The Annals of Applied Statistics
2017, Vol. 11, No. 2, 504–526
DOI: 10.1214/16-AOAS974

RANDOM EFFECTS MODELS FOR IDENTIFYING THE MOST HARMFUL MEDICATION ERRORS IN A LARGE, VOLUNTARY REPORTING DATABASE

BY SERGIO VENTURINI*, JESSICA M. FRANKLIN†,‡, LAURA MORLOCK§ AND FRANCESCA DOMINICI¶,[1]

Università Commerciale Luigi Bocconi, Brigham and Women's Hospital†, Harvard Medical School‡, Johns Hopkins University§ and Harvard TH Chan School of Public Health¶*

Medical errors are a major source of preventable morbidity, mortality and healthcare costs. Voluntary reporting systems are useful data sources that collect detailed information on the circumstances of medical errors occurring in hospitals. Identifying the characteristics of errors that frequently result in patient harm when they occur would allow investigators to prioritize among the many sources of potential errors and design targeted prevention strategies. In this paper, we use data from MEDMARX, a large anonymous and voluntary reporting system for medication errors, to identify the combinations of error characteristics that are more likely to result in harm. To this end, we consider a Bayesian hierarchical model with crossed random effects and a flexible specification of the random effects distribution. We then provide a ranking of the errors using optimal Bayesian ranking based on their probability of harm. The use of optimal Bayesian ranking accounts for the varying amount of uncertainty across the random effects estimates. Finally, we examine the sensitivity of results to different specifications of the random effects distributions. The utility of flexible random effects assumptions is illustrated by empirically comparing results under several choices. We found that errors caused by mistakes in reconciling a patient's current medication list with the medications prescribed at hospital discharge have an estimated 10.5% probability of harm. These errors had the highest rate of harm of errors that occur during the prescribing stage of medication use. In addition, we found that the results are sensitive to the random effects distribution used in estimation. Thus, an approach that explores this sensitivity is important for accurately comparing the relative harm across errors.

Received October 2015; revised August 2016.

[1]Supported by Award Number P01 CA134294 (Statistical Informatics for Cancer Research) from NIH/NCI, Award Number K18 HS021991 (A Translational Framework for Methodological Rigor to Improve Patient Centered Outcomes in End of Life Cancer Research) from AHRQ, Award Number R01 GM111339 (Bayesian Methods for Comparative Effectiveness Research with Observational Data) from NIH and Award Number R35 CA197449 (Statistical Methods for Analysis of Massive Genetic and Genomic Data in Cancer Research) from NIH/NCI.

Key words and phrases. Bayesian hierarchical model, empirical Bayes, data mining, spontaneous reporting.

1. Introduction. Medical errors are a major source of preventable morbidity, mortality and healthcare costs [Brennan et al. (1991), Leape et al. (1991)]. Globally, it is estimated that 142,000 people died in 2013 from adverse effects of medical treatment up from 94,000 in 1990 [GBD 2013 Mortality and Causes of Death Collaborators (2015)]. In 2000, the National Academy of Medicine (formerly known as the Institute of Medicine) estimated that each year in U.S. hospitals 44,000 to 98,000 deaths and 1,000,000 excess injuries may be attributed to medical errors [Kohn, Corrigan and Donaldson (2000)]. Past efforts to reduce medical errors within the hospital have primarily relied on internal investigations of the causes of errors that have resulted in serious harm, such as death or permanent injury to the patient [Aspden et al. (2003)]. The Joint Commission (JC) has advised another approach, stating:

> The aggregation of data from many health care organizations about their medical/health care errors and the root causes of these errors is necessary to set priorities for error reduction activities; to identify priorities for system/process redesign in health care organizations; and to assess the effectiveness of the efforts to reduce errors over time [The Joint Commission (2000)].

Voluntary reporting systems of medical errors collect detailed information on the root causes and circumstances of errors across many hospitals and provide an opportunity to approach patient safety as suggested by the JC. In particular, MEDMARX, a national, anonymous, subscription-based reporting system for medication errors launched by the United States Pharmacopeia in 1998, is a useful data source for investigators interested in understanding the processes of error and harm in hospital medication use. MEDMARX is the world's largest comparative repository of medication error data that includes today around two million medication errors and adverse drug reaction records, with a growth rate of the records number of more than one percent every month. It provides detailed information on medication errors occurring in hospitals, includingj (1) the "node," defined as the step in the medication use process where the error has occurred (e.g., Prescribing, Documenting, Dispensing); (2) the "type" of error (e.g., Improper dose, Wrong patient, Wrong time); and (3) one or more "causes" of error (e.g., Communication, Computer software, Decimal points). In addition, reports submitted to MEDMARX contain a standardized categorization of the degree of harm to the patient that is associated with the event. Errors that resulted in patient harm, known as adverse events, and errors that did not result in patient harm, known as near misses, are both reported to MEDMARX, with near misses comprising the vast majority of reported events (approximately 98%).

Morlock et al. (2010) used data from MEDMARX to provide evidence that the odds of reporting a given cause when a near miss has occurred is highly correlated with the odds of reporting the same cause when an adverse event has occurred. In other words, they found that near misses and adverse events have similar causes and contributing factors, and, therefore, they suggest that data on near misses can

be used to design strategies for preventing harm. However, there may still exist some combinations of error characteristics in terms of node, type and causes that can be identified as having a higher risk of harm than expected when they co-occur. These errors are obviously dangerous, but they are also the most difficult to identify, since they are rarely reported unless a serious adverse event has occurred. Therefore, identifying these high-risk errors and their characteristics, regardless of how often they occur, is imperative for improving patient safety in the hospital and would significantly contribute to the field of patient safety. In particular, identifying the errors with the highest risk of harm in each node of medication use will allow investigators to target prevention strategies to each step of the medication delivery process and to the hospital staff members that put these steps into action.

In this paper, we focus on modeling the risk of harm associated with an error profile, which we define as the co-occurrence of an error type and two error causes. For example, a prescribing error (Type) caused by communication problems (Cause 1) and by workflow disruption (Cause 2) characterizes a unique error profile. We chose to define an error profile by its type and causes because this characterization is specific enough to recommend targeted interventions, but general enough to include many reported events so that we have sufficient statistical power to estimate and rank the log odds of harm for each profile, separately by node.

Our goal in this paper is to identify error profiles in MEDMARX data that have been reported with harm more often than expected under the assumption of independence between error profile and harm. To this end, we first develop a Bayesian hierarchical model (BHM) for estimating the log odds of harm for each error profile, separately by node of medication use, accounting for the varying amount of data available on each error profile. Second, to check the robustness of our findings, we adapt to our situation the Gamma Poisson Shrinker (GPS) approach developed by DuMouchel (1999), an empirical Bayes method currently used by the Food and Drug Administration (FDA) adverse event reporting system in the U.S.[2] While both methods aim at identifying unusually large cell counts, they differ with respect to the following characteristics: (1) BHM is a logistic regression with random effects for the error profiles and the hospital effects, while GPS assumes a Poisson distribution with unknown means for the observed counts; (2) in BHM we assume a flexible distribution for the error profile random effects, a skew-t distribution [Fernández and Steel (1998)], while GPS uses a mixture of two gamma distributions for the Poisson rates; (3) the BHM approach is embedded in a fully Bayesian framework, while GPS uses an empirical Bayes approach; this implies that the BHM provides an estimate of the ranks which also accounts for the posterior uncertainty in the estimates of the log odds of harm, while GPS only provides an ordering; (4) GPS requires the definition of "baseline" frequencies, while BHM

[2] http://www.fda.gov/ScienceResearch/DataMiningatFDA/ucm446239.htm.

doesn't require the definition of any baseline; (5) the BHM approach we present allows for a thorough sensitivity analysis of the random effects distribution by using importance link function estimation; (6) BHM uses MCMC for computing the parameter estimates, while GPS requires the maximization of the marginal likelihood, which provides less numerically stable results when using sparse data; (7) BHM is a regression model, thus it can be easily extended to include further hospital and error-dependent covariates as well as time trends or seasonal effects.

