Assessing treatments effects in multi-center clinical trials with Application to Scleroderma Lung Study: A Semiparametric Bayesian Approach

Man-Wai Ho[*], Pulak Ghosh[†], Robert M. Elashoff[‡] and Ram C. Tiwari[§]

September 11, 2009

[*]Assistant Professor, Department of statistics and Applied Probability, National University of Singapore, 6 Science Drive 2, Singapore 117546 (Email: stahmw@nus.edu.sg)

[†]Associate Professor, Department of Quantitative Management and Information Sciences, Indian Institute of Management, Bannerghatta road, Bangalore 560076 India (Email: pulakghosh@gmail.com)

[‡]Professor, Department of Biostatistics, School of Public Health, University of California at Los Angeles, Los Angeles, CA, USA.

[§]Associate Director, Office of Biostatistics, Center for Drug Evaluation & Research, FDA (Email:ram.tiwari@fda.hhs.gov) The views expressed in this article are those of this author and not those of the United States Food and Drug Administration

Abstract

In this article we explore the testing of non-inferiority and equivalence hypotheses arising from multiple centers when the assumption of normality is violated. In a multi-center study, the trials are typically conducted at different centers which vary in terms of location, environment, demographics among others, leading to substantial amount of heterogeneity in the patient population. This unexplained variation in a multi-center clinical study is usually modeled using a random effects model, where the centers are assumed to be a random sample from the population of centers. Most research in this direction uses a parametric normal distribution which can be restrictive and may lead to biased result if the actual distribution is nonnormal. In this article, we overcome this parametric assumption by considering a broader class of random effects distribution for the centers. In particular, we develop a novel nested Dirichlet process (nDP) model to explore the sensitivity of the fixed treatment effects under various hypotheses, in the presence of nonnormality. Additional advantage of our proposed method is that it facilitates a hierarchical clustering structure. At one hand it clusters the centers according to their effects, and hence outlying centers can be identified. Simultaneously, subjects from the clustered centers are again clustered together enabling a borrowing of information across similar centers. Further, we present the methodology to test between the models with nDP versus a normal random center effects models. We discuss the results of our proposed methodology in a real example of a multi-center clinical trial on Scleroderma lung study. The results of the analysis along with the extensive simulation study show the advantage of our method when the center effects distribution is not normal.

Keywords: treatment effect; nested Dirichlet process; multi-center; Scleroderma lung study;

1 Introduction

A common goal in clinical trial is to compare several treatments conducted quite often at different centers. Generally, multi-center trials are designed with the objective of demonstrating an overall treatment effect from the combined contributions of all centers. Multi-center trials are thus very common in the field of drug development. The ICH E9 (1998) guidance outlines two main reason for the popularity of multi-center trials. First, it helps to enroll required number of patients in a time bound fashion. Second, multi-center trials provide a better basis for the generalization of the findings as it represents a broader class of patient populations. As noted by Freeman (1998) multi-center trials consist of many sources of variability due to various factors, viz, location, environment, demographics, etc. Due to this heterogeneity in multi-center trials there are two major sources of variation in treatment response that can be accounted for (Anello et al., 2005): the variation within and between centers. To account for these variability several researchers assumed a random center effects model to capture the heterogeneity inherent among different centers. Traditionally, a parametric normal distribution are assumed for these random center effects. Since the particular distributions of these latent effect measures can have an impact on conclusions of the trial, routine use of normal distribution would be rather a strong assumption (Higgins et al., 2009). In this article, we review and illustrate the danger of using a normal distribution in the absence of proper justification. To protect the model from distributional misspecifications, we develop a broader class of flexible nonparametric distribution using the recently developed nested Dirichlet Process (nDP; Rodríguez et al., 2008).

There has been a wide amount of literature of the mixed model approach to multi-center clinical trials with fixed treatment effects and random center effects. See Patel (2002) for a review. Some other work in a similar direction are in Khatri and Patel (1992), Rashid (2003), Thompson (1994) and Gould (2005). Most of the existing methods assumes a normal distribution. While this rather strong assumption makes the model easy to apply in widely used softwares such as SAS, the accuracy of this assumptions is difficult to check and the routine use of normality in mixed model is routinely questioned by many authors (Rashid, 2003; Ohlssen et al., 2007; Branscum et al., 2008; Higgins et al., 2009). Normality assump-

tion is too restrictive as it suffers from the lack of robustness particularly when the effects across centers show multi-modality and/or skewness, and thus may not provide an accurate estimation of between–center variation. Furthermore, inference on individual center effects can be misleading when the random center effects distribution deviates from normality. The ICH E9 addresses the issue and possible effects of having outliers in multi-center trials. Thus, it is of practical interest to develop statistical model with considerable flexibility in the distributional assumptions of the random effects as well as measurement error. Rashid (2003) developed a rank-based procedure for testing a non-inferiority and equivalence hypothesis for multi-center trials using mixed model. The R estimates are obtained by minimizing a sum of Jackel (1972) type dispersion functions based on intra-center ranks of residuals. However, this method has too much reliance on the central limit theorem and thus may not be realistic when there are fewer studies. Recently Lee and Thompson (2007) used a skewed distribution to reduce the effects of outlying centers, and a mixture distribution have been advocated to account for studies belonging to unknown groupings (Bohning, 2000). Although the use of a heavier-tailed distribution such as t-distribution provides some robustness, it may not be sufficient to represent the actual distribution of effects. For example, even a heavy-tailed distribution, such as the t, has a unimodal and a symmetric shape and restrictive in the sense that it fails to allow multi-modality, which may arise due to latent sub-populations. Bayesian semiparametric approach offers a useful alternative in this direction. There have been few work (Burr et al., 2003; Burr and Doss, 2005; Ohlssen, et al., 2007) on using a Dirichlet process (DP) prior in a multi-center clinical trials. A Dirichlet process consists of a control parameter and a baseline distribution which can be normal. A discrete mass points are drawn from this baseline distribution and how close the discrete distribution is to the baseline depends on the value of the control parameter. Thus the fitted random effects distribution using DP is flexible enough and has the potential to be robust to departures from a normal distribution while having good performance if the actual distribution is normal. Recently, Branscum et al. (2007) developed a Pólya tree method in a meta-analytic framework.

We consider a broader class of random effects distribution for the centers. In particular, we develop a novel nested Dirichlet process (nDP) model to explore the sensitivity of the

fixed treatment effects under various hypotheses, in the presence of non-normality. Additional advantage of our proposed method is that it allows a hierarchical clustering structure, whereby the centers clusters according to similarity of their effects, and hence outlying centers can be identified, and at the same time subjects from the clustered centers are also cluster borrowing information from similar centers. Further, we present the methodology to test nDP model versus a normal random center effects model. As mentioned, although semiparametric Bayesian models have been previously used in multi-center clinical trial data, to our knowledge this is the first systematic attempt to use the nDP for this kind of mixed model.

1.1 Motivating Data: Scleroderma Lung study

Our method is primarily motivated by the Scleroderma lung study (Tashkin et al., 2006), which is a double blinded, randomized clinical trial. The aim of the trial was to evaluate effectiveness of oral cyclophosphamide (CYC) versus placebo in the treatment of lung disease due to scleroderma. Scleroderma is an autoimmune connective-tissue disorder that is characterized by microvascular injury, excessive fibrosis of the skin, and distinctive visceral changes that can involve the lungs, heart, kidneys, and gastrointestinal tract. A number of agents have been evaluated as treatments for scleroderma-related interstitial lung disease, but none have been proven effective. Only CYC has shown promise in slowing down the decrease or even improve the forced vital capacity (FVC) over time. In this study our primary outcome is forced vital capacity (FVC), as percentage predicted) determined a 3-month intervals from baseline. At 13 clinical centers throughout the United States, the study enrolled 158 patients with scleroderma, restrictive lung physiology, dyspnea, and evidence of inflammatory interstitial lung disease on examination of bronchoalveolar-lavage fluid, thoracic high resolution computed tomography, or both. Patients received oral CYC ($\leq$ 2 mg per kilogram of body weight per day) or matching placebo for one year and were followed for an additional one year. Pulmonary function was assessed in every three months.

We are interested in evaluating whether oral CYC can either improve %FVC scores. The study enrolled 158 patients with scleroderma-related interstitial lung disease, who were randomized to receive either CYC ($2mg/kg$) or identical-appearing placebo for 18 months.

Since the study was conducted across 13 centers, it is important to assess the treatment effect when adjusted for the random center effects. Thus, in this paper we develop a model to test the effectiveness of the treatment CYC over placebo and assume a nDP for the random center effects. One of the scientific interest is to find the centers whose patients populations behave similarly. Out of the 153 patients, 145 completed at least six months of treatment and were included in the analysis.

The rest of the paper is organized as follows. In Section 2, we present the basic model, normal random center effects, the nDP preliminaries and state the hypothesis of interest. Section 3, gives the nDP model as a generalization of the basic random effects model, and Section 4 gives the posterior distributions of the parameters. Section 4 describes the simulation study and Section 5 described the analysis of the data from Scleroderma Lung Study. Section 6 have the discussion.

2 Background

2.1 Basic Model

In the following we describe the basic model with the existing normality assumption to put our new model in perspective. Let Y_{ijt} be the response of the i-th subject from j-th center under t-th treatment; $i = (1, 2, \cdots, n_j)$, $j = (1, 2, \cdots, C)$, $t = (1, 2, \cdots, T)$. Rashid et al. (2003) assume the following normal random center effects model (without covariate) for multi-center clinical trials:

$$Y_{ijt} = \theta_t + \beta_j + e_{ijt} \tag{1}$$

where θ_t is the fixed t-th treatment effect, β_j is the j-th center effect, $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_r)^\top$ is a $r \times 1$ dimensional vector of coefficients associated with $r \times 1$ dimensional vector of covariates $\mathbf{w}_{ij} = (w_{ij1}, w_{ij2}, \ldots, w_{ijr})^\top$ and e_{ijt} are the random measurement errors. The basic assumptions in (1) is that the random center effects and the errors are independent, and are both normally distributed with zero means.

Although in model (1), the basic assumption of the random center effects is Gaussian, as we discussed this assumption is questionable and inference can be biased under possible

misspecification of this normality assumption. Thus, the problem we address in this article is to broaden the class of distribution of the random center effects. In particular, we assume that the center effects come from some unspecified distributions. This allows more flexibility and robustness in the modeling of the observations from different centers when they do not seem to have come from a common distribution, but may have come from a mixture of normal distributions, a distribution with heavier tails, or from some other distributions which cannot be easily specified. Use of nDP is a robust generalization as it has the potential to capture these departures from a normal distribution while having good performance if the actual distribution is normal.

Based on the above model (1) an important question is to assess the efficacy of the treatment effect by pooling the data across the centers. There are some comments in this regard in the ICH E9 guidance (1999) which is described in detail and Annello et al. (2005). Based on these documents, there are two main categories of hypotheses in assessing treatment efficacy: one is testing equivalence between treatments, where the null hypothesis is that the difference between the active comparator and new drug is within a pre-specified limit; while the other one is testing for non-inferiority where the aim is to show that the new drug is not less effective than the control by more than a defined margin.

Although not exhaustive, we list the following three potential hypotheses:

a) $H_0 : \theta_1 = \theta_2 = \cdots = \theta_T = 0 \ \Rightarrow$ Equality

This is a general hypothesis in multi-center trials. Accepting H_0 in this hypothesis implies that there is no significant treatment effect in the study and thus treatments are not heterogeneous.

b) $H_{0t} : -\Delta < \theta_t - \theta_{t'} < \Delta \ \Rightarrow$ Equivalence

This hypothesis assesses the equivalence of any two treatment (including placebo) within a given range.

c) $H_{0t} : \theta_t \leq \theta_{t+1} - \Delta_0 \ \Rightarrow$ Non-inferiority

Accepting this hypothesis demonstrates a new treatment is not worse than an active control by more than a specified margin.

