Received: 19 December 2018 | Revised: 9 January 2020 | Accepted: 9 January 2020

DOI: 10.1002/sim.8484

RESEARCH ARTICLE

Statistics in Medicine WILEY

Comparing a multivariate response Bayesian random effects logistic regression model with a latent variable item response theory model for provider profiling on multiple binary indicators simultaneously

Peter C. Austin[1,2,3] | **Douglas S. Lee**[1,2,4,5] | **George Leckie**[6]

[1]ICES, Toronto, Canada

[2]Institute of Health Management, Policy and Evaluation, University of Toronto, Toronto, Canada

[3]Schulich Heart Research Program, Sunnybrook Research Institute, Toronto, Canada

[4]Department of Medicine, University of Toronto, Toronto, Canada

[5]Peter Munk Cardiac Centre and Joint Department of Medical Imaging, and University Health Network, Toronto, Canada

[6]Centre for Multilevel Modeling, University of Bristol, Bristol, UK

Correspondence
Peter C. Austin, ICES, G106, 2075 Bayview Avenue, Toronto, ON M4N 3M5, Canada.
Email: peter.austin@ices.on.ca

Funding information
Canadian Institutes of Health Research, Grant/Award Numbers: CRT43823, CTP79847, MOP 86508; Heart and Stroke Foundation of Canada, Grant/Award Number: Mid-Career Investigator Award; UK Economics and Social Research Council, Grant/Award Number: ES/R010285/1; Ontario Ministry of Health and Long-Term Care; ICES

Provider profiling entails comparing the performance of hospitals on indicators of quality of care. Many common indicators of healthcare quality are binary (eg, short-term mortality, use of appropriate medications). Typically, provider profiling examines the variation in each indicator in isolation across hospitals. We developed Bayesian multivariate response random effects logistic regression models that allow one to simultaneously examine variation and covariation in multiple binary indicators across hospitals. Use of this model allows for (i) determining the probability that a hospital has poor performance on a single indicator; (ii) determining the probability that a hospital has poor performance on multiple indicators simultaneously; (iii) determining, by using the Mahalanobis distance, how far the performance of a given hospital is from that of an average hospital. We illustrate the utility of the method by applying it to 10 881 patients hospitalized with acute myocardial infarction at 102 hospitals. We considered six binary patient-level indicators of quality of care: use of reperfusion, assessment of left ventricular ejection fraction, measurement of cardiac troponins, use of acetylsalicylic acid within 6 hours of hospital arrival, use of beta-blockers within 12 hours of hospital arrival, and survival to 30 days after hospital admission. When considering the five measures evaluating processes of care, we found that there was a strong correlation between a hospital's performance on one indicator and its performance on a second indicator for five of the 10 possible comparisons. We compared inferences made using this approach with those obtained using a latent variable item response theory model.

KEYWORDS

Bayesian analysis, health services research, logistic regression, multilevel data, provider profiling, random effects models

1 | INTRODUCTION

Provider profiling entails comparing the performance of healthcare providers on one or more indicators of quality of care. While provider profiling most frequently compares quality of care indicators across hospitals, comparisons of quality of care across physicians or surgeons are also done. Indicators of healthcare quality include patient outcomes (eg, death, occurrence of surgical complications, hospital length of stay) or measures of processes of care (eg, prescribing of appropriate medication, provision of smoking cessation counselling to patients who are current smokers).

Examples of reporting of patient outcomes include hospital report cards produced by the American states of New York, Pennsylvania, Massachusetts, New Jersey, as well as the Canadian province of Ontario, that reported hospital-specific mortality rates for patients undergoing coronary artery bypass graft surgery.[1-5] Similarly, Pennsylvania, California, and Ontario have publicly reported hospital-specific mortality rates for patients hospitalized with acute myocardial infarction (AMI).[6-8] Examples of reporting on process of care measures include The Enhanced Feedback for Effective Cardiac Treatment (EFFECT) Study that reported hospital-specific rates of prescribing of evidence-based medications for patients hospitalized with AMI or heart failure.[9,10]

Provider profiling has historically focused on single indicators of quality in isolation. Most of the cardiovascular report cards described above focused on mortality as the primary indicator of quality. Those studies that considered multiple quality indicators tended to examine each indicator in isolation. Thus, variation in hospital performance for one indicator is examined separately from the examination of variation in hospital performance for the other indicators. Examination of each indicator in isolation precludes a formal examination of whether specific hospitals perform poorly on multiple indicators or whether hospitals that perform poorly on one indicator tend to perform poorly on a second indicator.

A small number of studies have described statistical methods for summarizing provider performance on multiple indicators of quality of care. The most commonly described methods are based on latent variable models.[11-13] These methods are motivated, at least in part, by item response theory models from the psychometrics literature.[14] These models assume the existence of a single latent, or unmeasured, variable denoting the underlying quality of the healthcare provider. The performance of each hospital on each of the indicators is assumed to be related to this latent variable denoting hospital quality. Estimation of the value of the latent variable for each hospital permits identification of hospitals with excellent or poor quality of care.

The presence of multiple indicators of quality of care suggests that multivariate distributions and models may also be of use for studying variation in hospital performance on a set of multiple indicators. Dunson[15] developed Bayesian latent variable models for clustered mixed outcomes. These models allow for the simultaneous analysis of binary, categorical, and continuous outcomes. O'Malley et al[16] developed models for multivariate outcomes that allowed for the joint modeling of binary and continuous outcomes. While there is a small literature on multivariate models for the simultaneous analysis of multiple outcomes, we are unaware of their previous use in provider profiling.

The objective of the current study is 2-fold. First, to develop Bayesian multivariate response random effects logistic regression models for modeling between-hospital variation in performance on multiple binary indicators simultaneously. Second, to contrast this approach with the latent variable approach similar to those that have been described previously.[11-13] The article is structured as follows: In Section 2 we describe the statistical models to be used. In Section 3 we illustrate the application and interpretation of these methods when applied to a large sample of patients hospitalized with AMI. Finally, in Section 4 we summarize our findings and place them in the context of the literature.

2 | STATISTICAL METHODS FOR PROVIDER PROFILING ON MULTIPLE BINARY INDICATORS

In this section we describe two statistical methods for provider profiling on multiple binary outcomes. We first describe a Bayesian multivariate response random effects logistic regression model and how it can be used for provider profiling. We then describe a latent variable approach that is motivated by item response theory models.

2.1 | Bayesian multivariate response random effects logistic regression models

This method is based on fitting a separate random effects logistic regression model for each of the binary indicators. However, the random effects for the separate logistic regression models are drawn from a multivariate normal distribution. Thus, the provider-specific random effects for the different indicators can be correlated with one another.

