Journal of Biopharmaceutical Statistics

ISSN: 1054-3406 (Print) 1520-5711 (Online) Journal homepage: http://www.tandfonline.com/loi/lbps20

Bayesian Hierarchical Modeling for Detecting Safety Signals in Clinical Trials

H. Amy Xia , Haijun Ma & Bradley P. Carlin

To cite this article: H. Amy Xia , Haijun Ma & Bradley P. Carlin (2011) Bayesian Hierarchical Modeling for Detecting Safety Signals in Clinical Trials, Journal of Biopharmaceutical Statistics, 21:5, 1006-1029, DOI: 10.1080/10543406.2010.520181

To link to this article: https://doi.org/10.1080/10543406.2010.520181

Journal of Biopharmaceutical Statistics, 21: 1006–1029, 2011

ISSN: 1054-3406 print/1520-5711 online
DOI: 10.1080/10543406.2010.520181

Taylor & Francis
Taylor & Francis Group

BAYESIAN HIERARCHICAL MODELING FOR DETECTING SAFETY SIGNALS IN CLINICAL TRIALS

H. Amy Xia[1], Haijun Ma[1], and Bradley P. Carlin[2]
[1]*Amgen, Inc., Thousand Oaks, California, USA*
[2]*University of Minnesota, Twin Cities, Minnesota, USA*

Detection of safety signals from clinical trial adverse event data is critical in drug development, but carries a challenging statistical multiplicity problem. Bayesian hierarchical mixture modeling is appealing for its ability to borrow strength across subgroups in the data, as well as moderate extreme findings most likely due merely to chance. We implement such a model for subject incidence (Berry and Berry, 2004) using a binomial likelihood, and extend it to subject-year adjusted incidence rate estimation under a Poisson likelihood. We use simulation to choose a signal detection threshold, and illustrate some effective graphics for displaying the flagged signals.

Key Words: Bayesian hierarchical models; Clinical trials; Drug safety; Multiplicity; Signal detection.

1. INTRODUCTION

Safety assessment is critically important in drug development. It has been attracting increasing attention among industry sponsors and regulatory agencies in identifying various ways of enhancing safety planning and evaluation. The Safety Planning, Evaluation and Reporting Team (SPERT) was formed in 2006 by PhRMA (the Pharmaceutical Research and Manufacturers of America) to recommend a pharmaceutical industry standard for safety planning, data collection, evaluation and reporting (Crowe et al., 2002; Gould, 2002). One of the recommendations from the SPERT is to use a three-tiered system in analyzing the clinical trial adverse event (AE) data.

A key feature of the approach is to classify AEs into three tiers: Tier 1 AEs are the events for which a hypothesis has been prespecified and the most comprehensive statistical analyses are performed; Tier 2 and Tier 3 AEs are those that are not prespecified and the goal for analyzing those AEs is signal detection. The distinction between Tier 2 and Tier 3 is that the former are "common" events for which statistical inference can play a more important role in helping screen potential safety concerns, whereas the latter are "rare" events for which medical judgment usually prevails. However, the Bayesian approach discussed in this paper models the entire AE data set, so the distinction between Tier 2 and Tier 3 AEs may not be necessary. In fact, the Bayesian approach is advantageous in dealing with

Received 5 January 2010; Accepted 27 August 2010
Address correspondence to H. Amy Xia, Director Biostatistics, Amgen, Inc., One Amgen Center Drive, Thousand Oaks, CA 91320, USA; E-mail:hxia@amgen.com

rare events. Of course, signal detection is a process: Once a signal is identified, it needs to be further investigated, hypothesized, characterized, verified, and quantified through a multidisciplinary effort.

Hence, detection of safety signals from routinely collected, not prespecified Tier 2, and even Tier 3 adverse event data in clinical trials, is critical in drug development, as an initial step to understand the safety profile of the product and therefore not only protect the patients enrolled in the trial but also prevent unsafe drugs from being approved. One of the statistical challenges in this setting is the multiplicity problem. In clinical trials, because the number of types of adverse events is very large (typically in hundreds or even thousands in late-phase clinical trials), the current frequentist-based approaches of flagging AEs based on unadjusted p-values or confidence intervals (CIs) can result in an excessive number of false positive signals. On the other hand, if we adjust for multiplicity in traditional ways (e.g., the Bonferroni method), this may lead to an excessive rate of false negatives, whence important safety signals may be missed. Therefore, we need to have a flagging device or a screening tool that can strike a proper balance between "no adjustment" versus "too much adjustment." There was a consensus in the SPERT discussion that multiplicity should be addressed in analyzing AEs that were not prespecified.

Several Bayesian methods have been proposed in drug safety assessment since the late 1990s. In analyzing postmarketing spontaneous reports, a gamma–Poisson shrinkage algorithm in analyzing the Food and Drug Administration (FDA) Adverse Event Reporting System (AERS) database was proposed (DuMouchel, 1999), while a Bayesian Confidence Propagation Neural Network approach to analyze a Word Health Organization (WHO) database of adverse drug reactions was used in Bate et al. (1998). In the clinical trial arena, a novel Bayesian hierarchical modeling approach in analyzing binary outcomes in AE data was proposed (Berry and Berry, 2004), work we extend in this paper. In light of all aforementioned Bayesian work in drug safety assessment, Chi et al. (2002) commented from a regulatory perspective, "Safety assessment is one area where frequentist strategies have been less applicable. Perhaps Bayesian approaches in this area have more promise." From the frequentist side, false discovery rate (FDR) control methods (Benjamini and Hochberg, 1995; Benjamini and Yekutieli, 2001) have been used for multiple comparisons. A double false discovery rate (DFDR) approach was developed (Mehrotra and Heyse, 2004) to address the multiplicity issue for the clinical trial AE data.

So why are Bayesian models helpful? As stated in Berry and Berry (2004), there are at least four considerations in determining whether to flag an AE as a signal: (1) the actual significance levels; (2) the total number of types of AEs; (3) the rates of AEs not being flagged including their similarity with those being flagged; and (4) the biological relationships among various AEs. The first two are standard considerations in the frequentist approach to multiplicity adjustment. The last two are not, but they can be considered in Bayesian modeling. For example, AEs are now routinely coded in Medical Dictionary for Regulatory Activities (MedDRA) terms with a hierarchical structure. An AE can be coded in a lower level term (LLT), preferred term (PT), high-level term (HLT), high-level group term (HLGT), and system organ class (SOC). Biological relationships among various AEs are reflected in this intrinsic medical coding structure. For example, a reported AE "oedema of

veins" can be coded to the LLT "edema vascular," PT "angioneurotic oedema," HLT "angioedemas," and HLGT "angioedema and urticaria" under the SOC "skin and subcutaneous tissue disorders." Bayesian hierarchical models allow for explicitly modeling AEs with the existing coding structure, such as SOC and PT in MedDRA, which are routinely reported in drug clinical trials. AEs in the same SOC are more likely to be similar, so they can sensibly borrow strength from each other. The model also allows borrowing across SOCs, but does not impose it, depending on the actual data. In fact, clinical and safety people would (informally) consider the similarity of the AEs within SOCs when they review AE tables. For example, if differences in several cardiac vascular events were observed, then each would be more likely to be causal than if differences came from medically unrelated areas (say, if they arose from skin, neurological, thrombosis, and cancer). Bayesian hierarchical modeling allows a scientific, explicit, and more formal way to take this into consideration. Multiplicity is handled by Bayesian modeling of AE structure.

In this paper, we construct a three-level Bayesian hierarchical mixture model for binary responses, following the model of Berry and Berry (2004). We also apply a hierarchical Poisson mixture model under the framework of generalized linear models, which permits use of an offset that allows for differential periods of risk among patients while having better statistical properties for rare events. More importantly, we provide guidance on how to choose a signal detection threshold to achieve a fair balance between false positive error rates and false negative error rates via a simulation study. Wide application of the Berry and Berry approach has been hampered by lack of readily available software, so we implement our models using the most commonly used Bayesian software, WinBUGS 1.4 (**http://www.mrc-bsu.cam.ac.uk/bugs/welcome.shtml**). In addition, we create effective graphical tools for deciphering data, presenting results, and making inference when many AEs are analyzed.

