Memorial Sloan-Kettering Cancer Center
Memorial Sloan-Kettering Cancer Center, Dept. of Epidemiology
& Biostatistics Working Paper Series

Year 2010 *Paper* 19

Assessing noninferiority in a three-arm trial using the Bayesian Approach

Pulak Ghosh* Farouk S. Nathoo†
Mithat Gonen‡ Ram C. Tiwari**

*pulak.ghosh@iimb.ernet.in
†University of Victoria, nathoo@math.uvic.ca
‡Memorial Sloan-Kettering Cancer Center, gonenm@mskcc.org
**Federal Drug Administration, ram.tiwari@fda.hhs.gov

http://biostats.bepress.com/mskccbiostat/paper19

Assessing noninferiority in a three-arm trial using the Bayesian Approach

Pulak Ghosh, Farouk S. Nathoo, Mithat Gonen, and Ram C. Tiwari

Abstract

Non-inferiority trials, which aim to demonstrate that a test product is not worse than a competitor by more than a pre-specified small amount, are of great importance to the pharmaceutical community. As a result, methodology for designing and analyzing such trials is required, and developing new methods for such analysis is an important area of statistical research. The three-arm clinical trial is usually recommended for non-inferiority trials by the Food and Drug Administration (FDA). The three-arm trial consists of a placebo, a reference, and an experimental treatment, and simultaneously tests the superiority of the reference over the placebo along with comparing this reference to an experimental treatment. In this paper, we consider the analysis of noninferiority trials using Bayesian methods which incorporate both parametric as well as semi-parametric models. The resulting testing approach is both flexible and robust. The benefit of the proposed Bayesian methods is assessed via simulation, based on a study examining Home Based Blood Pressure Interventions.

Assessing noninferiority in a three-arm trial using the Bayesian Approach

Pulak Ghosh, Farouk Nathoo, Mithat Gönen, and Ram C. Tiwari *

April 27, 2010

Abstract

Non-inferiority trials, which aim to demonstrate that a test product is not worse than a competitor by more than a pre-specified small amount, are of great importance to the pharmaceutical community. As a result, methodology for designing and analyzing such trials is required, and developing new methods for such analysis is an important area of statistical research. The three-arm clinical trial is usually recommended for non-inferiority trials by the Food and Drug Administration (FDA). The three-arm trial consists of a placebo, a reference, and an experimental treatment, and simultaneously tests the superiority of the reference over the placebo along with comparing this reference to an experimental treatment. In this paper, we consider the analysis of noninferiority trials using Bayesian methods which incorporate both parametric as well as semi-parametric models. The resulting testing approach is both flexible and robust. The benefit of the proposed Bayesian methods is assessed via simulation, based on a study examining Home Based Blood Pressure Interventions.

Keywords: Bayesian methods; Gold Standard Design; Markov Chain Monte Carlo; noninferiority; Home-Based Blood Pressure Interventions

*Pulak Ghosh (Email: pulak.ghosh@iimb.ernet.in) is Associate Professor, Department of Quantitative Methods and Information Sciences, Indian Institute of Management, Bangalore. Farouk Nathoo (nathoo@math.uvic.ca) is Assistant Professor, Department of Mathematics and Statistics, University of Victoria. Mithat Gonen (gonenm@mskcc.org) is Associate Attending Biostatistician, Memorial Sloan Kettering Cancer Center. Ram C. Tiwari (Email: Ram.Tiwari@fda.hhs.gov) is Associate Director, Office of Biostatistics, CDER, FDA. The views expressed by Dr. Tiwari is his own and do not necessarily reflect those of FDA.

1 Introduction

Recently, there has been a growing interest in drug development to demonstrate whether a new treatment is not worse than that of an active control by more than a specified margin (Snapinn, 2000). This helps in assessing whether a less toxic, easier to administer, or less expensive treatment is clinically non-inferior to a standard treatment. This kind of clinical trial, where the intention is to investigate whether a new treatment is not inferior to the standard treatment by more than a small predefined margin, is usually known as non-inferiority trial (EMEA, 2005). There have been a series of articles on this topic; see for example, special issues of *Statistics in Medicine* (Volume 47, Issue 1, 2005) and *Journal of Biopharmaceutical Statistics* (Volume 14, Number 2, 2004). It is clear that a new treatment might be preferred to a standard therapy despite it not being better than a standard treatment. For example, the new treatment may be less invasive and less debilitating, or it may be less expensive, and hence preferable. For these reasons, once noninferiority with respect to the primary end point has been demonstrated, the new treatment would be an attractive option for patients, and this is of benefit to the health care system in general.

The statistical literature dealing with inference for noninferiority for two treatments has grown substantially in the last two decades (D'Agostino et al., 2003; Munk et al., 2005; Koti, 2007). Two-arm noninferiority trials of a test treatment and a well established reference treatment are an attractive option in that, in certain settings, there is no need to expose patients to a placebo. Nevertheless, two-arm noninferiority trials exhibit some major challenges in terms of design, analysis and interpretation (Jones et al., 1996; Rohmel, 1998; Temple et al., 2000; D'Agostino et al., 2003; Koch and Rohmel, 2004). Most two-arm trials lack the support of the assay sensitivity resulting in an inability of the trial to distinguish between test and active control treatments. As a result, the inclusion of a placebo group into trials comparing active treatments is useful, whenever this is ethical (Kieser and Friede, 2007; Koti 2007).

Recently, Pigeot et al. (2003) and Koch and Rohmel (2004) considered three-arm trials with both a known effective active standard treatment/drug and placebo as control groups. These three-arm noninferiority trials are useful as they avoid the difficulties described above. In this case, efficacy of the test treatment can be demonstrated by direct comparison to the placebo; however, a major limitation of these methods is the assumption of a homogenous variance in the response variables collected across the treatment arms. Along these lines, while noninferiority trials have generated considerable research in last few decades, there have been few attempts to address noninferioity

under a heteroscedastic variance assumption. Recently, Hasler et al. (2007) and Koti (2007) have considered the analysis of noninferiority trials in three treatment arms in the presence of heteroscadesticity. Hasler et al. (2007) used a t-distribution to test the noninferiority hypothesis; whereas, Koti (2007) has developed a new test procedure based on the Fieller-Hinkley distribution.

