A comparison of a Bayesian vs. a frequentist method for profiling hospital performance

Peter C. Austin PhD[1,2], C. David Naylor MD, DPhil, FRCPC[1,2,3,4] and Jack V. Tu MD PhD, FRCPC[1,2,3,4]
[1]Institute for Clinical Evaluative Sciences, Toronto, Ontario, Canada
[2]Department of Public Health Sciences, University of Toronto, Toronto, Ontario, Canada
[3]Department of Medicine, University of Toronto, Toronto, Ontario, Canada
[4]Division of General Internal Medicine, Clinical Epidemiology Unit and Health Care Research Program, Sunnybrook and Women's College Health Sciences Centre, Toronto, Ontario, Canada

Correspondence:
Dr Peter Austin
Institute for Clinical Evaluative Sciences
G-160, 2075 Bayview Avenue
North York, Ontario
M4N 3M5
Canada

Keywords: Bayesian statistics, hierarchical models, hospital classification, hospital performance, provider profiling

Accepted for publication:
11 May 2000

Abstract

The objective of this study was to compare the classification of hospitals as outcomes outliers using a commonly implemented frequentist statistical approach vs. an implementation of Bayesian hierarchical statistical models, using 30-day hospital-level mortality rates for a cohort of acute myocardial infarction patients as a test case. For the frequentist approach, a logistic regression model was constructed to predict mortality. For each hospital, a risk-adjusted mortality rate was computed. Those hospitals whose 95% confidence interval, around the risk-adjusted mortality rate, excludes the mean mortality rate were classified as outliers. With the Bayesian hierarchical models, three factors could vary: the profile of the typical patient (low, medium or high risk), the extent to which the mortality rate for the typical patient departed from average, and the probability that the mortality rate was indeed different by the specified amount. The agreement between the two methods was compared for different patient profiles, threshold differences from the average and probabilities. Only marginal agreement was shown between the Bayesian and frequentist approaches. In only five of the 27 comparisons was the kappa statistic at least 0.40. The remaining 22 comparisons demonstrated only marginal agreement between the two methods. Within the Bayesian framework, hospital classification clearly depended on patient profile, threshold and probability of exceeding the threshold. These inconsistencies raise questions about the validity of current methods for classifying hospital performance, and suggest a need for urgent research into which methods are most meaningful to clinicians, managers and the general public.

Introduction

There is an increasing demand for accountability in all health-care systems. Providers are facing increased scrutiny from the government, patients and health-care purchasers or managers. As part of this trend, many jurisdictions have released public report cards, comparing outcomes across hospitals or practice groups. Pennsylvania and California have both released hospital-specific reports for 30-day mortality following admission for acute myocardial infarction (AMI) (Luft *et al*. 1993; Pennsylvania Health Care Cost Containment Council 1996). In Ontario, Canada, report cards comparing 30-day and 1-year AMI mortality rates across all acute care hospitals have also recently been released (Tu *et al*. 1999a).

New York State has published hospital and surgeon-specific report cards for mortality following coronary artery bypass graft (CABG) surgery (New York State Department of Health 1992).

On each report card, those providers were highlighted whose performance was significantly different than expected. Those providers deemed 'high outliers' face public and professional scrutiny over the reasons for deviation from expected performance (Chassin *et al*. 1996; Zinman 1991).

Implicit in the production of such reports is the need to adjust for patient case-mix, so that hospitals treating sicker patients are not penalized unfairly. There has been much discussion in the medical literature on the need for risk-adjustment, and over what variables need to be included in risk-adjustment models (Dubois *et al*. 1988; Iezzoni 1994a, 1994b; Mueller *et al*. 1992; Normand *et al*. 1995; Volpi *et al*. 1993; Suarez *et al*. 1995; The Multicenter Postinfarction Research Group 1983). Iezzoni *et al*. (1995, 1996a, 1996b) have shown that models for classifying outcomes outliers or appraising outcome, at both the institutional and patient levels, show sharp differences in results depending on the severity measures used. All these comparisons have been carried out using traditional frequentist statistical methods, involving fitting a logistic regression model with age, gender and a given measure of disease severity to predict mortality.

Several frequentist methods have been implemented for institutional profiling. One method used frequently in AMI report cards is to compare the ratio of observed to expected mortality at each institution. Hospitals whose ratio differs significantly from one are classified as outliers. A second approach is to compare the difference between observed and expected mortality at each hospital. Those hospitals whose difference is significantly different from zero are classified as performance outliers. A third approach, which has been implemented in the Scottish AMI report cards (Scottish Office 1995), is to model mortality using logistic regression, with indicator variables for each hospital, in addition to patient-level risk factors as regressors. For each hospital, the odds of mortality, relative to the Scottish average, was computed. Those hospitals whose odds of mortality differed significantly from unity were classified as outliers.

DeLong *et al*. (1997) compared the performance of eight different frequentist approaches to hospital profiling for coronary artery bypass grafting surgery in a sample of 28 hospitals. They studied both fixed-effects models and models that incorporated random provider effects. One hospital was labelled as a high outlier by all eight methods. Twelve of the hospitals were not classified as outliers by any of the eight methods. The remaining 15 hospitals were labelled as outliers by at least one method, but not by all methods. All the analyses were carried out from a frequentist perspective.