Let $Y_{ij}^{(k)}$ denotes the kth binary indicator measured on the ith subject in the jth provider ($k = 1,\dots,K$). We make the assumption that $Y_{ij}^{(k)} = 1$ denotes a successful outcome or treatment for the ith patient in the jth provider. Let $\mathbf{X}_{ij}^{(k)}$ denote a vector of subject characteristics used for risk-adjustment when modeling variation in the kth indicator. For reasons of model interpretation, we will assume that any continuous variables have been centered around the sample average. Note that the vector of subject characteristics can vary across the indicator-specific logistic regression models (ie, we are not assuming that the same set of risk factors or subject characteristics will be used for each of the indicators). Where applicable, the vector may also include hospital characteristics (in some settings the analyst may want to account for immutable hospital characteristics that are beyond the control of the hospital, such as location). For each of the K binary indicators a random effects logistic regression model is fit:

$$\text{logit}(\Pr(Y_{ij}^{(k)} = 1)) = \text{logit}(p_{ij}^{(k)}) = \alpha_{0j}^{(k)} + \alpha^{(k)}\mathbf{X}_{ij}^{(k)}. \tag{1}$$

Note that for the kth indicator, the intercept varies across providers, while the regression slopes for the subject characteristics are fixed across providers. We note that allowing the regression slopes to vary across providers allows the performance of hospitals to differ for different patient groups, for example, male and female patients; however, we do not explore this further here. A multivariate normal distribution is then assumed for the distribution of the provider-specific intercepts for the K regression models:

$$\begin{pmatrix} \alpha_{0j}^{(1)} \\ \alpha_{0j}^{(2)} \\ \vdots \\ \alpha_{0j}^{(K)} \end{pmatrix} \sim \text{MVN}\left(\mu = \begin{pmatrix} \alpha_{0}^{(1)} \\ \alpha_{0}^{(2)} \\ \vdots \\ \alpha_{0}^{(K)} \end{pmatrix}, \Sigma = \begin{pmatrix} \sigma_{11}^2 & \sigma_{12} & \cdots & \sigma_{1K} \\ \sigma_{21} & \sigma_{22}^2 & \cdots & \sigma_{2K} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{K1} & \sigma_{K2} & \cdots & \sigma_{KK}^2 \end{pmatrix} \right). \tag{2}$$

The vector μ that parametrizes the multivariate normal distribution is the mean of the multivariate normal distribution, which has K components. The matrix Σ that parametrizes the multivariate normal distribution is the variance-covariance matrix and has $K(K+1)/2$ components.

For a given indicator, the parameter $\alpha_0^{(k)}$ denotes the log-odds of the indicator being present (ie, of a successful patient outcome/treatment) at an average hospital for a reference subject whose covariates are all equal to zero. For a given indicator, hospitals whose random effects are greater than $\alpha_0^{(k)}$ (ie, $\alpha_{0j}^{(k)} > \alpha_0^{(k)}$) are hospitals at which the odds of the indicator being present are higher than at an average hospital, while hospitals whose random effects are less than $\alpha_0^{(k)}$ (ie, $\alpha_{0j}^{(k)} < \alpha_0^{(k)}$) are hospitals at which the odds of the indicator being present are lower than at an average hospital. A hospital for which $\alpha_{0j}^{(k)} < \alpha_0^{(k)}$, $k = 1,\dots,K$ is a hospital with poorer performance than average on all K indicators. In the case study in the subsequent section we will illustrate different parameters that can be derived from the model to quantify provider performance on an array of binary quality indicators.

Fitting multivariate response random effects logistic regression models via maximum likelihood estimation (adaptive quadrature) would very likely prove computationally infeasible given the presence of six correlated random effects and the large number of providers in these data. We therefore fit all models using Markov Chain Monte Carlo (MCMC) methods.[17] We specify diffuse normal prior distributions for the fixed effects ($\alpha^{(k)}$, $k = 1,\dots,K$), a diffuse multivariate normal prior distribution for the mean of the multivariate normal distribution of the random effects μ, and a diffuse Wishart prior distribution for the precision matrix of the multivariate normal distribution of the random effects (inverse of the variance-covariance matrix), Σ^{-1}.

2.2 | Latent variable approach for use with multiple indicators

As noted in Section 1, a small number of studies have described statistical methods for summarizing provider performance on multiple indicators of quality of care that are based on latent variable models.[11-13] The methods described below are

motivated by methods described in these articles. The proposed methods are inspired by item response theory models which are used in the psychometric literature. Let θ denotes a latent or unmeasured variable that denotes the hospital's underlying quality of care. We use the convention that higher levels of θ denote higher quality, while lower levels of θ denote lower quality. We use the same notation as in Section 2.1 and, as above, assume that all of the indicators of quality of care are binary.

K regressions are fit, one for each of the K binary indicators. The regression model for the jth indicator is:

$$\text{logit}(\Pr(Y_{ij}^{(k)} = 1)) = \text{logit}(p_{ij}^{(k)}) = \alpha^{(k)} + \beta^{(k)}\theta_j + \gamma^{(k)}\mathbf{X}_{ij}^{(k)}. \quad (3)$$

If $\gamma^{(k)} = 0$ (ie, if the vector of covariates is excluded), then this is the conventional two-parameter item response theory model (albeit with a slightly different parameterization).[14] Specifically, the "difficulty" parameter $\delta^{(k)} = -\alpha^{(k)}/\beta^{(k)}$ measures the level of hospital quality needed to observe a 50% prevalence on the given indicator, while the "discrimination" parameters $\beta^{(k)}$ measure the degree to which the given indicator can distinguish between different levels of hospital quality. For indicators for which risk-adjustment is necessary, the model is an extension or modification of the two-parameter item response theory model. The sign of $\beta^{(k)}$ is not identifiable, so the constraint that $\beta^{(k)} > 0$ is added to the model.[13] This is consistent with our convention that higher levels of θ denote higher quality.

We fit these models using MCMC methods.[14,17] Diffuse normal distributions were assumed for two of the sets of regression coefficients ($\alpha^{(k)}$, $\gamma^{(k)}$, $k = 1, \ldots, K$), while diffuse half-normal priors were assumed for the $\beta^{(k)}$, due to the identifiability constraint described above ($\beta^{(k)} \sim N(0, \sigma^2)\text{I}(0,)$, where I(0,) denotes the indicator function that takes the value zero for negative values $\beta^{(k)}$ and 1 for positive values). Finally, we assumed that $\theta \sim N(0, 1)$.

3 | CASE STUDY

In this section we provide a case study to illustrate the application of the methods described in the previous section. We use data on patients hospitalized with AMI at a large set of hospitals in the Canadian province of Ontario.

3.1 | Data sources

We used data from the EFFECT study, which was designed to improve the quality of care provided to patients with cardiovascular disease in Ontario.[9] The sample for this case study consisted of 10 881 patients admitted with a diagnosis of AMI to one of 102 hospital sites in Ontario, Canada between 1 April 1999 and 31 March 2001. Data on patient characteristics, outcomes, and processes of care were obtained from patients' medical records by retrospective chart review.