In this paper, we put forth a novel Bayesian approach for the analysis of noninferiority trials under three treatments in the presence of heteroscadesticity. A fully Bayesian approach can have important advantages in accounting fully for various sources of uncertainty, and incorporating prior information (Gill, 2002; Gelman *et al.* , 2004). Posterior distributions can be computed efficiently and accurately using simulation based methods, and inference relating to non-inferiority testing can proceed without resorting to asymptotics, which is useful with small sample studies. In addition, as a non-inferiority trial involves treatments that have been well-studied in the past, it is plausible that prior information is available, and the ability to incorporate such information is an advantage. Finally, the Bayesian approach circumvents the difficulties encountered with traditional methods for hypothesis testing (Ghosh and Gönen, 2008), as the hypotheses of interest are assessed based on the posterior probability distribution, and not on p-values which are often misinterpreted.

Another potential drawback associated with existing methodology for non-inferiority testing is the assumption of normally distributed response variables. In general, inferences based on the normality assumption can be misleading when this assumption is not adequate (Ghosh and Gönen, 2008), and more flexible methods would be useful in many settings. More specifically, methods based on scale mixtures of the normal distribution are useful to consider, allowing for heavier-tailed distributions and leading to robust procedures. Thus, it is of practical interest to develop statistical models for noninferiority trials that move beyond the traditional parametric normal model. We develop here robust parametric and semiparametric Bayesian modeling approaches to assess noninferiority. To develop this approach we use mixtures of Dirichlet processes (MDP) (MacEachern, 1994; Escober and West, 1995; MacEachern and Muller, 1998; Ghosh, Basu and Tiwari, 2009) which lead to flexible models for data exhibiting non-normal behavior. Aside from gaining flexibility, our use of the MDP also facilitates an implementation in standard software for Bayesian computing, and this is a practical advantage.

In Section 2 we review the three arm noninferiority trial, and in Section 3 we describe parametric methods for analysis under heteroscedasticity. These methods are then extended in Section 4, we present our semi-parametric Bayes approach, which allows for flexibility under a wide range of

distributions. In Section 5 we describe the Home-Based Blood Pressure Intervention trial, which forms the basis of a simulation study conducted in Section 6. Finally, Section 7 draws conclusions and provides an outlook on future research.

2 Hypothesis testing in a three-arm noninferiority trial

There have been two main approaches adopted for testing in noninferiority trials. The traditional approach first defines a noninferiority margin δ, and then demonstrates that the effect of the experimental treatment is not worse than the effect of the control by more than this amount. This is referred to as the fixed-margin approach (Koch and Rohmel, 2004). The second approach involves directly combining the point estimate and variance from the noninferiority trial with those from historical trials, and is referred to as the synthesis approach (Koch and Rohmel, 2004). In this work we shall follow the traditional fixed-margin approach as described below.

Let $X_{E,i},\ X_{R,j},\ X_{P,k},\ (i = 1, 2, \cdots, n_E;\ j = 1, 2, \cdots, n_R;\ k = 1, 2, \cdots, n_P)$ denote the random variables corresponding to observations taken from the experimental, reference, and placebo groups respectively. We assume that these random variables are mutually independent and that

$$X_{E,i} \overset{\text{i.i.d.}}{\sim} N(\mu_E, \sigma_E^2),\ X_{R,j} \overset{\text{i.i.d.}}{\sim} N(\mu_R, \sigma_R^2),\ X_{P,k} \overset{\text{i.i.d.}}{\sim} N(\mu_P, \sigma_P^2) \tag{1}$$

so that we allow for heteroscedasticity across treatment arms, but assume normally distributed response variables, an assumption that will be relaxed in Section 4. Commonly, for a two-arm trial, the noninferiority hypothesis is formulated as

$$H_0:\ \mu_E - \mu_R \leq \delta \quad \text{vs} \quad H_a:\ \mu_E - \mu_R > \delta \tag{2}$$

where $\delta < 0$ denotes the pre-specified maximal clinically irrelevant amount, and is called the amount of noninferiority margin. The choice of δ in a clinical trial depends on a combination of statistical reasoning and clinical judgement. See the Concept Paper on the development of a Committee for Proprietary Medicinal Products (CPMP, 2005) guidelines for more on the choice of the noninferiority margin δ. Essentially, a rejection of the null hypothesis is required to demonstrate noninferiority support of the experimental treatment to the reference treatment.

When a placebo group is included in the trial, one can formulate δ as a negative fraction f of the unknown difference in mean response between the reference and placebo (Pigeot et al., 2003), that is $\delta = f(\mu_R - \mu_P)$, where f is a fraction ranging between $0 < f < 1$. Assuming, $\mu_R - \mu_P > 0$ and

employing this expression for δ in the hypothesis (2) we obtain

$$H_0 : \mu_E - \mu_R \leq f(\mu_R - \mu_P) \quad \text{vs} \quad H_a : \mu_E - \mu_R > f(\mu_R - \mu_P). \tag{3}$$

Next we let $\theta = (1 + f)$ so that (3) can be written as

$$H_0 : \frac{\mu_E - \mu_P}{\mu_R - \mu_P} \leq \theta \quad \text{vs} \quad H_a : \frac{\mu_E - \mu_P}{\mu_R - \mu_P} > \theta, \tag{4}$$

where θ is the prespecified fraction of the effect of the reference drug (relative to placebo) that we require for the effect of the test drug (relative to placebo), in order to declare noninferiority. According to CPMP (1999), reasonable choices of f include $-\frac{1}{2}, -\frac{1}{3}, -\frac{1}{5}$; however, it should also be mentioned that while making this choice, a clinical consideration should be applied in practice.

The alternative hypothesis in (4) implies that the test treatment achieves more than $\theta \times 100\%$ of the efficacy of the reference treatment, each compared to placebo. Pigeot et al. (2003) have shown that different choices of θ are chosen for different purposes. In particular, noninferiority of test treatment to the reference treatment is evaluated through a test of H_0 in (4) with $1 + f < \theta < 1$; and $0.5 \leq \theta < 1$ (Koch and Tangen, 1999).

