Original Article

Research Synthesis Methods

Received 27 April 2015, Revised 26 February 2016, Accepted 08 March 2016 Published online 10 June 2016 in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/jrsm.1209

Random-effects meta-analysis for systematic reviews of phase I clinical trials: Rare events and missing data

Mi-Ok Kim,[a]* Xia Wang,[b] Chunyan Liu,[a] Kathleen Dorris,[c] Maryam Fouladi[d] and Seongho Song[b]

Phase I trials aim to establish appropriate clinical and statistical parameters to guide future clinical trials. With individual trials typically underpowered, systematic reviews and meta-analysis are desired to assess the totality of evidence. A high percentage of zero or missing outcomes often complicate such efforts. We use a systematic review of pediatric phase I oncology trials as an example and illustrate the utility of advanced Bayesian analysis. Standard random-effects methods rely on the exchangeability of individual trial effects, typically assuming that a common normal distribution sufficiently describes random variation among the trial level effects. Summary statistics of individual trial data may become undefined with zero counts, and this assumption may not be readily examined. We conduct Bayesian semi-parametric analysis with a Dirichlet process prior and examine the assumption. The Bayesian semi-parametric analysis is also useful for visually summarizing individual trial data. It provides alternative statistics that are computed free of distributional assumptions about the shape of the population of trial level effects. Outcomes are rarely entirely missing in clinical trials. We utilize available information and conduct Bayesian incomplete data analysis. The advanced Bayesian analyses, although illustrated with the specific example, are generally applicable. © 2016 The Authors. *Research Synthesis Methods* Published by John Wiley & Sons Ltd.

Keywords: meta-analysis; systematic review; sparse outcomes; missing data; semi-parametric Bayesian analysis

1. Introduction

Systematic reviews and meta-analyses have grown in popularity as the need to base decisions on the totality of relevant and sound evidence in medicine has been increasingly recognized (Sutton and Higgins, 2008). They are particularly useful for phase I trials as individual trials are generally underpowered to appropriately establish clinical and statistical parameters in order to guide future drug development. A high percentage of zero or missing outcomes often complicate such efforts as illustrated below. In this paper, we use a case study and describe the utility of advanced Bayesian analysis, specifically semi-parametric analysis with a Dirichlet process prior and incomplete data analysis. We are concerned with random-effects analysis in this paper. In one treatment group studies like phase I without within trial comparators, discrepancies in population, outcomes, exposures/interventions, design and/or conduct across individual trials may affect how the effects of the single treatment arms are realized across individual trials directly. The assumptions of random individual trial effects are well justified.

The case study example comes from a systematic review of pediatric phase I oncology trials in patients with relapsed or refractory solid tumors (Dorris *et al*., 2015). This study reviewed publications that evaluated the safety and efficacy of molecularly targeted and cytotoxic agents. Molecularly targeted agents, the focus of recent drug

[a]*Division of Biostatistics and Epidemiology, Cincinnati Children's Hospital Medical Center, Cincinnati, OH 45229, USA*
[b]*Department of Mathematical Sciences, University of Cincinnati, Cincinnati, OH 45221, USA*
[c]*Center for Cancer and Blood Disorders, Children's Hospital Colorado, Aurora, CO 80045, USA*
[d]*Division of Oncology, Cincinnati Children's Hospital Medical Center, Cincinnati, OH 45229, USA*
**Correspondence to: Mi-Ok Kim, Division of Biostatistics and Epidemiology, Cincinnati Children's Hospital Medical Center, Cincinnati, OH 45229, USA.*
E-mail: miok.kim@cchmc.org

development efforts, primarily inhibit tumor cell growth without necessarily killing tumor cells. In contrast, cytotoxic agents kill tumor cells. This systematic review aimed to establish appropriate clinical and statistical parameters to guide future drug development for novel molecularly targeted drugs by examining whether the previously identified efficacy and toxicity rates of phase I cytotoxic trials can be generalized to studies of molecularly targeted agents. The efficacy and safety outcomes were defined by rare events, reporting zero counts in 66% and 48% of the molecularly targeted and the cytotoxic agent trials, respectively. Incomplete toxicity data further complicated meta-analysis of this systematic review.

The difficulty of handling zero counts is well known in the meta-analysis literature when two treatment group studies are concerned (e.g. Friedrich *et al.*, 2007 and references therein). Random effects meta-analysis for one treatment group studies is similarly affected. Standard response rate measures (e.g. the odds on the log scale for binary outcomes) may become undefined along with their variances, and methods relying on such statistics may not be directly applicable. Generalized linear mixed models (GLMM) method and full Bayesian (FB) methods have been introduced specifically to address this issue (Kooiman *et al.*, 2012, Hamer *et al.*, 2012, Appelman-Dijkstra *et al.*, 2011, Cai *et al.*, 2010, Singh *et al.*, 2009, Sweeting *et al.*, 2004, Platt *et al.*, 1999, Smith *et al.*, 1995). However, these methods rely on the exchangeability of individual trial effects, typically assuming that a common normal distribution sufficiently describes random variation among them. The assumption may not hold when a high percentage of zero events exist.

We conduct Bayesian semi-parametric analysis with a Dirichlet process prior and examined the assumption. The Dirichlet process prior is assumed for the population of individual trial level effects and allows a general shape for the population, such as a heavy tailed or multi-modal distribution. Under this prior, we obtain the posterior predictive distributions of the individual trial effects and estimate the population of individual trial effects.

The semi-parametric analysis is also useful for visually summarizing individual trial data. Summary statistics of individual trial data may become undefined with zero counts and not available, for example, for constructing forest plots. The semi-parametric analysis provides alternative statistics based on the posterior distributions of individual trial effects. The statistics are obtained free of parametric distribution assumptions about the population of individual trial level effects and are only minimally affected by the estimation of the population means. In contrast, individual trial estimates from GLMM or FB analysis are pulled toward the population means under the exchangeability of individual trial effects assumption and are not appropriate for the inspection of potentially outlying trial level effects. Similar Bayesian semi-parametric models have been considered in the literature (Burr and Doss, 2005, Branscum and Hanson, 2008), but their utility for improving model diagnostics was not examined, particularly as an enabling tool of visual inspection of data with many zero counts as described.

Not all trials report the same type of information, and systematic reviews uncommonly include a substantial portion of missing data. In the given case study, toxicity data were incomplete with dose limiting toxicity (DLT) outcomes missing in 23 trials (21%) and grade 3 or 4 toxicity outcomes missing in 67 trials (60%). However, only 2 of 89 total studies had toxicity information completely missing. We use a parametric Bayesian approach, and model the relationship between the overall grade 3 or 4 toxicities and the DLT events and other auxiliary information. This Bayesian incomplete analysis only excluded the two studies with completely missing toxicity information, salvaging 21 trials with missing DLT outcomes.

The rest of this paper is structured as follows: Section 2 presents the example phase I systematic review in detail. This serves as a real life application, of which the analysis results by different random effects meta-analysis methods are compared. Section 3 describes the application of standard random-effects meta-analysis methods to the case example. Section 4 compares the case study results. Section 5 presents the semi-parametric Bayesian analysis and the Bayesian incomplete DLT data analysis. Section 6 concludes the paper.

