Are (the log-odds of) Hospital Mortality Rates Normally Distributed in Ontario? Implications for Studying Variations in Outcomes of Medical Care

Peter C Austin, Institute for Clinical Evaluative Sciences
Are (the log-odds of) hospital mortality rates normally distributed? Implications for studying variations in outcomes of medical care

Peter C. Austin PhD

Senior Scientist, Institute for Clinical Evaluative Sciences, Department of Public Health Sciences, and Department of Health Management, Policy and Evaluation, University of Toronto, Toronto, ON, Canada

Abstract

Rationale Hierarchical regression models are increasingly being used to examine variations in outcomes following the provision of medical care across providers. These models frequently assume a normal distribution for the provider-specific random effects. The appropriateness of this assumption for examining variations in health care outcomes has never been explicitly tested.

Aims and Objectives To compare hierarchical logistic regression models in which the provider-specific random effects were either a normal distribution or a mixture of three normal distributions.

Methods We used data on 18,825 patients admitted to 109 hospitals in Ontario with a diagnosis of acute myocardial infarction. We used the Deviance Information Criterion, Bayes factors and predictive distributions to compare the evidence between the two competing models.

Results There was strong evidence that the distribution of hospital-specific log-odds of mortality was a mixture of three normal distributions compared to the evidence that it was normal. In some scenarios, the hospital-specific posterior tail probabilities of unacceptably high mortality were lower when a logistic-normal model was fit compared to when a logistic-mixture of normal distributions model was fit. Additionally, in these same scenarios, fewer hospitals were classified as having higher than acceptable mortality when the logistic-mixture of three normal distributions was used.

Conclusions These findings have important consequences for those who use hierarchical models to examine variations in outcomes of medical care across providers since the mixture of three normal distributions model indicated that variations in outcomes across providers was greater than indicated by the logistic-normal model.

Introduction

Hierarchical regression models are increasingly being used to examine variations in outcomes following provision of medical care across providers (doctors or hospitals) or across regions (states or counties). Bayesian hierarchical models have been used by Normand et al. [1] and Merlo et al. [2] for studying variations in hospital-specific mortality rates following acute myocardial infarction (AMI) and heart failure respectively and by Burgess et al. [3] for examining variations in intensive care unit readmission rates. Similarly, Gatsonis et al. used a Bayesian hierarchical regression model to analyse inter-state variability in access to coronary angiography following AMI [4,5]. Miller et al. examined the use of Empirical Bayes methods to estimate doctor charges [6]. Both Goldstein and Spiegelhalter [7] and DeLong et al. [8] used random effects models for analysing variations in mortality following coronary artery bypass graft surgery across providers. Leyland and Boddy used a random effects logistic regression model to analyse variations in AMI mortality rates across hospitals [9], while Langford and Bentham used random effects models to examine regional variability in mortality rates [10]. Advantages to models incorporating random provider effects include the elimination of regression to the mean bias and that providers with small numbers of patients can be included [1]. Furthermore, some investigators have suggested that the use of Bayesian methods allow profiling to be guided by clinical rather than statistical criteria [1,11].
It has been said, ‘Everyone believes in the normal law, the experimenters because they imagine it a mathematical theorem, and the mathematicians because they think it an experimental fact’ [12]. The majority of the random effects models discussed above assumed that the provider-specific random effects followed a normal distribution. However, the validity and implications of this assumption for modelling outcomes of cardiac care has never been explicitly examined. The objective of the current study was twofold: first, to explicitly examine whether, in 1 year in one jurisdiction, hospital-specific random effects for 30-day mortality were normally distributed; second, to examine the impact of model choice on provider profiling – the identification of hospitals with higher than acceptable mortality. To do so, we examined a range of Bayesian model selection methods. We examined the use of the Deviance Information Criterion (DIC), Bayes factors and predictive modelling to compare a hierarchical logistic regression model in which the provider-specific random effects followed a normal distribution with a model in which the random effects followed a mixture of three normal distributions.