For the derivation of the statistical test procedures for the test problem (4), it is helpful to express (4) as:

$$H_0 : \mu_E - \theta\mu_R - (1 - \theta)\mu_P \leq 0 \quad \text{vs} \quad H_a : \mu_E - \theta\mu_R - (1 - \theta)\mu_P > 0 \tag{5}$$

Pigeot et al. (2003) derived a Student-t statistic based on Fieler's confidence interval. They also considered a bootstrap percentile interval as an alternative to Fieler's method in case the assumption of normality does not hold. More recently, Hasler et al. (2007) extended the results to the situation where the group variances are heterogenous.

3 Bayesian Analysis

3.1 Prior distribution

Our first approach is based on a fully parametric model, where we specify a prior distribution for location and scale parameters $(\mu_l, \sigma_l^2),\ l \in \{E,\ P,\ R\}$ using a normal-inverse-gamma distribution $\mu_l|\sigma_l^2 \sim N(\mu_{0l}, \sigma_l^2|\kappa_{0l})$, and

$$\sigma_l^2 \sim \text{Inv-gamma}(\nu_{0l}/2, \sigma_{0l}^2\,\nu_{0l}/2),\ \ l \in \{E,\ R,\ P\}$$

where μ_{0l}, κ_{0l}, ν_{0l}, σ^2_{0l} are fixed hyperparameters. The appearance of σ^2_l in the conditional distribution of $\mu_l|\sigma^2_l$ calibrates the prior information on μ_l to the scale of measurement in the observed data, with $\kappa_{0l} \geq 0$ representing the number of prior observations on this scale. The prior sample size κ_{0l} can be adjusted with reference to the observed sample size n_l, and is an intuitive measure characterizing the degree of prior information on μ_l. The hyperparameters ν_{0l} and σ^2_{0l} are chosen to reflect prior information on the scale parameters σ^2_l. These variance components are typically nuisance parameters and a noninformative prior can be obtained by letting $\nu_{0l} \to 0$, resulting in $p(\sigma^2_l) \propto \sigma_l^{-2}$. Next we condition on $\mu_R - \mu_P > 0$, leading to a truncated prior having density of the form

$$p(\boldsymbol{\mu}, \boldsymbol{\sigma}) \propto I\{\mu_R > \mu_P\} \prod_{l \in \{E,P,R\}} p(\mu_l, \sigma^2_l | \mu_{0l}, \kappa_{0l}, \nu_{0l}, \sigma^2_{0l}) \tag{6}$$

where $I\{\cdot\}$ denotes the indicator function and $p(\mu_l, \sigma^2_l|\mu_{0l}, \kappa_{0l}, \nu_{0l}, \sigma^2_{0l}) = p(\mu_l|\mu_{0l}, \kappa_{0l}, \sigma^2_l)p(\sigma^2_l|\nu_{0l}, \sigma^2_{0l})$, with $p(\mu_l|\mu_{0l}, \kappa_{0l}, \sigma^2_l)$ the density of the $N(\mu_{0l}, \sigma^2_l/\kappa_{0l})$ distribution and $p(\sigma^2_l|\nu_{0l}, \sigma^2_{0l})$ the density of the Inv-gamma$(\frac{\nu_{0l}}{2}, \frac{\sigma^2_{0l}\nu_{0l}}{2})$ distribution.

In practice, analysis can be performed under several choices of prior parameters. Two such extreme choices are often called skeptical and enthusiastic priors. In a clinical trial of superiority, for example, one might center the prior for treatment difference in favor of the experimental arm to represent an enthusiastic prior (and vice versa for the skeptical prior). In the case of a non-inferiority trial with three arms such choices are not immediate. Here we have some suggestions for the practicing Bayesian statistician.

3.2 Posterior Distribution

The non-inferiority hypothesis (**??**) is evaluated under the marginal posterior distribution $[\boldsymbol{\mu}|\boldsymbol{X}]$, where $\boldsymbol{X} = \{X_{E,i},\ X_{R,j},\ X_{P,k}, i = 1, 2, \cdots, n_E;\ j = 1, 2, \cdots, n_R;\ k = 1, 2, \cdots, n_P\}$ denotes the observed data. Under the Gaussian assumption for the response variables, and under the prior (**??**), the density of this posterior arises through the product of three student-t densities, where again, the distribution is truncated so that $\mu_R > \mu_P$

$$p(\mu_E, \mu_P, \mu_R|\boldsymbol{X}) \propto I\{\mu_R > \mu_P\} \prod_{l \in \{E,P,R\}} t_{\nu_{nl}}(\mu_l|\mu_{nl}, \sigma_{nl}) \tag{7}$$

where $t_\nu(x|\mu, \sigma) \propto (1 + \frac{1}{\nu}(\frac{x-\mu}{\sigma})^2)^{-(\nu+1)/2}$ denotes the density function of the student-t distribution on ν degrees of freedom, with location μ and scale σ. For each $l \in \{E, P, R\}$, the parameters of the

posterior distribution are obtained in closed form as $\nu_{nl} = \nu_{0l} + n_l$; $\mu_{nl} = \frac{\kappa_{0l}}{\kappa_{0l}+n_l}\mu_{0l} + \frac{n_l}{\kappa_{0l}+n_l}\bar{X}_l$, and

$$\sigma^2_{nl} = \frac{\nu_{0l}\sigma^2_{0l}}{(\nu_{0l}+n_l)(\kappa_{0l}+n_l)} + \frac{(n_l-1)S^2_l}{(\nu_{0l}+n_l)(\kappa_{0l}+n_l)} + \frac{\kappa_{0l}n_l(\mu_{0l}-\bar{X}_l)^2}{(\nu_{0l}+n_l)(\kappa_{0l}+n_l)^2}$$

where $\bar{X}_l$ and S^2_l, $l \in \{E, P, E\}$ are the corresponding sample mean and variance from group l. Inference and, in particular, calculation of the posterior probability of H_1 in (**??**) is based on drawing samples from the posterior distribution (**??**), which is easily accomplished by drawing independent student-t random variables $t_{\nu_{nl}}(\mu_l|\mu_{nl}, \sigma_{nl})$, in conjunction with rejection sampling to ensure that $\mu_R > \mu_P$.