2. A case study: systematic review of pediatric phase I oncology trials

This systematic review included publications from 1990 to 2010 that studied the safety and efficacy of molecularly targeted and cytotoxic agents in pediatric phase I oncology trials. Molecularly targeted drugs target key molecular pathways that are disrupted or unregulated in specific cancers and may inhibit tumor cell growth without necessarily killing tumor cells. These drugs have been the focus of drug development increasingly over the past decade with the hope that rationally targeted therapies may be the key to finding cures for cancer in contrast to traditional cytotoxic agents that use nonspecific mechanisms to kill tumor cells. The systematic review aimed to establish appropriate clinical and statistical parameters to guide future clinical trial designs for novel molecularly targeted drugs by examining whether the previously identified efficacy and toxicity rates of phase I cytotoxic trials are generalizable to studies of molecularly targeted agents. We refer to Dorris *et al.* (2015) for the details.

The study identified 89 phase I studies with 30 studies investigating 26 molecularly targeted drugs and 59 studies evaluating 37 cytotoxic agents. Accounting for multiple strata, a total of 111 trials were included. A substantial number of the trials had small sample size. Fewer than 20 patients were available for assessment of the efficacy endpoints in 47.5% of the trials. In 10.1% of the trials, fewer than 10 were available. Citation information about individual trials is included in Supplemental Table 1.

The primary efficacy outcome was overall objective response that indicates complete resolution of all radiographic evidence of tumors or regression of the primary tumor greater than 25%. The primary safety or toxicity outcomes were dose-limiting toxicities (DLT). Both were defined by rare events; for example, zero events were reported for the primary efficacy outcome in 66% and 48% of the molecularly targeted and cytotoxic agent trials, respectively. Toxicity data contained a substantial portion of missing outcomes. DLT outcomes were missing in 23 trials (21%). Secondary toxicity outcomes, grade 3 or 4 toxicity events, were missing in 67 trials (60%). However, only two studies out of the 89 total had toxicity information completely missing.

3. Standard random-effects methods and applications to the case study

We model the binary primary efficacy outcome of patient j in a trial $i(Y_{ij})$ using a Bernoulli distribution:

$$Y_{ij}|p_i \sim Bernoulli\ (p_i)\text{for} j = 1, \ldots, n_i$$

where p_i is the true rate of the efficacy outcome in the trial i and n_i is the sample size. The individual trial data are summarized as follows: for each of $i = 1, \ldots, N$ trials,

$$\sum_j Y_{ij}|p_i, n_i \sim Binomial\,(n_i, , p_i). \tag{1}$$

True rates of binary outcomes are often considered on the logit scale and we denote the logit transformation by $\eta(\cdot)$, so that $\eta(p_i) = \log(p_i/(1-p_i))$. Given $\widehat{p}_i = \left(\sum_j Y_{ij}\right)/n_i$ standard random-effects methods assume

$$\eta(\widehat{p}_i) \sim N\left(\eta(p_i), \sigma_i^2\right), i = 1, \ldots, N \tag{2}$$

where σ_i^2 denote the variances of the trial level statistics on the logit scale. We let $X_i = x(x = 0, 1)$ denote the drug group of the single treatment arm tested in each trial with $x = 1$ for trials in the molecularly targeted drug group and $x = 0$ for trials in the cytotoxic agent drug group. Standard methods further assume that true response rates vary from trial to trial, which typically is described by normal distributions with the means μ_x and the variances τ_x^2;

$$\eta(p_i)\big|X_i = x \sim N\left(\mu_x, \tau_x^2\right)\text{for } \ x = 0, 1. \tag{3}$$

The between trial variances are also typically assumed same $\left(\tau_x^2 = \tau^2\right)$.

3.1. *DerSimonian and Laird's random effects model method*

DerSimonian and Laird's method (DerSimonian and Laird, 1986) assumes a simple random effects model, which is the model (3) not specifying the distributions of the true responses to be normal distributions. With the normal distribution specification the method can be explained as an empirical Bayes (EB) method in the sense that the computation and estimators are equivalent to those of EB method replacing the true parameter values in the model (2) and (3) with corresponding sample estimates. With $\widehat{\eta}_i = \eta(\widehat{p}_i) = \log(\widehat{p}_i/(1-\widehat{p}_i))$ we have

$$Var(\widehat{\eta}_i) = 1/[n_i p_i(1-p_i)] \tag{4}$$

by the delta method. The EB method replaces σ_i^2 with $\widehat{\sigma}_i^2 = 1/[n_i\widehat{p}_i(1-\widehat{p}_i)]$. It then computes the between-trial variance estimate $(\widehat{\tau}_x^2)$ by the method of moments based on $\widehat{\eta}_i$, $i = 1, \ldots, N$. Given $\widehat{\tau}_x^2$, σ_i^2 and $\widehat{\eta}_i$, $i = 1, \ldots, N$, the EB method provides estimates of the population mean (μ_x) given by

$$\widehat{\mu}_x = \sum_{X_i = x} \upsilon_i \widehat{\eta}_i / \sum_{X_i = x} \upsilon_i,$$

where $\upsilon_i = 1/\left[\widehat{\tau}_x^2 + \widehat{\sigma}_i^2\right]$. For the individual trial response rates on the transformed scale, the EB method provides $E\left[\eta(p_i)|data\right] = \left(\frac{\widehat{\eta}_i}{\widehat{\sigma}_i^2} + \frac{\widehat{\mu}_x}{\widehat{\tau}_x^2}\right) / \left(\frac{1}{\widehat{\sigma}_i^2} + \frac{1}{\widehat{\tau}_x^2}\right)$ as estimates.

With zero events, $\widehat{p}_i = 0$, and the quantities $\widehat{\eta}_i$ and $\widehat{\sigma}_i^2$ become undefined. This problem can be avoided by adding a small constant. We used an R package called *metafor* (Viechtbauer, 2010) with a default value of 0.5 in the analysis of the case study.

One may consider different transformations and avoid correcting zeros by adding a small constant arbitrarily. Variance stabilizing transformations are an option. We considered Freeman–Tukey's transformation (Freeman and Tukey, 1950). With the transformation we have

$$\widehat{\eta}_i = \eta(\widehat{p}_i) = \frac{1}{2}\left(\arcsin\sqrt{\frac{\widehat{p}_i}{1+1/n_i}} + \arcsin\sqrt{\frac{\widehat{p}_i + 1/n_i}{1+1/n_i}}\right).$$

We estimated the within-trial variance with the asymptotic variance, so that

$$\widehat{\sigma}_i^2 = 821/(n_i + 1/2). \tag{5}$$

The EB random effects method was applied similarly.

We note that the EB method is subject to bias inherently: it does not account for the source of variability when substituting the parameters with the estimates. In this particular case it does not make allowance for the imprecision in $\widehat{\mu}_x$, $\widehat{\tau}_x^2$ or $\widehat{\sigma}_i^2$'s and therefore underestimates the variance.

3.2. GLMM method

GLMM method also often considers the true response rate on the logit transformed scale and assumes the normal models (3). Unlike the EB method, it assumes model (1) instead of (2) and readily admits zero counts. The specific GLMM model under consideration in this case study is generalized linear logistic regression model with random study effects. We used the pseudo-likelihood method as in Wolfinger and Oconnell (1993) and Breslow and Clayton (1993) with doubly iterative algorithm to fit the model. This method finds modes of the log-posterior distribution as random effects estimates. We used the SAS GLIMMIX procedure for computation with the sample size $n_i' s$ treated as offset terms.