2.2 The nested Dirichlet process

Since Ferguson (1973) described the Dirichlet process (DP) as a random probability measure that can be viewed as a distribution on distributions, use of DP has become popular in the literature of nonparametric Bayes estimation (see, for example, Antoniak 1974; Lo 1978, 1984; Escobar 1988, 1994; Escobar and West 1995, Ghosh et al., 2009). Let $\texttt{DP}(\alpha H)$ denote a DP with base measure H and precision $\alpha > 0$. Replacing H by another DP, Rodríguez (2007) and Rodríguez et al. (2008) introduced the nested Dirichlet process (nDP), which, from a similar perspective, can be characterized as *a distribution on the space of distributions on distributions*. The nDP provides a framework to model collections of dependent distributions utilizing clustering features of the DP. A collection $\{F_j, j = 1, \ldots, C\}$ of distributions on any complete and separable metric space Θ such that $F_j \sim Q$ with $Q \equiv \texttt{DP}(\alpha\texttt{DP}(\rho H))$, for $\alpha, \rho > 0$ and H being a probability measure on Θ, is said to follow a nDP. Write $\{F_1, \ldots, F_C\} \sim \texttt{nDP}(\alpha, \rho, H)$. The stick-breaking characterization of the DP (Sethuraman 1994; Sethuraman and Tiwari 1982) implies that

$$F_j(\cdot) \sim Q \equiv \sum_{k=1}^{\infty} \pi_k^* \delta_{F_k^*(\cdot)} \tag{2}$$

and

$$F_k^*(\cdot) \equiv \sum_{l=1}^{\infty} \omega_{lk}^* \delta_{\beta_{lk}^*}(\cdot), \tag{3}$$

where $\beta_{lk}^* \overset{\text{iid}}{\sim} H$,

$$\omega_{lk}^* = u_{lk}^* \prod_{s=1}^{l-1}(1 - u_{sk}^*), \qquad u_{lk}^* \sim \texttt{beta}(1, \rho)$$

and

$$\pi_k^* = v_k^* \prod_{s=1}^{k-1}(1 - v_s^*), \qquad v_k^* \sim \texttt{beta}(1, \alpha),$$

with $\texttt{beta}(a, b)$ representing a beta probability distribution with parameters a and b on the $(0, 1)$ interval. The nDP naturally induces clustering in the space of distributions as a consequence of the almost surely discreteness feature of Q, illustrated by (2). Specifically, there is a non-zero probability $\mathbb{P}(F_j = F_{j'}|H) = 1/(1 + \alpha)$ that two distributions F_j and $F_{j'}$ follow the same random distribution F_k^* defined by (3). Furthermore, the nDP enables clustering between samples from the distributions in the collection. That is, samples from

one single F_j, or from F_j and $F_{j'}$, $j \neq j'$, are correlated, and possibly identical. As F_k^* is almost surely discrete as defined in (3), samples β_{ij} and $\beta_{i'j}$ from F_j may be identical to some β_{lk}^* if $F_j = F_k^*$, while the correlation is given by $\text{corr}(\beta_{ij}, \beta_{i'j}) = 1/(1+\rho)$. In analogy, respective samples β_{ij} and $\beta_{i'j'}$ from two different distributions F_j and $F_{j'}$, $j \neq j'$, may be identical to some β_{lk}^* if $F_j = F_{j'} = F_k^*$, while the correlation can be shown to be $\text{corr}(\beta_{ij}, \beta_{i'j'}) = 1/[(1+\alpha)(1+\rho)]$, which is always less than the correlation $1/(1+\rho)$ between two samples from the same F_j. See more discussion on nDP in Rodríguez (2007) and Rodríguez et al. (2008).

3 nDP Model and Methods

Here we generalize the model as described in (1) with covariates and use nDP to better model the heterogeneity among centers. We assume that

$$Y_{ijt} = \theta_t + \beta_{ij} + \mathbf{w}_{ij}^\top \boldsymbol{\gamma} + e_{ijt}, \tag{4}$$

where β_{ij} denotes the effect of the i-th subject at the j-th center, thus allowing for a nested subject effect, and $e_{ijt} \overset{\text{iid}}{\sim} \mathtt{N}(0, \tau^{-1})$. Generalizing the normality assumption for the center effects in (1), we assume that

$$(\beta_{ij}|F_j) \overset{\text{ind}}{\sim} F_j, \qquad j = 1, \ldots, C, i = 1, \ldots, n_j, \tag{5}$$

$$(\{F_1, \cdots, F_C\}|\alpha, \rho, H) \sim \mathtt{nDP}(\alpha, \rho, H), \qquad \text{with } H = \mathtt{N}(0, \sigma_\beta^2). \tag{6}$$

This formulation for the center effects has the following interpretations:

(i) (Heteroscedasticity) Different subjects i in different centers j may be influenced by different center effects.

(ii) (Exchangeability) For different subject i in the same center j, the center effects β_{ij} are independent and identically distributed for all treatments $t = 1, \ldots, T$.

(iii) Centers are clustered according to their effects on the response, and hence, outlying centers can be identified.

(iv) Simultaneously, subjects from similar centers are clustered together according to the effects attributed by centers. That is, being clustered together this allows borrowing information across centers that are similar.

We assumed standard choices for the prior distributions. Hence, given the covariates $(\mathbf{w}_{11}, \ldots, \mathbf{w}_{n_1 1}, \ldots, \mathbf{w}_{1C}, \ldots, \mathbf{w}_{n_C C})$, the observed data $\mathbf{Y} = (Y_{11t}, \ldots, Y_{n_1 1t}, \ldots, Y_{1Ct}, \ldots, Y_{n_C Ct})$ are assumed to satisfy the following hierarchical model hereafter referred to as *the nDP model*.

$$
\begin{aligned}
(Y_{ijt}|\theta_t, \beta_{ij}, \boldsymbol{\gamma}, \tau) &\overset{\text{ind}}{\sim} \mathtt{N}(\theta_t + \beta_{ij} + \mathbf{w}_{ij}^{\top}\boldsymbol{\gamma}, \tau^{-1}), && j = 1, \ldots, C, i = 1, \ldots, n_j, \\
\theta_t &\overset{\text{iid}}{\sim} \mathtt{N}(0, \sigma_\theta^2), && t = 1, \ldots, T, \\
(\beta_{ij}|F_j) &\overset{\text{iid}}{\sim} F_j, && i = 1, \ldots, n_j \\
(\{F_1, \cdots, F_C\}|\alpha, \rho, H) &\sim \mathtt{nDP}(\alpha, \rho, H) && \text{with } H = \mathtt{N}(0, \sigma_\beta^2), \\
\alpha &\sim \mathtt{gamma}(a_\alpha, b_\alpha), \\
\rho &\sim \mathtt{gamma}(a_\rho, b_\rho), \\
\boldsymbol{\gamma} &\sim \mathtt{MN}_r(\mathbf{0}, \Sigma_{\boldsymbol{\gamma}}), \\
\tau &\sim \mathtt{gamma}(a_\tau, b_\tau),
\end{aligned}
\tag{7}
$$

where $\sigma_\theta, \sigma_\beta, a_\alpha, b_\alpha, a_\rho, b_\rho, a_\tau, b_\tau$ are fixed positive constants, and $\Sigma_{\boldsymbol{\gamma}}$ is a known $r \times r$ variance-covariance matrix $\mathtt{MN}_r(\mathbf{0}, \Sigma_{\boldsymbol{\gamma}})$ represents an r-variate normal distribution with zero mean vector and covariance matrix $\Sigma_{\boldsymbol{\gamma}}$, and $\mathtt{gamma}(a, b)$ denotes a gamma distribution with shape parameter a and scale parameter b such that its mean is a/b.

As a consequence of the unique characterization of a DP in terms of the Pólya urn distribution of Blackwell and MacQueen (1983) the posterior distribution of the above nDP model can be represented in the form of a hierarchy of two layers, in which there is a Pólya urn distribution in each layer and the Pólya urn distribution at the top layer depends on that at the bottom layer. Though there exist explicit expressions for the two Pólya urns, handling of two such nested Pólya urns turns out to be quite cumbersome due to their complicated dependence structure, resulting in extreme difficulties implementing the Pólya urn Gibbs sampler in Escobar (1988, 1994), which is one of the most popular Markov chain Monte Carlo method for sampling from the posteriors in nonparametric models involving DP, for computations of posterior quantities in this model. For the explicit expressions of the Pólya urns for a nDP, one may refer to Rodríguez (2007).

4 Posteriors

Following Rodríguez (2007) and Rodríguez et al. (2008), we replace the stick-breaking representations of the DP priors for both F_j and F_k^* given in (2) and (3) by their almost sure truncation approximations which are finite sums of K and L elements, respectively. That is,

$$F_j(\cdot) \approx \sum_{k=1}^{K} \pi_k^* \delta_{F_k^*(\cdot)}, \tag{8}$$

where $\pi_k^* = v_k^* \prod_{s=1}^{k-1}(1 - v_s^*)$ with $v_k^* \sim \texttt{beta}(1, \alpha)$, for $k = 1, \ldots, K-1$, and $v_K^* = 1$, and, for $k = 1, \ldots, K$,

$$F_k^*(\cdot) \approx \sum_{l=1}^{L} \omega_{lk}^* \delta_{\beta_{lk}^*}(\cdot), \tag{9}$$

where $\beta_{lk}^* \overset{\text{iid}}{\sim} H$, $\omega_{lk}^* = u_{lk}^* \prod_{s=1}^{l-1}(1 - u_{sk}^*)$ with $u_{lk}^* \sim \texttt{beta}(1, \rho)$, for $l = 1, \ldots, L-1$, and $u_{Lk}^* = 1$. Write $\boldsymbol{\pi}^* = (\pi_1^*, \ldots, \pi_K^*)$, $\boldsymbol{\omega}^* = (\omega_{11}^*, \ldots, \omega_{L1}^*, \ldots, \omega_{1K}^*, \ldots, \omega_{LK}^*)$, and $\boldsymbol{\beta}^* = (\beta_{11}^*, \ldots, \beta_{L1}^*, \ldots, \beta_{1K}^*, \ldots, \beta_{LK}^*)$.

The finite dimensionalities of (8) and (9) allow us to express the Bayesian semiparametric model (7) entirely in terms of a finite number of random variables. Because of the nature of their prior distributions, these random variables can be drawn from some standard multivariate distributions. We assign them into the following four groups or *blocks of parameters*, namely, $(\boldsymbol{\zeta}, \boldsymbol{\xi}, \boldsymbol{\pi}^*, \boldsymbol{\omega}^*, \boldsymbol{\beta}^*, \alpha, \rho)$, $\boldsymbol{\gamma}$, τ, and $(\theta_1, \ldots, \theta_T)$, where $\boldsymbol{\zeta}$ and $\boldsymbol{\xi}$ are two vectors of classification variables describing clustering behavior of the center effects defined as follows. Let $\boldsymbol{\zeta} = (\zeta_1, \ldots, \zeta_C)$ denote a classification vector describing the center effects by setting $\zeta_j = k$, for $k = 1, \ldots, K$, if and only if the center effect for the j-th center, for $j = 1, \ldots, C$, β_{ij} is distributed as $F_j = F_k^*$. Furthermore, define classification variables ξ_{ij}, for $j = 1, \ldots, C$ and $i = 1, \ldots, n_j$, and set $\xi_{ij} = l$, for $l = 1, \ldots, L$, if and only if the center effect for the observation Y_{ijt} is given by $\beta_{ij} = \beta_{l\zeta_j}^*$. As shown below, the knowledge of these two vectors of classification variables provide an equivalent expression of the likelihood of the observed data. According to model (7), the likelihood of the observed response $Y_{ijt} = y_{ijt}$ from the i-th subject, receiving t-th treatment in j-th center, which is associated with a covariate vector $\mathbf{w}_{ij}$, is denoted by

$$f(y_{ijt}|\theta_t, \beta_{ij}, \boldsymbol{\gamma}, \tau, \mathbf{w}_{ij}) = \sqrt{\frac{\tau}{2\pi}} \exp\left[-\frac{\tau}{2}(y_{ijt} - \theta_t - \beta_{ij} - \mathbf{w}_{ij}^\top \boldsymbol{\gamma})^2\right]. \tag{10}$$

It can be equivalently expressed as

$$f(y_{ijt}|\theta_t, \beta^*_{\xi_{ij},\zeta_j}, \boldsymbol{\gamma}, \tau, \mathbf{w}_{ij}) = \sqrt{\frac{\tau}{2\pi}} \exp\left[-\frac{\tau}{2}(y_{ijt} - \theta_t - \beta^*_{\xi_{ij},\zeta_j} - \mathbf{w}_{ij}^\top \boldsymbol{\gamma})^2\right]. \tag{11}$$

Generalizing the idea of the blocked Gibbs algorithm of Ishwaran and James (2001), based on this equivalence relation of the likelihood, an iterative algorithm (discussed in Appendix) cycling through four steps, in which each step draws one of the four desirable blocks of parameters conditioning on all the other variables, can be derived for sampling random variates of the four blocks of parameters from their joint posterior distribution for evaluating posterior estimates of any quantity of interest in the problem.