Methods

Data sources

The data for the current study consisted of all 18 825 patients discharged from hospital with a most responsible diagnosis of AMI (International Classification of Disease ninth Revision, (ICD-9) code 410) between 1 April 2000 and 31 March 2001 from the 109 acute care hospitals in Ontario, Canada with an AMI volume of at least 30 patients over the time of the study. The creation of this Ontario Myocardial Infarction Database, a linked population-based administrative database is described elsewhere [13,14]. Adjustments for differences in case-mix were done by using the Ontario AMI mortality prediction rule for 30-day mortality [15]. The variables comprising the prediction rule consist of age, gender, cardiac severity (e.g. congestive heart failure, cardiogenic shock, cardiac dysrhythmias and pulmonary oedema) and co-morbid status (e.g. diabetes mellitus with complications, stroke, acute and chronic renal disease and malignancy), as derived from the ICD-9 codes present in the 15 secondary diagnostic fields of the hospitalization database. The derivation of the prediction rule in Ontario and its subsequent validation in California and Manitoba are described elsewhere [15].

Hierarchical models for provider performance

Each patient had an observed outcome (i.e. death within 30 days of hospital admission or survival to 30 days following admission) and a vector of demographic and risk factors that will be used for case-mix adjustment. Let $X_i$ denote a vector of demographic and co-morbidity variables for the $i^{th}$ patient treated at the $j^{th}$ hospital. This vector comprises age, gender and the nine co-morbidities that form the Ontario AMI mortality prediction model. Age was treated as a continuous variable that was centred around the cohort average, while the remaining predictor variables were dichotomous. Let $P_{ij}$ denote the probability of death within 30 days of admission for the $i^{th}$ patient treated at the $j^{th}$ hospital.

We examined two hierarchical logistic regression models to predict mortality within 30 days of hospital admission. These models incorporate random hospital effects and account for the possible homogeneity in outcomes for patients treated at the same hospital. The random effects model is characterized as

$$\logit(P_{ij}) = \beta_{0j} + \beta_{j}X_{i}$$

where $\beta_{0j} \sim F$, and where $F$ denotes a specified probability distribution. We examined two distributions for the hospital-specific random effects: first, a normal distribution, and second, a mixture of three normal distributions.

We considered other distributions for the hospital-specific random effects. In particular, we considered $t$ distributions with low degrees of freedom. However, the conventional Student’s $t$ distribution has variance fixed as a function of its degrees of freedom [16]. Furthermore, Student’s $t$ distributions are symmetric, and we wanted to allow for the possibility of asymmetry in the distribution of the hospital-specific log-odds of death. It is possible that there would be heavier tails on the upper end of the distribution than on the lower end of the distribution, because mortality rates are bounded below by 0% and are bounded above by 100%, while the model-based mortality rate for a reference patient was 7.3% in 1999, the year prior to our study (see section Prior distributions under Methods). We chose not to examine mixture of normal distributions with more than three components because of the computational burden of fitting these models using Markov Chain Monte Carlo (MCMC) methods. Because of the length of the manuscript, and because the mixture of two normal distributions is a sub-model of the mixture of three normal distributions, we did not examine the use of a logistic-mixture of two normal distributions hierarchical regression model.

Logistic-mixture of normal distributions hierarchical model

While many phenomena can be modelled probabilistically by a simple probability distribution (e.g. normal, log-normal, poisson and binomial), there are many phenomena that cannot easily be modelled probabilistically by a simple probability distribution. One solution is to choose a more elaborate parametric model. One can approximate any distribution by a mixture of simple component distributions. Given a phenomenon that follows a probability distribution $f(x)$, then one can approximate the probability distribution as $f(x) \equiv \sum_{i=1}^{k} p_{i}f_{i}(x|\theta)$. Here $f_{i}(x|\theta)$, the $i^{th}$ component of the mixture, is a simple probability distribution with parameter vector $\theta_{i}$, for $i = 1, \ldots , k$. The scalar $p_{i}$ is the $i^{th}$ mixing proportion, the probability that a given observation comes from the $i^{th}$ component of the mixture. The mixing proportions are such that $\sum_{i=1}^{k} p_{i} = 1$. If the number of components is finite, such a probability model is called a finite mixture model. Many complex phenomena can be modelled probabilistically as a finite mixture of a small number of simple distributions [17].

