{ik}^{\mathrm{T}}, 1 \le k \le K)^{\mathrm{T}}$, and $\mathbb{C} = (\mathbb{C}_1, \ldots, \mathbb{C}_n)^{\mathrm{T}}$. It can be proved by Bernstein's inequality in Bosq (1961) that $\|\mathbf{D}_n(\nu^0)\|_\infty = O_p\left(\sqrt{hn^{-1}\log(n)}\right)$. Also, by (A.4), $\left\| \left\{ -n^{-1}\partial^2 L_n(\lambda^0,\nu^0)/\partial\lambda\partial\lambda^{\mathrm{T}} \right\}^{-1} \right\|_\infty = O_p\left(h^{-1}\right)$. Thus for $a \in R^{P_n}$ with $\|a\|_2 = 1$,

$$\begin{aligned} & a^{\mathrm{T}}\left[\left\{ -n^{-1}\partial^2 L_n(\lambda^0,\nu^0)/\partial\lambda\partial\lambda^{\mathrm{T}} \right\}^{-1} \left\{ n^{-1}\partial L_n(\lambda^0,\nu^0)/\partial\lambda \right\} - \mathbf{V}_n(\nu^0)^{-1}\mathbf{D}_n(\nu^0) \right] \\ \le \; & \|a\|_\infty \left\| \left\{ -n^{-1}\partial^2 L_n(\lambda^0,\nu^0)/\partial\lambda\partial\lambda^{\mathrm{T}} \right\}^{-1} \right\|_\infty \left\| n^{-1}\partial L_n(\lambda^0,\nu^0)/\partial\lambda - \mathbf{D}_n(\nu^0) \right\|_\infty \\ & + \|a\|_\infty \left\| \left\{ -n^{-1}\partial^2 L_n(\lambda^0,\nu^0)/\partial\lambda\partial\lambda^{\mathrm{T}} \right\}^{-1} - \mathbf{V}_n(\nu^0)^{-1} \right\|_\infty \left\| \mathbf{D}_n(\nu^0) \right\|_\infty \\ = \; & O_p\left(h^q\right) + O_p\left(h^{q-1}\right) O_p\left(\sqrt{hn^{-1}\log(n)}\right). \end{aligned} \tag{A.12}$$

Let $\widehat{e} = \mathbf{V}_n(\nu^0)^{-1}\mathbf{D}_n(\nu^0)$. By Central Limit Theorem, $\left[B_r^{\mathrm{T}}(u)\mathrm{var}\left(\widehat{e}\,|\mathbb{C}\right)B_r(u)\right]^{-1/2} B_r^{\mathrm{T}}(u)\widehat{e} \to$ Normal$(0,1)$, where $\mathrm{var}(\widehat{e}\,|\mathbb{C}) = \{n\mathbf{V}_n(\nu^0)\}^{-1}$ and $B_r^{\mathrm{T}}(u)\mathrm{var}(\widehat{e}\,|\mathbb{C})\,B_r(u) = \widehat{\sigma}^2(u,\nu^0)$. By Lemma 2 and (A.2), there are constants $0 < c_v' < C_v' < \infty$, such that with probability approaching 1, $c_v' h^{-1} \le \|\mathbf{V}_n(\nu^0)^{-1}\|_2 \le C_v' h^{-1}$, and

$$\|\mathbf{V}_n(\nu^0)^{-1} - \mathbf{V}_n^0(\nu^0)^{-1}\|_2 = O_p(h^{-2}\sqrt{hn^{-1}\log(n)}). \tag{A.13}$$

Therefore, there exist constants $0 < c_\sigma \le C_\sigma < \infty$ such that with probability approaching 1 and for large enough n,

$$c_\sigma (nh)^{-1/2} \le \inf_{u\in[a_0,b_0]} \widehat{\sigma}(u,\nu^0) \le \sup_{u\in[a_0,b_0]} \widehat{\sigma}(u,\nu^0) \le C_\sigma (nh)^{-1/2}. \tag{A.14}$$

Thus $B_r^{\mathrm{T}}(u)\widehat{e} = O_p\left\{(nh)^{-1/2}\right\}$ uniformly in $u \in [a_0,b_0]$, and

$$B_r^{\mathrm{T}}(u)\left\{ -\partial^2 L_n(\lambda^0,\nu^0)/\partial\lambda\partial\lambda^{\mathrm{T}} \right\}^{-1} \left\{ \partial L_n(\lambda^0,\nu^0)/\partial\lambda \right\} = O_p\left\{(nh)^{-1/2} + h^q\right\}$$

uniformly in $u \in [a_0, b_0]$. By Taylor's expansion,

$$\widehat{\lambda}(\nu^0) - \lambda^0 = \left\{-\partial^2 L_n(\lambda^0, \nu^0)/\partial\lambda\partial\lambda^{\mathrm{T}}\right\}^{-1} \left\{\partial L_n(\lambda^0, \nu^0)/\partial\lambda\right\} \left\{1 + o_p(1)\right\}. \tag{A.15}$$

Thus by (A.12), (A.14), and Condition (C3),

$$\begin{aligned} & \sup_{u\in[a_0,b_0]} \left|\widehat{\sigma}(u, \nu^0)^{-1} \left[B_r^{\mathrm{T}}(u) \left\{\widehat{\lambda}(\nu^0) - \lambda\right\} - B_r^{\mathrm{T}}(u)\widehat{e}\right]\right| \\ = \; & O_p\left\{(nh)^{1/2}\right\} O_p\left\{(h^q) + O_{a.s.}\left(h^{q-1}\right) O_{a.s.}\left(\sqrt{hn^{-1}\log(n)}\right)\right\} \\ & + O_p\left\{(nh)^{1/2}\right\} o_p\{(nh)^{-1/2} + h^q\} \\ = \; & o_p(1). \end{aligned}$$

Therefore by Slutsky's theorem $\widehat{\sigma}^{-1}(u, \nu^0) \left\{\widehat{m}(u, \nu^0) - \widetilde{m}(u)\right\} \rightarrow \text{Normal}(0, 1)$ and $\widehat{m}(u, \nu^0) - \widetilde{m}(u) = O_p\left\{(nh)^{-1/2}\right\}$ uniformly in $u \in [a_0, b_0]$. By $\sup_{u\in[a_0,b_0]} |m(u) - \widetilde{m}(u)| = o(h^q)$, we have $|\widehat{m}(u, \nu^0) - m(u)| = O_p\{(nh)^{-1/2} + h^q\}$ uniformly in $u \in [a_0, b_0]$. By Slutsky's theorem, we have

$$\widehat{\sigma}^{-1}(u, \nu^0) \left\{\widehat{m}(u, \nu^0) - m(u)\right\} \rightarrow \text{Normal}(0, 1).$$

Since $\widehat{m}'(u, \nu) = B_{r-1}^{\mathrm{T}}(u) D_1 \widehat{\lambda}(\nu)$ and $B_{r-1}(u)$ are B-spline basis functions with one order lower than $B_r(u)$, by the same argument as in Zhou and Wolfe (2000) and the proof for $\widehat{m}(u, \nu^0)$, we have the result (b) in Theorem 1. Then the proof is complete.

