Biometrics 69, 661–672
September 2013
DOI: 10.1111/biom.12051

Bayesian Approach for Clinical Trial Safety Data Using an Ising Prior

Bradley W. McEvoy,[1,*] Rajesh R. Nandy,[2,3] Ram C. Tiwari[1]

[1]Office of Biostatistics, CDER, FDA, 10903 New Hampshire Ave, Silver Spring, Marryland 20993, U.S.A.
[2]Department of Psychology, University of California at Los Angeles, Los Angeles, California 90095, U.S.A.
[3]Department of Biostatistics, University of California at Los Angeles, Los Angeles, California 90095, U.S.A.
*email: bradley.mcevoy@fda.hhs.gov

Summary. In drug safety, development of statistical methods for multiplicity adjustments has exploited potential relationships among adverse events (AEs) according to underlying medical features. Due to the coarseness of the biological features used to group AEs together, which serves as the basis for the adjustment, it is possible that a single adverse event can be simultaneously described by multiple biological features. However, existing methods are limited in that they are not structurally flexible enough to accurately exploit this multi-dimensional characteristic of an adverse event. In order to preserve the complex dependencies present in clinical safety data, a Bayesian approach for modeling the risk differentials of the AEs between the treatment and comparator arms is proposed which provides a more appropriate clinical description of the drug's safety profile. The proposed procedure uses an Ising prior to unite medically related AEs. The proposed method and an existing Bayesian method are applied to a clinical dataset, and the signals from the two methods are presented. Results from a small simulation study are also presented.

Key words: Bayes Factor; Drug safety; High-dimensional data analysis; Ising prior; Markov random field; MedDRA.

1. Introduction

Accurately characterizing risks for adverse events (AEs) is a challenging yet important feature in all aspects of drug development. In some instances the observed imbalances in risks between study arms for selected safety endpoints are expected based on known pharmacological properties or class effects. An example of this is the association of congestive heart failure (CHF) with the thiazolidinedione (TZD) class of antidiabetic drugs (FDA, 2007). In other instances, the differential risks that are observed are not expected. The challenge, particularly in the latter scenario, is in identifying associations arising by chance. The inability to differentiate a false signal from a true signal can have profound consequences. In a regulatory environment this may include an inaccurate characterization of the risks in the product label or not being able to establish the safety requirements needed for drug approval.

In order to decrease the likelihood of mischaracterizing the risks associated with a drug, several statistical adjustment strategies are available. In some of these approaches the statistical adjustment exploits AEs with common biological features by grouping them together, where an individual AE is characterized by a single feature; see, for example, Mehrotra and Heyse (2004), Xia, Ma, and Carlin (2011), and Rosenkranz (2010), among others. The most popular of these methods is the three-level nested Bayesian hierarchical model proposed by Berry and Berry (2004; henceforth referred to as BB). In this model, for multiplicity adjustment, the risks for the AEs that share a biological feature or a body system are assumed exchangeable, and the body systems themselves are assumed exchangeable at the next level of hierarchy.

Utilizing biological dependencies to inform statistical adjustment has an obvious clinical appeal. However, it is unlikely that the aforementioned statistical methods accurately capture complex correlations present in clinical safety data due to multiple ways an AE can be described or characterized, such as by site of manifestation or etiology. To illustrate the potential implication, consider an AE grouped by site of manifestation. This AE is assumed to be related, in some sense, to other AEs with the same site of manifestation. In a hierarchical approach the problem becomes when an AE may also be related, in other meaningful ways, to one or more AEs belonging to possibly different groups. When this happens, the influence of AEs within a group is different than the influence of the AEs not in the group, irrespective of whether they have a biological relationship or not. Making the groups more granular would not resolve this problem due the complex dependencies in drug safety data as well as data sparseness issues.

The structural rigidness of the hierarchical formulation led Berry et al. (2010) to question the appropriateness of the model in BB. Therefore, the development of statistical methods that adjust for multiplicity while preserving the multidimensional characterization of an AE, and thereby more accurately representing the safety profile of the drug, is warranted. To accommodate the multi-dimensional characterization of an AE, we propose a Bayesian hypothesis testing approach. Each AE is assigned a binary indicator encoding the hypothesis test that the risk is non-differential between the treatment and comparator arms versus the hypothesis that the risk is differential. Repeating the hypothesis evaluation across a total of K AEs, results in binary surface of K risk differential indicators. The assumed biological dependencies among AEs are then incorporated into the hypothesis evaluations by using a binary Markov random field (MRF) prior, namely the Ising prior (Ising, 1925), for the binary surface.

The proposed application of the Ising model was first used by Smith et al. (2003) and Smith and Fahrmeir (2007) for functional magnetic resonance imaging (fMRI) data, wherein they extended the seminal work of George and McCullough (1993) to accommodate variable selection in spatially dependent simultaneous regression models. To our knowledge, the use of Ising model for high-dimensional hypothesis evaluation across multiple endpoints has not been applied beyond fMRI data. Therefore, the proposed method serves as an introduction of the use of the Ising model to the arena of clinical safety data.

The proposed framework differs from the nested hierarchical approach in important ways, beyond its ability to incorporate the multi-dimensional nature of clinical safety data. One difference is the proposed approach shares information between AEs through the risk differential indicators, instead of amplitude and precision of the comparative metric (e.g., odds ratios) as done in the hierarchical approach. Another difference is the proposed approach can be applied in various settings, ranging from a few AEs that are of clinical interest to all or most of the AEs observed in a study. The hierarchical approach is typically reserved for investigations of the latter type as variance components within and across features have to be estimated.

In the next section, we provide preliminaries on the Ising model. The underlying statistical model, the prior and posterior distributions are presented in Section 3. In Section 4, we apply the proposed method to the dataset presented in Mehrotra and Heyse (2004). In this section, we also contrast results from the proposed methodology with those in BB. Results from a limited simulation study are presented in Section 5. We conclude in Section 6 with a brief discussion including potential extensions of the methodology for drug safety data. The Appendix outlines the Markov chain Monte Carlo (MCMC) sampling algorithm.

2. Ising Model Preliminaries

The classical Ising model is defined on a lattice, where the neighbors of a site on the lattice are defined by adjacency. However, because relationships between AEs may not be sufficiently captured by a lattice representation of the Ising model, we consider its more general formulation on an undirected graph G. The graph G describes a set of connections or edges between a finite set of K vertices or nodes V, $\mathrm{V} = 1,\ldots,K$. An edge (or connection) exists between two vertices if they are considered neighbors, with the exception that a vertex is not a neighbor with itself. Relationships between vertices are required to be symmetric, so that if vertex i is a neighbor with vertex j, then vertex j is a neighbor with vertex i. We denote the set of vertices that share an edge with the kth vertex by D_k.

For every vertex $k \in \mathrm{V}$ in the graph, let z_k take on values in {0,1} with $\mathbf{z} = (z_1, \ldots, z_K)$. To help illustrate what z and z_k represent by an example, consider the image reconstruction problem, where $\mathbf{z}$ represents the entire black and white picture, and z_k describes whether the kth pixel is black ($z_k = 0$) or white ($z_k = 1$). Let $\mathbf{Z}$ denote the space of the 2^K possible configurations of z. A probability distribution for z is the Gibbs or Boltzmann distribution (Liu, 2004), which has the form

$$P(\mathbf{z}) = \frac{\exp\{-H(\mathbf{z})\}}{\sum_{\mathbf{z}\in\mathbf{Z}} \exp\{-H(\mathbf{z})\}}$$

In the case of the Ising model, H is defined by

$$H(\mathbf{z}) = -\rho \sum\nolimits_k z_k - \theta \sum\nolimits_k \sum\nolimits_{j\in D_k} I(z_k = z_j) \qquad (1)$$

where $-\infty < \rho < \infty$, $\theta \geq 0$, and $I(A)$ is an indicator function with $I(A) = 1$ if the argument A is true, and 0 else. The first term in (1) is referred to as the "external field," where the parameter ρ, tends to orient elements of $\mathbf{z}$ in the direction of the external force (or field), thereby influencing the sparsity of $\mathbf{z}$; $\mathbf{z}$ is less sparse for larger values of ρ. The second term in (1) is referred to as the "interaction term" denoting the interaction between the neighboring vertices, with the parameter θ controlling the amount of spatial smoothing among the neighboring vertices. Larger values of θ result in greater levels of smoothing, with elements of $\mathbf{z}$ being independent if $\theta = 0$ (Smith et al., 2003).