The rest of the paper is organized as follows. Section 2 describes five Bayesian hierarchical binomial and Poisson mixture models for the binary and subject-year adjusted outcomes. Section 3 presents the results from our case study. We also address the issue of model selection and comparison using the Deviance Information Criterion (DIC) (Spiegelhalter et al., 2002), as well as how to draw inferences from the different models. Section 4 describes a simulation study to demonstrate why false positives and false negatives are better controlled by our Bayesian procedure. Finally, section 5 summarizes our approach, and suggests directions for future work in this area. We also attach the WinBUGS program for Bayesian hierarchical binomial and Poisson mixture models and the S-Plus program for the volcano plot in Appendix I and Appendix II, respectively.

2. BAYESIAN HIERARCHICAL MODELING

In analyzing AE data for drug trials, two measures are commonly used in dealing with binary and count data: subject incidence rate (or "crude rate"), and subject-year adjusted incidence rate, respectively. Subject incidence is defined as the number of subjects experiencing a certain event divided by the number of subjects initially exposed to the drug, regardless of duration of use. It is most appropriate where all subjects are treated and followed for the same period of time or for very short-term drug exposure, or for acute events following closely in time after

exposure. When the subjects in a trial have different durations of drug exposure or follow-up, or in the presence of censoring, subject incidence is not appropriate (O'Neill, 1998). To adjust for potential differences in duration of drug exposure, the subject-year adjusted incidence rate may be used. This is defined as the number of subjects experiencing a specific event divided by their total subject-time at risk, which is a summation of the time from first drug exposure to first event for subjects with at least one event or the total observation time for subjects without the event. The underlying assumption with this measure is that the risk of having an event is constant over time. This assumption may not hold for certain types of adverse events. Other time-to-event analysis methods that relax this assumption, such as Cox regression and Kaplan–Meier estimation, may be used, but they usually require individual subject-level information.

In this section, we describe five Bayesian models. The first three use logistic regression for subject incidence models, while the other two use Poisson regression for subject-year adjusted incidence models.

- Model 1a: three-stage model with normal prior on log-OR (logarithm of odds ratio).
- Model 1b: three-stage model with mixture prior on log-OR.
- Model 1c: nonhierarchical one-stage Bayesian mixture model.
- Model 2a: three-stage model with normal prior on log-RR (logarithm of relative risk).
- Model 2b: three-stage model with mixture prior on log-RR.

2.1. Model 1a: Bayesian Logistic Regression Model with Normal Prior on Log-OR

For $b = 1, \ldots, B$ and $j = 1, \ldots, k_b$, let Y_{bj} and X_{bj} be the number of subjects with an AE with PT j under SOC b in treatment and placebo groups. N_t and N_c are the number of subjects in treatment and control groups, respectively.

We assume a binomial likelihood for the AE counts, i.e., $Y_{bj} \sim Binom(N_t, t_{bj})$ and $X_{bj} \sim Binom(N_c, c_{bj})$, where t_{bj} and c_{bj} are the probability of AE for PT j and SOC b in treatment and control groups, respectively. We consider a logistic regression mean structure: $\log it(c_{bj}) = \log(c_{bj}/(1 - c_{bj})) = \gamma_{bj}$; $\log it(t_{bj}) = \gamma_{bj} + \theta_{bj}$. Note that $\theta_{bj} = \log \frac{t_{bj}(1-c_{bj})}{c_{bj}(1-t_{bj})}$ is the logarithm of odds ratio (log-OR).

We set the stage 1 prior distributions to be

$$\gamma_{bj} \sim N(\mu_{\gamma b}, \sigma^2_{\gamma b}) \quad \text{and} \quad \theta_{bj} \sim N(\mu_{\theta b}, \sigma^2_{\theta b})$$

The stage 2 prior distributions are as follows:

$$\mu_{\gamma b} \sim N(\mu_{\gamma 0}, \tau^2_{\gamma 0}) \quad \sigma^2_{\gamma b} \sim IG(\alpha_\gamma, \beta_\gamma)$$
$$\mu_{\theta b} \sim N(\mu_{\theta 0}, \tau^2_{\theta 0}) \quad \sigma^2_{\theta b} \sim IG(\alpha_\theta, \beta_\theta)$$

The fully Bayesian hierarchical model involves the following stage 3 prior distributions:

$$\mu_{\gamma 0} \sim N(\mu_{\gamma 00}, \tau^2_{\gamma 00}) \quad \tau^2_{\gamma 0} \sim IG(\alpha_{\gamma 00}, \beta_{\gamma 00})$$
$$\mu_{\theta 0} \sim N(\mu_{\theta 00}, \tau^2_{\theta 00}) \quad \tau^2_{\theta 0} \sim IG(\alpha_{\theta 00}, \beta_{\theta 00})$$

The hyperparameters $\mu_{\gamma 00}$, $\tau^2_{\gamma 00}$, $\mu_{\theta 00}$, $\tau^2_{\theta 00}$, $\alpha_{\gamma 00}$, $\beta_{\gamma 00}$, $\alpha_{\theta 00}$, $\beta_{\theta 00}$, α_γ, β_γ, α_θ, and β_θ are considered fixed constants. In our data analysis, we use the following values: $\mu_{\gamma 00} = \mu_{\theta 00} = 0$, $\tau^2_{\gamma 00} = \tau^2_{\theta 00} = 10$, $\alpha_{\gamma 00} = \alpha_{\theta 00} = \alpha_\gamma = \alpha_\theta = 3$, and $\beta_{\gamma 00} = \beta_{\theta 00} = \beta_\gamma = \beta_\theta = 1$.

2.2. Model 1b: Bayesian Logistic Regression Model with Mixture Prior on Log-OR

For Model 1b, we take the same likelihood and mean structure as Model 1a, but alter the prior distribution for the log-OR to a mixture distribution:

$$\theta_{bj} \sim \pi_b \delta(0) + (1 - \pi_b) N(\mu_{\theta b}, \sigma^2_{\theta b})$$

where $\delta(0)$ is a distribution having unit point mass at 0. This is the model proposed in Berry and Berry (2004). This mixture allows a point mass on equality of the treatment and the control rates because many AEs may be completely unaffected by treatment.

The same prior distributions are adopted for the common parameters as in Model 1a. For the new hyperparameters, the prior distributions are as follows:

$$\begin{aligned} \pi_b &\sim Beta(\alpha_\pi, \beta_\pi) \\ \alpha_\pi &\sim Exp(\lambda_\alpha) I[\alpha_\pi > 1] \\ \beta_\pi &\sim Exp(\lambda_\beta) I[\beta_\pi > 1] \end{aligned}$$

where the truncated exponential prior distributions for α_π and β_π are chosen to minimize the prior mass associated with 0 and 1, the boundary values for π_b. We take the same fixed values for the hyperparameters $\mu_{\gamma 00}$, $\tau^2_{\gamma 00}$, $\mu_{\theta 00}$, $\tau^2_{\theta 00}$, $\alpha_{\gamma 00}$, $\beta_{\gamma 00}$, $\alpha_{\theta 00}$, $\beta_{\theta 00}$, α_γ, β_γ, α_θ, and β_θ as for Model 1a. In addition, we treat λ_α and λ_β as fixed constants, where $\lambda_\alpha = \lambda_\beta = 0.1$.

2.3. Model 1c: Nonhierarchical One-Stage Bayesian Model with Mixture Prior

For Model 1c, following the same binomial likelihood and logistic regression mean structure as in Model 1a and Model 1b, we assume $\gamma_{bj} \sim N(0, 10^2)$ and $\theta_{bj} \sim 0.5\delta(0) + 0.5N(0, 10^2)$. That is to say, no information is borrowed across different AEs within the same SOC and all PTs are treated independently. With vague prior information, this model should deliver results similar to frequentist approaches unadjusted from multiplicity.

2.4. Model 2a: Bayesian Log-Linear Regression Model with Normal Prior on Log-RR

Let T and C be the total subject-time at risk in treatment and control groups, respectively, where $T = \sum_{i=1}^{N_t} \xi_i$ and $C = \sum_{i=1}^{Nc} \zeta_i$ with ξ_i being the time at risk for patient i in the treatment group and ζ_i in the control group, respectively. We now assume a Poisson likelihood for the AE counts, i.e., $Y_{bj} \sim Pois(t_{bj}T)$ and $X_{bj} \sim$

$Pois(c_{bj}C)$, where t_{bj} and c_{bj} are the hazard rates per subject-year in the treatment and control groups, respectively. We consider a log-linear regression mean structure $\log(c_{bj}) = \gamma_{bj}$, $\log(t_{bj}) = \gamma_{bj} + \theta_{bj}$. Note that $\theta_{bj} = \log(t_{bj}/c_{bj})$ is the logarithm of relative risk (log-RR).

