Biometrics 000, 000–000 (2017, to appear) DOI: 000
000 0000

Web-based Supplementary Materials for "Nonparametric Bayesian covariate-adjusted estimation of the Youden index" by Vanda Inácio de Carvalho, Miguel de Carvalho and Adam Branscum

This paper has been submitted for consideration for publication in *Biometrics*

In this supplement to the main paper we present computational and technical details, along with supporting figures and other results. Specifically, the full conditional distributions for the Gibbs sampler used to fit the nonparametric regression model are presented in Section A; we fit the unconditional (i.e., without covariates) nonparametric model in Section 2.1 of the paper using analogous computational methods. Section B contains proofs of Theorems 1 and 2. Details on the nonparametric kernel based method of Zhou and Qin (2015) and additional figures are provided in Section C. Finally, in Section D we give details on a data-driven prior and provide additional figures.

Let D and $\bar{D}$ denote the diseased and nondiseased populations, respectively. For $d \in \{D, \bar{D}\}$, the cumulative distribution functions are denoted $F_d(\cdot)$ in the absence of covariates and $F_d(\cdot|\mathbf{x})$ in the regression model with covariates. Following the notation used in the main paper, we write $\text{YI} = \max_{c\in\mathbb{R}}\{F_{\bar{D}}(c) - F_D(c)\}$ to denote the Youden index induced by the Dirichlet Process Mixture (DPM) model in Section 2.1, and $\text{YI}(\mathbf{x}) = \max_{c\in\mathbb{R}}\{F_{\bar{D}}(c \mid \mathbf{x}) - F_D(c \mid \mathbf{x})\}$ to denote the covariate-dependent Youden index induced by the Dependent Dirichlet Process (DDP) model in Section 2.2. Below, we use the notations $\|\cdot\|_\infty$ and $\|\cdot\|_1$ to denote the sup-norm and the L_1-norm, respectively, so that $\|F_{\bar{D}} - F_D\|_\infty = \sup_{c\in\mathbb{R}} |F_{\bar{D}}(c) - F_D(c)|$ and $\|f_{\bar{D}} - f_D\|_1 = \int_{\mathbb{R}} |f_{\bar{D}}(u) - f_D(u)|\mathrm{d}u$.

For the proof of Theorem 1 in Appendix B we use the following auxiliary lemma often know in mathematical language as the *reverse triangle inequality*.

LEMMA 1: *Let V be a normed vector space. Then, $|\,\|\boldsymbol{a}\| - \|\boldsymbol{b}\|\,| \leqslant \|\boldsymbol{a} - \boldsymbol{b}\|$, for all $\boldsymbol{a}$, $\boldsymbol{b}$ in V.*

Proof. See Christensen (2010, Lemma 2.1.2).

Section A: Blocked Gibbs sampler algorithm

The Gibbs sampling algorithm that we used to fit our nonparametric DDP regression model is essentially a covariate-dependent version of the Blocked Gibbs sampler algorithm in Ishwaran and James (2002), where iterative sampling is conducted using the full conditional distributions catalogued below. We omit the subscripts D and $\bar{D}$ in order to present a general setting that is applicable to both the diseased and nondiseased populations. We describe an algorithm for the non-spline version in which the sampling models for the data from both populations have the form $F(\cdot|\mathbf{x}) = \int \Phi(\cdot|\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}, \sigma^2)\, \mathrm{d}G(\boldsymbol{\beta}, \sigma^2)$, where $G \sim \mathrm{DP}(\alpha, G^*)$ and $G^*(\boldsymbol{\beta}, \sigma^{-2}) = N_p(\boldsymbol{m}_{\boldsymbol{\beta}}, \boldsymbol{S}_{\boldsymbol{\beta}})\Gamma(a, b)$ with $\boldsymbol{m}_{\boldsymbol{\beta}} \sim N_p(\boldsymbol{m}_0, \boldsymbol{S}_0)$ and $\boldsymbol{S}_{\boldsymbol{\beta}}^{-1} \sim \mathrm{Wishart}_p(\nu, (\nu\boldsymbol{\psi})^{-1})$. The truncation

$$G^L(\cdot) = \sum_{\ell=1}^{L} p_\ell \delta_{(\boldsymbol{\beta}_\ell, \sigma_\ell^2)}(\cdot)$$

is used to approximate G (Ishwaran and James, 2001; Ishwaran and Zarepour, 2000, 2002), where L is chosen to be large, e.g., $L = 20$ (Chung and Dunson, 2009). The weights are obtained from the stick-breaking equation

$$p_\ell = q_\ell \prod_{r<\ell}(1 - q_r),$$

where $q_1, \ldots, q_{L-1} \mid \alpha \overset{\text{iid}}{\sim} \mathrm{Beta}(1, \alpha)$ and $q_L = 1$. Upon introducing membership indicators (Diebolt and Robert, 1994) such that $z_i = \ell$ when y_i comes from $N(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}_\ell, \sigma_\ell^2)$, the full conditional distributions are as follows:

$$q_\ell \mid \text{else} \sim \mathrm{Beta}\left(n_\ell + 1, \alpha + \sum_{r=\ell+1}^{L} n_r\right),$$

where $n_\ell = \sum_{i=1}^{n} I(z_i = \ell)$ is the number of observations from component ℓ; in addition, note that $P(z_i = \ell \mid \text{else}) \propto p_\ell \phi(y_i|\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}_\ell, \sigma_\ell^2)$ and

$$\begin{cases} \boldsymbol{\beta}_\ell \mid \text{else} & \sim N_p\left(\boldsymbol{V}_\ell\left(\boldsymbol{S}_{\boldsymbol{\beta}}^{-1}\boldsymbol{m}_{\boldsymbol{\beta}} + \sigma_\ell^{-2}\sum_{\{i:z_i=\ell\}}\mathbf{x}_i y_i\right), \boldsymbol{V}_\ell\right), \quad \boldsymbol{V}_\ell = \left(\boldsymbol{S}_{\boldsymbol{\beta}}^{-1} + \sigma_\ell^{-2}\sum_{\{i:z_i=\ell\}}\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}\right)^{-1}, \\ \sigma_\ell^{-2} \mid \text{else} & \sim \Gamma\left(a + \frac{n_\ell}{2}, b + \frac{1}{2}\sum_{\{i:z_i=\ell\}}(y_i - \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}_\ell)^2\right), \end{cases}$$

where

$$\begin{cases} \boldsymbol{m}_{\boldsymbol{\beta}} \mid \text{else} & \sim N_p\left(\boldsymbol{V}\left(\boldsymbol{S}_0^{-1}\boldsymbol{m}_0 + \boldsymbol{S}_{\boldsymbol{\beta}}^{-1}\sum_{\ell=1}^{L}\boldsymbol{\beta}_\ell\right), \boldsymbol{V}\right), \quad \boldsymbol{V} = (\boldsymbol{S}_0^{-1} + L\boldsymbol{S}_{\boldsymbol{\beta}}^{-1})^{-1}, \\ \boldsymbol{S}_{\boldsymbol{\beta}}^{-1} \mid \text{else} & \sim \text{Wishart}_p\left(\nu + L, \left(\nu\boldsymbol{\psi} + \sum_{\ell=1}^{L}(\boldsymbol{\beta}_\ell - \boldsymbol{m}_{\boldsymbol{\beta}})(\boldsymbol{\beta}_\ell - \boldsymbol{m}_{\boldsymbol{\beta}})^{\mathrm{T}}\right)^{-1}\right), \end{cases}$$

and

$$\alpha \mid \text{else} \sim \text{Gamma}\left(a + L, b - \sum_{\ell=1}^{L-1}\log(1-q_\ell)\right).$$

The full conditional distributions for the spline version follows immediately by replacing $\mathbf{x}^{\mathrm{T}}$ by $\mathbf{z}^{\mathrm{T}} = (B_1(x), \ldots, B_Q(x))$ and by using $\boldsymbol{\beta}_l = (\beta_{l1}, \ldots, \beta_{lQ})^{\mathrm{T}}$.

Section B: Proofs of Theorems 1 and 2

Proof of Theorem 1. The proof entails simple manipulations and a result from Lijoi et al. (2004, §3). Note that

$$\begin{aligned} |\mathrm{YI}_\omega - \mathrm{YI}| &= \left|\max_{c\in\mathbb{R}}\{F_{\bar{D}}^{\omega}(c) - F_D^{\omega}(c)\} - \max_{c\in\mathbb{R}}\{F_{\bar{D}}(c) - F_D(c)\}\right| \\ &= |\,\|F_{\bar{D}}^{\omega} - F_D^{\omega}\|_\infty - \|F_{\bar{D}} - F_D\|_\infty\,| \\ &\leqslant \|(F_{\bar{D}}^{\omega} - F_D^{\omega}) - (F_{\bar{D}} - F_D)\|_\infty \\ &= \|(F_{\bar{D}}^{\omega} - F_{\bar{D}}) - (F_D^{\omega} - F_D)\|_\infty \end{aligned} \tag{1}$$

where the penultimate step follows from the reverse triangle inequality. It follows from (1) that

$$|\mathrm{YI}_\omega - \mathrm{YI}| \leqslant \|F_{\bar{D}}^{\omega} - F_{\bar{D}}\|_\infty + \|F_D^{\omega} - F_D\|_\infty \leqslant \|f_{\bar{D}}^{\omega} - f_{\bar{D}}\|_1 + \|f_D^{\omega} - f_D\|_1, \tag{2}$$

where the last step follows from

$$\begin{aligned} \|F_{\bar{D}}^{\omega} - F_{\bar{D}}\|_\infty = \sup_{c\in\mathbb{R}}\left|\int_{-\infty}^{c}\{f_{\bar{D}}^{\omega}(u) - f_{\bar{D}}(u)\}\,\mathrm{d}u\right| &\leqslant \sup_{c\in\mathbb{R}}\int_{-\infty}^{c}|f_{\bar{D}}^{\omega}(u) - f_{\bar{D}}(u)|\,\mathrm{d}u \\ &= \int_{-\infty}^{\infty}|f_{\bar{D}}^{\omega}(u) - f_{\bar{D}}(u)|\,\mathrm{d}u = \|f_{\bar{D}}^{\omega} - f_{\bar{D}}\|_1, \end{aligned}$$

and analogously $\|F_D^{\omega} - F_D\|_\infty \leqslant \|f_D^{\omega} - f_D\|_1$.

Hence, as it can be noticed from (2), to have $|\mathrm{YI}_\omega - \mathrm{YI}| < \varepsilon$, it would suffice having $\|f_{\bar{D}}^{\omega} - f_{\bar{D}}\|_1 < \varepsilon/2$ and $\|f_D^{\omega} - f_D\|_1 < \varepsilon/2$, thus implying that

$$\{\omega\in\Omega : |\mathrm{YI}_\omega - \mathrm{YI}| < \varepsilon\} \supseteq \{\omega\in\Omega : \|f_{\bar{D}}^{\omega} - f_{\bar{D}}\|_1 < \varepsilon/2, \|f_D^{\omega} - f_D\|_1 < \varepsilon/2\},$$

from where it finally follows that

$$P(\omega \in \Omega : |\mathrm{YI}_\omega - \mathrm{YI}| < \varepsilon) \geqslant P(\omega \in \Omega : \|f_D^\omega - f_{\bar{D}}\|_1 < \varepsilon/2) \times P(\omega \in \Omega : \|f_D^\omega - f_D\|_1 < \varepsilon/2) > 0,$$

for every $\varepsilon > 0$, given that as a consequence of Lijoi et al. (2004, §3), it holds that for every $\varepsilon > 0$,

$$P(\omega \in \Omega : \|f_{\bar{D}}^\omega - f_{\bar{D}}\|_1 < \varepsilon/2) > 0, \quad P(\omega \in \Omega : \|f_D^\omega - f_D\|_1 < \varepsilon/2) > 0.$$

■

Proof of Theorem 2. The proof is analogous to that of Theorem 1, but we need to use Theorem 4 in Barrientos et al. (2012), which is essentially a predictor-dependent version of the results in Lijoi et al. (2004, §3). Similar arguments as in the proof of Theorem 1 yield,

$$|\mathrm{YI}_\omega(\boldsymbol{x}_i) - \mathrm{YI}(\boldsymbol{x}_i)| \leqslant \|f_{\bar{D}}^\omega(\cdot \mid \boldsymbol{x}_i) - f_{\bar{D}}(\cdot \mid \boldsymbol{x}_i)\|_1 + \|f_D^\omega(\cdot \mid \boldsymbol{x}_i) - f_D(\cdot \mid \boldsymbol{x}_i)\|_1, \quad i = 1, \ldots, n.$$