Leyland & Boddy (1998) compared the performance of a frequentist hierarchical model with that of a logistic regression model with indicator variables for each hospital. They found that the hierarchical modelling approach was much more conservative in labelling hospitals as outliers.

There is a growing interest in the use of Bayesian methods for institutional profiling. However, there is no unique Bayesian approach to this problem. Normand *et al*. (1997) developed a Bayesian hierarchical regression model to model 30-day mortality following acute myocardial infarction. They examined the probabilities of hospital-specific mortality rates exceeding a given threshold. They chose to use thresholds that were determined as functions of the data, but comment that one could alternatively use externally defined thresholds. Similarly, Christiansen & Morris (1997) advocate the use of Bayesian hierarchical regression models. Gatsonis *et al*. (1995) developed a Bayesian hierarchical model to examine geographic variation in access to coronary angiography following acute myocardial infarction. An alternative approach would be to fit hierarchical regression models, and then to rank hospitals according to their risk-adjusted mortality rates. A 95% credible interval can be constructed around each hospital's ranking. Those hospitals whose 95% credible interval lay entirely at the top or bottom quartile of ranks would be labelled as an outlier. This method is described by Marshall & Spiegelhalter (1998). In this paper, we study the approach advocated by Normand *et al*. (1997).

To date much research has been conducted on comparing different severity adjustment measures, in terms of the classification of hospitals or individual patients. These comparisons were all conducted

within the same modelling framework. There has also been extensive research on comparing the performance of different frequentist approaches to hospital classification, for the purpose of classifying hospitals as outliers. Bayesian approaches to institutional profiling have been studied separately in the literature. Given the previous research on comparisons of methods for institutional profiling, the next step is to compare a frequentist and a Bayesian approach to hospital classification. The purpose of this paper is to compare one commonly used frequentist approach to hospital classification, with one Bayesian method for hospital profiling, on a large cohort of patients hospitalized with AMI. We chose the given frequentist approach since it has been implemented in many of the AMI report cards discussed above. We chose a Bayesian method for hospital profiling that has also been used in the literature. This paper presents one of the first comparison of methods for institutional profiling using two different statistical paradigms: Bayesian and frequentist.

Methods

Background to frequentist and Bayesian methods

Given observed data, and a model containing parameters, the likelihood function is the likelihood of observing the given data, conditional on a particular set of parameter values. Both the frequentist and Bayesian approaches to statistical analysis make use of the likelihood function.

The frequentist approach to statistics (Casella & Berger 1990) assumes that the available data are a randomly generated subset from a larger population. Parameters (e.g. means, variances, regression coefficients) are assumed to be fixed but unknown values in the larger population. The analyst seeks to generate estimates of these true, but unknown population parameters, and computes sample statistics accordingly. Frequently, statistical estimation is carried out using maximum likelihood methods (Casella & Berger 1990), where the parameter estimates are those that maximize the likelihood (those parameter values under which the data were most likely to arise). Any statistical test of hypotheses has two components: the null hypothesis (e.g. the mean is zero, the regression slope is zero), and an alternative hypothesis (e.g. the mean is not zero, the regression slope is not zero). A *P*-value is calculated for each statistical test. A *P*-value is the long-term probability of obtaining a test statistic, at least as large as the one observed, if data were to be repeatedly generated under identical conditions from the larger population in which the null hypothesis is true. For a small *P*-value, we reject the null hypothesis. Traditionally, most statistical analyses of report cards have been carried out from a frequentist perspective.

The Bayesian paradigm (Lee 1997) allows one to combine previous beliefs about underlying parameters, with the observed data to obtain probability distributions of the parameters. The Bayesian perspective views both the data, as well as the underlying parameters which generated the data, as random variables. Bayes's theorem provides a method of combining the likelihood function with previous beliefs about the parameters' probability distribution to obtain the posterior probability distribution. The posterior probability distribution is the probability distribution of the unobserved parameters, conditional on the observed data, given one's previous beliefs about this probability distribution. Once the posterior distribution has been determined, one is able to make probabilistic statements about the underlying, unobserved parameters, such as the probability that a given parameter exceeds some threshold.

Traditionally, it has been difficult to develop closed-form expressions of the posterior distribution, except in the simplest of cases. However, with the advent of Markov Chain Monte Carlo (MCMC) methods (Gilks *et al*. 1996), Bayesian methods are being implemented with increasing frequency. MCMC methods are computer-intensive methods that allow one to simulate draws from the posterior distribution, without having to calculate the posterior distribution.

Data sources

Data from the Ontario Myocardial Infarction Database (OMID) (Tu 1999b) were used in this study. Creation of this linked population-based administrative database is described in detail elsewhere (Tu 1999b). For the current study, all 17 818 AMI admis-

sions from 1 April 1996 to 31 March 1997 to the 139 acute care hospitals in Ontario, with an AMI volume of at least 20 patients over the time of the study, were included. Information on patient demographics, comorbidities and 30-day mortality was available for all patients.