3.3. Full Bayesian (FB) method

A full Bayesian (FB) model also typically includes models (1) and (3) and readily admits zero counts similarly as the GLMM. The FB method differs from the GLMM in that the model parameters are also considered random quantities. Without loss of generality, a prior distribution $\pi(\phi)$ is assumed for the joint distribution of the population mean and variance parameters as follows:

$$\mu_1, \mu_0, \tau_1^2, \tau_0^2 \sim \pi(\phi), \tag{6}$$

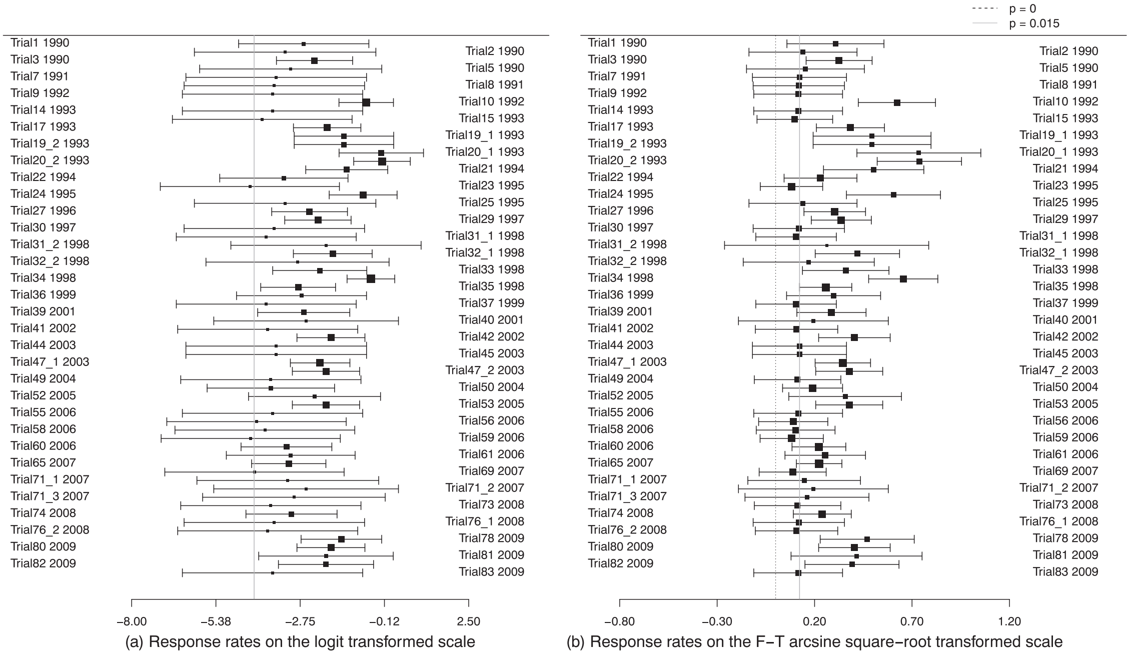
where ϕ denotes a vector of appropriately specified hyper-parameters.

A prior distribution is typically based on evidence external to the study in question or on subjective *a priori* beliefs. In the case of systematic reviews, evidence unrelated to the individual trials that systematic reviews examine is qualified as external evidence. The difficulty of and potentially subjective decisions involved in specifying the prior distribution are often cited as a major disadvantage of Bayesian methods (Sutton and Abrams, 2001). The empirical Bayes approach described in the earlier section avoids this difficulty by substituting parameters with sample estimates. We refer to Sutton and Abrams (2001) for the discussion of the importance and influence of the prior specification and choices available specifically for a meta-analysis.

In the given case study, we used vague priors, a normal prior N (0, 10^6) for μ_0 and μ_1, and a uniform prior U(0, 10) for τ_0 and τ_1. Vague or defused priors mitigate the impact of subjective prior specification (Gelman *et al.*, 2013). Gibbs sampling can be conveniently carried out for computation and posterior inference. Some details are as follows; three independent chains of 100 000 Markov chain Monte Carlo (MCMC) simulation samples were generated after 50 000 of burn-ins and by retaining every 20th iteration. WinBUGS implemented in R package called R2WinBUGS version 2.1-18 (Sturtz *et al.*, 2005) was used. The posterior mean values were used as summary estimates with 95% central credibility intervals. The posterior probability that the mean response rate of the cytotoxic agents is higher than that of the molecularly targeted agents was calculated using the MCMC samples and was provided as an evidence for the significance of the drug group difference. Sampling traces and distributions and Gelman–Rubin diagnostics were obtained by using the coda package for R, version 0.15-2 (Plummer *et al.*, 2006). No evidence against convergence was identified.

4. Case study analysis results

We first discuss results by the EB method on the two different transformation scales. Figure 1 presents forest plots of the response rate data of the cytotoxic agent group on the logit transformed scale (Figure 1a) and the Freeman–Tukey's (F–T) arcsine square-root transformed scale (Figure 1b). With the logit transformation, the reported values are the log of the odd of having objective response. Confidence intervals were calculated using normal approximation with the within trial variance estimates defined respectively by (4) and (5). The plots are not directly comparable as the scales are different. Grey vertical reference lines are drawn for comparison at points that correspond to the efficacy response rate of 0.015 on the respective transformed scales. We observe that several trials are found on the left of the reference line on the F–T transformed scale (Figure 1b), whereas only



(a) Response rates on the logit transformed scale

(b) Response rates on the F−T arcsine square−root transformed scale

Figure 1. Response rates of the individual trials that tested cytotoxic agents: (a) on the logit transformed scale and (b) on the Freeman–Tukey's (F–T) arcsine square-root transformed scale

two trials are found on the left of the reference line on the logit transformed scale with the 0.5 constant correction applied to the logit transformation (Figure 1a). The EB method also weighted studies with low odds less, which resulted in overestimating the overall means, as compared with the GLMM and the FB results that are on the same logit transformed scale (see Figure 2). Although both groups were affected, the small constant correction affected the molecularly targeted drug group more as it has the higher percentage of zero observed observations (66% vs. 48%): compared to the GLMM and FB results, the downward bias in the EB estimated between trial variance is much greater in the molecularly targeted drug group than in the cytotoxic group.