Hence, to have $|\mathrm{YI}_\omega(\boldsymbol{x}_i) - \mathrm{YI}(\boldsymbol{x}_i)| < \varepsilon$, it would suffice having

$$\|f_{\bar{D}}^\omega(\cdot \mid \boldsymbol{x}_i) - f_{\bar{D}}(\cdot \mid \boldsymbol{x}_i)\|_1 < \varepsilon/2, \quad \|f_D^\omega(\cdot \mid \boldsymbol{x}_i) - f_D(\cdot \mid \boldsymbol{x}_i)\|_1 < \varepsilon/2, \quad i = 1, \ldots, n.$$

By similar arguments as in the proof of Theorem 1,

$$\begin{aligned} &P(\omega \in \Omega : |\mathrm{YI}_\omega(\boldsymbol{x}_i) - \mathrm{YI}(\boldsymbol{x}_i)| < \varepsilon,\, i = 1, \ldots, n) \\ &\geqslant P(\omega \in \Omega : \|f_{\bar{D}}^\omega(\cdot \mid \boldsymbol{x}_i) - f_{\bar{D}}(\cdot \mid \boldsymbol{x}_i)\|_1 < \varepsilon/2,\, i = 1, \ldots, n) \\ &\quad \times P(\omega \in \Omega : \|f_D^\omega(\cdot \mid \boldsymbol{x}_i) - f_D(\cdot \mid \boldsymbol{x}_i)\|_1 < \varepsilon/2,\, i = 1, \ldots, n) > 0, \end{aligned}$$

for every $\varepsilon > 0$, given that as a consequence of results on the Hellinger support of the DDP (Barrientos et al., 2012, Theorem 4), it holds that for every $\varepsilon > 0$,

$$\begin{cases} P(\omega \in \Omega : \|f_{\bar{D}}^\omega(\cdot \mid \boldsymbol{x}_i) - f_{\bar{D}}(\cdot \mid \boldsymbol{x}_i)\|_1 < \varepsilon/2,\, i = 1, \ldots, n) > 0, \\ P(\omega \in \Omega : \|f_D^\omega(\cdot \mid \boldsymbol{x}_i) - f_D(\cdot \mid \boldsymbol{x}_i)\|_1 < \varepsilon/2,\, i = 1, \ldots, n) > 0. \end{cases}$$

■

Section C: Simulation study details and figures

In this section we present additional details and supporting figures to the statistical analysis conducted in the simulation study section of the main paper.

Nonparametric kernel method (Zhou and Qin, 2015). This method is based on modeling test outcomes through nonparametric heteroscedastic regression models

$$y_D = \mu_D(x) + \sigma_D(x)\varepsilon_D, \qquad y_{\bar{D}} = \mu_{\bar{D}}(x) + \sigma_{\bar{D}}(x)\varepsilon_{\bar{D}},$$

where μ_D and $\mu_{\bar{D}}$ are the regression functions and σ_D^2 and $\sigma_{\bar{D}}^2$ are the variance functions. Here ε_D and $\varepsilon_{\bar{D}}$ are independent random variables with zero mean, variance one, and distribution functions F_{ε_D} and $F_{\varepsilon_{\bar{D}}}$, respectively. Both the regression and variance functions in each group are estimated using local constant estimators; that is, using Nadaraya–Watson estimators (Fan and Gijbels, 1996, Section 2). The bandwidths involved in the computation of these estimators were selected sequentially and by expected Kullback–Leibler cross-validation (Hurvich et al., 1998) as implemented in the `R` function `npregbw` from the `np` package. Once we have the estimates $\widehat{\mu}_d(x)$ and $\widehat{\sigma}_d^2(x)$, $d \in \{D, \bar{D}\}$, we can estimate the standardized residuals

$$\widehat{\varepsilon}_{Di} = \frac{y_{Di} - \widehat{\mu}_D(x_{Di})}{\widehat{\sigma}_D(x_{Di})}, \qquad \widehat{\varepsilon}_{\bar{D}j} = \frac{y_{\bar{D}j} - \widehat{\mu}_{\bar{D}}(x_{\bar{D}j})}{\widehat{\sigma}_{\bar{D}}(x_{\bar{D}j})}, \quad i = 1, \ldots, n_D, \quad j = 1, \ldots, n_{\bar{D}}.$$

Denoting by $\widehat{F}_{\widehat{\varepsilon}_D}$ and $\widehat{F}_{\widehat{\varepsilon}_{\bar{D}}}$, the empirical distribution functions of $\hat{\varepsilon}_D$ and $\widehat{\varepsilon}_{\bar{D}}$, respectively, we have

$$\widehat{\mathrm{YI}}_{\mathrm{kernel}}(x) = \max_{c \in \mathbb{R}} \left\{ \widehat{F}_{\widehat{\varepsilon}_{\bar{D}}}\left(\frac{c - \widehat{\mu}_{\bar{D}}(x)}{\widehat{\sigma}_{\bar{D}}(x)} \right) - \widehat{F}_{\widehat{\varepsilon}_D}\left(\frac{c - \widehat{\mu}_D(x)}{\widehat{\sigma}_D(x)} \right) \right\},$$
$$\hat{c}^*_{\mathrm{kernel}}(x) = \mathrm{argmax}_{c \in \mathbb{R}} \left\{ \widehat{F}_{\widehat{\varepsilon}_{\bar{D}}}\left(\frac{c - \widehat{\mu}_{\bar{D}}(x)}{\widehat{\sigma}_{\bar{D}}(x)} \right) - \widehat{F}_{\widehat{\varepsilon}_D}\left(\frac{c - \widehat{\mu}_D(x)}{\widehat{\sigma}_D(x)} \right) \right\}.$$

To obtain pointwise confidence bands, a bootstrap of the residuals (see, for instance, Gonzalez-Manteiga et al., 2011, Section 5) is employed.

[Figure 1 about here.]

[Figure 2 about here.]

[Figure 3 about here.]

[Figure 4 about here.]

[Figure 5 about here.]

[Figure 6 about here.]

[Figure 7 about here.]

[Figure 8 about here.]

[Figure 9 about here.]

[Figure 10 about here.]

[Table 1 about here.]

Section D: Application—Data-driven prior and supplementary figures

A sensitivity analysis was conducted using the following prior specification suggested by Inácio de Carvalho et al. (2013) and Zhou et al. (2015)

$$a_d = 3, \quad b_d = \widehat{\sigma}_d^2,$$

$$\mathbf{m}_{d0} = (\mathbf{z}_d'\mathbf{z}_d)^{-1}\mathbf{z}_d'\mathbf{y}_d, \quad \mathbf{S}_{d0} = \widehat{\sigma}_d^2(\mathbf{z}_d'\mathbf{z}_d)^{-1},$$

$$\nu_d = Q + 2, \quad \Psi_d = 30\mathbf{S}_{d0},$$

with $\widehat{\sigma}_d^2 = \|\mathbf{y}_d - \mathbf{z}_d\mathbf{m}_{d0}\|^2/(n_d - Q - 1)$. The results are presented in Figures 15 and 16 and in Table 2.

[Figure 11 about here.]

[Figure 12 about here.]

[Figure 13 about here.]

[Figure 14 about here.]

[Figure 15 about here.]

[Table 2 about here.]

Section E: `R` code for implementing our methods

Below, we discuss some `R` code for illustrating how to implement the nonparametric Bayes estimator for the covariate-adjusted youden index and corresponding optimal cutoff. Before running the code chunks below, start by cleaning workspace and install the following packages (if not installed).

```
rm(list = ls())
if (!require("splines")) install.packages("splines")
if (!require("Hmisc")) install.packages("Hmisc")
if (!require("MASS")) install.packages("MASS")
```

For reproducibility reasons, this pdf file has been prepared using `knitr` (Xie, 2015); we fix `setseed` and list below the information about `R`, the OS, and loaded packages:

```
set.seed(1)
sessionInfo()
```

```
## R version 3.3.2 (2016-10-31)
## Platform: x86_64-apple-darwin13.4.0 (64-bit)
## Running under: macOS Sierra 10.12.1
##
## locale:
## [1] en_US.UTF-8/C/en_US.UTF-8/C/C/en_US.UTF-8
##
## attached base packages:
## [1] splines   stats     graphics  grDevices utils     datasets  base
##
## other attached packages:
## [1] MASS_7.3-45     Hmisc_3.17-4    ggplot2_2.1.0   Formula_1.2-1
## [5] survival_2.39-5 lattice_0.20-34 knitr_1.15
##
## loaded via a namespace (and not attached):
##  [1] Rcpp_0.12.5         cluster_2.0.5       magrittr_1.5
##  [4] munsell_0.4.3       colorspace_1.2-6    highr_0.6
##  [7] stringr_1.0.0       plyr_1.8.4          tools_3.3.2
```

```
## [10] nnet_7.3-12          grid_3.3.2           data.table_1.9.6
## [13] gtable_0.2.0         latticeExtra_0.6-28  Matrix_1.2-7.1
## [16] gridExtra_2.2.1      RColorBrewer_1.1-2   acepack_1.3-3.3
## [19] rpart_4.1-10         evaluate_0.10        stringi_1.1.1
## [22] methods_3.3.2        scales_0.4.0         chron_2.3-47
## [25] foreign_0.8-67
```

In the code chunks below, we follow the 80 characters per line standard. The key function for fitting the covariate-adjusted Youden index is:

```
bsplinesddp <- function(y, x, grid, xpred, m, S, nu, psi, atau, btau,
                        alpha, L, nsim, knots) {
  yt <- y / sd(y)
  n <- length(y)
  ngrid <- length(grid)
  npred <- length(xpred)
  X <- bs(x, degree = 3, knots = knots, intercept = TRUE)
  k <- ncol(X)
  Xpred <- predict(bs(x, degree = 3, knots = knots, intercept = TRUE),
                   xpred)

  p <- ns <- rep(0, L)
  v <- rep(1 / L, L)
  v[L] <- 1

  beta <- matrix(0, nrow = L, ncol = k)
  tau <- rep(1 / var(yt), L)

  prop <- prob <- matrix(0, nrow = n, ncol = L)

  P <- Tau <- Sigma2 <- matrix(0, nrow = nsim, ncol = L)
  Beta <- Beta1 <- array(0, c(nsim, L, k))
  Beta[1, , ] <- beta
  Tau[1, ] <- tau

  mu <- matrix(0, nrow = nsim, ncol = k)
  Sigmainv <- array(0, c(nsim, k, k))
  mu[1, ] <- mvrnorm(1, mu = m, Sigma = S)
  Sigmainv[1, , ] <- rWishart(1, df = nu, solve(nu * psi))

  Dens <- array(0, c(nsim, ngrid, L, npred))
  Densm <- array(0, c(nsim, ngrid, npred))
  Fdist <- array(0, c(nsim, ngrid, L, npred))
```

```
Fdistm <- array(0, c(nsim, ngrid, npred))

## 1) ALLOCATE EACH OBSERVATION TO A COMPONENT MIXTURE
for(i in 2:nsim) {
  cumv <- cumprod(1 - v)

  p[1] <- v[1]
  for(l in 2:L)
    p[l] <- v[l] * cumv[l - 1]
  for(l in 1:L)
    prop[, l] <- p[l] * dnorm(yt, mean = X %*% beta[l, ],
                              sd = sqrt(1 / tau[l]))
  prob <- prop / apply(prop, 1, sum)
  z <- rMultinom(prob, 1)
  P[i, ] <- p

  for(l in 1:L)
    ns[l] <- length(which(z == l))

  ## 2) UPDATE STICK-BREAKING WEIGHTS
  for(l in 1:(L - 1))
    v[l] <- rbeta(1, 1 + ns[l], alpha + sum(ns[(l + 1):L]))

  ## 3) UPDATE PARAMETERS OF EACH COMPONENT MIXTURE
  for(l in 1:L) {
    tX <- matrix(t(X[z == l, ]), nrow = k, ncol = ns[l])
    V <- solve(Sigmainv[i - 1, ,] + tau[l] * tX %*% X[z == l, ])
    mu1 <- V %*% (Sigmainv[i - 1, , ] %*% mu[i - 1, ] + tau[l] *
                  tX %*% yt[z == l])
    Beta[i, l, ] <- beta[l,] <- mvrnorm(1, mu = mu1, Sigma = V)
    Beta1[i, l, ] <- sd(y) * Beta[i, l, ]

    Tau[i, l] <- tau[l] <- rgamma(1, shape = atau + (ns[l] / 2),
                                  rate = btau + 0.5 * (t(yt[z == l] -
                                  X[z == l, ] %*% beta[l, ]) %*%
                                  (yt[z == l] - X[z == l, ]
                                  %*% beta[l, ])))
    Sigma2[i, l] <- var(y) * (1 / Tau[i, l])
  }

  Vaux <- solve(solve(S) + L * Sigmainv[i - 1, , ])
  meanmu <- Vaux %*% (solve(S) %*% m + Sigmainv[i - 1, , ] %*%
                    t(t(apply(Beta[i, , ], 2, sum))))
  mu[i, ] <- mvrnorm(1, mu = meanmu, Sigma = Vaux)
```

```
    Vaux1 <- 0
    for(l in 1:L)
      Vaux1 <- Vaux1 + (Beta[i, l, ] - mu[i, ]) %*%
               t((Beta[i, l, ] - mu[i, ]))
    Sigmainv[i, , ] <- rWishart(1, nu + L, solve(nu * psi + Vaux1))

    ## 4) COMPUTE DENSITY AND DISTRIBUTION FUNCTION TRAJECTORIES
    for(l in 1:L) {
      for(j in 1:npred) {
        Dens[i, ,l , j] <- P[i, l] * dnorm(grid, Xpred[j, ] %*%
                                           Beta1[i, l, ],
                                           sqrt(Sigma2[i, l]))
        Fdist[i, , l, j] <- P[i, l] * pnorm(grid, Xpred[j, ] %*%
                                              Beta1[i, l, ],
                                              sqrt(Sigma2[i, l]))
      }
    }

    for(j in 1:ngrid) {
      for(l in 1:npred) {
        Densm[i, j, l] <- sum(Dens[i, j, , l])
        Fdistm[i, j, l] <- sum(Fdist[i, j, , l])
      }
    }
  }
  return(list(P, Beta1, Sigma2, Densm, Fdistm))
}
```