The purpose of this paper is not to compare variables in risk-adjustment models, but rather to compare methods for profiling hospitals, once a risk model has been constructed. However, for face validity we include some information on our risk-adjustment model, which is described in more detail elsewhere (Tu 1999a, 1999b). The OMID cohort was created from administrative databases, and thus did not contain clinical variables such as blood pressure, heart rate or type of infarct. Coded comorbidities used in the risk-adjustment model consisted of shock, diabetes with complications, congestive heart failure, cancer, cerebrovascular disease, pulmonary oedema, acute renal failure, chronic renal failure and cardiac dysrhythmias. Age (in four categories) and gender were also entered in the risk-adjustment model. This risk-adjustment model was developed in a 1994–96 AMI cohort. The area under the Receiver Operating Characteristic (ROC) curve (Hanley & McNeil 1982) on the derivation dataset was 0.775. The model was applied to a validation dataset of 1997 AMI patients, and the area under the ROC curve was 0.779.

Statistical models

The frequentist approach to hospital classification that we have chosen, involves fitting a logistic regression model to the entire cohort to model the probability of mortality given the chosen risk factors (DeLong *et al.* 1997). Once this model has been fitted, each patient has a predicted probability of mortality. These probabilities can then be summed up within each hospital to produce the number of deaths that one would expect at this hospital, given its case-mix. The ratio of the observed to expected number of deaths for each hospital is multiplied by the overall cohort mortality rate to produce the risk-adjusted mortality rate. This is interpreted as being the mortality rate that would have been observed at the hospital if its case-mix had been similar to that of the average case-mix in the province. One can construct confidence intervals around the risk-adjusted mortality rate (DeLong *et al.* 1997; Hosmer & Lemeshow 1995). Those hospitals whose confidence intervals exclude the cohort mortality rate are labelled as either high or low outliers. We chose to use the ratio of observed-to-expected mortality since this method has been implemented in most of the report cards discussed in the introduction. However, as noted in the introduction, this is not the only frequentist approach available for hospital profiling. We chose to use 95% confidence intervals in our implementation of the frequentist approach. We used SAS version 6.12 (SAS Institute Inc. 1997) to implement the frequentist logistic regression model.

The Bayesian method that we have chosen to implement involves fitting a Bayesian hierarchical model (Wong & Mason 1985) as advocated by Normand *et al.* (1997). This approach fits a separate regression model to the patients from each hospital. Let p_{ij} denote the probability of mortality for the i^{th} patient at the j^{th} hospital. Let $x_{1ij}, \ldots, x_{kij}$ denote the values of the k predictor variables measured on the i^{th} patient at the j^{th} hospital. We fit the model $\log(p_{ij}/(1-p_{ij})) = \alpha_{0j} + \alpha_{1j}x_{1ij} + \ldots + \alpha_{kj}x_{kij}$. The regression coefficients may be fixed across hospitals, or they may be allowed to vary across hospitals (or a combination of the two). In the case where they vary across hospitals, we assume that $\alpha_{0j} \sim N(\mu_0,\sigma_o), \ldots,$ $\alpha_{kj} \sim N(\mu_k,\sigma_k)$. Here μ_0 is the mean baseline log-odds of mortality for a patient, all of whose covariates equal 0, in the population of hospitals. We define $p_0 = \exp(\mu_0)/(\exp(\mu_0) + (1)$ to be the mean probability of mortality for this patient in the population of hospitals. Similarly, μ_1 is the mean effect of the first predictor in the population of hospitals. By convention, one then assumes diffuse or non-informative priors on the parameters to be estimated. This implies that our prior beliefs about the parameters' probability distribution are vague and imprecise, and that the parameter can assume values over a large range, with approximately equal likelihood. The previous probability distribution for the regression coefficients was assumed to be a normal distribution, with mean zero and a large variance (we chose $\sigma^2 = 1000$). The previous probability distribution for the variance components was assumed to follow a diffuse, inverse gamma distribution. These are standard choices in fitting hierarchical regression

models (Spiegelhalter *et al*. 1996). The model was then estimated using MCMC methods, using the BUGS (Bayesian inference Using Gibbs Sampling) software package (Gilks *et al*. 1994). In our implementation, we fit a random intercept model, where the intercept was allowed to vary across hospitals, while the remaining coefficients were fixed across hospitals. Hence, the hospital-specific mortality rate was allowed to vary across hospitals. By constraining the remaining coefficients to be equal across models, we are assuming that the relationship between a given predictor and mortality is the same at all hospitals. Allowing the effects of predictors to vary across institutions raises the issue of differences in quality of care, which we want to examine through variability in the intercept term, and not through variability in the effect of predictor variables. It also makes the resultant model more difficult to interpret.

With the Bayesian approach, three decisions need to be made: the patient profile, the threshold and the probability of exceeding the threshold. A hospital will be classified as an outlier if the probability of mortality for a specific patient profile exceeds a given threshold with at least a specified probability level. In the Bayesian framework each parameter is assumed to follow a probability distribution, allowing one to make probabilistic statements about parameters or linear combinations of parameters. One can calculate p_j, the probability of mortality for a specified patient, at the j^{th} hospital, and the probability with which this hospital-specific probability of mortality exceeds a given threshold. Hospitals whose probabilities of mortality exceed a given threshold with a certain probability are then classified as outliers. For example, given a threshold level c (here c is the relative proportion that the threshold mortality rate is above or below the average mortality rate for the given profile), and a probability level of 0.5, then a hospital is classified as a high outlier if the probability that $p_j > (1 + c)\ p_0$ is at least 0.5. Similarly, a hospital is classified as a low outlier if the probability that $p_j < (1\text{-}c)\ p_0$ is at least 0.5. In this formulation, p_0 is the mean mortality rate for the given patient profile. This concept is illustrated graphically in Fig. 1, for c = 15%. This can be repeated with different patient profiles, with different thresholds, and with different probabilities of exceeding the threshold. We have chosen our thresholds internally, as a function of the model, using an approach similar to that of Normand *et al*. (1997). However, one can also choose external thresholds. For a given patient profile, one can define the highest acceptable mortality rate. Similarly, one can define what mortality rate constitutes excellent quality of care for a given patient profile and use this for the lower threshold. To put our analysis into a clinical context, the overall 30-day mortality rate in our cohort was 14.7%.