MEDMARX data presents several challenges to this objective. First, the data are nested within hospitals; error reports are submitted from multiple hospitals, and the types of errors, as well as the frequency of reporting, may vary across hospitals. Second, the data are high dimensional. There are many reports of error (we consider here a subset of approximately 1.1 million), and for each report, the data contain indicators of each of 67 possible causes and 14 possible types, in addition to the categorical variable for node, resulting in 30,954 distinct error profiles possible in each node. Third, the number of occurrences of each error profile varies widely across profiles. Some profiles may be cited on thousands of error reports, while other profiles are very rare. Therefore, comparisons across error profiles must account for the fact that there are different amounts of information available for each profile.

The objective and related challenges of this study are similar, but distinct, to what is faced in analyses of spontaneous reporting systems (SRSs), the most common form of pharmacovigilance [see, e.g., Ahmed, Bégaud and Tubert-Bitter (2015), DuMouchel (1999), Gibbons et al. (2008), Madigan et al. (2010)]. SRS databases collect reports of adverse reactions believed to be associated with medication use, regardless of whether or not error was involved. SRSs are typically used for early detection of signals of new, rare or serious adverse drug reactions (ADRs). These reactions may not have been detected by the relatively small numbers of patients included in pre-marketing clinical trials or by larger post-marketing surveillance studies. The analysis of the SRS databases usually lead to further confirmatory investigations or sometimes regulatory warnings and changes of product information leaflets. In Section 2, we describe the MEDMARX data and illustrate some important data features. In Section 3, we present the BHM and methods for estimation and inference. In Section 4, we present the empirical Bayes method adapted from DuMouchel (1999). In Section 5, we apply the methods to the MEDMARX data to investigate the most harmful error profiles in the prescribing node. Finally, in Section 6, we discuss the statistical and scientific findings and their impact on the field of medication safety.

2. The MEDMARX database. MEDMARX is a national, anonymous, subscription-based reporting system for medication errors launched by United States Pharmacopeia in 1998, that today has collected more than two million medication errors [Huckels-Baumgart and Manser (2014), Santell et al. (2003), Schiff et al. (2015)]. In this paper we consider a subset of the MEDMARX database, the

TABLE 1

Number and percent of error reports in each harmscore category. The harmscore categories are defined by The National Coordinating Council for Medication Error Reporting and Prevention (NCC MERP). We do not include reports in category A because no error has occurred. Near misses include reports in the categories B, C and D. Adverse events include reports in the categories E, F, G, H and I

Category	Description	Number	Percent
No Error			
A	Circumstances or events that have the capacity to cause error.	–	–
Error, no harm			
B	An error occurred but the error did not reach the patient.	490,638	44.56
C	An error occurred that reached the patient but did not cause patient harm.	506,589	46.01
D	An error occurred that reached the patient and required monitoring to confirm that it resulted in no harm to the patient and/or required intervention to preclude harm.	83,771	7.61
Error, harm			
E	An error occurred that may have contributed to or resulted in temporary harm to the patient and required intervention.	15,908	1.44
F	An error occurred that may have contributed to or resulted in temporary harm to the patient and required initial or prolonged hospitalization.	3354	0.30
G	An error occurred that may have contributed to or resulted in permanent patient harm.	183	0.02
H	An error occurred that required intervention necessary to sustain life.	376	0.03
Error, death			
I	An error occurred that may have contributed to or resulted in the patient's death.	136	0.01
Total		1,100,955	100.00

reports collected over the period January 1, 1999, to December 31, 2007, corresponding to 1,100,955 reports of medication errors collected from 688 participating hospitals. Each report is categorized according to the "harmscore" developed by the National Coordinating Council for Medication Error Reporting and Prevention (NCC MERP) (2001). The harmscore identifies the degree of harm to the patient caused by the reported error. Table 1 summarizes the number and percent of errors in each harmscore category for the data we consider. The great majority of reported errors (98.2%) did not harm the patient. Each error report also contains information on several variables describing and characterizing the circumstances of the error. There are 14 predefined types of error that may be selected on a report, and multiple types may be cited simultaneously on a single report of error.

Reports may also cite multiple causes simultaneously out of the 67 predefined choices available. Definitions for all potential error types and causes are available in Venturini et al. (2017a).

We created a new dataset containing the number of times each error profile was reported by each hospital and the number of times it was reported with patient harm in each hospital, separately by the node of medication use where the error occurred. We excluded reports from hospitals that submitted fewer than 100 total reports because these hospitals have the least experience in reporting, and, therefore, they are likely to have the poorest quality of error reports. The 578 remaining hospitals submitted 1,097,259 reports between 1999 and 2007. Although there were 30,954 distinct error profiles possible, only a small subset of these possibilities were ever reported in each node because not all error profiles can occur in each node. Also note that a single report may contribute to the counts for more than one profile, since, for example, a report citing two types would be included in the profiles for both types.

Table 2 shows separately by node: (1) where the medication error occurred, (2) the average probability of harm (p) across error profiles weighted by the number of times each error profile occurred, (3) the number of error profiles that are reported at least once with harm ($y > 0$), (4) the number of error profiles that are reported at least once overall ($N > 0$), and the number of error profiles that are reported at least twice, both (5) with harm ($y > 1$) and (6) overall ($N > 1$). The row in Table 2 labeled "Not recorded" includes information from error reports where a node was not marked. These reports make up a very small subset of submitted reports and do not include any reports of errors resulting in harm.

We restrict the analysis to the error profiles that were reported with harm at least twice ($y > 1$) because we are not interested in those error profiles that never result in harm, and those profiles resulting in harm only once may be due to a

TABLE 2

Separately by the node of medication use where the error occurred, the average probability of harm (p) across all error profiles, weighted by the number of times each error profile occurred, the number of error profiles reported at least once both with harm (y) and overall (N), and the number of error profiles reported at least twice, both with harm and overall. The row "Not recorded" contains information from reports where the node was not marked

Node	**100 × p**	**$y > 0$**	**$N > 0$**	**$y > 1$**	**$N > 1$**
Administering	3.6045	3679	15,414	1798	11,379
Dispensing	0.8897	2423	18,970	1028	14,010
Documenting	1.3496	2827	16,497	1328	11,914
Monitoring	8.0305	591	3302	232	1376
Prescribing	3.0714	2826	12,444	1115	8435
Procurement	1.3363	50	1967	4	554
Not recorded	0.0000	0	60	0	1

singular misreporting. In addition, we stratify the ranking of the error profiles by node to provide targeted recommendations for improving safety in each step of the medication use process. Indeed, exploratory analyses we performed have shown that the effects of both hospital and error profile on the log odds of harm varies widely across nodes. In addition, Table 2 confirms the need for differing mean and variance parameters for the log odds of harm in each node. We present here results for the prescribing node only, while the results for the other nodes are provided in Venturini et al. (2017a).

3. A Bayesian Hierarchical Model (BHM) with flexible random effects. In this section, we introduce a BHM for identifying the medication error profiles with the largest log odds of harm.

3.1. *Model definition.* Let N_{ij} be the number of times that the co-occurrence of the three events (error type, first cause, second cause) that define the error profile i is cited on a report from hospital j. Let y_{ij} be the corresponding number of times that profile i in hospital j was reported with harm. To estimate the log odds of harm for each error profile, we introduce the following BHM with crossed random effects:

$$
\begin{aligned}
&\text{Level I:} && y_{ij}|N_{ij}, p_{ij} \sim \text{Bin}(N_{ij}, p_{ij}), \\
&&& \text{logit}(p_{ij}) = \gamma + \theta_i + \delta_j, \\
&\text{Level II:} && \theta_i|\sigma, \eta, k \sim \text{St}(0, \sigma, k, \eta), \qquad i = 1, \ldots, n, \\
&&& \delta_j|\tau^2 \sim N\big(0, \tau^2\big), \qquad j = 1, \ldots, J, \\
&\text{Priors:} && \gamma \sim N(g, G), \\
&&& \tau^2 \sim \text{IG}(a_2, b_2), \\
&&& \sigma^2 \sim \text{IG}(a_1, b_1), \\
&&& k \sim \text{Unif}(0, \infty), \\
&&& \eta \sim \text{Unif}(0, \infty).
\end{aligned} \tag{3.1}
$$

In this model, p_{ij} is the probability of harm for the errors with profile i in hospital j. We model the logit of p_{ij} in terms of a fixed effect γ and two sets of cross-classified random effects [Gelman and Hill (2007)]: (1) an effect for error profile, $\{\theta_i, i = 1, \ldots, n\}$, and (2) an effect for hospital, $\{\delta_j, j = 1, \ldots, J\}$. Because our primary interest lies in the estimation of the error profile random effects, we follow the recommendation in Lee and Thompson (2008) and consider a highly flexible skew-t distribution on the profile random effects, where σ is the scale parameter, η parameterizes the amount of skewing and k is the degrees of freedom for the t distribution. This distribution is based on introducing skewing into the

symmetric scaled t distribution, as described in Fernández and Steel (1998). The parameters characterizing the center (in our case, set at 0) and the spread (σ) refer to the mean and standard deviation of the underlying symmetric distribution. In the skew-t distribution, the centrality parameter defines the mode of the distribution, but it is no longer either the mean or the median. Similarly, in the skew-t distribution, σ still characterizes the spread, but it can no longer be interpreted directly as the standard deviation of the distribution. The posterior distributions on σ, k and η provide evidence on the shape of the random effects distribution that is supported by the data.

