Statistics
in Medicine

Research Article

Received 29 May 2014, Accepted 9 January 2015 Published online 26 January 2015 in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.6438

A Dirichlet process mixture model for survival outcome data: assessing nationwide kidney transplant centers

Lili Zhao,*† Jingchunzi Shi, Tempie H. Shearon and Yi Li

Mortality rates are probably the most important indicator for the performance of kidney transplant centers. Motivated by the national evaluation of mortality rates at kidney transplant centers in the USA, we seek to categorize the transplant centers based on the mortality outcome. We describe a Dirichlet process model and a Dirichlet process mixture model with a half-cauchy prior for the estimation of the risk-adjusted effects of the transplant centers, with strategies for improving the model performance, interpretability, and classification ability. We derive statistical measures and create graphical tools to rate transplant centers and identify outlying groups of centers with exceptionally good or poor performance. The proposed method was evaluated through simulation and then applied to assess kidney transplant centers from a national organ failure registry.

Keywords: Dirichlet process mixture; stick-breaking process; mixture model; clustering; survival data; transplant

1. Introduction

Monitoring and tracking the performance of health care providers, such as hospitals, nursing homes, dialysis facilities, or surgical wards, ensure the delivery of high quality care to the vulnerable patient population [1]. This article is in response to the urgent need for the evaluation of kidney transplant centers in the USA with respect to their mortality rates after transplantation. The data include patients in the Scientific Registry of Transplant Recipients who received their kidney from 2008 to 2011. A total of 56,455 kidney transplants were performed at 242 transplant centers.

There is a large amount of literature describing methods for the evaluation of center performances and identification of outlying centers with extremely good or poor performance. Data used to evaluate performance include binary (standardized) mortality data, counts of adverse events, or continuous data measuring quality of life. Some examples of using parametric approaches can be found in [2–5]. Among these articles, Liu *et al.* [2] and Jones and Spiegelhalter [3] used a normal hierarchical (random effects) model for the center effects. As we know, random effects models improve estimation by borrowing information across transplant centers and thus shrinking estimates of the center effects toward the overall mean and leading to a reduced variation of the estimates. However, the smaller variance is achieved at the cost of bias, and inappropriate shrinkage could prevent the centers with exceptionally good or poor performance from being identified. For this reason, He *et al.* [5] and Kalbfleisch and Wolfe [4] prefer a model with center effects being considered as fixed, leading to independent (no shrinkage) center estimates. It seems that a desirable model would combine the advantages of both the fixed and random effects models, in the sense that it would allow borrowing strength across similar centers, but avoid shrinking outlying centers toward the population mean.

Moreover, in both random or fixed effects models, it is not immediately clear how unusual centers, that is, any with exceptionally good or poor performance, can be identified. A common strategy is to measure

Department of Biostatistics, University of Michigan, Ann Arbor, MI, U.S.A.
**Correspondence to: Lili Zhao, Department of Biostatistics, University of Michigan, Ann Arbor, MI, U.S.A.*
†E-mail: zhaolili@umich.edu

the deviation of each transplant center relative to the population average using a p-value or a (adjusted) Z-score derived from an assumed parametric or empirical null distribution [4, 6]. However, ideally, a model would provide an in-built diagnostic measure for centers with unusual outcomes.

Ohlssen *et al.* [7] applied a Dirichlet process (DP) model and a Dirichlet process mixture (DPM) model to the problem of hospital comparisons using mortality rates. The nonparametric Bayesian approach satisfies the aforementioned requirements. It allows for a more flexible distribution of hospital effects and accommodates outlying hospitals. Furthermore, the embedded clustering feature in DP models provides inherent diagnostic measures to identify outlying centers.

However, Ohlssen *et al.* [7] considered binary mortality data. In our application, majority of the data are censored; as of January 31, 2013, 93% patients were still alive. Therefore, we extend the work of Ohlssen *et al.* to estimate center effects with survival outcomes (i.e., time-to-death), in which center effects are represented as random effects (frailties) in a Cox proportional hazards model. The model with a mixture of points (DP) or mixture of normals (DPM) provides a more flexible distribution for the center effects compared to a parametric frailty model. More importantly, the model classifies centers into different subgroups (clusters), and centers within the same cluster have similar mortality rates (performance). To our knowledge, this would be the first attempt to apply such an approach to evaluate survival outcomes of nationwide transplant centers.

The remaining of the article is organized as follows. In Section 2, we describe a DP model and a DPM model for the estimation of center effects in a Cox proportional hazards model, propose strategies to improve model performance, and create graphical tools to evaluate centers. In Section 3, we present simulation studies to investigate the clustering performance and shrinkage effects for data with different clustering structures. In Section 4, we illustrate the analysis on the kidney transplant data. Section 5 is the concluding discussion.

2. Model

2.1. Cox proportional hazards model

The data are denoted by $\{(t_{ij}, \delta_{ij}, x_{ij}), i = 1, \cdots, N; j = 1, \cdots, n_i\}$, where t_{ij} is the observed event time for patient j in transplant center i; $\delta_{ij} = 1$ if t_{ij} is an observed failure time and 0 if the failure time is right censored at t_{ij}, and x_{ij} is a p-dimensional vector of covariates.

Under the proportional hazards model, we have

$$\lambda(t_{ij}) = \lambda_0(t_{ij}) \exp\{\alpha x_{ij} + \beta_i\}$$

where α denotes the effects of patient-level covariates and β_i is the effect associated with center i after adjusting for patient-level covariates. The baseline hazard, $\lambda_0(t)$, is assumed to be piecewise constant on a partition composed of K disjoint intervals, yielding the piecewise exponential model [8, 9]. That is, $\lambda_0(t) = \sum_{k=1}^{K} \lambda_k \mathrm{I}(a_{k-1} < t \leqslant a_k)$, where $a_0 = 0$ and $a_K = \max\{t_{ij}\}$. Let $\lambda = (\lambda_1, \cdots, \lambda_K)$, the likelihood for (α, λ, β) is given by

$$L(\alpha, \lambda, \beta) = \prod_{i=1}^{N} \prod_{j=1}^{n_i} f\left(N_{ij}, x_{ij}, \boldsymbol{\Delta}_{ij}; \alpha, \lambda, \beta\right)$$

and

$$f(\boldsymbol{N}_{ij}, x_{ij}, \boldsymbol{\Delta}_{ij}; \alpha, \lambda, \beta) = \prod_{k=1}^{K} \exp\{-\exp\{\log \lambda_k + \alpha x_{ij} + \beta_i\}\Delta_{ijk}\} \{\exp\{\log\{\lambda_k\} + \alpha x_{ij} + \beta_i\}\Delta_{ijk}\}^{N_{ijk}} \tag{1}$$

where $\boldsymbol{N}_{ij} = (N_{ij1}, \cdots, N_{ijK})$ and N_{ijk} takes a value of one if $t_{ij} \in (a_{k-1}, a_k]$ and $\delta_{ij} = 1$ and N_{ijk} is zero otherwise. Define $\boldsymbol{\Delta}_{ij} = (\Delta_{ij1}, \cdots, \Delta_{ijK})$, and $\Delta_{ijk} = (\min\{a_k, t_{ij}\} - a_{k-1})_+$ with x_+ as $\max(x, 0)$.