Since the Ising model is a binary MRF, the conditional distribution of z_k, given the rest of the z_k's, depends only on its neighbors (Marin and Robert, 2007), which makes it easy to simulate $\mathbf{z}$ using a single-site Gibbs sampler.

3. Statistical Model

Let y_k^i be the number of subjects, among the N^i subjects randomized to treatment $i = T, C$, who experienced the kth AE, (viz, $\mathrm{AE_k}$ $), k = 1,\ldots,K$, where $i = T$ refers to experimental treatment arm, and $i = C$ the comparator arm. Assume that $y_k^i|\pi_k^i \sim \text{Binomial}(N^i, \pi_k^i)$. We adopt a hypothesis evaluation framework to characterize each AE according to whether the risk between the treatment and comparator arms is non-differential (i.e., supports the null hypothesis, H_{0k}), $\pi_k^T = \pi_k^C = \pi_k$ or differential (i.e., supports the alternative hypothesis, H_{1k}), $\pi_k^T \neq \pi_k^C$. These competing hypotheses are formally evaluated by introducing, for $\mathrm{AE_k}$, an indicator variable, $\gamma_k = \{0, 1\}$. If $\gamma_k = 1$, the two treatment groups for $\mathrm{AE_k}$ have a non-differential risk (NDR) encoded by the model:

$$y_k^i|\pi_k \sim \text{Binomial}(N^i, \pi_k) \text{ with } \pi_k \sim \text{Beta}(\alpha_k, \beta_k) \qquad (2)$$

If $\gamma_k = 0$, the two treatment groups for $\mathrm{AE_k}$ have a differential risk encoded by the models:

$$y_k^i|\pi_k^i \sim \text{Binomial}(N^i, \pi_k^i) \text{ with } \pi_k^i \sim \text{Beta}(\alpha_k^i, \beta_k^i), \;\; i = T, C. \qquad (3)$$

By repeating the hypothesis set-up for each of the K AEs, the collection of risk differential indicators $\boldsymbol{\gamma} = (\gamma_1, ..., \gamma_K)$ characterizes the safety profile for the particular investigation. We refer to $\boldsymbol{\gamma}$ as the *treatment adverse-event hypothesis surface* (TAEHS).

Although there is no theoretical basis to exclude the AEs that did not occur in the study, in what follows here we restrict K to only those AEs that have occurred. It would be very difficult to identify possible signals if such a restriction

was not made. Refer to Crowe et al. (2009) for additional considerations for observed AEs. This restriction may be relaxed when the methodology is applied to a select number of AEs that are of clinical interest, such as the components of the major adverse cardiovascular event (MACE) composite endpoint, even if there were no, say, cardiovascular deaths observed.

3.1. *Ising Prior*

The TAEHS, with 2^K possible configurations of 0's and 1's, is assigned the Ising prior given by

$$f(\boldsymbol{\gamma}|\theta, \rho_k, k = 1, ..., K) \propto \exp\left(\sum_{k=1}^{K} \rho_k \gamma_k + \theta \sum_{k=1}^{K} \sum_{j \in \mathrm{D}_k} I(\gamma_k = \gamma_j)\right) \quad (4)$$

Unlike (1), the Ising model (4) has an AE specific parameter of the external field, ρ_k, to accommodate for differential preference for favoring the null or alternative hypothesis across AEs. These prior beliefs may come from previous studies or known class effects. The inner summation of the interaction term (i.e., the second term in (4)) is limited to a subset of the K-1 AEs, considered a priori related to $\mathrm{AE_k}$, denoted by the set $\mathrm{D_k}$. If desired, θ can be assigned a prior distribution such as a Uniform, $\theta \sim U(0, \theta_{\max})$, where $\theta_{\max}$ is specified. In practice, however, θ is often fixed due to challenges (discussed below) that arise when it is assigned a prior distribution. In the next section, we outline some strategies for specifying the Ising model hyperparameters.

To automate the process of defining relationships between the K AEs we propose using the Medical dictionary for regulatory activities (MedDRA) structure. See Mozzicato (2009) for an overview of MedDRA. Two AEs, defined by their preferred terms (PTs), may be considered neighbors if they have at least one system organ class (SOC) in common. Other levels of the MedDRA could alternatively be used. As MedDRA is a tool for the classification and coding of AEs, not for defining relationships between AEs, we consider it critically important that the assumed relationships are refined and scrutinized based on clinical insight into the disease, drug, and study population being investigated.

It is worth noting that when $\theta = 0$ in (4), the expression for the probability that $\gamma_k = 1$ is the same as if we assumed an independence prior for $\boldsymbol{\gamma}$, that is, $f(\boldsymbol{\gamma}) = \prod_k \delta_k^{\gamma_k}(1-\delta_k)^{1-\gamma_k}$, where $\delta_k = f(\gamma_k = 1)$. This relationship is apparent by defining ρ_k as follows and plugging it into (6). Under independence, the Ising prior has the joint distribution $f(\boldsymbol{\gamma}) \propto \exp\left(\sum \rho_k \gamma_k\right)$ with marginal probability $(1 + \exp(-\rho_k))^{-1}$. If marginal probabilities are set to δ_k, it follows that $\rho_k = \mathrm{logit}(\delta_k)$, where $\mathrm{logit}(x) = \log(x/(1-x))$. This exercise also reveals that the interpretation of the parameter ρ_k is not foreign under independence; ρ_k is the prior belief, on the log-odds scale, that the risks in the treatment and comparator arms for $\mathrm{AE_k}$ are non-differential.

3.1.1. Ising prior hyperparameters. Before describing strategies to specify the Ising prior hyperparameters, we begin by commenting on the practicality of having an AE specific parameter of the external field. Despite our mention of this flexibility in the previous section, in practice it is unlikely that each parameter can be specified in a clinically meaningful way, particularly for large K. This challenge leads us to make simplifying assumptions for ρ_k across the K AEs, which are detailed below.

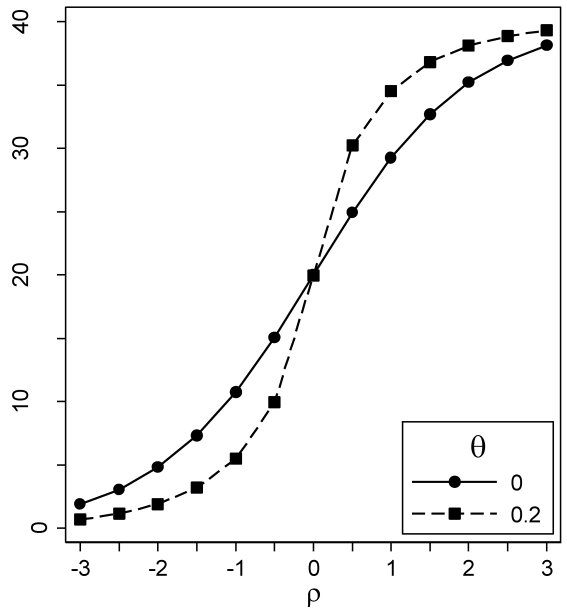

Figure 1. Expected a priori number of AEs with non-differential risk for the primary analysis.