Under the skew-t distribution, we interpret γ as the mode of the distribution of the log odds of harm across all error profiles and hospitals. The random effect θ_i, our main parameter of interest, is the additional log odds of harm with respect to γ that is associated with error profile i, and σ characterizes the heterogeneity in the true log odds of harm across error profiles, controlling for the clustering of data within hospital. The random effect δ_j is the additional log odds of harm with respect to γ that is associated with hospital j, and τ characterizes the heterogeneity in the true log odds of harm across hospitals, controlling for error profile. The values for the hyperparameters were chosen to induce noninformative priors, including $g = -4$ (corresponding to the logit of the overall probability of harm among all reports, that is, 1.8%), $G = 1000$, and $a_1 = b_1 = a_2 = b_2 = 0.001$. Improper uniform priors were used for k and η.

3.2. *Model estimation*. A full Bayesian estimation of model (3.1) requires the implementation of a complicated Markov Chain Monte Carlo simulation algorithm with data augmentation. The data augmentation approach is motivated by the representation of a Student t-distribution as a scale mixture of normals [Fernández and Steel (1998)]. Therefore, to alleviate the computational burden of the algorithm, we approximate the calculation of the marginal posterior distribution of the θ_i parameters by adopting a two-step approach. First, we fix $k = \infty$ and $\eta = 1$, forcing a symmetric, normal distribution on the θ_i. With k and η fixed, we obtain a sample, $\pi^{\infty,1}$, from the joint posterior distribution of all other parameters via MCMC with adaptive Metropolis steps for each set of random effects [Haario, Saksman and Tamminen (2001)]. More specifically, the model is reparameterized (hierarchically centered) so that we sample from the posteriors of $\gamma + \delta_j$ rather than directly from the posteriors of δ_j to improve efficiency [Browne (2004)]. We use adaptive random-walk Metropolis–Hastings steps for the random effects, $\boldsymbol{\theta}_i$ and $\gamma + \boldsymbol{\delta}_j$, to achieve acceptance rates between 20–50%. The initial proposal distribution for each random effect is taken to be a normal distribution with a small standard deviation of 0.25. This standard deviation was updated every 100 iterations of the chain. The chain is checked for convergence every 2500 samples and then the Monte Carlo error is computed [see Flegal, Haran and Jones (2008)]. After 15,000 iterations, with the first 2500 discarded as burn-in, we achieved the

desired Monte Carlo error (< 0.05 for all parameters except for the very noisy hospital-specific random effects for which the error was < 0.2).

In the second step, we consider deviations from the normal random effects distribution, including $k = \{3, 6, 10, 30, 60, \infty\}$ and $\eta = \{0.5, 0.8, 1, 1.25, 2\}$, corresponding to an extreme left skew, moderate left skew, no skewing, moderate right skew and extreme right skew, respectively. For each pair of (k, η) values, we use importance link function estimation [MacEachern and Peruggia (2000)] based on the chain $\pi^{\infty,1}$ to obtain new posterior samples under these values. Then we approximate the marginal posterior distribution of (k, η) using importance resampling. We briefly review the main idea of importance link function estimation in Section 2 of Venturini et al. (2017a).

In particular, we first transform the sample $\pi^{\infty,1}$ by the link function, which should be chosen to yield a transformed sample, $\tilde{\pi}^{k,\eta}$, that more closely resembles the desired posterior distribution. In this case, the only parameters that we expect to have significantly changed posteriors under the new values of (k, η) are the θ_i. Therefore, we use the identity transformation for all other parameters (γ, $\boldsymbol{\delta}$, σ, τ), and choose the transformation for each θ_i as the corresponding unnormalized (conditional) posterior distribution

$$f_i(\theta_i; \gamma, \boldsymbol{\delta}, \sigma, k, \eta) = \prod_{j=1}^{J} \mathrm{Bin}(y_{ij}|N_{ij}, p_{ij}) \,\mathrm{St}(\theta_i|0, \sigma, k, \eta). \tag{3.2}$$

We let $\tilde{\theta}_i^{k,\eta} = \hat{\theta}_i^{k,\eta} + A_i^{k,\eta}(\theta_i - \hat{\theta}_i^{\infty,1})$, where

$$\hat{\theta}_i^{k,\eta} = \max_{\theta} f_i(\theta; \hat{\gamma}, \hat{\boldsymbol{\delta}}, \hat{\sigma}, k, \eta),$$

$$\hat{\theta}_i^{\infty,1} = \max_{\theta} f_i(\theta; \hat{\gamma}, \hat{\boldsymbol{\delta}}, \hat{\sigma}, k = \infty, \eta = 1),$$

$$A_i^{k,\eta} = \sqrt{\frac{\frac{\partial^2}{\partial\theta^2}\log\{f_i(\theta; \hat{\gamma}, \hat{\boldsymbol{\delta}}, \hat{\sigma}, \infty, 1)\}|_{\theta=\hat{\theta}_i^{\infty,1}}}{\frac{\partial^2}{\partial\theta^2}\log\{f_i(\theta; \hat{\gamma}, \hat{\boldsymbol{\delta}}, \hat{\sigma}, k, \eta)\}|_{\theta=\hat{\theta}_i^{k,\eta}}}},$$

and θ_i is a sample from $\pi^{\infty,1}$. The quantities, $\hat{\gamma}$, $\hat{\boldsymbol{\delta}}$ and $\hat{\sigma}$ are the posterior mean estimates of these parameters from $\pi^{\infty,1}$. Moreover, $\hat{\theta}_i^{k,\eta}$ and $\hat{\theta}_i^{\infty,1}$ are the modes of the posterior distribution (3.2) under the corresponding two cases. Last, the quantity $A_i^{k,\eta}$ is related to the Jacobian of the transformation defined above [see Example 2 in MacEachern and Peruggia (2000)].

Finally, we take a 10% resample (without replacement) from the transformed chain, $\tilde{\pi}^{k,\eta}$, with a sampling probability proportional to the importance ratio

$$IR = \prod_{i=1}^{n} \frac{f_i(\tilde{\theta}_i^{k,\eta}; \gamma, \boldsymbol{\delta}, \sigma, k, \eta)}{f_i(\theta_i; \gamma, \boldsymbol{\delta}, \sigma, \infty, 1)/A_i^{k,\eta}}. \tag{3.3}$$

Under uniform priors for k and η, we estimate the unnormalized posterior probability of each combination of (k, η) values by summing the corresponding importance ratios [see, e.g., Gelman et al. (2014), Section 13.5].

Even if it does not take fully into account the uncertainty related to the estimation of k and η, our approach presents two main advantages. The first one regards the computational speed of the algorithm, which is fairly reasonable even in our implementation in plain `R` (see Section 7), that is, without relying on more efficient programming languages such as C/C++ or Julia. The second advantage is the possibility to perform a thorough sensitivity analysis of the results by choosing any specific values of k and η. Furthermore, one could easily modify the equations above and consider a prior specification for the θ_i's which differs from the one we chose.