Implementation of the iterative algorithm for M, some large number, cycles results in a Markov chain of realizations of the four blocks of parameters. Suppose that a Markov sequence $(\theta_1^{(1)}, \ldots, \theta_T^{(1)}), \ldots, (\theta_1^{(M)}, \ldots, \theta_T^{(M)})$ is generated for the treatment effects $(\theta_1, \ldots, \theta_T)$. The posterior probabilities for different hypotheses about any relationship between the treatments, equality, equivalence, or non-inferiority, can be approximated by sample probabilities of the events of interest obtained from the posterior samples. For instance, the probability of equivalence of any two treatment $\mathbb{P}(-\Delta < \theta_t - \theta_{t'} < \Delta)$, for some small $\Delta > 0$, is approximated by

$$\frac{1}{M}\sum_{i=1}^{M} I_{\{-\Delta < \theta_t^{(i)} - \theta_{t'}^{(i)} < \Delta\}},$$

and the probability of non-inferiority of the t'-th treatment effect $\theta_{t'}$ to the t-th treatment effect θ_t, denoted by $\mathbb{P}(\theta_t \leq \theta_{t'} - \Delta)$, is approximated by

$$\frac{1}{M}\sum_{i=1}^{M} I_{\{(\theta_t^{(i)} \leq \theta_{t'}^{(i)} - \Delta)\}}.$$

For purpose of investigation of accuracy in the estimation or drawing prediction of any new observation, one can make use of density estimates of any observation y associated with t-th treatment in the j-th center and covariate vector $\mathbf{w}$, that are in general computed as

$$\frac{1}{M}\sum_{k=1}^{M} f_t^{(k)}(y|\mathbf{w}), \tag{12}$$

where $f_t^{(k)}$ is defined as in (10) according to posterior samples of the unknown parameters in the k-th iteration, denoted by $\theta_t^{(k)}, \beta_{ij}^{(k)}, \boldsymbol{\gamma}^{(k)}$, and $\tau^{(k)}$, for subjects associated with the t-th

treatment from the same center j. In particular, for the nDP model, suppose that in the k-iteration, $\{\beta^*_{\xi_{ij}^{(k)},\zeta_j^{(k)}}\}^{(k)}$ denotes the posterior draw of the center effect β_{ij} for the i-th observation in the j-th center according to classification variables $\xi_{ij}^{(k)}$ and $\zeta_j^{(k)}$ in the same iteration, $f_t^{(k)}(y|\mathbf{w})$ is given by

$$\frac{1}{N_j(t)}\sqrt{\frac{\tau^{(k)}}{2\pi}}\sum_{i=1}^{n_j}\exp\left\{-\frac{\tau^{(k)}}{2}\left(y-\theta_t^{(k)}-\{\beta^*_{\xi_{ij}^{(k)},\zeta_j^{(k)}}\}^{(k)}-\mathbf{w}^\top\boldsymbol{\gamma}^{(k)}\right)^2\right\}I_{\{A_t(i,j)\}},$$

where $I_{\{A_t(i,j)\}}$ is an indicator function for the event that the i-th observation in j-th center is associated with t-th treatment, and $N_j(t)\equiv\sum_{i=1}^{n_j}I_{\{A_t(i,j)\}}\leq n_j$ is the total number of observations associated with t-th treatment among all n_j observations in j-th center. For the normal model, $\beta_{ij}^{(k)}$ for all $i=1,\ldots,n_j$ are identical, say, denoted by $\beta_j^{(k)}$, then $f_t^{(k)}(y|\mathbf{w})$ equals

$$\frac{1}{N_j(t)}\sqrt{\frac{\tau^{(k)}}{2\pi}}\sum_{i=1}^{n_j}\exp\left\{-\frac{\tau^{(k)}}{2}(y-\theta_t^{(k)}-\beta_j^{(k)}-\mathbf{w}^\top\boldsymbol{\gamma}^{(k)})^2\right\}I_{\{A_t(i,j)\}},$$

which reduces to $f(y|\theta_t^{(k)},\beta_j^{(k)},\boldsymbol{\gamma}^{(k)},\tau^{(k)},\mathbf{w})$ as the summand is constant for any i and the total number of summands equals $N_j(t)$.

4.1 Model Comparison

To our knowledge, the random effects model (7) is the only direct generalization of the normal/Gaussian model considered in the literature, defined as in (1), using nDP. Thus, it is important to formally test the utility of nDP over simple normal model. However, developing a formal Bayes factor for this purpose can be tough as in general, it is difficult to compute a Bayes factor in any Bayesian nonparametric mixture model involving DP since exact evaluation of the marginal likelihood/density of the observations requires performing a multi-fold integration with respect to the Pólya urn distribution or calculation of a finite sum with total number of summands roughly of magnitude of the Bell's number. Basu and Chib (2003) proposed a non-iterative algorithm based on the collapsed sequential importance sampler developed in MacEachern et al. (1999), which is also discussed in the context of weighted Chinese restaurant processes by Lo et al. (1996), to approximate the latter sum. See also Hayakawa et al. (2002) who applied the same algorithm to evaluate a Bayes factor in Bayesian mixture hazard models involving gamma and weighted gamma processes. To

the best of our knowledge, no one has proposed any iterative algorithm for these proposes. Furthermore, since an extension of such an non-iterative algorithm for posterior inference of models involving nDP is not available yet, it is practically impossible to approximate the marginal likelihood of the nDP model (7), and thus, in turn, to evaluate a Bayes factor in the model.

However, the Bayes factor has several other potential problems (Gelfand and Dey, 1994), the most significant being numerical instability. Therefore we consider an alternative predictive measure of model performance, introduced by Geisser and Eddy (1979) as a predictive criterion termed the *log pseudo marginal likelihood* (LPML). LPML has been used extensively in problems of Bayesian model selection (see, for example, Chen et al. 2000, Chapter 10; Brown and Ibrahim 2003, Ghosh et al. 2009) as a useful summary statistic for comparing model fits. Models with greater LPML values represent a better fit. The LPML is defined based on estimates of the conditional predictive ordinate (CPO; Gelfand et al. 1992; Chen et al. 2000) for all observations

$$\text{LPML} = \sum_{j=1}^{C} \sum_{i=1}^{n_j} log\left(\widehat{\text{CPO}}_{ij}\right), \tag{13}$$

where

$$\widehat{\text{CPO}}_{ij} = \left[\frac{1}{M}\sum_{k=1}^{M} \frac{1}{f(y_{ijt}|\theta_t^{(k)}, \beta_{ij}^{(k)}, \boldsymbol{\gamma}^{(k)}, \tau^{(k)}, \mathbf{w}_{ij})}\right]^{-1},$$

with f defined in (10), is the estimate of the CPO for the i-th observation from j-th center. Specifically, for the nDP model,

$$\begin{aligned}
\widehat{\text{CPO}}_{ij}^{\text{nDP}} &= \left[\frac{1}{M}\sum_{k=1}^{M} \frac{1}{f(y_{ijt}|\theta_t^{(k)}, \{\beta^*_{\xi_{ij}^{(k)},\zeta_j^{(k)}}\}^{(k)}, \boldsymbol{\gamma}^{(k)}, \tau^{(k)}, \mathbf{w}_{ij})}\right]^{-1} \\
&= \frac{M}{\sqrt{2\pi}}\left[\sum_{k=1}^{M} \frac{1}{\sqrt{\tau^{(k)}}} \exp\left\{\frac{\tau^{(k)}}{2}(y_{ijt} - \theta_t^{(k)} - \{\beta^*_{\xi_{ij}^{(k)},\zeta_j^{(k)}}\}^{(k)} - \mathbf{w}_{ij}^\top \boldsymbol{\gamma}^{(k)})^2\right\}\right]^{-1}.
\end{aligned}$$

For the normal model, it takes the same form as $\widehat{\text{CPO}}_{ij}^{\text{nDP}}$ with $\{\beta^*_{\xi_{ij}^{(k)},\zeta_j^{(k)}}\}^{(k)}$ replaced by $\beta_j^{(k)}$.

5 Simulation Study

In this section, we present numerical examples designed to demonstrate the ability of the nDP model in providing accurate estimates for all of the fixed treatment effects, the random center effects, and the covariate effects. Simulation results based on the nDP model are obtained by implementing the introduced iterative algorithm with truncation levels in (8) and (9) set as $K = 35$ and $L = 55$, respectively. Posterior estimates of parameters and other quantities of interest are computed based on $M = 10,000$ samples taken from the Markov chain once every 5 iterations after discarding 50000 burn-in samples. Hyperprior parameters in (7) are set as follows unless otherwise stated: To deflate the priors, we set $\sigma_\theta = \sigma_\beta = \sigma_\gamma = 100$, and $a_\tau = b_\tau = 0.001$. Furthermore, we set $a_\alpha = b_\alpha = a_\rho = b_\rho = 3$, implying that $E(\alpha) = E(\rho) = 1$, which is a common choice in the literature, and $P(\alpha > 3) = P(\rho > 3) \approx 0.006$.

5.1 Simulated Data

Six different sets of simulated data are generated according to (4) based on the following set-up. There are $T = 2$ different treatments with known effects $\theta_1 = -\theta_2$, 4 different centers with random effects β_{ij} distributed according to different mixtures of known distributions, and the error term e_{ijt} follows a normal, or a Student's t distribution, or their mixtures. Except the last dataset, the number of independent observations from each center is given by $n_j = 50$, and hence, the sample sizes of all datasets are 200. For purpose of comparison, these datasets are also analyzed by the normal random center effects model, which differs from (7) with β_{ij} replaced by β_j, for $i = 1, \ldots, n_j$, and $F_j = H = \mathtt{N}(0, \sigma_\beta^2)$. This alternative model is referred here as the normal model.

In the first dataset, $\theta_1 = -\theta_2 = 0.5$, there are no covariates, and both β_{ij} and e_{ijt} follow mixtures of normals, where

$$\begin{array}{lcllll} \beta_{i1} & \sim & 0.6\mathtt{N}(0, 2^2) & + & 0.4\mathtt{N}(3, 1), & i = 1, \ldots, n_1, \\ \beta_{i2} & \sim & 0.5\mathtt{N}(0, 2^2) & + & 0.5\mathtt{N}(3, 1), & i = 1, \ldots, n_2, \\ \beta_{i3} & \sim & 0.8\mathtt{N}(5, 1) & + & 0.2\mathtt{N}(10, 1), & i = 1, \ldots, n_3, \\ \beta_{i4} & \sim & 0.8\mathtt{N}(5, 1) & + & 0.18\mathtt{N}(10, 1) + 0.02\mathtt{N}(-1, 2), & i = 1, \ldots, n_4, \end{array} \tag{14}$$

with all $n_j = 50$, and $e_{ijt} \sim 0.3\mathtt{N}(-2,1) + 0.4\mathtt{N}(0,1) + 0.3\mathtt{N}(2,1)$.