A consideration and feature of the Ising model used to guide hyperparameter selection are the phase transition points. This strategy may also be used to limit or restrict the range of hyperparameters when conducting a prior sensitivity analysis, which is encouraged. Phase transition is a well-known feature of the Ising model and occurs when the underlying state of the system being analyzed, in our case the TAEHS, undergoes a transition between being ordered and disordered at or near the phase transition boundaries (Li and Zhang, 2010). Figure 1 illustrates this concept using simulated data from the prior distribution based on the primary analysis assuming a common external field parameter, $\rho_k = \rho$. For $\theta = 0.2$ there is a dramatic change in the expected number of the 40 AEs with a NDR when ρ is set to -1, and when it is set to 1. This change characterizes a phase transition. From the information conveyed in such a plot, our strategy is to select hyperparameters that match our a priori belief of how treatment is anticipated to collectively affect the K AEs being analyzed.

For the specification of external field parameters, one strategy is to assume a common parameter across AEs, $\rho_k = \rho$. If we consider Figure 1 coupled with the clinical plausibility that it is much more likely that a small number of the AEs, compared to a large number of AEs, would exhibit a difference in the response between treatment groups, reasonable values of ρ to consider may be between 1 and 2. Selecting $\rho \leq 0$ may be difficult to clinically justify since it is unlikely that risk imbalances exist on a large number of safety endpoints.

Another scenario is when a subset of the K AEs are expected to have a differential risk. This expectation may come from previous completed trials or known class drug effects.

One strategy is to assign the AEs that are expected to differ between groups a common external field parameter; AEs without this expectation are assigned a common but different (and larger) external field parameter. This task of selecting hyperparameters for this scenario is more complicated than the previous strategy. One approach is to contrast expectations across the K AEs assuming two distinct values of the external field with expectations assuming a common value across all AEs.

In addition to assessing the expected number AEs with a NDR, it may also be informative to evaluate the marginal prior probability that $\gamma_k = 1$, $P(\gamma_k = 1)$. Importantly, assuming $\rho_k = \rho$ does not necessarily lead to the same marginal probability across the K endpoints since probabilities may depend on the number of neighbors. In the special case where $\rho_k = 0$ for all k, it can be shown that, regardless of the choice of θ, the marginal prior probability that $\gamma_k = 1$ is 0.5 (see Smith and Fahrmeir, 2007). While the hyparameters that yield equal marginal prior odds may be difficult to justify per the strategy detailed above, they may, nonetheless, be worth considering for a sensitivity analysis.

An important consideration when selecting a fixed θ, or $\theta_{\max}$ when θ assigned a uniform prior distribution, is that a large value of $\theta_{\max}$ can lead to a homogeneous TAEHS during MCMC sampling (i.e., all elements of $\boldsymbol{\gamma}$ are either 0 or 1) (Marin and Robert, 2007). One strategy to gauge the potential degree of smoothing is to assess the influence of θ (or $\theta_{\max}$) on the conditional prior probabilities, $f(\gamma_k = 1|\gamma_j, j \neq k) \propto \exp(-\rho_k + \theta \sum_{j \in \mathrm{D}_k} I(\gamma_k = \gamma_j))^{-1}$, assuming all the neighbors exhibit a NDR (i.e., $\gamma_j = 1$ for each $j \in \mathrm{D_k}$), that is, the maximal probability. Web Figure 1 in Web Appendix A illustrates the relationship for different θ (with $\rho_k = 0$). Using the information from this Figure and paying particular attention to AEs with the most number of neighbors, we recommend not choosing θ (or $\theta_{\max}$) that leads to conditional probabilities near 1 or a change from independence that is substantial.

3.2. *Posterior Distribution*

MCMC methods are used to sample from the posterior distribution (Robert and Cassela, 2004). Due to the conjugacy of models (2) and (3), the sampling algorithm is not computationally intricate, and the event probabilities follow a Beta distribution with updated parameters. The sampling algorithm becomes more computationally involved when the smoothing parameter θ is assigned a prior distribution. The difficulty arises from the normalizing constant having to be evaluated at each iteration of the sampler, which is computationally prohibitive except for small K. Web Appendix B details a strategy for approximating the normalizing constant as well as provides the full conditionals for the other model parameters.

To gain more insight about this model, we describe the posterior inferences in terms of the canonical Bayesian hypothesis test summary, namely the Bayes Factor (Carlin and Louis, 2009). Consider the full conditional for γ_k, which is Bernoulli with probability $1/(1 + h_k)$, where

$$h_k = \frac{f(y_k|\gamma_k = 0)}{f(y_k|\gamma_k = 1)} \exp\left(-\rho_k + \theta \sum_{j \in \mathrm{D}_k} (1 - 2\gamma_j)\right) \quad (5)$$

The term h_k is the Bayes factor, where the ratio of marginal densities is the posterior-odds and exponent term represents the prior-odds. The prior-odds expression does not resemble the more familiar form, $f(\gamma_k = 0)/f(\gamma_k = 1)$, due to the allowance for dependencies between AEs; however, in the special case when $\theta = 0$, the prior odds does take this form. From this simplification, under independence, it becomes straightforward to visualize how the model adjusts for the treatment effect amongst its neighbors. Specifically, compared to independence, the probability for favoring a non-differential effect will increase if more than half of the AEs neighbors favor a non-differential effect, and decrease otherwise.

Several posterior summaries can be evaluated to assess the safety of the drug being investigated. Two of possible interest are the marginal posterior probability of a NDR between treatment arms, $f(\gamma_k = 1|y)$, and a model averaged comparative summary.

For the marginal posterior probability that $\gamma_k = 1$, AEs with small values should be flagged for being associated with treatment. The challenge lies in defining "small;" we propose a two-stage flagging scheme since clinical trial data tend to be underpowered for safety endpoints. At the first stage, AEs with posterior estimates of 0.5 or less can be flagged for possible association. At the second stage, more stringent threshold can be set for associations that are stronger. One possibility is to threshold estimates based on calibrating the marginal posterior probabilities with the Frequentist p-value. This rule, described in Smith et al. (2003), makes use of the fact that $-2\log((1 - f(\gamma_k = 0|y))/f(\gamma_k = 0|y))$ is on the same scale as a likelihood ratio test, and is approximately distributed as χ^2_1, a Chi-squared with 1 degree of freedom (Raferty, 1996). A p-value of 0.05 gives a critical value of 3.841, so that solving for the posterior probability, $f(\gamma_k = 0|y) = 0.8722$ at this critical value. Then, AEs with $f(\gamma_k = 0|y) > 0.8722$ (or $f(\gamma_k = 1|y) < 0.1278$) can be classified as having a differential risk between the treatment and comparator arms.

For the model averaged estimate, $E(g(\pi_k^T, \pi_k^C)|y)$, comparative summaries can be taken to be the log-odds ratio, log-relative risk, or risk difference, corresponding to $g(\pi_k^T, \pi_k^C)$ defined, respectively as $\log(\pi_k^T(1 - \pi_k^C)/\pi_k^C(1 - \pi_k^T))$, $\log(\pi_k^T/\pi_k^C)$, or $\pi_k^T - \pi_k^C$. Note that the expectation for these quantities can be expressed, by conditioning on γ_k, as

$$E(g(\pi_k^T, \pi_k^C)|y) = E(g(\pi_k^T, \pi_k^C)|y, \gamma_k = 0)(1 - f(\gamma_k = 1|y)) \quad (6)$$

The term associated with $E(g(\pi_k^T, \pi_k^C)|y, \gamma_k = 1)$ is omitted from (6) since it is zero, thus revealing that the estimate is shrunken to the null value 0. More shrinkage occurs for large $f(\gamma_k = 1|y)$, coinciding with a NDR between treatment groups. The strength of the association can be captured by $f(\pi_k^T > \pi_k^C|y)$, with probabilities larger than 0.85 being considered significant. This threshold was obtained by discounting the significance threshold if the data were analyzed under (3) by the decision rule for classifying differential risks. That is, to account for the fact that $\pi_k^T = \pi_k^C$ when $\gamma_k = 1$, we discount the commonly used 0.975 threshold by 0.8722, that is, the probability for classifying an AE as having a differential risk.