3.3. *Optimal Bayesian ranking.* Using the posterior samples of the θ_i, we estimate the ranks of the log odds of harm of the various error profiles. We choose to rank profiles based on the log odds of harm because a high log odds of harm could indicate inadequate safeguards for that combination and ample opportunity for system improvement. We desire estimates that produce high ranks when the log odds of harm is high, appropriately accounting for the varying amount of uncertainty across estimates of θ_i. We use optimal Bayesian ranking as described in Shen and Louis (1998) and Louis and Shen (1999), which gives estimates of rank for profile i

$$\hat{R}_i = \sum_{k=1}^{n} \widehat{\mathbb{P}}(\theta_k \le \theta_i | \boldsymbol{y}, \boldsymbol{N}), \tag{3.4}$$

where $\widehat{\mathbb{P}}(\theta_k \le \theta_i | \boldsymbol{y}, \boldsymbol{N})$ denotes the estimate of the posterior probability that $\{\theta_k \le \theta_i\}$ based on the sampled values from the MCMC simulation. Typically, the optimal ranks $\hat{R}_i$ are not integers.

3.4. *Posterior predictive model checking.* In the previous sections, we outlined a model and an estimation technique that are focused on ranking error profiles with respect to their log odds of harm, accounting for the heterogeneity across hospitals. We allow the data to provide evidence on the preferred shape of the random effects distribution for error profiles (i.e., on the preferred values of σ, k and η), but restrict the hospital random effects to a normal distribution. Therefore, investigators may be interested in checking the quality of model fit for the hospital random effects and determining the improvement in fit from the resampled model in (3.1) compared to the model with normally distributed random effects for error profiles.

These questions may be addressed using the posterior predictive checking strategy described in Gelman, Meng and Stern (1996) and Gelman et al. (2000). We create a sample of replicated data, $Y_{ij}^{\text{rep}} \sim \text{Bin}(N_{ij}, p_{ij})$, where the values of p_{ij}

are taken from the MCMC chain from a fitted model. Let the test statistic, T, be some function of the data. We compare the values of T in the replicated data, T^{rep}, to the value of T in the observed data, T^{obs}, via the posterior predictive p-value

$$p\text{-value}_T = \min\{\widehat{\mathbb{P}}(T^{\text{rep}} \leq T^{\text{obs}} | \boldsymbol{y}, \boldsymbol{N}), \widehat{\mathbb{P}}(T^{\text{rep}} \geq T^{\text{obs}} | \boldsymbol{y}, \boldsymbol{N})\}. \tag{3.5}$$

In order to examine the adequacy of model fit separately for the error profile random effects and the hospital random effects, we consider two sets of test statistics: $Y_{i+} = \sum_{j=1}^{J} Y_{ij}$ and $Y_{+j} = \sum_{i=1}^{n} Y_{ij}$. Small p-values for the Y_{i+} test statistics indicate error profiles with observed data that is not predicted well by the model and possible problems in the error profile random effects distribution. Small p-values for the Y_{+j} test statistics indicate hospitals with observed data that is not predicted well by the model and possible problems in the hospital random effects distribution. We calculate posterior predictive p-values for each error profile (Y_{i+}) and hospital (Y_{+j}) before and after the resampling step described in Section 3.2.

4. Empirical Bayes Data Mining (EBDM). As a way to check the robustness of the ranking determined by the method presented in the previous sections, we now describe how to adapt the approach discussed in DuMouchel (1999) to our medication error data. This method is one of the most widely used approaches for signal detection in SRSs, and it is used by the FDA in its adverse event reporting system. In DuMouchel (1999), investigators sought to determine likely adverse drug reactions from post-marketing data. Statistically, this objective requires the identification of "interestingly large" cell counts in a large, sparse frequency table of medication-event combinations. As in DuMouchel (1999), our data can also be represented as a large, multidimensional contingency table, given by (Harm $\times$ Type $\times$ Cause 1 $\times$ Cause 2). We would like to identify the "interestingly large" cell counts, particularly those cells containing information on harmful errors that have large counts compared to the count that would be expected if error profile and harm were independent. In addition, we want to stratify the statistical analysis by hospital to control for potential reporting biases by hospital.

We illustrate how to apply this approach to our data in Venturini et al. (2017a).

5. Identifying the most harmful error profiles in the prescribing node. There were a total of $n = 1115$ different profiles of medication error in the prescribing node that were reported as resulting in harm at least twice ($y > 1$) between 1999 and 2007 from 533 hospitals that reported more than 100 total medication errors. The top panels of Figure 1 show boxplots of the raw probability of harm for each error profile (on the left) and each hospital (on the right) by binned sample size (N), with the raw probability of harm defined as $\sum_{j=1}^{J} y_{ij} / \sum_{j=1}^{J} N_{ij}$ for the error profiles and $\sum_{i=1}^{n} y_{ij} / \sum_{i=1}^{n} N_{ij}$ for the hospitals. The lower panels of Figure 1 present the frequencies of error profiles and hospitals in each bin. These plots

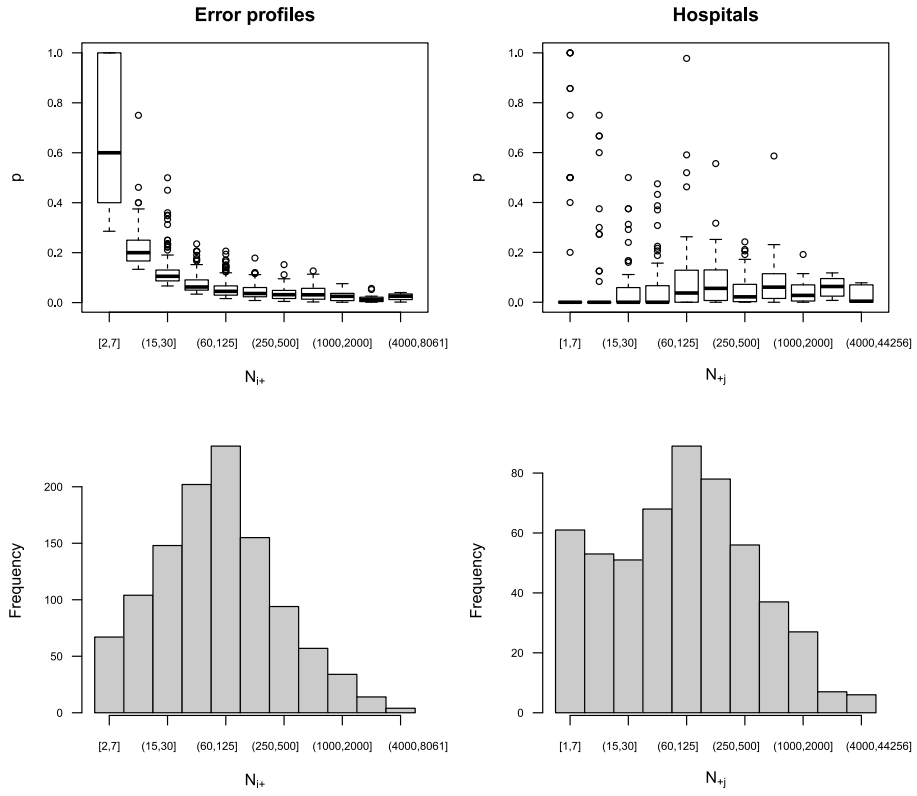

FIG. 1. *The top panel contains boxplots of the raw rates of harm for error profiles* (*left*) *and hospitals* (*right*), *binned according to the total sample size* (*total number of reports*) *for each profile or hospital. The lower panel contains the number of error profiles* (*left*) *or hospitals* (*right*) *in each bin of sample size.*

show that the error profiles with the highest estimated probabilities of harm are also those with the smallest sample size, that is, those that occurred less frequently. This finding indicates that the highest observed rates of harm are due to sampling variability, rather than strong effects for these error profiles, and estimation will benefit from a random effects model that shrinks the less precisely estimated log odds of harm toward the mean. The highest observed rates of harm for error profiles with moderately large sample sizes (> 30) are in the range of 10–20%. These rates are still large compared to the overall rate of harm (1.8%), but are much smaller than the observed rates for the least frequently cited error profiles. On the other hand, many hospitals with large sample sizes report a high proportion of harmful errors. For example, one hospital that reported between 501 and 1000 errors cited harm on nearly 60% of reports. Clearly, much of the variability in the reporting of harm can be explained by the varying tendency to report harmful events across hospitals, and this variation must be accounted for when ranking error profiles.