We examined a hierarchical model in which the hospital-specific random effects were assumed to follow a mixture of three normal distributions. Given a finite mixture model $\sum_{i=1}^{k} p_{i}f_{i}(x|\mu_{i},\sigma_{i})$, with three components, and where $\phi_{i}$ is the normal density function, we use a modification of a parameterization suggested...
Are hospital mortality rates normally distributed?

by Mengersen and Robert [18–20]. We define two positive shift parameters \( \theta_1 \) and \( \theta_2 \) (\( \theta_i > 0 \) for all \( i \)). We parameterize the means of the normal components so that their location is defined relative to the mean of the first normal component. We then parameterize the model so that

\[
\mu_1 = \mu_i, \\
\mu_2 = \mu_i + \theta_1, \\
\mu_3 = \mu_i + \theta_1 + \theta_2.
\]

We replace the standard deviation (SD) of the normal distribution by the precision, \( \tau_i \), defined as \( 1/\sigma_i^2 \). Thus a logistic-mixture of three normals contains the following parameters: \( \mu_i, \theta_1, \theta_2, \tau_1, \tau_2, \tau_3, p_1, p_2, p_3 \) and \( \beta \), where \( \beta \) is the vector of regression coefficients. The conventional logistic-normal model contained the following parameters: \( \mu_i, \tau_i \), and \( \beta \).

**Model estimation**

Markov Chain Monte Carlo methods were used to compute the posterior probability distributions of the model parameters [21]. Proper prior distributions were assumed for each parameter in each model. Because of the known sensitivity of Bayes factors to the prior distribution, we examined four different sets of prior distributions for each model. The four different sets of prior distributions are described in section Prior distributions under Methods.

Three parallel MCMC chains were constructed for each model and for each set of prior distributions by using sets of initial values drawn from an over-dispersed distribution. For each model, and for each set of prior distributions, an initial number of ‘burn-in’ iterations were allowed in order to allow the process to converge to the posterior distribution. The number of burn-in iterations for the logistic-normal hierarchical was 55 000, while the number of burn-in iterations for the logistic-mixture of three normal distributions was 230 000. The Gibbs samplers for the logistic-normal models were then monitored for an additional 10 000 iterations. For the mixture models, once the Gibbs sampler had achieved stationarity, the Gibbs samplers were then monitored for an additional 100 000 iterations, with a thinning interval of 100 – resulting in 1000 iterations for analysis from each of the three parallel chains (previous work with complex mixture models had indicated that a high thinning interval was necessary [22]). Convergence of the Gibbs sampler was assessed by using the Gelman–Rubin criterion for parallel runs from a Gibbs sampler [23]. Using the above methods, the Gelman and Rubin shrink factors for the median and 97.5th percentiles for all model parameters were all at most 1.01 and 1.05 respectively. All subsequent analyses were conducted by using the monitored samples from each of the three parallel chains combined. The Gibbs sampler was implemented by using the BUGS software package [24].

**Prior distributions**

A conventional logistic regression model was fit to comparable data from 1999 (the year prior to our study) by using maximum likelihood estimation. The model contained variables for age, gender and the nine variables contained in the Ontario AMI mortality prediction model: the same variables that would be used in the hierarchical logistic regression models. The parameter estimates along with the standard errors of these estimates for the intercept, and the 11 regression parameters were noted. We constructed four sets of prior distributions. The first set consisted of prior distributions that were felt to be the most likely, whereas the other three sets were used for purposes of sensitivity analyses.

The prior distributions for the regression parameters for age, gender and each of the nine variables in the Ontario AMI mortality model were based upon the logistic regression model fit by using 1999 data (the year prior to the current study). The prior distributions for the regression parameters were assumed to be normal with mean equal to estimated regression coefficient and with SDs equal to the standard error of estimated regression coefficient.

The prior distribution for the mean of the distribution of the random effects in the logistic-normal model was taken to be a normal distribution with mean equal to the estimated intercept in 1999 and SD equal to the standard error of this estimated regression coefficient. The estimated intercept in the conventional logistic regression model estimated by using 1999 data was \(-2.54\). This implies that the 30-day mortality rate among male patients of average age with no co-morbidities was 7.3%. We believed that it was likely that the true 30-day mortality rate for this set of reference patients lay between 2% and 18%. This corresponds to placing an \( N(-2.7, SD = 0.6) \) prior distribution on the model intercept for the logistic-normal hierarchical model. The prior distribution for \( \tau \), the precision (inverse of the variance) of the random effects distribution was chosen to be \( \Gamma(0.05, 0.05) \), where \( \Gamma \) denotes the gamma distribution. Thus, the mean of the prior distribution on the precision is 1. A mean of 1 would imply that most random effects lay within 4 of each other (95% of the mass of the normal distribution lies within \( \pm 2 \) SDs of the mean). However, by using a \( \Gamma(0.05, 0.05) \) distribution, we are also allowing the true SD to lay between 0.08 and 1.4E45 with probability 0.95. Thus, there is a wide range of plausible SDs for the random effects.