Table 1
Clinical trial data from Mehrotra and Heyse (2004).

AE	Body system	AE name	Treatment, $N^T = 148$	Control, $N^C = 132$	Fisher's exact p
			y_k^T (%)	y_k^C (%)	
1	1	Astenia/fever	57 (38.5)	40 (30.3)	0.167
2	1	Fever	34 (23.0)	26 (19.7)	0.561
3	1	Infection, fungal	2 (1.4)	0 (0.0)	0.5
4	1	Infection, viral	3 (2.0)	1 (0.8)	0.625
5	1	Malaise	27 (18.2)	20 (15.2)	0.525
6	2	Anorexia	7 (4.7)	2 (1.5)	0.179
7	2	Cendisiasis, oral	2 (1.4)	0 (0.0)	0.5
8	2	Constipation	2 (1.4)	0 (0.0)	0.5
9	2	Diarrhea	24 (16.2)	10 (7.6)	0.029
10	2	Gastroenteritis	3 (2.0)	1 (0.8)	0.625
11	2	Nausea	2 (1.4)	7 (5.3)	0.089
12	2	Vomiting	19 (12.8)	19 (14.4)	0.73
13	3	Lymphadenopathy	3 (2.0)	2 (1.5)	1
14	4	Dehydration	0 (0.0)	2 (1.5)	0.221
15	5	Crying	2 (1.4)	0 (0.0)	0.5
16	5	Insomnia	2 (1.4)	2 (1.5)	1
17	5	Irritability	75 (50.7)	43 (32.6)	0.002
18	6	Bronchitis	4 (2.7)	1 (0.8)	0.375
19	6	Congestion, nasal	4 (2.7)	2 (1.5)	0.687
20	6	Congestion, respiratory	1 (0.7)	2 (1.5)	0.603
21	6	Cough	13 (8.8)	8 (6.1)	0.497
22	6	Infection, upper respiratory	28 (18.9)	20 (15.2)	0.431
23	6	Laryngotracheobronchitis	2 (1.4)	1 (0.8)	1
24	6	Pharyngitis	13 (8.8)	8 (6.1)	0.497
25	6	Rhinorrhea	15 (10.1)	14 (10.6)	1
26	6	Sinusitis	3 (2.0)	1 (0.8)	0.625
27	6	Tonsillitis	2 (1.4)	1 (0.8)	1
28	6	Wheezing	3 (2.0)	1 (0.8)	0.625
29	7	Bite/sting	4 (2.7)	0 (0.0)	0.125
30	7	Eczema	2 (1.4)	0 (0.0)	0.5
31	7	Pruritis	2 (1.4)	1 (0.8)	1
32	7	Rash	13 (8.8)	3 (2.3)	0.021
33	7	Rash, diaper	6 (4.1)	2 (1.5)	0.288
34	7	Rash, measles/rubella-like	8 (5.4)	1 (0.8)	0.039
35	7	Rash, varicella-like	4 (2.7)	2 (1.5)	0.687
36	7	Urticaria	0 (0.0)	2 (1.5)	0.221
37	7	Viral exanthema	1 (0.7)	2 (1.5)	0.603
38	8	Conjunctivitis	0 (0.0)	2 (1.5)	0.221
39	8	Otorrhea	18 (12.2)	14 (10.6)	0.711
40	8	Otitis media	2 (1.4)	1 (0.8)	1

It is worth noting that the direction of the shrinkage in (6) is different from the population average shrinkage in the hierarchical modeling approach. We consider this null-shrinkage reasonable for two reasons. First, the population average makes the effect of less frequent (and possibly severe or serious) AEs sensitive to the effect of the more frequent (and possibly less severe) AEs. The clinical appropriateness or implication of this in practice is not clear. Second, in a study that is likely to be under powered for the endpoints being analyzed, null-shrinkage is conservative.

4. Application

We apply the methods developed above to vaccine trial data described in Mehrotra and Heyse (2004; henceforth, referred to as MH). Refer to MH or BB for study design details. Table 1 lists information from 40 AEs classified according to one of eight body systems for 148 and 132 toddlers randomized to the treatment and control arms, respectively. Four AEs identified to have statistically significant differences based on Fisher's exact test were diarrhea (p-value = 0.029), irritability (p-value = 0.003), rash (p-value = 0.021), and measles/rubella-like rash (p-value = 0.039).

For all analyses the Beta distribution hyperparameters were set as follows: $\alpha_k = \alpha_k^i = 0.25$, and $\beta_k = \beta_k^i = 0.75$ for all k, $k = 1,\ldots,K$, and i, $i = T,C$. We used this prior, as opposed to Jefferys or a uniform prior, since the likelihood of experiencing an AE tends to be small. For the Ising prior, we assumed a constant external field parameter, that is, $\rho_k = \rho$. Also, because results were qualitatively similar when the smoothing

Table 2
Posterior summaries of the primary analysis.

AE #	Body system	AE name	$P(\gamma_k = 1\|y)$	$P(\pi_k^T > \pi_k^C\|y)$	$E(\log(OR)\|y)$
1	1	Astenia/fever	0.961	0.035	0.016
2	1	Fever	0.976	0.016	0.002
3	1	Infection, fungal	0.845	0.148	0.736
4	1	Infection, viral	0.947	0.043	0.062
5	1	Malaise	0.975	0.019	0.005
6	2	Anorexia	0.917	0.078	0.107
7	2	Cendisiasis, oral	0.878	0.117	0.584
8	2	Constipation	0.877	0.117	0.574
9	2	Diarrhea	0.832	0.165	0.148
10	2	Gastroenteritis	0.953	0.038	0.045
11	2	Nausea	0.866	0.005	-0.198
12	2	Vomiting	0.984	0.006	-0.008
13	3	Lymphadenopathy	0.938	0.041	0.009
14	4	Dehydration	0.681	0.010	-1.526
15	5	Crying	0.742	0.247	1.201
16	5	Insomnia	0.934	0.032	-0.011
17	5	Irritability	0.294	0.706	0.534
18	6	Bronchitis	0.973	0.025	0.041
19	6	Congestion, nasal	0.989	0.008	0.010
20	6	Congestion, respiratory	0.988	0.003	-0.016
21	6	Cough	0.990	0.009	0.011
22	6	Infection, upper respiratory	0.991	0.007	0.004
23	6	Laryngotracheobronchitis	0.989	0.009	0.009
24	6	Pharyngitis	0.989	0.009	0.005
25	6	Rhinorrhea	0.993	0.004	0.001
26	6	Sinusitis	0.983	0.014	0.027
27	6	Tonsillitis	0.988	0.009	0.002
28	6	Wheezing	0.983	0.013	0.023
29	7	Bite/sting	0.710	0.286	1.572
30	7	Eczema	0.890	0.105	0.522
31	7	Pruritis	0.966	0.025	0.039
32	7	Rash	0.744	0.255	0.378
33	7	Rash, diaper	0.950	0.046	0.042
34	7	Rash, measles/rubella-like	0.739	0.259	0.578
35	7	Rash, varicella-like	0.970	0.024	0.028
36	7	Urticaria	0.864	0.005	-0.665
37	7	Viral exanthema	0.967	0.007	-0.027
38	8	Conjunctivitis	0.760	0.008	-1.182
39	8	Otorrhea	0.967	0.023	0.000
40	8	Otitis media	0.939	0.043	0.025

parameter, θ, was fixed at $\theta_{\max}$ or assigned a $U(0, \theta_{\max})$ prior distribution, we only present results for the case where θ is fixed. Specific values of the hyperparameters are detailed below. For the MCMC sampling, we took 10,000 draws from the posterior distribution after a burn-in of 50,000 and having the chain thinned by 50. Data and program code are available with the Supplementary Web Material.