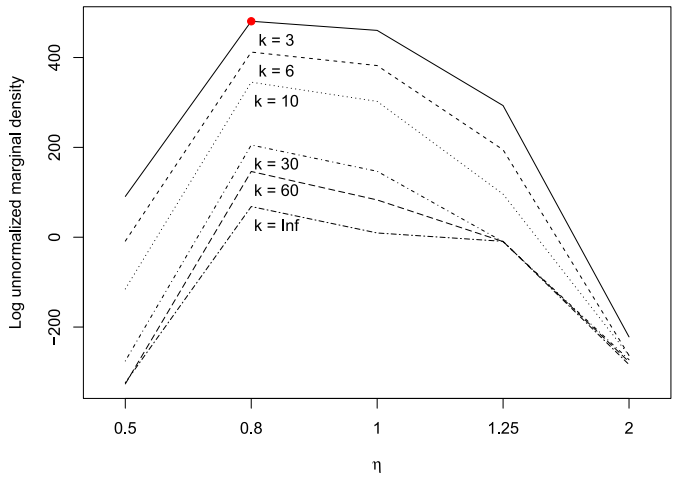

FIG. 2. *The log unnormalized marginal posterior distributions for k and η, calculated at the six values of k and five values of η considered. The dot indicates the maximum at $k = 3$ and $\eta = 0.8$.*

Figure 2 displays the log of the unnormalized marginal posterior densities for k and η, calculated from the importance ratios (3.3) for the six values of k and five values of η considered. This figure shows that the data support a moderate left skew ($\eta = 0.8$) in the distribution of random effects across error profiles. Figure 2 also shows that the data support small values of k ($k = 3$), corresponding to distributions with much heavier tails and less shrinkage than the normal distribution. Therefore, we focus the remainder of the paper on results from the model estimated with a skew-t random effects distribution for error profile with skew parameter $\eta = 0.8$ and degrees of freedom $k = 3$.

Figures C1 and C2 in Venturini et al. (2017c) provide details about the BHM estimation, while Figure C3 in the same document compares the posterior predictive p-values under the model with normally distributed error profile random effects and under the resampled model with skew-t random effects. The latter figure shows that prediction is substantially improved after resampling for most profiles.

Table 3 summarizes the fifteen error profiles within the prescribing node that have the highest estimated optimal Bayesian ranks based on the log odds of harm. Table 3 also summarizes for each of the most harmful profiles: (1) the ordering based on $EBGM_{htc_1c_2}$, the estimated geometric mean of the empirical Bayes posterior distribution as defined in Venturini et al. (2017c), (2) the probability of harm as estimated by the BHM, (3) the number of times the profile resulted in harm, (4) the number of times that the profile occurred, and (5) the type and causes defining the error. For example, the top-kranked error profile in the prescribing node is a "prescribing error" (type) caused by "performance (human) deficit" (cause 1) and "reconciliation-discharge" (cause 2). It has an estimated 10.48% probability of being reported with harm, and it was reported 23 times overall with 8 of those reports

TABLE 3

The fifteen medication error profiles in the prescribing node with the highest estimated optimal Bayesian ranks. The maximum possible rank is $n = 1115$. *The first column* (*BHM*) *summarizes the estimated optimal Bayesian rank based on the estimated probability of harm from the BHM in the third column* ($100\hat{p}$); *the second column* (*EBDM*) *contains the ordering based on the estimated empirical Bayes grand mean* $EBGM_{htc_1c_2}$ *from the EBDM model*; *the fourth column* (y) *contains the number of times the event resulted in harm*; *the fifth column* (N) *summarizes the total number of times the event occurred. The remaining columns describe the error profile definition* (*type* $\times$ *cause* 1 $\times$ *cause* 2)

	BHM	**EBDM**	$\mathbf{100\hat{p}}$	y	N	**Type**	**Causes**
1	1058.5	1073	10.48	8	23	Prescribing error	Performance (human) deficit Reconciliation-discharge
2	1049.1	1107	4.93	38	213	Improper dose/quantity	Knowledge deficit System safeguard(s)
3	1045.7	1103	5.19	25	262	Improper dose/quantity	Knowledge deficit Monitoring inadequate/lacking
4	1044.4	1114	4.48	75	493	Prescribing error	Knowledge deficit System safeguard(s)
5	1009.8	1104	4.47	29	422	Improper dose/quantity	Monitoring inadequate/lacking Performance (human) deficit
6	1009.4	1110	4.18	49	592	Prescribing error	Knowledge deficit Monitoring inadequate/lacking
7	979.6	1091	4.76	16	94	Prescribing error	Communication Contraindicated in disease
8	970.5	1093	4.36	19	214	Improper dose/quantity	Monitoring inadequate/lacking Procedure/protocol not followed
9	944.2	1112	3.56	76	1193	Prescribing error	Monitoring inadequate/lacking Performance (human) deficit
10	942.2	1109	3.58	64	504	Prescribing error	Communication System safeguard(s)
11	922.8	1086	4.27	13	90	Prescribing error	Dispensing device involved Contraindicated, drug allergy
12	906.8	1101	3.46	36	487	Prescribing error	Documentation Contraindicated, drug allergy
13	892.5	1094	3.50	24	226	Prescribing error	Monitoring inadequate/lacking System safeguard(s)
14	890.1	1105	3.35	43	534	Prescribing error	Monitoring inadequate/lacking Procedure/protocol not followed
15	880.2	1097	3.35	33	295	Improper dose/quantity	Performance (human) deficit System safeguard(s)

citing harm. The estimated optimal Bayesian rank for this error profile is 1058.5, and the ordering from the EBDM method is 1073 out of $n = 1115$. Therefore, both the BHM and the EBDM method indicate that preventing occurrences of this error profile is one of the top priorities for reducing harm due to medication errors in the hospital.

The cause "reconciliation-discharge" refers to a process of reviewing a patient's medications at the time of discharge and determining if discrepancies between the current and ordered medication list is due to error. Medication reconciliation has been shown to reduce prescribing errors in discharges from the ICU [Pronovost et al. (2003)]. However, the high probability of harm in this top-ranked error profile indicates that when mistakes are made in reconciliation, the resulting prescribing errors will often result in harm to the patient. Therefore, double checking the reconciliation review may be a useful intervention for hospitals that do not currently have a system in place for preventing these errors.

Other error profiles listed in Table 3 vary widely with respect to the specificity of the error profile definition and the number of occurrences. We find that most highly ranked error profiles with moderate to small sample sizes are associated with highly specific causes. For example, the 7th ranked error profile is a "prescribing error" caused by "communication" and "contraindicated in disease." This error profile directly indicates the kinds of interventions that might be useful for preventing it—in this case, interventions to improve communication among healthcare providers about patients' comorbidities during prescribing. Conversely, the majority of error profiles with very large sample sizes relate to combinations of nonspecific error causes and types, resulting in noninformative intervention recommendations. For example, the error profile with the highest ordering from the EBDM model is a "prescribing error" caused by "knowledge deficit" and "communication." It occurred a total of 2808 times and resulted in harm 147 times. While this error profile clearly represents many harmful events, its definition is very broad, making the development of a targeted prevention strategy for these errors very challenging. This error profile is given a rank of 836.4 out of 1115 by the BHM.

In general, the BHM and EBDM method provided qualitatively similar orderings of the error profiles. Of the error profiles with a Bayesian rank in the top fifteen, nine have an $EBGM_{htc_1c_2}$ value in the top 15, and all have an $EBGM_{htc_1c_2}$ in the top 50 out of the 1115 error profiles considered. However, important quantitative differences between the two methods are observed in this analysis. Specifically, the BHM generally gives higher estimated ranks to error profiles with higher raw rates of harm and smaller sample sizes and, therefore, highly specific error profile definitions. The EBDM model gives higher orderings to error profiles with lower raw rates of harm and larger sample sizes and, therefore, less specific error profile definitions.