3.2 | Indicators of quality of care

We considered six binary indicators of quality of care for patients hospitalized with AMI: (i) reperfusion (ie, reopening blocked coronary arteries using either thrombolysis or percutaneous coronary intervention) in patients with ST-segment elevation myocardial infarction (STEMI); (ii) assessment of left ventricular ejection fraction (LVEF); (iii) measurement of cardiac troponin levels; (iv) use of acetylsalicylic acid (ASA) within 6 hours of hospital arrival; (v) use of beta-blockers within 12 hours of hospital arrival; (vi) survival to 30 days from hospital admission (both in-hospital and out-of-hospital deaths were captured). The first five indicators assess patient-specific processes of care, while the sixth indicator is a patient outcome. The sixth indicator denotes survival to 30 days from hospital admission. While a more common indicator is death within 30 days of hospital admission, we changed the indicator from death to survival so that a positive response for each of the indicators denotes a good outcome or process of care, while a negative response for each indicator denotes a poor outcome or process of care. The first indicator is only for use in the subset of patients with STEMI, while the remaining five indicators can be assessed in all patients. In the analytic dataset, the value of the reperfusion indicator was set to missing for those subjects who did not present with a STEMI. For all other indicators, the value of the indicator was set to either 0 or 1, denoting that the indicator was not satisfied (not performed) or satisfied (performed), respectively.

The first five indicators are processes of care measures and thus one would expect that they would be performed for all eligible patients (eg, LVEF should be assessed in all patients). Thus, in the subsequent models, no risk-adjustment will be

used for these five outcomes. However, comparison of mortality (or survival) across hospitals requires risk-adjustment, as the illness severity may vary across hospitals. Thus, in the subsequent models, risk-adjustment will be used for the survival outcome.

3.3 | Multivariate response random effects logistic regression models

We fit a multivariate response Bayesian random effects logistic regression as described in the previous section. The regression model for 30-day survival had a single explanatory variable, the GRACE score, which is a validated model for predicting mortality in patients with acute coronary syndromes (the components of the GRACE score are age, history of myocardial infarction, history of heart failure, increased pulse rate at presentation, lower systolic blood pressure at presentation, elevated initial serum creatinine level, elevated initial serum cardiac biomarker levels, ST-segment depression on presenting electrocardiogram, and not having a percutaneous coronary intervention performed in hospital).[18] The other five logistic regression models had no explanatory variables. These five indicators denote processes of care that should be provided to all patients, regardless of illness severity. Accordingly, no covariates were included in these five regression models.

MCMC methods were used to estimate the posterior distribution of the model parameters. Three chains were run, each using different initial values for the model parameters. Diffuse prior distributions were assumed for the model parameters. Each chain used an initial run of 500 000 "burn-in" iterations (a high number of burn-in iterations was determined to be necessary through a trial and error approach and examination of the subsequent trace plots for the sampled parameters), and was then monitored for an additional 50 000 iterations, with a thinning interval of 10 (ie, 5000 monitored iterations were retained from each of the three chains). Thus a total of 15 000 monitored iterations were used to determine the posterior distributions of the parameters of interest. The Gibbs sampler was implemented using OpenBUGS version 3.2.3 using the R2OpenBUGS package for R.

Diffuse prior distributions were specified for all parameters. The prior distribution for the fixed slope for the GRACE score in the model for 30-day survival was specified to be a normal distribution with mean zero and variance 100. The prior distribution for μ, the mean of the multivariate normal distribution of the random effects, was specified to be a multivariate normal distribution with mean zero and a variance-covariance matrix equal to $100 \times \mathrm{I}_{6\times 6}$, where I denotes the 6×6 identity matrix. Finally, the prior distribution for Σ^{-1}, the precision matrix of the multivariate normal distribution of the random effects, was specified to be the Wishart distribution $W_6\left(\frac{1}{6}\mathrm{I}_{6\times 6}, 6\right)$.

A total of 640 model parameters were monitored: one fixed effect slope for the GRACE score in the model for 30-day survival, six parameters for the mean of the multivariate distribution of the random effects, 21 parameters for the precision matrix (inverse of the symmetric variance-covariance matrix) for the multivariate distribution of the random effects, and 612 hospital-specific random effects for the six indicators (102 hospitals $\times$ 6 indicators). Convergence of the Gibbs sampler was assessed by visual inspection of the trace plots for 28 parameters (the six components of the mean of the multivariate normal distribution, the 21 components of the variance-covariance matrix, and the one fixed effect for the effect of the GRACE score on patient survival) and for the 30 random effect parameters for the first five hospitals. The three separate chains starting at different starting values mixed well and displayed no lack of convergence. The convergence of each chain was also assessed using Geweke's statistic,[19] by which we tested the equality of the means of the sampled parameters in the first 25% of the chain with that in the last 25% of the chain. If the sampled values of a given parameter are drawn from the same stationary distribution, then the two means are equal and the resultant test statistic will have a standard normal distribution. For each of the three chains, there was no evidence that the distribution of Geweke's test statistic was not normal across the 640 model parameters when using visual inspection of a normal quantile-quantile plot.

3.4 | Latent variable approach

We fit a multivariate response Bayesian random effects logistic regression model. MCMC methods were used to estimate the posterior distribution of the model parameters. Three chains were run, each using different initial values for the model parameters. Diffuse prior distributions were assumed for the model parameters. Each chain used an initial run of 500 000 "burn-in" iterations, and was then monitored for an additional 3 500 000 iterations, with a thinning interval of 700 (ie, 5000 monitored iterations were retained from each of the three chains). A very high thinning interval (and therefore number of monitoring iterations) was used due to the high degree of autocorrelation in the monitored chains

when lower thinning intervals were used. Thus a total of 15 000 monitored iterations were used to determine the posterior distributions of the parameters of interest. The Gibbs sampler was implemented using PROC MCMC in SAS (SAS/STAT 14.3). Different statistical software was used for the two models as both models were not able to be fit using only one of the programs.

The prior distribution for the intercept term ($\alpha^{(k)}$) in each of the six models was specified to be a normal distribution with mean zero and variance 100. The prior distribution for the slope associated with the latent variable ($\beta^{(k)}$) in each of the six models was specified to be a half-normal distribution with mean zero and variance 100. The prior distribution for the fixed slope for the GRACE score in the model for 30-day survival was specified to be a normal distribution with mean zero and variance 100.

A total of 115 model parameters were monitored: one fixed effect slope for the GRACE score in the model for 30-day survival, six intercept parameters, six slope parameters, and 102 hospital-specific values of the latent variable denoting hospital quality. Convergence of the Gibbs sampler was assessed by visual inspection of the trace plots for 13 parameters (the intercepts and slopes in the six regression models) and for the five values of the latent variable for the first five hospitals. The three separate chains starting at different starting values mixed well and displayed no lack of convergence. The convergence of each chain was also assessed using Geweke's statistic,[19] by which we tested the equality of the means of the sampled parameters in the first 25% of the chain with that in the last 25% of the chain. In the first and third set of chains, none of the 115 applications of Geweke's test resulted in a rejection of stationarity. In the second set of chains, only one of the 115 applications of Geweke's test resulted in a rejection of stationarity. If one were to apply a Bonferroni correction to each set of 115 applications, then none of the tests would have resulted in a rejection of stationarity.