Model 2a is the Poisson likelihood counterpart of Model 1a, whose prior distributions we again adopt.

2.5. Model 2b: Bayesian Log-Linear Regression Model with Mixture Prior on Log-RR

Model 2b is the Poisson likelihood counterpart of Model 1b; i.e., it has the same likelihood and mean structure as Model 2a, and the same prior distributions as Model 1b.

3. DATA ANALYSIS

In this section, we apply our five Bayesian hierarchical models to aggregated AE data from four double-blind, placebo-controlled, Phase II/III clinical trials of a compound that, for reasons of confidentiality, we refer to as Drug X. All four studies are of 12 or 24 weeks in duration, and are fairly similar in design and population. As such, without loss of generality we pool the AE data. The resulting aggregated data set has 1245 subjects in the treatment group and 720 subjects in the control group. The reported AEs are coded to 465 PTs under 24 SOCs. An excerpt of the data is given in Table 1 for illustration.

In comparing the different models, we use the Deviance Information Criterion (DIC) as a model selection method. Inferences from the selected model are discussed in section 3.2. Since statistical graphs are useful tools in deciphering data and presenting results, use-specific graphical tools are used to illustrate the analysis results. All Bayesian models in this paper were implemented using WinBUGS 1.4.2. For each model, posterior summaries were obtained from 20,000 draws from two initially overdispersed chains after a 10,000-iteration burn-in. The simulation study was done via BRugs, an interface for calling BUGS from R; see the website http://www.mirrorservice.org/sites/lib.stat.cmu.edu/R/CRAN/src/contrib/Descriptions/BRugs.html. All graphs were produced using S-Plus and all other computations were done using SAS.

3.1. Model Selection

DIC is a hierarchical modeling generalization of the AIC (Akaike Information Criterion). It is widely used in Bayesian model selection due to its simplicity and generality (and perhaps its ready implementation within BUGS). Deviance is defined as $D(\theta) = -2\log(p(y \mid \theta)) + h(y)$, where y are the data, θ are the unknown parameters, $p(y \mid \theta)$ is the likelihood function, and $h(y)$ is a normalizing function of the data alone. Thus, $\overline{D} = E^{\theta}[D(\theta)]$ is a measure of how well the model fits the data, and $p_D = \overline{D} - D(\hat{\theta})$ can be shown to be interpretable as an effective number of parameters. In a Bayesian hierarchical model, p_D is typically smaller than the number of parameters in the model due to Bayesian shrinkage of the random effects

Table 1 Sample data from the example

b	j	SOC	Preferred term	Ctrl n	Trt n	Ctrl r (%)	Trt r (%)	Ctrl n/E	Trt n/E
1	1	Blood and lymphatic system disorders	Lymphadenopathy	6	10	0.83	0.80	0.04	0.03
...	...	...	...			...			
8	8	General disorders and administration site conditions	Fatigue*	9	37	1.25	2.97	0.06	0.13
8	9		Feeling cold	1	0	0.14	0.00	0.01	0.00
8	10		Feeling hot	1	3	0.14	0.24	0.01	0.01
...	...	...	...			...			
11	34	Infections and infestations	Herpes simplex*	3	19	0.42	1.53	0.02	0.07
	...		...			...			
11	64		Sinusitis*	12	46	1.67	3.69	0.08	0.16
...	...	...	...			...			
12	12	Injury, poisoning, and procedural complications	Excoriation*	0	8	0.00	0.64	0.00	0.03
...	...	...	...			...			
22	15	Skin and subcutaneous tissue disorders	Ecchymosis*	0	12	0.00	0.96	0.00	0.04
...	...	...	...			...			

Note. r = subject incidence (%). n/E = subject-year adjusted incidence rate (per subject-year), where n is observed count and E is total subject-time at risk.

*Fisher's two-sided exact test unadjusted p-value ≤ 0.05 with higher risk on treatment arm.

toward a common value, as encouraged by the model structure. Here $\hat{\theta}$ is a plug-in estimate of θ, usually the posterior mean. Then DIC is defined as $DIC = p_D + \overline{D}$. Like AIC and BIC, models with smaller DIC are preferred.

Tables 2 and 3 give the DIC scores for our five Bayesian hierarchical binomial and Poisson models, respectively. For short-term studies, like our case, differences

Table 2 DIC scores for binomial likelihood Bayesian models

	$\overline{D}$	p_D	DIC
Model 1a	2604.86	382.09	2986.95
Model 1b	2644.86	324.03	2968.89
Model 1c	2574.86	602.38	3177.24

Table 3 DIC scores for Poisson likelihood Bayesian models

	$\overline{D}$	p_D	DIC
Model 2a	2605.15	382.00	2987.15
Model 2b	2644.15	332.49	2976.65

in exposure or follow-up duration among the treatment groups are often not big, so a Poisson model may not be needed to adjust for different follow-up times. On the other hand, for rare AEs in studies with a relatively large sample size, the binomial and Poisson models give similar results. As data used for the two sets of models are not the same, we cannot directly compare between binomial likelihood models and Poisson likelihood models using DIC. However, comparisons within models with the same likelihood indicate that models with mixture priors (i.e., Models 1b and 2b) are preferred. With the presence of the point mass at zero, the mixture prior reduces the number of effective parameters. In section 4.2, we discuss the inferences from the Binomial likelihood model having a mixture prior (i.e., Model 1b and Model 1c).

3.2. Inferences

Fisher's exact test is the common choice in clinical studies for flagging safety signals in AE data sets. However, due to the large number of AE terms, multiplicity is an issue that must be addressed. As we stated previously, unadjusted methods will inevitably identify false positive signals and lead to high type I error rates.

Figure 1 is a volcano plot, also called a "p-risk plot" (O'Connell, 2007), which displays the raw risk differences between treatment and placebo groups versus the unadjusted two-sided Fisher's exact test p-values, where the latter is on a negative base-10 logarithm scale. To look for potential safety signals, we need to focus on the upper right panel, where PTs with higher risk on treatment group and smaller p-values are located. PTs on the upper left corner can be viewed as AEs with protective drug effects. Different symbols combinations are used in Fig. 1 to represent different SOCs, so that clustering of PTs from same SOC is easier to notice. Figure 1 shows that five out of the 465 PTs have unadjusted p-values smaller than 0.05 with higher risk on the treatment group.

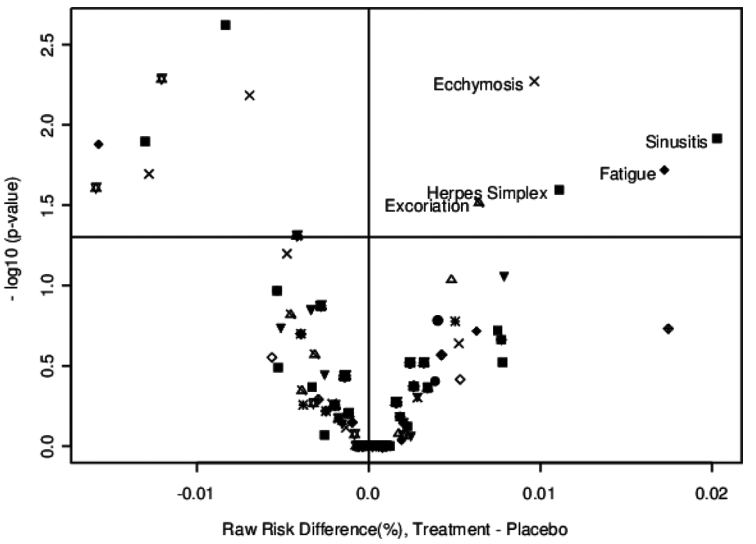

Figure 1 Volcano plot of risk differences versus unadjusted Fisher's exact test p-values.