6. Discussion. Databases such as MEDMARX contain information on many sources of potential harm in the medication-use process. Even hospitals with a sophisticated safety culture have neither the time nor the resources to understand and intervene on all of them simultaneously. Moreover, events that are rare but very harmful may be difficult to detect within a single hospital until they have already caused significant patient harm. Using data from many hospitals to prioritize

medication events for intervention allows us to consider the vast array of potential events in each step of the medication-use process and identify those that are lacking or have inadequate safeguards. Therefore, a robust approach for prioritization and characterization of events from these data is very important for improving medication safety. The statistical model and data analysis presented here allow us to characterize the medication errors that pose the highest risk of harm across many hospitals. In addition, considering that in recent years many hospitals have implemented systematic programs for medication reconciliation [see, e.g., Boockvar et al. (2006), Kwan et al. (2013), Mueller et al. (2012), Ramjaun et al. (2015)], another application of the methodology we present could be to investigate the effectiveness of such large-scale interventions through a comparison of the results before and after the introduction of such programs.

The error profiles that have a high log odds of harm in the MEDMARX data and that will be identified by the methods presented in this paper must result from one of two possible patterns, both of which indicate a need for interventions: (1) the error profile results in harm frequently when it occurs, and is generally reported accurately regardless of the resulting harm, or (2) the error profile is rarely caught and reported unless harm has occurred, causing the nonharmful occurrences of the error profile to be underreported. In the first case, these profiles clearly have large opportunities for improvements in safety and are in need of additional safeguards to prevent the errors from causing harm. In the second case, the profiles are in need of interventions to catch the errors *before* they cause harm. Therefore, the error profiles identified with the highest log odds of harm will be important for setting intervention priorities, regardless of error profile-specific biases in reporting. In this work we focused exclusively on the error profiles, while we disregarded completely the hospital effects, denoted as δ_j in (3.1). We did it on purpose because these parameters strongly depend on each hospital's reporting practices. Different hospitals use the system in different ways, and so we adjust for hospital effects to account for this, but don't want to interpret those parameters as indicative of hospital quality. A further decision we took is to define an error profile by its type and causes. This characterization may appear subjective because it may not include all relevant aspects of an error, or even overemphasize some of them. However, this definition provides a good comprise for recommending targeted interventions while including in the analysis many reported events so that we do not lose too much statistical power.

We considered two methods for ranking the error profiles with respect to their probability of resulting in harm: a BHM for the log odds of harm and the EBDM model for the ratio of observed to expected rates of reporting harm that was adapted from the SRS literature [see Ahmed, Bégaud and Tubert-Bitter (2015), Gibbons and Amatya (2016)]. Both methods produced qualitatively similar orderings of the error profiles by shrinking estimates for profiles with extreme observed rates of harm and small sample sizes. In addition, both methods account for the variation

in the rates of reporting harm across hospitals. However, there were some important differences between the two methods. First, for each error profile, the BHM provides an estimate of the rank that accounts for the posterior uncertainty in our estimates of the log odds of harm. The EBDM method provides only an ordering, based on the posterior means of the error profile parameters. This ordering does not incorporate the uncertainty in the profile parameters and does not accurately represent the distance between the estimated parameters across error profiles. Understanding both the distance between parameter estimates and our uncertainty about those estimates is important for determining which error profiles are most in need of intervention.

Second, the EBDM model is computationally more difficult than the BHM because it involves the maximization of a five-dimensional likelihood function. In the data presented in this analysis, finding the likelihood maximum required an expectation-maximization (EM) algorithm using several randomly chosen starting points to ensure that the maximum identified with this algorithm was global. An unexpected feature we found is that the parameter estimates we obtained are larger as compared to those typically arising in the pharmacovigilance context [e.g., DuMouchel (1999)]. As a consequence, this makes the posterior distribution of the λs rather insensitive to the data themselves except for very large counts. One solution to these computational problems may be to embed the EBDM approach in a fully Bayesian framework by specifying hyperprior distributions on the parameters in the gamma mixture distribution. Estimating the model in this way would further allow for optimal Bayesian ranking of the error profile parameters, λ, as mentioned by DuMouchel (1999), as well as exploration of other random effects distributions for λ, as done in the BHM. However, we found that, in some datasets, estimating the EBDM model would be very challenging, regardless of the estimation procedure, because one or more of the expected counts $E_{htc_1c_2}$ were equal to zero. In this case, the ratio $\lambda_{htc_1c_2}$ will be fixed at infinity, no matter how much shrinkage is applied to $\mu_{htc_1c_2}$. Therefore, in these data, we prefer to model the probability of harm directly because it leads to more interpretable results and simpler computations.

The model presented in this paper used the notion of the importance link function [MacEachern and Peruggia (2000)] to estimate the optimal Bayesian rank for error profiles under a flexible class of random effects distributions and then to determine which of those distributions is preferred by the data. Although a normal distribution for the random effects usually provides the simplest MCMC estimation and is often reasonable, it can be restrictive. The use of a skew-t distribution for the error profile random effects combined with computation via the importance link function resampling allowed us to fully explore the sensitivity of results, including the relative ranks of error profiles, to the specification of a wide range of random effects distributions without running multiple MCMC chains. In particular, the identification of error profiles with extremely large (or small) log odds of harm was facilitated by the t distribution, which produces less shrinkage on

random effects estimates. Incorporating skewing into the random effects distribution allowed for different amounts of shrinkage in the two tails of the distribution, which is desirable since we had no a priori knowledge that the error profile random effects should be symmetric.

In particular, using the importance link function resampling to estimate a model under multiple random effects distributions provided greatly improved estimation compared to regular importance resampling. Although regular importance resampling has been used to reestimate models under varying random effects distributions [Gelman et al. (2014)], we argue that this method rarely performs well. If the posterior distributions of the random effects change considerably under the new hierarchical distribution, then the importance weights will be low for most samples, reflecting poor resampling properties. If the posterior distributions of the random effects do not change much under the new hierarchical distribution, then resampling is unneeded because results are invariant to the hierarchical distribution. In this analysis, transforming the random effects parameters prior to resampling provided a better pool of samples, especially for the random effects distributions that were best supported by the data.

Another popular Bayesian approach frequently implemented in practice to flexibly specify the random effects distributions in a hierarchical model is through nonparametric priors, for example, Dirichlet process mixtures [Escobar and West (1995), Ferguson (1973), Kyung, Gill and Casella (2010), MacEachern and Müller (1998), Müller et al. (2015)]. In recent years, there has been an explosion of proposals within this framework. The most recent ones also allow for dependence across random distributions. For example, in De Iorio et al. (2004) the random effects distributions F_x are indexed by a q-dimensional vector of categorical covariates, $x = (x_1, \ldots, x_q)$. A nonparametric probability model is then defined for F_x using an ANOVA-type structure [similar to our specification in (3.1) for $\text{logit}(p_{ij})$] such that marginally for each x the random measure F_x follows a Dirichlet process. Although these models are more difficult to implement in practice, they have proven to provide a more general approach for random effects distribution modeling, and we plan to further study their application within our context in the future.

In addition, many other approaches have been proposed for the checking of random effects distributions in generalized linear mixed models besides the posterior predictive checking strategy employed here. Frequentist approaches generally focus on diagnostic tests for goodness of fit of a null distribution [Abad, Litière and Molenberghs (2010), Huang (2011), Tchetgen and Coull (2006), Waagepetersen (2006)]. While these tests can be useful for rejecting a distribution, they do not provide guidance on alternative distributions or how to improve the model. Most Bayesian approaches are closely related to the posterior predictive check [Bayarri and Castellanos (2007), Dey et al. (1998), Sinharay and Stern (2003), Stern and Cressie (2000)], but are modified to avoid the conservatism usually associated with posterior predictive checking (in the sense that the p-values calculated from this method will not generally have a uniform distribution under the null). For example,

prior predictive checking, partial posterior predictive checking, and other modifications on posterior predictive checking focus on invalidating false random effects distributions. We use classic posterior predictive checks because we are not interested in the true shape of the random effects distributions or in the values of the hierarchical parameters from the random effects distributions; we are focused only in accurately characterizing the evidence on the comparative risks of error profiles.