In this article, the attention is focused on modeling random center effects, $\beta_1, \cdots, \beta_N$, which characterize heterogeneities of transplant centers. A large value of β_i corresponds to a high mortality rate (poor performance) associated with center i. Often, $\beta_i's$ are assumed to be generated from some parametric distribution such as log-normal, gamma, and positive stable. In the next section, we propose to estimate the random effects by a nonparametric Bayesian model.

2.2. Truncated stick-breaking process

To model the random center effects of $\beta_1, \cdots, \beta_N$, we first consider a model of a mixture of point masses using a DP prior,

$$\beta_1, \cdots, \beta_N \sim G$$
$$G \sim \mathrm{DP}(a, G_0),$$

where G_0 corresponds to a best guess for G as a priori and a expresses confidence in this guess.

The stick-breaking representation [10] implies that $G \sim \mathrm{DP}(a, G_0)$ is equivalent to

$$G = \sum_{h=1}^{\infty} \pi_h \delta_{\beta_h}, \quad \beta_h \sim G_0, \quad \text{and} \quad \sum_{h=1}^{\infty} \pi_h = 1 \tag{2}$$

where G is a mixture of countably but infinite atoms, and these atoms are drawn independently from the base distribution G_0, and δ_β is a point mass at β. In our study, an atom is like a cluster (i.e., a subgroup of centers), π_h is the probability assigned to the h^{th} cluster, β_h is the value of that cluster, and all transplant centers in a cluster share the same β_h. In (2), $\pi_h = V_h \prod_{l<h}(1 - V_l)$, which is formulated from a stick-breaking process, with $V_h \sim \mathrm{Beta}(1, a)$ for $h = 1, \cdots, \infty$. In this stick-breaking process, V_1 is the proportion of the stick broken off and assigned to β_1, and V_2 is the proportion of the remaining $1 - V_1$ length stick assigned to β_2, and so on. For values of a close to zero, $V_1 \approx 1$, so $\pi_h \approx 1$, which essentially assign all probability weight to a single cluster, while for large a, each of the clusters is assigned a vanishingly small weight, so the number of clusters could be as many as the number of centers. In this article, we fix $\alpha = 1$, a widely used choice in applications that favors a few clusters [11].

One potential issue with this representation of a mixture of point masses is that it assumes a discrete distribution for the random effects so that different centers in a cluster have exactly the same random effect values. It may be more realistic to assume that centers in a cluster have similar, but not identical, random effect values. To accomplish this, let $\beta_i \sim N\left(\mu_h, \sigma_h^2\right)$ and $\left(\mu_h, \sigma_h^2\right) \sim G$. That is, (2) becomes

$$G = \sum_{h=1}^{\infty} \pi_h \mathrm{N}\left(\mu_h, \sigma_h^2\right), \quad \left(\mu_h, \sigma_h^2\right) \sim G_0, \quad \text{and} \quad \sum_{h=1}^{\infty} \pi_h = 1 \tag{3}$$

In this case, the random distribution, G, is characterized as a DPM of normals [12]. A mixture of normals allow a flexible continuous random effects distribution of the center effects. (Readers can refer to a book by Dunson [11] for a detailed review of the DP and DPM model).

Recent research has focussed on using the constructive definition of the DP to produce practical MCMC algorithms [13]. The principle is to approximate the full process by truncating the DP(M) at a maximum number of clusters H, so that

$$G = \sum_{h=1}^{H} \pi_h \delta_{\beta_h} \text{ in DP and } G = \sum_{h=1}^{H} \pi_h \mathrm{N}\left(\mu_h, \sigma_h^2\right) \text{ in DPM}$$

A large H provides an accurate approximation to the full DP(M) but requires a large computation effort. Strategies have been proposed to specify H [7, 14]. In this study, we are interested in detecting subgroups (clusters) of centers with exceptionally good or poor performance compared to the population average, so we set the maximum number of clusters to be 5 (i.e., $H = 5$) in both simulation studies and the real case example. We also evaluate the sensitivity of the model with a larger H.

2.3. Classify centers into different clusters

The blocked sampler of Ishwaran and James [13] is used to allocate each center to one of the clusters by sampling the label Z_i $(i = 1, \cdots, N)$ from a multinomial conditional posterior. In the DP model, probabilities in the multinomial distribution are as follows:

$$Pr(Z_i = h|-) = \frac{\left\{V_h \prod_{l<h}(1 - V_l)\right\} \prod_{j=1}^{n_i} f(\boldsymbol{N}_{ij}, \boldsymbol{\Delta}_{ij}, x_{ij}; \alpha, \lambda, \beta_h)}{\sum_{r=1}^{H} \left\{V_r \prod_{l<r}(1 - V_l)\right\} \prod_{j=1}^{n_i} f(\boldsymbol{N}_{ij}, \boldsymbol{\Delta}_{ij}, x_{ij}; \alpha, \boldsymbol{\lambda}, \beta_r)},$$

where $f(\boldsymbol{N}_{ij}, \boldsymbol{\Delta}_{ij}, x_{ij}; \alpha, \boldsymbol{\lambda}, \beta)$ is defined in (1).