A.4 Proof of Theorem 2

Define $L_{ik}(\nu) = (\beta_{k1} + \beta_{k2} G_{ik}) m(X_{ik0} + X_{ik}^{\mathrm{T}}\alpha) + Z_{ik}^{T}(\theta_{k1} + \theta_{k2} G_{ik}) + G_{ik} W_{ik}^{\mathrm{T}} \theta_{k3}$. It is straightforward to prove that $\partial L_{i1}(\nu)/\partial\nu = Q_{i1}(\nu)$, and for $k = 2, \ldots, K$, $\partial L_{ik}(\nu)/\partial\nu = Q_{ik}(\nu)$. Then by (A.15) and Condition (C3) and by the same arguments as the proof for proposition 4.1 in Ai and Chen (2003), we have

$$\begin{aligned} \partial L_n(\nu^0)/\nu \; = \; & \textstyle\sum_{k=1}^{K}\sum_{i=1}^{n_k} [(Y_{ik} - H_{ik}(\nu^0)) \times \\ & \{\partial L_{ik}(\nu^0)/\partial\nu - (\beta_{k1}^0 + \beta_{k2}^0 G_{ik})\eta^0(U_{ik}(\alpha^0))\}] \left\{1 + o_p(1)\right\} \\ = \; & \textstyle\sum_{k=1}^{K}\sum_{i=1}^{n_k} [(Y_{ik} - H_{ik}(\nu^0)) \times \\ & \{Q_{ik}(\nu^0) - (\beta_{k1}^0 + \beta_{k2}^0 G_{ik})\eta^0(U_{ik}(\alpha^0))\}] \left\{1 + o_p(1)\right\}, \end{aligned}$$

$$\begin{aligned} & \partial^2 L_n(\nu^0)/\partial\nu\partial\nu^{\mathrm{T}} \\ = \; & -\textstyle\sum_{k=1}^{K}\sum_{i=1}^{n_k} [H_{ik}(\nu^0)(1 - H_{ik}(\nu^0)) \times \\ & \{\partial L_{ik}(\nu^0)/\partial\nu - (\beta_{k1}^0 + \beta_{k2}^0 G_{ik})\eta^0(U_{ik}(\alpha^0))\}^{\otimes 2}] \left\{1 + o_p(1)\right\} \\ = \; & -\textstyle\sum_{k=1}^{K}\sum_{i=1}^{n_k} [V_{ik}(\nu^0)\{Q_{ik}(\nu^0) - (\beta_{k1}^0 + \beta_{k2}^0 G_{ik})\eta^0(U_{ik}(\alpha^0))\}^{\otimes 2}] \left\{1 + o_p(1)\right\}. \end{aligned}$$

By Taylor's expansion, we have

$$\nu - \nu^0 = -\{\partial^2 L_n(\nu^0)/\partial\nu\partial\nu^{\mathrm{T}}\}^{-1}\{\partial L_n(\nu^0)/\nu\} \left\{1 + o_p(1)\right\}.$$

By the above result, we have (8). Then the asymptotic normality in Theorem 2 follows from the Central Limit Theorem and (8).

A.5 Proof of Theorem 3

Here we show that our method for estimating ν is semiparametric efficient when $(Y_{i1}, ..., Y_{iK})$ are independent given $\mathbb{C}_i$. We have that

$$\log\{\mathrm{pr}(Y_i = y_i \mid \mathbb{C}_i)\} = \textstyle\sum_{k=1}^{K} \{y_{ik}\log(H_{ik}) + (1 - y_{ik})\log(1 - H_{ik})\}\,.$$

The i^{th} score with respect to ν is $S_{\nu i} = \sum_{k=1}^{K}(Y_{ik} - H_{ik})Q_{ik}$. The nuisance tangent space is

$$\Lambda = \left\{\textstyle\sum_{k=1}^{K}(Y_{ik} - H_{ik})(\beta_{k1}^0 + \beta_{k2}^0 G_{ik})\eta(X_{ik0} + X_{ik}^{\mathrm{T}}\alpha^0) : \eta(\cdot) \in \mathbb{R}^{J+2K+2Kd+Ka-2}\right\}.$$

We decompose $S_{\nu i}$ as $S_{\nu i} = S_{\mathrm{eff},i} + S_{1i}$, where

$$\begin{aligned}
S_{\mathrm{eff},i} &= \textstyle\sum_{k=1}^{K}(Y_{ik} - H_{ik})\left\{Q_{ik} - (\beta_{k1}^0 + \beta_{k2}^0 G_{ik})\eta_{0i}\right\},\\
S_{1i} &= \textstyle\sum_{k=1}^{K}(Y_{ik} - H_{ik})(\beta_{k1}^0 + \beta_{k2}^0 G_{ik})\eta_{0i},\\
\eta_{0i} &= \frac{E\left\{\sum_{k=1}^{K}V_{ik}Q_{ik}(\beta_{k1}^0 + \beta_{k2}^0 G_{ik}) \mid X_{ik0} + X_{ik}^{\mathrm{T}}\alpha^0\right\}}{E\left\{\sum_{k=1}^{K}V_{ik}(\beta_{k1}^0 + \beta_{k2}^0 G_{ik})^2 \mid X_{ik0} + X_{ik}^{\mathrm{T}}\alpha^0\right\}}.
\end{aligned}$$

Obviously, $S_{1i} \in \Lambda$. For any element $S_i \in \Lambda$, say $S_i = \sum_{k=1}^{K}(Y_{ik} - H_{ik})(\beta_{k1}^0 + \beta_{k2}^0 G_{ik})\eta(X_{ik0} + X_{ik}^{\mathrm{T}}\alpha^0)$, we can easily verify that $E(S_{\mathrm{eff},i}^{\mathrm{T}}S_i) = 0$.

Thus, $S_{\mathrm{eff},i}$ is the residual of the orthogonal projection of $S_{\nu i}$ onto Λ, hence it is the efficient score. The minimum variance bound for estimating ν is therefore

$$\mathrm{cov}_{opt}\{n^{1/2}(\widehat{\nu} - \nu)\} = \{E(S_{\mathrm{eff},i}S_{\mathrm{eff},i}^{\mathrm{T}})\}^{-1} = [E\textstyle\sum_{k=1}^{K}V_{ik}\left\{Q_{ik} - (\beta_{k1}^0 + \beta_{k2}^0 G_{ik})\eta_{0i}\right\}^{\otimes 2}]^{-1}.$$

Since S_{1i} is the orthogonal projection of $S_{\nu i}$ onto Λ, it minimizes the covariance matrix of $S_{\nu i} - S_i$ among all the functions $S_i \in \Lambda$, i.e., η_{0i} minimizes

$$\begin{aligned}
&\mathrm{cov}[\textstyle\sum_{k=1}^{K}(Y_{ik} - H_{ik})\{Q_{ik} - (\beta_{k1}^0 + \beta_{k2}^0 G_i)\eta(X_{ik0} + X_{ik}^{\mathrm{T}}\alpha^0)\}]\\
&\quad = E[\textstyle\sum_{k=1}^{K}V_{ik}\{Q_{ik} - (\beta_{k1}^0 + \beta_{k2}^0 G_i)\eta(X_{ik0} + X_{ik}^{\mathrm{T}}\alpha^0)\}^{\otimes 2}]
\end{aligned}$$

among all possible $\eta(X_{ik0} + X_{ik}^{\mathrm{T}}\alpha^0) \in \mathbb{R}^{J+2K+2Kd+Ka-2}$. This shows that $\mathbf{\Sigma}$ in Theorem 2 reaches the semiparametric efficiency bound, as claimed.

A.6 Proof of Theorem 4

Let $\varepsilon_{ik} = (\beta_{k1}^0 + \beta_{k2}^0 G_i)(Y_{ik} - H_{ik})$ and $\varepsilon_i = (\varepsilon_{i1}, \ldots, \varepsilon_{iK})^{\mathrm{T}}$. Following the same procedure as the proof of Theorem 1, we have that $||\widehat{\lambda}_w(\nu^0) - \lambda^0||_2 = O_p(n^{-1/2}P_n + P_n^{-q+1/2})$. By this result and Taylor's expansion, we have

$$0 = \textstyle\sum_{i=1}^{n}B_r(U_i)\mathbf{A}_i\mathbf{B}_i^{-1}\mathbf{A}_i^{\mathrm{T}}B_r^{\mathrm{T}}(U_i)\{\lambda^0 - \widehat{\lambda}_w(\nu^0)\} + \textstyle\sum_{i=1}^{n}B_r(U_i)\mathbf{A}_i\mathbf{B}_i^{-1}\varepsilon_i + o_p(n^{1/2}).$$

Thus

$$\begin{aligned}\widehat{\lambda}_w(\nu^0) - \lambda^0 \;&=\; \{n^{-1}\textstyle\sum_{i=1}^n B_r(U_i)\mathbf{A}_i\mathbf{B}_i^{-1}\ \mathbf{A}_i^{\rm T}B_r^{\rm T}(U_i)\}^{-1} \\ &\qquad\qquad \times\{n^{-1}\textstyle\sum_{i=1}^n B_r(U_i)\mathbf{A}_i\mathbf{B}_i^{-1}\varepsilon_i\}\{1+o_p(1)\}. \qquad (\text{A}.16)\end{aligned}$$

Then with probability approaching 1,var$\{\widehat{\lambda}_w(\nu^0) - \lambda^0 | \mathbb{C}_i\}$ approaches $\Pi_n^{-1}\Xi_n\Pi_n^{-1}$. Theorem 4 can be proved following the same methods as in the proof of Theorem 1.