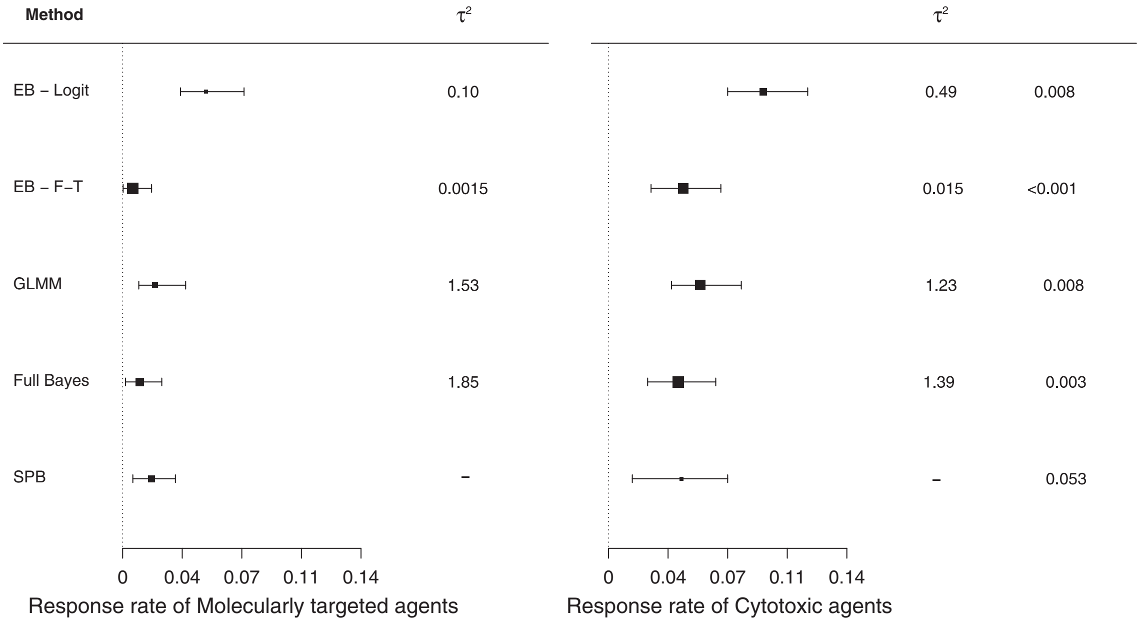


Figure 2. Estimates of the population mean response rates of each drug group and their comparisons: EB-Logit = Empirical Bayes (DerSimonian and Laird's method) analysis with the logit transformation, EB-F-T = Empirical Bayes (DerSimonian and Laird's method) analysis the Freeman–Tukey's arcsine square-root transformation, and SPB = semi-parametric Bayes analysis with a Dirichlet process prior. τ^2 indicates the between trial variance estimates

The EB method applied with the F–T transformation has its own shortcomings. It uses large sample approximation to define the within trial variances shown in (5), which did not work well for the given data. The approximation is known not to perform well for small or large p_i if n_i is not large (Mosteller and Youtz, 1961). Figure 2 in Mosteller and Youtz (1961) specifically shows that with $n = 10$, the large sample approximation overestimates the variance (or the within trial variances in this meta-analysis case) by ≥125% if the response rate is <0.07. A number of molecularly targeted and cytotoxic trials reported response rates <0.07 with 66% and 48% respectively reporting zero counts. Also about 10% of the individual trials have sample size <10. The poorly estimated within trial variances affect the overall mean estimation as the overall means are weighted averages with the weights depending on the within trial variances. The inappropriateness of the large sample approximation is also indicated in the forest plot (see Figure 1b): the dotted vertical line indicates the lower limit of a valid range corresponding to the support of efficacy rate. The lower limits of many confidence intervals stretch beyond the valid range.

The GLMM and FB methods are not subject to the aforementioned limitations, and reported comparable results.

5. Advanced Bayesian analysis

5.1. *Semi-parametric Bayesian analysis*

For a meta-analysis to be valid, individual trial results should be sufficiently similar to be compared and combined for a common pooled estimate. In standard random-effects analysis, this "similarity" assumption typically requires that a common normal distribution sufficiently describes random variation among trial level response rates. The earlier sections showed that the GLMM and the FB methods assumed the simple homoscedastic normal model (3). In the presence of a high percentage of zeros, however, the "similarity" assumption may not readily examined. We use a Bayesian semi-parametric GLMM to examine the assumption and sensitivity of the GLMM and FB results.

We employed the Dirichlet process GLMM proposed in Mukhopadhyay and Gelfand (1997) specifically. The model assumes the populations of trial level response rates of each drug group are mixture distributions and replaces the model (3) and the prior (6) with the following:

$$\eta(p_i)|G, x \sim G_x, \tag{7.a}$$

$$G_x|\alpha_x,\ G_0 \sim DP(\alpha_x G_0), \tag{7.b}$$

$$\alpha_x|\mathrm{a}_0,\ \mathrm{b}_0 \sim \mathrm{Gamma}(\mathrm{a}_0,\ \mathrm{b}_0), \tag{7.c}$$

$$G_0 = N(\cdot|\mu,\ \Sigma), \tag{7.d}$$

$$\Sigma|v_0,\ T \sim IW(v_0, T), \tag{7.e}$$

where G denotes the population distribution of individual trial response rates, and $DP(\alpha G_0)$ denotes the mixtures of Dirichlet processes with a precision parameter α and a normal base distribution G_0 (Antoniak, 1974). With the model (7.b) this semi-parametric model allows a general family of distributions denoted by $DP(\alpha G_0)$ and assumes the population distribution G is drawn from this family of distributions. In Bayesian terminology, this means placing a prior on the population distribution. This contrasts with the FB method which assumes a completely specified distribution up to a fixed number of unknown parameters as in model (3) and place priors on the parameters as in (6).

The mixture of Dirichlet processes (DP) specifically means that G follows an infinite mixture of normal distributions instead of one normal distribution. It does not *priori* fix the number of normal distributions and hence is distribution free, allowing more general distributional shapes for G such as heavy tailed or multi-modal distributions. In the literature, simulation studies have reported that Bayesian approaches with the DP prior well approximate an unknown distribution, whether the unknown distribution is a simple, homoscedastic normal, a mixture of normal distributions, or a skewed or heavy-tailed distribution (Gelfand and Mukhopadhyay, 1995, Pati and Dunson, 2014, Xu *et al*., 2015).

We used a gamma prior Gamma($\mathrm{a}_0 = 1$, $\mathrm{b}_0 = 1$) on α for the given case study. The gamma prior allowed the data to inform more strongly about the number of normal distributions needed to model G as a mixture distribution and about how tight neighborhood of G_0 the true population distribution G is in. For the base distribution G_0, a vague normal prior, $N(0, 10^6)$, is placed on μ, and a vague inverse Wishart prior distribution, $IW(3, 1)$, was placed on Σ. Similar to the FB analysis, we ran independent chains and confirmed the convergence of MCMC chains. Inference was performed also similarly as with the FB, using the posterior means, the 95% central credible intervals, and the posterior probability computed by the MCMC samples (see Figure 3a). This kind of model has been used successfully to model random effects in many situations (Dey *et al*., 1998, Kleinman and Ibrahim, 1998).