4.1. *Results*

This primary analysis assumes AEs within the same body system are neighbors; hence, information from AEs in different body systems is not shared. The Ising prior hyperparameters were set as follows: $\rho = 1$ and $\theta = 0.2$; these values coincide with a priori belief that 34.5 out of the 40 AEs will exhibit a non-differential effect, which we did not considered unreasonable.

Table 2 displays the AE specific results. None of the 40 AEs had an estimate of $f(\gamma_k = 1|y)$ below the proposed 0.1278 threshold for declaring a differential risk between groups. The estimate of $f(\gamma_k = 1|y)$ is henceforth referred to as the estimate of a NDR. Irritability (AE #17) had the smallest NDR estimate, 0.294, and was the only estimate below 0.5. The other 39 AEs, including the three AEs with statistically significant p values, had a NDR estimate > 0.68, thus suggesting that treatment had no association with these events.

To evaluate the magnitude and strength of the association, we also evaluated the model averaged log-odds ratio (logOR). Among the AEs with at least one event in both treatment arms, measles/rubella-like rash (AE #34) had the largest estimate (logOR $= 0.58$) followed by irritability (logOR $= 0.53$). Irritability, however, had a stronger association with treatment than measles/rubella-like rash as measured

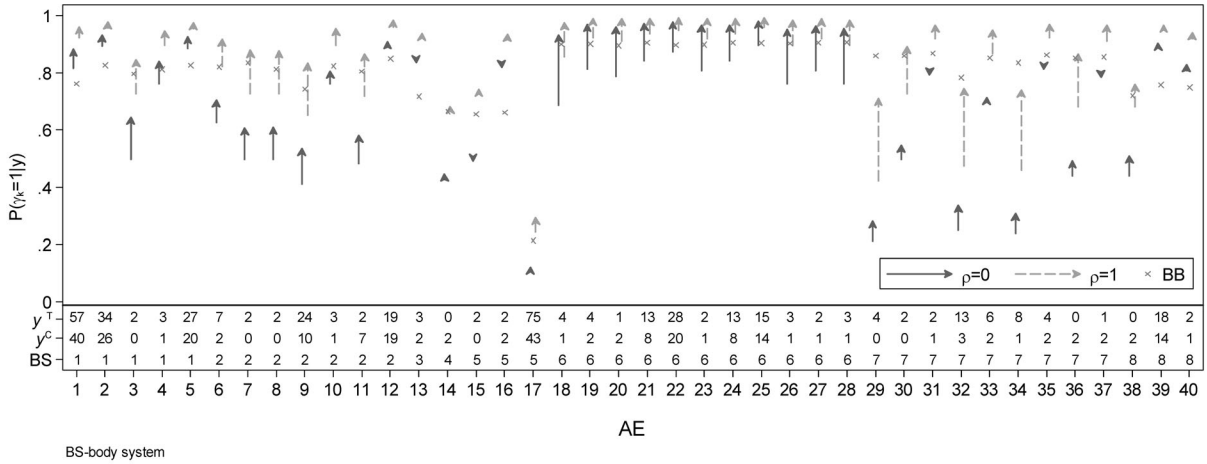

Figure 2. Posterior probability of non-differential risk for MH data listed in Table 1 for the proposed method and BB. The tail of the line segment corresponds to the posterior estimate assuming AEs are independent. The arrow is the posterior estimate after imposing the model's respective structure to the AEs. The estimate for BB is the log-odds ratio equals 0.

by the posterior probability that $\pi_k^T > \pi_k^C$ (0.706 vs. 0.259). For the other AEs, with the possible exception of rash (AE # 32), the model averaged estimates clustered around zero.

Irritability is the only AE from this analysis with evidence that appears to be associated with treatment. This association, while compelling, is not considered strong, given that the estimate of $f(\gamma_k = 1|y)$ is not below the 0.1278 threshold or the posterior probability that $\pi_k^T > \pi_k^C$ is below 0.85.

4.2. *Prior Sensitivity Analysis and BB*

We investigated the sensitivity of the above results to different Ising prior hyperparameters values: $\rho = 0{,}1$ and $\theta = 0$, 0.2. Results are illustrated in Figure 2 along with results from BB. The tail of the two line segments, one for $\rho = 0$ (solid line) and the other for $\rho = 1$ (dashed line), corresponds to the NDR estimate when $\theta = 0$, with the head of the arrow being the estimate when $\theta = 0.2$. The BB estimate displayed is the posterior estimate that the log-odds ratio is exactly 0, corresponding to the risks being equal between groups. We do not consider it appropriate to present results for the unadjusted model presented in BB, that is, the "Solo" model. Our reasoning is BB does not reduce to the unadjusted approach by modifying one structural element of the model, like setting $\theta = 0$ in the proposed approach.

For $\rho = 0$ (and with $\theta = 0$ or 0.2), irritability (AE #17) is the only AE with an NDR estimate below the 0.1278 threshold. The model averaged log(OR) estimate for irritability (with $\theta = 0.2$) was 0.66, with a posterior probability that $\pi_k^T > \pi_k^C$ of 0.88, which exceeds the 0.85 threshold for a strong association. This finding highlights a general dependency of results on the ρ. Larger NDR estimates occurred for $\rho = 1$ compared to $\rho = 0$, which is not surprising since this parameter controls the sparsity of the TAEHS.

Several AEs within body system 7 (AE #29 to AE #38) may be identified as possible signals when $\rho = 0$. In particular, AEs 29, 32, and 34 had an estimated NDR above the 0.1278 threshold but notably below 0.5, suggesting some preference for a differential effect. Two of these AEs had significant p values (AEs 32, 34), while the AE bite/sting (AE #29) is noteworthy because of the rule of four argument (Crowe et al., 2009). Note that the fourth AE with a significant p-value (diarrhea, AE #9) did not stand out in theses analyses.

Results in BB were generally aligned with results from our primary analysis. Both approaches identified irritability (AE #17) as being most likely having a differential effect, and the other AEs strongly favoring a non-differential effect.

Also, apparent from the plot is that there was not much variability in the BB estimates except for AEs 13–17 (body systems 3, 4, and 5). Two of these AEs, lymphadenopathy (AE #13) and dehydration (AE #14), are interesting as they highlight underlying structural differences in modeling approaches. In the proposed approach, these AEs had no neighbors and, therefore, no information to share between AEs; hence, the NDR estimates when either $\theta = 0$ or 0.2 were identical. In BB, theses AEs are strictly influenced by AEs in different body systems through the estimate of the variance of the distribution of the overall mean across body systems. While this difference, although subtle, is important to consider when interpreting adjusted results.

4.3. *Multiple Body Systems*

The sensitivity of results to the different neighbor structures is investigated on a subset of 20 AEs (Table 3). The first structure, referred to as N1, uses the same strategy in the above investigations. The second structure, referred to as N2, is based on the MedDRA classification. In N2 two AEs are considered neighbors if the have at least one SOC in common. The four MedDRA SOCs used to describe the AEs are gastrointestinal disorders (A), respiratory, thoracic, and mediastinal disorders (B), ear, and labyrinth disorder (C), and infections and infestation (D). Importantly, the relationships assumed by the N2 algorithm were not clinically evaluated as this analysis is intended to convey attributes of the methodology.