In this paper, we examined three different patient profiles: (1) a male aged 50–64 years, with no additional risk factors, (2) a patient aged 65–74 years, all of whose risk factors are set to the cohort average and (3) a female aged at least 75 years, with chronic renal failure, heart failure and cardiac dysrhythmias. These were termed as low-risk, medium-risk and high-risk patient profiles, respectively. It should be noted that the medium-risk patient is a fictitious patient profile. It represents a person whose risk factors are all set to the cohort average. Therefore, this patient is defined as being 66% male, and similarly for the other risk factors. A similar approach was used by Gatsonis *et al*. (1995). For each patient profile, the probability of mortality at each hospital is calculated based on the fitted model. The chosen thresholds are 10%, 15% and 20% above and below the mean hospital-specific mortality rate for the given patient profile in the population of hospitals. Three different probabilities of exceeding the given threshold were chosen: 0.333, 0.50 and 0.666. Hospitals whose probabilities of mortality lie below the lower threshold with at least the given probability level are classified as low outliers, whereas hospitals whose probabilities of mortality exceed the upper threshold with at least the given probability level are classified as high outliers. The three thresholds were chosen to represent high, moderate and low demands for quality of medical care. We chose our three levels of probability to represent weak, moderate and strong evidence that the hospital had exceeded the given threshold. Thus, we have 27 scenarios for classifying a hospital as an outlier. For a given patient profile, the possible scenarios are described in Table 1.

Using each of the two methods, a hospital was classified as a high outlier, a low outlier or neither. The agreement between the Bayesian and frequentist methods of hospital classification was assessed using