We checked the normality assumption of the GLMM and the FB methods, using the posterior predictive distributions of the individual trial effects under the semi-parametric model. The predictive distributions were

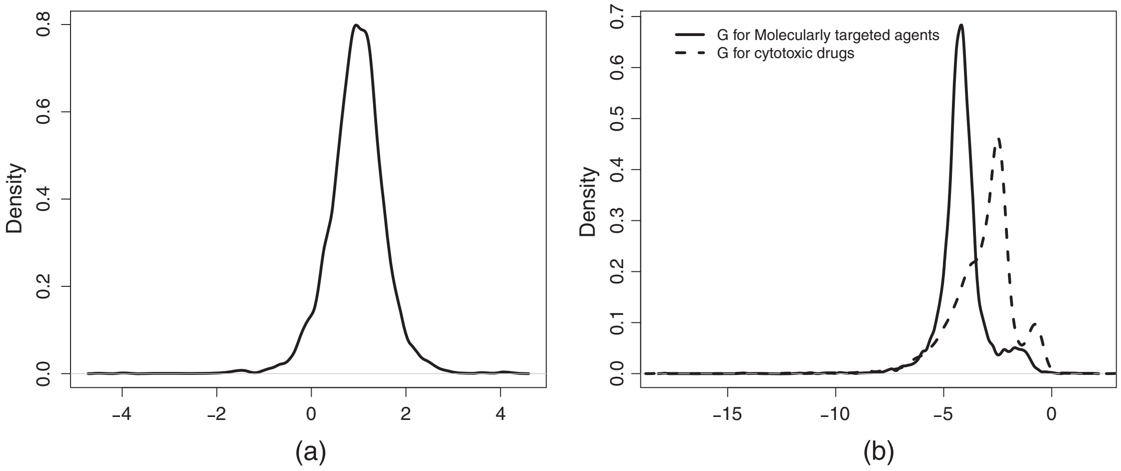


Figure 3. (a) Posterior distribution of population mean response rate difference on the logit transformed scale; (b) posterior predictive distributions of trial level response rates on the logit transformed scale

estimated based on 5000 MCMC samples from 200 000 runs with 100 000 burn-ins and by retaining every 20th iteration. In 63.6% of the 5000 MCMC samples, the DP model selected a mixture of two to five normal distributions to describe the population of trial level response rates for the molecularly targeted drug group. The three most frequent choices were mixing two to four normal distributions. The model selected a mixture of five to ten for the cytotoxic agent group in 64.1% with six to eight as the three most frequent choices. The predictive distributions suggest significant deviations from the assumed normality, particularly in the right tails (see Figure 3b). The small bumps in the right tails suggest the existence of sub populations that may be qualitatively different. We further investigate this possibility below.

Despite the deviation from the normality assumption, other inference results of the semi-parametric analysis agree well with those of the GLMM and the FB analysis in Figure 2. This implies that the population means and drug group comparison results of the GLMM and the FB analysis are robust to the deviations from their respective assumptions. No single parameter corresponds to the between trial variance in this semi-parametric Bayes analysis and no explicit estimate is available.

The semi-parametric analysis is also useful in checking the presence of outlying trial effects. In a meta-analysis, individual trials with extreme or outlying observations are not uncommon (Ohlssen *et al*., 2007). As the main objective of a meta-analysis is to provide a reasonable summary, the presence of such outliers may question whether the outlying trials are inherently different from and are inappropriate to be combined with the rest. Forest plots are an essential tool for graphically summarizing individual trial data and visually inspecting for potential outliers. With zero observed counts, however, standard summary measures may become undefined, in which case forest plots are not applicable. This is a lesser known problem but is not trivial.

The trial level estimates from the semi-parametric analysis can be used instead. The trial level estimates are pooled estimates (toward the population means) but are obtained free of distributional assumptions. Because of the DP mixture prior, they are only minimally affected from the population mean estimation in contrast to the GLMM or FB method. In the GLMM or FB analysis, the individual trial estimates are obtained under the normality assumption, under which outlying individual trial effects, if exist, are much more strongly pooled toward the population means. The forest plot in Figure 4 presents the individual trial estimates from the semi-parametric analysis in chronological order within each drug group, molecularly targeted drug studies first and cytotoxic studies later. We suspect two individual trial estimates in the molecularly targeted drug group and five in the cytotoxic group may be qualified as extreme observations. These potential outlying trials correspond to the small bumps in the right tails of the semi-parametric Bayesian estimated individual trial effect population distributions (Figure 3b). Whether the bumps suggest sub populations that may be qualitatively different can be investigated by whether these studies are qualitatively different from the rest.

We investigated this question by examining the tested agents and the characteristics of the patient samples of the ostensibly outlying two molecularly targeted drug trials and five cytotoxic drug trials. We found that each of the tested agents was considered "winners" and continually tested in later studies as a single agent or in combination. Such later studies also reported relatively high response rates, and in comparison to these later studies, the seemingly outlying trial effects did not look extreme. Also, three of the seven trials were designed for a single tumor type based on knowledge of the biology of the relevant tumor while the other trials allowed for multiple relapsed tumor types. This single versus multiple tumor type mix was similarly observed in the rest

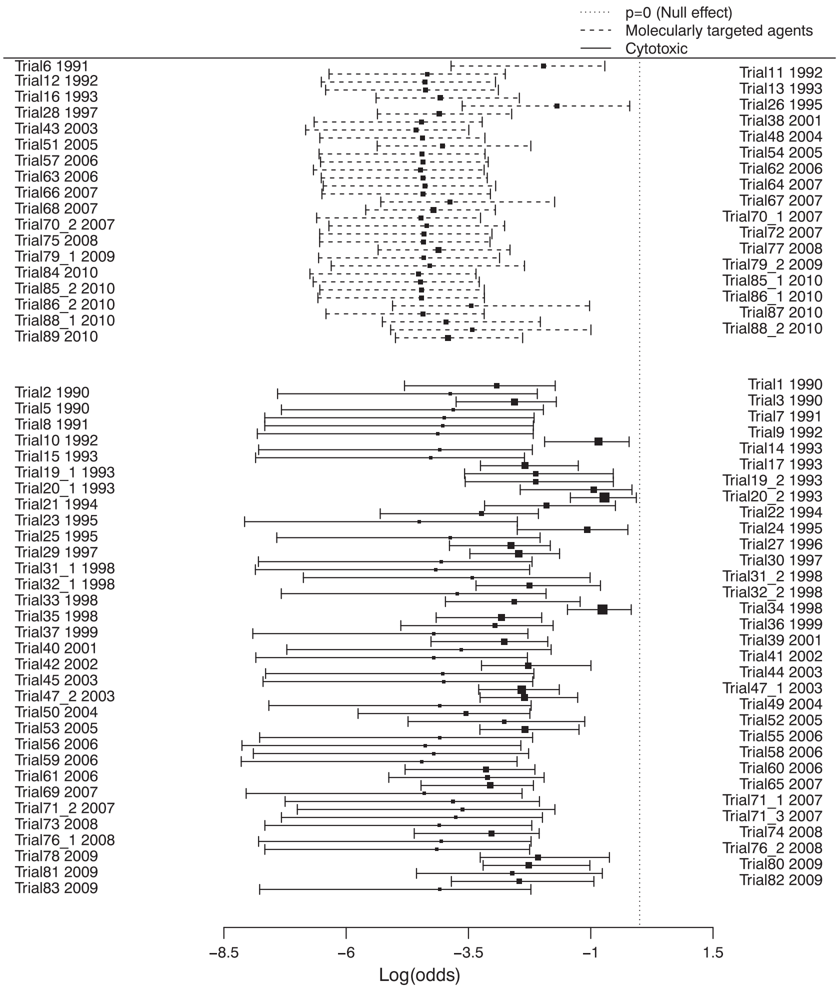
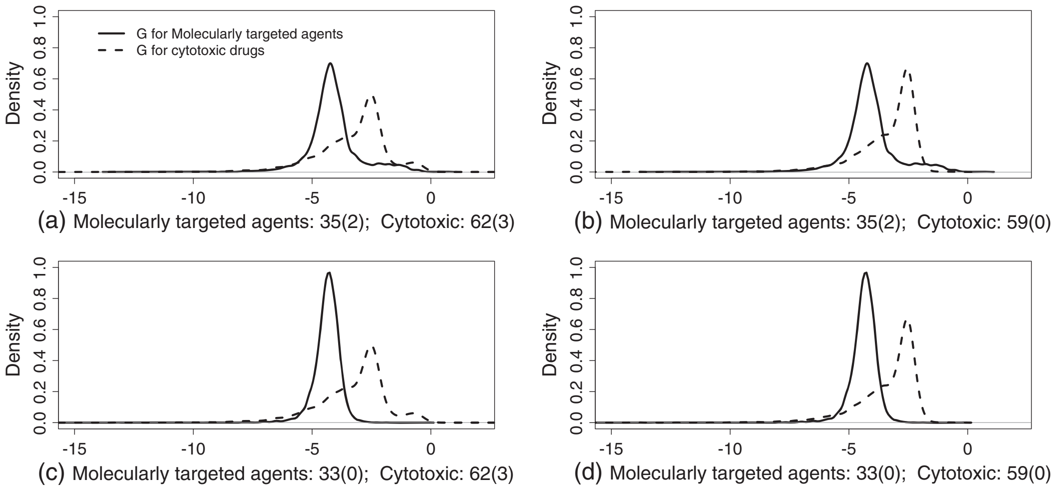