A.7 Proof of Theorem 5

Let ζ_i be the $d_\nu K \times 1$ vector formed by K length d_ν vectors. The kth, $k = 1, \ldots, K$ vector component is $(Y_{ik} - H_{ik}(\nu))\{\widehat{Q}_{ik}(\nu) + (\beta_{k1} + \beta_{k2}G_i)\{\widehat{\lambda}'_w(\nu)\}^{\rm T}B_r(X_{i0} + X_i^{\rm T}\alpha)\}$. Following the same outline as the proof of Theorem 2, it can be proved that

$$\sqrt{n}(\widehat{\nu}_w - \nu^0) = \sqrt{n}(\textstyle\sum_{i=1}^n \mathbf{C}_i\mathbf{D}_i^{-1}\mathbf{C}_i^{\rm T})^{-1}(\sum_{i=1}^n \mathbf{C}_i\mathbf{D}_i^{-1}\zeta_i) + o_p(1).$$

Therefore,

$$\begin{aligned}\text{var}(\sqrt{n}(\widehat{\nu}_w - \nu^0)\,|\mathbb{C}_i\,) \;&=\; n(\textstyle\sum_{i=1}^n \mathbf{C}_i\mathbf{D}_i^{-1}\mathbf{C}_i^{\rm T})^{-1}(\sum_{i=1}^n \mathbf{C}_i\mathbf{D}_i^{-1}\mathbf{D}_i^{*}\mathbf{D}_i^{-1}\mathbf{C}_i^{\rm T}) \\ &\qquad \times(\textstyle\sum_{i=1}^n \mathbf{C}_i\mathbf{D}_i^{-1}\mathbf{C}_i^{\rm T})^{-1} + o_p(1),\end{aligned}$$

and the asymptotic normality of $\sqrt{n}(\widehat{\nu}_w - \nu^0)$ given in Theorem 5 follows from the Central Limit Theorem.

A.8 Extending to Multiple Study Centers

Here we indicate briefly the necessary changes needed if there are multiple study centers, and multiple dependent disease outcomes within each study center. Suppose that there are $k = 1, ..., K$ study centers, with $\ell = 1, ..., L_k$ binary disease outcomes in each center, and with $i = 1, ..., n_k$ observations at the k^{th} center. Write the outcomes at $\mathbf{Y}_{ik} = (Y_{ik1}, ..., Y_{ikL_k})$, and write the covariates as $\mathbb{C}_{ik} = (G_{ik}, X_{ik0}, X_{ik}, Z_{ik}, G_{ik}W_{ik})$. The model is

$$\begin{aligned}&\text{pr}(Y_{ik\ell} = 1 \mid \mathbb{C}_{ik}) = H_{ik\ell} \qquad\qquad (\text{A}.17)\\ &\qquad = H_{ik\ell} = H\{(\beta_{k\ell 1} + \beta_{k\ell 2}G_{ik})m(X_{ik0} + X_{ik}^{\rm T}\alpha) + Z_{ik}^{\rm T}(\theta_{k\ell 1} + \theta_{k\ell 2}G_{ik}) + G_{ik}W_{ik}^{\rm T}\theta_{k\ell 3}\}.\end{aligned}$$

We make the same assumptions as in Section A.2, but in addition we assume that $\lim_{n_1,\ldots,n_K\to\infty}(\max n_k/\min n_k) = c$ with $0 < c < \infty$.

From the above model, we can see that in different centers, because different physical populations are studied, the same disease occurrence is modeled with different parameters. Thus, we can simply view the L_k diseases in $k = 1, ..., K$ centers as $\sum_{k=1}^K L_k$ different diseases from a single center, and all our analyses formulated for data from one center applies.

Component	Units	HEI-2005 score calculation
Total Fruit	cups	$\min\{5, 5 \times (\text{density}/.8)\}$
Whole Fruit	cups	$\min\{5, 5 \times (\text{density}/.4)\}$
Total Vegetables	cups	$\min\{5, 5 \times (\text{density}/1.1)\}$
DOL	cups	$\min\{5, 5 \times (\text{density}/.4)\}$
Total Grains	ounces	$\min\{5, 5 \times (\text{density}/3)\}$
Whole Grains	ounces	$\min\{5, 5 \times (\text{density}/1.5)\}$
Milk	cups	$\min\{10, 10 \times (\text{density}/1.3)\}$
Meat and Beans	ounces	$\min\{10, 10 \times (\text{density}/2.5)\}$
Oil	grams	$\min\{10, 10 \times (\text{density}/12)\}$

Saturated Fat	% of energy	if density $\geq$ 15 score = 0
		else if density $\leq$ 7 score = 10
		else if density $>$ 10 score $= 8 - \{8 \times (\text{density} - 10)/5\}$
		else, score $= 10 - \{2 \times (\text{density} - 7)/3\}$

Sodium	milligrams	if density $\geq$ 2000 score=0
		else if density $\leq$ 700 score=10
		else if density $\geq$ 1100
		score $= 8 - \{8 \times (\text{density} - 1100)/(2000 - 1100)\}$
		else score $= 10 - \{2 \times (\text{density} - 700)/(1100 - 700)\}$

SoFAAS	% of energy	if density $\geq$ 50 score = 0
		else if density $\leq$ 20 score=20
		else score $= 20 - \{20 \times (\text{density} - 20)/(50 - 20)\}$

Table 1: Description of the HEI-2005 scoring system. Except for saturated fat and SoFAAS, density is obtained by multiplying intake by 1000 and dividing by intake of kilo-calories. For saturated fat, density is 9×100 saturated fat (grams) divided by calories, i.e., the percentage of calories coming from saturated fat intake. For SoFAAS, the density is the percentage of intake that comes from intake of calories, i.e., the division of intake of SoFAAS by intake of calories. Here, "DOL" is dark green and orange vegetables and legumes. Also, "SoFAAS" is calories from solid fats, alcoholic beverages and added sugars. The total HEI-2005 score is the sum of the individual component scores.

	Men		Women	
Description	# Cases	Percentages	# Cases	Percentages
Sample size	294,673		199,285	
Breast cancer			7,736	3.88%
Ovarian cancer			759	0.38%
Prostate cancer	23,477	7.97%		
Colorectal cancer	4,693	1.59%	2,291	1.15%
Lung cancer	6,135	2.08%	3,630	1.82%

Table 2: Summary of the NIH-AARP data.

	Colorectal Cancer			Lung Cancer		
	Estimate	se	p-value	Estimate	se	p-value
Total Fruit	**0.27**	0.87	0.40	**2.18**	0.34	0.00
Whole Fruit	0.54	0.81	0.57	1.33	0.33	0.32
Total Grains	2.58	0.85	0.06	2.96	0.33	0.00
Whole Grains	**2.44**	0.85	0.09	**0.53**	0.27	0.08
Total Vegetables	0.01	1.02	0.33	0.99	0.36	0.98
DOL Vegetables	1.33	0.72	0.65	0.99	0.26	0.96
Dairy	**2.44**	0.42	0.00	**0.42**	0.10	0.00
Meat and Beans	0.00	0.53	0.06	0.00	0.18	0.00
Oils	0.58	0.32	0.20	0.33	0.11	0.00
Sodium	0.80	0.45	0.65	1.12	0.16	0.45
Saturated Fat	0.53	0.31	0.13	0.94	0.13	0.65
Empty Calories	0.49	0.21	0.02	0.21	0.08	0.00

Table 3: Results for the analysis of Section 5.3, where lung cancer and colorectal cancer were analyzed separately, thus each analysis has one disease and 2 independent populations. The weights of the component scores were normalized so that their sum = 12, thus placing the weights on the same scale as the HEI-2005 total score, whose weights all = 1. The p-values for the test that the individuals components = 1 are also displayed.