The arguments of the `bsplinesddp` function are self-explanatory from the article. To analyze the diabetes data we proceed as follows:

```
## setwd("Add your working directory here")
load("diabetes.Rdata")
ind0 <- which(diabetes[, 2] == 0)
ind1 <- which(diabetes[, 2] == 1)
n0 <- length(ind0)
n1 <- length(ind1)
y0 <- diabetes[ind0, 1]
y1 <- diabetes[ind1, 1]
x0 <- diabetes[ind0, 3]
x1 <- diabetes[ind1, 3]
var0 <- var1 <- 1
```

```
knots0 <- c()
knots1 <- c()
nk <- length(knots0)
```

Next, we apply the workhorse function `bsplinesddp` to the pair (glucose levels, age) for diseased (`x1`, `y1`) and nondiseased subjects (`x0`, `y0`):

```
res0np <- bsplinesddp(y = y0, x = x0, grid = seq(50, 500, len = 200),
                      xpred = seq(32, 76, by = 2), m = rep(0, nk + 4),
                      S = 100 * diag(nk + 4), nu = nk + 6,
                      psi = diag(nk + 4), atau = 0.1, btau = 0.1, alpha = 1,
                      L = 20, nsim = 5000, knots = knots0)
res1np <- bsplinesddp(y = y1, x = x1, grid = seq(50, 500, len = 200),
                      xpred = seq(32, 76, by = 2), m = rep(0, nk + 4),
                      S = 100 * diag(nk + 4), nu = nk + 6,
                      psi = diag(nk + 4), atau = 0.1, btau = 0.1, alpha = 1,
                      L = 20, nsim = 5000, knots = knots1)
```

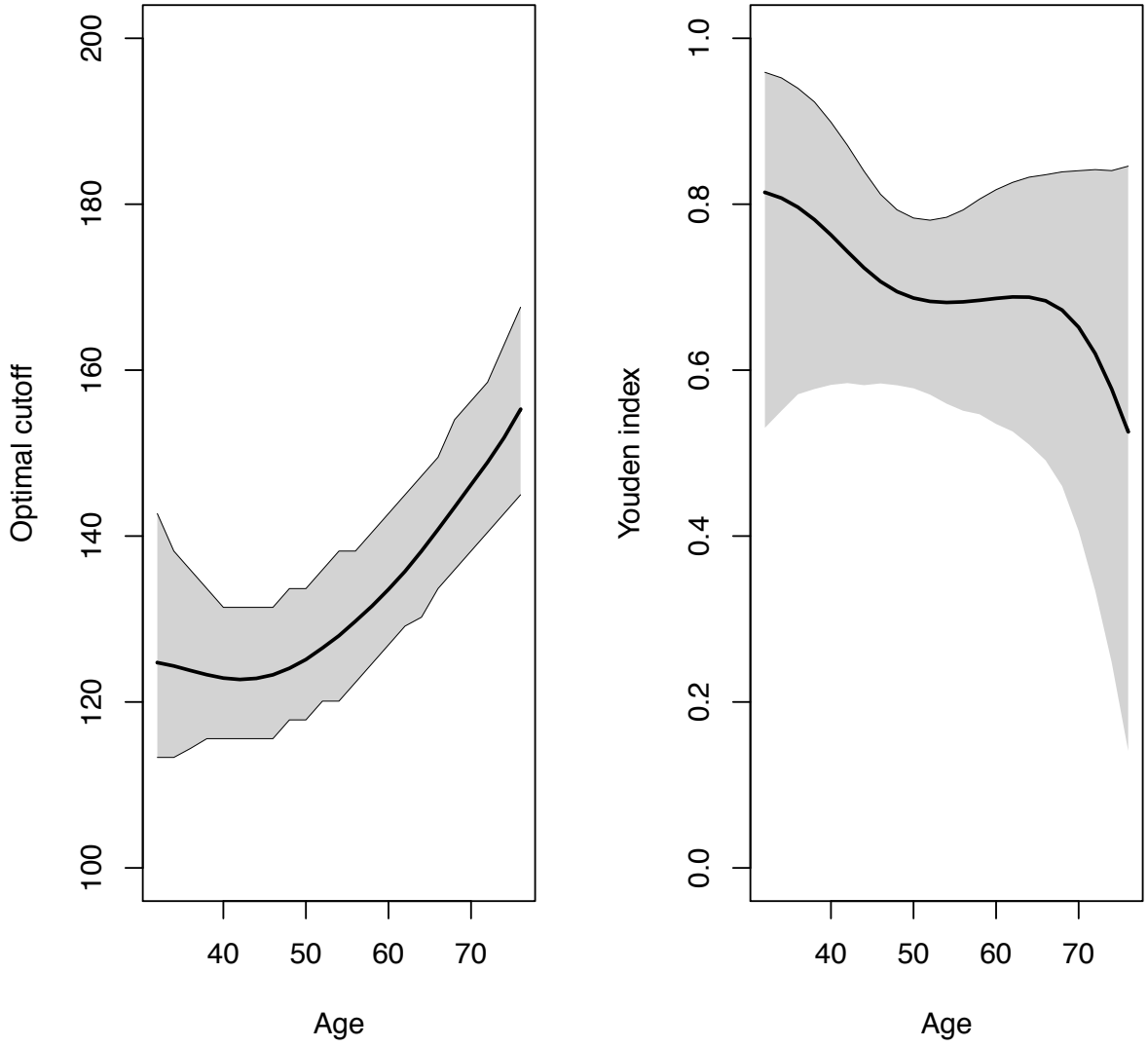
We compute the covariate-adjusted optimal cutoff and corresponding Youden index as follows:

```
grid <- seq(50, 500, len = 200)
xpred <- seq(32, 76, by = 2)
ngrid <- length(grid)
npred <- length(xpred)
nsim <- 5000
nburn <- 1500

difcnp <- array(0, c(nsim - nburn, ngrid, npred))
for(k in 1:npred)
  for(j in 1:ngrid)
  difcnp[, j, k] <- res0np[[5]][(nburn + 1):nsim, j, k] -
                    res1np[[5]][(nburn + 1):nsim, j, k]

coptcnp <- matrix(0, nrow = nsim - nburn, ncol = npred)
for(k in 1:npred)
  for(j in 1:(nsim - nburn))
    coptcnp[j, k] = mean(grid[which(difcnp[j, , k] == max(difcnp[j, , k]))])

coptcrnp <- matrix(nrow = npred, ncol = 3)
for(j in 1:npred) {
```

```
  coptcrnp[j, 1] <- quantile(coptcnp[, j], 0.025)
  coptcrnp[j, 2] <- mean(coptcnp[, j])
  coptcrnp[j, 3] <- quantile(coptcnp[, j], 0.975)
}

yicnp <- matrix(0, nrow = nsim - nburn, ncol = npred)
for(k in 1:npred)
  for(j in 1:(nsim - nburn))
    yicnp[j,k] <- max(difcnp[j, , k])

yicrnp <- matrix(nrow = npred,ncol = 3)
for(j in 1:npred) {
  yicrnp[j, 1] <- quantile(yicnp[, j], 0.025)
  yicrnp[j, 2] <- mean(yicnp[, j])
  yicrnp[j, 3] <- quantile(yicnp[, j], 0.975)
}
```

The first column of Figure 3 in the paper can then be produced using the following lines of code.

```
par(mfrow = c(1, 2))
plot(xpred, coptcrnp[, 2], type = "l", xlim = c(32, 76), ylim = c(100, 200),
     xlab = "Age", ylab = "Optimal cutoff", lwd = 2)
polygon(x = c(rev(xpred), xpred), y = c(rev(coptcrnp[, 1]), coptcrnp[, 3]),
        border = NA, col = "lightgray")
lines(xpred, coptcrnp[, 1], lwd = 0.3)
lines(xpred, coptcrnp[, 2], lwd = 2)
lines(xpred, coptcrnp[, 3], lwd = 0.3)

plot(xpred, yicrnp[, 2], type = "l", xlim = c(32, 76), ylim = c(0, 1),
     xlab = "Age", ylab = "Youden index", lwd = 2)
polygon(x = c(rev(xpred), xpred), y = c(rev(yicrnp[, 1]), yicrnp[, 3]),
        border = NA, col = "lightgray")
lines(xpred, coptcrnp[, 1], lwd = 0.3)
lines(xpred, yicrnp[, 2], lwd = 2)
lines(xpred, yicrnp[, 3], lwd = 0.3)
```



The Log PseudoMarginal Likelihood (LPML) was computed as follows:

```
L <- 20
X0 <- bs(x0, degree = 3, knots = knots0, intercept = TRUE)

term0 <- array(0, c(nsim - nburn, L, n0))
for(i in 1:n0)
  for(k in 1:(nsim-nburn))
    for(l in 1:L)
      term0[k, l, i] <- res0np[[1]][k + nburn,l] *
                        dnorm(y0[i], mean = X0[i,] %*%
```

```
                                res0np[[2]][k + nburn, 1,],
                                sd = sqrt(res0np[[3]][k + nburn, 1]))

termsum0 <- matrix(0, nrow = nsim - nburn, ncol = n0)
for(i in 1:n0)
  for(k in 1:(nsim-nburn))
    termsum0[k, i] <- sum(term0[k, , i])

cpoinv0 <- numeric(n0)
for(i in 1:n0)
  cpoinv0[i] <- mean(1 / termsum0[, i])

cpo0 <- 1 / cpoinv0
lpml0 <- sum(log(cpo0))
```

And we thus get that the value of `lpml0` is −878.18.

References

Barrientos, A. F., Jara, A., and Quintana, F. (2012). On the support of MacEachern's dependent Dirichlet processes and extensions. *Bayesian Analysis* **7**, 277–310.

Christensen, O. (2010). *Functions, Spaces, and Expansions: Mathematical Tools in Physics and Engineering*. Berlin: Birkhäuser.

Chung, Y., and Dunson, D. B. (2009). Nonparametric Bayes conditional distribution modeling with variable selection. *Journal of the American Statistical Association* **104**, 1646–1660.

Diebolt, J., and Robert, C. P. (1994). Estimate of finite mixture distributions through Bayesian sampling. *Journal of the Royal Statistical Society*, Ser. B **56**, 363–375.