Figure 4. Forest plot using the semi-parametric Bayes estimates of individual trial response rates on the logit transformed scale. The dotted vertical line is a reference line drawn at 0 on the logit transformed scale

of the studies. The seven trials were also similar in the patient characteristics such as the median patient age and prior therapy exposure. This led us to conclude the ostensibly outlying seven trials are merely a result of random sampling.

On the contrary, forest plots using sample estimates of individual trials, not pooled estimates, are not appropriate for this kind of investigation. The forest plot on the logit transformed scale (Figure 1a) is subject to the upward bias because of the constant correction applied to zero counts. In the presence of the upward bias, the five cytotoxic trials indicated by the semi-parametric analysis as ostensibly outlying do not appear clearly distinctive from the rest. The distinction was relatively clear in the forest plot on the double arcsine transformation scale (Figure 1b).

We note that the semi-parametric analysis is robust to perturbation in the prior specification (7.c)–(7.e). Nieto-Barajas and Prunster (2009) performed a sensitivity analysis for a wide class of Bayesian nonparametric density estimators, including the mixture of DP, by perturbing the prior. Comparison of the resulting posterior density estimates found that the density estimation is robust. We also note that slight changes in the posterior distribution do not have much impact on the results. We performed simulation studies by arbitrarily removing the ostensibly outlying trials corresponding to the small bumps in the posterior predictive distributions. We removed 2 ostensibly outlying trials from each drug group or all. When not all outlying trials were removed, the resulting posterior predictive distributions showed small bumps corresponding to the remaining outlying trials, whereas they did not when all were removed (see Figure 5). The change in the number of outlying trials, however, minimally affected the posterior probability based group comparison. The posterior probability was 0.053 with the original data and ranged from 0.03 to 0.09 when some or all outlying trials were removed. We note that the semi-parametric analysis results, although robust, shall be interpreted with caution. Qualitative investigation should follow for subpopulations indicated by small bumps in the resulting posterior density estimates.



(a) Molecularly targeted agents: 35(2); Cytotoxic: 62(3)

(b) Molecularly targeted agents: 35(2); Cytotoxic: 59(0)

(c) Molecularly targeted agents: 33(0); Cytotoxic: 62(3)

(d) Molecularly targeted agents: 33(0); Cytotoxic: 59(0)

Figure 5. Examples of posterior predictive distributions of trial level response rates when some or all outlying trials were removed. Numbers inside parentheses denote the numbers of outlying individual trials remained after arbitrarily removing some or all

5.2. *Full Bayesian analysis of incomplete toxicity data*

The primary toxicity outcome of the example case study was dose-limiting toxicities (DLT). DLT were defined as grade 3 or 4 toxicity that occurred during the first cycle of drug. This outcome was missing in 23 trials. However, only 2 of 89 trials included toxicity information completely missing. Secondary toxicity outcomes and other information were available in the rest. We used a full Bayesian method and modeled the auxiliary information, salvaging the 21 trials with missing DLT.

The auxiliary information included number of overall grade 3 or 4 toxicities, numbers of patients evaluable for efficacy and toxicity assessment (n_i), and total number of courses of therapy (m_i). The overall grade 3/4 toxicity events are sums of hematologic and non-hematologic toxicities that occurred during all courses of therapy and included DLT events. The count generally increases along with the total number of treatment courses given per trial while the total number of treatment courses given per trial increases with the number of evaluable patients. The count also increases with the efficacy response rate observed in the trial, as patients were allowed to continue on receiving additional treatment courses if they responded.

We modeled each toxicity event count using a Poisson distribution but constrained them to be added up to certain totals respectively. For example, hematologic DLT event counts $\left(Z_{i,}^{DLT-Hem}\right)$ and hematologic non-DLT event counts need to be added up to the overall grade 3 or 4 hematologic toxicity event counts $\left(Z_i^{G3/G4-Hem}\right)$. This relationship suggested the following model:

$$Z_i^{G3/G4-Hem}\Big|\lambda_{i1},\lambda_{i3},n_i,m_i\sim Poisson\ (n_i\lambda_{i1}+m_i\lambda_{i3}),$$

$$Z_i^{DLT-Hem}\Big|Z_i^{G3/G4-Hem},\lambda_{i1},\lambda_{i3},n_i,m_i\sim Binomial\left(Z_i^{G3/G4-Hem},\ {}^{n_i}\lambda_{i1}/{}_{n_i}\lambda_{i1}+m_i\lambda_{i3}\right),$$

where $\lambda_{i1}, \lambda_{i3}$ denote respectively the event rates of the hematologic DLT event count and the non-DLT, grade 3 or 4 hematologic toxicity event count. In the model the DLT event counts increase with the number of patients (n_i), whereas the overall grade 3/4 toxicity count increases with the total number of treatment courses given per trial (m_i). Similar relationships hold for non-hematologic event variables. For missing m_i, we imputed it based on the following assumptions: we first consider m_{ij}, the number of treatment courses given to the subject j in the trial i, and assume it depends on the patient's response to the treatment as follows: $m_{ij}\sim Poisson\left(\lambda_{ij}^*\right)$ independently for all trials $i = 1, \ldots, N$ and $j = 1, \ldots, n_i.$,where $\lambda_{ij}^* = \lambda_1^*$ if the patient responded to the treatment of the trial or $\lambda_{ij}^* = \lambda_2^*$ otherwise. With T_i and n_i denote the total number of responders and patients for trial i, we have

$$m_i\big|T_i, n_i\sim Poisson\ \left[T_i\lambda_1^* + (n_i - T_i)\lambda_2^*\right]\text{for}i = 1, \ldots, N.$$

A joint prior distribution was specified for all the parameters.