In our Bayesian hierarchical models for safety signal detection, we use the posterior exceedance probability to identify potential signals. That is, an AE of PT j in SOC b is flagged if the exceedance probability $\Pr(\mathrm{OR}_{bj} > d^* | \text{ Data}) > p$, where d^* and p are prespecified constants, and $\mathrm{OR}_{bj} = \exp(\theta_{bj})$ where θ_{bj} is log-OR in binomial models and log-RR in Poisson models for PT j in SOC b. For safety signal detection, we usually choose $d^* = 1$, which indicates higher risk on the treatment arm. However, an effect could have a high probability of OR > 1 but be clinically unimportant. Therefore, different values for $d^* > 1$ (e.g., $d^* = 1.5$ or 2) can be applied to reflect a clinically meaningful effect. Under the Bayesian framework, it is very straightforward and flexible to compute $\Pr(\mathrm{OR}_{bj} > d^* | \text{ Data})$ with various values of d^*.

The exceedance probabilities for other statistics such as risk differences (RD) are also easily obtainable in a Bayesian framework. For example, an AE can be flagged if $\Pr(t_{bj} - c_{bj} > d^{**} \mid Data) > p$, where d^{**} and p are again prespecified constants.

Table 4 presents various posterior exceedance probabilities from Model 1b for the five PTs that have two-sided Fisher's exact test p-value less than 0.05 and higher observed risks on the treatment arm than the placebo. The fourth column gives the posterior probability of OR = 1, i.e., no difference between treatment and placebo arms. The following three columns give the posterior exceedance probabilities based on OR using different clinically meaningful cutoffs. The last two columns give the exceedance probabilities based on RD. The table shows that smaller p-values do not necessarily imply higher posterior exceedance probabilities. For example, ecchymosis has the smallest p-value, but does not have the largest probability of OR > 1. In our Bayesian hierarchical model, PTs in the same SOC are considered more alike than those across different SOCs. Thus, PTs within the same SOC will borrow more strength from each other. But the traditional way of carrying out Fisher's exact test for all PTs does not take such a hierarchy into consideration, treating all PTs as independent from each other, leading to a loss of information. Moreover, a few random outliers may lead to erroneous conclusions when we do not take other correlated observations into consideration.

Table 4 Inferences of binomial hierarchical model with mixture prior (Model 1b)

SOC	PT	Exact p-value	Post prob OR = 1	Post prob OR > 1	Post prob OR > 1.2	Post prob OR > 2	Post prob RD > 2%	Post prob RD > 5%
General disorders and administration site conditions	Fatigue	0.019	0.430	0.564	0.546	0.319	0.099	0.000
Infections and infestations	Herpes simplex	0.025	0.459	0.532	0.513	0.357	0.000	0.000
Infections and infestations	Sinusitis	0.012	0.302	0.697	0.687	0.422	0.280	0.000
Injury, poisoning, and procedural complications	Excoriation	0.030	0.680	0.296	0.276	0.175	0.000	0.000
Skin and subcutaneous tissue disorders	Ecchymosis	0.005	0.457	0.535	0.524	0.437	0.000	0.000

In Table 4, we notice that the p-value for ecchymosis is smaller than that for sinusitis, indicating more evidence of risk difference for ecchymosis. However, as shown in Table 5, the AEs in the skin and subcutaneous tissue disorders SOC, to which ecchymosis belongs, do not show a consistent pattern of adverse effect. In fact, about half of the PTs had negative treatment differences (i.e., the placebo rate was higher than the treatment rate). Thus, there seems to be a signal when we consider ecchymosis alone, but it will not be flagged in the context of the other AEs within the same SOC. The Bayesian model adjusts for this accordingly. On the other hand, in the infections and infestations SOC, to which sinusitis belongs, most AEs show higher risk on treatment arm. This supports what we have observed for sinusitis. Thus, the posterior exceedance probability for sinusitis is higher than that for ecchymosis after taking into consideration the other PTs in the same SOC. Similarly, excoriation, examined alone, seems to present safety concerns with a

Table 5 AEs in PT for SOC "skin and subcutaneous tissue disorders"

Preferred term	Ctrl r (%)	Trt r (%)	Diff (Trt – Ctrl)	Preferred term	Ctrl r (%)	Trt r (%)	Diff (Trt – Ctrl)
1 Acne	0.28	0.56	0.28	27 Onychorrhexis	0.00	0.08	0.08
2 Actinic keratosis	0.00	0.16	0.16	28 Pain of skin	0.69	0.00	−0.69
3 Alopecia	0.42	0.24	−0.18	29 Periorbital edema	0.14	0.00	−0.14
4 Alopecia areata	0.14	0.00	−0.14	30 Photosensitivity reaction	0.28	0.08	−0.20
5 Angioneurotic edema	0.00	0.08	0.08	31 Precancerous skin lesion	0.00	0.08	0.08
6 Aquagenic pruritus	0.14	0.00	−0.14	32 Pruritus	2.08	0.80	−1.28
7 Cold sweat	0.14	0.00	−0.14	33 Pruritus, allergic	0.14	0.08	−0.06
8 Decubitus ulcer	0.14	0.00	−0.14	34 Pruritus,, generalized	0.14	0.16	0.02
9 Dermal cyst	0.00	0.32	0.32	35 Psoriasis	0.69	0.48	−0.21
10 Dermatitis	0.28	0.16	−0.12	36 Pustular psoriasis	0.14	0.08	−0.06
11 Dermatitis, allergic	0.28	0.08	−0.20	37 Rash	0.69	0.56	−0.13
12 Dermatitis, contact	0.69	0.80	0.11	38 Rash, generalized	0.00	0.08	0.08
13 Dry skin	0.00	0.08	0.08	39 Rash, maculopapular	0.00	0.08	0.08
14 Dyshidrosis	0.00	0.08	0.08	40 Rash, papular	0.28	0.08	−0.20
15 Ecchymosis	0.00	0.96	0.96	41 Rosacea	0.28	0.00	−0.28
16 Eczema	0.14	0.24	0.10	42 Seborrhoeic dermatitis	0.28	0.00	−0.28
17 Erythema nodosum	0.14	0.00	−0.14	43 Skin burning sensation	0.00	0.08	0.08
18 Guttate psoriasis	0.00	0.08	0.08	44 Skin fissures	0.14	0.08	−0.06
19 Hair growth abnormal	0.00	0.08	0.08	45 Skin hemorrhage	0.00	0.08	0.08
20 Heat rash	0.14	0.08	−0.06	46 Skin hyperpigmentation	0.14	0.00	−0.14
21 Hyperhidrosis	0.14	0.08	−0.06	47 Skin lesion	0.00	0.16	0.16
22 Ingrowing nail	0.00	0.08	0.08	48 Skin ulcer	0.56	0.08	−0.48
23 Intertrigo	0.14	0.24	0.10	49 Stasis dermatitis	0.14	0.00	−0.14
24 Milia	0.00	0.08	0.08	50 Urticaria	0.28	0.80	0.53
25 Night sweats	0.14	0.24	0.10	51 Vitiligo	0.14	0.00	−0.14
26 Onychomadesis	0.00	0.16	0.16				

Note. r = subject incidence (%).

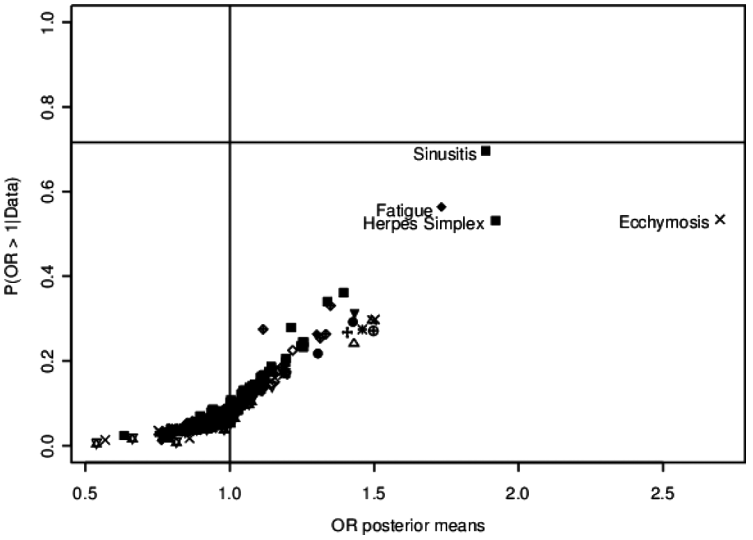

Figure 2 Volcano plot of posterior estimates from binomial hierarchical model with mixture prior.