In the DPM, the probabilities are

$$Pr(Z_i = h|-) = \frac{\left\{V_h \prod_{l<h}(1-V_l)\right\} \prod_{j=1}^{n_i} \left\{N\left(\beta_i; \mu_h, \sigma_h^2\right)\right\}^{\eta}}{\sum_{r=1}^{H} \left\{V_r \prod_{l<r}(1-V_l)\right\} \prod_{j=1}^{n_i} \left\{N\left(\beta_i; \mu_h, \sigma_h^2\right)\right\}^{\eta}} \tag{4}$$

Motivated by the work of Hofmann [15], we introduce a tempering parameter η in (4). Similar strategies have been used in simpler mixture models for efficient Gibbs sampling [16]. When $\eta = 1$, the allocation probability is weighted by the regular likelihood; when $\eta > 1$, the contribution of each observation is strengthened. Our simulation studies reveal that an η of 2 leads to significantly improved clustering performance compared to an η of 1, especially when the prior for cluster parameters (such as $\sigma_1^2 = \sigma_2^2 = \cdots = \sigma_H^2$) are weak. In different contexts, such as in adaptive randomization trials [17], a similar annealing parameter ($\eta < 1$) has been found useful in making randomization more balanced between treatment groups.

2.4. *Centered stick-breaking process*

In parametric hierarchical models, it is a standard practice to place a mean constraint on the latent variable distribution for the sake of identifiability and interpretability [18, 19]. In this article, we center the DP to have zero mean. Following Yang *et al.* [18], we estimate the mean of the process, μ_G^m, at the m^{th} MCMC iteration as

$$\mu_G^m = \sum_{h=1}^{H} V_h^m \prod_{l<h} \left(1 - V_l^m\right) \beta_h^m$$

where V_h^m and β_h^m are the posterior samples from the uncentered process defined in (2) and $\beta_i^m - \mu_G^m$ $(h = 1, \cdots, H)$ is the 'centered' estimate for center i at the m^{th} iteration. The same idea applies to the DPM model.

Centering the process improves model performance in two aspects. First, it improves MCMC convergence and mixing rates. Second, a 'centered' estimate can be interpreted as a deviation from the population average.

2.5. *Prior specification*

In models (2) and (3), G_0 is often chosen to have a normal distribution. In *DP* model, $\beta_h \sim N\left(\mu_0, \sigma_0^2\right)$ $(h = 1, \cdots, H)$. The hyperparameters $\left(\mu_0, \sigma_0^2\right)$ can be fixed or assigned a normal-inverse gamma hyperprior. A hyperprior would allow the base distribution having unknown mean and variance and provide a shrinkage of center effects toward the overall mean. In *DPM* model, we assumed that $\mu_h \sim N\left(\mu_0, \sigma_0^2\right)$, $(h = 1, \cdots, H)$, with a normal hyperprior for μ_0 and a half-cauchy prior for σ_0 [20], that is, $f(\sigma_0) \propto \left(1 + \left(\frac{\sigma_0}{A}\right)^2\right)^{-1}$, with a smaller A indicating a stronger prior information and a greater shrinkage. This Cauchy prior behaves well for a small number of clusters, and it restricts σ_0^2 away from very large values and has better behavior near zero, compared to the inverse-gamma family [20]. We also assume that $1/\sigma_h^2 \sim$ Gamma(e_0, f_0) and fix hyperparameters (e_0, f_0) to be weakly informative, because fully noninformative priors are not possible in a mixture context [21].

In the Cox proportional hazards model, the gamma process is used as a prior for the cumulative baseline hazard function Λ_0 [22], that is, $\Lambda_0 \sim \mathcal{GP}\left(c_0 \Lambda_0^*, c_0\right)$, where Λ_0^* is often assumed to be a known parametric function. For example, $\Lambda_0^* = b_0 y^{k_0}$ corresponds to the Weibull distribution, and c_0 represents the degree of confidence in this prior guess. Normal priors are used for the effect of covariates, α, and gamma priors are used for precision parameters.

2.6. *Statistical measures to rate centers*

In this section, we propose metrics to evaluate and cluster N transplant centers by modeling random effects $\beta_i \cdots, \beta_N$. A simple metric to rate transplant centers is their *ranks*. At each MCMC iteration, β_i $(i = 1, \cdots, N)$ is ranked; without ties, the smallest β_i has rank 1, and the largest β_i has rank N. Over all MCMC iterations, we obtain a distribution of ranks for each center. Another useful metric to assess *pairwise clustering* between centers is an $N \times N$ matrix of posterior probabilities of two centers being

classified into the same cluster [7, 23]. This posterior probability between any two centers is calculated as the number of times two centers are assigned into the same cluster across all MCMC iterations. We combine the aforementioned two measures and graphically represent the $N \times N$ probability matrix using a heat map where transplant centers are ordered by their posterior means of the *ranks*. This heat map reveals a clustering structure of the studied centers that facilitates rating centers, as well as identifying outlying groups of centers.

Additionally, in order to visually detect outlying centers, we calculate the proportion of centers in the same cluster as center i, denoted by PS. Together with the *rank* (percentile) statistics, we create a graph that helps identify centers that are in isolated small clusters with exceptionally low or high ranks.

3. Simulation studies

In this section, we conduct simulation studies to investigate performance of the proposed DP and DPM models, in terms of accurately estimating center effects $\beta_1, \cdots, \beta_N$ and correctly identifying true clustering structures. We generate patient survival data for $N = 48$ transplant centers. Specifically, survival times are generated from a Cox model [24], $S(t|\text{center}_i) = \exp[-\Lambda_0(t)\exp(\beta_i)]$, where Λ_0 is the cumulative hazard function of a Weibull distribution, with a scale parameter of one and a shape parameter of 0.8, suggesting the population mortality rate decreases over time, which is observed in the kidney transplant data. For illustration purposes, we do not include covariates in the simulation. Center effects $\beta_1, \cdots, \beta_{48}$ are generated from a normal distribution, and a value of β_i larger (smaller) than zero represents the mortality rate of center i below (above) the population average. Different values of β_i allocate 48 centers into three subgroups (true clusters) as shown in Table I. For example, in scenario I, the first 16 centers, $\beta_1, \cdots, \beta_{16}$, are simulated from a normal distribution with a mean -0.69 and a standard deviation (SD) of 0.2, which form a cluster (denoted by C1) with the above-average performance; likewise, the next 16 centers form a cluster (C2) with the population-average performance and the last 16 centers form a cluster (C3) with the below-average performance. β_i of -0.69 and 0.69 correspond to a hazard ratio of 0.5 and 2 relative to the population average, respectively. These clinically meaningful ratios are expected to be detected in the real data analysis. In *scenario IV*, we generate all β_i $(i = 1, \cdots, 48)$ from $N(0, 0.3)$ to investigate if the model can correctly assign all centers into a single cluster. In all scenarios, within each cluster, the first half centers have 20 patients $(n = 20)$, and the other half centers have 40 patients $(n = 40)$.