This full Bayesian model assumed missing at randomness (MAR) within each drug group. We considered other possibilities of letting the probability of missingness dependent on the efficacy rate, the sample size and the total number of treatment courses but did not find any notable association.

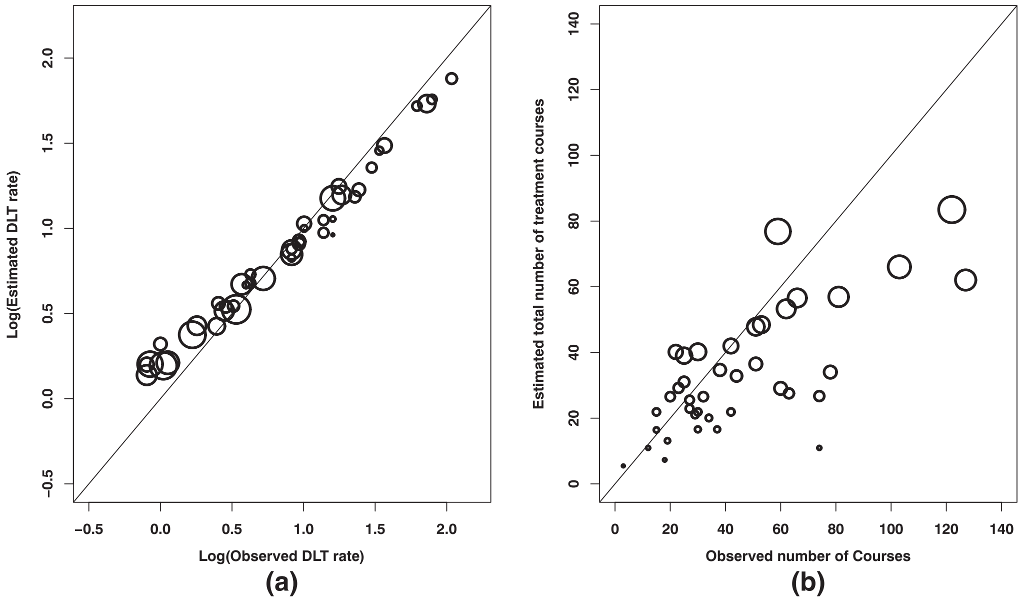


Figure 6. (a) Scatter plot of observed versus estimated trial level dose limiting toxicity (DLT) rates of the cytotoxic drug group; (b) scatter plot of observed versus estimated total number of treatment courses given per trial. The size of the bubbles indicates the sample size of individual trials

We examined the goodness of the fit of the full Bayesian model first. Figure 6a presents a plot of estimated versus observed DLT event rates of non-missing cases for the cytotoxic group. The agreement between the estimated versus observed trial level DLT rates follows a tilted 45 degree line with the Bayesian estimates, larger in the lower range while smaller in the upper range. This is expected as the Bayesian estimates are pooled toward the population mean estimate. The tight point cloud along the tilted 45 degree line suggests a good agreement. Figure 6b represents a plot of estimated versus observed total numbers of treatment courses given per trial. It suggests a reasonable agreement considering 60% of the data is missing.

We conducted complete cases (CC) alone analyses for comparison. Inference results of different methods are presented in Figure 7. Significantly higher toxicity rates were unanimously reported for the cytotoxic agent group. The empirical Bayes (EB) results refer to the log transformation results that used a constant correction by 1/2 for

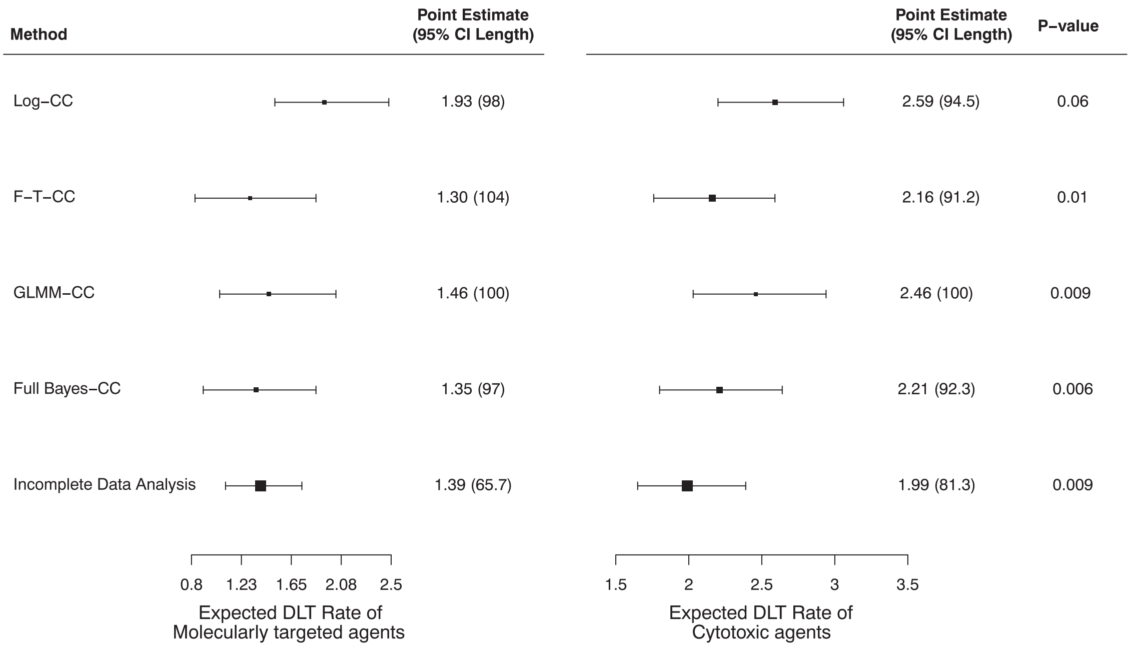


Figure 7. Population mean dose limiting toxicity (DLT) rate estimates and comparison of drug groups. Incomplete data analysis refers to the Bayes analysis that utilized auxiliary information

zero counts. We similarly note an upward bias in the estimates because of the constant correction. The estimates of the Bayesian incomplete data analysis are similar to the GLMM and FB estimates. This is expected as the standard methods provide valid results under MAR. The incomplete data analysis differs in that instead of excluding trials with the missing outcomes, it included them and utilized the data more efficiently. Although other possibilities were considered, the incomplete data analysis may have introduced bias, however, and hence the efficiency gained comes with the potential of bias.

The length of confidence intervals or credible intervals in case of Bayesian analyses can be used to compare the efficiency of each method. As longer intervals mean lower efficiency, the ratio of two interval lengths indicates the relative efficiency of the associated analysis methods. We used 95% confidence intervals or 95% central credible intervals. Numbers in the parentheses in Figure 6 indicate relative efficiency of each method in percent as compared with the GLMM complete case analysis. The incomplete analysis utilized the auxiliary information and is the most efficient. The 95% credibility intervals were 34.3% and 18.7%, shorter than the respective confidence intervals of the GLMM complete case alone analysis.