Fisher's exact test p-value of 0.03. However, the Bayesian hierarchical model shows that it is no longer alarming (with the probability of OR $>$ 1 only being 0.296) after incorporating information of other PTs in the same SOC. Table 4 also shows that the posterior probabilities of RD $>$ 2% are all very low for these five AEs, further indicating that none of these events are of very much concern in terms of sheer risk difference.

Figure 2 is the volcano plot of results from hierarchical Model 1b. It can be seen that sinusitis has the highest posterior probability of OR $>$ 1, about 0.7, followed by fatigue, herpes simplex, and ecchymosis, which are all higher than 0.5. Similar to Fig. 1, this plot also uses different symbols combinations to distinguish PTs from different SOCs.

3.3. Effect on CIs by Model

Figure 3 compares the 95% CIs from different models for the five PTs for which the Fisher's two-sided test p-values are less than 0.05. The first group of CIs are the frequentist approach exact CIs for OR (Thomas, 1971), unadjusted for multiplicity. For both ecchymosis and excoriation, ORs are not well defined, so the upper limits of CI are set to be infinity and only the lower limits are computed. In general, Bayesian models deliver narrower CIs, indicating higher precision with respect to the estimates. This is mainly due to the information-borrowing feature of Bayesian hierarchical models. As we have pointed out, the nonhierarchical one-stage Bayes model (Model 1c) treats all PTs as independent and does not allow borrowing of information. Thus, with vague prior information, Model 1c should be very close to the unadjusted frequentist approach. This similarity can be easily seen in Fig. 3. The other two Bayesian models both have three-level hierarchies. Their results are similar, but Model 1b allows for a point mass at no treatment effects and results

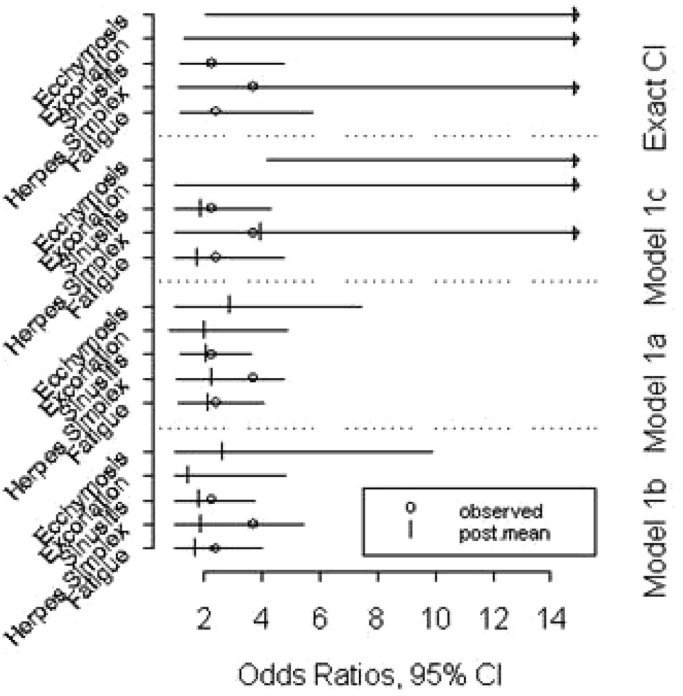

Figure 3 Confidence intervals or credible intervals by model.

in slightly wider CIs. It can be seen that the hierarchical Bayesian models still yield narrower CIs compared with the frequentist approach, as the biologic structure of the data has been taken into consideration.

4. SIMULATION STUDY

In this section, we describe a simulation study that evaluates the performance of the Bayesian hierarchical model with mixture prior (Model 1b) under a null scenario and other scenarios with elevated risks. Simulation studies are conducted to better understand the operating characteristics of the models (Berry et al., 2010, sec. 2.5.4). The importance of type I error in terms of reducing false signals has been well explained elsewhere. However, type II error may be of more importance in safety assessment since missing an actual drug–AE interaction is probably a much greater public health risk than falsely identifying a potential drug–AE interaction. The nonhierarchical one-stage Bayes binomial model (Model 1c), Fisher's two-sided exact test unadjusted for multiple comparisons, controlled for FDR using the Benjamini–Hochberg approach (referred to as BH), and using the DFDR approach were also used as references in our investigation. Using simulation, our goal is to investigate whether modeling the SOC/PT hierarchy is helpful and what operating characteristics should be chosen to achieve acceptable FDR and power for different risk sizes.

Our original data set has 1965 subjects, 720 on the placebo arm and 1245 on the treatment arm, with 1166 subjects having at least one AE. As such, to create our simulated null settings we randomly assign 1245 subjects to treatment group and the rest to control. The AEs of each subject are still retained to preserve the SOC/PT hierarchy structure and intra-individual correlation of the AEs. We create 500 such null data sets.

The parameters we would like to investigate in terms of impact on the power of the models in detecting safety signals are the effect sizes of the risks, whether the AEs are relatively common, and the number of PTs within a SOC. We identified two SOCs, one having 3 PTs with AEs and one having 21 PTs with AEs. SOCs with this number of AEs are fairly typical in clinical trials. Within each SOC, we set the AE rates in the placebo arm to be 1%, 5%, and 10%, respectively, to represent relatively infrequent to common AEs. For moderate- and high-risk scenarios we set the log OR to be normally distributed with mean log 2 or log 5, respectively, with standard deviation 0.25. We then get the posterior exceedance probabilities using Model 1b and Model 1c on the simulated data sets, using different combinations of d^* and p in $\Pr(RR_{bj} > d^* \mid Data) \geq p$ for inference.

In Table 6, we summarize the results of the simulation study by presenting the FDR and power of different methods under the null and "moderate elevated risk" (OR = 2) in both the SOCs with 3 PTs and 21 PTs scenarios. The FDR is calculated as the proportion of falsely identified signals among all signals identified. The power is calculated as the proportion of correctly identified signals among all true signals.

Table 6 Summary of simulation study

		Null scenario	OR = 2 with placebo risk 1%		OR = 2 with placebo risk 5%		OR = 2 with placebo risk 10%	
		FWER	FDR	Power	FDR	Power	FDR	Power
Nonadjusted Fisher's exact test	Two-sided test, p-value <= 0.05	1	0.35	0.36	0.19	0.83	0.18	0.92
BH	q-value ≤ 0.1	0.01	0.01	0.06	0.01	0.66	0.01	0.83
DFDR	$p^* = 0.025$, $p^{\#} = 0.05$	0.06	0.03	0.21	0.01	0.82	0.01	0.92
Nonhierarchical Bayes model* (Model 1c)	$d^* = 1,\ p = 0.6$	1	0.48	0.23	0.24	0.67	0.21	0.82
	$d^* = 1,\ p = 0.7$	0.91	0.35	0.19	0.14	0.64	0.11	0.80
	$d^* = 1,\ p = 0.8$	0.65	0.22	0.16	0.07	0.61	0.06	0.77
Bayesian hierarchical model* (Model 1b)	$d^* = 1,\ p = 0.6$	0.17	0.07	0.76	0.08	0.95	0.08	0.97
	$d^* = 1,\ p = 0.7$	0.11	0.05	0.69	0.05	0.93	0.05	0.96
	$d^* = 1,\ p = 0.8$	0.07	0.03	0.60	0.02	0.90	0.02	0.95

Note. FDR (false discovery rate) is calculated as the proportion of falsely flagged signals among those flagged signals. Power is calculated as the proportion of correctly flagged signals among true signals that should be flagged. For BH method, i.e., Benjamini Hochberg's FDR method, the q-value is the FDR cutoff value. For DFDR method, p^* is the SOC level FDR cutoff value and $p^{\#}$ is the PT level FDR cutoff value. For Bayesian Model 1c and Model 1b, the inference was based on $\Pr(\mathrm{OR}_{bj} > d^* \mid Data) \geq p$

Overall, the Bayesian hierarchical Model 1b outperforms both the nonhierarchical one-stage Bayesian model and the unadjusted Fisher's exact test in the sense that the Bayesian hierarchical model has lower FDR and higher power for all scenarios examined. It is worth mentioning that for the null scenario, the unadjusted Fisher's exact test has an FDR equal to 1. That is, this method always claims detection of signals when there are actually no elevated risks in the data set. The nonhierarchical one-stage Bayesian model also has very high FDR. The BH method has lowest FDR for all scenarios examined. However, the power is also much lower, especially when risk on the placebo arm is less frequent (1% and 5%). Using the DFDR method, the FDR is similar to but the power is lower than those of the Bayesian hierarchical Model 1b (with $d^* = 1$ and $p = 0.8$). The reduction of power is more noticeable when the background placebo rate is less frequent. The performance of all models is better when the AE types are more common (i.e., when the background placebo risk is higher).