3.5 | Results

The overall prevalence of the six indicators were 62% (reperfusion in patients with STEMI), 46% (LVEF assessment), 55% (measurement of cardiac troponin), 51% (ASA within 6 hours of admission), 21% (beta-blockers within 12 hours of admission), and 89% (survival to 30 days after admission). The hospital-specific prevalences of the indicators ranged from 41% to 100% (reperfusion), 0% to 87% (LVEF), 0% to 100% (troponin), 25% to 83% (ASA), 3% to 61% (beta-blockers), and 80% to 97% (survival to 30 days).

3.5.1 | Multivariate response random effects logistic regression models

The posterior mean of the variances of the hospital-specific random effects were 0.20 (reperfusion in patients with STEMI), 0.94 (LVEF assessment), 5.40 (measurement of cardiac troponin), 0.02 (ASA within 6 hours of admission), 0.05 (beta-blockers within 12 hours of admission), and 0.03 (survival to 30 days after admission). These are equivalent to variance partition coefficients (VPCs) of 0.06, 0.22, 0.62, 0.01, 0.02, and 0.01, respectively (using the latent variable formulation of the VPC).[20-22] Thus, 6% of the variation in use of reperfusion therapy is due to systematic differences between hospitals, while 62% of the variation in measurement of cardiac troponin was due to systematic differences between hospitals. Some of the indicators displayed only minor between-hospital variation, while others displayed very strong between-hospital variation.

Correlation of hospital-specific random effects

Within each iteration of the Gibbs sampler, the sampled precision matrix was inverted to obtain the variance-covariance matrix of the distribution of the hospital-specific random effects. From this matrix we obtained the correlation matrix for the hospital-specific random effects. We also computed Bayesian one-sided *P*-values for the hypothesis that the correlation was negative (these were computed as the proportion of the sampled correlations that were negative). The posterior mean of the correlation matrix and the corresponding Bayesian one-sided *P*-values are reported in Table 1. The scatterplot matrix plotting the pairwise relationship between hospital-specific posterior means of the different random effects is presented in Figure 1.

In interpreting the magnitude of specific correlations, we will use the following criteria, which are based on Cohen's discussion of effect sizes: $0.1 < \rho \leq 0.3$ denotes weak correlation; $0.3 < \rho \leq 0.5$ denotes moderate correlation; $\rho > 0.5$ denotes strong correlation.[23,24]

TABLE 1 Posterior mean of the correlation coefficients and Bayesian one-sided *P*-values

Indicator	Reperfusion	LVEF	Troponins	ASA	Beta-blockers	Survival to 30 days
Reperfusion	1.00	0.80 (<0.001)	0.84 (<0.001)	0.38 (0.037)	0.61 (0.001)	0.18 (0.167)
LVEF	0.80 (<0.001)	1.00	0.41 (<0.001)	0.53 (0.018)	0.69 (0.004)	0.32 (0.056)
Troponin	0.84 (<0.001)	0.41 (<0.001)	1.00	0.09 (0.359)	0.30 (0.105)	−0.01 (0.533)
ASA	0.38 (0.037)	0.53 (0.018)	0.09 (0.359)	1.00	0.47 (0.042)	0.37 (0.073)
Beta-blockers	0.61 (0.001)	0.69 (0.004)	0.30 (0.105)	0.47 (0.042)	1.00	0.37 (0.069)
Survival to 30 days	0.18 (0.167)	0.32 (0.056)	−0.01 (0.533)	0.37 (0.073)	0.37 (0.069)	1.00

Note: Each cell contains the posterior mean of the correlation coefficient (Bayesian one-sided *P*-value).
Abbreviations: ASA, acetylsalicylic acid; LVEF, left ventricular ejection fraction.

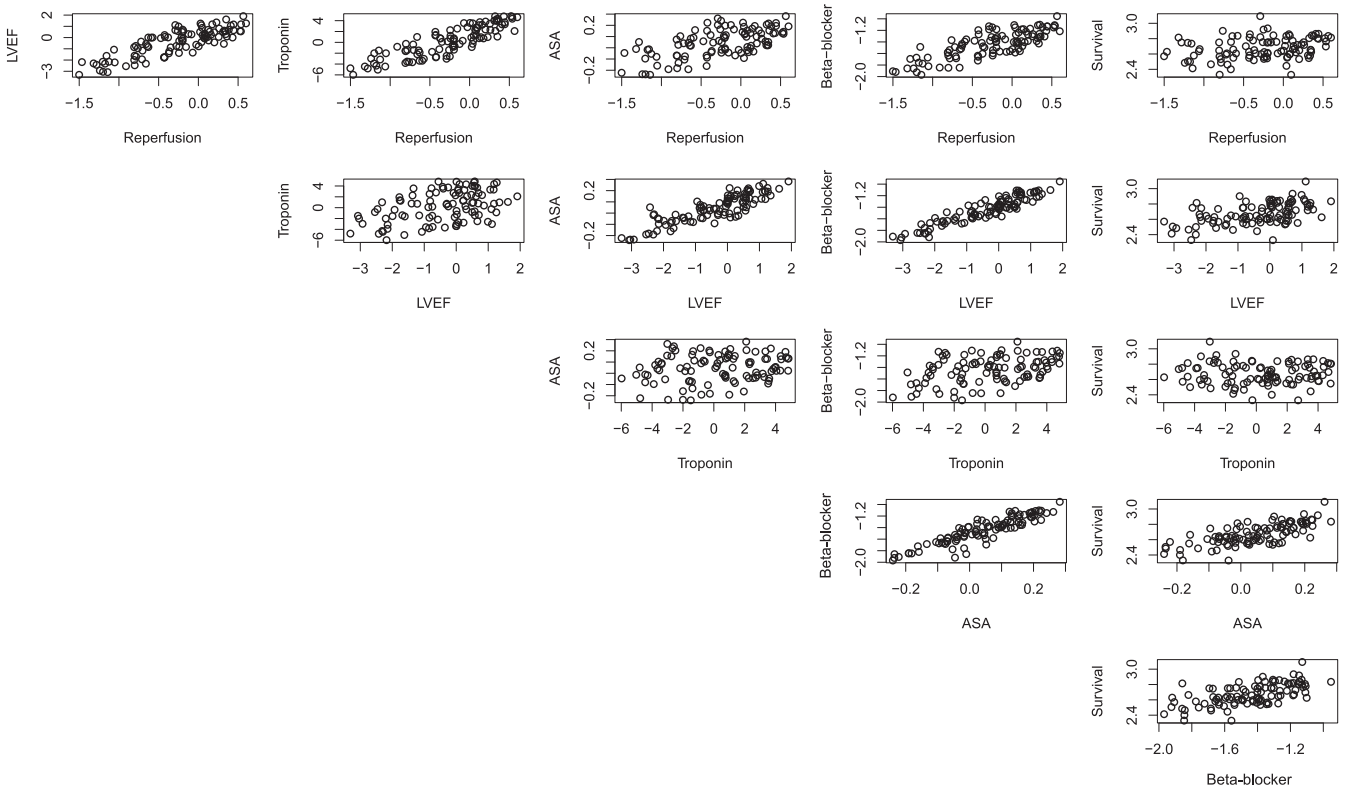

FIGURE 1 Correlation between hospital-specific random effects for the six indicators

There was a strong correlation between a hospital's use of reperfusion therapy in patients with STEMI (reperfusion) and its conducting of LVEF assessments ($\rho = 0.80$), measurement of cardiac troponins ($\rho = 0.84$) and its of use beta-blockers within 12 hours of hospital arrival ($\rho = 0.61$). Similarly, there was a strong correlation between a hospital's measurement of LVEF and use of ASA within 6 hours ($\rho = 0.53$) and its use of beta-blockers within 12 hours ($\rho = 0.69$).