We use three models described in Section 2.5 to estimate center effects for the aforementioned simulated data:

(1) *DP*: a DP model with fixed hyperparameters, that is, $\left(\mu_0, \sigma_0^2\right)$ are fixed.
(2) *DP-HP*: a DP model with a random normal-inverse gamma hyperprior for $\left(\mu_0, \sigma_0^2\right)$.
(3) *DPM*: a DPM model with $\mu_h \sim \left(\mu_0, \sigma_0^2\right)$, where μ_0 has a normal hyperprior and σ_0 has a half-cauchy prior with $A = 1$.

The DP model, with fixed hyperparameters, does not induce shrinkage between clusters but shrinks centers within the same cluster to a single estimate. In contrast, DPM allows shrinkage between and within clusters with a smaller A indicating a stronger shrinkage (A of 1 is chosen to be higher than we expect for the SD of the underlying $\mu_h's$ so that the model will constrain σ_0 only weakly). Intuitively, DP-HP could have a stronger shrinkage than DPM because DP-HP has the strongest shrinkage within cluster

Table I. Data generated for 48 transplant centers with three clusters.

Scenario		C1	C2	C3
I	♯ centers[a]	16	16	16
	β_i	$N(-0.69, 0.2^2)$	$N(0, 0.2^2)$	$N(0.69, 0.2^2)$
II	♯ centers	8	24	16
	β_i	$N(-0.69, 0.2^2)$	$N(0, 0.2^2)$	$N(0.69, 0.2^2)$
III	♯ centers	4	40	4
	β_i	$N(-0.69, 0.1^2)$	$N(0, 0.2^2)$	$N(0.69, 0.1^2)$

[a]♯ centers denotes number of centers; the first half centers consist of 20 patients and the other half centers consist of 40 patients.

by forcing all centers within a cluster having the same estimate, as well as a between-cluster shrinkage that is induced by a hyperprior for (μ_0, σ_0^2).

We compute three Bayesian model comparison criteria for selecting the best model: modified deviance information criterion (DIC_3) [25], Watanabe–Akaike information criterion (WAIC) [26], and log-pseudo marginal likelihood (LPML) [9]. DIC_3 is preferred in our setting over the standard DIC proposed by Spiegelhalter *et al.* [27] because it correctly reflects the effective number of parameters in mixture models. WAIC was proposed recently and can also be viewed as an improvement over the standard DIC, and it also approximates Bayesian cross-validation. It is invariant to parametrization and also works for singular model [28]. LPML is a cross-validated leave-one-out measure of a model's ability to predict the data. It is valid for small and large samples and does not suffer from a heuristic justification based on large sample normality. The best model should have the smallest DIC_3 and WAIC and the largest LPML.

We also calculate the mean classification error to evaluate the clustering performance. In Section 2.6, we define a $N \times N$ matrix of posterior probabilities of two centers being classified into the same cluster. If the true clustering structure is known (such as in the simulation studies), a $N \times N$ (0, 1) matrix would represent a true probability matrix with 1 indicating a pair of centers in the same cluster and 0 indicating they are not in the same cluster. The deviation of the posterior probability matrix from the true probability matrix depicts a $N \times N$ matrix with classification errors. First, we calculate the average error for all pairs of centers in the same cluster and for all pairs of centers not in the same cluster separately, and then we average these two types of errors to obtain the mean classification error (MCE). The smaller the mean classification error ($0 \leqslant \text{MCE} \leqslant 1$), the better the clustering performance.

Posterior computation is presented in the Appendix (item 2–8). The models are implemented in **R**. All normal priors are assumed to have a mean of zero and a variance of 100. The baseline hazard is assumed to have an exponential distribution with $b_0 = 1$ and $c_0 = 0.1$ (a robust prior used in [29]), and the time axis is partitioned into five intervals based on the observed quantiles. All the priors are set to be quite weak. With a burn-in of 1000 iterations, an additional 2000 iterations are used for inference. In calculation of allocation probabilities in the DPM model of (4), we observe that $\eta = 2$ significantly improved clustering performance compared to $\eta = 1$; when $\eta = 2$, the mean classification error is 0.29, 0.27, and 0.25 in

Table II. Parameter estimation with respect to the absolute bias (Bias), standard deviation (SD), and mean square error (MSE), based on 1000 simulated datasets.

			DP			DPM		
Scenario		n	C1	C2	C3	C1	C2	C3
I	Bias	20	0.15	0.03	0.12	0.11	0.02	0.08
		40	0.08	0.02	0.06	0.06	0.02	0.02
	SD	20	0.24	0.28	0.25	0.30	0.27	0.27
		40	0.23	0.21	0.22	0.26	0.18	0.24
	MSE	20	0.08	0.08	0.08	0.10	0.07	0.09
		40	0.06	0.04	0.05	0.07	0.03	0.06
II	Bias	20	0.11	0.09	0.25	0.05	0.09	0.21
		40	0.02	0.11	0.17	0.01	0.10	0.15
	SD	20	0.31	0.21	0.26	0.33	0.22	0.28
		40	0.29	0.15	0.25	0.28	0.17	0.25
	MSE	20	0.11	0.05	0.13	0.11	0.06	0.12
		40	0.09	0.04	0.09	0.08	0.04	0.08
III	Bias	20	0.35	0.00	0.30	0.28	0.01	0.25
		40	0.21	0.01	0.15	0.16	0.01	0.15
	SD	20	0.30	0.12	0.27	0.34	0.14	0.26
		40	0.31	0.09	0.26	0.30	0.12	0.25
	MSE	20	0.21	0.01	0.16	0.20	0.02	0.13
		40	0.14	0.01	0.09	0.12	0.01	0.09

DP, Dirichlet process; DPM, Dirichlet process mixture.

scenarios I–III, respectively, which are much smaller than the that of 0.49, 0.49, and 0.47 observed when $\eta = 1$. The mean classification error with $\eta = 3$ is very similar to $\eta = 2$. Additionally, estimates of center effects are very close for different values of η. Because the setting of our simulation mimics the real example, we present the results based on $\eta = 2$ in both simulation studies and the real case study as follows.