First, the posterior probability of equivalence of the two treatments, that is, $\theta_1 = \theta_2$, is approximated by the nDP model as

$$\frac{1}{M}\sum_{k=1}^{M} I_{\{-\Delta<2\theta_1^{(k)}<\Delta\}} = \begin{cases} 0.0005, & \Delta = 0.01, \\ 0.0039, & \Delta = 0.05, \end{cases}$$

providing strong evidence that the two treatments are not equivalent, where $\theta_1^{(1)}, \ldots, \theta_1^{(M)}$ are posterior draws of the treatment effect θ_1 generated by the iterative algorithm. Second, the non-inferiority of θ_1 to θ_2 is justified by the posterior probability, approximated by

$$\frac{1}{M}\sum_{k=1}^{M} I_{\{\Delta<2\theta_1^{(k)}\}} = \begin{cases} 0.9906, & \Delta = 0.01, \\ 0.9886, & \Delta = 0.05. \end{cases}$$

These posterior probabilities are comparable with their corresponding probabilities approximated by the normal model, which are not reported here. In short, the nDP model seems to be working fine in estimating the treatment effects in this scenario when the errors have a mixture of normal distribution.

Based on the second and the third simulated datasets, we aim at providing an in-depth study of the performance of the nDP model and at demonstrating the superiority of the nDP model over the normal model when dealing with data involving probably some extreme values. In these two datasets, the treatment effect θ_1 remains as 0.5, there are again no covariates, and β_{ij} are distributed as in the first dataset except with zero standard deviations in all the components of the mixture distributions defined in (14) (that is, for instance, β_{i1} is distributed as a two-point mixture at 0 and 3 with respective weights 0.6 and 0.4). The error distributions in the two datasets from which e_{ij} are generated are chosen to be the Student t distributions with 5 degrees of freedom and 1 degree of freedom, respectively, which possess thicker tails than the mixture of normals in the case of the previous dataset. Estimates of posterior probability of equivalence of the two treatments by both methodologies, not reported here, are all close to zero. Table 1 summarizes the posterior probability estimates of the non-inferiority of θ_1 to θ_2, for some $\Delta > 0$, from the two methodologies. The probability estimates, produced by the proposed nDP model for the center effects distributions, roughly equal to 99% in all cases and, are always larger than those produced by the normal model.

Moreover, it seems that the normal model fails to provide as strong evidence as the nDP model in supporting the non-inferiority of θ_1 to θ_2, as the resulting probability estimates are only $\approx 86\%$ when the error distribution is a thick-tailed Cauchy distribution (*i.e.*, for the third dataset). Figure 1 displays boxplots of posterior samples of θ_1 from which the reported probability estimates are computed. When the error distribution is the Student t distribution with 5 degrees of freedom (in the upper graph of Figure 1), the median of the posterior samples from the nDP model (given by 0.447) is closer to the true value of $\theta_1 = 0.5$, compared with the median based on the normal model (given by 0.358). Based on the third dataset (with Cauchy error), the lower graph of Figure 1 depicts that the median of the posterior samples from the nDP model is 0.794, which is close to the true value 0.5, and the range of the samples is comparable to those in the upper plot taking into account that the Cauchy distribution has a thicker tail than the Student t distribution with degrees of freedom greater than 1. On the contrary, corresponding posterior samples from the normal model are totally non-sensible, with median as 1.659 and a much larger range than all the other cases in the same figure. In summary, the flexible nDP model seems to be more powerful in estimating the treatment effects than the normal model when the data are generated from distributions with thicker tails than normal.

[Table 1 about here.]

[Figure 1 about here.]

The superiority of the nDP model over the normal model can be further demonstrated in Figures 2 and 3 constructed based on the second dataset (with error distributed as the Student t distribution with 5 degrees of freedom). Figure 2 depicts empirical distributions of posterior samples of both τ and the center effects β_{ij} from the two methodologies, of which the latter graphs of β_{ij} provide estimates of the center effects distributions F_j. The top two graphs of Figure 2 show that posterior estimates of τ from the nDP model are more likely to take larger values closer to 1, compared with those from the normal model that cluster between 0.1 and 0.3. We argue that the magnitude of τ can serve as an indicator of accuracy in estimation of the treatment effects, as variances of the full conditionals of θ_1 or θ_s, displayed in (17) or (16) in Appendix, are roughly inversely proportional to τ through B_1

or B_s. That is, θ_1 or θ_s is less variable when τ is large. Furthermore, histograms for the 4 center effects are much closer to the true center effects distributions from the nDP model (left column) compared with those from the normal model (right column). For instance, the lower two graphs at the left column clearly show two modes at 5 and 10, respectively.

[Figure 2 about here.]

Figure 3 shows density estimates of any observation associated with the two treatments (left to right) in the four centers (top to bottom) based on the second dataset, wherein solid and dashed lines represent estimates from the nDP and the normal models, respectively, and histograms of the simulated data are displayed in the respective settings according to treatments and centers (roughly 25 observations in each histogram). All solid and dashed lines in the four upper plots, which display roughly unimodal-shaped histograms, are closer to each other. However, in the four lower plots wherein the histograms of the data are somehow bimodal, all the dashed lines fail to capture either the major mode or the small mode, while the solid lines not only capture the small mode to the right more clearly but also emphasize the major mode more precisely. Hence, the nDP model, but not the normal model, is flexible enough in dealing with data that exhibit multimodality together with probably extreme observations.

[Figure 3 about here.]

Next, we look at the performance of the nDP model when the data depend on some covariates. The fourth and fifth datasets differ from the third dataset, which has a Cauchy error distribution, in two aspects. First, there is one covariate from which the observations are generated according to (4) with coefficient $\gamma = -5$. The covariates w_{ij} follow a uniform distribution on $(-1, 1)$ and a normal distribution $\mathtt{N}(0, 1.5^2)$, respectively, in the two datasets. Second, the center effects β_{ij} follow the distributions in (14). Figure 4 presents the empirical distributions of posterior samples of θ_1 and γ based on both datasets from the two methodologies. Histograms of the samples from the nDP model are displayed. Their corresponding density estimates by kernel smoothing techniques are represented by solid lines. Dashed lines are kernel density estimates based on posterior samples from the normal model. The

four histograms, produced by the nDP model, are well-centered and concentrated at the true values of $\theta_1 = 0.5$ and $\gamma = -5$, respectively. However, kernel density estimates produced by the normal model (dashed lines) either do not peak at the true values or spread over a larger range than the histograms in every graph. In addition, we computed the LPML for the two models. Values of LPML for the nDP model and the normal model are -655.126 and -934.999 based on the fourth dataset (with uniform covariates), and -628.022 and -744.6 based on the fifth dataset (with normal covariates), respectively. The irrefutable conclusion that the nDP model outperforms the normal model can be further consolidated by the magnitudes of posterior estimates of τ by the two methods, of which those from the nDP model range from 0 to 0.4 but those from the normal model range from 0 to 0.02 only.

[Figure 4 about here.]

To better differentiate the two methods, we take a closer look at the simulation results based on the fifth dataset wherein the covariates are normally distributed, as results from the normal model shown in the right column of Figure 4 seem more comparable with those from the nDP model. In the top right plot there, 95% posterior interval estimates of θ_1 are given by $(0.001, 0.982)$ for the nDP model and $(-0.308, 1.825)$ for the normal model. We simulated more data based on the same settings such that there are 200, instead of 50, observations from each of the 4 centers, and obtained 95% posterior interval estimates of θ_1 as $(0.295, 0.710)$ for the nDP model and $(0.126, 5.732)$ for the normal model. This shows that the nDP model leads to less variable estimate of θ_1 which is closer to the true value of $\theta_1 = 0.5$ than the normal model, as sample sizes increase.

Finally, we scrutinize for how the special clustering features of the nDP model benefit inference in this context of meta-analysis with the aid of the last simulated dataset, which is a variant of the second dataset. This sixth dataset is identical to the second one in terms of involving no covariate, and same distributions of β_{ij} and of the errors e_{ij}, but the treatment effects are chosen to be smaller as $\theta_1 = 0.05$ in a way to illustrate the ability of the methods in estimating treatment effects of negligible magnitudes compared with magnitudes of the center effects and the errors. Furthermore, the sample size n_j from each center is increased from 50 to 400. According to the probability estimates of both the equivalence

of the two treatments and the non-inferiority of θ_1 to θ_2 defined with $\Delta = 0.01$, given in Table 2, it seems that the nDP model outperforms the normal model by a small margin. However, medians (resp. means) of the posterior samples of θ_1 from the nDP model and the normal model are 0.0316 (resp. 0.0304) and -0.0005 (resp. -0.0005), respectively, wherein the former estimates are much closer to the true value $\theta_1 = 0.05$ than the latter ones. Further, similar empirical distributions of posterior samples of the center effects β_{ij} from the two methodologies to those given in Figure 2, are observed (not included here). That is, estimates of individual center effect from the normal model spread over a small range of values in different iterations. For instance, posterior samples drawn from F_1 for the normal model concentrate over the interval $(0, 2.5)$. On the contrary, for the nDP model, there are different values of estimates of center effects over a much larger range for different observations from the same center in each iteration, wherein some of them may be identical to each other and some of them are substantially larger or smaller than the others, as displayed in the left column of Figure 2. For example, in each iteration, there are two major clusters of center effects with values roughly equalling 5 and 10, respectively, for both centers 3 and 4. Indeed, inherited from the distinct clustering features of nDP, center effects of some particular observations are often estimated to take the same value as center effects of other observations from the same center or different centers throughout different iterations of the proposed algorithm. Among all different M iterations, with possibly different collections of estimates of the center effects, we selected the "best", or the most representative, iteration that corresponds to the largest value of a proxy of LPML, denoted by $\text{LPML}^{(k)}$, which is defined as in (13) with $\widehat{\text{CPO}}_{ij}$ replaced by

$$\widetilde{\text{CPO}}_{ij}^{(k)} \equiv f(y_{ijt}|\theta_t^{(k)}, \beta_{ij}^{(k)}, \boldsymbol{\gamma}^{(k)}, \tau^{(k)}, \mathbf{w}_{ij}).$$

From the resulting "best" iteration, the estimate of θ_1 equals 0.032, which is close to either the median (0.0316) or the mean (0.0304) given above. Eight histograms of center effects estimates are constructed in Figure 5 with respect to both the center they belong to and their true values, $0, 3, 5$, and 10 (from top row to bottom row), during the generation of the dataset. Class widths for these histograms are chosen as small as 0.1 such that estimates of center effects for different observations are stacked into the same bar if and only if their

values are identical. Consequently, for instance, the two longest bars from both graphs at the top row contain only duplicates of center effects estimates as -0.511 and 0.553, respectively, from centers 1 and 2. This shows that the center effects distributions for centers 1 and 2, F_1 and F_2, are clustered together by the nDP prior, which justifies that the two center effects distributions, or equivalently, the two centers, are similar according to definition of this last dataset. Moreover, most ($> 95\%$ of) center effects with true values as 0 are estimated to cluster with one another into two major clusters, and their estimates are either of the two above values that are close to 0. That is, a large number of observations in center 1 are estimated to have the same center effects estimate, either -0.511 or 0.553, as many observations in center 2, demonstrating borrowing of information both within center and across centers that are similar. Analogous interpretations based on the other six histograms at the other 3 rows of Figure 5 can be made. For instance, most individual center effects for different observations are estimated to be close to the true values, 3, 5, and 10, as displayed in the 3 respective rows. In summary, this demonstrates that the special clustering features of the nDP result in accurate estimations of both the treatment effects and the random center effects.

[Table 2 about here.]

[Figure 5 about here.]