	Estimate	se	p-value
Total Fruit	1.89	0.31	0.00
Whole Fruit	1.32	0.30	0.27
Total Grains	2.94	0.30	0.00
Whole Grains	0.70	0.26	0.32
Total Vegetables	0.97	0.34	0.93
DOL Vegetables	0.93	0.24	0.81
Dairy	0.61	0.09	0.00
Meat and Beans	0.00	0.17	0.00
Oils	0.39	0.11	0.00
Sodium	1.13	0.15	0.36
Saturated Fat	0.89	0.12	0.40
Empty Calories	0.23	0.07	0.00

Table 4: Results for estimated weights $\widehat{\alpha}$ in the analysis of Section 5.4, with two populations (men and women), three diseases for men (lung, colorectal and prostate cancer) and four diseases for women (lung, colorectal, breast and ovarian cancer). The weights of the component scores were normalized so that their sum = 12, thus placing the weights on the same scale as the HEI-2005 total score, whose weights all = 1. The p-values for the test that the individuals components = 1 are also displayed. The actual estimated weights for Meat and Beans was actually negative, but we have set it = 0 for nutritional purposes.

	Estimate	se	p-value
Men, Lung	-1.00	NA	NA
Men, Colorectal	-0.39	0.06	0.00
Men, Prostate	0.00	0.06	0.07
Women, Lung	-0.91	0.07	0.00
Women, Colorectal	-0.28	0.08	0.00
Women, Breast	-0.07	0.04	0.11
Women, Ovarian	0.00	0.14	0.90

Table 5: Results for $\widehat{\beta}$ for the analysis of Section 5.4, with two populations (men and women), three diseases for men (lung, colorectal and prostate cancer) and four diseases for women (lung, colorectal, breast and ovarian cancer). The weights of the component scores were normalized so that their sum = 12, thus placing the weights on the same scale as the HEI-2005 total score, whose weights all = 1. The p-values for the test that the individual $\widehat{\beta}$ terms = 0 are also displayed. The actual estimated coefficients for Prostate and Ovarian cancers were positive, but we have set them = 0 for nutritional purposes, with the constraint that a better diet is not a risk factor for either disease.

	Mean	Estimated se	Actual se	Coverage
Total Fruit	1.03	0.32	0.34	93.00
Whole Fruit	1.04	0.55	0.59	95.40
Total Grains	1.00	0.55	0.58	95.00
Whole Grains	0.99	0.37	0.38	94.10
Total Vegetables	1.01	0.44	0.45	94.20
DOL Vegetables	0.99	0.38	0.38	94.90
Dairy	1.01	0.27	0.28	95.10
Meat and Beans	1.01	0.29	0.31	94.00
Oils	1.00	0.28	0.29	94.10
Sodium	0.98	0.35	0.37	94.30
Saturated Fat	0.98	0.40	0.42	94.20
Empty Calories	0.96	0.46	0.47	94.40

Table 6: Results of the simulation study when $n = 3000$, the binary responses have correlation 0.05 and where the actual values of α all = 1.00. Here *Estimate* is the mean of the estimates, *Estimated se* is the mean of the estimated standard errors, *Actual se* is the actual standard deviation of the estimates, and *Coverage* is the actual coverage of a nominal 95% confidence interval. The actual estimates $\widehat{\alpha}$ were normalized to sum to 12.

	Mean # Cases	True β	Mean $\widehat{\beta}$	Estimated se	Actual se	Coverage
Population 1, Disease 1	826	-1.00	-1.00	NA	NA	NA
Population 1, Disease 2	1052	-0.60	-0.61	0.12	0.11	97.00
Population 1, Disease 3	1347	-0.20	-0.21	0.10	0.09	97.70
Population 2, Disease 1	957	-0.80	-0.80	0.14	0.14	94.40
Population 2, Disease 2	1104	-0.57	-0.57	0.12	0.13	94.50
Population 2, Disease 3	1265	-0.33	-0.34	0.11	0.11	94.00
Population 2, Disease 4	1422	-0.10	-0.11	0.10	0.09	98.40

Table 7: Simulation results for β when $n = 3000$ and the binary responses have correlation 0.05. Estimate is the mean, Estimated se is the mean of the estimated standard errors, Actual se is the actual standard deviation of the estimates, and Coverage is the actual coverage of a nominal 95% confidence interval. The average total number of cases across the simulation = 7,975.

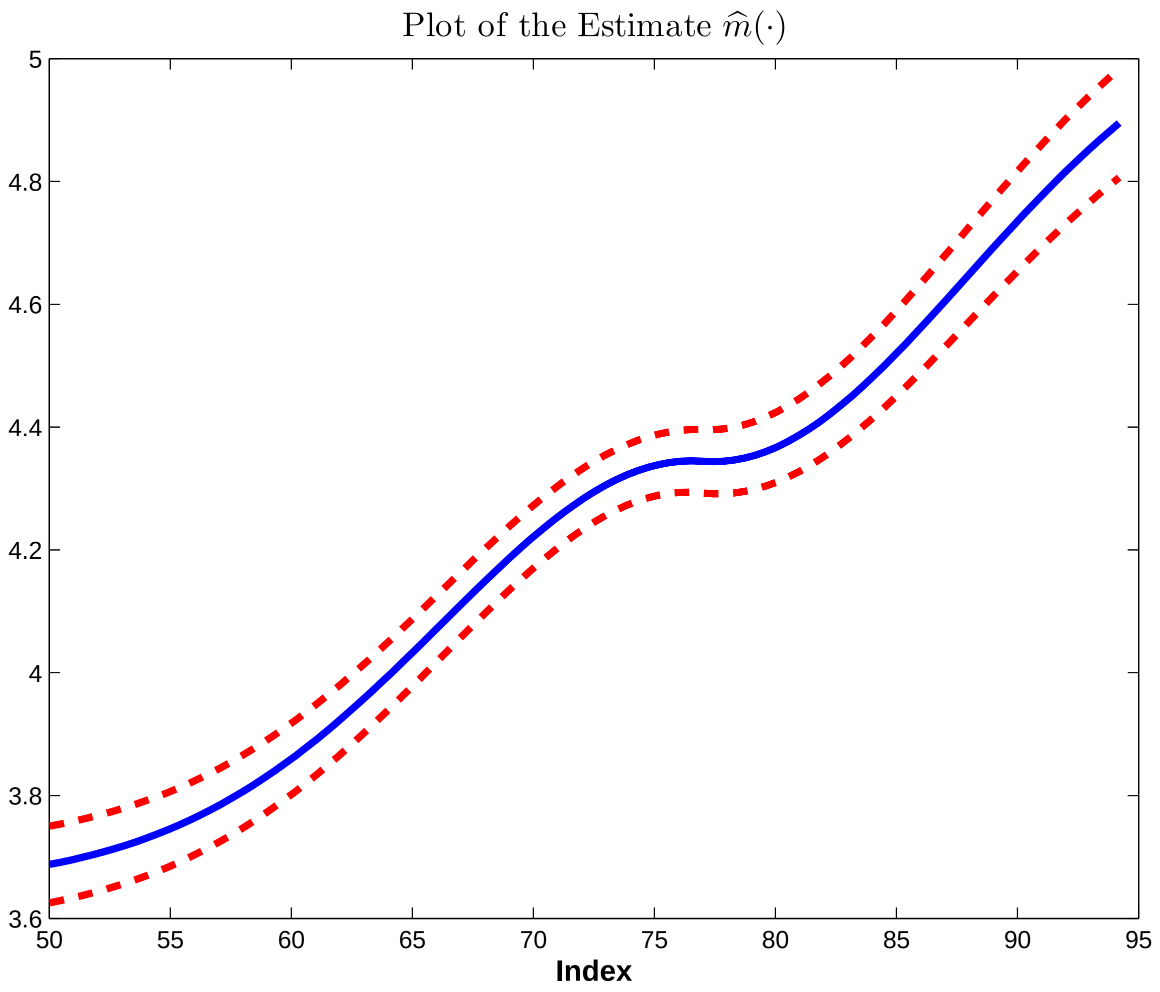

Figure 1: Analysis of multiple diseases as in Section 5.4. The function $\widehat{m}(X^{\mathrm{T}}\widehat{\alpha})$ along with its pointwise 95% confidence interval.