Fan, J., and Gijbels, I. (1996). *Local Polynomial Modelling and its Applications*. Chapman & Hall, London.

Gonzalez-Manteiga, W., Pardo-Fernandez, J. C., and Van Keilegom, I. (2011). ROC curves in non-parametric location-scale regression models. *Scandinavian Journal of Statistics* **38** 169–184.

Hurvich, C. M., Simonoff, J. S. and Tsai, C. L. (1998). Smoothing parameter selection in nonparametric regression using an improved Akaike information criterion. *Journal of the Royal Statistical Society*, Ser. B **60**, 271–293.

Inácio de Carvalho, V., Jara, A., Hanson, T. E., and de Carvalho, M. (2013). Bayesian nonparametric ROC regression modeling. *Bayesian Analysis* **8**, 623–646.

Ishwaran, I., and James, L. F. (2001). Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association* **96,** 161–173.

Ishwaran, H., and James, L. F. (2002). Approximate Dirichlet process computing in finite normal mixtures: smoothing and prior information. *Journal of Computational and Graphical Statistics* **11**, 508–532.

Ishwaran, I., and Zarepour, M. (2000). Markov chain Monte Carlo in approximate Dirichlet and Beta two-parameter process hierarchical models. *Biometrika* **87**, 371–390.

Ishwaran, I., and Zarepour, M. (2002). Exact and approximate sum representation for the Dirichlet process. *Canadian Journal of Statistics*, **30**, 269–283.

Lijoi, A., Prünster, I., and Walker, S. G. (2004). Extending Doob's consistency theorem to nonparametric densities. *Bernoulli* **10**, 651–663.

Xie, Y. (2015). *Dynamic Documents with R and knitr*. Boca Raton, FL: Chapman & Hall/CRC.

Zhou, H., and Qin, G. (2015). Nonparametric covariate adjustment for the Youden index. In *Applied Statistics in Biomedicine and Clinical Trials Design*. Cham: Springer.

Zhou, A., Hanson, T., and Knapp, R. (2015). Marginal Bayesian nonparametric model for time to disease arrival of threatened amphibian populations. *Biometrics* **71**, 1101–1110.

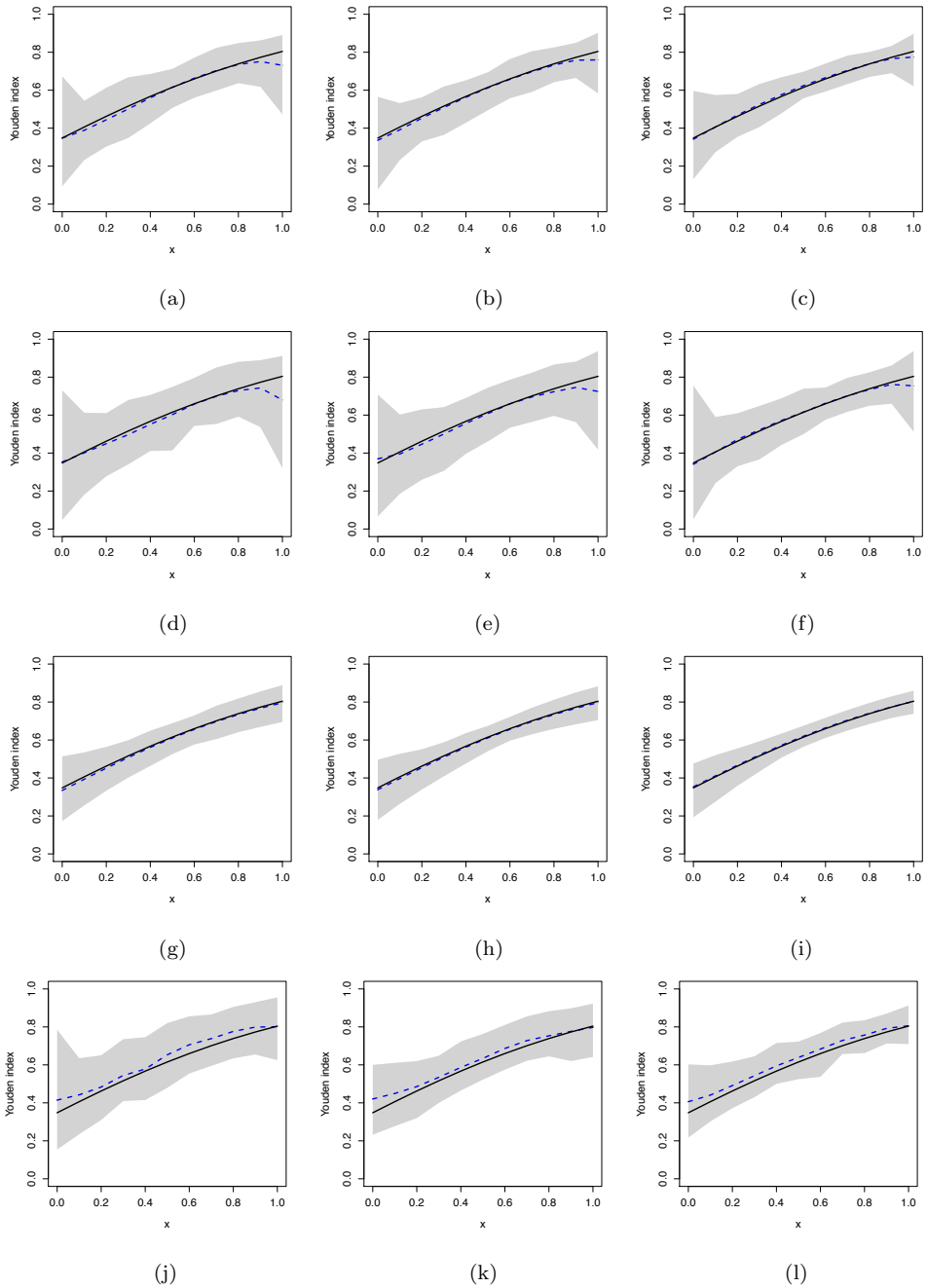


Figure 1. True (solid black line) and mean across simulations (dashed blue line) of the posterior mean (for the Bayesian estimators) of the Youden index function under Scenario 1. A band constructed using the pointwise 2.5% and 97.5% quantiles across simulations is presented in gray. Row 1: B-splines DDP estimator with $Q = 4$. Row 2: B-splines DDP estimator with $Q = 7$. Row 3: Normal estimator. Row 4: Kernel estimator. Panels (a), (d), (g), and (j) show the results for $(n_D, n_{\bar{D}}) = (100, 100)$, panels (b), (e), (h), and (k) for $(n_D, n_{\bar{D}}) = (100, 200)$, and panels (c), (f), (i), and (l) for $(n_D, n_{\bar{D}}) = (200, 200)$.

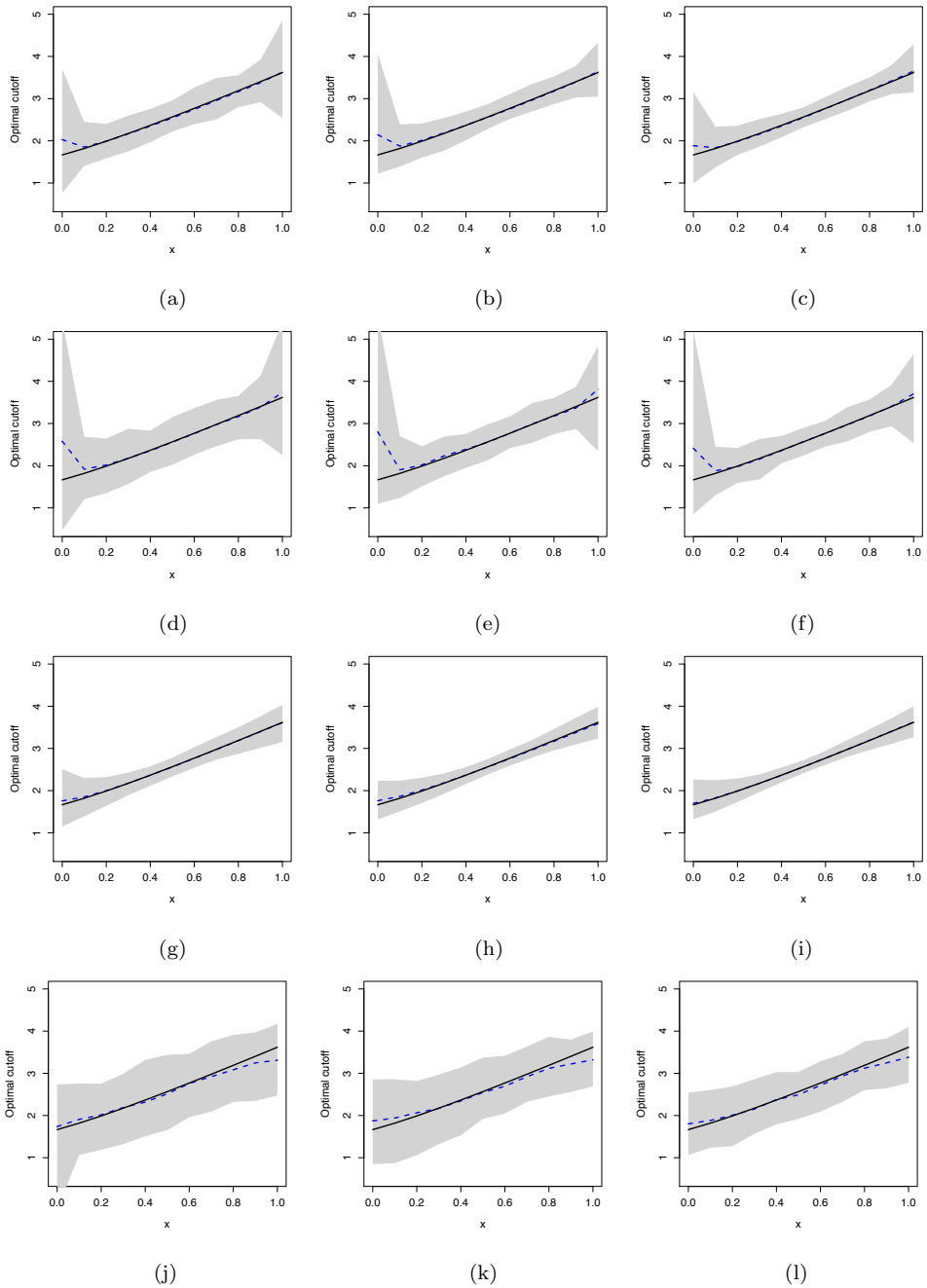


Figure 2. True (solid black line) and mean across simulations (dashed blue line) of the posterior mean (for the Bayesian estimators) of the optimal cutoff function under Scenario 1. A band constructed using the pointwise 2.5% and 97.5% quantiles across simulations is presented in gray. Row 1: B-splines DDP estimator with $Q = 4$. Row 2: B-splines DDP estimator with $Q = 7$. Row 3: Normal estimator. Row 4: Kernel estimator. Panels (a), (d), (g), and (j) show the results for $(n_D, n_{\bar{D}}) = (100, 100)$, panels (b), (e), (h), and (k) for $(n_D, n_{\bar{D}}) = (100, 200)$, and panels (c), (f), (i), and (l) for $(n_D, n_{\bar{D}}) = (200, 200)$.