5.2 Application to Scleroderma Lung Data

We analyze the Scleroderma lung study as described in Section 1.1. Our main goal here is to assess the efficacy of the oral CYC treatment over the placebo while accounting for the center effects. We take the difference of FVC at baseline from FVC values at week 18th as the endpoint here, and fit the following model without any covariate (Inclusion of covariate can be done in a straightforward way),

$$\text{FVC}_{ijt} = \theta_t + \beta_{ij} + e_{ijt};\ i = 1, 2, \cdots, 145;\ j = 1, 2, \cdots, 13;\ t = \text{oral CYC, placebo.} \quad (15)$$

Analogous to Figures 4 and 2, Figure 6 presents the empirical distributions of posterior samples of the treatment effect θ_1, τ, and the 13 center effects from both the nDP model

and the normal model. Histograms of the samples from the nDP model are displayed, with their corresponding density estimates by kernel smoothing techniques represented by solid lines. Dashed lines are kernel density estimates based on posterior samples from the normal model. First of all, in the plot of estimates of τ, the dashed line centers on the interval $(0.005, 0.01)$, while the solid line spreads over a wider range on $(0.005, 0.03)$. Based on our argument justified with respect to Figure 2, that large magnitude of τ indicates high accuracy in estimation of the treatment effects, this suggests that inference by the nDP model is more accurate than that by the normal model. This is further supported by values of LPML of the two models given as -637.024 and -643.326, respectively. In addition, more than half of the 13 plots of the estimated center effects distributions reveal that the dashed line spreads over a smaller range of values compared with the solid line on the same plot. That is, the normal model tends to result in estimated center effects distributions on a smaller range of values, compared with the nDP model. Such a phenomenon that also appears when the second simulated dataset is analyzed (referred to the right column of Figure 2) may not be desirable as this implies that the center effects distributions are not estimated properly.

[Figure 6 about here.]

From the nDP model, the treatment effect θ_1 is estimated to be 1.365, and τ is estimated to be 0.014. The posterior probability of equivalence of the two treatments, that is, $\theta_1 = \theta_2$, is approximated as

$$\frac{1}{M}\sum_{k=1}^{M} I_{\{-\Delta<2\theta_1^{(k)}<\Delta\}} = \begin{cases} 0.0006, & \Delta = 0.005, \\ 0.0019, & \Delta = 0.01, \\ 0.0088, & \Delta = 0.05, \end{cases}$$

providing strong evidence that the two treatments are not equivalent. Non-inferiority of θ_1 to θ_2 is also well supported by the posterior probability, which is approximated as

$$\frac{1}{M}\sum_{k=1}^{M} I_{\{\Delta<-2\theta_1^{(k)}\}} = \begin{cases} 0.9003, & \Delta = 0.005, \\ 0.8998, & \Delta = 0.01, \\ 0.8965, & \Delta = 0.05. \end{cases}$$

However, these posterior probabilities approximated by the normal model is around 0.85, giving not as strong evidence as the nDP model. In addition, the normal model results in a smaller estimated treatment effect as 1.152.

Lastly, we demonstrate that the nDP model gives a good fit of the FVC data with the aid of density estimates of new observations based on the "best" iteration selected according to $\text{LPML}^{(k)}$, the previously introduced proxy of LPML. Some of histograms of the FVC data with respect to the two treatments (left to right) and the 13 centers (left to right; top to bottom) except center 9 as there is no observation collected based on treatment 2, shown in Figure 7, suggest that the FVC data exhibit multimodality together with probably some extreme observations. Similar to what are displayed in Figure 3, all density estimates from the normal model (not presented here) are unimodal bell-shaped curves, failing in capturing either multi-modes or possibly outlying observations. Most density estimates from the nDP model plotted in Figure 7 seem fitting the corresponding histograms quite well, demonstrating good account of capturing multimodality and dealing with outlying observations by the proposed methodology. Parameter estimates from this "best" iteration, say, k^*-th iteration, include $\theta^{(k^*)} = 1.309$ and $\tau^{(k^*)} = 0.058$. All center effects distributions are clustered together, except centers 2 and 10. Most estimated individual center effects in the major cluster are among a list of values given as $-24.405, -9.298, -1.173, 0.331, 14.922$ and 26.505. Center effects from centers 2 and 10 are estimated to be among $-22.229, -10.19$ and -1.334.

[Figure 7 about here.]

6 Conclusion

Multi-center clinical trial has become a popular and useful tool for quantitative synthesizing and summarizing information in the medical literature. Given the availability of reliable data, a multi-center trial should employ robust methods. However, there is empirical evidence to suggest that the use of robust methods is low. The random effects multi-center model is a parsimonious way of accounting for within-center and between-center variation. In this research, we have provided a general modeling framework to analyze the multi-center clinical

trial in a mixed model framework. This mixed model framework provides a useful way of describing typical data from multi-center trials. While it is argued that the routine use of normal distribution may bias the inference on treatment effects, it was also of interest to cluster the centers behaving similarly interms of patient population. To achieve this goal, we developed a novel Bayesian semiparametric model where we account for the nested center effects using the nDP.

Using a thorough simulation study, and application to a real dataset, we have demonstrate the ability of the nDP model in providing accurate estimates of the parameters of interest particularly when the random center effects is not coming from normal. Since our model can provide a way to evaluate the treatment effects correctly even under the distributional misspecification, our research can serve as a useful tool for deriving better analysis of multi-center clinical trials. The insensitivity to outliers and the nice clustering behavior of the center effects make our nDP approach an important tool in detecting outlying centers and a robust alternative to the traditional parametric analysis.

References

Anello, C., O' Neill, R., and Dubey, S. (2005), Multicentre trials: a US regulatory perspective, *Statistical Methods in Medical Research*, **14**, 303-318.

Antoniak, C. E. (1974), Mixtures of Dirichlet processes with applications to nonparametric problems, *The Annals of Statistics*, **2**, 11521174.

Blackwell, D. and MacQueen, J. B. (1973), Ferguson distributions via Pólya urn schemes, *The Annals of Statistics*, **1**, 353-355.

Basu S. and Chib S. (2003), Marginal likelihood and Bayes factors for Dirichlet process mixture models, *Journal of the American Statistical Association*, **98**, 224-235.

Burr, D. and Doss, H. (2005), A Bayesian semiparametric model for random-effects meta-analysis, *Journal of the American Statistical Association*, **100**, 242251.

Böhning, D. (2000), *Computer-assisted Analysis of Mixtures and Applications: Meta-analysis, Disease mapping and Others*. Boca Raton: Chapman and Hall-CRC.

Branscum, A. and Hanson, T. (2008), Bayesian nonparametric meta-analysis using Pólya tree mixture models, *Biometrics*, **64**, 825-833.

Chen, M. H., Shao, Q. M., and Ibrahim, J. G. (2000), *Monte Carlo methods in Bayesian computation*, Springer-Verlag Inc (Berlin; New York).

Escobar, M. D. (1988), Estimating the means of Several Normal Populations by Nonparametric Estimation of the Distribution of the Means, unpublished Ph.D. thesis, Yale University, Dept. of Statistics.

Escobar, M. D. (1994), Estimating normal means with a Dirichlet process prior, *Journal of the American Statistical Association*, **89**, 268-277.

Escobar, M. D. and West, M. (1995), Bayesian density estimation and inference using mixtures *Journal of the American Statistical Association*, **90**, 577-588.

Ferguson, T. S. (1973), A Bayesian analysis of some nonparametric problems, *The Annals of Statistics*, **1**, 209-230.

Gelfand, A. E., Dey, D. K., and Chang, H. (1992), Model determination using predictive distributions with implementation via sampling-based methods (with discussion). In *Bayesian Statistics*, Volume 4, J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith (eds), 147-159. Oxford: Oxford University Press.

Geisser, I., and Eddy, W. (1979), A predictive approach to model selection, *Journal of the American Statistical Association*, **74**, 153-160.

Gould, A. L. (2005), Bayesian analysis of multicenter trial outcomes, *Statistical Methods in Medical Research*, **14**, 249-280.

Hayakawa, Y., Zukerman, J., Paul, S., and Vignaux, T. (2001), Bayesian nonparametric testing of constant versus nondecreasing hazard rates. In *System and Bayesian Reliability: Essays in Honor of Professor Richard E. Barlow on His 70^{th} Birthday* (Y Hayakawa, T Irony and M Xie Eds), 391-406, World Scientific.

Higgins, J. P. T., Thompson, S. G., and Spiegelhalter, D. J. (2009), A re-evaluation of random-effects meta-analysis, *Journal of the Royal Statistical Society: Series A*, **172**, 139-159.

International Conference on Harmonization, E9. Guidance on choice of Control Group and Related Issues in Clinical Trials. Federal Register. *Food and Drug Administration*, 1998; **66**, 1-33.

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

Jackel, L. A. (1972), Estimating regression coefficients by minimizing the dispersion of residuals, *Annals of Mathematical Statistics*, **43**, 1449-1458.

Khatri, C. G., and Patel, H. I. (1992), Analysis of a multicenter trial using a multivariate approach to a mixed linear model, *Communications in Statistics: Theory and Methods*, **21**, 21-39.

Lee, K. J., and Thompson, S. G. (2007), Flexible parametric models for random-effects distributions, *Statistics in Medicine*, **27**, 418-434.

Lo, A. Y. (1978). Bayesian nonparametric density methods. Technical Report, University of California, Berkeley.

Lo, A. Y. (1984), On a class of Bayesian nonparametric estimates. 1. Density estimates, *The Annals of Statistics*, 12, 351-357.

MacEachern, S. N., Clyde, M. and Liu, J. S. (1999), Sequential importance sampling for nonparametric Bayes models: The next generation, *Canadian Journal of Statistics*, **27**, 251-267.

Ohlssen, D. I., Sharples, L. D., and Spiegelhalter, D. J. (2007), Flexible random-effects models using Bayesian semi-parametric models: applications to institutional comparisons, *Statistics in Medicine*, **26**, 2088-2112.

Patel, H. I. (2002), Robust analysis of a mixed-effect model for a multicenter clinical trial, *Journal of Biopharmaceutical Statistics*, **12**, 21-37.

Rashid, M. M. (2003), Rank-based test for non-inferiority and equivalence hypotheses in multi-center clinical trials using mixed models, *Statistics in Medicine*, **22**, 291-311.

Rodríguez, A., Dunson, D. B., and Gelfand, A. E. (2008), The nested Dirichlet process, with discussion, *Journal of the American Statistical Association*, **103**, 1131-1154.

Sethuraman, J. (1994), A constructive definition of Dirichlet priors, *Statistica Sinica*, **4**, 639-650.

Sethuraman, J. and Tiwari, R. C. (1982), Convergence of Dirichlet measure and the interpretation of their parameters, In *Statistical Decisions Theory and Related Topics III*, vol. 2 (Gupta, S, and Berger, J. O. Eds.), Academic Press, 305-315.

Tashkin, D. P., Elashoff, R. M., and Clements, P. J. (2006), Cyclophosphamide versus placebo in scleroderma lung disease, *The New England Journal of Medicine*, **354**, 2655- 2666.

Thompson, S. G. (1994), Why sources of heterogeneity in meta-analysis should be investigated,

7 Appendix

An iterative algorithm for sampling random variates of $(\boldsymbol{\zeta}, \boldsymbol{\xi}, \boldsymbol{\pi}^*, \boldsymbol{\omega}^*, \boldsymbol{\beta}^*, \alpha, \rho)$, $\boldsymbol{\gamma}$, τ, and $(\theta_1, \ldots, \theta_T)$ from their joint posterior distribution cycles through the following four steps.