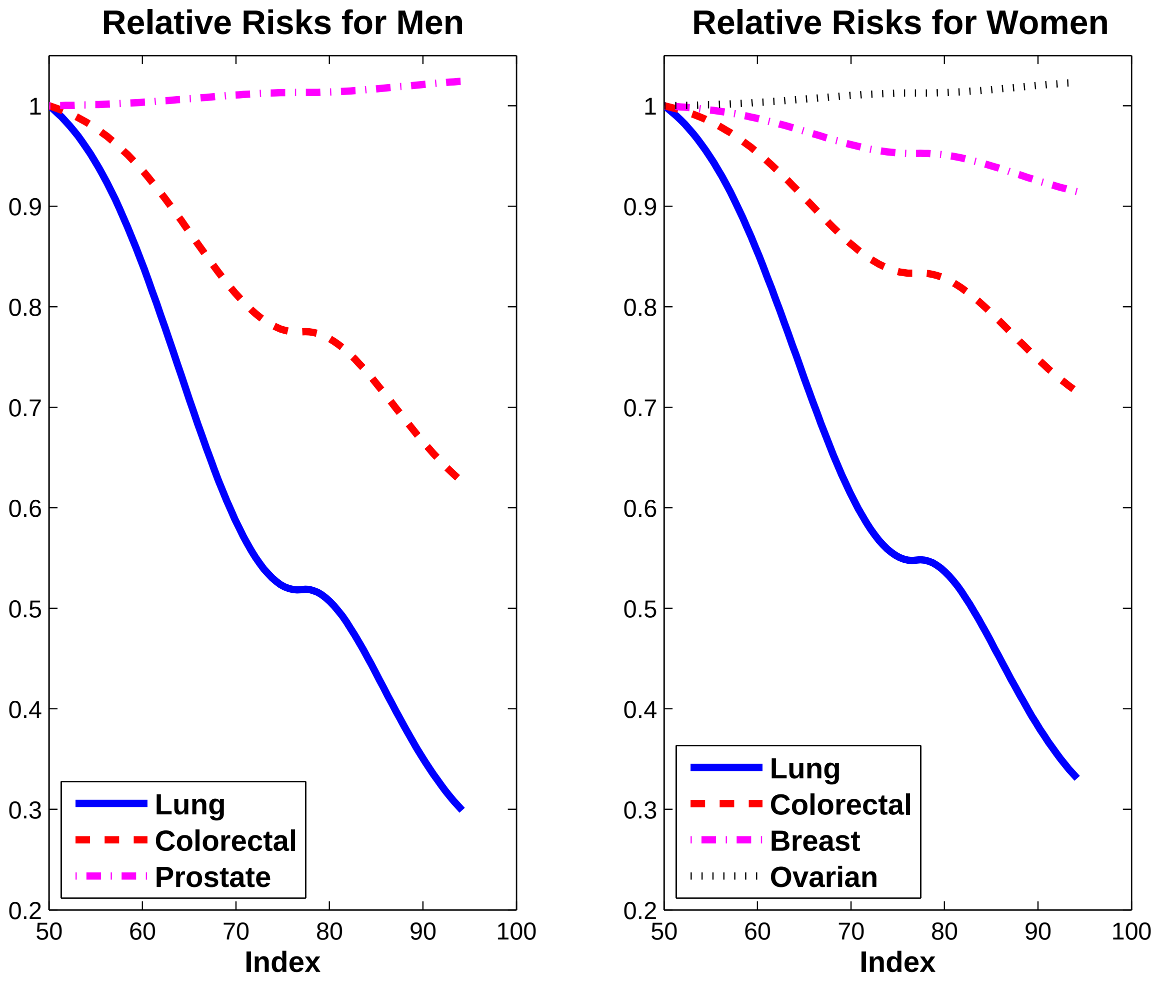

Figure 2: Analysis of multiple diseases as in Section 5.4. Relative risks for men and women on a grid between the 3rd and 97th percentile of the index. Left panel is for men: solid blue line is the relative risk for lung cancer, while the dashed red line is for colorectal cancer. The right is for women: solid blue line is the relative risk for lung cancer, dashed red line is for colorectal cancer and the dot-dashed magenta line is for breast cancer.

Supplementary Material to *A Semiparametric Single-Index Risk Score Across Populations*

Shujie Ma
Department of Statistics, University of California at Riverside, Riverside, CA 92521,
shujie.ma@ucr.edu

Yanyuan Ma
Department of Statistics, University of South Carolina, Columbia SC 29208,
ma44@mailbox.sc.edu

Yanqing Wang
Fred Hutchinson Cancer Research Center, Seattle WA 98109, ywang237@fredhutch.org

Eli S. Kravitz
Department of Statistics, Texas A&M University, 3143 TAMU, College Station, TX
77843-3143, kravitze@tamu.edu

Raymond J. Carroll
Department of Statistics, Texas A&M University, 3143 TAMU, College Station, TX
77843-3143, and School of Mathematical Sciences, University of Technology Sydney,
Broadway NSW 2007, carroll@stat.tamu.edu

S.1 Correlations in the HEI Component Scores

Table S.1 gives correlations of the HEI-2005 component scores.

Men

	FTot	FWhl	GTot	GWhl	VTot	DOL	Milk	Meat	Oil	SFat	Sodi	SoFS
FTot	1.00	0.74	0.12	0.20	0.20	0.22	0.07	-0.06	-0.09	0.33	-0.02	0.43
FWhl	0.74	1.00	0.18	0.23	0.26	0.27	0.08	-0.02	-0.04	0.26	-0.11	0.42
GTot	0.12	0.18	1.00	0.49	0.15	0.11	0.00	0.12	0.04	0.12	-0.49	0.44
GWhl	0.20	0.23	0.49	1.00	0.09	0.13	0.09	-0.07	-0.07	0.24	-0.25	0.39
VTot	0.20	0.26	0.15	0.09	1.00	0.66	-0.09	0.22	0.17	0.13	-0.49	0.41
DOL	0.22	0.27	0.11	0.13	0.66	1.00	-0.08	0.15	0.04	0.20	-0.35	0.40
Milk	0.07	0.08	0.00	0.09	-0.09	-0.08	1.00	-0.13	-0.13	-0.11	-0.07	0.16
Meat	-0.06	-0.02	0.12	-0.07	0.22	0.15	-0.13	1.00	0.24	-0.20	-0.42	0.13
Oil	-0.09	-0.04	0.04	-0.07	0.17	0.04	-0.13	0.24	1.00	-0.16	-0.20	0.16
SFat	0.33	0.26	0.12	0.24	0.13	0.20	-0.11	-0.20	-0.16	1.00	0.07	0.27
Sodi	-0.02	-0.11	-0.49	-0.25	-0.49	-0.35	-0.07	-0.42	-0.20	0.07	1.00	-0.54
SoFS	0.43	0.42	0.44	0.39	0.41	0.40	0.16	0.13	0.16	0.27	-0.54	1.00

Women

	FTot	FWhl	GTot	GWhl	VTot	DOL	Milk	Meat	Oil	SFat	Sodi	SoFS
FTot	1.00	0.72	0.02	0.12	0.17	0.22	0.08	-0.10	-0.14	0.34	0.06	0.38
FWhl	0.72	1.00	0.09	0.15	0.24	0.26	0.10	-0.04	-0.09	0.27	-0.04	0.37
GTot	0.02	0.09	1.00	0.49	0.05	0.02	-0.05	0.08	-0.01	0.15	-0.39	0.30
GWhl	0.12	0.15	0.49	1.00	0.03	0.09	0.07	-0.09	-0.11	0.24	-0.22	0.33
VTot	0.17	0.24	0.05	0.03	1.00	0.65	-0.08	0.19	0.12	0.15	-0.44	0.38
DOL	0.22	0.26	0.02	0.09	0.65	1.00	-0.06	0.12	-0.02	0.23	-0.32	0.38
Milk	0.08	0.10	-0.05	0.07	-0.08	-0.06	1.00	-0.17	-0.19	-0.01	-0.03	0.14
Meat	-0.10	-0.04	0.08	-0.09	0.19	0.12	-0.17	1.00	0.20	-0.14	-0.40	0.09
Oil	-0.14	-0.09	-0.01	-0.11	0.12	-0.02	-0.19	0.20	1.00	-0.17	-0.18	0.09
SFat	0.34	0.27	0.15	0.24	0.15	0.23	-0.01	-0.14	-0.17	1.00	-0.02	0.41
Sodi	0.06	-0.04	-0.39	-0.22	-0.44	-0.32	-0.03	-0.40	-0.18	-0.02	1.00	-0.46
SoFS	0.38	0.37	0.30	0.33	0.38	0.38	0.14	0.09	0.09	0.41	-0.46	1.00

Table S.1: The correlations of the HEI-2005 component scores for men and women separately. Here ”FTot” is Total Fruit, ”FWhl” is Whole Fruit, ”GTot” is Total Grains, ”GWhl” is Whole Grains, ”VTot” is Total Vegetables, ”DOL” is DOL, ”Milk” is Milk, ”Meat” is Meat and Beans, ”Oil” is Oil, ”SFat” is Saturated Fat, ”Sodi” is Sodium and ”SoFS” is SoFAAS.