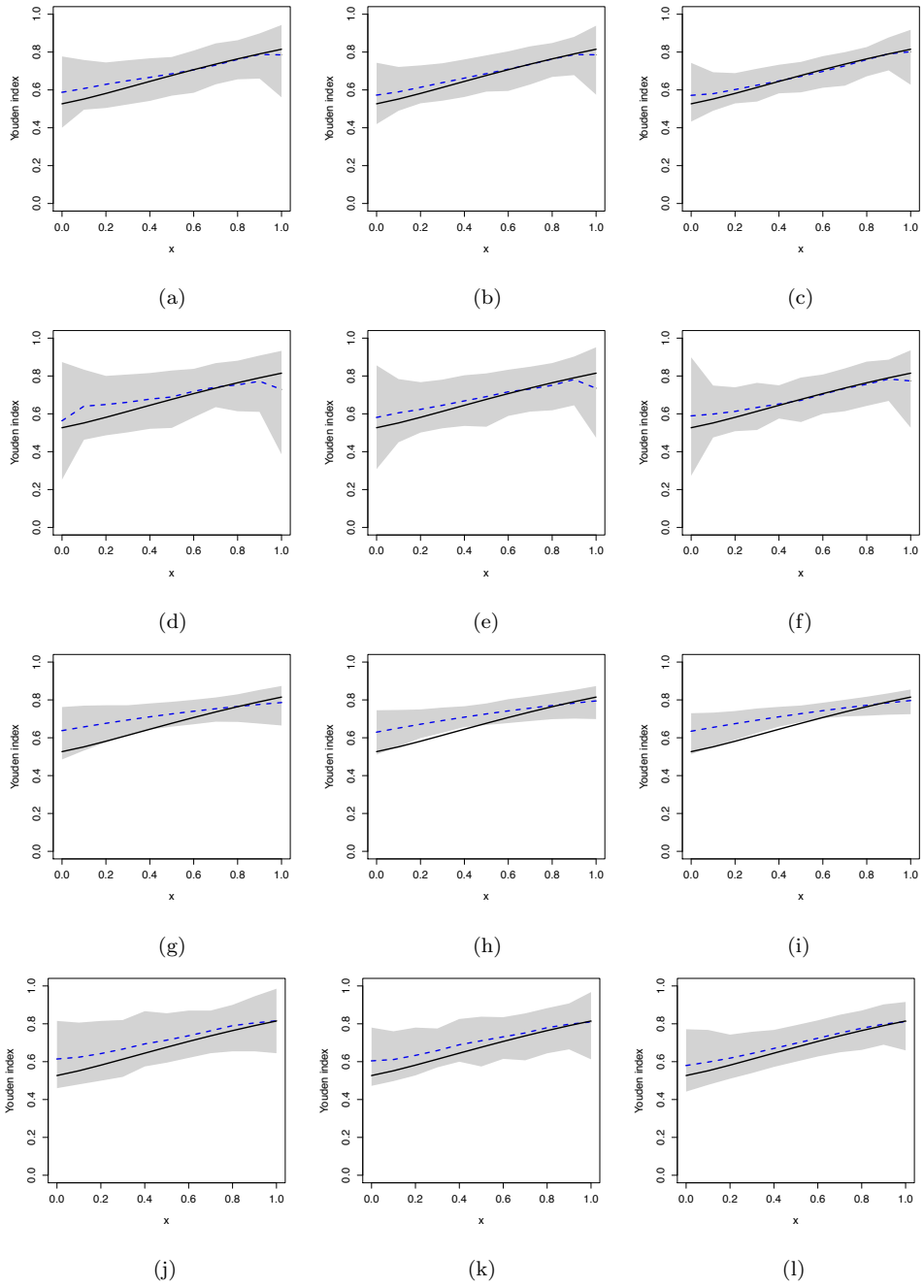


Figure 3. True (solid black line) and mean across simulations (dashed blue line) of the posterior mean (for the Bayesian estimators) of the Youden index function under Scenario 2. A band constructed using the pointwise 2.5% and 97.5% quantiles across simulations is presented in gray. Row 1: B-splines DDP estimator with $Q = 4$. Row 2: B-splines DDP estimator with $Q = 7$. Row 3: Normal estimator. Row 4: Kernel estimator. Panels (a), (d), (g), and (j) show the results for $(n_D, n_{\bar{D}}) = (100, 100)$, panels (b), (e), (h), and (k) for $(n_D, n_{\bar{D}}) = (100, 200)$, and panels (c), (f), (i), and (l) for $(n_D, n_{\bar{D}}) = (200, 200)$.

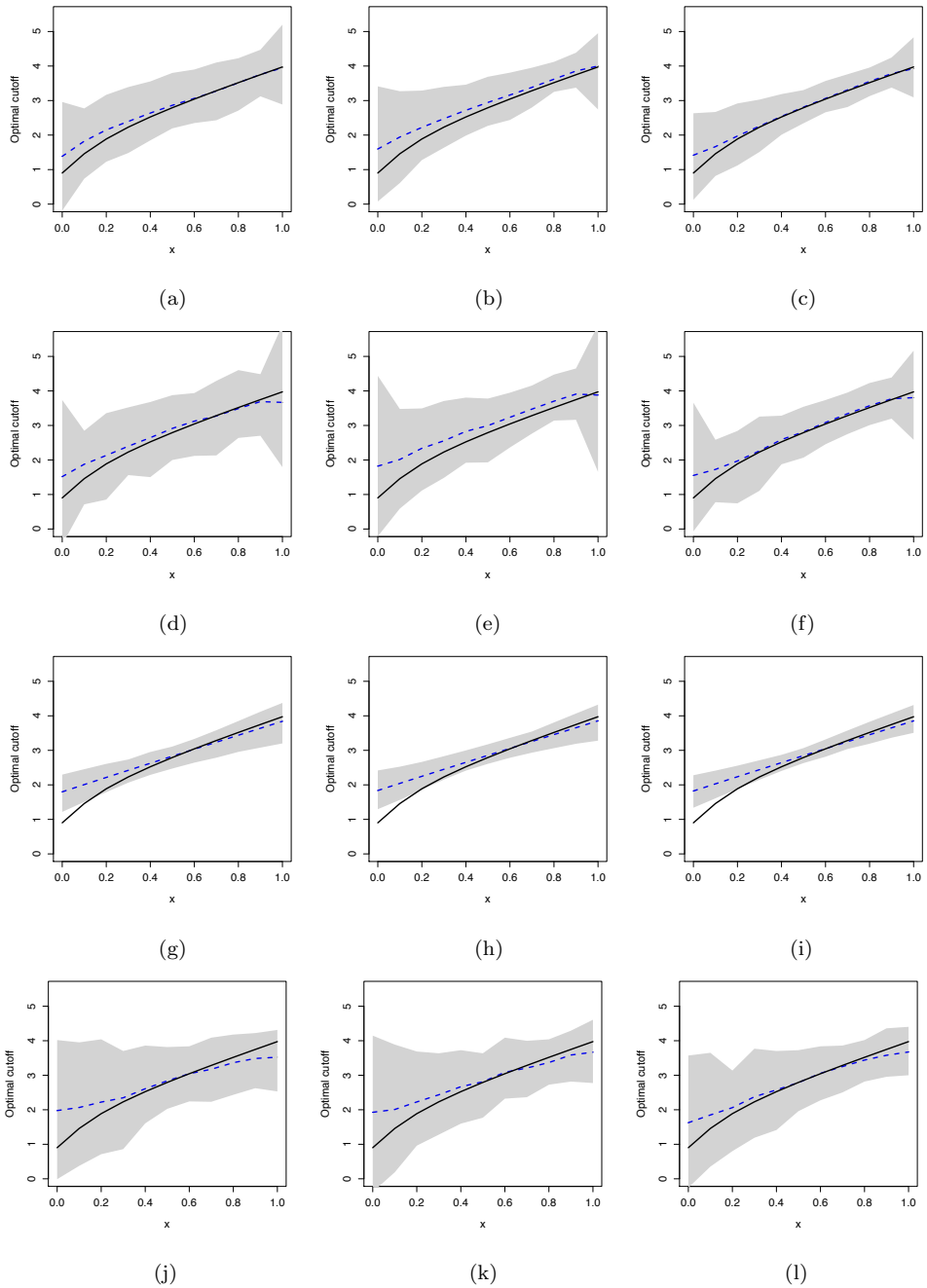


Figure 4. True (solid black line) and mean across simulations (dashed blue line) of the posterior mean (for the Bayesian estimators) of the optimal cutoff function under Scenario 2. A band constructed using the pointwise 2.5% and 97.5% quantiles across simulations is presented in gray. Row 1: B-splines DDP estimator with $Q = 4$. Row 2: B-splines DDP estimator with $Q = 7$. Row 3: Normal estimator. Row 4: Kernel estimator. Panels (a), (d), (g), and (j) show the results for $(n_D, n_{\bar{D}}) = (100, 100)$, panels (b), (e), (h), and (k) for $(n_D, n_{\bar{D}}) = (100, 200)$, and panels (c), (f), (i), and (l) for $(n_D, n_{\bar{D}}) = (200, 200)$.

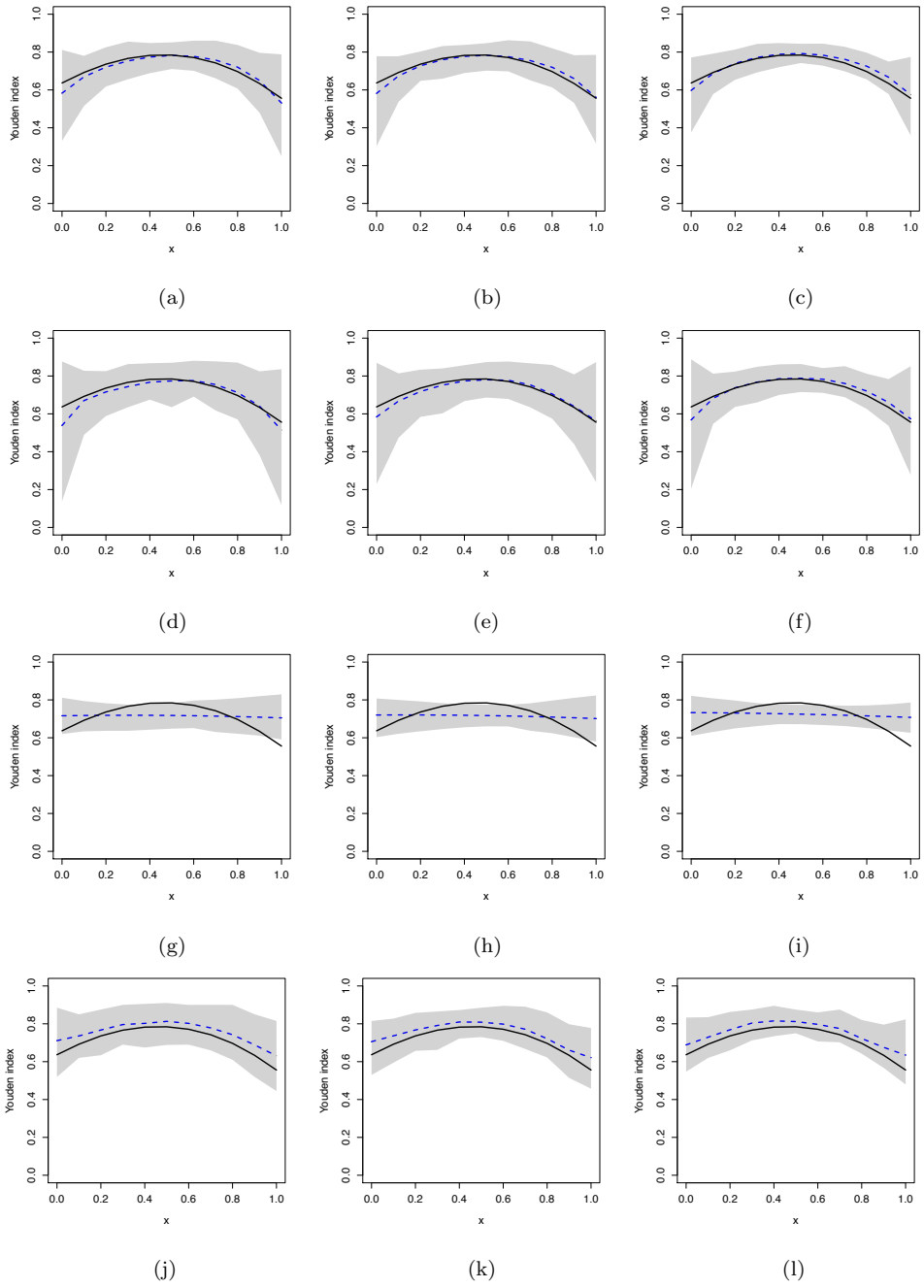


Figure 5. True (solid black line) and mean across simulations (dashed blue line) of the posterior mean (for the Bayesian estimators) of the Youden index function under Scenario 3. A band constructed using the pointwise 2.5% and 97.5% quantiles across simulations is presented in gray. Row 1: B-splines DDP estimator with $Q = 4$. Row 2: B-splines DDP estimator with $Q = 7$. Row 3: Normal estimator. Row 4: Kernel estimator. Panels (a), (d), (g), and (j) show the results for $(n_D, n_{\bar{D}}) = (100, 100)$, panels (b), (e), (h), and (k) for $(n_D, n_{\bar{D}}) = (100, 200)$, and panels (c), (f), (i), and (l) for $(n_D, n_{\bar{D}}) = (200, 200)$.