3.3 Test Procedure

In determining whether the experimental drug is non-inferior or not, it is necessary for an investigator to pick a value for θ which is the required cut-off point in order for the experimental drug to be non-inferior than the active control. Then the clinician finds the posterior probability of the hypothesis (**??**). The experimental drug is said to be non-inferior to the active control if this posterior probability is greater than the some pre-specified cut-off point, say, R_{NI}. Thus, one will declare the experimental drug to be non-inferior if

$$P(H_1 : \frac{\mu_E - \mu_P}{\mu_R - \mu_P} > \theta|Data) \ > R_{\text{NI}}$$

Calculation of the above posterior probability in our case is straightforward. If we can draw T values from the posterior distribution of $\frac{\mu_E-\mu_P}{\mu_R-\mu_P}$ we can estimate this probability by

$$\widehat{P}(H_1 : \frac{\mu_E - \mu_P}{\mu_R - \mu_P} > \theta|Data) = \frac{1}{T}\sum_{l=1}^{T} I(\frac{\mu^l_E - \mu^l_P}{\mu^l_R - \mu^l_P} > \theta)$$

where, $\mu^l_E, \mu^l_P, \mu^l_R$ are respectively the values of μ_E, μ_P, μ_R in the lth iteration of the algorithm.

The choice of R_{NI} is highly consequential. A reasonable default value might be 0.5, which essentially means that, between the null and alternative, the hypothesis with the higher posterior probability is retained. However, in some contexts it may be useful to consider higher values of R_{NI}, depending on the operating characteristics needed.

4 Semiparametric Extensions

There could be instances where X_E, X_R, X_P are skewed or multimodal and thus far from normal. While a natural procedure is to use some ad-hoc transformation to achieve normality, a normalizing

transformation on the original data should be avoided if a more suitable parametric model can be found. This is mainly because transformations sometimes may not be useful as it changes the original unit of the data, which, in turn, makes it difficult to interpret and communicate the findings. In addition, it is often not straightforward, or impossible, to discover the right transformations. To address this need we also consider modeling these responses by a Dirichlet process mixture (DPM) model. This new approach creates a model and inference procedure that is robust to departures from the assumption of normality.

The DPM models have recently become computationally feasible with development of MCMC methods for sampling from the posterior distribution of the Dirichlet process (Escobar, 1994; Escobar and West, 1998; MacEachern, 1994; Ishwaran and James, 2001). The DPM models are by far the most widely used semiparametric Bayesian models, mainly because of the ease of computation, and ability to characterize different shapes.

Suppose, as before, the observed response is $\{X_i\}$. We drop the treatment subscript $l \in \{E, P, R\}$ for the different treatment for the time being. Under the error-DPM model we assume that X_i follows a scale mixture of normal DPM whose density (with respect to Lebesgue measure) is given by

$$f(X_i|\mu_i, G) = \int \phi(X_i|\mu_i, \xi_i\,\zeta)\, dG(\xi_i). \tag{8}$$

Here $\phi(X|\mu, \tau)$ denotes the density of the $N(\mu, \zeta^{-1})$ distribution. The key feature of the model is the assumption that the scale mixing distribution G is unknown, and is modeled by a Dirichlet process (DP) prior with concentration parameter ν and specified base probability measure $G_0(\cdot|\kappa)$ that depends on an unknown parameter vector κ (G and G_0 here denote probability measures although we often refer to them as distributions). This model can be expressed hierarchically as

$$\begin{aligned}
X_i|\,\mu_i,\ \xi_i,\ \zeta &\overset{\text{indep}}{\sim} N\left(\mu_i,\ \text{variance}{=}\xi_i^{-1}\zeta^{-1}\right),\ i = 1, \ldots, n \\
\xi_1, \ldots, \xi_n|\,G &\overset{\text{iid}}{\sim} G \\
G\,|\,\nu, \kappa, G_0 &\sim \text{DP}(\nu, G_0(\cdot|\kappa)) \\
(\zeta, \kappa, \nu) &\sim \pi(\zeta)\,\pi(\kappa)\,\pi(\nu)\,,
\end{aligned} \tag{9}$$

with the mean μ. The Bayesian model specification for the error-DPM model is completed by assigning prior probability models for the hyperparameter vector κ of G_0 and the concentration parameter ν. We will use a $Gamma(s/2, s/2)$ distribution for the base measure, $G_0(.)$ We note here that the class of normal- scale mixtures is quite broad and includes many popular heavier tailed

distributions such as the Logistic family and the t-family of distributions; distributions which are often used in robust statistical procedures.

There are several ways to implement a DPM prior. Recent research has focused on using the following constructive definition of the DP (Sethuraman and Tiwari, 1982; Sethuraman, 1994) to produce MCMC algorithms

$$G(\cdot) = \sum_{r=1}^{\infty} p_r \delta_{Z_r}(\cdot); \quad \text{where } Z_r \overset{\text{iid}}{\sim} G_0(\cdot|\kappa), \tag{10}$$

$$\text{with } p_1 = V_1, p_r = V_r \prod_{j=1}^{r-1}(1 - V_j), \text{ and } V_r \overset{\text{iid}}{\sim} \text{Beta}(1, \nu), \ r \geq 1 \tag{11}$$

If we truncate the sum in (**??**) at a large integer $R > 0$ we obtain the models considered in Ishwaran and Zarepour (2002), Ishwaran and James (2001, 2002). This reduces $G(\cdot)$ into finite dimensional form as $G = \sum_{r=1}^{R} p_r \delta_{Z_r}(\cdot)$. The model in (**??**) can then be expressed hierarchically as

$$\begin{aligned} X_i|\mathbf{Z} = (Z_1, \cdots, Z_R), s, \mu_i &\overset{\text{ind}}{\sim} \phi(X_i|\mu_i, \zeta\, Z_{s_i}), \quad i = 1, \ldots, n \\ s_i|p &\overset{\text{iid}}{\sim} \sum_{r=1}^{R} p_r \delta_r(\cdot), \end{aligned} \tag{12}$$

where s_i is the latent mixture component indicator for the i^{th} observation, $s = (s_1, \ldots, s_n)$, $Z_r \overset{\text{iid}}{\sim} G_0(\cdot|\kappa)$, and the distribution of $p = (p_1, \ldots, p_R)$ is specified by the stick-breaking construction. The so-called blocked Gibbs sampler updates $\mathbf{Z}, s$ and p in multivariate blocks. Another advantage here is that since the DPM structure is reduced to a finite mixture model by this truncation and a non-conjugate structures can be more easily handled now. The effect of truncation on the distribution of functionals of a Dirichlet process has been studied by Ohlssen, Sharples, and Spiegelhalter (2007), and Ishwaran and Zarepour (2002). Ishwaran and Zarepour (2002) suggest taking $R = \sqrt{n}$ for large n, and $R = n$ for small n. We follow the suggestion of Ishwaran and Zarepour (2002) in choosing R.