1. Sampling of $(\boldsymbol{\zeta}, \boldsymbol{\xi}, \boldsymbol{\pi}^*, \boldsymbol{\omega}^*, \boldsymbol{\beta}^*, \alpha, \rho)$ for the center effects are carried out through the following ste4ps:

   (a) Sample the classification variables ζ_j for $j = 1, \ldots, C$ from a multinomial distribution with probabilities

   $$\mathbb{P}(\zeta_j = k | \cdots) \propto \pi_k^* \prod_{i=1}^{n_j} \sum_{l=1}^{L} f(y_{ijt} | \theta_t, \beta_{lk}^*, \boldsymbol{\gamma}, \tau, \mathbf{w}_{ij}), \qquad k = 1, \ldots, K.$$

   (b) Sample the classification variables ξ_{ij} for $j = 1, \ldots, C$ and $i = 1, \ldots, n_j$ from a multinomial distribution with probabilities

   $$\mathbb{P}(\xi_{ij} = l | \cdots) \propto \omega_{l\zeta_j}^* f(y_{ijt} | \theta_t, \beta_{l\zeta_j}^*, \boldsymbol{\gamma}, \tau, \mathbf{w}_{ij}), \qquad l = 1, \ldots, L.$$

   (c) Sample $\boldsymbol{\pi}^*$ by generating

   $$\begin{aligned} (u_k^* | \cdots) &\overset{\text{ind}}{\sim} \texttt{beta}\left(1 + m_k, \alpha + \sum_{s=k+1}^{K} m_s\right), \qquad k = 1, \ldots, K-1, \\ u_K^* &= 1, \end{aligned}$$

   where $m_k = \sum_{j=1}^{C} \mathbf{I}_{\{\zeta_j = k\}}$ is the number of distributions among $F_1, \ldots, F_C$ assigned to component k in (8), and constructing $\pi_k^* = u_k^* \prod_{s=1}^{k-1}(1 - u_s^*)$ for $k = 1, \ldots, K$.

   (d) Sample $\boldsymbol{\omega}^*$ by generating, for $k = 1, \ldots, K$,

   $$\begin{aligned} (v_{lk}^* | \cdots) &\overset{\text{ind}}{\sim} \texttt{beta}\left(1 + n_{lk}, \rho + \sum_{s=l+1}^{L} n_{lk}\right), \qquad l = 1, \ldots, L-1, \\ v_{Lk}^* &= 1, \end{aligned}$$

   where $n_{lk} = \sum_{j=1}^{C} \sum_{i=1}^{n_j} \mathbf{I}_{\{\zeta_j = k, \xi_{ij} = l\}}$ is the number of center effects assigned to atom l of distribution k in (9), and constructing $\omega_{lk}^* = v_{lk}^* \prod_{s=1}^{l-1}(1 - v_{sk}^*)$ for $l = 1, \ldots, L$.

(e) Sample β^*_{lk}, for $k = 1, \ldots, K$ and $l = 1, \ldots, L$, according to

$$(\beta^*_{lk}|\cdots) \sim \mathtt{N}\left(\frac{\tau \sum\limits_{\{i,j|\zeta_j=k,\xi_{ij}=l\}} (y_{ijt} - \theta_t - \mathbf{w}_{ij}^\top \boldsymbol{\gamma})}{n_{lk}\tau + \sigma_\beta^{-2}}, \frac{1}{n_{lk}\tau + \sigma_\beta^{-2}}\right).$$

(f) Sample

$$(\alpha|\cdots) \sim \mathtt{gamma}\left(a_\alpha + (K-1), b_\alpha - \sum_{k=1}^{K-1} \log(1-u^*_k)\right)$$

and

$$(\rho|\cdots) \sim \mathtt{gamma}\left(a_\rho + K(L-1), b_\rho - \sum_{l=1}^{L-1}\sum_{k=1}^{K} \log(1-v^*_{lk})\right),$$

where $\mathtt{gamma}(a, b)$ represents a gamma random variable X with density $h(x|a, b) \propto x^{a-1}e^{-bx}$, $x > 0$.

2. Sample $\boldsymbol{\gamma}$ from its full conditional distribution,

$$p(\boldsymbol{\gamma}|\cdots) \propto \left[\prod_{j=1}^{C}\prod_{i=1}^{n_j} f(y_{ijt}|\theta_t, \beta^*_{\xi_{ij},\zeta_j}, \boldsymbol{\gamma}, \tau, \mathbf{w}_{ij})\right] \phi_r(\boldsymbol{\gamma}|\mathbf{0}, \Sigma_{\boldsymbol{\gamma}}).$$

where $\phi_r(\cdot|\mathbf{0}, \Sigma_{\boldsymbol{\gamma}})$ is a r-variate normal density with mean vector $\mathbf{0}$ and variance-covariance matrix $\Sigma_{\boldsymbol{\gamma}}$. For instance, when $\boldsymbol{\gamma} = \gamma$ is univariate and is distributed as $\mathtt{N}(0, \sigma^2_\gamma)$,

$$(\gamma|\cdots) \sim \mathtt{N}\left(\frac{\tau \sum\limits_{j=1}^{C}\sum\limits_{i=1}^{n_j} w_{ij}(y_{ijt} - \theta_t - \beta^*_{\xi_{ij},\zeta_j})}{\tau \sum\limits_{j=1}^{C}\sum\limits_{i=1}^{n_j} w^2_{ij} + \sigma_\gamma^{-2}}, \frac{1}{\tau \sum\limits_{j=1}^{C}\sum\limits_{i=1}^{n_j} w^2_{ij} + \sigma_\gamma^{-2}}\right),$$

where w_{ij} is the covariate for i-th subject from j-th center.

3. Sample τ from its full conditional distribution,

$$p(\tau|\cdots) \propto \left[\prod_{j=1}^{C}\prod_{i=1}^{n_j} f(y_{ijt}|\theta_t, \beta^*_{\xi_{ij},\zeta_j}, \boldsymbol{\gamma}, \tau, \mathbf{w}_{ij})\right] h(\tau|a_\tau, b_\tau).$$

That is,

$$(\tau|\cdots) \sim \mathtt{gamma}\left(a_\tau + \frac{1}{2}\sum_{j=1}^{C} n_j, b_\tau + \frac{1}{2}\sum_{j=1}^{C}\sum_{i=1}^{n_j}(y_{ijt} - \theta_t - \beta^*_{\xi_{ij},\zeta_j} - \mathbf{w}_{ij}^\top\boldsymbol{\gamma})^2\right).$$

4. For identifiability issue, assume that sum of all θ_i's equals zero, that is, $\theta_T = -\sum_{i=1}^{T-1}\theta_i$. For $s = 1, \ldots, T-1$, let $\tilde{\theta}_s = \sum_{i=1, i\neq s}^{T-1}\theta_i$. Sample θ_s for $s = 1, \ldots, T-1$ from its full conditional distribution,

$$(\theta_s|\cdots) \sim \mathtt{N}\left(\frac{B_s\sigma_\theta^2\widetilde{m}_s}{B_s\sigma_\theta^2+1}, \frac{\sigma_\theta^2}{B_s\sigma_\theta^2+1}\right), \tag{16}$$

where $B_s = \tau[\sum_{j=1}^{C}\sum_{i=1}^{n_j}\mathbf{I}_{\{t=s\}} + \sum_{j=1}^{C}\sum_{i=1}^{n_j}\mathbf{I}_{\{t=T\}}] \equiv M\tau$ with M being the total number of observations among $N = \sum_{j=1}^{C} n_j$ satisfying the events $\{t=s\}$ or $\{t=T\}$, and

$$\widetilde{m}_s = \frac{\tau}{B_s}\left[\sum_{j=1}^{C}\sum_{i=1}^{n_j}(y_{ijt} - \beta^*_{\xi_{ij},\zeta_j} - \mathbf{w}_{ij}^\top\boldsymbol{\gamma})\mathbf{I}_{\{t=s\}} + \sum_{j=1}^{C}\sum_{i=1}^{n_j}(-y_{ijt} - \tilde{\theta}_s + \beta^*_{\xi_{ij},\zeta_j} + \mathbf{w}_{ij}^\top\boldsymbol{\gamma})\mathbf{I}_{\{t=T\}}\right].$$

4* When there are $T = 2$ treatments, we assume that $\theta_1 = -\theta_2 \equiv \theta$ for identifiability issue. We sample θ_1 from its full conditional distribution,

$$(\theta_1|\cdots) \sim \mathtt{N}\left(\frac{B_1\sigma_\theta^2\widetilde{m}_1}{B_1\sigma_\theta^2+1}, \frac{\sigma_\theta^2}{B_1\sigma_\theta^2+1}\right), \tag{17}$$

where $B_1 = (\sum_{j=1}^{C} n_j)\tau = N\tau$, and

$$\widetilde{m}_1 = \frac{1}{N}\left[\sum_{j=1}^{C}\sum_{i=1}^{n_j}(y_{ijt} - \beta^*_{\xi_{ij},\zeta_j} - \mathbf{w}_{ij}^\top\boldsymbol{\gamma})\mathbf{I}_{\{t=1\}} + \sum_{j=1}^{C}\sum_{i=1}^{n_j}(-y_{ijt} + \beta^*_{\xi_{ij},\zeta_j} + \mathbf{w}_{ij}^\top\boldsymbol{\gamma})\mathbf{I}_{\{t=2\}}\right].$$

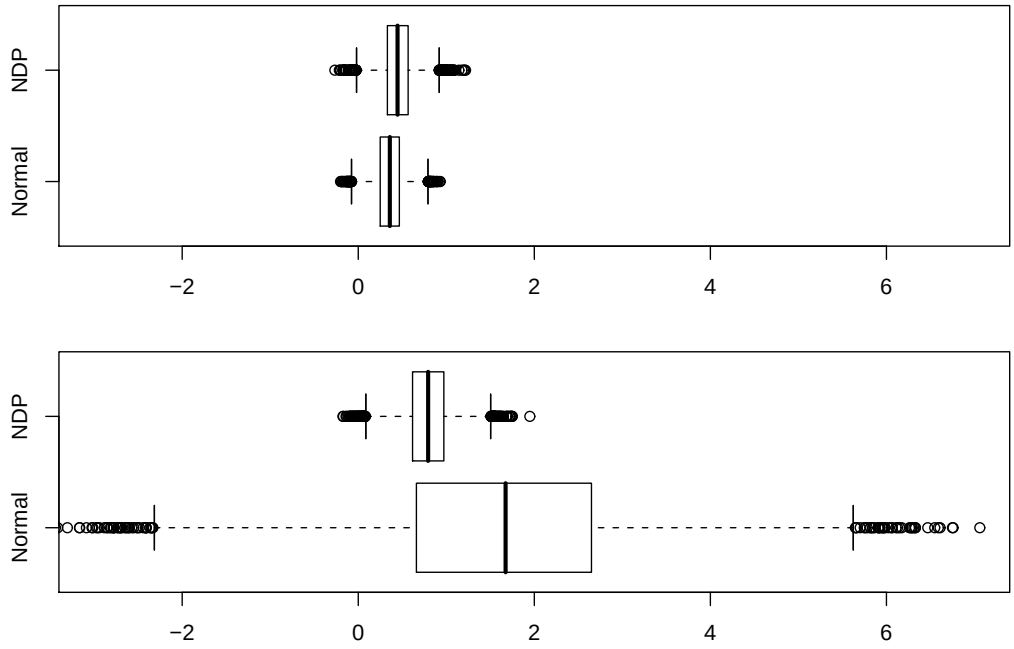

Figure 1: Boxplots of posterior samples of θ_1 based on the second and the third simulated datasets with respective error distributions as t with 5 degrees of freedom (upper graph) and Cauchy distribution (lower graph) assuming nDP and normal distributions on center effect distributions.