In total, seven AEs can be described by multiple SOCs; all of these AEs are described by the SOC infections and infestation. The relationship among AEs implied by structure N1 and N2 is illustrated in Figure 3, with the (i, j) element

Table 3
NDR estimates for a subset of 20 AEs with assuming different relationships between AEs (N1, N2) using unaltered and altered data.

			Unaltered data			Altered data		
AE	MH	SOC	$\theta = 0$	N1	N2	$\theta = 0$	N1	N2
6	2	A	0.628	0.706	0.781	0.628	0.573	0.673
7	2	A	0.496	0.616	0.689	0.496	0.446	0.564
8	2	A	0.496	0.601	0.696	0.496	0.457	0.556
9	2	A	0.410	0.530	0.614	0.013	0.013	0.024
10	2	A, D	0.762	0.809	0.938	0.218	0.210	0.478
11	2	A	0.483	0.588	0.693	0.483	0.440	0.538
12	2	A	0.897	0.913	0.940	0.897	0.856	0.898
18	6	B, D	0.687	0.929	0.951	0.687	0.933	0.941
19	6	B	0.815	0.965	0.964	0.815	0.967	0.965
20	6	B	0.789	0.956	0.962	0.789	0.956	0.960
21	6	B	0.843	0.971	0.971	0.843	0.971	0.970
22	6	B	0.875	0.977	0.978	0.875	0.977	0.977
23	6	B, D	0.808	0.962	0.973	0.808	0.960	0.968
24	6	A, B, D	0.843	0.970	0.987	0.843	0.969	0.975
25	6	B	0.896	0.982	0.979	0.896	0.976	0.979
26	6	B, D	0.762	0.949	0.967	0.762	0.947	0.957
27	6	B, D	0.808	0.962	0.975	0.808	0.961	0.966
28	6	B	0.762	0.949	0.950	0.762	0.953	0.952
39	8	C	0.891	0.899	0.907	0.188	0.204	0.215
40	8	C, D	0.808	0.837	0.937	0.808	0.795	0.902

The altered data modified event counts for diarrhea (AE#9), gastroenteritis (AE #10), and othorrhea (AE #39).
MH-classification of AEs from MH.
SOCs: A, gastrointestinal disorders; B, respiratory, thoracic, and mediastinal disorders; C, ear and labyrinth disorder; D, infections and infestations MH-classification of AEs from MH.

of the matrix black if the AE $_i$ is a neighbor with AE $_j$, and not otherwise. The algorithm for constructing an adjacency matrix is detailed in Web Appendix C.

The more complicated relationships between AEs in N2 is illustrated by pharyngitis (AE #24, row 14 in plots). Pharyngitis is a neighbor with all AEs except otorrhea (AE #39, row 19); these two AEs would be neighbors if either pharyngitits could be also described by SOC C, or otorrhea could be described by either SOC A, B, or D. Another important way the two structures differ is the neighbors of the neighbors. A good example of this is otorrhea. Under N1 and N2, otorrhea has the same and only one neighbor, otitis media (AE #40, row 20). Otitis media, however, under N1 is only neighbors with otorrhea; under N2, otitis media is a neighbor with otorrhea

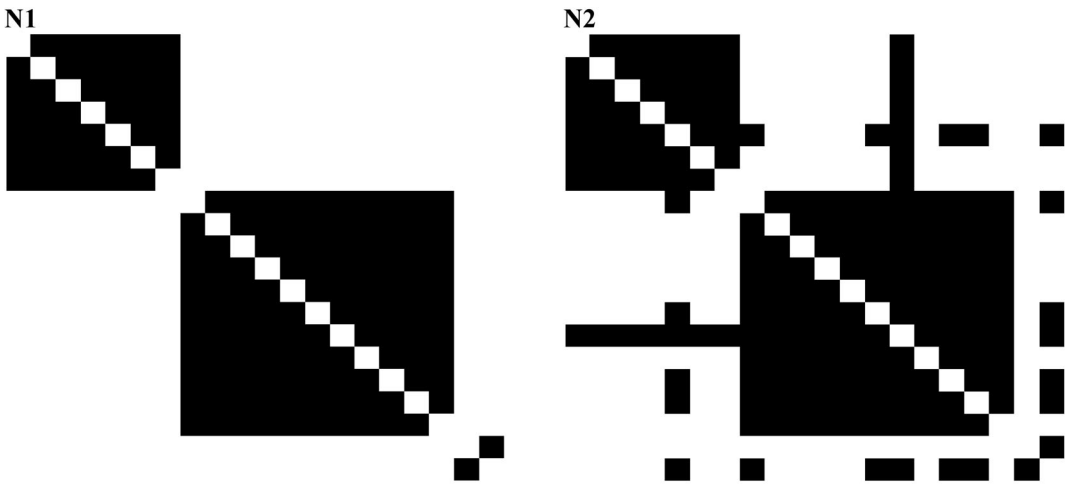

Figure 3. Assumed neighborhood structures in subset of AEs. The left panel (N1) portrays relationships per MH. The right panel (N2) allows two AEs to be neighbors if the have an SOC in common. The i, j component of the plot is black if AE i and j are assumed neighbors, and not otherwise.

and 6 other AEs (i.e., those that can be described by SOC D). The influence of otitis media on otorrhea will therefore depend, in part, on the neighbors of otitis media; thus, in N2, there will be a (indirect) dependence between AEs that are not neighbors.

In this investigation, we set the Ising prior hyperparameters as follows: $\theta = 0, 0.2$, and $\rho = 0$. Results from the analyses are given in Table 3 (under unaltered data). Under independence all of the NDR estimates were about 0.5 or larger. Based on the magnitude of the estimates and how information is shared between AEs, it is not surprising that NDR estimates are larger for N1 compared to independence ($\theta = 0$), and larger for N2 compared to N1 given the additional neighbors in N2.

Cough (AE #21) and pharyngitis (AE #24) illustrate the impact of different neighborhood structures. Both these AEs are described by body system 6 in MH and have 13 and 8 events in the treatment and control arms, respectively. Under N1 the estimates for the two AEs should be identical; any difference is attributable to simulation error. Under N2 the posterior estimates for the two AEs are different, with the NDR estimate slightly larger for pharyngitis (0.987 vs. 0.971 for cough), which can be attributed to it having more neighbors.

To further illustrate the potential impact of N1 and N2 we analyzed a modified dataset by altering event counts for diarrhea (AE #9), gastroenteritis (AE #10), and otorrhea (AE #39). The modified counts (treatment, control) were set as follows: diarrhea (24, 5), gastroenteritis (11, 2), and otorrhea (18, 5). Results from this analysis are displayed in Table 3 (under altered data). Under independence, the association between treatment and diarrhea was strong (NDR = 0.013), and less than 0.5 (but > 0.1278) for both gastroenteritis and otorrhea.

There are two noteworthy findings from this analysis. First, AEs classified in body system 2 in MH that did not have their event counts manipulated had, compared to independence, different trends in their NDR estimates. Under N2, the NDR estimates were larger compared to assuming independence, but smaller under N1. For instance, the NDR estimate for nausea (AE #11) decreased from 0.483 under independence to 0.446 for N1, but increased to 0.543 for N2. The other noteworthy finding is that the influence of otorrhea's neighbors was not substantial. While a difference in the NDR estimate for otorrhea was expected under the different structures, the difference in NDR estimates under N1 (0.204) and N2 (0.215) was not large. This difference, although small, can be attributed to the larger NDR estimate for otitis media under N2 (0.902) resulting from its additional neighbors than under N1 (0.795). For the other AEs with the same neighbors under N1 and N2 but with different neighbors of neighbors (i.e., AEs 19, 20, 21, 22, 25, and 28), the difference in NDR estimates under N1 and N2 were also very small.