Based on the above idea, we now put a DPM prior on the distribution of the experimental drug, active control and the placebo as follows:

$$X_{E,i} \sim N(\mu_E, \gamma_{1i}^{-1}\sigma^2), \quad X_{R,j} \sim N(\mu_R, \gamma_{2j}^{-1}\sigma^2), \quad X_{P,k} \sim N(\mu_P, \gamma_{3k}^{-1}\sigma^2) \tag{13}$$

$$\gamma_{Ei} \sim G_1, \quad , \gamma_{Rj} \sim G_2, \quad \gamma_{Pk} \sim G_3 \tag{14}$$

$$G_1 \sim DP(\nu_1, G_{01}), \quad G_2 \sim DP(\nu_2, G_{02}), \quad G_3 \sim DP(\nu_3, G_{03}) \tag{15}$$

$$G_{0k} \sim \text{Gamma}(a, b), \quad k = 1, 2, 3 \tag{16}$$

$$\nu_k \sim \text{Gamma}(c, d), \quad k = 1, 2, 3 \tag{17}$$

5 Example: Home-Based Blood Pressure Interventions

The method was motivated by the design of a randomized trial to assess the effectiveness of organizational interventions at improving blood pressure (BP). Home health care is a non-institutional setting that provides services to a high-risk population characterized by multiple chronic conditions and significant needs for both medical and self-care management. Several patients in home health care have essential hypertension (HTN), but for various reasons management of blood pressure for patients with HTN has traditionally received less attention than the management of other chronic conditions such as diabetes and chronic obstructive pulmonary disease. For this reason a randomized trial is proposed that will examine the effectiveness of two organizational interventions aimed at improving BP control among a high-risk home care population. The two interventions to be tested include (i) a "basic" intervention delivering key "just-in-time" information to nurses, physicians and patients while the patient is receiving traditional post-acute home health care; and (ii) an "augmented" intervention transitioning patients to a Home-Based HTN Support Program that extends the information, monitoring and feedback available to patients and primary care physicians for an 18-month period beyond an index home care admission. Usual care is included as a third arm. The primary goal is to see if the basic intervention is at least as good as the augmented one, relative to the usual care. In the terminology of Section 3, usual care is the placebo group, augmented care is the reference group and basic care is the experimental group.

As this trial has been designed but not completed the simulation studies in the next section are motivated by this example. Note that, because of the nature of the delivery of home health care, it was decided to randomize the caregivers (nurses) instead of patients, thus making this a cluster-randomized study.

5.1 Simulation Studies

We consider data generation under two scenarios, normally distributed values of the logarithm of the blood pressure as well as a heavier-tailed distribution (Laplace distribution) as the second case, to assess the robustness of our method and its semiparametric extension. The Laplace distribution has density $f(x) = \frac{1}{2\sigma}\exp(-\frac{1}{b}|x-\mu|)$.

Since the trial is 1:1:1 randomized we are taking $n_E = n_R = n_P = 150$. Using the preliminary data that was gathered to design the trial we are estimating $\mu_P = 1$, $\mu_R = 4.9$, $\sigma_E^2 = 5$, $\sigma_R^2 = 3$ and $\sigma_P^2 = 1$. We consider two clinically different scenarios: a non-inferior scenario where $\mu_E = 5$ and an inferior scenario on were $\mu_E = 3$. Also we considered a range of values for μ_E as described next.

When clustering is taken into account, the entire data generation process is replicated with 50 caregivers randomized to each arm and each caregiver is assumed to have three patients keeping the sample size at 150 per arm. Patients within a caregiver are assumed exchangeable with a correlation of 0.1.

The values of the hyperparameters associated with means and variances are listed in the setup below. For the DP models, we take the base-measures, G_{0k} to be $Gamma(s/2, s/2)$ with $s \sim uniform(1, 100)$ and the concentration parameter $\nu_k \sim Gamma(\sqrt{n}, 1)$ where n is the sample in the given group.

Here we present the results of two simulation studies examining the performance of the proposed methods. In each case, for a given set of parameter values μ_l, σ_l, $l \in \{E, P, R\}$; sample sizes n_l, $l \in \{E, P, R\}$; and effectiveness threshold θ; the performance is evaluated through the expected posterior probability of non-inferiority $E[P(H_1|Data)]$, where the expectation is taken with respect to the sampling distribution, generating repeated realizations of the data. In the first study, we assume a Gaussian sampling distribution, applying the model described in Section 2, and illustrate the impact of prior information under various sample size assumptions. In the second study, heavier tailed sampling distributions are considered, and we examine the performance of the Gaussian model under misspecification, and compare this performance with that of the semiparametric scale mixture model proposed in Section 3.

Design of simulations 1:

- Assume $n_E = n_P = n_R = n$ and $\theta = 0.8$
- Set $\sigma_E^2 = 5$, $\sigma_R^2 = 3$, $\sigma_P^2 = 1$
- Set $\mu_P = 1$, $\mu_R = 4.9$ and let μ_E vary from $\mu_E = 2.95, 2.96, \ldots, 6.01$ so that $\frac{\mu_E - \mu_P}{\mu_R - \mu_P}$ covers a range of values from 0.5 to 1.2.
- For a given value of $\frac{\mu_E - \mu_P}{\mu_R - \mu_P}$, generate data, and compute $P(H_1|Data)$ under the three priors considered. Repeat $n_{sim} = 1000$ times and compute the average as a Monte Carlo estimate of $E[P(H_1|Data)]$ for each of the three priors.
- For the three priors considered, we set $\nu_{0l} = 0$ so that they are all uninformative with respect to the variance parameters. The three priors vary according to hyperparameters κ_{0l} and μ_{0l}, which represent, respectively, the prior sample size and prior mean of μ_l
  1. noninformative prior: $\kappa_{0l} = 0,\ \mu_{0l} = 0$, $l \in \{E, R, P\}$
  2. enthusiastic prior: $\kappa_{0l} = 10$, $l \in \{E, R, P\}$, $\mu_{0E} = 4$, $\mu_{0R} = 3$, $\mu_{0P} = 1$
  3. skeptical prior: $\kappa_{0l} = 10$, $l \in \{E, R, P\}$, $\mu_{0E} = \mu_{0P} = 1$, $\mu_{0E} = 3$
- Repeat over the entire range of values for $\frac{\mu_E - \mu_P}{\mu_R - \mu_P}$ in order to generate a curve for each prior.
- Generate curves for four different sample sizes: $n = 20, 50, 100, 150$