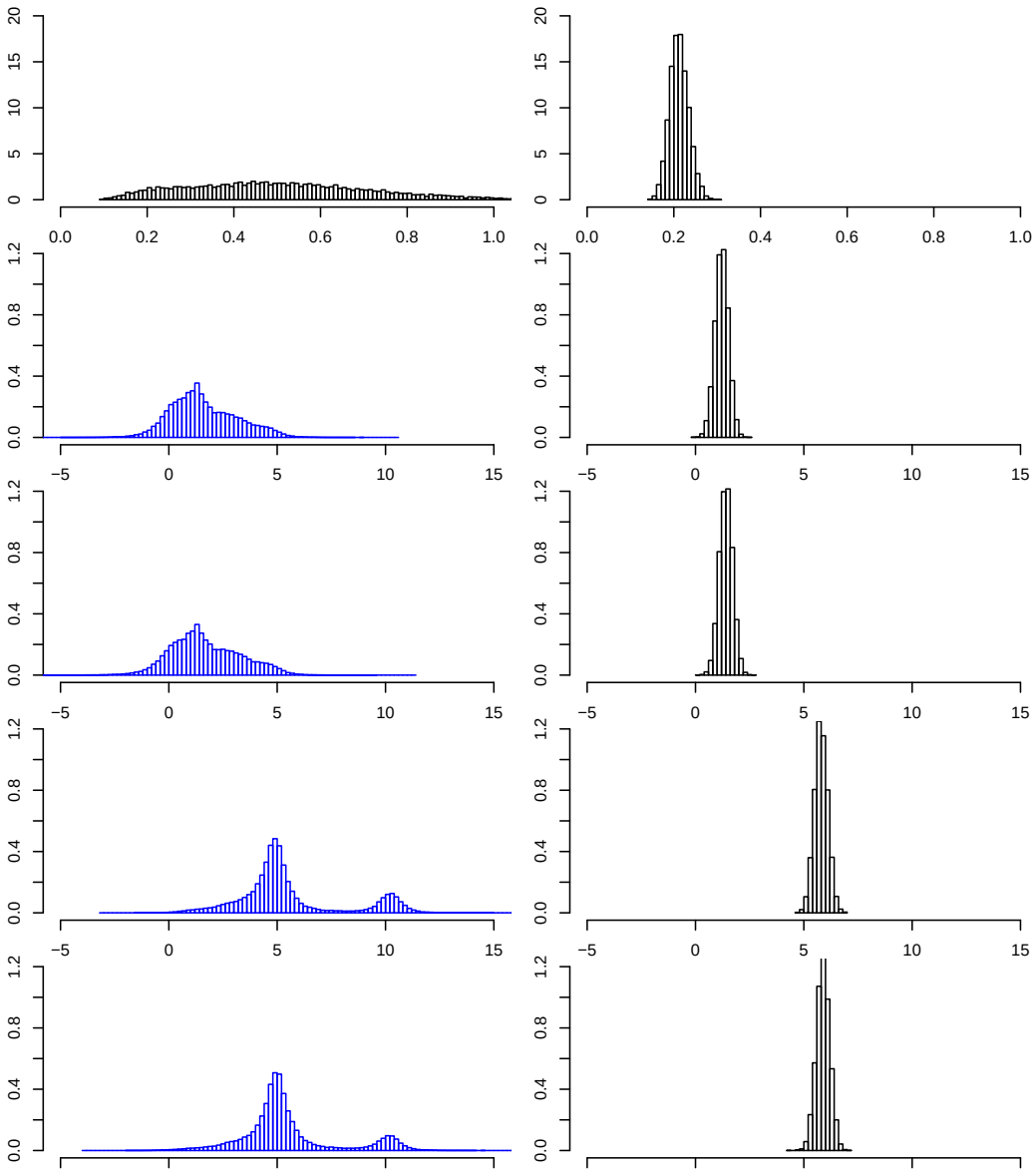

Figure 2: Empirical distributions of posterior samples of τ, and β_{ij} for $j = 1, \ldots, 4$ (from top to bottom) based on the second simulated dataset with error distribution as Student's t distribution with 5 degrees of freedom assuming nDP (left) and normal (right) distributions on center effect distributions.

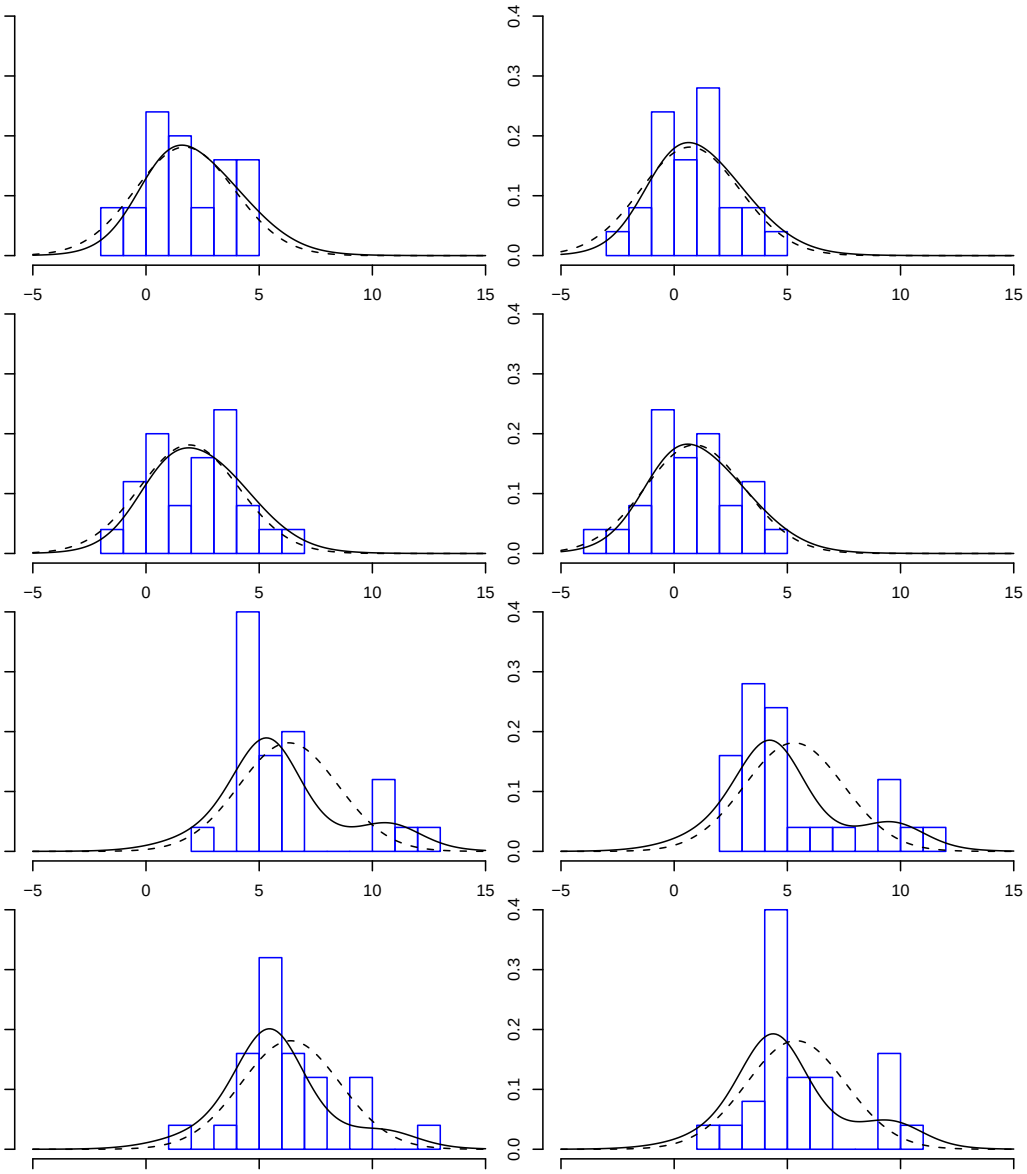

Figure 3: Density estimates (solid lines: from the nDP model; dashed lines: from the normal model) of observations associated with the 2 treatments (left to right) in the 4 centers (top to bottom) based on the second simulated dataset with error distribution as Student's t distribution with 5 degrees of freedom.

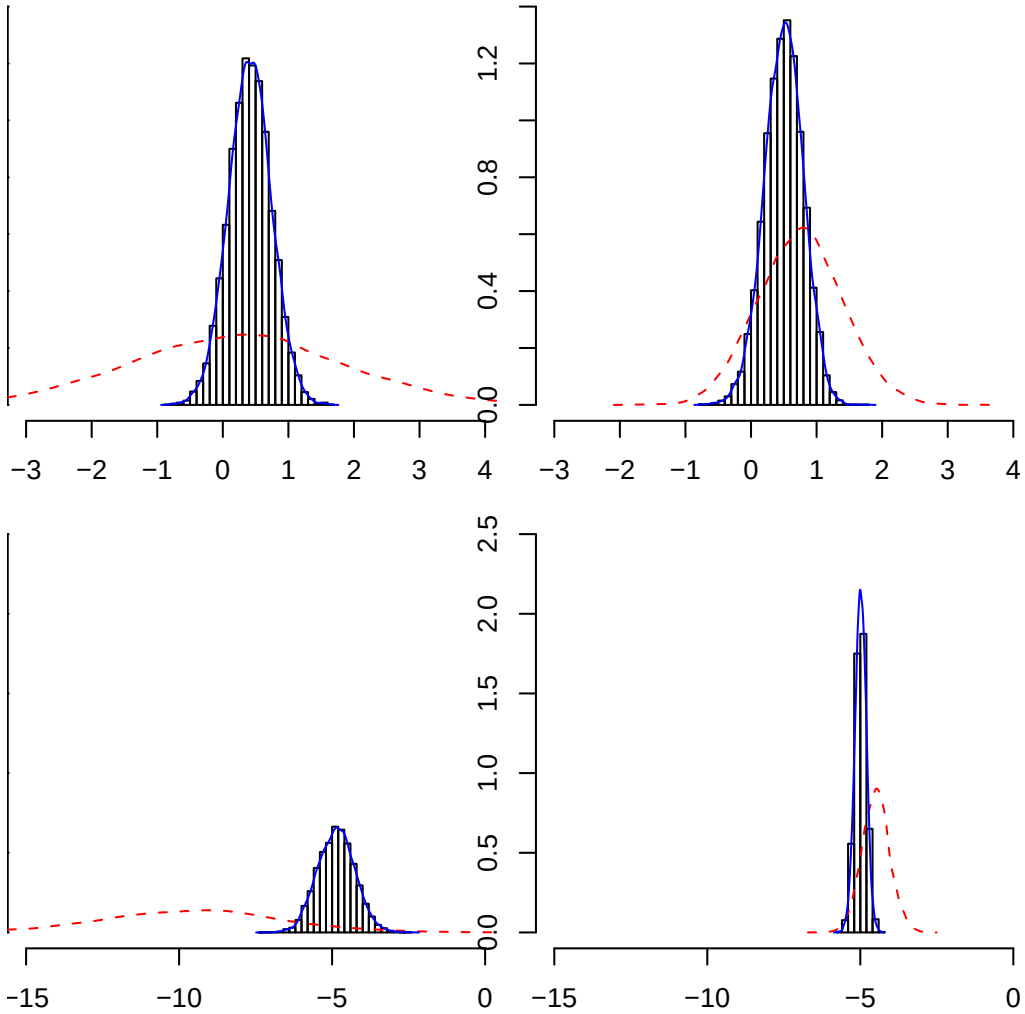

Figure 4: Empirical distributions of posterior samples of θ_1 (top row) and γ (bottom row) based on the fourth and the fifth simulated dataset with respective uniform (left column) and normal (right column) covariates assuming nDP (histograms and solid lines) and normal (dashed lines) distributions on center effects distributions.