Table II shows the parameter estimations in scenarios I–III, with respect to the absolute bias (Bias), SD, and mean square error (MSE), based on 1000 repeated datasets. As expected, centers with $n = 40$ have more accurate estimates compared to centers with $n = 20$, as evidenced by a smaller bias, SD, and MSE. Surprisingly, parameter estimations are very similar between DP-HP and DP models, so we only present the DP for illustration. It is also interesting to note that, in both DP and DPM models, estimate

Table III. Diagnosis statistics under four studied scenarios.

Scenario		DP	DP-HP	DPM
I	DIC_3	6157	6154	**6142**
	WAIC	6170	6167	**6155**
	LPML	−3062	−3061	**−3055**
II	DIC_3	6414	6418	**6394**
	WAIC	6430	6433	**6407**
	LPML	−3189	−3196	**−3181**
III	DIC_3	**6311**	6318	6316
	WAIC	**6323**	6433	6327
	LPML	**−3139**	−3142	−3142
IV	DIC_3	6382	6386	**6380**
	WAIC	**6382**	6386	6383
	LPML	−3182	−3184	**−3179**

DIC_3, deviance information criterion; WAIC, Watanabe–Akaike information criterion; LPML, log-pseudo marginal likelihood; DP, Dirichlet process; DPM, Dirichlet process mixture. Best statistics are in bold.

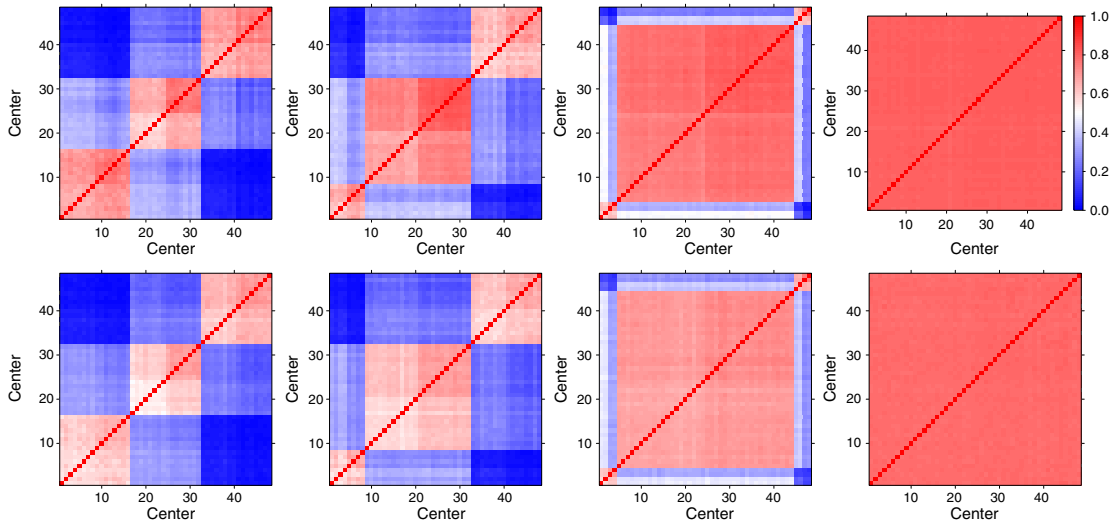


Figure 1. Pairwise posterior probabilities of two centers assigned to the same cluster under four scenarios using the Dirichlet process (DP) model (the first row) and the Dirichlet process mixture model (the second row), with $H = 5$ and $\eta = 2$. White, red, and blue color corresponds to a probability of equal to, larger, and less than 0.5, respectively. The darker the red, the closer the probability is to 1; the darker the blue, the closer the probability is to 0. The heat map of DP-HP is very similar to DP, so their heat maps are not presented.

Table IV. Mean classification errors under four studied scenarios.

Scenario	n	DP	DPM
I	20	0.32	0.36
	40	0.26	0.31
II	20	0.29	0.31
	40	0.24	0.26
III	20	0.27	0.29
	40	0.22	0.25
IV	20	0.19	0.21
	40	0.19	0.21

DP, Dirichlet process; DPM, Dirichlet process mixture.

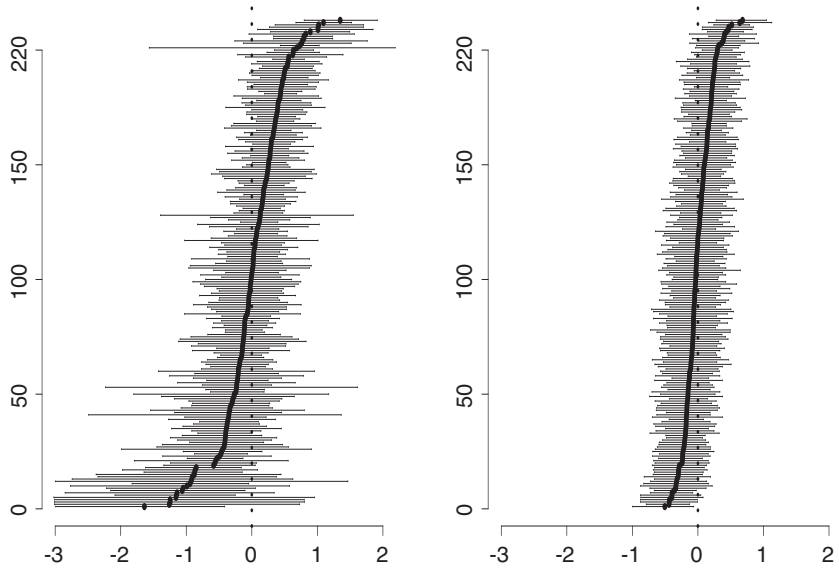
in a small clusters can be significantly biased toward a large cluster. In scenario IV, the mean estimate of center effect is very close to the true mean of zero (−0.0002 and 0.0004 for centers with $n = 20$ and $n = 40$, respectively). Similarly, centers with a larger sample size have smaller SD and MSE (results not shown here). As indicated by the diagnostic statistics in Table III, when data consist of a few clusters, and each with a decent number of centers (such as scenarios I and II), DPM is a better choice. A DP model is slightly better if a cluster consisting of a large number of centers is accompanied by a few small outlying clusters (such as scenario III). In the last scenario, when there is only a single cluster, the DP and DMP models perform similarly.