FDR and power were also calculated by SOC, though these conditional values are not shown in Table 6. These results show power to detect PTs in the larger SOC (with 21 PTs) is usually higher than that to detect those in the smaller SOC (with 3 PTs). Similar patterns are seen for the scenarios when OR = 5, but the performance of all models is better with the higher odds ratio. Clearly the hierarchical modeling is helpful in reducing FDR and improving power. Based on the simulation results, the combination of cutoffs $d^* = 1$ and $p = 0.8$ seem to have preferable statistical properties for all scenarios examined. Thus, we recommend using this pair of threshold values for subsequent data analysis.

5. DISCUSSION AND FUTURE WORK

The current practice of flagging routinely collected AEs based on unadjusted p-values or CIs is not satisfactory, because it can lead to an excessive number of false positive signals. This in turn can cause undue concern for drug approval and labeling, and perhaps even result in unnecessary postmarketing commitment at the time of product registration. In this paper, we approached this problem via a Bayesian hierarchical mixture modeling framework for analyzing binary and count data. There are a few advantages of our Bayesian approach. First, in contrast to considering each type of AE independently, it allows for explicitly modeling AEs with the existing MedDRA coding structure, so that strength can be borrowed within and across SOCs. As a result, the CIs shrink, as opposed to being too wide (conservative) as in the traditional approach. A nice feature of the Bayesian hierarchical modeling is that data will tell how much borrowing should take place. That is, if an AE has a different pattern compared to other AEs in the same SOC, it will essentially lead to less borrowing. While hierarchical models are widely used in practice and can be fitted using non-Bayesian approaches (Davidian and Giltinan, 1995; Diggle et al., 1994), Bayesian approaches are more flexible to accommodate complex modeling requirements. Second, the Bayes approach is attractive in dealing with rare AE data, because the model adaptively modulates the extremes. In addition, our inferences are based on the full posterior distributions, relaxing the need to assume normality, which may not be sensible for rare events. Next, the use of a mixture prior allows a point mass on equality of the treatment and control rates, sensible for AEs not causally related to the drug. Finally, it is straightforward

to assess the posterior probability of clinically important differences on different scales (risk difference, OR, or RR), to avoid detecting medically unimportant signals.

Our simulation study results showed that compared with the Bayesian hierarchical approach, the BH method successfully controls for FDR but does not have enough power to detect drug–AE interactions, especially when elevated risk is moderate and event rate is less frequent. The DFDR method has an overall satisfactory performance. However, as rare events (count of event less than five in both groups combined) are removed as a preprocessing step, the power is lower when event are not common (e.g., 1%).

Although in this paper we modeled AEs with the SOC/PT structure under the MedDRA, our approach could also be used to model a more granular structure (e.g., SOC/HLGT/PT) or a different structure entirely (e.g., HLT/PT) if justified by the biologic relationship. Ideally, a group of PTs allowed to borrow strength from each other should reflect a clinical entity or concept. One might be concerned that SOC is too broad and general. For instance, PTs under a SOC may not be medically related; e.g., the SOC "injury, poisoning, and procedural complications" may cover PTs ranging from "transplant rejection" to "alcohol poisoning," which are medically unrelated. It has been suggested that HLGT/PT or HLT/PT may be better than SOC/PT because an HLGT/HLT would typically reflect a medical/clinical concept more closely than an SOC. However, for certain data sets like our motivating example, where more than 75% of the HLTs contain 1 or 2 PTs with AEs, there might not be much smoothing (shrinkage) allowed if we model HLT/PT. One of our future tasks is to model HLT/PT instead of SOC/PT if there is a suitable data set. Furthermore, predefined clinical events of interest for a specific product can be modeled if the coding search strategies for these events are available. An example in such a setting could be to use a modeling strategy based on the Standard MedDRA Query (SMQ) so the model would allow PTs under an SMQ to borrow more strength from each other.

Our approach in this paper has been to use hierarchical modeling to smooth AE response rates within and across subgroups. One may be concerned that this may have the effect of masking outliers, which may turn out to be of greatest interest. Our view is that our goal is only to detect "real" signals based on reliably large sample sizes and conforming to our view of how AEs worthy of identification tend to cluster. By contrast, our methods seek to limit the identification of outlying "false" signals that do not fit our conceptual model for how safety signals arise (e.g., we would not want to flag an AE that was unlike every member of its SOC). Thus, while we view our hierarchical shrinkage of outliers as a virtue of the method, it is important to note that other models may be more appropriate (e.g., shrinking defined independently of SOCs). Clearly, our current model is just a first attempt, and other forms of shrinkage informed by medical judgment or using published SMQs could also be considered.

Graphical displays like the volcano plot are very effective in displaying flagged signals when analyzing the hundreds or even thousands of AE types in a trial. In particular, the volcano plot can illustrate both the magnitude of the treatment effect and strength of statistical evidence in the same graph, so one can visualize the flagged signals and potential clustered events stemming from the same biologic

system, such as a system of organ class. A drill-down feature can be developed for those AEs identified in the upper right quadrant to facilitate further investigation.

Our application analyzed AE data in a single trial setting. However, multiple clinical trials are usually conducted in a typical new drug development program. One future direction is to incorporate individual subject data from multiple studies into the model so an "individual patient data meta-analysis" of comparing the AE rates between the treatment and control groups can be performed, in order to gain more power and perhaps more insight as to whether there is a consistency in flagging the same signal among the studies. In addition, if there are subject-level characteristics or subgroups, it will be very straightforward to add covariates to our Bayesian hierarchical regression models.

A full Bayesian analysis of data sets like ours requires study of the robustness of our conclusions to changes in our model assumptions, especially informative prior specifications (Carlin and Louis, 2008). In section 2.1, we set a reasonably noninformative hyper prior N(0, 10) for mean log-OR (roughly 95% sure the odds ratio for treatment vs. control is between 0.002 and 492, with an estimate of the mean OR of 1.0) and IG (3,1) for the variance parameters at stage 3. The hyper parameters of the beta prior distribution for the point mass probability π_b are assigned to be $\alpha_\pi \sim Exp(0.1)I[\alpha_\pi > 1]$ and $\beta_\pi \sim Exp(0.1)I[\beta_\pi > 1]$, which correspond to a flat prior for π_b. In order to evaluate robustness of the posterior inferences to the prior specifications, we carried out sensitivity analysis with the following hyper priors: a skeptical prior N(0.7, 10) for mean log-OR (roughly 95% sure the odds ratio for treatment vs. control is between 0.004 and 990, with an estimate of the mean OR of 2.0), an IG(1, 0.1) for the variance parameters, a Unif(0, 2) for the standard deviation parameters, and $\alpha_\pi \sim Exp(1)I[\alpha_\pi > 1]$ and $\beta_\pi \sim Exp(1)I[\beta_\pi > 1]$, which correspond to a prior for π_b that is more centered around 0.5. Overall the effects of these changes were minor. Therefore, our results appear to be robust to different prior specifications.

A key feature of the Bayesian approach to the analysis of clinical data is its ability to permit continuous monitoring of the data, rather than restricting attention to a prespecified series of looks, possibly determined by an alpha-spending function (Lan and DeMets, 1983). This is because Bayesian decisions are not made using p-values, but simply by summarizing the posterior or predictive distribution, which in accordance with the likelihood principle (Birnbaum, 1962) can be done at any time. Thus, as an extension of our current approach we could permit earlier stopping if the available data indicated an elevated AE rate. Bayesian stopping rules of this sort are commonly used; see Berry et al. (2010) for a variety of examples that simulate the probabilities of early stopping and the numbers of patients saved in such cases. Of course, such flexible stopping rules will alter the frequentist operating characteristics (power; type I and II error) of the procedure , so such summaries are typically simulated as well Berry et al. (2010).