For the logistic-mixture of three normal distributions, the prior distribution for \( \mu_i \) was chosen to be \( N(-3.7, SD = 1) \). This choice was motivated by similar arguments used for the prior on the mean of the distribution of the random effects for the logistic-normal model. Our belief was that the three normal components would have different means, and that the mean of the first normal component would be below that of the single normal distribution in the logistic-normal model. The prior distributions for the two positive shift parameters \( \theta_i \) were chosen to be truncated versions of \( N(0, SD = 1.3) \). This distribution has mean 1.04 and SD 1.26 [25]. Thus, the mean of the second component will on average be approximately \(-2.7 \), and the mean of the third component will on average be approximately \(-1.7 \). As in the logistic-normal model, the prior distributions for the precisions of the normal components were assumed to be \( \Gamma(0.05, 0.05) \). Finally, the mixing probabilities for the mixture models were assumed to follow a uniform Dirichlet distribution.

The primary prior distributions are described above. We also examined three other sets of prior distributions. Prior distributions for the parameters of each distribution of random effects are described in Table 1. In each of the three sensitivity analyses, the prior distributions for the 11 regression coefficients were normal distributions with means equal to the maximum likelihood estimate of the coefficient and with SDs equal to the corresponding standard errors.
Bayesian methods of model selection

Deviance Information Criterion

Spiegelhalter et al. proposed the DIC as a Bayesian measure of model complexity and fit [26]. Given data y, and parameters θ, the model deviance is computed at each iteration of a long MCMC run: \( D(\theta^*) = D(\theta^t) \) at iterations \( t = 1, \ldots, T \). Let \( D(\bar{\theta}) \) denote the deviance evaluated at the posterior mean of the parameters. Then the effective number of parameters is defined as:

\[
p_e = \frac{1}{T} \sum_{t=1}^{T} D(\theta^t) - D(\bar{\theta})
\]

Then, the DIC is defined as \( \text{DIC} = D(\bar{\theta}) + 2p_e \). Spiegelhalter et al. suggest that models having DIC values within 1–2 of the model with the lowest DIC deserve consideration, while models having DIC values within 3–7 of the model with the lowest DIC have considerably less support.

Bayes factors

Given two competing models \( M_1 \) and \( M_2 \), and observed data \( D \), one can compute the Bayes factor \( B_{21} \) for model 2 against model 1. Given a probability model \( M \), the quantity \( P(\text{DIM}) \) is called the marginal likelihood of the model and is equal to the integral of the product of the likelihood function of the model and the prior probability distribution of the model parameters. We define \( B_{21} = P(\text{DIM}_2)/P(\text{DIM}_1) \). That is, the Bayes factor is the ratio of the marginal likelihood for the second model to the marginal likelihood of the first model.

We used the compound Laplace-Metropolis estimator of the logarithm of the marginal likelihood, proposed by Lewis and Raftery for models incorporating random effects [27]. Let \( P \) denote the number of model parameters and \( m \) denote the number of hospitals. We define \( \eta \) to be the vector of fixed effects parameters and \( \Sigma \) to be the vector of hyperparameters for the distribution of the random effects. We define \( \theta = (\eta, \Sigma) \). We denote by \( \theta^m \) the posterior mode of \( \theta \), which we estimated using the L1 centre of \( (\eta, \Sigma) \), samples of which were obtained by using the Gibbs sampler. Letting

\[
h(\theta) = \log \{ f(\theta) f(Y | \theta) \}
\]

we define \( H^* \) to be minus the Hessian of \( h \) evaluated at \( \theta^m \).

Then the compound Laplace-Metropolis estimator is defined as:

\[
LM_m = \frac{1}{2} \log(2\pi) + \frac{1}{2} \log \{|H^*|\} + \log \{ f(\theta^m) \} + \sum_{i=1}^{m} \hat{L}_i
\]

Where

\[
\hat{L}_i = \log \{ \int f(Y_i | \alpha_i, \eta_i, \Sigma) f(\alpha_i | \eta_i, \Sigma) d\alpha_i \}
\]

Here \( f(\alpha_i | \eta_i, \Sigma) \) denotes the conditional density function of the random effects. Finally,

\[
f(Y_i | \eta_i, \alpha_i) = \prod_{r=1}^{\infty} \exp \{ X_i \eta_i + \alpha_i \} \frac{1}{1 + \exp \{ X_i \eta_i + \alpha_i \}}
\]

where \( Y_i \) is a vector of observed outcomes taking on only the values 0 or 1 (1 denoting death within 30 days, and 0 denoting survival to 30 days) for the patients treated at the \( i \)th hospital.