Next, we look at the clustering performance. Figure 1 displays the estimated clustering structure for the DP and DPM model under four scenarios over 1000 repeated datasets. Each heat map is created based on 48×48 matrix, containing pairwise posterior probabilities between centers as defined in Section 2.6. Because we know the true cluster status for each center, centers are ordered by their true IDs. It is apparent that the true clustering structure is well represented in all scenarios in both DP and DPM models. For example, in scenario I, the model correctly identifies three clusters (subgroups) as characterized by three red squares, and each red square consists 16 centers having high probabilities of being classified into the same cluster; the blue square represents small probabilities of the 16 below-average centers being classified into the same cluster as the 16 population-average centers, and the probabilities are even smaller (darker blue) between the 16 below-average and the 16 above-average centers. Furthermore, centers with a large sample size ($n = 40$) are more likely to be classified correctly than centers with a small sample size ($n = 20$), as evidenced by smaller mean classification errors in scenarios I–III as shown in Table IV. It is interesting to note that the DP model seems to have (slightly) better clustering performance compared to the DPM model. In scenario IV, all centers are correctly assigned to a single cluster (see the last column of Figure 1), but the clustering performance is similar with different sample sizes (mean classification errors are the same for $n = 20$ and $n = 40$).

4. Application

We apply our model to evaluate nationwide transplant centers in the USA. We exclude all centers with less than 10 patients in total, leaving data for 213 transplant centers. The number of patients per center has a median of 198 and an interquartile range of (111, 356). Survival outcome is defined from the time of kidney transplantation to death; patients who are alive at the last follow-up time point were considered to be right centered. A total of nine patient-level covariates are selected using a forward selection algorithm and also per relevant medical literature, including cold ischemia time, peak renal reactive antibody level, body mass index, time on renal replacement therapy, donor race, recipient race, donor history of diabetes, previous solid organ transplant, and recipient diagnosis. Because of the retrospective nature of the analysis, values are missing for some of those characteristics. For instance, there are 16.47% missing data in time on renal replacement therapy and 2.14% missing in peak renal reactive antibody. In order to include patients with partially missing covariates while reserving the original covariate distributions, we

create a binary variable for each covariate indicating if the data are missing for each subject. For example, a continuous covariate is created into two variables with one variable containing the original value and the other variable containing one if the data are missing and zero otherwise. By doing so, we create 18 covariates. Because of the large number of transplants (>50,000) and the large dimension of patient-level covariates, Kalbfleisch and Wolfe [4] used a two-stage approach to obtain the risk-adjusted center effects. In the first stage, they estimated patient-level covariates from a Cox model stratified by transplant centers; in the second stage, they derived center effects by fixing the covariate effects obtained from the first stage. However, we apply a fully Bayesian approach, in which we use a Gibbs algorithm that alternates between (1) updating effects of covariates with a Metropolis–Hastings algorithm conditional on estimated center effects and then (2) updating center effects conditional on estimated covariate effects in a DP or DPM model. Further details on MCMC sampling is given in the Appendix.



(a) Fixed effects (b) Normal random effects

(c) DP (d) DPM

Figure 2. Caterpillar plots of 95% credible intervals for estimates of the 213 center effects in kidney transplant data. The transplant centers were placed in the order of their posterior means. The dotted vertical line corresponds to the population average.

The priors used in the application are the same as in the simulation studies except that $1/\sigma_h^2 \sim$ Gamma$(3, 0.5)$, which is also weak relative to the likelihood. Again, $H = 5$ and $\eta = 2$ are used. With a burn-in of 10,000 iterations, an additional 20,000 iterations were used for posterior inference. It takes about 4 h for data to run on an Intel Xeon 3.10 GHz 4 GB RAM, x64 Linux computer. We observe that the chain mixes well and the results are robust to different choices of the initial values.

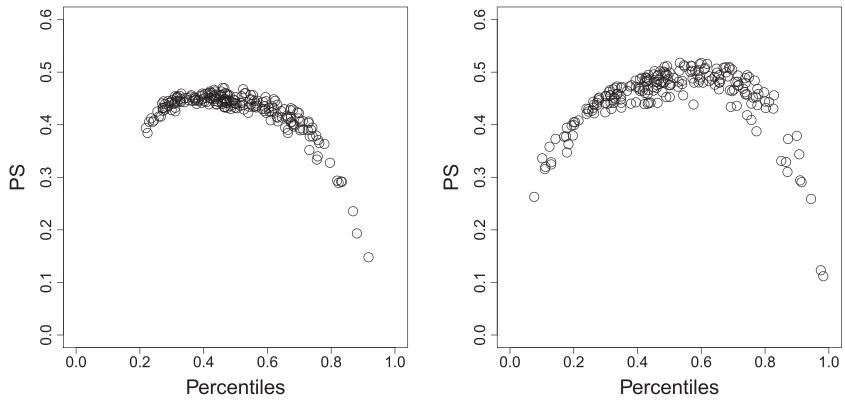
As a comparison, we also fit a fixed effects model and a normal random effects (NRE) model to the transplant data. In the NRE model, $\beta_1, \cdots, \beta_N$ are assumed to be generated from $N\left(\mu_\beta, \sigma_\beta^2\right)$, and μ_β has a normal prior with mean zero and variance 100 and $1/\sigma_\beta^2 \sim$ Gamma$(3, 0.5)$ (same priors as in the DPM model). Figure 2 presents caterpillar plots of posterior estimates of the 213 centers. Among the four models, the NRE, DP, and DPM models have very similar ordering of the 213 posterior means; for example, three centers with the largest β_i and two centers with the smallest β_i are exactly the same for the three models. In Table V, diagnostic statistics show that a NRE a model is better than the fixed effects model, and the DPM model is the best among four models. Compared to the fixed and NRE model, an important feature of the DP and DPM models is to classify centers into subgroups (clusters), and centers in the same cluster have similar performance (similar $\beta_i's$). Based on the statistical measures described in Section 2.6, we present discovery steps in the succeeding text for the detection of outlying subgroups of centers.

Figure 3 depicts outlying centers at two tails, that is, centers with very low and high percentiles and small probabilities of being in the same cluster as other centers. In both DP and DPM models, two transplant centers (with id 116 and 178) have the worst outcomes, that is, the two centers have high ranks (percentile>0.8) and having small probabilities ($PS<0.2$) of being assigned to the same cluster as other centers. It is also interesting to note that a few centers with exceptionally good performance are observed in the DPM model but not in the DP model.

Table V. Diagnosis statistics for the kidney transplant data.