In this example, posterior predictive checking indicated that the resampled model with skew-t parameters $k = 3$ and $\eta = 0.8$ has improved prediction and, therefore, improved the accuracy of rankings for many error profiles. However, there were still several error profiles with small posterior predictive p-values in the resampled model. Upon closer examination, we noticed that these error profiles correspond to the error profiles with the smallest sample sizes in the dataset ($2 \leq N \leq 8$), and the small p-values result from consistent under-prediction by the model. Recall that we restricted the dataset to only include error profiles that were reported with harm at least twice, and so, for all of these error profiles, $y_{i+}^{\text{obs}} \geq 2$. Therefore, even the resampled model may be shrinking the effects for these profiles too much. We could potentially solve this problem by considering additional values of k and η, for example, $k = 1$ to achieve a Cauchy distribution, which would produce even less shrinkage on the random effects. We don't consider this solution here, since these error profiles have very small sample sizes and may be viewed as outliers. For the majority of error profiles, especially error profiles with moderate sample sizes, we found that flexible random effects models are useful tools for accurately characterizing the relative risks of harm in a hospital.

As a final note, we would like to focus on one limitation of our analysis. Since MEDMARX data are collected on a voluntary basis, they share the same weaknesses that are typical of the SRSs, and in particular the uncontrolled collection of self-reported entries riddled with systematic under-reporting, over-reporting and duplicate reporting. Many methods for analyzing SRS data have been proposed in the literature, all focusing on the development of automatic signal detection strategies [for an overview see Ahmed, Bégaud and Tubert-Bitter (2015), Gibbons and Amatya (2016)]. More recently, various techniques have emerged whose primary aim is the minimization of the false discovery rate [Ahmed et al. (2009, 2010, 2012)] and the identification of drug-drug interactions [Eugéne et al. (2000)], but no single method can claim absolute superiority. However, to the best of our knowledge, no method exists to assess the extent and the consequences of the biases mentioned above in voluntary databases. Our belief, but we don't have any proof, is that the *anonymity* nature of the MEDMARX data may act as a partial relief from those biases (especially from under-reporting). At any rate, all these limitations present both challenges and opportunities for further methodological developments in the future.

7. Software. All the routines developed during the preparation of this manuscript are available as an `R` package called `mederrRank` freely downloadable

from the Comprehensive R Archive Network (http://cran.r-project.org). The package also contains a subset of the MEDMARX data we used, which are provided for illustrative purposes.

Acknowledgments. The content is solely the responsibility of the authors and does not necessarily represent the official views of the Institutions. We would like to thank the Editors and the reviewers for their helpful comments that greatly improved the paper.

SUPPLEMENTARY MATERIAL

Supplement A: Error Definitions and Results for the Other Nodes (DOI: 10.1214/16-AOAS974SUPPA; .pdf). This supplement contains the list of definitions for all potential error types and causes for MEDMARX reports and the results for the Bayesian hierarchical model (BHM) and empirical Bayes data mining (EBDM) approach applied to the data from each of the four other nodes of medication use: documenting, dispensing, administering and monitoring.

Supplement B: Empirical Bayes Data Mining Approach (DOI: 10.1214/16-AOAS974SUPPB; .pdf). Section 1 of this supplement shows how to adapt the GPS method developed by DuMouchel (1999) and briefly described in Section 4 to the MEDMARX data. Moreover, in Section 2 we provide a brief description of the importance link function estimation as described in MacEachern and Peruggia (2000).

Supplement C: Bayesian Hierarchical Model Estimates (DOI: 10.1214/16-AOAS974SUPPC; .pdf). This supplement reports more details about the estimation of the BHM described in Section 3.

REFERENCES

ABAD, A. A., LITIÈRE, S. and MOLENBERGHS, G. (2010). Testing for misspecification in generalized linear mixed models. *Biostat.* **11** 771–786.

AHMED, I., BÉGAUD, B. and TUBERT-BITTER, P. (2015). Evaluation of post-marketing safety using spontaneous reporting databases. In *Statistical Methods for Evaluating Safety in Medical Product Development* (A. L. Gould, ed.). Wiley, New York.

AHMED, I., HARAMBURU, F., FOURRIER-RÉGLAT, A., THIESSARD, F., KREFT-JAIS, C., MIREMONT-SALAMÉ, G., BÉGAUD, B. and TUBERT-BITTER, P. (2009). Bayesian pharmacovigilance signal detection methods revisited in a multiple comparison setting. *Stat. Med.* **28** 1774–1792. MR2751597

AHMED, I., DALMASSO, C., HARAMBURU, F., THIESSARD, F., BRÖET, P. and TUBERT-BITTER, P. (2010). False discovery rate estimation for frequentist pharmacovigilance signal detection methods. *Biometrics* **66** 301–309.

AHMED, I., THIESSARD, F., MIREMONT-SALAMÉ, G., HARAMBURU, F., KREFT-JAIS, C., BÉGAUD, B. and TUBERT-BITTER, P. (2012). Early detection of pharmacovigilance signals with automated methods based on false discovery rates. *Drug Saf.* **35** 495–506.

ASPDEN, P., CORRIGAN, J. M., WOLCOTT, J. and ERICKSON, S. M. (2003). *Patient Safety: Achieving a New Standard for Care*. The National Academy Press, Washington, DC.

BAYARRI, M. J. and CASTELLANOS, M. E. (2007). Bayesian checking of the second levels of hierarchical models. *Statist. Sci.* **22** 322–343. MR2416808

BOOCKVAR, K. S., CARLSON LACORTE, H., GIAMBANCO, V., FRIDMAN, B. and SIU, A. (2006). Medication reconciliation for reducing drug-discrepancy adverse events. *Am. J. Geriatr. Pharmacother.* **4** 236–243.

BRENNAN, T., LEAPE, L., LAIRD, N., HEBERT, L., LOCALIO, A., LAWTHERS, A., NEWHOUSE, J., WEILER, P. and HIATT, H. (1991). Incidence of adverse events and negligence in hospitalized patients. Results of the Harvard medical practice study I. *N. Engl. J. Med.* **324** 370–376.

BROWNE, W. J. (2004). An illustration of the use of reparameterisation methods for improving MCMC efficiency in crossed random effect models. *Multilevel Model. Newsl.* **16** 13–25.

DEY, D. K., GELFAND, A. E., SWARTZ, T. B. and VLACHOS, P. K. (1998). A simulation-intensive approach for checking hierarchical models. *TEST* **7** 325–346.

DE IORIO, M., MÜLLER, P., ROSNER, G. L. and MACEACHERN, S. N. (2004). An ANOVA model for dependent random measures. *J. Amer. Statist. Assoc.* **99** 205–215. MR2054299

DUMOUCHEL, W. (1999). Bayesian data mining in large frequency tables, with an application to the FDA spontaneous reporting system. *Amer. Statist.* **53** 177–190.

ESCOBAR, M. D. and WEST, M. (1995). Bayesian density estimation and inference using mixtures. *J. Amer. Statist. Assoc.* **90** 577–588. MR1340510

EUGÉNE, V. P., EGBERTS, A., HEERDINK, E. R. and LEUFKENS, H. (2000). Detecting drug–drug interactions using a database for spontaneous adverse drug reactions: An example with diuretics and non-steroidal anti-inflammatory drugs. *Eur. J. Clin. Pharmacol.* **56** 733–738.

FERGUSON, T. S. (1973). A Bayesian analysis of some nonparametric problems. *Ann. Statist.* **1** 209–230. MR0350949

FERNÁNDEZ, C. and STEEL, M. F. J. (1998). On Bayesian modeling of fat tails and skewness. *J. Amer. Statist. Assoc.* **93** 359–371. MR1614601

FLEGAL, J. M., HARAN, M. and JONES, G. L. (2008). Markov chain Monte Carlo: Can we trust the third significant figure. *Statist. Sci.* **23** 250–260. MR2516823

GBD 2013 MORTALITY AND CAUSES OF DEATH COLLABORATORS (2015). Global, regional, and national age–sex specific all-cause and cause-specific mortality for 240 causes of death, 1990–2013: A systematic analysis for the Global Burden of Disease Study 2013. *Lancet* **385** 117–171.

GELMAN, A. and HILL, J. (2007). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge Univ. Press, Cambridge.

GELMAN, A., MENG, X.-L. and STERN, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Statist. Sinica* **6** 733–807. MR1422404

GELMAN, A., GOEGEBEUR, Y., TUERLINCKX, F. and VAN MECHELEN, I. (2000). Diagnostic checks for discrete data regression models using posterior predictive simulations. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **49** 247–268. MR1821324

GELMAN, A., CARLIN, J. B., STERN, H. S., DUNSON, D. B., VEHTARI, A. and RUBIN, D. B. (2014). *Bayesian Data Analysis*, 3rd ed. CRC Press, Boca Raton, FL. MR3235677

GIBBONS, R. D. and AMATYA, A. K. (2016). *Statistical Methods for Drug Safety*. CRC Press, Boca Raton, FL.