We evaluated the integral in (5) numerically, using the ‘integrate’ function in R [28].

Kass and Raftery [29] provide the following guidelines for interpreting the evidence for a given model, against that for a competing model, using Bayes factors. A Bayes factor of less than 1 denotes negative evidence for \( M_2 \). A Bayes factor between 1 and 3 denotes evidence that is barely worth mentioning for \( M_2 \). A Bayes factor between 3 and 20 denotes positive evidence for \( M_2 \). A Bayes factor between 20 and 150 denotes strong evidence for \( M_2 \). Finally, a Bayes factor greater than 150 denotes very strong evidence for \( M_2 \).

Predictive distributions

Bernardo and Smith argue that the fundamental purpose of a model is a ‘predictive probability specification for observables’ [30]. Using this paradigm, one would choose the model that best predicted observable quantities. In our setting, the random effects, which are of primary interest, are latent parameters that are unobservable. However, the number of hospital-specific deaths is an observable that is related to the random effects. Predictive distributions of the number of deaths at each hospital were generated within the MCMC analyses. The mean posterior number of deaths at each hospital was compared with the actual number of deaths at each hospital. Predictive accuracy was summarized in two manners:

Squared prediction error (SPE):

\[
\sum_{i}^{n}(Y_{obs, i} - \bar{Y}_{pred, i})^2
\]
Absolute prediction error (APE): \[ YY_{\text{obs}, j} - YY_{\text{pred}, j} \sum_{j=1}^{109} | \]

Model selection was based on choosing the model with the lowest SPE and APE.

**Bayesian methods for provider profiling**

One objective of provider profiling is to identify hospitals with higher than acceptable mortality rates \[31\]. A proposed Bayesian method of provider profiling is to determine the posterior probabilities of unacceptably high mortality \[1\]. Using each model, for each hospital we computed the probability that the hospital-specific log-odds of death was 50% higher than that of an average hospital. We fit the following model to the data:

\[
\text{logit}(P_X) = \beta_0 + \beta_1 X_i \quad \text{where} \quad F_j \sim F
\]

where \( F \) was either a normal distribution or a mixture of three normal distributions. \( \beta_0 \) denotes the hospital-specific random effect for the \( j \)th hospital. This is the log-odds of mortality for a reference patient (a man of average age with no co-morbidity) at the given hospital. Let \( \beta_0 \) denote the log-odds of death at an average hospital. When \( F \) is a normal distribution, \( \beta_0 \) corresponds to the mean of this distribution, whereas when \( F \) is a mixture of three normal distributions, \( \beta_0 \) corresponds to the weighted average of the means of the three normal components with weights equal to the mixing proportions. Then, for each hospital, we determined:

\[ P(\beta > 1.5\beta_0) \]

Thus, for each hospital, we determined the probability that the odds of mortality were 50% higher than at an average hospital within the supra-population of hospitals. We then assessed the correlation of these posterior tail probabilities between the two classes of models and between different sets of prior distributions within the same class of models.

Our model for hospital profiling does not depend on the nature of the distribution of the hospital-specific random effects, but rather on the relative difference between the odds of death for a reference patient at an average hospital compared with at a specific hospital. The degree of variation permitted by the distribution of the random effects can affect the number of hospitals that are defined to be high-mortality outliers, but it does not alter the definition of what constitutes a high-mortality hospital.

**Results**

**Posterior distribution of random effects**

Figure 1 depicts the posterior distribution of the random effects for each of the two classes of hierarchical models under the four different sets of prior distributions. The posterior normal distribution is relatively insensitive to the assumed prior distributions, whereas the posterior mixture of three normal distributions was more sensitive to the choice of prior distributions. Under all four sets of prior distributions, the mixture of three normal distributions had substantially heavier tails than the simple normal distribution. This indicates that there is greater variability in 30-day mortality between hospitals under the logistic-mixture of three normal distributions model than under the logistic-normal model. The posterior mean of each parameter in the random effects distribution are reported in Table 2.