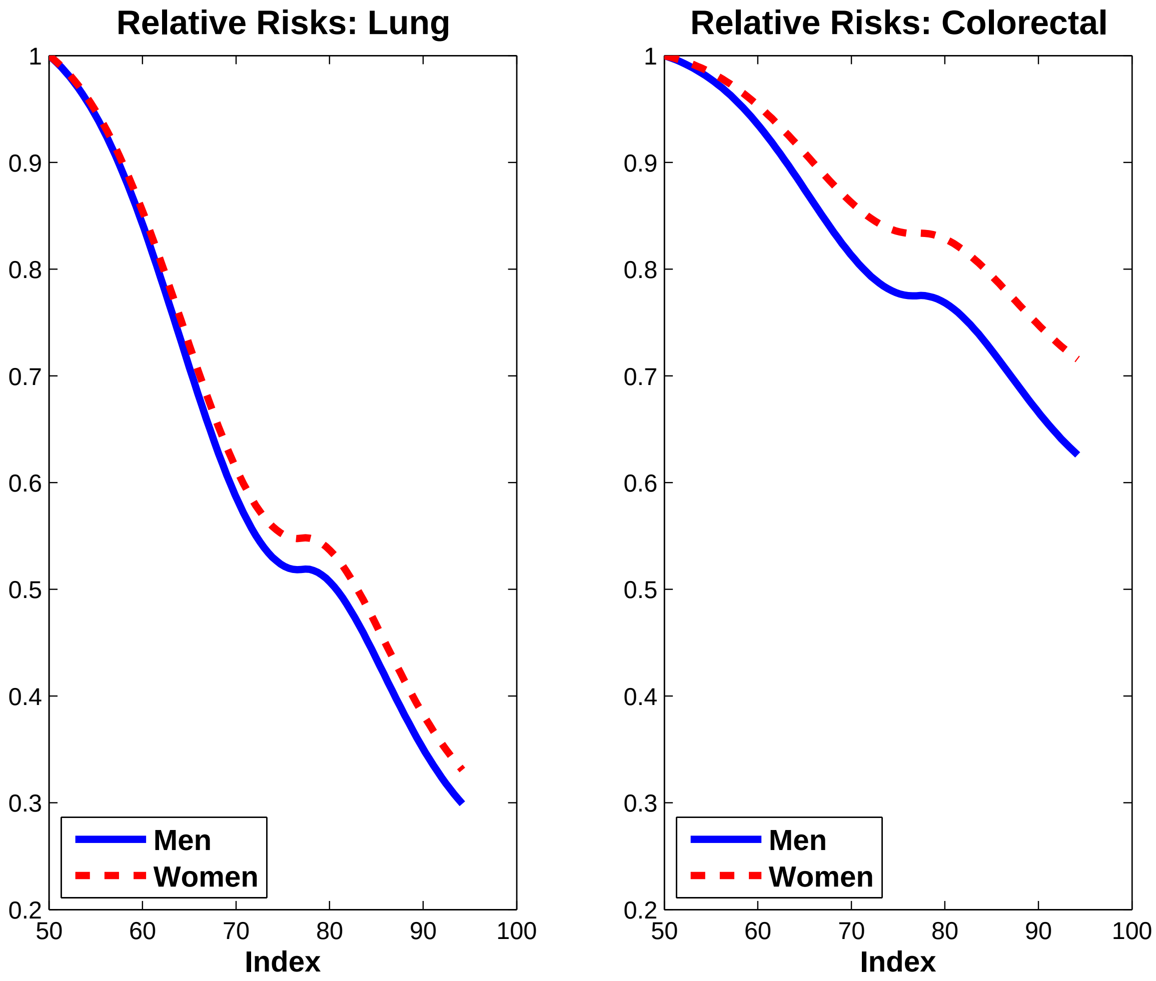

Figure S.1: Analysis of multiple diseases as in Section 5.4. Relative risks for lung and colorectal cancer. Solid blue lines are for men, dashed red lines are for women. Left panel is lung cancer, and right panel is colorectal cancer.

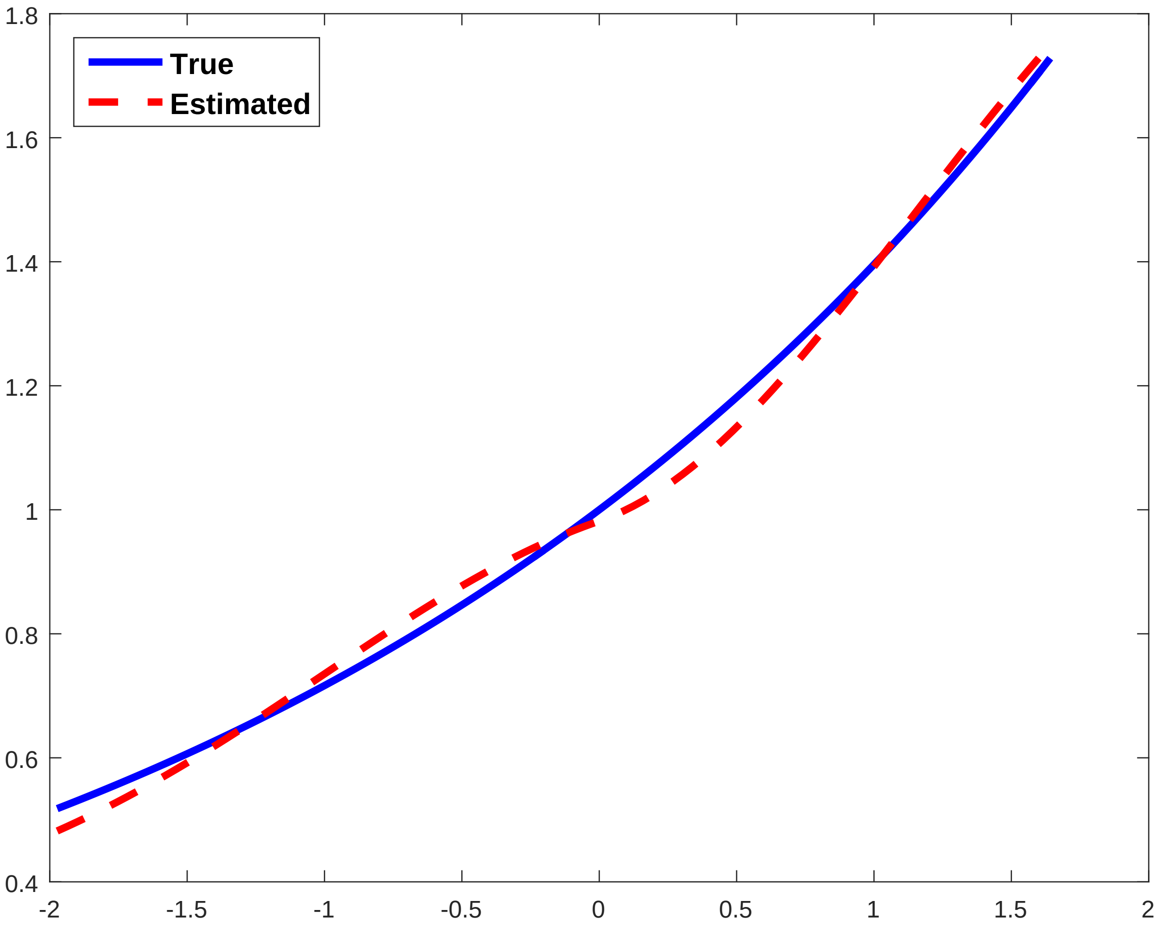

Figure S.2: Results of the simulation study in Section 6 when $n = 3000$. The solid blue line is the true function $m(u)$, while the dashed red line is the mean estimates from the simulation.