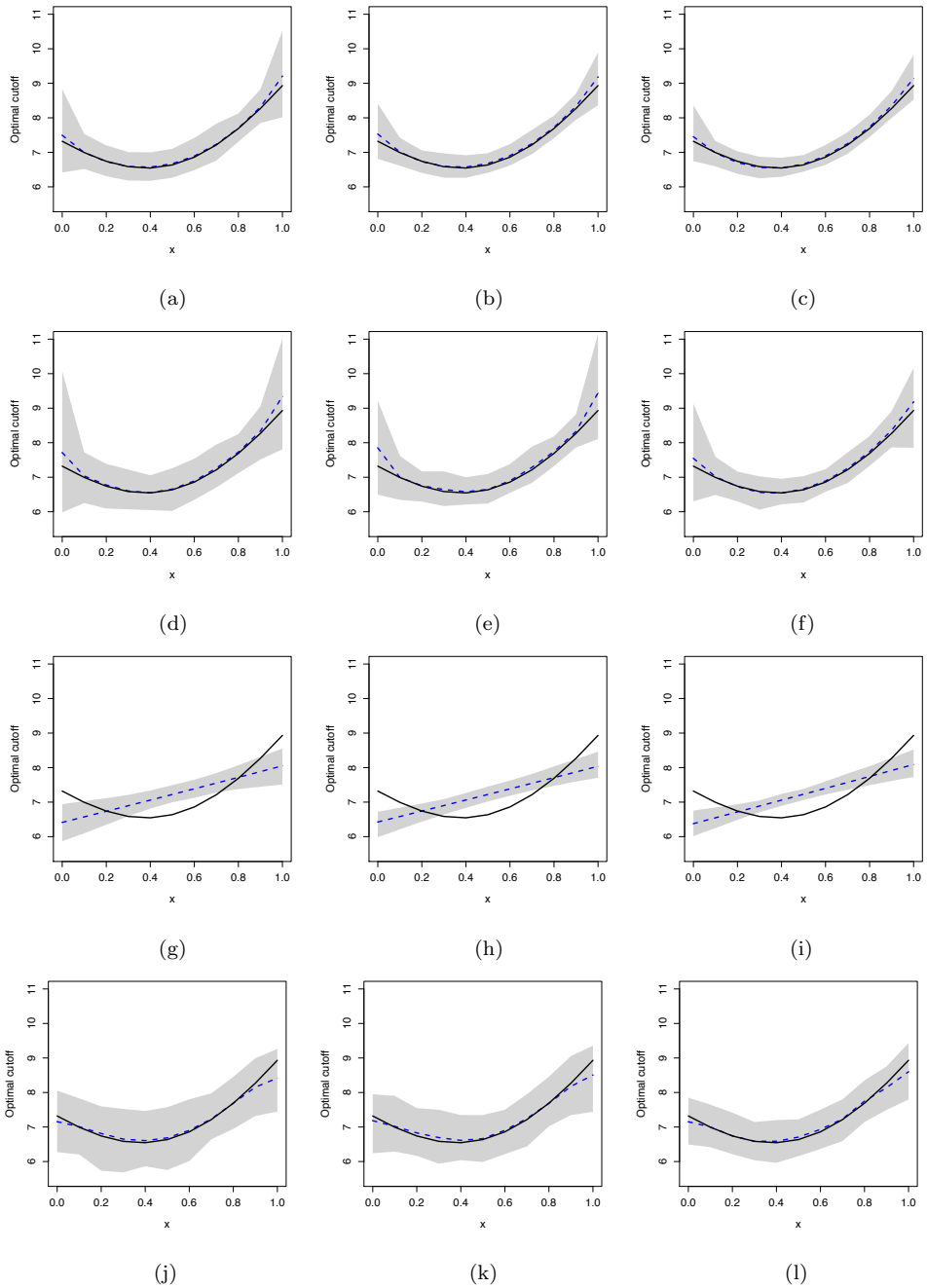


Figure 6. True (solid black line) and mean across simulations (dashed blue line) of the posterior mean (for the Bayesian estimators) of the optimal cutoff function under Scenario 3. A band constructed using the pointwise 2.5% and 97.5% quantiles across simulations is presented in gray. Row 1: B-splines DDP estimator with $Q = 4$. Row 2: B-splines DDP estimator with $Q = 7$. Row 3: Normal estimator. Row 4: Kernel estimator.Panels (a), (d), (g), and (j) show the results for $(n_D, n_{\bar{D}}) = (100, 100)$, panels (b), (e), (h), and (k) for $(n_D, n_{\bar{D}}) = (100, 200)$, and panels (c), (f), (i), and (l) for $(n_D, n_{\bar{D}}) = (200, 200)$.

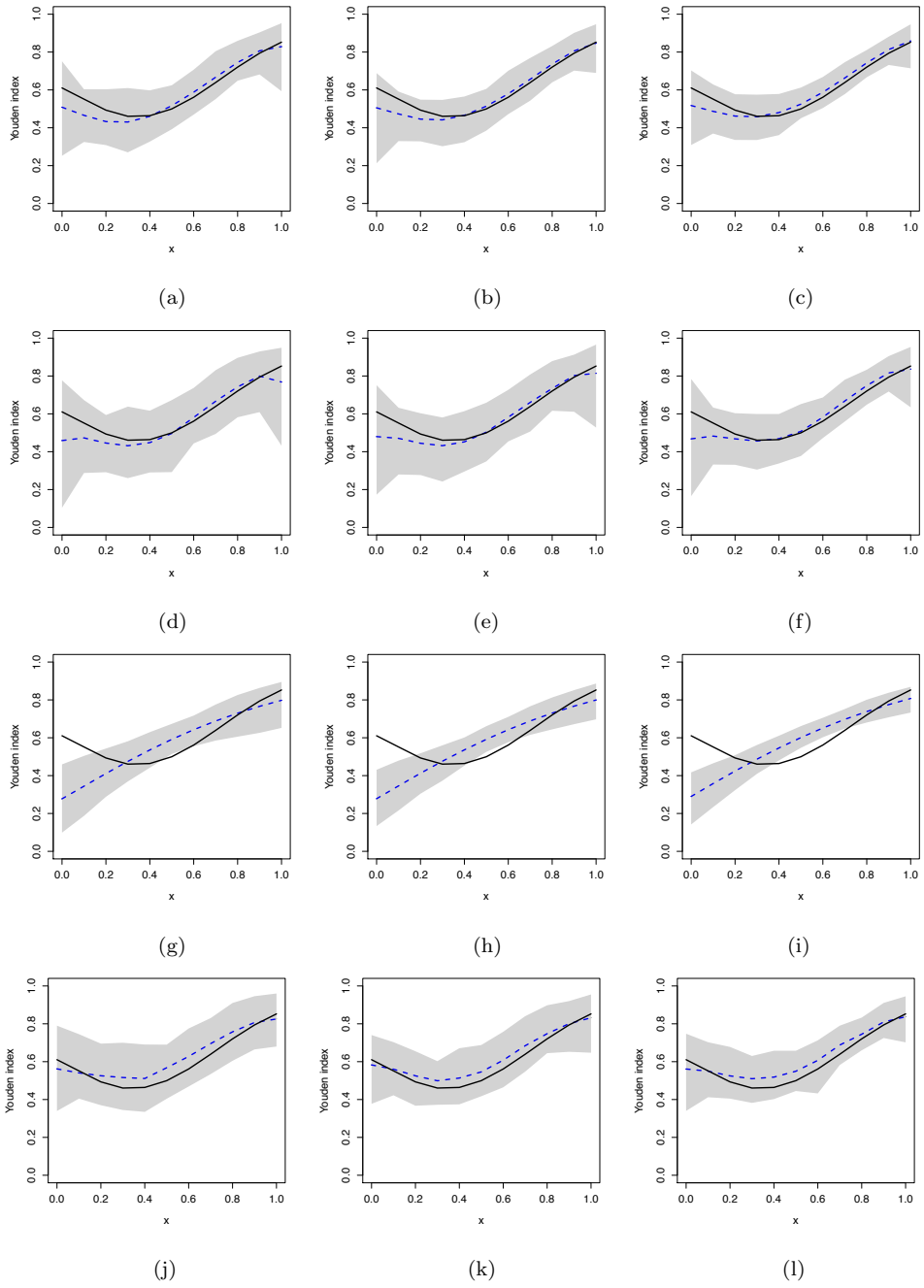


Figure 7. True (solid black line) and mean across simulations (dashed blue line) of the posterior mean (for the Bayesian estimators) of the Youden index function under Scenario 4. A band constructed using the pointwise 2.5% and 97.5% quantiles across simulations is presented in gray. Row 1: B-splines DDP estimator with $Q = 4$. Row 2: B-splines DDP estimator with $Q = 7$. Row 3: Normal estimator. Row 4: Kernel estimator. Panels (a), (d), (g), and (j) show the results for $(n_D, n_{\bar{D}}) = (100, 100)$, panels (b), (e), (h), and (k) for $(n_D, n_{\bar{D}}) = (100, 200)$, and panels (c), (f), (i), and (l) for $(n_D, n_{\bar{D}}) = (200, 200)$.

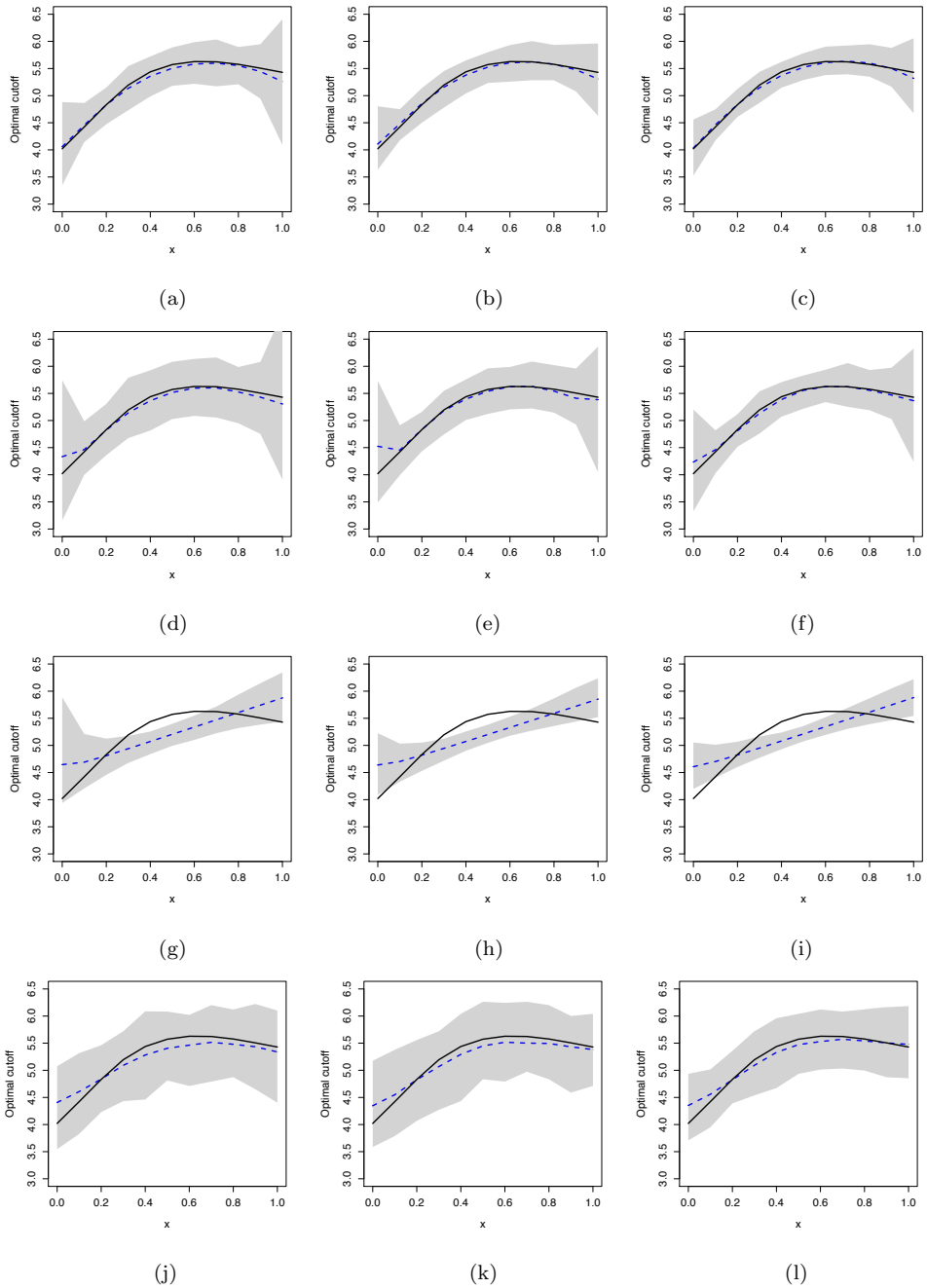


Figure 8. True (solid black line) and mean across simulations (dashed blue line) of the posterior mean (for the Bayesian estimators) of the optimal cutoff function under Scenario 4. A band constructed using the pointwise 2.5% and 97.5% quantiles across simulations is presented in gray. Row 1: B-splines DDP estimator with $Q = 4$. Row 2: B-splines DDP estimator with $Q = 7$. Row 3: Normal estimator. Row 4: Kernel estimator. Panels (a), (d), (g), and (j) show the results for $(n_D, n_{\bar{D}}) = (100, 100)$, panels (b), (e), (h), and (k) for $(n_D, n_{\bar{D}}) = (100, 200)$, and panels (c), (f), (i), and (l) for $(n_D, n_{\bar{D}}) = (200, 200)$.