**Model selection criteria**

**Deviance Information Criterion**

The DIC values for each model and for each set of prior distributions are reported in Table 3. The DIC values for the logistic-normal models demonstrated relatively little variability, ranging from a low of 11 871.7 to a high of 11 877.4. The logistic-normal model under the primary set of prior distributions had the lowest DIC value (11 871.7). The DIC values for the logistic-mixture of three normal distributions models displayed greater variability,
ranging from a low of 11 855.3 to a high of 11 886.7. The model under the primary set of prior distributions had the lowest DIC value (11 855.3). The difference in DIC values between the logistic-normal model and the logistic-mixture of three normal distributions model was 16.4, indicating strong evidence in favour of the logistic-mixture of three normal distributions model. Thus, using the DIC for choosing between competing models, one would choose the regression model in which the hospital-specific random effects followed a mixture of three normal components.

Bayes factors
The marginal log-likelihood for each model and for each set of prior distributions is reported in Table 3. For the logistic-normal hierarchical models, the marginal log-likelihoods ranged from a low of $-5956.0$ to a high of $-5951.5$. When the random effects were assumed to follow a mixture of three normal distributions, the marginal log-likelihoods ranged from a low of $-5959.6$ to a high of $-5948.2$. The choice of prior distribution had a modestly greater impact upon the marginal log-likelihood for the logistic-mixture of three normal distributions models than it did for the logistic-normal models.

Within each class of models, we then chose the model with the highest marginal likelihood and used Bayes factors to compare between competing models. Among the logistic-normal models, the marginal log-likelihood was highest under the second set of prior distributions. Similarly, among the logistic-mixture of three normal distributions models, the marginal log-likelihood was highest under the second set of prior distributions. The Bayes factors comparing the logistic-mixture of three normal distributions model with the logistic-normal hierarchical model was 27.1 ($e^{3.3}$). Thus, using Bayes factors and the guidelines of Kass and Raftery, there was strong evidence for the logistic-mixture of three normal distributions regression model compared with the conventional logistic-normal regression model.

Predictive distributions
Measures of predictive accuracy for each model and for each set of prior distributions are reported in Table 3. SPE for the logistic-normal models ranged from a low of 623.8 to a high of 744.5. SPE for the logistic-mixture of three normal distributions model ranged from a low of 279.6 to a high of 593.2. APE for the logistic-normal regression models ranged from a low of 204.5 to a high of 222.9. APE for the logistic-mixture of three normal distributions hierarchical models ranged from a low of 138.3 to a high of 211.9.

Using predictive accuracy as a criterion for model selection, one would choose the model in which the random effects followed a mixture of normal distributions.
Summary of model selection

Regardless of the model selection method employed, the logistic-mixture of three normal distributions hierarchical model was chosen over the logistic-normal regression model. However, it should be noted that while the logistic-mixture of three normal distributions model was always selected over the logistic-normal regression model, different model selection methods favoured the logistic-mixture of three normal distributions model under different sets of prior distributions. For instance, the logistic-mixture of three normal distributions hierarchical model was selected under the first set of prior distributions by using DIC but was selected under the third set of prior distributions by using SPE and APE.

Impact of assumed distribution of random effects on hospital profiling

We compared the posterior tail probabilities of unacceptable performance within the same class of models assuming different prior distributions as well as between the two classes of hierarchical models. Figure 2 displays plots of the posterior tail probabilities of unacceptable performance within and between the classes of hierarchical regression models. The six plots below the main diagonal display the correlation between hospital-specific probabilities of unacceptable performance under the logistic-normal model under the four different sets of prior distributions. One observes that for the logistic-normal hierarchical model, the choice of prior distribution had at most a negligible impact upon the hospital-specific posterior probability of unacceptable performance. However, under the primary set of prior distributions (priors 1 and the first two sensitivity analyses (priors 2 and 3), it is evident that the posterior probabilities of unacceptable performance were shrunk towards zero when the logistic-normal model was used compared with when the logistic-mixture of three normal distributions model was used.

Under the primary set of prior distributions, the mean hospital-specific posterior probability of unacceptable performance was 0.098 higher under a logistic-mixture of three normal distributions than it was under the logistic-normal model (range from 0.036 to 0.134). However, the mean difference in the hospital-specific probability of unacceptable performance under the logistic-mixture of three normal distributions (second sensitivity analysis) and the logistic-normal model (primary prior distributions) was