Model	DIC_3	WAIC	LPML
Fixed effects	41752	41834	−20787
Normal random effects	41718	41764	−20788
DP	41755	41793	−20803
DPM	**41692**	**41723**	**−20781**

DIC_3, deviance information criterion; WAIC, Watanabe–Akaike information criterion; LPML, log-pseudo marginal likelihood; DP, Dirichlet process; DPM, Dirichlet process mixture.



(a) (b)

Figure 3. The x-axis is the mean percentile, and the y-axis is the mean percentage of centers being in the same cluster as center i for the kidney transplant data using the Dirichlet process model (a) and Dirichlet process mixture model (b). Isolated data points in the lower left (right) corner depict outlying centers with exceptionally good (poor) performances.

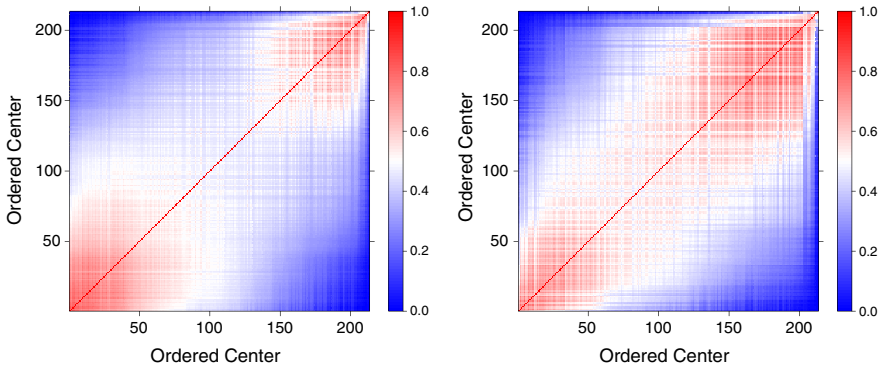


Figure 4. Heat maps representing pairwise posterior probabilities of the two centers are classified into the same cluster when $H = 5$; centers are ordered based on their mean ranking scores. (a) and (b) are from the DP and DPM model, respectively.

Next, we illustrate a way to detect outlying groups of centers using the heat map (Figure 4). This figure is based on pairwise probabilities and ordered by *rank* statistics as described in Section 2.6. Using a threshold of 0.5 on the probabilities, m centers form a cluster if all probabilities in the $m \times m$ matrix are larger than 0.5. At the upper right corner, centers 178 and 116, in both the DP and DPM models, appear to form a cluster (pairwise probability is 0.63 in DP and 0.71 in the DPM model), which performs significantly worse than the population average; at the lower left corner, 71 centers in the DP model and 24 centers in the DPM model appear to form a cluster that performs above average (all pairwise probabilities in the 71×71 and 24×24 matrix are larger than 0.5). Additionally, another cluster seems to arise that performs better than the previously mentioned small outlying cluster at the upper right corner but still worse than the population average (37 such centers in the DP model and 57 in the DPM model). Compared to the DP model, the DPM model has more centers that do not clearly belong to a unique cluster, as demonstrated by many overlapping (light) red squares along the 45 degree diagonal.

We also did sensitivity analysis for the parameters in the model. In this application, we tried a larger H ($H = 20$ and $H = 50$) and considered a random a. We find that a large H does not improve the model performance (DIC_3, WAIC, and LPML are the same as fixing $H = 5$), and data seem to contain little information about estimating the parameter a, leading to the same, or slightly worse diagnostic statistics, compared to a fixed a (results not shown). We also increased the A to 5 in the half-cauchy prior; this weaker prior provides similar model performance statistics as with $A = 1$ (DIC_3, WAIC, and LPML are similar).

5. Discussion

We proposed a nonparametric Bayesian approach to model random transplant center effects using the DP and DPM models. Random center effects are estimated using either a model of mixture of points (DP) or a model of mixture of normals (DPM). Compared to parametric frailty models, the proposed DPM (DP) model is capable of classifying centers into subgroups (clusters), and centers within the same cluster have similar (same) estimates of center effects. The work has been applied to evaluate long-term mortality rates of kidney transplant centers from a national organ failure registry. In a Cox proportional hazards model, we update patient-level covariates effects and risk-adjusted center effects in a fully Bayesian framework while classifying centers with different performance. Using statistical measures and graphical tools (such as rank statistics and heat maps), we first depict outlying individual centers with exceptional good and poor performance and then make use of the clustering feature of the model to detect outlying groups of centers with unusual outcomes. In the transplant study, we found that a small cluster of two centers has the worst performance and a bigger cluster of about 30 centers has the best performance. These findings will help policy makers make detailed inspection of the outlying groups and propose strategies to improve the performance.

To improve the model performance and interpretability for the survival data, we center the DP. Without centering, MCMC chains exhibit very high autocorrelation, which has no hope of yielding any meaningful estimates. Centering the DP process by constraining the mean to zero dramatically improved model

convergence and interpretability of the estimates of the center effects. To increase model's classification ability, we introduce a tempering parameter in calculating the allocation probability and observe that $\eta = 2$ is a good choice in studied cases.

During the study, we also tried to implement the proposed method in WinBUGS and JAGS. However, it is not straightforward to incorporate the parameter η in the modeling. We will continue working on it, and meanwhile, the R codes will be made available to the public through the author's web site once it is published.

Appendix

The Markov chain Monte Carlo procedure for estimating the posterior distributions is implemented by repeatedly drawing samples from the full conditional distributions of the parameters.

(1) The vector of covariates was divided into three groups with six covariates per group, and α was updated by groups. Within each group, the corresponding α was updated using the adaptive Metropolis–Hastings algorithm [30]. The initial estimates of α was calculated from a Cox model stratified by centers. The multivariate normal proposal density centered at the previous value, and the covariance in the proposal, was 'refined' by using the empirical covariance from an extended burn-in period.

(2) Update baseline hazard in interval $k(k = 1, \cdots, 5)$ from Gamma$(1\times0.1+D_k, 0.1+\sum_{i\in R_k} exp\{\alpha x_{ij}+\beta_i\}\Delta_{ijk})$, and D_k and R_k represent the number of death and the number subjects at risk in interval k, respectively.

Updates specific to DP model:

(3) Update cluster indicator Z_i as specified in Section 2.3.

(4) Update the stick-breaking weights from conditionally conjugate beta posterior distributions:

$$V_h|- \sim \text{Beta}\left(1 + \sum_{i=1}^{N} \text{I}(Z_i = h), a + \sum_{i=1}^{N} \text{I}(Z_i > h)\right), h = 1, \cdots, H$$

(5) Given the centers with labels specific to cluster h, update β_h by the adaptive rejection algorithm and $\beta_h \sim N(0, 100)$ as a priori [9, 31].