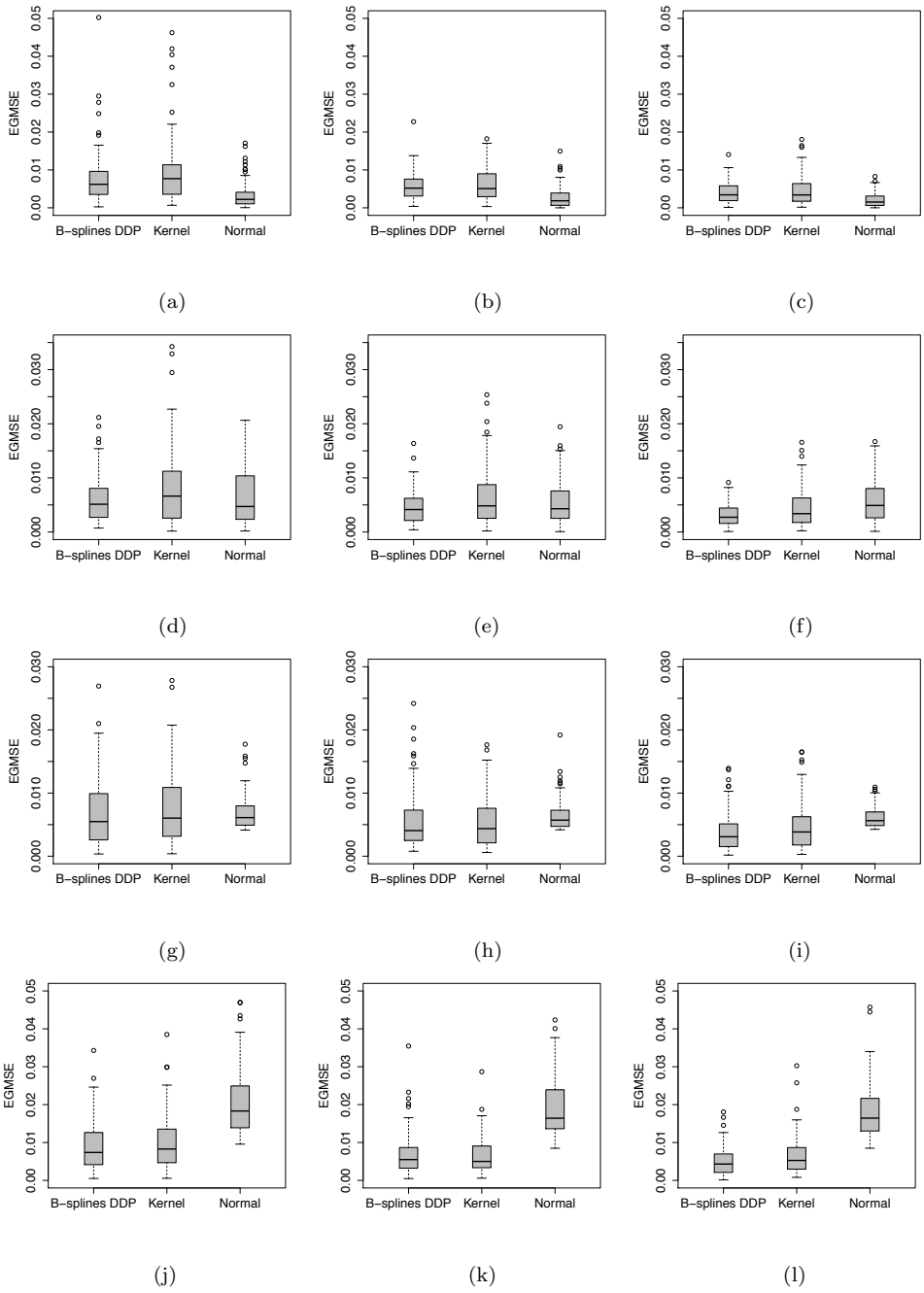


Figure 9. Boxplots of the empirical global mean squared error (EGMSE) of the Youden index across simulations for the B-splines DDP estimator ($Q = 4$), kernel estimator, and normal estimator. Panels (a)–(c), (d)–(f), (g)–(i), and (j)–(l) display the results under Scenario 1, 2, 3, and 4, respectively. Panels (a), (d), (g), and (j) show the results for $(n_D, n_{\bar{D}}) = (100, 100)$, panels (b), (e), (h), and (k) for $(n_D, n_{\bar{D}}) = (100, 200)$, and panels (c), (f), (i), and (l) for $(n_D, n_{\bar{D}}) = (200, 200)$.

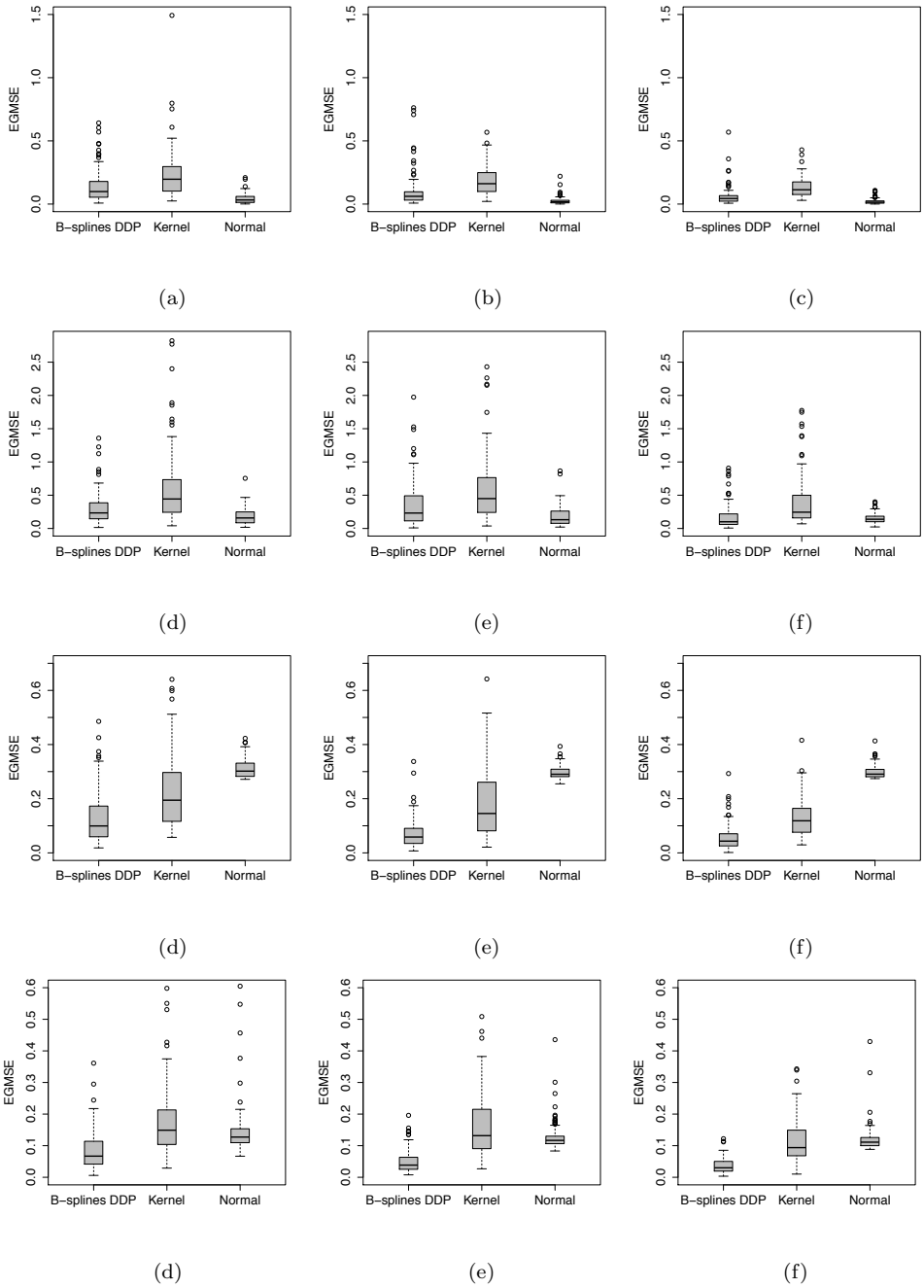
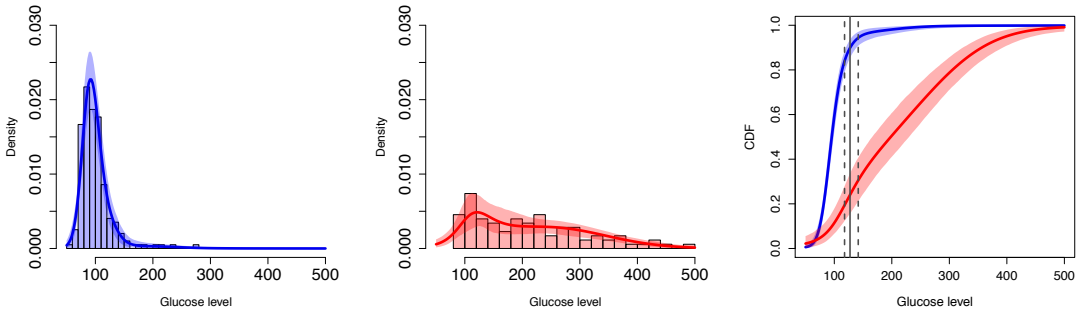
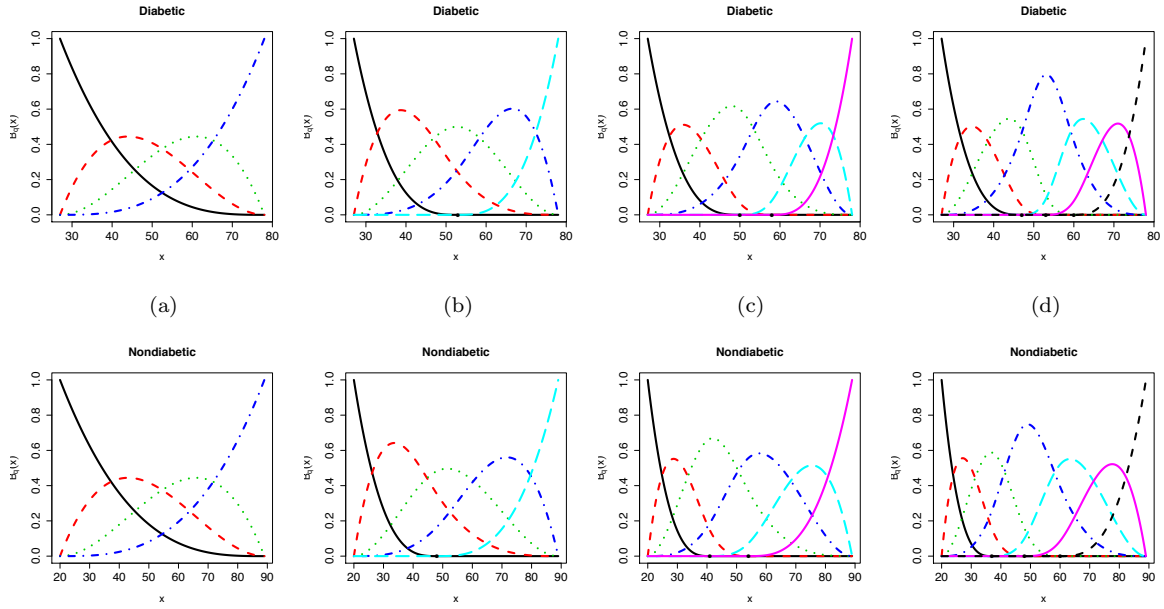


Figure 10. Boxplots of the empirical global mean squared error (EGMSE) of the optimal cutoff across simulations for the B-splines DDP estimator ($Q = 4$), kernel estimator, and normal estimator. Panels (a)–(c), (d)–(f), (g)–(i), and (j)–(l) display the results under Scenario 1, 2, 3, and 4, respectively. Panels (a), (d), (g), and (j) show the results for $(n_D, n_{\bar{D}}) = (100, 100)$, panels (b), (e), (h), and (k) for $(n_D, n_{\bar{D}}) = (100, 200)$, and panels (c), (f), (i), and (l) for $(n_D, n_{\bar{D}}) = (200, 200)$.