Five of our indicators denote processes of care. Of the 10 comparisons of hospital performance on pairs of process of care indicators, five demonstrated strong correlation ($\rho > 0.50$). Of the remaining five pairwise comparisons, there was a moderate correlation between a hospital's use of reperfusion therapy in patients with STEMI and its use of ASA within 6 hours of arrival ($\rho = 0.38$), there was a moderate correlation between a hospital's assessment of LVEF and measurement of troponins ($\rho = 0.41$), and there was a moderate correlation between a hospital's use of ASA within 6 hours of arrival and its use of beta-blockers within 12 hours of arrival ($\rho = 0.47$). There was a weak correlation between a hospital's measurement of cardiac troponins and its use of beta-blockers within 12 hours of arrival ($\rho = 0.30$). Finally, the correlation between a hospital's measurement of cardiac troponins and its use of ASA within 6 hours of arrival was negligible ($\rho = 0.09$).

Of our six binary indicators, patient survival was the sole indicator that was a patient outcome and not a process of care. Hospital performance on patient survival was not strongly correlated with hospital performance on any of the five processes of care indicators. There was a moderate correlation between a hospital's performance on patient survival and its assessment of LVEF ($\rho = 0.32$), its use of ASA within 6 hours of arrival ($\rho = 0.37$), and its use of beta-blockers within 12 hours of arrival ($\rho = 0.37$). There was a weak correlation between a hospital's performance on patient survival and its use of reperfusion therapy ($\rho = 0.18$). There was no correlation between a hospital's performance on patient survival and its measurement of cardiac troponins ($\rho = -0.01$).

Probability of having negative random effects for all six indicators simultaneously

A hospital with a negative random effect for a given indicator is hospital whose performance on that indicator is worse than at an average hospital (since the indicators were structured so that a value of zero denoted a poor outcome or process of care and a value of one denoted a good outcome or process of care). At each monitored iteration of the Gibbs sampler, we constructed a binary variable for each hospital that denoted whether all six random effects for that hospital in that iteration were negative. The mean of this indicator variable across the 15 000 monitored iterations denotes the posterior probability that all six random effects were simultaneously negative for the given hospital. This posterior probability ranged from 0 to 0.85 across the 102 hospitals. The median posterior probability across the 102 hospitals was 0, while the 75th percentile was 0.12. The 90th percentile was 0.52. Thus, for the 10% most extreme hospitals, the posterior probability that all six random effects were negative was at least 0.52. There were 11 (11.0%) hospitals for which this posterior probability was at least 0.5. The left panel of Figure 2 depicts a "snake plot" in which the posterior probability of below-average performance on all six indicators is plotted against the hospital's rank on these probabilities.

The above process was repeated to determine the posterior probability that a given hospital had positive random effects for each of the six indicators simultaneously (and thus had performance superior to that of an average hospital on each of the six indicators). This posterior probability ranged from 0 to 0.77 across the 102 hospitals. The median posterior probability was 0.01, while the 75th percentile was 0.19. The 90th percentile was 0.50. There were 11 (11%) hospitals for which this posterior probability was at least 0.5. The right panel of Figure 2 depicts a "snake plot" in which the posterior probability of above-average performance on all six indicators is plotted against the rank of these probabilities. The Spearman rank correlation between a hospital's probability of having below-average performance on all six indicators and the hospital's probability of having above-average performance on all six indicators was −0.87.

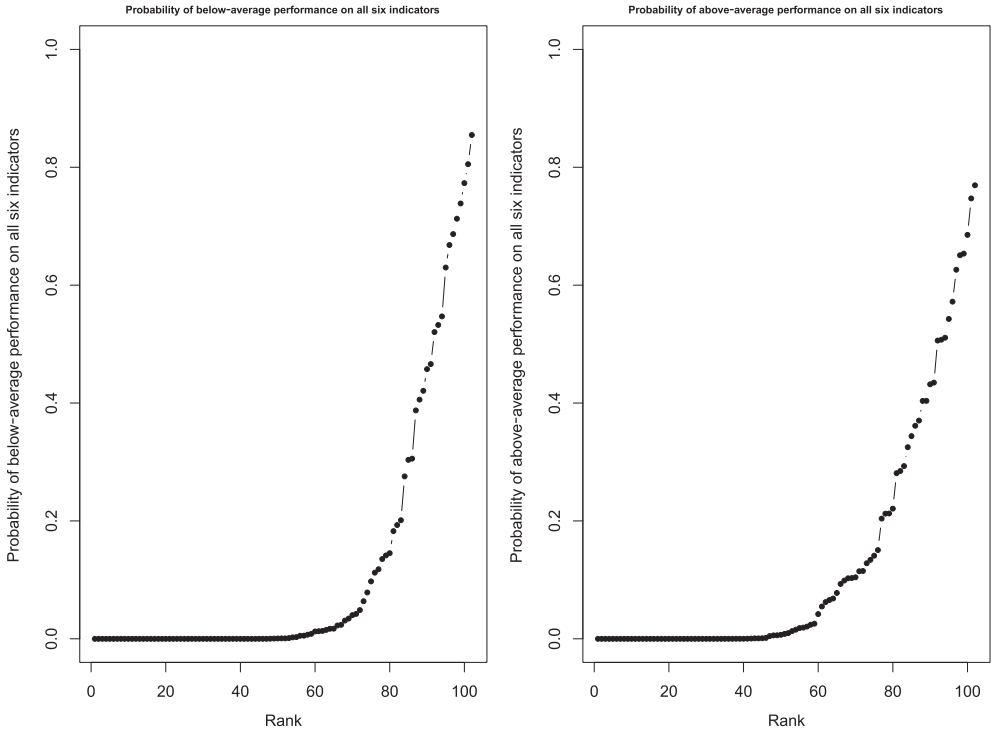

FIGURE 2 Probability of below/above-average performance on all six indicators

Mahalanobis distance of each hospital from the center of the distribution of the random effects

The metric discussed in the previous section involved the probability that a hospital had worse performance than average on all six indicators. While this metric involves directionality, it does provide information on how "far" the performance of a given hospital is from an average hospital. One could contrast two hospitals that both have a very high case load and who both have a high probability of having poor performance on all indicators. The first hospital has performance on each indicator that is slightly worse than average. However, due to its very high case load, the probability of poor performance on all six indicators is high. The second hospital has performance on each indicator that is substantially worse than average. Its estimated probability of poor performance on all six indicators is also very high. One would want to distinguish between these two hospitals. To do so, one would want a metric that conveys information about how far a given hospital is from an average hospital.