(6) Update a from a Gamma distribution.

$$a \sim \text{Gamma}\left(1, a_0 + H - 1, b_0 - \sum_{r=1}^{H-1} \log(1 - V_r)\right) \text{I}(0.3, 10)$$

The prior for a is gamma with hyperparameters a_0 and b_0, which are constrained in the range from 0.3 to 10.

(7) In DP-HP, given $\beta_1, \cdots, \beta_h, \cdots, \beta_H$, update (μ_0, σ_0^2) using the normal-inverse gamma conjugacy form Carlin and Louis[32].

Compared to the DP model, there are some changes in the DPM model,

(8) Update β_i by the adaptive rejection algorithm and with $\beta_i \sim \left(\mu_{Z_i}, \sigma_{Z_i}^2\right)$ as a priori.

(9) Gibbs sampling of cluster-specific parameters and hyperparameters in G_0 using the half-cauchy prior can be found in [20].

Acknowledgements

The authors gratefully acknowledge the constructive comments of referees and thank Dr. David Dunson for helpful discussions.

References

1. Conway PH, Clancy C. Transformation of health care at the front line. *JAMA* 2009; **301**:763–765.
2. Liu J, Louis TA, Pan W, Ma JZ, Collins AJ. Methods for estimating and interpreting provider-specific standardized mortality ratios. *Health Serv Outcomes Res Methodol* 2003; **1754**:135–149.
3. Jones HE, Spiegelhalter DJ. The identification of unusual heath-care providers from a hierarchical model. *Lifetime Data Anal* 2011; **65**:154–163.

4. Kalbfleisch JD, Wolfe RA. On monitoring outcomes of medical providers. *Stat Biosciences* 2013; **40**:1–30.
5. He K, Kalbfleisch JD, Li Y, Li Y. Evaluating hospital readmission rates in dialysis facilities with and without adjusting for hospital effects. *Lifetime Data Anal* 2013; **19**:490–512.
6. Spiegelhalter D, Sherlaw-Johnson C, Bardsley M, Blunt I, C CW, Grigg O. Statistical methods for healthcare regulation: rating, screening and surveilliance. *J Royal Stat Soc A* 2012; **175**:1–47.
7. Ohlssen DI, Sharples LD, Spiegelhalter DJ. Flexible random-effects models using bayesian semi-parametric models: applications to institutional comparisons. *Statistics in Medicine* 2007; **26**:2088–112.
8. Jara A, Hanson T, Quintana FA, Muller P, Rosner GL. Ppackage: Bayesian semi and nonparametric modeling in R. *Journal of Statistical Software* 2011; **40**:1–30.
9. Ibrahim JG, Chen MH, Sinha D. *Bayesian survival analysis*. Springer: New York, 2001.
10. Sethuraman J. A constructive definition of Dirichlet priors. *Statistica Sinica* 1994; **4**:639–650.
11. Dunson D. *Nonparametric Bayes Applications to Biostatistics*. Cambridge: Cambridge University Press, 2010.
12. Escobar MD, West M. Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association* 1995; **90**:577–588.
13. Ishwaran H, James LF. Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association* 2001; **101**:179–194.
14. Ishwaran H, Zarepour M. Markov chain Monte Carlo in approximate Dirichlet and beta two-parameter process hierarchical models. *Biometrika* 2000; **87**:371–390.
15. Hofmann T. Unsupervised learning by probabilistic latent semantic analysis. *Machine Learning* 2001; **42**:177–196.
16. Dunson D, Das S. *Bayesian distribution regression via augmented particle sampling*, Department of Statistical Science, Duke University, 2009.
17. Lee JJ, Gu X, Liu S. Bayesian adaptive randomization designs for targeted agent development. *Clinical Trials* 2010; **7**: 584–596.
18. Yang M, Dunson DB, Baird D. Semiparametric Bayes hierarchical models with mean and variance constraints. *Computational Statistics and Data Analysis* 2010; **54**:2172–2186.
19. Yisheng Li PM, Lin X. Center-adjusted inference for a non-parametric Bayesian random effect distribution. *Statistica Sinica* 2011; **21**:1201–1223.
20. Gelman A. Prior distributions for variance parameters in hierachical models. *Bayesian Analysis* 2006; **1**:515–533.
21. Richardson S, Green P. On Bayesian analysis of mixtures with an unknown number of components. *Journal of the Royal Statistical Society, Series B* 1997; **59**:731–792.
22. Kalbfleisch JD. Non-parametric Bayesian analysis of survival time data. *Journal of the Royal Statistical Society. Series B* 1978; **40**:214–221.
23. Dunson DB, Xue Y, Carin L. The matrix stick-breaking process: flexible Bayes meta analysis. *Journal of the American Statistical Association* 2008; **103**:317–327.
24. Bender R, Augustin T, Blettner M. Generating survival times to simulate Cox proportional hazards models. *Statistics in Medicine* 2005; **24**:1713–1723.
25. Celeux G, Forbes F, Robert CP, Titterington DM. Deviance information criteria for missing data models. *Bayesian Analysis* 2006; **1**:651–674.
26. Watanabe S. Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research* 2010; **11**:3571–3594.
27. Spiegelhalter DJ, Best NG, Carlin BP, de Linde AV. Bayesian measures of model complexity and fit (with discussion). *Journal of the Royal Statistical Society* 2002; **64**:583–639.
28. Gelman A, Carlin JB, Stern HS, Dunson DB, Vehtari A, Rubin DB. *Bayesian data analysis* 3nd edition. Chapman & Hall/CRC: Boca Raton, 2013.
29. Zhao L, Feng D, Bellile LE, Taylor JMG. Bayesian random threshold estimation in a Cox proportional hazards cure model. *Statistics in Medicine* 2014; **33**:650–661.
30. Haario H, Saksman S, Tamminen J. An adaptive Metropolis algorithm. *Bernoulli* 2001; **7**:223–242.
31. Gilks WR, Best NG, Tan KKC. Adaptive rejection Metropolis sampling within Gibbs sampling. *Applied Statistics* 1995; **44**:455–472.
32. Carlin B, Louis T. *Bayes and empirical Bayes methods for data analysis.* 2nd edition. Chapman & Hall/CRC: Boca Raton, 2000.