6. Conclusion

The importance of decision making based on the totality of relevant and sound evidence is increasingly emphasized in medicine. More systematic reviews of phase I trials will likely be conducted to establish appropriate clinical and statistical parameters to guide future clinical trials. We focus on the handling of sparse and incomplete data. We used a systematic review of a pediatric phase I oncology trial and showed that standard random effects methods may not be sufficient or adequate. A high percentage of zeros may question the exchangeability of individual trial level effects under a Gaussian distribution which is typically assumed in standard random effects analyses. A semi-parametric analysis with a DP prior estimates the population of trial level effects free of parametric distributional assumptions on the population and enables examining the assumption. A semi-parametric analysis with a DP prior, however, may not be feasible or desirable unless the number of individual studies included is large. We also showed that incomplete toxicity data can be addressed by an advanced Bayesian model that utilizes auxiliary information. The missing outcome case illustrated by the case study is rather ideal in that a clear and plausible imputation mechanism exists from observed data. Such knowledge is required for the advanced incomplete analysis.

Acknowledgements

Dr. Dorris' effort on this project was supported in part by Ruth L. Kirschstein National Research Service Award (NIEHS ES010957). Dr. Kim's effort was partially supported by an Agency for Healthcare Research and Quality's (AHRQ) Career Enhancement Award (K18HS022437) and a National Science Foundation (NSF) award (DMS-1007535), and Dr. Wang by the Association for Women in Mathematics-National Science Foundation Travel Grant.

Reference

Antoniak CE 1974. Mixtures of dirichlet processes with applications to Bayesian nonparametric problems. *Annals of Statistics* **2**: 1152–1174.

Appelman-Dijkstra NM, Kokshoorn NE, Dekkers OM, Neelis KJ, Biermasz NR, Romijn JA, Smit JWA, Pereira AM 2011. Pituitary dysfunction in adult patients after cranial radiotherapy: systematic review and meta-analysis. *Journal of Clinical Endocrinology & Metabolism* **96**: 2330–2340.

Branscum AJ, Hanson TE 2008. Bayesian nonparametric meta-analysis using Polya tree mixture models. *Biometrics* **64**: 825–833.

Breslow NE, Clayton DG 1993. Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association* **88**: 9–25.

Burr D, Doss H 2005. A Bayesian semiparametric model for random-effects meta-analysis. *Journal of the American Statistical Association* **100**: 242–251.

Cai TX, Parast L, Ryan L 2010. Meta-analysis for rare events. *Statistics in Medicine* **29**: 2078–2089.

Dersimonian R, Laird N 1986. Meta-analysis in clinical trials. *Controlled Clinical Trials* **7**: 177–88.

Dey D, Müller P, Sinha D 1998. Practical Nonparametric and Semiparametric Bayesian Statistics. New York: Springer.

Dorris, K, Li, D, Liu, C, Wang, X, Hummel, T, Perentesis, J, Ingle, A, Kim, M & Fouladi, M 2015. A comparison of safety and efficacy of cytotoxic versus molecularly-targeted drugs in pediatric phase I solid tumor oncology trials. *Manuscript to be submitted to Journal of Clinical Oncology*.

Freeman MF, Tukey JW 1950. Transformations related to the angular and the square root. *Annals of Mathematical Statistics* **21**: 607–611.

Friedrich JO, Adhikari NKJ, Beyene J 2007. Inclusion of zero total event trials in meta-analyses maintains analytic consistency and incorporates all available data. *BMC Medical Research Methodology* **7**: 1–6.

Gelfand AE, Mukhopadhyay S 1995. On nonparametric Bayesian inference for the distribution of a random sample. *Canadian Journal of Statistics-Revue Canadienne De Statistique* **23**: 411–420.

Gelman A, Carlin JB, Stern HS, Dunson DB, Vehtari A, Rubin DB 2013. Bayesian Data Analysis. CRC Press: Boca Raton.

Hamer PCD, Robles SG, Zwinderman AH, Duffau H, Berger MS 2012. Impact of intraoperative stimulation brain mapping on glioma surgery outcome: a meta-analysis. *Journal of Clinical Oncology* **30**: 2559–2565.

Kleinman KP, Ibrahim JG 1998. A semiparametric Bayesian approach to the random effects model. *Biometrics* **54**: 921–938.

Kooiman J, Pasha SM, Zondag W, Sijpkens YWJ, Van Der Molen AJ, Huisman MV, Dekkers OM 2012. Meta-analysis: serum creatinine changes following contrast enhanced CT imaging. *European Journal of Radiology* **81**: 2554–2561.

Mosteller F, Youtz C 1961. Tables of the Freeman–Tukey transformations for the binomial and Poisson distributions. *Biometrika* **48**: 433–440.

Mukhopadhyay S, Gelfand AE 1997. Dirichlet process mixed generalized linear models. *Journal of the American Statistical Association* **92**: 633–639.

Nieto-Barajas LE, Prunster I 2009. A sensitivity analysis for Bayesian nonparametric density estimators. *Statistica Sinica* **19**: 685–705.

Ohlssen DI, Sharples LD, Spiegelhalter DJ 2007. Flexible random-effects models using Bayesian semi-parametric models: applications to institutional comparisons. *Statistics in Medicine* **26**: 2088–2112.

Pati D, Dunson DB 2014. Bayesian nonparametric regression with varying residual density. *Annals of the Institute of Statistical Mathematics* **66**: 1–31.

Platt RW, Leroux BG, Breslow N 1999. Generalized linear mixed models for meta-analysis. *Statistics in Medicine* **18**: 643–54.

Plummer M, Best N, Cowles K, Vines K 2006. CODA: convergence diagnosis and output analysis for MCMC. *R News* **6**: 7–11.

Singh JA, Christensen R, Wells GA, Suarez-Almazor ME, Buchbinder R, Lopez-Olivo MA, Ghogomu ET, Tugwell P 2009. A network meta-analysis of randomized controlled trials of biologics for rheumatoid arthritis: a Cochrane overview. *Canadian Medical Association Journal* **181**: 787–796.

Smith TC, Spiegelhalter DJ, Thomas A 1995. Bayesian approaches to random-effects meta-analysis: a comparative study. *Statistics in Medicine* **14**: 2685–2699.

Sturtz S, Ligges U, Gelman A 2005. R2WinBUGS: a package for running WinBUGS from R. *Journal of Statistical Software* **12**: 1–16.

Sutton AJ, Abrams KR 2001. Bayesian methods in meta-analysis and evidence synthesis. *Statistical Methods in Medical Research* **10**: 277–303.

Sutton AJ, Higgins JPI 2008. Recent developments in meta-analysis. *Statistics in Medicine* **27**: 625–650.

Sweeting MJ, Sutton AJ, Lambert PC 2004. What to add to nothing? Use and avoidance of continuity corrections in meta-analysis of sparse data. *Statistics in Medicine* **23**: 1351–75.

Viechtbauer W 2010. Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software* **36**: 1–48.

Wolfinger R, Oconnell M 1993. Generalized linear mixed models—a pseudo-likelihood approach. *Journal of Statistical Computation and Simulation* **48**: 233–243.

Xu ZG, Maceachern S, Xu XY 2015. Modeling non-Gaussian time series with nonparametric Bayesian model. *Ieee Transactions on Pattern Analysis and Machine Intelligence* **37**: 372–382.

Supporting information

Additional Supporting Information may be found with the online version of this article at the publisher's web site.