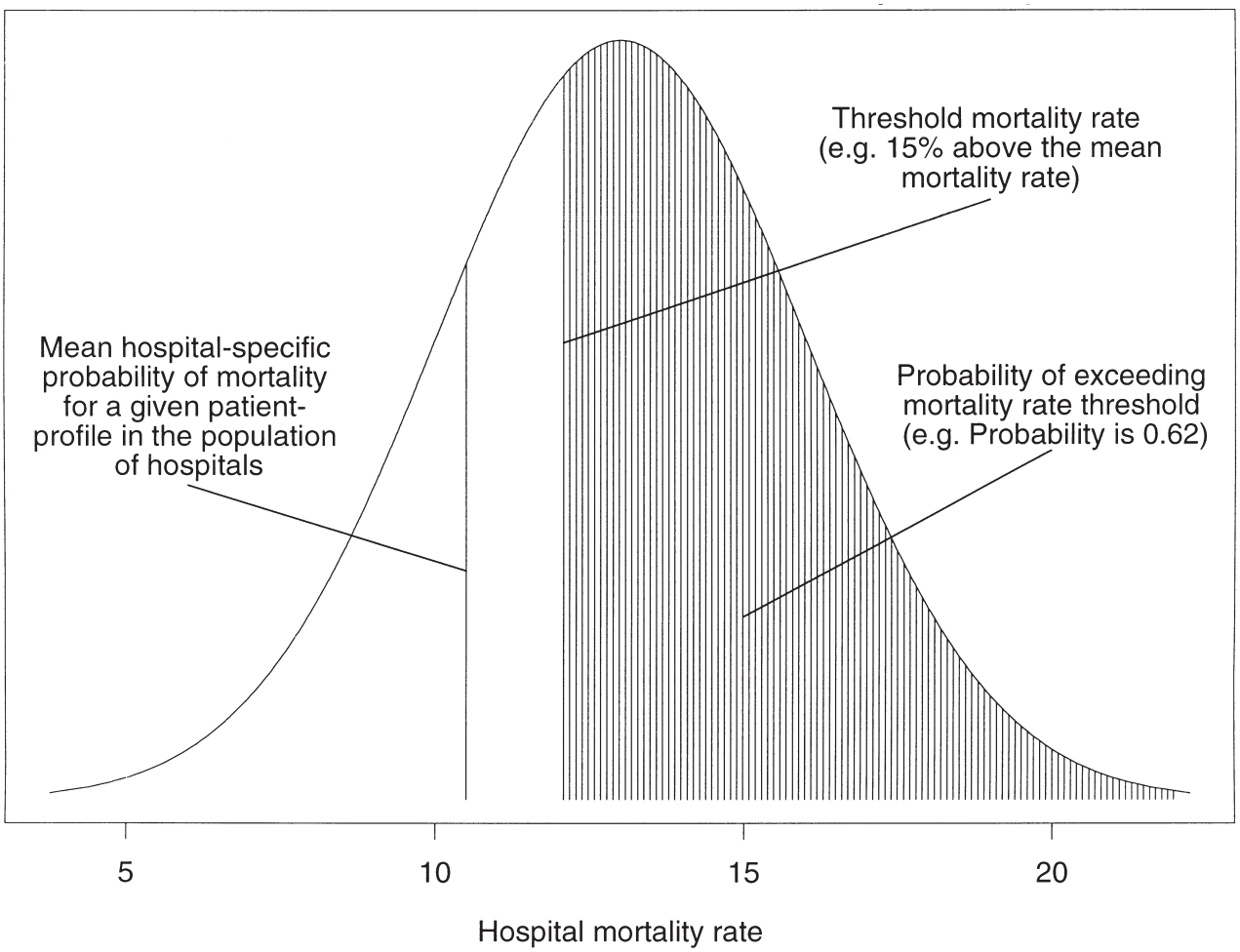

Figure 1 Probability distribution of the hospital-specific mortality rate for a given hospital and for a given patient profile. In the Bayesian paradigm, for each hospital, the mortality rate for a specific patient profile follows a probability distribution. In the above figure, for a given hospital, the mortality rate distribution for a medium risk patient is normally distributed as N(m = 13.0, s = 3). The mean mortality rate for this patient profile in the hospital population is 10.5%. The threshold 15% above the mean mortality rate is 12.1%. The probability that the given hospital's mortality rate exceeds this threshold is 0.62%.

the kappa statistic (Fleiss 1981). We used the convention (Landis & Koch 1977) that a kappa statistic of greater than 0.75 denotes excellent agreement, a kappa statistic between 0.4 and 0.75 denotes good agreement, whereas a kappa statistic of less than 0.4 denotes marginal agreement. We used a weighted kappa, since that allows weighting different levels of disagreement. We used the weights suggested by Cicchetti and Allison (Fleiss 1981). We repeated the analysis using a traditional kappa and obtained similar results.

Results

Either approach classifies each hospital as a low outlier, a high outlier or neither. Table 2 compares the performance of the traditional approach to the Bayesian hierarchical models for a low-risk patient profile. Tables 3 and 4 compare the performance of the traditional approach to the Bayesian hierarchical models for a medium-risk patient profile and a high-risk patient profile, respectively. Within each table, for a given threshold and a given probability of

Table 1 Hospital classification for specific scenarios, for a given patient profile

Threshold	*Probability of exceeding threshold* 0.333	0.50	0.666
10%	Hospital is classified as an outlier if: $P(p_i > 1.1p_o) \geq 0.333$ or $P(p_i < 0.90p_o) \geq 0.333$	Hospital is classified as an outlier if: $P(p_i > 1.1p_o) \geq 0.50$ or $P(p_i < 0.90p_o) \geq 0.50$	Hospital is classified as an outlier if: $P(p_i > 1.1p_o) \geq 0.666$ or $P(p_i < 0.90p_o) \geq 0.666$
15%	Hospital is classified as a outlier if: $P(p_i > 1.15p_o) \geq 0.333$ or $P(p_i < 0.85p_o) \geq 0.333$	Hospital is classified as a outlier if: $P(p_i > 1.15p_o) \geq 0.50$ or $P(p_i < 0.85p_o) \geq 0.50$	Hospital is classified as an outlier if: $P(p_i > 1.15p_o) \geq 0.666$ or $P(p_i < 0.85p_o) \geq 0.666$
20%	Hospital is classified as an outlier if: $P(p_i > 1.2p_o) \geq 0.333$ or $P(p_i < 0.80p_o) \geq 0.333$	Hospital is classified as an outlier if: $P(p_i > 1.2p_o) \geq 0.50$ or $P(p_i < 0.80p_o) \geq 0.50$	Hospital is classified as an outlier if: $P(p_i > 1.2p_o) \geq 0.666$ or $P(p_i < 0.80p_o) \geq 0.666$

p_i is the hospital-specific mortality rate for a given patient profile. p_0 is the mean mortality rate in the population of hospitals for the given patient profile. Interpretation of the first cell: a hospital is classified as a high outlier if the probability is 33.3% or higher that its mortality rate (for the given patient profile) is at least 10% above the mean hospital-specific mortality rate (for the given profile). A hospital is classified as a low outlier if the probability is 33.3% or higher that its mortality rate (for the given patient profile) is at least 10% below the mean hospital-specific mortality rate (for the given profile).

Table 2 A comparison of hospital classification using a frequentist statistical model, and a Bayesian hierarchical model with a low-risk patient profile

	Bayesian hierarchical model: low-risk patient profile								
	10% Threshold			*15% Threshold*			*20% Threshold*		
Frequentist risk model	*Low*	*Neither*	*High*	*Low*	*Neither*	*High*	*Low*	*Neither*	*High*
Probability of exceeding threshold at least 0.333									
Low	7	0	0	6	1	0	1	6	0
Neither	17	88	17	1	118	3	0	122	0
High	0	0	10	0	5	5	0	9	1
		Kappa = 0.45			Kappa = 0.67			Kappa = 0.20	
Probability of exceeding threshold at least 0.5									
Low	4	3	0	0	7	0	0	7	0
Neither	0	122	0	0	122	0	0	122	0
High	0	9	1	0	10	0	0	10	0
		Kappa = 0.44			Kappa = 0			Kappa = 0	
Probability of exceeding threshold at least 0.666									
Low	0	7	0	0	7	0	0	7	0
Neither	0	122	0	0	122	0	0	122	0
High	0	10	0	0	10	0	0	10	0
		Kappa = 0			Kappa = 0			Kappa = 0	