As a final caveat, the improvements in FDR and power arising from our Bayesian hierarchical model manifest under the assumption that the statistical model is correctly specified. Future work looks to more careful investigation of what may happen when the model is misspecified (Begg and Lagakos, 1990).

In conclusion, the field of clinical trial signal detection is still in its infancy. More research and more experience with existing models are needed. Our tool ultimately serves as a flagging device, and as such can never replace medical

judgment. As safety evaluation embraces multidisciplinary collaborations, further developments await statisticians working closely with clinicians and safety scientists to advance this field.

ACKNOWLEDGMENTS

We are grateful to Dr. George Rochester for helpful discussions, and to Drs. Steven Snapinn and George Williams for constructive comments and suggestions. We also thank two referees for their critical reading of this material and for their insightful comments that greatly improved the paper.

REFERENCES

Bate, A, Lindquist, M., Edwards, I. R., Olsson, S., Orre, R., Lansner, A., De Freitas, R. M. (1998). A Bayesian neural network method for adverse drug reaction signal detection. *European Journal of Clinical Pharmacology* 54:315–321.

Begg, M. D., Lagakos, S. (1990). On the consequences of model misspecification in logistic regression. *Environmental Health Perspectives* 87:69–75.

Benjamini, Y., Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B (Methodological)* 57(1):289–300.

Benjamini, Y., Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics* 29(4):1165–1188.

Berry, S., Berry, D. (2004). Accounting for multiplicities in assessing drug safety: A three-level hierarchical mixture model. *Biometrics* 60:418–426.

Berry, S. M., Carlin, B. P., Lee, J. J., Muller, P. (2010). *Bayesian Adaptive Methods for Clinical Trials*. Boca Raton, FL: Chapman and Hall/CRC Press.

Birnbaum, A. (1962). On the foundations of statistical inference (with discussion). *Journal of the American Statistical Association* 57:269–326.

Carlin, B. P., Louis, T. A. (2008). *Bayesian Methods for Data Analysis*. 3rd ed. Boca Raton, FL: Chapman & Hall/CRC Press.

Chi, G., Hung, H. M. J., O'Neill, R. (2002). Some comments on "Adaptive Trials and Bayesian Statistics in Drug Development" by Don Berry. *Pharmaceutical Report* 9:1–11.

Crowe, B., Xia, H. A., Berlin, J., Watson, D., Shi, H., Lin, S., Kuebler, J., et al. (2009). Recommendations for safety planning, data collection, evaluation and reporting during drug, biologic and vaccine development: A report of the Safety Planning, Evaluation and Reporting Team (SPERT). *Clinical Trials* 6:430–440.

Davidian, M., Giltinan, D. M. (1995). *Nonlinear Models for Repeated Measurement Data*. Boca Raton, FL: Chapman and Hall/CRC Press.

Diggle, P. J., Liang, K.-Y., Zeger, S. L. (1994). *Analysis of Longitudinal Data*. Oxford, UK: Oxford University Press.

DuMouchel, W. (1999). Bayesian data mining in large frequency tables, with an application to the FDA Spontaneous Reporting System (with discussion). *American Statistician* 53:177–202.

Gould, A. L. (2002). Drug safety evaluation in and after clinical trials. Presented at the Deming Conference, Atlantic City, NJ, 3 December.

Lan, K. K. G., DeMets, D. L. (1983). Discrete sequential boundaries for clinical trials. *Biometrika* 70:659–663.

Mehrotra, D. V., Heyse, J. F. (2004). Multiplicity considerations in clinical safety analysis. *Statistical Methods in Medical Research* 13:227–238.

O'Connell, M. (2007). *Statistical graphics for the design and analysis of clinical development studies*. http://www.insightful.com/news_events/webcasts/2006/07clinical/default.asp

O'Neill, R. T. (1998). Assessment of safety. In: *Biopharmaceutical Statistics in Drug Development*. Chap. 13. New York: Marcel Dekker.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., van der Linde, A. (2002) Bayesian measures of model complexity and fit (with discussion). *Journal of the Royal Statistics Society B* 64:583–640.

Thomas, D. G. (1971). Algorithm AS-36. Exact confidence limits for the odds ratio in a 2×2 table. *Applied Statistics* 20:105–110.

APPENDIX I. SAMPLE WinBUGS CODE

```
## Data example:

list(Nae = 465, Nc = 720, Nt = 1245, B = 24, b = c(1, 2, 2, 2,
2, 2, 2, 2, 2, 3, 3, 4, 4, 4, 4, 4, 4, 4, 4, 4, 5, 6, 6, 6, 6,
6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 7, 7, 7, 7, 7, 7, 7, 7, 7, 7, ...),
j = c(1, 1, 2, 3, 4, 5, 6, 7, 8,
1, 2, 1, 2, 3, 4, 5, 6, 7, 8, 9, 1, 1, 2, 3, 4, 5, 6, 7, 8, 9,
10, 11, 12, 13, 14, 15, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, ...),
Y = c(10, 0, 1, 1, 1, 1, 2, 0, 0, 0,
0, 0, 1, 5, 0, 0, 0, 2, 4, 2, 0, 1, 1, 2, 5, 1, 1, 3, 1, 1, 1,
0, 1, 1, 0, 1, 1, 2, 1, 6, 2, 7, 2, 0, 1, 0, 0, 4, 29, 4, 15, ...),
X = c(6, 1, 0, 0, 0, 1, 4, 1, 1,
1, 1, 1, 0, 4, 1, 1, 1, 0, 1, 0, 1, 1, 0, 0, 1, 0, 0, 0, 0, 0,
1, 1, 0, 0, 1, 1, 3, 0, 0, 2, 3, 2, 1, 1, 0, 1, 2, 6, 15, 1, ...))


###################################
## M1b: Binomial model           ###
##      with mixture prior       ###
###################################

model{

for (i in 1:Nae) {

X[i] ~ dbin(c[b[i], j[i]], Nc)
       Y[i] ~ dbin(t[b[i], j[i]], Nt)

       logit(c[b[i], j[i]]) <- gamma[b[i], j[i]]
       logit(t[b[i], j[i]]) <- gamma[b[i], j[i]] + theta[b[i],
       j[i]]

       gamma[b[i], j[i]] ~ dnorm(mu.gamma[b[i]], tau.gamma
       [b[i]])
```

```
        p0[i] ~ dbern(pi[b[i]] ) # prob of point mass
theta1[b[i], j[i]]~ dnorm(mu.theta[b[i]], tau.theta[b[i]])

        # theta=0 w.p. pi[i] and theta=theta1 w.p. 1-pi[i]
        theta[b[i], j[i]] <- (1- p0[i]) * theta1[b[i], j[i]]

        OR[b[i],j[i]] <- exp(theta[b[i],j[i]] )
ORpv2[b[i], j[i]] <- step(OR[b[i],j[i]] - 2 )     # OR >= 2
        ORpv2[b[i], j[i]] <- step(OR[b[i],j[i]] -1.2 )
        # OR >= 1.2
        ORpv[b[i], j[i]] <- 1- step(-OR[b[i],j[i]])     # OR >1

        RD[b[i], j[i]]<- t[b[i], j[i]] -c[b[i], j[i]]
        RDpv[b[i], j[i]]<- 1- step(c[b[i], j[i]] -t[b[i], j[i]])
             # RD>0
        RDpv2[b[i], j[i]]<- step(t[b[i], j[i]] -c[b[i],
        j[i]]- 0.02)   # RD>=2%
        RDpv5[b[i], j[i]]<- step(t[b[i], j[i]] -c[b[i],
        j[i]]- 0.05)   # RD>=5%

        D[i]<-X[i]*log(c[b[i], j[i]])+(Nc-X[i])*log(1-c[b[i],
        j[i]]) + Y[i]*log(t[b[i], j[i]])+(Nt-
Y[i])*log(1-t[b[i],j[i]])
        }

         Dbar<- -2* sum(D[])     # -2logL without normalizing
         constant

        # SOC level parameters

for(k in 1:B){
        pi[k] ~ dbeta(alpha.pi, beta.pi)

        mu.gamma[k] ~ dnorm(mu.gamma.0, tau.gamma.0)
        tau.gamma[k] ~ dgamma(3,1)

        mu.theta[k] ~ dnorm(mu.theta.0, tau.theta.0)
        tau.theta[k] ~ dgamma(3,1)

}

        # hyperpriors for gamma's;
        mu.gamma.0 ~ dnorm(0, 0.1)
        tau.gamma.0 ~ dgamma(3,1)

        # hyperpriors for theta's;
        mu.theta.0 ~ dnorm(0, 0.1)
        tau.theta.0 ~ dgamma(3,1)
```

```
        # hyperpriors for pi's;
        alpha.pi ~ dexp(0.1) I(1, )
        beta.pi ~ dexp(0.1) I(1, )
}

##############################
## M2b: Poisson model       ###
##      with mixture prior  ###
##############################
model{
for (i in 1:Nae)
 {
 X[i] ~ dpois(c[ b[i], j[i] ] )
 Y[i] ~ dpois( t[ b[i], j[i] ] )

 c[ b[i], j[i] ] <- lambda_c[ b[i], j[i] ] * dur0[i ]
 t[ b[i], j[i] ] <- lambda_t[ b[i], j[i] ] * dur1[i ]

 log(lambda_c[b[i], j[i]]) <- gamma[b[i], j[i]]
 log(lambda_t[b[i], j[i]]) <- gamma[b[i], j[i]] + theta[b[i],
  j[i]]

 tgamma[i]~ dnorm(mu.gamma[b[i]], tau.gamma[b[i]])
 gamma[b[i], j[i]] <- tgamma[i]

 p0[i] ~ dbern(pi[b[i]] )
 z[i] <- 1- p0[i]
 ttheta[i]~ dnorm(mu.theta[b[i]], tau.theta[b[i]])
 theta1[b[i], j[i]] <- ttheta[i]

 theta[b[i], j[i]] <- (1- p0[i] ) * theta1[b[i], j[i]]

 D[i]<- X[i]*log(c[ b[i], j[i] ]) - c[ b[i], j[i] ]-t[ b[i], j[i]]
  + Y[i]*log(t[ b[i], j[i] ])

 RR[b[i],j[i]] <- exp(theta[b[i],j[i]] )
 RRpv20[b[i], j[i]] <- step(RR[b[i],j[i]] -2 )   # RR >= 2
 RRpv2[b[i], j[i]] <- step(RR[b[i],j[i]] -1.2 )   # RR >=1.2
 RRpv[b[i], j[i]] <- 1- step(-RR[b[i],j[i]])      # RR >1


 RD[b[i], j[i]]<- lambda_t[b[i], j[i]] - lambda_c[b[i], j[i]] #
  unit: pt/yr
 RDpv[b[i], j[i]]<- 1 - step(lambda_c[b[i], j[i]] -
  lambda_t[b[i], j[i]] )  # Risk trt > Risk ctrl
 RDpv2[b[i], j[i]]<- step(lambda_t[b[i], j[i]] - lambda_c[b[i],
  j[i]]- 2) # RD>= 2/pat-yr
```

```
 RDpv5[b[i], j[i]]<- step(lambda_t[b[i], j[i]] - lambda_c[b[i],
  j[i]]- 2) # RD>= 5/pat-yr

       }

       Dbar<- -2* sum(D[])     # -2logL without normalizing
       constant

       # SOC level parameters
   for(k in 1:B){
       pi[k] ~ dbeta(alpha.pi, beta.pi)
       mu.gamma[k] ~ dnorm(mu.gamma.0, tau.gamma.0)
       tau.gamma[k] ~ dgamma(3,1)

       mu.theta[k] ~ dnorm(mu.theta.0, tau.theta.0)
       tau.theta[k] ~ dgamma(3,1)
       }

       # hyperpriors for gamma's;
       mu.gamma.0 ~ dnorm(0, 0.1)
       tau.gamma.0 ~ dgamma(3,1)

       # hyperpriors for theta's;
       mu.theta.0 ~ dnorm(0, 0.1)
       #mu.theta.0<-0.5
       tau.theta.0 ~ dgamma(3,1)

       # hyperpriors for pi's;
       alpha.pi ~ dexp(.1) I(1, )
       beta.pi ~ dexp(.1) I(1, )

}
```