Results are depicted in Figure 1. Within each panel of Figure 1, we see that as the prior moves from enthusiastic to skeptical, the posterior probability of H_1 decreases for each value of $(\mu_E - \mu_P)/(\mu_R - \mu_P)$, as expected. Moving from one panel to the other, we observe that the effect of increasing sample size is making each curve more steep (and hence more likely to reject H_1) and also closer to one another (hence robust to prior specification). Both of these behaviors are intuitive and expected. Furthermore, since $\theta = 0.8$, one would consider $(\mu_E - \mu_P)/(\mu_R - \mu_P) > 0.8$ corresponding to the case where the alternative is true. The non-informative prior yields approximately 0.5 posterior probability at that point, for all sample sizes considered, suggesting that it is well calibrated.

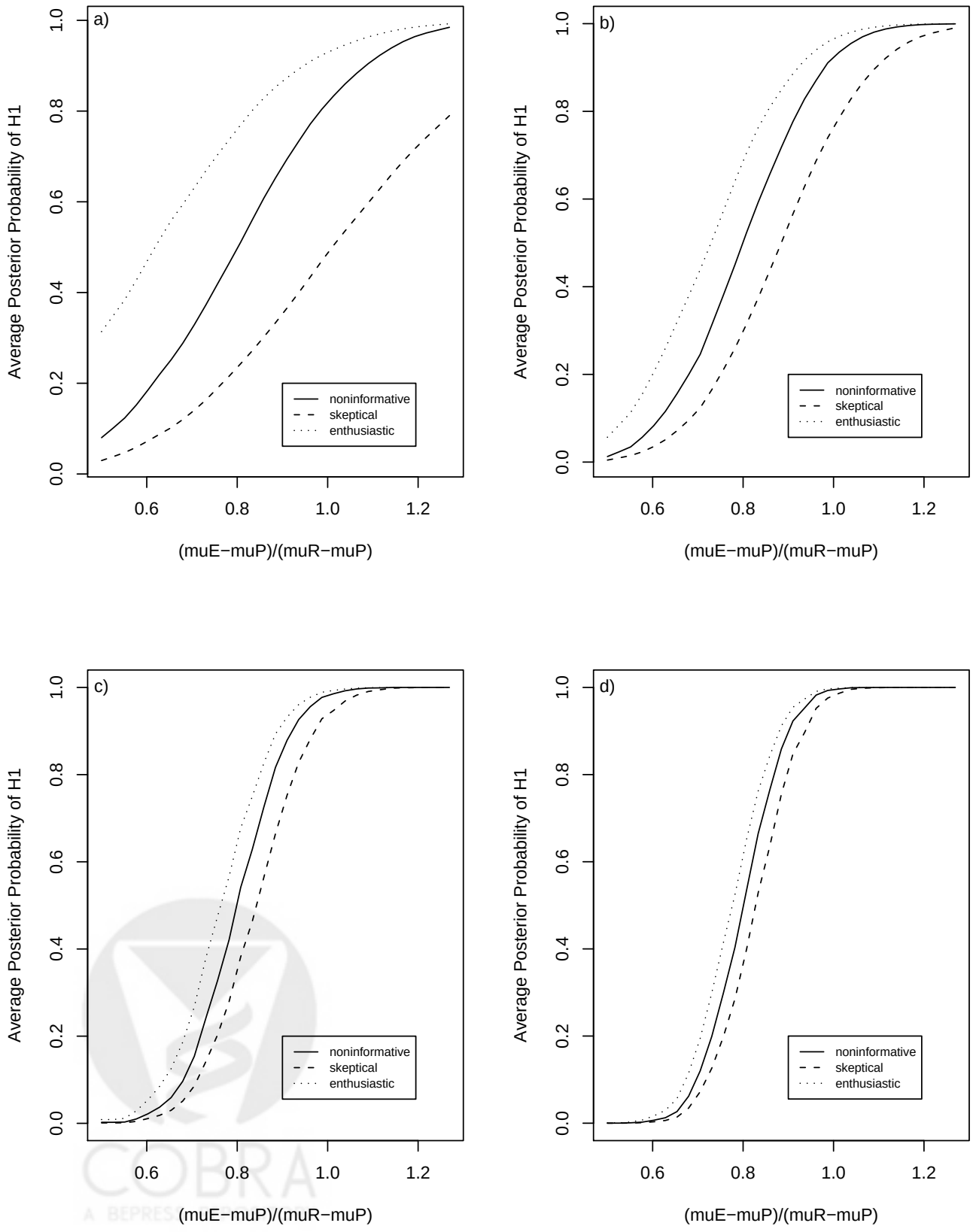

Figure 1: Study 1: Curves depicting expected posterior probability of non-inferiority $E[P(H_1|Data)]$ as a function of $\frac{\mu_E-\mu_P}{\mu_R-\mu_P}$ for each of the three priors considered. These based on $\theta = 0.8$, $\sigma_E^2 = 5$, $\sigma_R^2 = 3$,$\sigma_P^2 = 1$, $\mu_P = 1$, $\mu_R = 4.9$ and $n_E = n_P = n_R = n$ with (a) $n = 20$; (b) $n = 50$; (c) $n = 100$; (d) $n = 150$.