S.1 Simulation Study for Independent Populations, Single Disease

We simulate data from the marginal logit model,

$$\begin{aligned}&\text{logit}\{\text{pr}(Y_{ik}=1\mid G_{ik},X_{ik0},X_{ik},Z_{ik},G_{ik}W_{ik})\}\\&\quad=\beta_k m(X_{ik}^{\mathrm{T}}\alpha)+Z_{ik}^{\mathrm{T}}\theta_{k1}+G_{ik}W_{ik}^{\mathrm{T}}\theta_{k3},\end{aligned}$$

for $i=1,\ldots,n$ and $k=1,2$, so that we consider two independent populations, men and women as given in the real data example and let $G_{i1}=0$ for men and $G_{i2}=1$ for women. Moreover, we use the twelve HEI-2005 scores described in Table 1 as the variables $X_{ik}, k=0,\ldots,11$. We use dummy variables for age categories and body mass index categories as the variables Z_{ik}, and for women, we use the two dummy variables for hormone replacement therapy as the variables W_{ik} from the NIH-AARP Study of Diet and Health. We let each component in α be $1/\sqrt{12}=0.289$, and simulate β_k from Uniform$[0,1]$ and θ_{k1} and θ_{k3} from Uniform$[-0.5,0.5]$. The nonparametric function takes the form $m(u)=\exp(u/3)$.

The sample size were $n=1000,2000,5000$, respectively, and 200 simulation replications are run to draw summary statistics. Table S.1 reports the median value of the asymptotic standard error (ASE) calculated according to Theorem 5, the empirical standard error (ESE) and the absolute mean of the bias (Bias) among 200 replications for the estimate of each component in α. The biases for estimating α are small, and as expected from the theory, decrease with increasing sample size. In addition, the differences between the asymptotic standard errors for the estimates of α and the estimated standard errors are small and decrease with increasing sample size. In results not reported here, the function estimates were nearly unbiased, and became more precise as sample size increases.

n		α_1	α_2	α_3	α_4	α_5	α_6	α_7	α_8	α_9	α_{10}	α_{11}	α_{12}
	ASE	0.162	0.242	0.239	0.185	0.213	0.187	0.147	0.159	0.150	0.202	0.187	0.218
1000	ESE	0.129	0.201	0.204	0.162	0.184	0.170	0.119	0.131	0.122	0.146	0.173	0.205
	Bias	0.066	0.043	0.046	0.048	0.057	0.068	0.074	0.045	0.050	0.065	0.030	0.031
	ASE	0.121	0.200	0.197	0.148	0.169	0.142	0.112	0.125	0.115	0.156	0.148	0.158
2000	ESE	0.110	0.176	0.181	0.128	0.140	0.127	0.098	0.111	0.092	0.122	0.138	0.148
	Bias	0.033	0.039	0.021	0.039	0.028	0.036	0.038	0.017	0.040	0.031	0.018	0.041
	ASE	0.086	0.144	0.122	0.108	0.116	0.101	0.080	0.085	0.081	0.091	0.107	0.114
5000	ESE	0.077	0.128	0.112	0.091	0.098	0.088	0.072	0.069	0.064	0.073	0.094	0.108
	Bias	0.019	0.011	0.011	0.017	0.004	0.021	0.017	0.018	0.015	0.023	0.002	0.007

Table S.2: Results of the simulation of Section S.1 with two independent populations and one disease. The median value of the asymptotic standard error (ASE), empirical standard error (ESE) and the absolute mean of the bias (Bias) of the estimators for $\alpha = (\alpha_1, \ldots, \alpha_{12})^{\mathrm{T}}$ for $n = 1000, 2000, 5000$. Here, $||\alpha||_2 = 1$.

S.2 Simulation Study for Multiple Populations and Diseases

The data are simulated from the logistic model with multiple populations and diseases as described in Section 6. The tables below report the numerical results for simulation studies for $n = 2000$ with independent binary outcomes and the binary responses having correlation 0.05 and 0.1, and $n = 3000$ with independent binary outcomes and the binary responses having correlation 0.10. We observe similar results as given in the tables in Section 6.

	Mean	Estimated se	Actual se	Coverage
Total Fruit	1.03	0.37	0.40	93.40
Whole Fruit	1.05	0.63	0.70	96.30
Total Grains	1.04	0.63	0.70	95.90
Whole Grains	0.97	0.43	0.47	92.60
Total Vegetables	0.99	0.50	0.52	94.70
DOL Vegetables	1.00	0.43	0.46	94.60
Dairy	0.98	0.31	0.32	94.30
Meat and Beans	0.98	0.34	0.37	92.80
Oils	1.00	0.32	0.36	92.50
Sodium	0.99	0.40	0.42	94.70
Saturated Fat	0.99	0.46	0.51	91.00
Empty Calories	0.98	0.53	0.56	94.90

Table S.3: Results for α of the simulation study in Section 6 when $n = 2000$, the binary outcomes are independent, and where the actual of α all = 1.00. Here *Estimate* is the mean of the estimates, *Estimated se* is the mean of the estimated standard errors, *Actual se* is the actual standard deviation of the estimates, and *Coverage* is the actual coverage of a nominal 95% confidence interval. The actual estimates $\widehat{\alpha}$ were normalized to sum to 12.

	Mean	Estimated se	Actual se	Coverage
Total Fruit	1.01	0.39	0.41	94.50
Whole Fruit	1.01	0.66	0.72	96.40
Total Grains	1.08	0.66	0.77	94.20
Whole Grains	0.98	0.45	0.50	92.20
Total Vegetables	1.01	0.52	0.56	94.50
DOL Vegetables	0.98	0.45	0.47	93.70
Dairy	0.98	0.32	0.33	94.80
Meat and Beans	1.00	0.35	0.40	92.00
Oils	0.99	0.33	0.35	94.00
Sodium	0.99	0.42	0.43	94.50
Saturated Fat	0.99	0.47	0.53	91.50
Empty Calories	0.98	0.55	0.59	93.80

Table S.4: Results of the simulation study when $n = 2000$, the binary responses have correlation 0.05 and where the actual of α all = 1.00. Here *Estimate* is the mean of the estimates, *Estimated se* is the mean of the estimated standard errors, *Actual se* is the actual standard deviation of the estimates, and *Coverage* is the actual coverage of a nominal 95% confidence interval. The actual estimates $\widehat{\alpha}$ were normalized to sum to 12.

	Mean	Estimated se	Actual se	Coverage
Total Fruit	1.01	0.40	0.42	94.00
Whole Fruit	1.02	0.68	0.75	95.90
Total Grains	1.09	0.69	0.81	94.50
Whole Grains	0.98	0.46	0.52	91.90
Total Vegetables	1.01	0.54	0.58	95.10
DOL Vegetables	0.98	0.47	0.50	94.20
Dairy	0.98	0.33	0.35	93.40
Meat and Beans	0.99	0.37	0.42	91.80
Oils	0.99	0.35	0.38	93.40
Sodium	0.98	0.43	0.45	94.30
Saturated Fat	0.99	0.49	0.56	92.30
Empty Calories	0.98	0.57	0.63	93.10

Table S.5: Results for α of the simulation study in Section 6 when $n = 2000$, the correlation among the binary outcomes is 0.10, and where the estimates of α all $= 1.00$. Here *Estimate* is the mean of the estimates, *Estimated se* is the mean of the estimated standard errors, *Actual se* is the actual standard deviation of the estimates, and *Coverage* is the actual coverage of a nominal 95% confidence interval. The actual estimates $\widehat{\alpha}$ were normalized to sum to 12.