Table 3 A comparison of hospital classification using a frequentist statistical model, and a Bayesian hierarchical model with a medium-risk patient profile

	Bayesian hierarchical model: medium-risk patient profile								
		10% Threshold			*15% Threshold*			*20% Threshold*	
Frequentist risk model	*Low*	*Neither*	*High*	*Low*	*Neither*	*High*	*Low*	*Neither*	*High*
Probability of exceeding threshold at least 0.333									
Low	7	0	0	3	4	0	0	7	0
Neither	10	98	14	0	122	0	0	122	0
High	0	1	9	0	7	3	0	9	1
		Kappa = 0.52			Kappa = 0.51			Kappa = 0.10	
Probability of exceeding threshold at least 0.5									
Low	2	5	0	0	7	0	0	7	0
Neither	0	122	0	0	122	0	0	122	0
High	0	9	1	0	10	0	0	10	0
		Kappa = 0.29			Kappa = 0			Kappa = 0	
Probability of exceeding threshold at least 0.666									
Low	0	7	0	0	7	0	0	7	0
Neither	0	122	0	0	122	0	0	122	0
High	0	10	0	0	10	0	0	10	0
		Kappa = 0			Kappa = 0			Kappa = 0	

Table 4 A comparison of hospital classification using a frequentist statistical model, and a Bayesian hierarchical model with a high-risk patient profile

	Bayesian hierarchical model: high-risk patient profile								
		10% Threshold			*15% Threshold*			*20% Threshold*	
Frequentist risk model	*Low*	*Neither*	*High*	*Low*	*Neither*	*High*	*Low*	*Neither*	*High*
Probability of threshold exceeding at least 0.333									
Low	3	4	0	0	7	0	0	7	0
Neither	0	122	0	0	122	0	0	122	0
High	0	9	1	0	10	0	0	10	0
		Kappa = 0.37			Kappa = 0			Kappa = 0	
Probability of exceeding threshold at least 0.5									
Low	0	7	0	0	7	0	0	7	0
Neither	0	122	0	0	122	0	0	122	0
High	0	10	0	0	10	0	0	10	0
		Kappa = 0			Kappa = 0			Kappa = 0	
Probability of exceeding threshold at least 0.666									
Low	0	7	0	0	7	0	0	7	0
Neither	0	122	0	0	122	0	0	122	0
High	0	10	0	0	10	0	0	10	0
		Kappa = 0			Kappa = 0			Kappa = 0	

exceeding the threshold, are the kappa statistics for assessing agreement between the two methods.

The traditional method of classifying hospitals identified seven low outliers and 10 high outliers (at the 0.05 level). The number of outliers detected by the Bayesian hierarchical models varied with the patient profile, the threshold and the probability of exceeding the threshold.

With the low-risk patient profile, using the 10% threshold and a 0.333 probability of exceeding the threshold, the Bayesian hierarchical models classified 51 hospitals as outliers. In this scenario, the kappa statistic was 0.45. A hospital was more likely to be labelled an outlier using the Bayesian approach than by the frequentist approach.

With the medium-risk patient profile, the Bayesian hierarchical models classified several hospitals as outliers using the 10% and 15% thresholds with a 0.333 probability of exceeding the threshold. The scenarios with the 10% and 15% thresholds produced good agreement (kappa = 0.52 and kappa = 0.51, respectively) with the frequentist approach.

In only one of the nine scenarios involving the high-risk patient profile was the agreement between the two methods any greater than would be expected by chance alone (kappa > 0). In the remaining eight scenarios, no hospitals were identified as outliers using the Bayesian approach.

Overall, only five combinations of patient profile, threshold and probability level produced good agreement with the frequentist approach (0.40 < kappa < 0.75). In no scenarios did kappa exceed 0.75, denoting excellent agreement.

Discussion

We have compared the performance of two statistical paradigms that can be used to assess the outcomes of hospital services. Our test case compared hospital-level 30-day mortality outcomes for acute myocardial infarction in the Canadian province of Ontario, with a Bayesian as contrasted to a frequentist methodology. In most instances, the two approaches produce strikingly different conclusions. In only five instances was there any semblance of agreement between the two methods (kappa > 0.40).

The Bayesian hierarchical approach shrinks each hospital's estimate of mortality for a specific patient profile towards the average mortality rate in the population of hospitals, thus reducing the apparent variability between hospitals, compared to the implementation of a fixed-effects model. Hence, the probability that a hospital lies above or below a given threshold will be small. In 18 of the 27 scenarios examined, no hospitals were classified as outliers by the Bayesian approach.

The frequentist approach classifies hospitals as outliers by computing what their mortality rate would have been, had their case-mix been similar to that of the entire cohort. In contrast to this, the Bayesian hierarchical modelling approach computes the probability that the hospital's mortality rate for a specific patient profile lies above or below some threshold. The Bayesian approach is useful for finding those patient profiles for which there is a real difference between hospitals. The results indicate that using the Bayesian approach it is difficult to classify hospitals as outliers for medium and high-risk patient profiles, compared with the frequentist approach. However, one is able to classify hospitals as outliers for low-risk patient profiles using the Bayesian approach, when both the threshold and the required probability of exceeding the threshold are low. The Bayesian approach may indicate that it is among the low-risk patients that there is the greatest room for improvement in the quality of medical care provided, and thus the greatest variation between hospitals.