Design of simulations 2:

- Replace the Gaussian sampling distribution with data simulated from (1) a t-distribution with $DF = 2$ and (2) a Laplace distribution.
- Location and scale parameters are set as in study 1, with $\sigma_E^2 = 5$, $\sigma_R^2 = 3$, $\sigma_P^2 = 1$, $\mu_P = 1$, $\mu_R = 4.9$ and we let μ_E vary from $\mu_E = 2.95, 2.96, \ldots, 6.01$ so that $\frac{\mu_E - \mu_P}{\mu_R - \mu_P}$ covers a range of values from 0.5 to 1.2.
- Assume $n_E = n_P = n_R = 20$ and $\theta = 0.8$
- Fit the Gaussian model and generate curves depicting $E[P(H_1|Data)]$ for each of the three priors, based on $n_{sim} = 100$ data replications.
- Fit the DP scale mixture model and generate curves depicting $E[P(H_1|Data)]$ for each of the three priors, based on $n_{sim} = 100$ data replications. Compare to results obtained from Gaussian model.
- The noninformative, skeptical and enthusiastic priors on location and scale parameters are assumed to be the same as in study 1. For the DP model, winbugs does not allow improper priors, so we approximate the prior described in Section 3.1 by taking $\nu_{0l} = 0.001$ (as opposed to $\nu_{0l} = 0$) in the prior for variance components; and $\kappa_{0l} = 0.001$ (as opposed to $\kappa_{0l} = 0$) in the noninformative prior for the location parameters.

Results are depicted in Figure 2. The upper three panels correspond to the t_2 simulation for the three different priors: non-informative (a), skeptical (b) and enthusiastic (c). For the non-informative prior, we see that the posterior probabilities of DP and Gaussian cross around $(\mu_E - \mu_P)/(\mu_R - \mu_P) = 0.8$, the value of θ used in the simulations, again showing excellent prior calibration. For values $(\mu_E - \mu_P)/(\mu_R - \mu_P) > 0.8$ the DP model gives higher posterior probabilities, and hence more power, than the Gaussian model. For values $(\mu_E - \mu_P)/(\mu_R - \mu_P) < 0.8$ the DP model gives lower posterior probabilities which implies an appropriate level of conservatism under the null hypothesis. The results from the skeptical model are also similar, the only difference being that posterior probabilities for both models are proprtionately smaller than those observed with the non-informative prior. Hence the price of insisting on a parametric model when it is wrong, results in an increase in the number of both false positive and false negative decisions. While the

enthusiastic model shows a similar pattern, the curves cross earlier since it takes a smaller signal to convince the enthusiast.

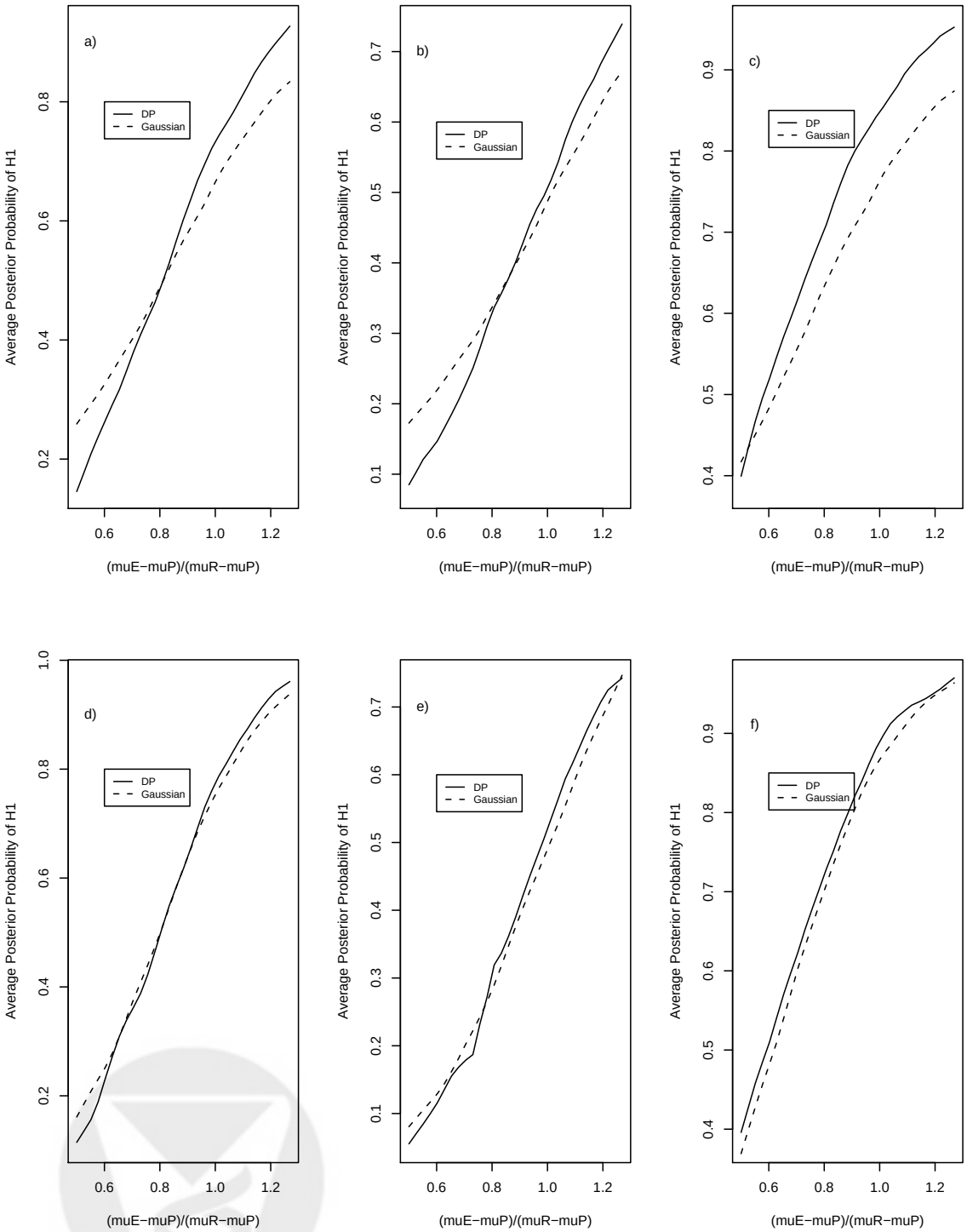

Figure 2: Study 2: Curves depicting expected posterior probability of non-inferiority $E[P(H_1|Data)]$ as a function of $\frac{\mu_E-\mu_P}{\mu_R-\mu_P}$ for the Gaussian and DP models. These based on $\theta = 0.8$, $\sigma_E^2 = 5$, $\sigma_R^2 = 3$,$\sigma_P^2 = 1$, $\mu_P = 1$, $\mu_R = 4.9$ and $n_E = n_P = n_R = n = 20$. Panels (a), (b) and (c) correspond to data simulated from a t_2 distribution and models based on the noninformative, skeptical and enthusiastic prior respectively. Panels (d), (e) and (f) correspond to data simulated from a Laplace distribution and models based on the noninformative, skeptical and enthusiastic prior respectively.