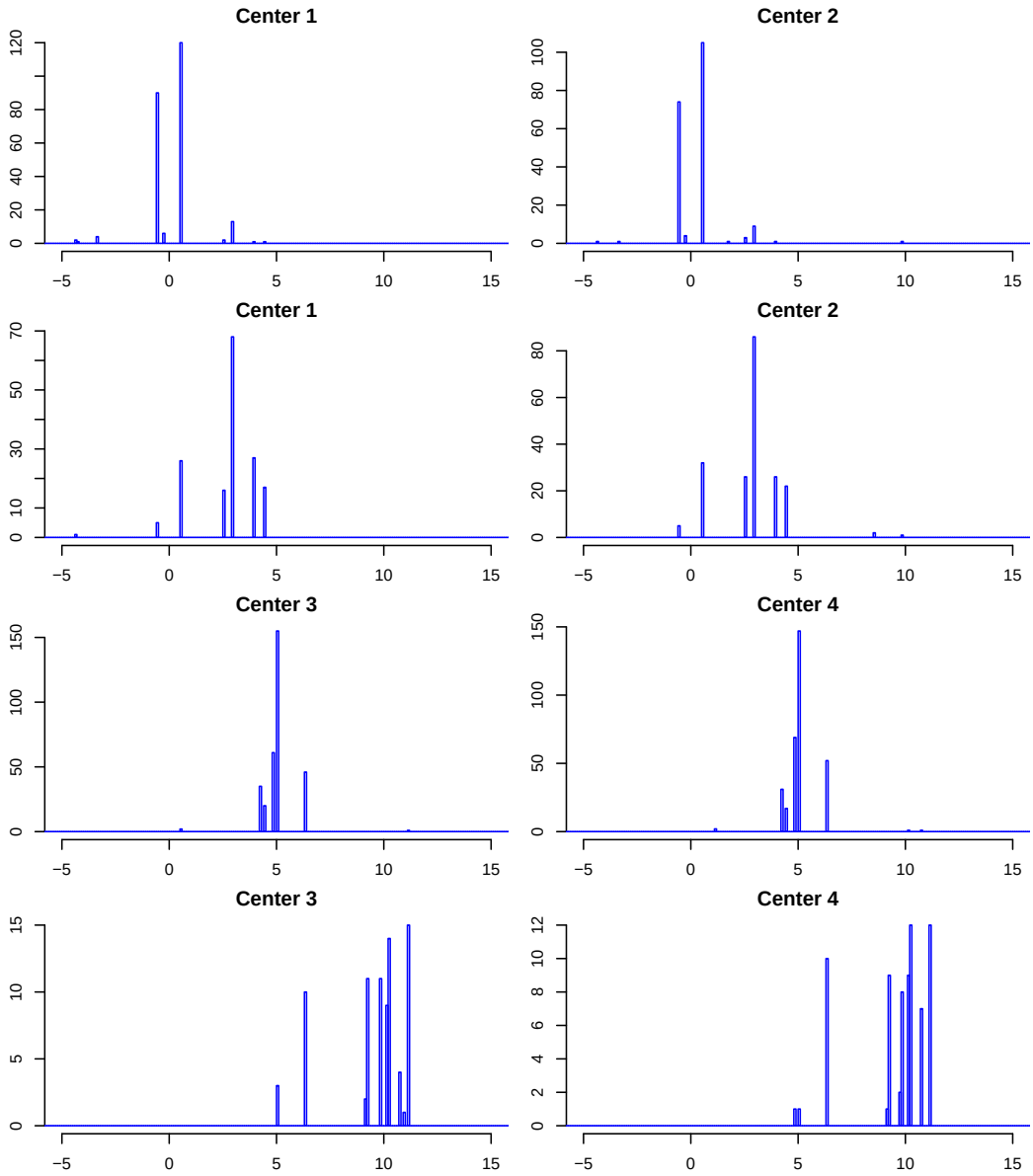

Figure 5: Frequency histograms of center effects from the "best" iteration based on the sixth simulated dataset. True values of the center effects are $0, 3, 5$, and 10 from top row to bottom row.

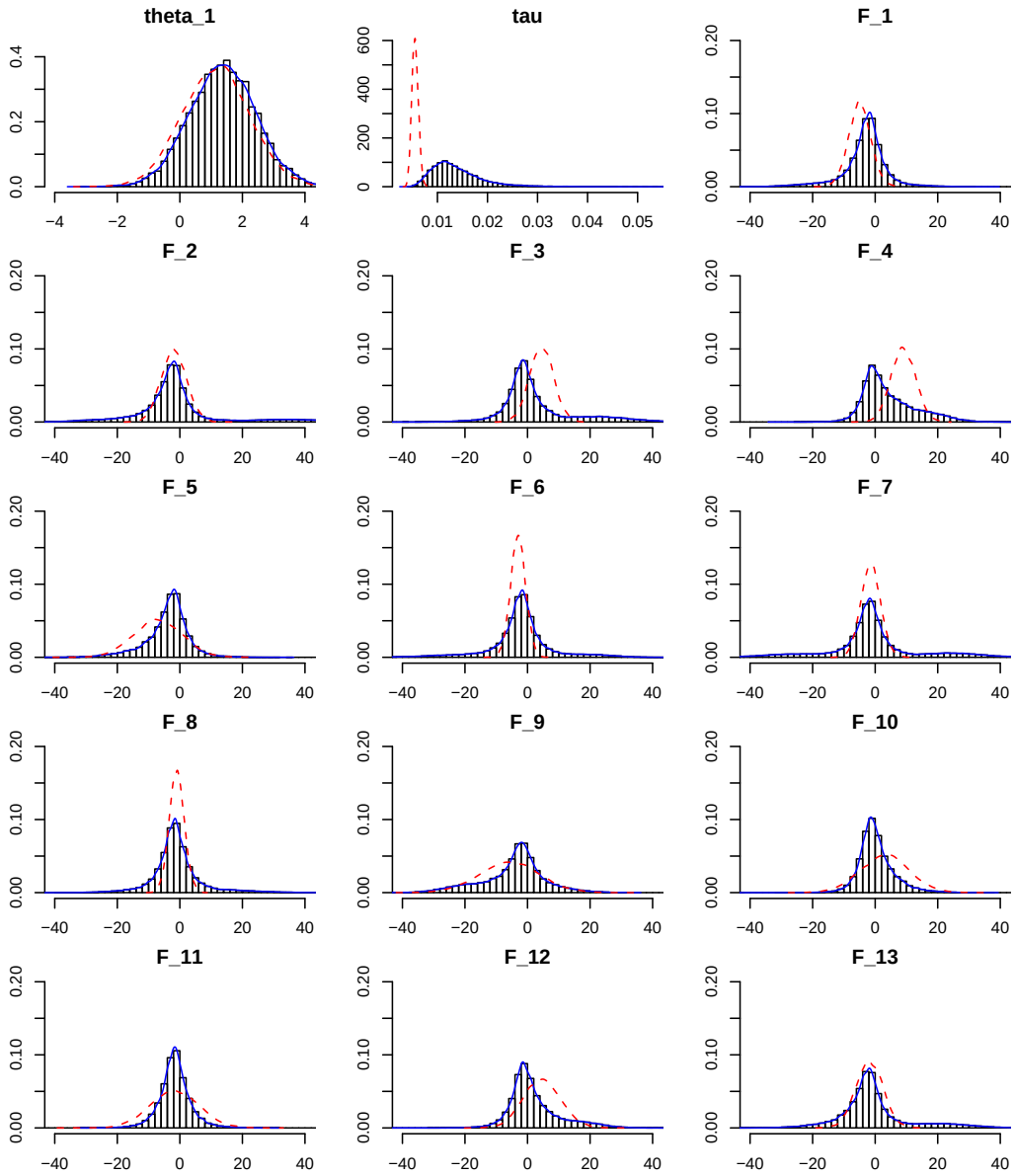

Figure 6: Empirical distributions of posterior samples of θ_1, τ, and the center effects based on the FVC data assuming nDP (histograms and solid lines) and normal (dashed lines) distributions on center effects distributions.

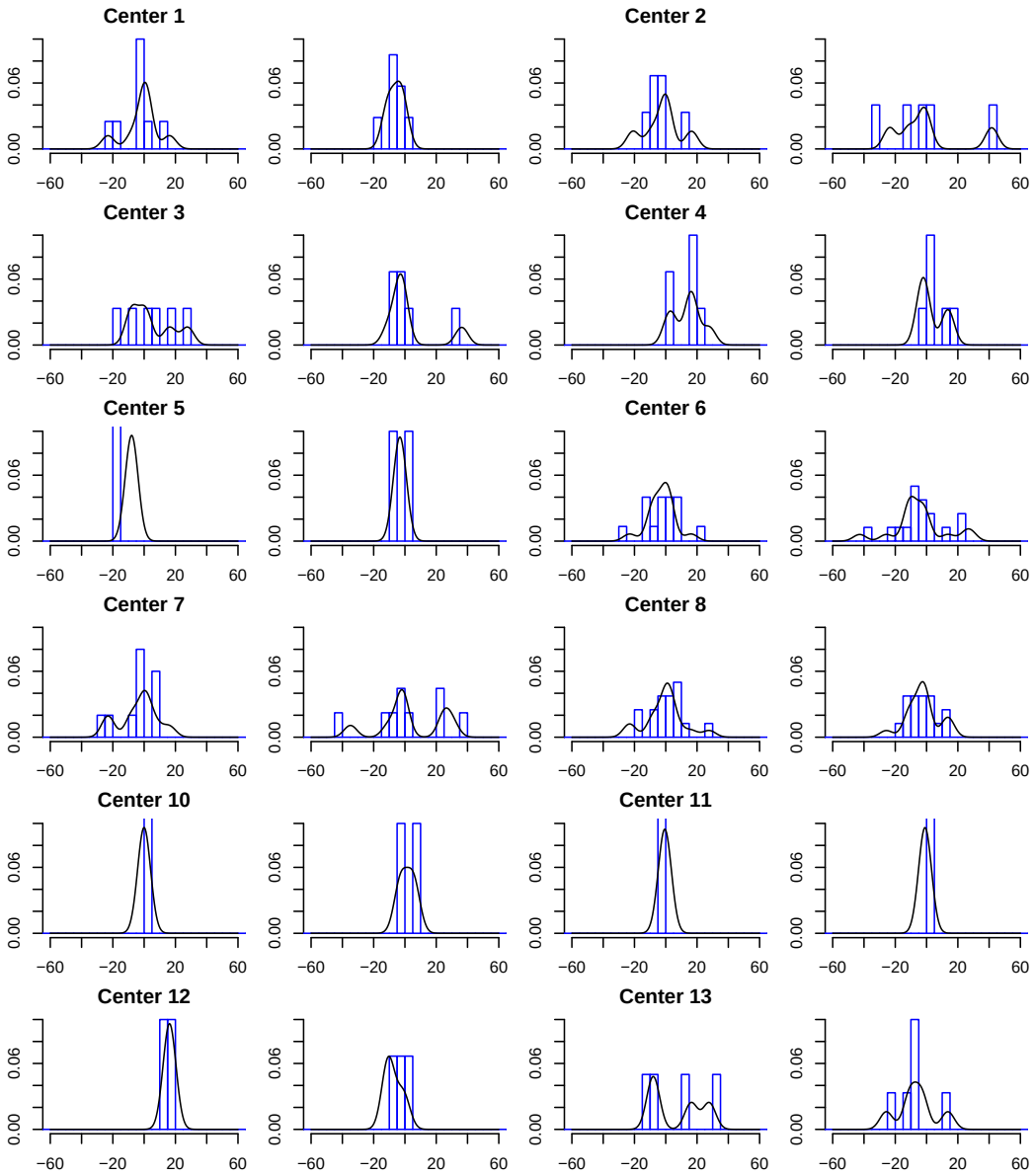

Figure 7: The FVC data (histograms) and density estimates (solid lines) of observations associated with the 2 treatments (left to right) in the 13 centers except center 9 (left to right; top to bottom) based on *estimates from the "best" iteration by assuming nDP on center effects distributions*.

Table 1: Probability estimates of non-inferiority of θ_1 to θ_2 based on the second and the third simulated datasets with respective error distributions as t and Cauchy distributions assuming nDP and normal distributions on center effect distributions.

		Center Effects Distribution	
Error distribution	Δ	nDP	Normal
t with 5 df	0.01	0.991	0.983
	0.05	0.989	0.977
Cauchy	0.01	0.997	0.870
	0.05	0.996	0.866

Table 2: Probability estimates of equivalence of treatments and non-inferiority of θ_1 to θ_2 (with $\Delta = 0.01$) based on the sixth simulated dataset.

	Center Effects Distribution	
Probability estimates	nDP	Normal
Equivalence of θ_1 and θ_2	0.052	0.072
Non-inferiority of θ_1 to θ_2	0.675	0.461