The Mahalanobis distance is a measure of distance in multivariate space.[25] Given a multivariate distribution with mean $\boldsymbol{\mu}$ and variance-covariance matrix $\boldsymbol{\Sigma}$, the Mahalanobis distance of a vector $\mathbf{x}$ from the mean of the distribution is defined as $D = \sqrt{(\mathbf{x}-\boldsymbol{\mu})^t\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})}$. In the context of multivariate provider profiling, this distance measure allows one to determine the distance of each provider from a hospital with average performance on each of the indicators. Note that the Mahalanobis distance is only a measure of distance and does not provide information on directionality, furthermore it implicitly gives equal weight to all six indicators (which is consistent with what we have been doing in these analyses). We will subsequently combine this distance measure with information about the probability of a hospital having poor quality of care.

At each monitored iteration of the Gibbs sampler, we used the sampled values of the random effects and the precision matrix (the inverse of the variance-covariance matrix) to determine the Mahalanobis distance of each hospital from the mean of the multivariate distribution of the random effects. By examining this quantity for each hospital across the 15 000 monitored chains we determined the posterior distribution of the Mahalanobis distance for each hospital. The mean of the hospital-specific posterior distribution of the Mahalanobis distance ranged from 1.73 to 3.11, with a median of 2.31 (25th and 75th percentiles: 2.12 and 2.51).

We made the admittedly subjective decision that a hospital whose Mahalanobis distance exceeded that of 75% of hospitals was far from the center of the multivariate distribution (ie, we defined those hospitals in the top 25% of distance to be far from the center of the distribution). Accordingly, at each iteration of the Gibbs sampler, we computed the Mahalanobis distance of each hospital from the mean of the multivariate distribution of the random effects. Within the given iteration of the Gibbs sample, we determined the 75th percentile of this distribution across the 102 hospitals. We then created an indicator variable for each hospital denoting whether, for the given iteration, that hospital's Mahalanobis distance exceeded the 75th percentile of the distribution of Mahalanobis distances. For each hospital we computed the mean of this indicator variable across the 15 000 iterations of the Gibbs sampler. This quantity is the posterior probability that the hospital's Mahalanobis distance exceeds the 75th percentile of the distribution of Mahalanobis distances across hospitals. This posterior probability ranged from 0.06 to 0.75 across the 102 hospitals, with a median of 0.21 (25th and 75th percentiles: 0.14 and 0.33).

A hospital with a large Mahalanobis distance is far from the center of the distribution of hospital performance. However, the Mahalanobis distance does not, on its own, provide information about the quality of care provided by that hospital. Figure 3 describes the relationship between Mahalanobis distance and the probability of having negative random effects for all six indicators (left panel) and the probability of having positive random effects for all six indicators (right panel). In each panel, we have used solid red circles to denote those hospitals whose probability of having below-average performance on all six indicators exceeded 0.5 and whose Mahalanobis distance exceeded the 75th percentile of such distances. In each panel, we have used solid blue triangles to denote those hospitals whose probability of having above-average performance on all six indicators exceeded 0.5 and whose Mahalanobis distance exceeded the 75th percentile of such distances. In each panel, hospitals in the top-right quadrant merit further examination. In the left panel, there are seven hospitals that have a high probability of having below-average performance on all six indicators and that are also far from the performance of an average hospital. These are hospitals that may merit focused quality improvement initiatives to improve the quality of care provided to patients with AMI. In the right panel, there are two hospitals that have a high probability of having above-average performance on all six indicators and that are also far from the performance of an average hospital. These two hospitals may merit focused attention so that the reasons for their high quality performance can be identified and disseminated to other hospitals.

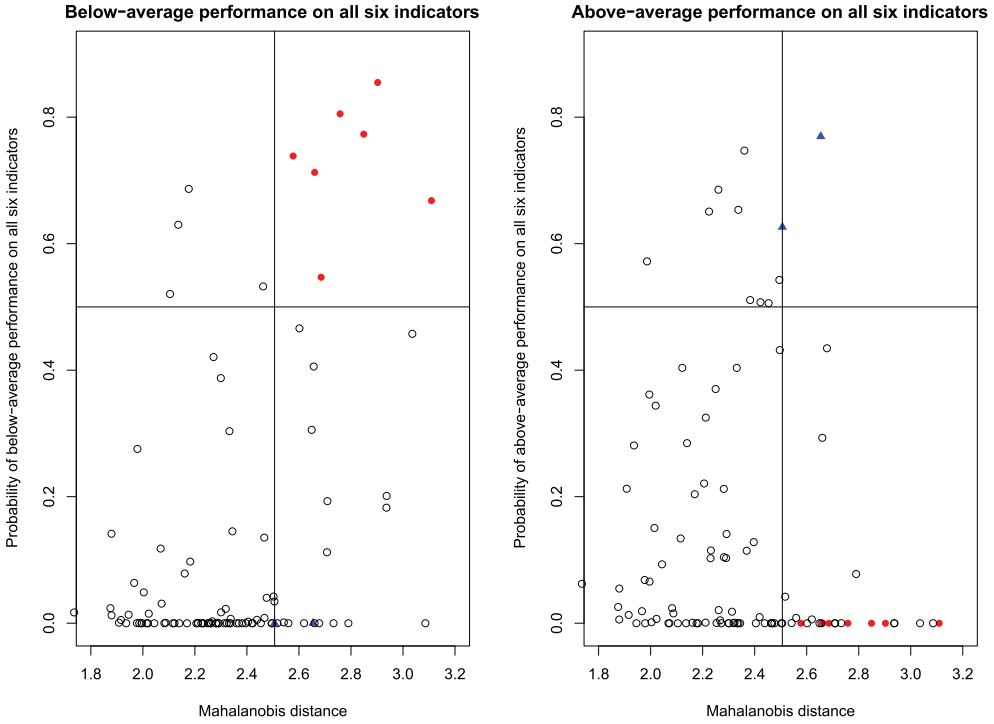

FIGURE 3 Relationship between Mahalanobis distance and below/above-average performance on all six indicators [Colour figure can be viewed at wileyonlinelibrary.com]

Profiling using clinical benchmarks

We conducted a set of analyses to illustrate the application of multivariate response multilevel logistic regression models for provider profiling when an external standard for acceptable performance is used. At the beginning of Section 3.5, we reported the overall performance on the six indicators. We made the arbitrary decision that hospitals whose performance on a given processes of care indicator was worse than the provincial average were deemed to have below-average performance on that indicator. Furthermore, we decided that hospitals whose adjusted survival (as determined by the hospital-specific random effect for the survival model) was worse than average had below-average performance on survival. At each iteration of the Gibbs sampler, we determined whether each hospital had below-average performance on all six indicators. We then determined the proportion of the iterations in which each hospital had below-average performance on all six indicators. We then repeated this process examining the probability that each hospital had above-average performance on all six indicators.