The traditional frequentist approach is relatively straightforward, with the only choice being the level of significance required to classify a hospital as an outlier. One drawback to the frequentist approach is its reliance on *P*-values. As such, they represent an artificial comparison – that is, the probability of the data given the hypothesis of no differences between hospitals, rather than the probability of the parameters, given the observed data.

With the Bayesian approach, three decisions need to be made: the patient profile, the threshold and the probability of exceeding the threshold. Each of these choices has an impact on the results, as Tables 2–4 demonstrate. In choosing a threshold, one defines what constitutes both excellence and mediocrity in medical care. The Bayesian approach allows one to quantify uncertainty about a hospital's performance. Differing degrees of uncertainty about performance

can be examined by allowing the required probability of exceeding the threshold to vary. The Bayesian approach also allows each hospital's performance to be examined for specific patient profiles. By examining different patient profiles, one can determine those profiles for which there is the greatest room for improvement in medical care. By computing each hospital's probability of exceeding the threshold defining quality of care, one can judge the strength of the evidence for either excellent or poor quality of medical care. We believe that the Bayesian approach allows one to gain a deeper understanding of how specific hospitals differ from the norm. The Bayesian method allows one to use medically informed criteria for assessing hospital performance. However, the Bayesian approach requires explication to those used to the frequentist paradigm, but may actually be more intuitive to non-statisticians.

Our study has certain limitations. For the Bayesian method, our choices of threshold levels and the required probabilities of exceeding these thresholds for classification as an outlier are somewhat subjective, as are our choices of patient profiles. Other, defensible, choices could also be examined. We limited our choice to three different thresholds, and three different probabilities, to limit the number of scenarios that were examined. In so doing, we demonstrated that classification using the Bayesian hierarchical models depended on each of the patient profile, the threshold and the probability of exceeding the threshold. There are alternative methods that we could have chosen to assess agreement. Using each method, we could have divided the hospitals into quartiles based upon the risk-adjusted mortality rate. We could then have seen how many hospitals changed quartiles when an alternative method of profiling was used.

In conclusion, for most of the scenarios that we examined, there was poor concordance between the Bayesian and frequentist methods. This research was performed on 17 818 patients hospitalized for AMI at 139 hospitals in Ontario. Correctly determining which hospitals deliver either excellent or mediocre medical care is an important clinical issue. We believe that the discordance in determining outlier status highlights the limitations of the current methods of measuring and reporting hospital performance. Each method operates under a different paradigm, and allows one to put different questions to the data. Our findings suggest a need for urgent research into which methods are most meaningful to clinicians, managers and the general public.

Acknowledgements

Dr Naylor was a Medical Research Council of Canada Senior Scientist at the time the project was completed. Dr Tu is supported by a Medical Research Council of Canada Scholar Award. This work was supported in part by an operating grant from the Medical Research Council of Canada. The views expressed herein are solely those of the authors and do not represent the views of any of the sponsoring organizations.

References

Casella G. & Berger R.L. (1990) *Statistical Inference*. Duxbury Press, Belmont, CA.

Chassin M.R., Hannan E.L. & DeBuono B.A. (1996) Benefits and hazards of reporting medical outcomes publicly. *New England Journal of Medicine* **334**, 394–398.

Christiansen C.L. & Morris C.N. (1997) Improving the statistical approach to health care provider profiling. *Annals of Internal Medicine* **127**, 764–768.

DeLong E.R., Peterson E.D., DeLong D.M., Muhlbaier L.H., Hackett S. & Mark D.B. (1997) Comparing risk-adjustment methods for provider profiling. *Statistics in Medicine* **16**, 2645–2664.

Dubois C., Pierard L.A., Albert A., Smeets J., Demoulin J., Boland J. & Kulbertis H.E. (1988) Short-term risk stratification at admission based on simple clinical data in acute myocardial infarction. *American Journal of Cardiology* **61**, 216–219.

Fleiss J.L. (1981) *Statistical Methods for Rates and Proportions*, 2nd edn. Wiley, New York, NY.

Gatsonis C.A., Epstein A.M., Newhouse J.P., Normand S.L. & McNeil B.J. (1995) Variations in the utilization of coronary angiography for elderly patients with an acute myocardial infarction. An analysis using hierarchical logistic regression. *Medical Care* **33**, 625–642.

Gilks W.R., Richardson S. & Spiegelhalter D.J. (1996) Introducing Markov chain Monte Carlo. In *Markov Chain Monte Carlo in Practice* (eds W.R. Gilks, S. Richardson & D.J. Spiegelhalter), pp. 1–19. Chapman & Hall, London.

Gilks W.R., Thomas A. & Spiegelhalter D.J. (1994) A language and program for complex Bayesian modelling. *The Statistician* **43**, 169–178.

Hanley J.A. & McNeil B.J. (1982) The meaning and use of the area under a Receiver Operating Characteristic (ROC) curve. *Radiology* **143**, 29–36.

Hosmer D.W. & Lemeshow S. (1995) Confidence interval estimates of an index of quality performance based on logistic regression models. *Statistics in Medicine* **14**, 2161–2172.