6 Discussion

In this paper we have developed two Bayesian approaches for hypothesis testing in non-inferiority trials under a three-arm design. One is a parametric model, relying on normality of the data and the other is nonparametric using a Dirichlet process prior. The latter has the advantage of accommodating data from skewed or thick-tailed distributions without requiring a transformation. Our method has the advantage of accommodating heteroscedasticity, a common simplifying assumption in the literature, which is unlikely to hold in many cases. Finally we take full advantage of the Bayesian framework both conceptually, by using the posterior probabilities for inference, and computationally, by using MCMC to accommodate the Dirichlet process prior.

We applied our method to an ongoing trial investigating an intervention to blood pressure management in the home care setting. Since the data are not yet available, we simulated under the conditions presumed in the study protocol. The results suggest that the method works well under a variety of priors. While the parametric method is efficient when correct, it may suffer considerably when there are substantial deviations. We recommend routine use of the DPM unless there is strong support for the parametric assumptions.

We are currently working on extensions to binary and censored data, where similar principles apply, although implementations may differ substantially.

References

[1] CPMP(1999) Concept Paper on the development of a Committee for Proprietary Medicinal Products (CPMP). Points to consider on biostatistical/methodological issues arising from recent CPMP discussions on licensing applications: choice of delta. Available at: http://www.emea.eu.int/pdfs/human/ewp/215899en.pdf

[2] DAgostino R, B, Massaro J, M, Sullivan L, M (2003). Non-inferiority trials: design concepts and issuesthe encounters of academic consultants in statistics, *Statistics in Medicine* ; **2** 2: 169186.

[3] Escober, M. D. and West, M. (1995), 'Bayesian density estimation and inference using mixtures, *Journal of the American Statistical Association*, 90, 577-580.

[4] EMEA European Medicines Agency (2005). Guideline on the Choice of the Non-Inferiority Margin. Availabel at http://www.ema.europa.eu/pdfs/human/ewp/215899en.pdf.

[5] Gelman, A. and Carlin, J.B. and Stern, H.S. and Rubin, D.B. (2004), Bayesian Data Analysis, 2nd ed, London: CRC Press.

[6] Gill, J. (2002), Bayesian Methods: A Social and Behavioral Sciences Approach. Chapman & Hall/CRC. New York.

[7] Ghosh P, and Gönen M. (2008). Bayesian modeling of multivariate average bioequivalence, *statisics in Medicine*, **27**: 2402-2419.

[8] Ghosh, P., Basu, S., and Tiwari, R.C. (2009), Bayesian Analysis of Cancer Rates from SEER Program Using Parametric and Semiparametric Joinpoint Regression Models, Journal of the American Statistical Association, 104, 439-452.

[9] Hasler, M, Vonk, R, Hothorn, L.A. (2008). Assessing non-inferiority of a new treatment in a three-arm trial in the presence of heteroscedasticity, *Statistics in Medicine*, 490503.

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

[11] Ishwaran, H. and James, L. (2002), Dirichlet process computing in finite normal mixtures: smoothing and prior information, *Journal of Computational and Graphical statistics* , 11, 508-532.

[12] Ishwaran, H. and Zarepour, M. (2002), Dirichlet prior sieves in finite normal mixtures, *Statistica Sinica*, 12, 941-963.

[13] Jones B, Jarvis P, Lewis J, A, Ebbutt A, F (1996). Trials to assess equivalence: The importance of rigorous methods, *British Medical Journal*, **313**: 3639.

[14] Kieser, M, and Friede, T (2007). Planning and analysis of three-arm non-inferiority trials with binary endpoints, *Statistics in Medicine*, **26**, 253273.

[15] Koch, A, and Rohmel, J. (2004). Hypothesis testing in the gold standard design for proving the efficacy of an experimental treatment relative to placebo and a reference, *Journal of Biopharmaceutical Statistics*, **14**, 315-325.

[16] Koti, K.M (2007). Use of the Fieller-Hinkley distribution of the ratio of random variables in testing for noninferiority, *Journal of Biopharmaceutical Statistics*, **1** 7: 215-228.

[17] MacEachern, S. N. (1994). Estimating normal means with a conjugate style Dirichlet process prior. *Communications in Statistics: Simulation and Computation* **2** 3, 727-741.

[18] MacEachern, S., and Müller, P. (1998), Estimating Mixture of Dirichlet Process Models *Journal of Computational and Graphical Statistics*, 7, 223-238.

[19] Mimi, K., and Xue, X. (2004). Likelihood ratio and a Bayesian approach were superior to standard noninferiority analysis when the noninferiority margin varied with the control event rate, *Journal of clinical epidemiology* ; **5** 7: 1253-1261.

[20] Munk A, Trampisch H, J(2005). Therapeutic equivalenceclinical issues and statistical methodology in noninferiority trials, *Biometrical Journal*; **4** 7: 79.

[21] 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.

[22] Pigeot, I, Schafer, J, Rohmel, J, Hauschke, D (2003). Assessing non-inferiority of a new treatment in a three-arm clinical trial including a placebo, *Statistics in Medicine*, **22**:883899.

[23] Rohmel, J. (1998). Therapeutic equivalence investigations: statistical considerations, *Statistics in Medicine*, **17**, 1703-1714.

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

[25] 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.), Acdemic Press, 305-315.

[26] Temple, R, and Ellenberg, S (2000). Placebo-controlled trials and active-control trials in the evaluation of new treatments. part 1: Ethical and scientific issues, *Annals of Internal Medicine*, **133**, 455-463.

[27] Snapinn, S, S. (2000). Noninferiority trials, *Current Controlled Trials in Cardiovascular Medicine*; **1** : 1921.