The hospital-specific probability of below-average performance on all six indicators ranged from 0 to 0.86, with a median of 0. There were 11 hospitals for which this probability exceeded 0.50. Further investigation for the reasons for poor performance at these 11 hospitals may be merited. The hospital-specific probability of above-average performance on all six indicators ranged from 0 to 0.64, with a median of 0. There was only one hospital for which this probability exceeded 0.50.

3.5.2 | Latent variable approach

The posterior means and the 95% highest probability density intervals for the 13 regression parameters (six intercepts, six slopes for the latent variable, and the slope for the GRACE score) are reported in Table 2. The slope parameter associated with the latent variable denoting hospital quality is referred to as a "discrimination" parameter in the context of item response theory models. The larger the slope associated with the latent variable, the greater the association between hospital quality and performance on a given indicator. Indicators associated with higher values of β have a greater ability to distinguish between different levels of hospital quality compared with indicators with lower values of β.

The relationship between the latent variable for hospital quality and performance on the six different indicators is described in Figure 4. In this figure, the latent variable was allowed to range from −2 to 2, (the range in which approximately 95% of hospitals would lie, given the assumption that this variable follows a standard normal distribution across

TABLE 2 Posterior means and 95% HPD intervals

Indicator	Posterior mean	95% HPD interval
Reperfusion	0.5019	(0.4378, 0.5669)
LVEF	−0.2108	(−0.3298, −0.0880)
Troponin	−0.1586	(−1.2560, 0.8923)
ASA	0.0251	(−0.0147, 0.0620)
Beta-blockers	−1.3504	(−1.3983, −1.3028)
30-day survival	2.6676	(2.5773, 2.7555)
Slope for latent variable (β)		
Reperfusion	0.0088	(0.0000, 0.0256)
LVEF	0.5953	(0.4910, 0.7016)
Troponin	5.4985	(4.4408, 6.5913)
ASA	0.0186	(0.0000, 0.0451)
Beta-blockers	0.0640	(0.0118, 0.1158)
30-day survival	0.0345	(0.0000, 0.0809)
Slope for GRACE score (γ)		
30-day survival	−0.0337	(−0.0356, −0.0319)

Abbreviations: ASA, acetylsalicylic acid; HPD, highest probability density; LVEF, left ventricular ejection fraction.

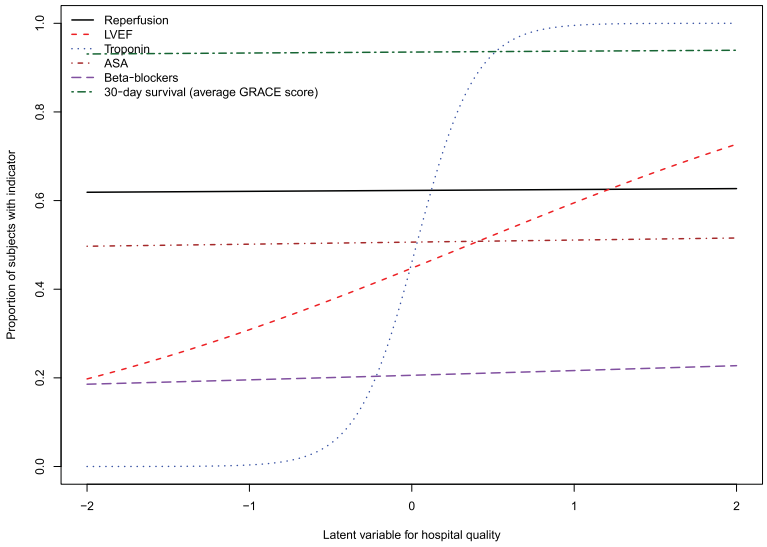

FIGURE 4 Relationship between latent variable and performance on the six indicators [Colour figure can be viewed at wileyonlinelibrary.com]

hospitals). There is one curve for each of the six indicators. The curve for survival denotes the probability of 30-day survival for a patient with an average GRACE score as a function of the underlying latent variable. The indicator denoting measurement of cardiac troponins had the greatest ability to discriminate between hospitals on the basis of quality of care. Measurement of LVEF had moderate ability to discriminate between hospitals on the basis of quality. The remaining four indicators had negligible ability to discriminate between hospitals.

3.5.3 | Comparison of the two approaches

The agreement between the Bayesian multivariate logistic regression approach and the latent variable approach is explored in Figure 5. This figure consists of four panels. In each panel we have plotted the estimated hospital-specific