APPENDIX II. SAMPLE S-PLUS CODE FOR THE VOLCANO PLOT

```
ppPlot<-function(indata = data4plot,
            X    = "risk difference",
            Y    = "pvalue",
   xlabel = "P(theta>0|Data)",
   ylabel = "P-value ( -log(p) )",
            socCol = "AESOC",
            ptCol  = "AEPT",
            hline  = -log(0.05),
            vline  = 0 ,
            xlimits=NULL,
            ylimits=NULL,
```

```
          plottitle="",
          displayLegend=T, legendcorner=c(0,0),
          labelquad=c(3,4),
          colorSOC=T) {
  op <- par()

  if(is.null(xlimits)) xlimits <- range(indata[,X])
  if(is.null(ylimits)) ylimits <- range(indata[,Y])

  plot(indata[,X], indata[,Y], type="n", xlab=xlabel,
   ylab=ylabel,main=plottitle, xlim=xlimits, ylim=ylimits)

  abline(h= hline)
  abline(v= vline)

  if(3 %in% labelquad & 4 %in% labelquad){
       indx <- indata[,X] > vline } else {
          indx <- indata[,X] > vline & indata[,Y] > hline}

  if(colorSOC == F) {
      points(indata[,X], indata[,Y], cex = 0.7)
      text(x = indata[indx,X], y =  indata[indx,Y], labels
           = paste(" ", labels = as.character(indata[[ptCol]])
           , sep =
          "")[indx], cex = 0.6, adj = 0)
  } else{
           unq.soc <- unique(as.character(indata[[socCol]] ))
       unq.soc <- unq.soc[is.na(unq.soc) == F]
       scol <- c(1, 2, 3, 4, 5, 7, 8)
       scol <- rep(scol, (length(unq.soc)/length(
         scol)) + 1)[1:length(unq.soc)]
       spch <- seq(1, 6)
       spch <- rep(spch, (length(unq.soc)/length(
         spch)) + 1)[1:length(unq.soc)]
       socrank <- match(as.character(indata[[socCol]]),
         unq.soc)
         points(indata[,X], indata[,Y], cex = 0.7,
                col=scol[socrank], pch=spch[socrank])
         text(x = indata[indx,X], y = indata[indx,Y],
              labels = paste(" ", labels = as.character(indata
              [[ptCol]]), sep = "")[indx],
              cex = 0.6, adj = 0, col=scol[socrank][indx])
if(displayLegend) {
    socLeg <- unq.soc
    socLeg <- sapply(unq.soc, function(x)
    {
      out <- x
      if(nchar(x) > 25)
```

```
          out <- paste(substring(
            x, 1, 25),
            "...", sep = ""
            )
      invisible(out)
    })

    key(corner = legendcorner, points = list(
      col = scol, pch = spch, cex =
      0.6), text = list(socLeg, col
      = scol, cex = 0.6),
      transparent = F, border = T)
  }

if(names(dev.cur()) == "java.graph") {
    metalabels <- as.character(indata[[ptCol]])
    if(length(socCol) != 0)
    metalabels <- paste(metalabels, "\nSOC=",
           as.character(indata[[socCol]]),
           sep = "")
if(length(numAETrtCol) != 0) {
    if(length(allnumTrt) != 0)
         metalabels <- paste(metalabels,
            "\n", xaxisTrtLabel, " n=",
            indata[[numAETrtCol]],
            "(", round(indata[[
            numAETrtCol]]/allnumTrt *
            100, digits = 2), "%)",
            sep = "")
      else metalabels <- paste(metalabels, "\n",
            xaxisTrtLabel, " n=",
            indata[[numAETrtCol]],
            sep = "")
  }
  if(length(numAECtrlCol) != 0) {
      if(length(allnumCtrl) != 0)
          metalabels <- paste(metalabels,
          "\n", xaxisCtrlLabel,
          " n=", indata[[
          numAECtrlCol]], "(", round(
          indata[[numAECtrlCol]]/
          allnumCtrl * 100, digits
           = 2), "%)", sep = "")
      else metalabels <- paste(metalabels, "\n",
            xaxisCtrlLabel, " n=",
            indata[[numAECtrlCol]],
            sep = "")
  }
```

```
  java.identify(xypoint[,1], xypoint[,2],labels = as.character
   (metalabels))
  }

}

  par(op)
  invisible()
  }
```