(a) (b) (c)

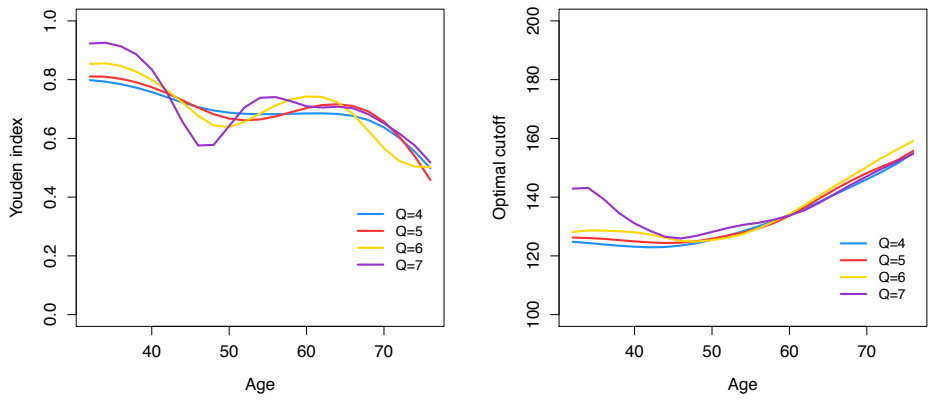
Figure 11. Histogram of the glucose levels in the non-diabetic (panel a) and diabetic group (panel b) along with the posterior mean and 95% pointwise probability interval of the density for each group under (independent) Dirichlet process mixtures of normals. Panel (c) presents the estimated distribution functions and optimal cutoff with its 95% probability interval.



(a) (b) (c) (d)

(e) (f) (g) (h)

Figure 12. Basis functions for the diabetes data. Panels (a) and (e): $Q = 4$. Panels (b) and (f): $Q = 5$. Panels (c) and (g): $Q = 6$. Panels (d) and (h): $Q = 7$.



(a) (b)

Figure 13. Diabetes data: posterior means of the Youden index (a) and optimal cutoff (b) for the different values of Q considered.

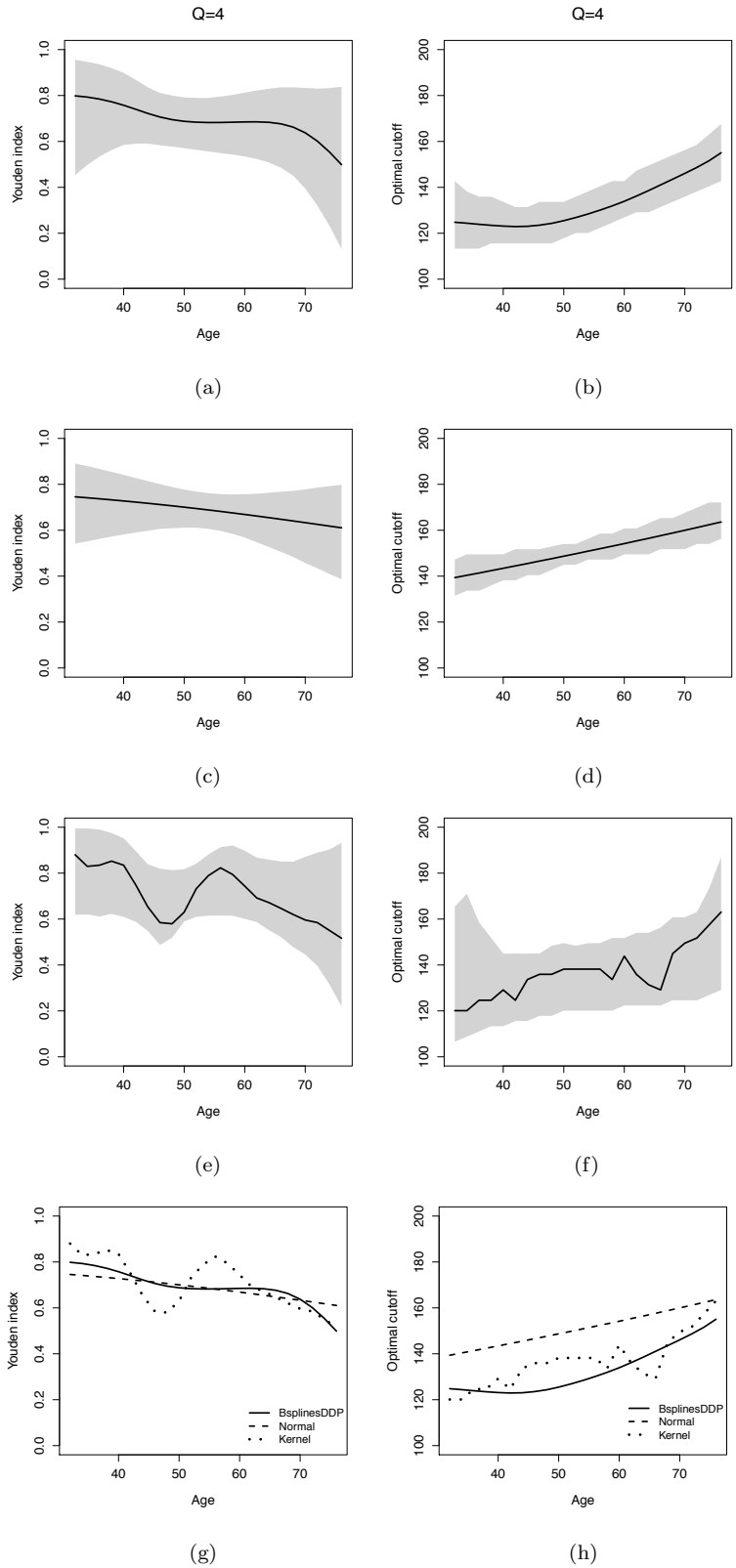


Figure 14. Estimated Youden index and optimal cutoff as a function of age. Panels (a) and (b) present results from the B-splines DDP estimator ($Q = 4$), panels (c) and (d) present results obtained under the normal linear model, while panels (e) and (f) show the results obtained under the nonparametric kernel model. For ease of comparison, panels (g) and (h) display the three estimators together.

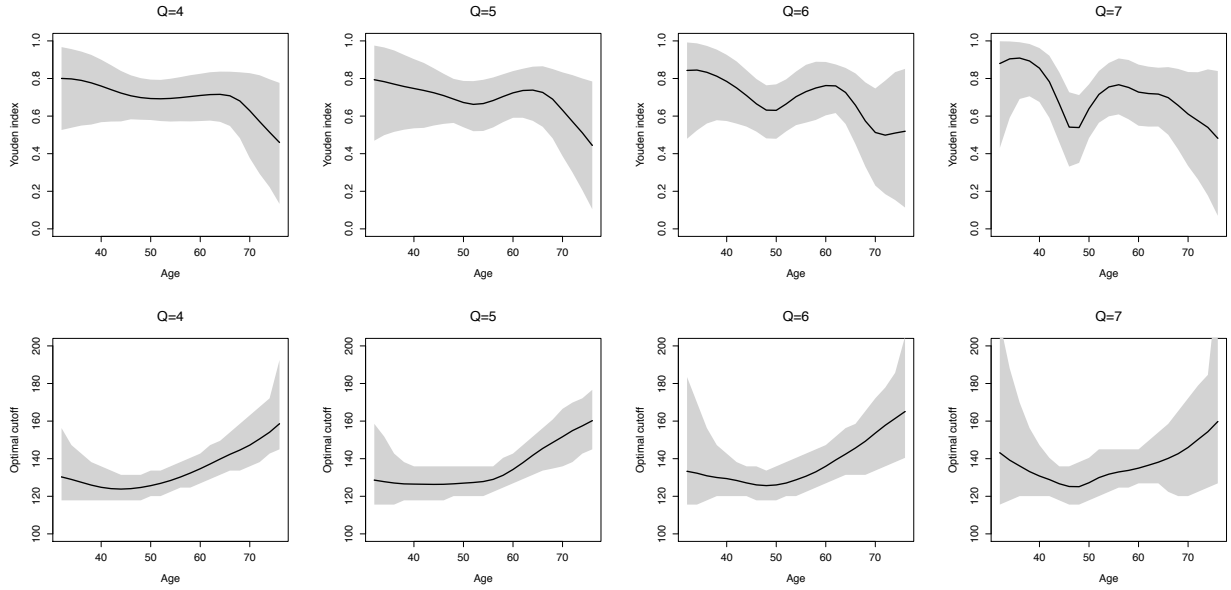


Figure 15. Estimated Youden index and optimal cutoff as a function of age, using the data-driven prior (Section D), for $Q = 4$, $Q = 5$, $Q = 6$, and $Q = 7$. Solid lines represent posterior means and the gray areas correspond to pointwise 95% posterior bands.

	Scenario 1		Scenario 2		Scenario 3		Scenario 4	
	$\text{LPML}_{\bar{D}}$	LPML_{D}	$\text{LPML}_{\bar{D}}$	LPML_{D}	$\text{LPML}_{\bar{D}}$	LPML_{D}	$\text{LPML}_{\bar{D}}$	LPML_{D}
$\mathbf{n_0 = n_1 = 100}$								
B-splines DDP	-187 $(-197, -174)$	-215 $(-224, -204)$	-224 $(-236, -213)$	-212 $(-224, -201)$	-187 $(-198, -174)$	-237 $(-246, -225)$	-178 $(-190, -165)$	-187 $(-197, -176)$
Normal	-185 $(-196, -173)$	-213 $(-222, -202)$	-235 $(-243, -227)$	-220 $(-227, -213)$	-193 $(-203, -180)$	-235 $(-244, -225)$	-182 $(-195, -170)$	-185 $(-195, -175)$
$\mathbf{n_0 = n_1 = 200}$								
B-splines DDP	-368 $(-383, -351)$	-427 $(-441, -411)$	-440 $(-452, -427)$	-414 $(-428, -399)$	-368 $(-382, -351)$	-471 $(-485, -456)$	-351 $(-368, -336)$	-371 $(-384, -356)$
Normal	-366 $(-382, -350)$	-425 $(-439, -409)$	-471 $(-480, -460)$	-436 $(-446, -424)$	-381 $(-395, -362)$	-470 $(-483, -453)$	-360 $(-376, -343)$	-369 $(-384, -353)$

Table 1

Average LPML and 90% *interval for each model in each simulation scenario.*

	Prior 1		Prior 2	
	$\text{LPML}_{\bar{D}}$	LPML_D	$\text{LPML}_{\bar{D}}$	LPML_D
$Q = 4$	-878	-530	-880	-530
$Q = 5$	-880	-532	-881	-531
$Q = 6$	-881	-531	-883	-531
$Q = 7$	-887	-532	-886	-532

Table 2

Diabetes data: LPML for the different values of Q considered under prior configuration 1 (the one used on the main manuscript) and under prior configuration 2 (the one described in Section D of this Supplementary Material).