Iezzoni L.I. (1994a) *Risk Adjustment for Measuring Health Care Outcomes*. Health Administration Press, Ann Arbor, MI.

Iezzoni L.I. (1994b) Using risk-adjusted outcomes to assess clinical practice: an overview of issues pertaining to risk adjustment. *Annals of Thoracic Surgery* **58**, 1822–1826.

Iezzoni L.I., Ash A.S., Shwartz M., Daley J., Hughes J.S. & Mackiernan Y. (1995) Predicting who dies depends on how severity is measured: implications for evaluating patient outcomes. *Annals of Internal Medicine* **123**, 763–770.

Iezzoni L.I., Shwartz M., Ash A.S., Hughes J.S., Daley J. & Mackiernan Y. (1996a) Severity measurement methods and judging hospital death rates for pneumonia. *Medical Care* **34**, 11–28.

Iezzoni L.I., Shwartz M., Ash A.S. & Mackiernan Y. (1996b) Predicting in-hospital mortality for stroke patients. *Medical Decision Making* **16**, 348–356.

Landis J.R. & Koch G.G. (1977) The measurement of observer agreement for categorical data. *Biometrics* **33**, 159–174.

Lee P.M. (1997) *Bayesian Statistics: an introduction*, 2nd edn. Arnold, London.

Leyland A.H. & Boddy F.A. (1998) League tables and acute myocardial infarction. *Lancet* **351**, 555–558.

Luft H.S., Romano P.S., Remy L.L. & Rainwater J. (1993) *Annual Report of the California Hospital Outcomes Project*. California Office of Statewide Health Planning and Development, Sacramento, CA.

Marshall E.C. & Spiegelhalter D.J. (1998) Reliability of league tables of *in vitro* fertilisation clinics: retrospective analysis of live birth rates. *British Medical Journal* **316**, 1701–1705.

Mueller H.S., Cohen L.S., Braunwald E., Forman S., Feit F., Ross A., Schweiger M., Cabrin H., Davison R., Miller D., Solomon R. & Knatterud G.L. for the TIMI Investigators. (1992) Predictors of early morbidity and mortality after thrombolytic therapy of acute myocardial infarction: analyses of patient subgroups in the Thrombolysis in Myocardial Infarction (TIMI) Trial, Phase II. *Circulation* **85**, 1254–1264.

New York State Department of Health (1992) *Coronary Artery Bypass Graft Surgery in New York State 1989–91*. New York State Department of Health, Albany, New York.

Normand S.L., Glickman M.E. & Gatsonis C.A. (1997) Statistical methods for profiling providers of medical care: issues and applications. *Journal of the American Statistical Association* **92**, 803–814.

Normand S.T., Morris C.N., Fung K.S., McNeil B.J. & Epstein A.M. (1995) Development and validation of a claims based index for adjusting for risk of mortality: the case of acute myocardial infarction. *Journal of Clinical Epidemiology* **48**, 229–243.

Pennsylvania Health Care Cost Containment Council (1996) *Focus on Heart Attack in Pennsylvania. Research methods and results*. Pennsylvania Health Care Cost Containment Council, Harrisburg, PA.

SAS Institute Inc. (1997) *SAS/STAT Software: changes and enhancements through release 6.12*. SAS Institute Inc., Cary, NC.

Scottish Office (1995) *Clinical Outcome Indicators, 1994*. Clinical Resource and Audit Group, Edinburgh.

Spiegelhalter D.J., Best N.G., Gilks W.R. & Inskip H. (1996) Hepatitis B: a case study in MCMC methods. In *Markov Chain Monte Carlo in Practice* (eds W.R. Gilks, S. Richardson & D.J. Spiegelhalter), pp. 21–43. Chapman & Hall, London.

Suarez C., Herrera M., Vera A., Torrado E., Ferriz J. & Arboleda J.A. (1995) Prediction on admission of in-hospital mortality in patients older than 70 years with acute myocardial infarction. *Chest* **108**, 83–88.

The Multicenter Postinfarction Research Group (1983) Risk stratification and survival after myocardial infarction. *New England Journal of Medicine* **309**, 331–336.

Tu J.V., Austin P., Naylor C.D., Iron K. & Zhang H. (1999a) Acute myocardial infarction outcomes in Ontario. *Cardiovascular Health Services in Ontario: an ICES atlas* (eds C.D. Naylor & P.M. Slaughter), pp. 83–110. Institute for Clinical Evaluative Sciences, Toronto, Canada.

Tu J.V., Naylor C.D. & Austin P. (1999b) Temporal changes in the outcomes of acute myocardial infarction in Ontario, 1992–96. *Canadian Medical Association Journal* **161**, 1257–1261.

Volpi A., De Vita C., Franzosi M.G., Geraci E., Maggioni A.P., Mauri F., Negri E., Santoro E., Tavazzi L. & Tognoni G. (1993) Determinants of 6-month mortality in survivors of myocardial infarction after thrombolysis: Results of the GISSI-2 data base. *Circulation* **88**, 416–429.

Wong G. & Mason W. (1985) The hierarchical logistic regression model for multilevel analysis. *Journal of the American Statistical Association* **80**, 513–524.

Zinman D. (1991) State takes docs' list to heart. *Newsday* December 18, p. 7. New York, NY.