logistic-mixture of three normal distributions. One observes that for the logistic-mixture of three normal distributions, the choice of prior distribution had a moderate impact upon the hospital-specific posterior probability of unacceptable performance. The four plots along the main diagonal of Fig. 2 contain plots of the posterior tail probabilities of unacceptable performance under the four logistic-mixture of three normal distributions against those arising from the logistic-normal model under the primary set of prior distributions (these were chosen because they were the most likely model within the logistic-normal class of models by using DIC and predictive distributions; additionally, the choice of prior distribution was shown above to have little impact upon the posterior probabilities of unacceptable mortality). In the third sensitivity analysis (prior 4), the logistic-mixture of three normal distributions model resulted in comparable posterior probabilities of unacceptable performance compared with the logistic-normal regression model. However, under the primary set of prior distributions (priors 1) and the first two sensitivity analyses (priors 2 and 3), it is evident that the posterior probabilities of unacceptable performance were shrunk towards zero when the logistic-normal model was used compared with when the logistic-mixture of three normal distributions model was used.

Figure 2 Posterior tail probabilities (within and between model classes). RE, random effect.
0.038 (range from -0.001 to 0.204). For 25% of hospitals, the difference exceeded 0.064. Thus, there is evidence that the choice of assumed distribution for the provider-specific random effects can result in moderate to substantial changes in the probability of unacceptable for a minority of hospitals. However, the degree to which this occurs is influenced by the choice of prior distributions.

In order to classify hospitals into those that have acceptable mortality and those that have unacceptably high mortality, one can specify a probability threshold. Those hospitals whose posterior probability of unacceptable mortality exceeds this probability threshold are classified as having unacceptably high mortality. A probability threshold of 0.5 has been proposed in the literature [1]. This is equivalent to assuming a 1-0 loss function in which false negatives and false positives are equally penalized [32]. Using this probability threshold, one hospital was classified as having higher than acceptable mortality by using the logistic-normal model under each of the four sets of prior distributions. The number of hospitals classified as having higher than acceptable mortality by using the logistic-mixture of three normal distributions was three, six, eight and one under the four sets of prior distributions respectively. In the one instance in which only one hospital was classified as a high-mortality outlier (prior 4: third sensitivity analysis), the same hospital was classified as a high-mortality outlier when the logistic-normal models were employed.

Discussion

There is an increasing interest in using hierarchical regression models to study variations in outcomes of medical care across providers or regions. These models frequently assume that the provider-specific random effects follow a normal distribution. However, the validity of this assumption has never been explicitly tested. We demonstrated that, in Ontario in 2000, there was greater evidence that the distribution of hospital-specific random effects for 30-day mortality was a mixture of three normal distributions than there was evidence that it was normal. The preference for the mixture of three normal distributions over the normal distribution was made regardless of the model selection method employed.

These findings have important implications for those using hierarchical logistic regression models to study variations in health care outcomes across providers. These models have been used to study variations in both outcomes of care and access to care [1–10]. However, in most instances, a normal distribution for the provider-specific random effects was assumed. Our findings indicate that this assumption may not be realistic, and that, in Ontario in 2000, there was strong evidence that the distribution of hospital-specific random effects had heavier tails than a normal distribution. This implies that variations in health care outcomes may be greater than those indicated by current modelling approaches.

The posterior tail probability of unacceptable mortality has been proposed as a measure of institutional performance [1]. We found that if in Ontario in 2000, the true distribution of log-odds of mortality is assumed to be a mixture of three normal distributions, then hospital-specific probabilities of unacceptable performance may be larger compared with if one assumes a logistic-normal hierarchical model. In particular, if one uses a logistic-normal hierarchical model, hospitals may seem to have a lower probability of unacceptable performance than that is warranted by the data. Thus, the use of an incorrect distribution for the hospital-specific random effects may result in incorrect conclusions about institutional performance for a minority of hospitals. Furthermore, in our data, under three different sets of prior distributions, the number of hospitals identified as having unacceptably high mortality when the logistic-mixture of three normal distributions model was greater than the number of hospitals identified when the logistic-normal model was used for profiling.