	Mean	Estimated se	Actual se	Coverage
Total Fruit	0.99	0.31	0.33	94.00
Whole Fruit	0.98	0.53	0.56	96.20
Total Grains	1.04	0.53	0.57	93.70
Whole Grains	1.00	0.36	0.37	95.80
Total Vegetables	1.01	0.42	0.45	93.40
DOL Vegetables	0.98	0.36	0.36	95.20
Dairy	1.01	0.26	0.26	94.70
Meat and Beans	1.00	0.28	0.29	94.70
Oils	1.00	0.27	0.27	96.00
Sodium	0.99	0.33	0.33	94.60
Saturated Fat	1.02	0.38	0.41	93.30
Empty Calories	0.98	0.44	0.45	95.00

Table S.6: Results of the simulation study when $n = 3000$ when the binary outcomes are independent and where the actual of α all $= 1.00$. Here *Estimate* is the mean of the estimates, *Estimated se* is the mean of the estimated standard errors, *Actual se* is the actual standard deviation of the estimates, and *Coverage* is the actual coverage of a nominal 95% confidence interval. The actual estimates $\widehat{\alpha}$ were normalized to sum to 12.

	Mean	Estimated se	Actual se	Coverage
Total Fruit	1.03	0.34	0.35	94.20
Whole Fruit	1.04	0.57	0.62	95.80
Total Grains	1.01	0.57	0.60	95.40
Whole Grains	1.00	0.39	0.40	94.10
Total Vegetables	1.01	0.45	0.47	93.90
DOL Vegetables	0.98	0.39	0.40	94.20
Dairy	1.01	0.28	0.29	95.40
Meat and Beans	1.01	0.30	0.32	93.30
Oils	1.00	0.29	0.30	94.20
Sodium	0.98	0.36	0.38	94.10
Saturated Fat	0.98	0.41	0.44	93.90
Empty Calories	0.96	0.48	0.49	95.60

Table S.7: Results of the simulation study when $n = 3000$,the binary responses have correlation 0.10 and where the actual of α all = 1.00. Here *Estimate* is the mean of the estimates, *Estimated se* is the mean of the estimated standard errors, *Actual se* is the actual standard deviation of the estimates, and *Coverage* is the actual coverage of a nominal 95% confidence interval. The actual estimates $\widehat{\alpha}$ were normalized to sum to 12.

	Mean #	True	Mean	Estimated	Actual	
	Cases	β	$\widehat{\beta}$	se	se	Coverage
Population 1, Disease 1	545	-1.00	-1.00	NA	NA	NA
Population 1, Disease 2	718	-0.60	-0.61	0.15	0.15	95.80
Population 1, Disease 3	898	-0.20	-0.20	0.12	0.12	97.10
Population 2, Disease 1	626	-0.80	-0.81	0.18	0.18	95.90
Population 2, Disease 2	738	-0.57	-0.57	0.16	0.15	94.30
Population 2, Disease 3	837	-0.33	-0.34	0.14	0.14	95.10
Population 2, Disease 4	954	-0.10	-0.12	0.13	0.11	98.10

Table S.8: Results for β of the simulation study in Section 6 when $n = 2000$ and the binary outcomes are independent. Here *True* is the true value, *Estimate* is the mean of the estimates, *Estimated se* is the mean of the estimated standard errors, *Actual se* is the actual standard deviation of the estimates, and *Coverage* is the actual coverage of a nominal 95% confidence interval.

	Mean # Cases	True β	Mean $\widehat{\beta}$	Estimated se	Actual se	Coverage
Population 1, Disease 1	531	-1.00	-1.00	NA	NA	NA
Population 1, Disease 2	721	-0.60	-0.60	0.15	0.15	95.70
Population 1, Disease 3	894	-0.20	-0.21	0.12	0.11	96.50
Population 2, Disease 1	636	-0.80	-0.81	0.18	0.19	94.20
Population 2, Disease 2	732	-0.57	-0.57	0.16	0.16	95.80
Population 2, Disease 3	830	-0.33	-0.34	0.14	0.14	95.70
Population 2, Disease 4	956	-0.10	-0.12	0.13	0.11	97.80

Table S.9: Simulation results for β when $n = 2000$ and the binary responses have correlation 0.05. Here *True* is the true value, *Estimate* is the mean of the estimates, *Estimated se* is the mean of the estimated standard errors, *Actual se* is the actual standard deviation of the estimates, and *Coverage* is the actual coverage of a nominal 95% confidence interval. The mean total number of cases is 5300.

	Mean # Cases	True β	Mean $\widehat{\beta}$	Estimated se	Actual se	Coverage
Population 1, Disease 1	531	-1.00	-1.00	NA	NA	NA
Population 1, Disease 2	720	-0.60	-0.61	0.14	0.14	96.10
Population 1, Disease 3	896	-0.20	-0.21	0.12	0.11	96.50
Population 2, Disease 1	636	-0.80	-0.81	0.18	0.19	94.30
Population 2, Disease 2	731	-0.57	-0.58	0.16	0.16	95.60
Population 2, Disease 3	829	-0.33	-0.35	0.14	0.14	95.10
Population 2, Disease 4	959	-0.10	-0.13	0.13	0.11	97.20

Table S.10: Results for β of the simulation study in Section 6 when $n = 2000$ and the correlations among the binary outcomes = 0.10. Here *True* is the true value, *Estimate* is the mean of the estimates, *Estimated se* is the mean of the estimated standard errors, *Actual se* is the actual standard deviation of the estimates, and *Coverage* is the actual coverage of a nominal 95% confidence interval. The mean number of cases overall is 5302.

	True β	Mean $\widehat{\beta}$	Estimated se	Actual se	Coverage
Population 1, Disease 1	-1.00	-1.00	NA	NA	NA
Population 1, Disease 2	-0.60	-0.61	0.12	0.12	96.00
Population 1, Disease 3	-0.20	-0.21	0.10	0.09	96.10
Population 2, Disease 1	-0.80	-0.80	0.15	0.14	95.30
Population 2, Disease 2	-0.57	-0.56	0.13	0.12	95.90
Population 2, Disease 3	-0.33	-0.33	0.11	0.11	95.90
Population 2, Disease 4	-0.10	-0.11	0.10	0.09	98.50

Table S.11: Simulation results for β when $n = 3000$ and the binary outcomes are independent. Estimate is the mean, Estimated se is the mean of the estimated standard errors, Actual se is the actual standard deviation of the estimates, and Coverage is the actual coverage of a nominal 95% confidence interval.

	True β	Mean $\widehat{\beta}$	Estimated se	Actual se	Coverage
Population 1, Disease 1	-1.00	-1.00	NA	NA	NA
Population 1, Disease 2	-0.60	-0.61	0.12	0.11	96.80
Population 1, Disease 3	-0.20	-0.21	0.10	0.09	97.30
Population 2, Disease 1	-0.80	-0.80	0.14	0.14	94.40
Population 2, Disease 2	-0.57	-0.58	0.12	0.13	95.10
Population 2, Disease 3	-0.33	-0.34	0.11	0.11	95.00
Population 2, Disease 4	-0.10	-0.11	0.10	0.09	97.80

Table S.12: Simulation results for β when $n = 3000$ and the binary responses have correlation 0.10.

	HEI Total Score			Our method, Table 5		
	Estimate	se	pvalue	Estimate	se	pvalue
Male Lung	-1.00	NA	NA	-1.00	NA	NA
Male Colorectal	-0.47	0.05	0.00	-0.39	0.06	0.00
Male Prostate	0.04	0.03	0.15	0.11	0.06	0.07
Female Lung	-0.95	0.06	0.00	-0.91	0.07	0.00
Female Colorectal	-0.33	0.08	0.00	-0.28	0.08	0.00
Female Breast	-0.08	0.05	0.09	-0.07	0.04	0.11
Female Ovarian	0.12	0.15	0.44	0.02	0.14	0.90

Table S.13: Results for β using the original HEI-2005 Total Score, but normalized as in the paper so that the coefficient for Male Lung Cancer = -1. This can be compared to the results of Table 5, repeated in the last 3 columns.