5. Simulation Study

We present results from a small simulation study to illustrate performance attributes of the proposed methodology. Also presented are results from two BB-type models to allow for a comparative assessment. The separate models are evaluated under data generated as follows. Events for $K = 5$ endpoints were independently simulated from a Binomial distribution with $N^T = N^C$ set to either 150 or 500. Two configurations of event probabilities were considered; one configuration, referred to as the null scenario, set $\pi_k^i = 0.05$ for $i = T, C$ and $k = 1, \ldots, K$; the other configuration, referred to as the signal scenario, was similar to null scenario except π_1^T was set to 0.1.

The first BB-type model considered is based on formulating the BB model to data for a single body system. We therefore refer to this model as the 2-level BB. Under this model, event probabilities are modeled on the log-odds scale, with $\text{logit}(\pi_k^C) = \lambda_{0k}$ and $\lambda_{1k} = \text{logit}(\pi_k^T) - \lambda_{0k}$. We assume independent Normal priors for λ_{0k}, $\lambda_{0k} \sim N(-2, 10^2)$ and a mixture hierarchical Normal prior for λ_{1k} given by $\lambda_{1k} \sim \varphi \text{I}_{[0]} + (1-\varphi)N(\mu_1, \tau^2)$ with $\mu_1 \sim N(0, 5^2)$, $\tau \sim U(0, 5)$ and $\varphi \sim \textit{Bernoulli}(0.5)$. The second BB-type model is the so-called Solo Bayesian model from BB. This model is similar to the previous model except there is no hierarchical distribution for the Normal component of the mixture prior; specifically, $\lambda_{1k} \sim \varphi \text{I}_{[0]} + (1-\varphi)N(0, 5^2)$.

We consider two separate formulations of the Ising prior to allow for reasonable comparisons between two BB-type models. Coinciding with the independence of the Solo Bayesian model, the first Ising prior formulation set $\theta = 0$. To allow for dependence between endpoints, thus being "comparable" to the 2-level BB, the second Ising prior formulation assumes $\theta \sim U(0, 0.6)$ with the relationship between the K endpoints characterized by a complete graph, meaning an endpoint is a neighbor with the other four endpoints. Both Ising prior formulations set $\rho_k = \rho = 0$, leading to $P(\gamma_k = 1) = 0.5$, and fixed hyperparameters of the Beta distribution as follows: $\alpha_k = \alpha_k^i = 0.25$, and $\beta_k = \beta_k^i = 0.75$ for all k, $k = 1, \ldots, K$, and i, $i = T, C$. A total of 1,000 simulated dataset were constructed for each simulation configuration. The MCMC sampling scheme for each simulated dataset took 5,000 draws from the posterior distribution after a burn-in of 10,000 and having the chain thinned by 10.

For each simulation configuration, the performance characteristics assessed are the mean squared error (MSE) for the posterior estimate of the log-odds ratio, and the median of the distribution of NDR estimates from the 1,000 simulated datasets. For the BB-type models the NDR correspond to the posterior estimate that the log-odds ratio is exactly 0. Use of the median for describing central tendency, compared to the mean, is favored since the distribution of posterior estimates tended to be skewed, with mass tending to be concentrated near 0 or 1. Lastly, we also present the likelihood that the NDR estimate is below 1/3 for purposing of evaluating a decision rule; we did not use the threshold 0.1278 due difficultly assessing trends. For the null scenario and signal scenario for $k \neq 1$, the decision rule provides an estimate of type-I error; for $k = 1$ in the signal scenario, applying the decision rule provides an estimate of the power. Simulation results are averaged across the K endpoints for the null scenario; for the signal scenario, simulation results are averaged all endpoints, where appropriate, and presented separately for $k = 1$ and $k \neq 1$.

Simulation results are provided in Table 4. For the null scenario, MSEs were reasonably similar for the Ising prior with $\theta = 0$ and Solo Bayesian model, as expected. When information was allowed to be shared across endpoints, the MSE for the Ising approach decreased while the MSE for the 2-level

Table 4
Simulation study results.

Statistic	N^T	Model	Null	Signal		
			Overall	Overall	$k \neq 1$	$k = 1$
MSE	150	Ising: $\theta = 0$	0.135	0.187	0.132	0.409
		Ising: $\theta \sim$ U(0,0.6)	0.094	0.169	0.098	0.453
		Solo Bayesian	0.125	0.180	0.121	0.415
		2-Level BB	0.122	0.169	0.119	0.370
	500	Ising: $\theta = 0$	0.013	0.040	0.013	0.147
		Ising: $\theta \sim$ U(0,0.6)	0.007	0.046	0.010	0.192
		Solo Bayesian	0.012	0.041	0.011	0.162
		2-Level BB	0.013	0.043	0.017	0.145
Median-NDR	150	Ising: $\theta = 0$	0.853	0.842	0.852	0.682
		Ising: $\theta \sim$ U(0,0.6)	0.930	0.906	0.911	0.824
		Solo Bayesian	0.876	0.868	0.876	0.722
		2-Level BB	0.806	0.788	0.798	0.635
	500	Ising: $\theta = 0$	0.915	0.902	0.915	0.115
		Ising: $\theta \sim$ U(0,0.6)	0.969	0.935	0.942	0.238
		Solo Bayesian	0.932	0.922	0.932	0.170
		2-Level BB	0.884	0.838	0.857	0.115
P(NDR < 1/3)	150	Ising: $\theta = 0$	0.027	—	0.026	0.241
		Ising: $\theta \sim$ U(0,0.6)	0.014	—	0.014	0.165
		Solo Bayesian	0.021	—	0.020	0.203
		2-Level BB	0.031	—	0.032	0.241
	500	Ising: $\theta = 0$	0.012	—	0.012	0.698
		Ising: $\theta \sim$ U(0,0.6)	0.004	—	0.007	0.561
		Solo Bayesian	0.009	—	0.008	0.639
		2-Level BB	0.013	—	0.020	0.699

BB model did not change. Systematic differences between approaches were also observed for the median NDR. With respect to the independence model, the Ising approach had larger median values under dependence, while the 2-level BB had smaller median values. Because of the larger NDR values, it is not surprising, when applying the decision rule, that the Ising prior with $\theta \sim U(0, 0.6)$ results in a lower type-I error rate compared to the 2-level BB model.

For the signal scenario there appeared to be systematic differences between approaches for the different performance metrics. The Ising approach had, for endpoints without a signal ($k \neq 1$), smaller MSE when the prior allowed for a relationship between the endpoints, but an increase in MSE for signal endpoint ($k = 1$). Conversely, for the two BB-type models there was not much difference in MSE for the null endpoints ($k \neq 1$), but a smaller MSE for the 2-level BB for the signal endpoint ($k = 1$). For the median NDR, the Ising approach had a larger value for both the signal endpoint ($k = 1$) and null endpoints ($k \neq 1$) for the prior that allowed dependencies between endpoints. For the BB-type models the trend was reversed, with median values being smaller in the 2-level BB for both the null endpoints and the signal endpoint. It is therefore not surprising that, based on the decision rule, the Ising approach is more conservative than 2-level BB model for detecting a signal.

6. Discussion

We presented a flexible Bayesian approach for analyzing drug safety data by applying an extension of hypothesis evaluations to multiple related endpoints. The primary advantage of the proposed method is in its ability to borrow information from medically related AEs, where the assumed relationships are not influenced by the structural rigidness of the underlying statistical model. Because of the model's ability to accommodate complex features of drug safety data, it is expected that the clinicians and statisticians, alike, would find the methodology and ensuing inferences attractive.