Bayesian model selection is an area of active research in the statistical literature. We used three different criteria for model selection: the DIC, Bayes factors and predictive model validation. There are strengths and limitations to each method. Bayes factors provide a formal framework for comparing the relative evidence for competing models. Furthermore, generally agreed upon guidelines have been proposed for quantifying the degree of evidence for one model over a competing model. Similar guidelines are less well developed for the DIC. However, Bayes factors are known to be sensitive to the assumed prior distribution. Furthermore, both Bernardo and Smith [30] and Spiegelhalter et al. [26] argue that Bayes factors are only appropriate in settings in which it is believed that one and only one of the competing models is the true model. In contrast to this, the use of DIC can be used to compare between competing models, at least one of which is adequate, even if none of the models is indeed the true model. While the logistic-mixture of three normal distributions is a flexible hierarchical model, there is no reason to believe that it represents the true model. Thus, in this setting, the use of DIC may be preferable over that of Bayes factors. Predictive validation allows one to directly quantify the predictive ability of each model. However, there are no agreed upon standards to allow one to quantify the degree of evidence for one model over that for a competing model. A limitation to the use of predictive modelling is that the provider-specific random effects are the parameters upon which inference is being made. However, these are latent parameters and are thus not truly observable. Thus, we had to rely on a proxy – the number of deaths at each hospital. This quantity is both observable and is also related to the provider-specific random effect. The reader is referred to Bernardo and Smith [30], who provide a more involved discussion of Bayesian model selection and to Kass and Raftery [29], who provide a more detailed comparison of Bayes factors with Akaike’s Information Criterion and Schwarz’s Bayesian Information Criterion. The preference for the mixture of three normal distributions over the normal distribution was made regardless of the model selection method employed. However, one must note that different sets of prior distributions were favoured by different model selection methods. This suggests that there is a need for improved methods to elucidate prior distributions for this type of hierarchical regression model. The sensitivity of the model selection method to the choice of prior distributions suggests that further research is warranted in this area prior to definitively favouring the logistic-mixture of three normal distributions hierarchical model.

There are certain limitations to the current study. The first limitation is that our data were limited to a single year in one jurisdiction, necessitating that our findings be replicated elsewhere. However, the current study provides a framework within which to compare the relative evidence for different distributions for the provider specific random effects. A second limitation is that we have reported on mortality as the only outcome measure. This decision was made because mortality is one of the most
frequently reported measures in cardiac report cards [1,2,8–11, 33–37]. However, variations in other outcomes, such as access to invasive diagnostic procedures, such as angiography, have also been examined in the literature [4,5]. Because of space and computational limitations, we could not examine whether our results extended to other outcomes. A third limitation to the current study is the impact of prior distributions on the calculated Bayes factors. Bayes factors are known to be sensitive to the assumed prior distributions. We used four sets of prior distributions for each of the hierarchical models examined and demonstrated relatively consistent findings across different sets of prior distributions. Fourth, we did not adjust for hospital-level covariates such as volume of AMI patients, presence of facilities for invasive cardiac procedures or the academic status of the hospital. Including these variables in the regression model would have resulted in hospitals being compared with hospitals with similar characteristics, rather than against all other hospitals [38]. However, arguments have been made that for mortality outcomes, hospitals should be compared with all other hospitals and not against those with similar characteristics [38,39]. Inclusion of provider-level characteristics is rarely done in provider profiling.

We examined the impact of the assumed distribution for the hospital-specific random effects on hospital profiling. In order to do so, we adopted one method of hospital profiling that has been suggested in the statistical literature: the use of posterior tail probabilities of unacceptable performance [1]. Other methods of profiling have been advocated, including methods based on rankings [40] and significance testing [41,42]. We focused on the use of posterior tail probabilities because it is the most commonly discussed Bayesian method in the profiling literature. This method requires that one specify a threshold and determine the posterior probability that the hospital’s mortality exceeds the threshold. We chose a relative threshold: the threshold was based on having a reference patient having an odds of mortality that was 50% greater than at an average hospital. Note that this definition of the threshold is not dependent upon the shape of the distribution of the random effects, but only on the overall mean of the distribution.

In conclusion, there was strong evidence that the distribution of hospital-specific log-odds of 30-day mortality was not normal in Ontario in 2000. In particular, there was strong evidence that the distribution of hospital-specific random effects was a distribution that had heavier tails than a normal distribution. Our conclusions were consistent regardless of whether model selection was based on DIC, Bayes factors or predictive validation. Our findings may have important consequences for all researchers using hierarchical models to examine variations in outcomes of medical care across providers.

Acknowledgements
The Institute for Clinical Evaluative Sciences (ICES) is supported in part by a grant from the Ontario Ministry of Health and Long-Term Care. The opinions, results and conclusions are those of the authors, and no endorsement by the Ministry of Health and Long-Term Care or by the Institute for Clinical Evaluative Sciences is intended or should be inferred. This research was supported in part by an operating grant from the Natural Sciences and Engineering Research Council (NSERC) of Canada. Dr Austin is supported in part by a New Investigator award from the Canadian Institutes of Health Research (CIHR).

References

522