GIBBONS, R. D., SEGAWA, E., KARABATSOS, G., AMATYA, A. K., BHAUMIK, D. K., BROWN, C. H., KAPUR, K., MARCUS, S. M., HUR, K. and MANN, J. J. (2008). Mixed-effects Poisson regression analysis of adverse event reports: The relationship between antidepressants and suicide. *Stat. Med.* **27** 1814–1833. MR2420347

HAARIO, H., SAKSMAN, E. and TAMMINEN, J. (2001). An adaptive Metropolis algorithm. *Bernoulli* **7** 223–242. MR1828504

HUANG, X. (2011). Detecting random-effects model misspecification via coarsened data. *Comput. Statist. Data Anal.* **55** 703–714. MR2736589

HUCKELS-BAUMGART, S. and MANSER, T. (2014). Identifying medication error chains from critical incident reports: A new analytic approach. *J. Clin. Pharmacol.* **54** 1188–1197.

KOHN, L. T., CORRIGAN, J. and DONALDSON, M. S. (2000). *To Err Is Human*: *Building a Safer Health System*. The National Academy Press, Washington, DC.

THE JOINT COMMISSION (2000). Reporting of Medical/Health Care Errors: A Position Statement of the Joint Commission.

KWAN, J. L., LO, L., SAMPSON, M. and SHOJANIA, K. G. (2013). Medication reconciliation during transitions of care as a patient safety strategy: A systematic review. *Ann. Intern. Med.* **158** 397–403.

KYUNG, M., GILL, J. and CASELLA, G. (2010). Estimation in Dirichlet random effects models. *Ann. Statist.* **38** 979–1009. MR2604702

LEAPE, L., BRENNAN, T., LAIRD, N., LAWTHERS, A., LOCALIO, A., BARNES, B., HEBERT, L., NEWHOUSE, J., WEILER, P. and HIATT, H. (1991). The nature of adverse events in hospitalized patients. Results of the Harvard medical practice study II. *N. Engl. J. Med.* **324** 377–384.

LEE, K. J. and THOMPSON, S. G. (2008). Flexible parametric models for random-effects distributions. *Stat. Med.* **27** 418–434. MR2418453

LOUIS, T. A. and SHEN, W. (1999). Innovations in Bayes and empirical Bayes methods: Estimating parameters, populations and ranks. *Stat. Med.* **18** 2493–2505.

MACEACHERN, S. N. and MÜLLER, P. (1998). Estimating mixture of Dirichlet process models. *J. Comput. Graph. Statist.* **7** 223–238.

MACEACHERN, S. N. and PERUGGIA, M. (2000). Importance link function estimation for Markov chain Monte Carlo methods. *J. Comput. Graph. Statist.* **9** 99–121. MR1819867

MADIGAN, D., RYAN, P., SIMPSON, S. and ZORYCH, I. (2010). Bayesian methods in pharmacovigilance. In *Bayesian Statistics*, *Vol.* 9 (J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. M. Smith and M. West, eds.). Oxford Univ. Press, London.

MORLOCK, L., DOMINICI, F., MYERS, J. A., SHORE, A. D., PRONOVOST, P. J., DY, S. M. and COUSINS, D. D. (2010). Comparing near miss and harmful medication errors: Testing the causal continuum hypothesis using data from the MEDMARX National Reporting System, Technical report, Johns Hopkins Univ., Baltimore, MD.

MUELLER, S. K., SPONSLER, K. C., KRIPALANI, S. and SCHNIPPER, J. L. (2012). Hospital-based medication reconciliation practices: A systematic review. *Arch. Intern. Med.* **172** 1057–1069.

MÜLLER, P., QUINTANA, F. A., JARA, A. and HANSON, T. (2015). *Bayesian Nonparametric Data Analysis*. Springer, Cham. MR3309338

NATIONAL COORDINATING COUNCIL FOR MEDICATION ERROR REPORTING AND PREVENTION (NCC MERP) (2001). *Medication Error Index*.

PRONOVOST, P., WEAST, B., SCHWARZ, M., WYSKIEL, R. M., PROW, D., MILANOVICH, S. N., BERENHOLTZ, S., DORMAN, T. and LIPSETT, P. (2003). Medication reconciliation: A practical tool to reduce the risk of medication errors. *J. Crit. Care* **18** 201–205.

RAMJAUN, A., SUDARSHAN, M., PATAKFALVI, L., TAMBLYN, R. and MEGUERDITCHIAN, A. N. (2015). Educating medical trainees on medication reconciliation: A systematic review. *BMC Med. Educ.* **15** 33.

SANTELL, J. P., HICKS, R. W., MCMEEKIN, J. and COUSINS, D. D. (2003). Medication errors: Experience of the United States Pharmacopeia (USP) MEDMARX Reporting System. *J. Clin. Pharmacol.* **43** 760–767.

SCHIFF, G. D., AMATO, M. G., EGUALE, T., BOEHNE, J. J., WRIGHT, A., KOPPEL, R., RASHIDEE, A. H., ELSON, R. B., WHITNEY, D. L., THACH, T.-T., BATES, D. W. and SEGER, A. C. (2015). Computerised physician order entry-related medication errors: Analysis of reported errors and vulnerability testing of current systems. *BMJ Qual. Saf.* **24** 264–271.

SHEN, W. and LOUIS, T. A. (1998). Triple-goal estimates in two-stage hierarchical models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **60** 455–471. MR1616061

SINHARAY, S. and STERN, H. S. (2003). Posterior predictive model checking in hierarchical models. *J. Statist. Plann. Inference* **111** 209–221. MR1955882

STERN, H. S. and CRESSIE, N. (2000). Posterior predictive model checks for disease mapping models. *Stat. Med.* **19** 2377–2397.

TCHETGEN, E. J. and COULL, B. A. (2006). A diagnostic test for the mixing distribution in a generalised linear mixed model. *Biometrika* **93** 1003–1010. MR2285086

VENTURINI, S., FRANKLIN, J. M., MORLOCK, L. and DOMINICI, F. (2017a). Supplement to "Random effects models for identifying the most harmful medication errors in a large, voluntary reporting database." DOI:10.1214/16-AOAS974SUPPA.

VENTURINI, S., FRANKLIN, J. M., MORLOCK, L. and DOMINICI, F. (2017b). Supplement to "Random effects models for identifying the most harmful medication errors in a large, voluntary reporting database." DOI:10.1214/16-AOAS974SUPPB.

VENTURINI, S., FRANKLIN, J. M., MORLOCK, L. and DOMINICI, F. (2017c). Supplement to "Random effects models for identifying the most harmful medication errors in a large, voluntary reporting database." DOI:10.1214/16-AOAS974SUPPC.

WAAGEPETERSEN, R. (2006). A simulation-based goodness-of-fit test for random effects in generalized linear mixed models. *Scand. J. Stat.* **33** 721–731. MR2300912

S. VENTURINI
CERGAS BOCCONI
UNIVERSITÀ COMMERCIALE LUIGI BOCCONI
VIA RÖNTGEN 1
MILANO 20136
ITALY
E-MAIL: sergio.venturini@unibocconi.it

J. M. FRANKLIN
DEPARTMENT OF MEDICINE
BRIGHAM AND WOMEN'S HOSPITAL
BOSTON, MASSACHUSETTS 02115
USA
AND
HARVARD MEDICAL SCHOOL
BOSTON, MASSACHUSETTS 02120
USA
E-MAIL: jmfranklin@partners.org

L. MORLOCK
DEPARTMENT OF HEALTH POLICY AND MANAGEMENT
JOHNS HOPKINS UNIVERSITY
BALTIMORE, MARYLAND 21205
USA
E-MAIL: lmorloc1@jhu.edu

F. DOMINICI
DEPARTMENT OF BIOSTATISTICS
HARVARD TH CHAN SCHOOL OF PUBLIC HEALTH
BOSTON, MASSACHUSETTS 02120
USA
E-MAIL: fdominic@hsph.harvard.edu