This methodology can be applied to wide-array of scenarios one may encounter in drug safety evaluations, ranging from investigations of an entire safety database to a few selected AEs that may be of clinical interest. Another attractive feature about the methodology is its simplicity. Subject matter experts without technical expertise should be able to fully comprehend why the estimate of $f(\gamma_k = 1|y)$ increased or decreased in magnitude (compared to independence) based on assessing the effect of treatment on the risk differentials for neighboring AEs. This is in contrast to the hierarchical modeling approach that has adjustments resulting from convoluted dependencies at the different model levels.

To realize the full potential of the proposed methodology requires close collaboration between statisticians and clinicians. Statisticians lack the clinical expertise needed to justify why two AEs may or may not have biological features in common, which warrants them being treated as "neighbors." The need for such clinical justification was illustrated with the modified data, where the relative magnitude (compared to independence scenario) of the estimate of $f(\gamma_k = 1|y)$ depended on assumed neighborhood structure. Because of the potential for such divergence, it is advisable to pre-specify the neighborhood structure.

This methodology should be used to complement, not replace other analytical approaches. In practice, it is likely that the different approaches will yield qualitatively similar conclusions about risk, despite their reliance on different assumptions. Understanding differences in methods and their ensuing impact is particularly important when differences do occur. Our approach and the ones referenced in this article make different and rely on (largely) untestable assumptions about the relationship between endpoints. While these statistical nuisances are important, it is important to keep in mind a thorough safety evaluation does not conclude with the identification of signals. A thorough evaluation includes assessing case report forms, concomitant medication use, and other important considerations, such as other AEs experienced. The latter point is worth highlighting since our approach and others do not exploit different AEs occurring within the same subject.

That said, we believe the general comparison of the proposed methodology with other methods, including the nested hierarchical approach, misses the scope of the proposed methodology. The motivation for our methodology is based on being able to make inferences that accounts for multidimensional relationships between AEs based on clinical and biological considerations. Since the other approaches are deficient in this regard, as described in Section 1, comparison of methods is not relevant. This is not to say, however, that there is no common ground between our approach and others. One place where common ground exists was explored in the simulation study, where the set-up had the inference for each endpoint influenced by the others endpoints. In the simulation study, we demonstrated the proposed methodology is conservative and does protect against type-I errors for the scenarios investigated. The approach's favorable performance overall and compared to the 2-level BB model in the null scenario is not surprising based on the Ising prior's ability to encourage the clustering of liked value binary variables (Higdon, 1993). Additional simulations are needed to establish the proposed methodology in different settings.

Due to the novelty of this methodology to clinical safety data, numerous extensions can be made. One possible extension is to account for the relatedness of neighboring AEs. For two AEs considered biologically related, say AE k and AE j, the relationship between them can be quantified by assigning their relationship a weight w_{kj}, $w_{kj} > 0$, where a larger weight corresponds to them being more related. The weights can then be incorporated in (4) by replacing the term $\sum_{j\in \mathrm{D}_k} I(\gamma_k = \gamma_j)$ with $\sum_{j\in \mathrm{D}_k} w_{jk} I(\gamma_k = \gamma_j)$. One strategy to specify weights is to consider a small number of weights, say 1/3, 2/3, or 1 corresponding to low, moderate, and high degree of relatedness, respectively.

Other extensions can be made by adopting a regression framework and performing variable selection on the covariate associated with treatment. In this set-up the methodology could easily be extended to data collected from multiple trials or for the purpose of vulnerable patient sub-populations. This latter application can be considered an alternative approach to multivariate Bayesian logistic regression proposed by DuMouchel (2012). In either set-up, the method can be extended to trials with more than two trial arms. Lastly, one can assume a Poisson likelihood if either one is interested in the total number of events instead the number of subjects that experienced an event, or to account for potential differences in follow-up time.

7. Supplementary Materials

Web Appendices and Figures referenced in Sections 3.1, 3.2, and 4.4, and program code referenced in Section 4 is available with this paper at the Biometrics website on Wiley Online Library.

Acknowledgements

The authors would like to thank Bob O'Neill, the Associate Editor and Referees for their invaluable comments. This article reflects the views of the authors and should not be construed to represent FDA's views or policies.

References

Berry, S. M. and Berry, D. A. (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., and Muller, P. (2010). *Bayesian Adaptive Methods for Clinical Trials*. Boca Raton: Chapman and Hall.

Carlin, B. P. and Louis, T. A. (2009). *Bayesian Methods for Data Analysis, 3 rd edition*. Boca Raton: Chapman and Hall.

Crowe, B. J., Xia, H. A., Berlin, J. A., Watson, D. J., Shi, H., Lin, S. L., Kuebler, J., Schriver, R. C., Santanello, N. C., Rochester, G., Porter, J. B., Oster, M., Mehrotra, D. V., Li, Z., King, E. C., Harpur, E. S., and Hall, D. B. (2009). Recommendations for the safety planning, data collection, evaluation and reporting during drug, biologic and vaccine development: A report of the safety planning, evaluation and reporting team. *Clinical Trials*, **6**, 430–440.

DuMouchel, W. (2012). Multivariate Bayesian logistic regression for analysis of clinical safety studies. *Statistical Science*, **27**, 319–339.

FDA, 08/14/2007, "Manufacturers of Some Diabetes Drugs to Strengthen Warning on Heart Failure Risk." Accessed on 4/23/2012 www.fda.gov/NewsEvents/Newsroom/PressAnnouncements/2007/ucm108966.htm.

GeorgeMcCulloch:1993 George, E. and McCulloch, R. (1993). Variable selection via Gibbs sampling. *Journal of the American Statistical Association* **88**, 881–889.

Higdon, D. M. (1993). Contribution: Spatial statistics and Bayesian computation (with discussion), by J. Besag and P. J. Green. *Journal of the Royal Statistical Society, Series B*, **55**, 25–37.

Ising, E. (1925). Beitrag zur theorie des ferromagnetismus. *Zeitschrift fur Physik* **31**, 253–258.

Li, F. and Zhang, N. R. (2010). Bayesian Variable Selection in Structured High-dimensional covariate spaces with application in genomics. *Journal of the American Statistical Association* **105**, 1202–1214.

Liu, J. S. (2004). *Monte Carlo Strategies in Scientific Computing*. New York: Springer.

Marin, J. M. and Robert, C. P. (2007). *Bayesian Core: A Practical Approach to Computational Bayesian Statistics*. New York: Springer.

Mehrotra, D. V. and Heyse, J. F., (2004). Use of the false discovery rate for evaluating clincal safety data. *Statistical Methods in Medical Research* **13**, 227–238.

Mozzicato, P. (2009). MedDRA: An overview of the medical dictionary for regulatory activities. *Pharmaceutical Medicine*, **23**, 65–75.

Raferty, A. E. (1996). Hypothesis testing and model selection. In *Markov Chain Monte Carlo in Practice*, W. R. Gilks, S. Richardson, and D. J. Spiegelhalter(eds), 163–187. London: Chapman and Hall.

Robert, C. P. and Casella, G. (2004). *Monte Carlo Statistical Methods*, 2nd edition. New York: Springer.

Rosenkranz, G. K. (2010). An Approach to the integrated safety analyses from clinical studies. *Drug Information Journal* **44**, 649–657.

Smith, M. and Fahrmeir, L. (2007). Spatial Bayesian variable selection with application to functional magnetic resonance imaging. *Journal of the American Statistical Association* **102**, 417–431.

Smith, M., Putz, B., Auer, D., and Fahrmeir, L. (2003). Assessing brain activity through spatial Bayesian variable selection. *NeuroImage* **20**, 802–815.

Xia, H. A., Ma, H., and Carlin, B. P., (2011). Bayesian hierarchical modeling for detecting safety signals in clinical trials. *Journal of Biopharmaceutical Statistics* **21**, 1006–1029.

Received May 2012. Revised March 2013. Accepted April 2013.