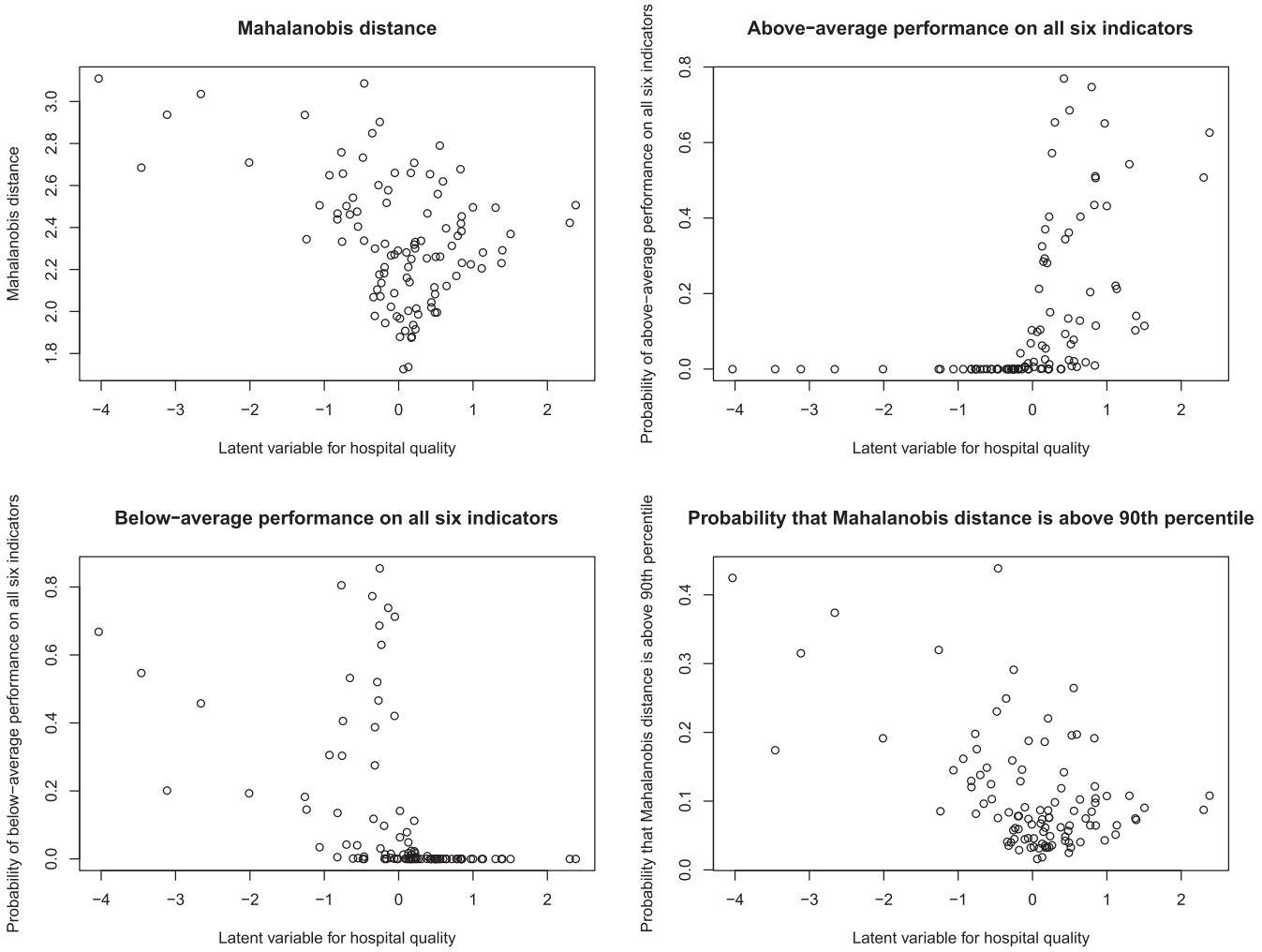

FIGURE 5 Comparison of latent variable approach and multivariate logistic regression model

latent variable denoting hospital quality vs a different metric based on the Bayesian multivariate logistic regression model. In the top left panel we have plotted the posterior mean of the hospital-specific Mahalanobis distance against the latent variable denoting hospital quality. As one moves away from a latent variable value of zero (denoting average quality), the Mahalanobis distance tends to increase, since we are moving away from the center of the multivariate distribution. In the bottom left panel we have plotted the probability of below-average performance on all six indicators against the latent variable denoting hospital quality. The Spearman rank correlation for these two quantities was −0.76, denoting, as expected, a strong negative correlation. Note that the large majority of hospitals with a positive latent variable (denoting above-average quality) tended to have a very low probability of below-average performance on all six indicators. In the top-right panel we have plotted the probability of above-average performance on all six indicators against the latent variable denoting hospital quality. The Spearman rank correlation for these two quantities was 0.82, denoting, as expected, a strong positive correlation. Finally, in the bottom right panel we have plotted the probability that the Mahalanobis distance is above the 90th percentile of distance against the latent variable denoting hospital quality. The Spearman rank correlation for these two quantities was −0.28, denoting a weak correlation.

4 | DISCUSSION

The primary objective of this article was to describe a multivariate response Bayesian random effects logistic regression model that allows for provider profiling simultaneously on multiple binary indicators of quality of care. Analyses based on this model allow for a formal assessment of whether performance on different quality indicators is correlated within given

healthcare providers. This allows for an examination of whether providers that have poor performance on one indicator also have poor performance on a second indicator, and conversely whether providers that have good performance on one indicator also have good performance on a second indicator. The use of this approach allows for a deeper examination of provider performance than does examining each indicator in isolation. A secondary objective was to compare inferences made using this model with those made using a previously described latent variable model for profiling on multiple indicators of hospital performance.

The latent variable model can be thought of as a simplification of the multivariate response logistic regression model. This can be seen in comparing formulas (2) and (3) above. Formula (3) is simply formula (2) where we have constrained all cross-equation correlations to equal one. The multivariate response logistic regression model allows for different pairwise correlations between different pairs of indicators. In contrast, the latent variable model assumes that there is only one underlying dimension of hospital quality. In so doing, it implicitly assumes that all the correlations in the multivariate model are equal to one. With this simplifying assumption comes greater ease of interpretation. This may be possible when the indicators are such that is reasonable to assume a unidimensional hospital quality model. If the indicators pertained to very different types of patients (eg, surgical site infections in patients undergoing colorectal surgery and troponin measurement in patients hospitalized with AMI), this assumption may not be realistic. Similarly, when the indicators reflect a mix of process, outcome, and structural measures, this assumption may not be realistic, and the multivariate response model may be preferred. In the current application we have five process measures and one outcome measure. More generally, many of the estimated correlations in the multivariate response models are substantially lower than 1, further suggesting that the multivariate response model would likely be statistically preferred to the latent variable model.

Christiansen and Morris[26] advocated for the use of Bayesian random effects regression models in provider profiling, suggesting that one of the advantages of this approach is the ability to make probabilistic statements about acceptable provider performance. Similarly, Normand et al[27] described the use of Bayesian random effects regression models for healthcare provider profiling. The use of these methods allows for determining the probability of unacceptable performance at individual hospitals. The statistical performance of different Bayesian methods for hospital profiling has been examined in a series of articles.[28-32] Importantly, all of these articles considered scenarios in which healthcare providers were being compared on a single outcome.

While there is a large literature on statistical methods for provider profiling for a single patient outcome (eg, mortality), there is a limited literature on methods to compare healthcare provider performance on multiple patient outcomes. The majority of previous studies used latent variable approaches to model hospital performance on multiple indicators. To the best of our knowledge, apart from these studies that a used latent variable approach, only one study has examined healthcare provider profiling for multiple patient outcomes simultaneously.[33] Robinson et al developed a multivariate regression model for patient medical costs. The two outcomes were primary care costs and specialty care costs. The underlying regression model for each of the two cost outcomes was a two-part model that accounted for the large number of patients with zero costs. Importantly, the outcomes considered in that study were continuous. The novelty of the current study is its focus on provider profiling on multiple binary indicators of healthcare quality. Binary indicators are arguably more common in healthcare provider profiling than are continuous outcomes.

While there is a paucity of research on methods for multivariate provider profiling in health research, there is limited research on this topic in the field of education research. The primary focus in education is on test scores, which are continuous outcomes. Goldstein[34] described multivariate response random effects linear models that can be used for modeling variation in multiple student test scores (eg, mathematics test scores, reading test scores, and writing test scores) across schools. Leckie[35] recently showed how these multivariate response models can then be extended to model multiple cohorts of students and so examine the stability of school effects over time as well as the consistency in school effects across different academic subjects. Our motivation is the same as that of Goldstein and Leckie. While their focus was on multiple continuous test scores, our focus is on performance on multiple binary indicators. However, the same method underlies both approaches: a multivariate response random effects regression model.

The focus of the current study was on provider profiling for multiple binary indicators simultaneously. Our focus was on binary indicators as these are the most common type of indicator in healthcare profiling. However, the methods that we described can be easily modified to be used in settings with a mixture of continuous and binary indicators.[36] A recent example from the school performance monitoring literature is a joint analysis of school effects on student attainment (continuous), absence (continuous), and exclusion (binary) indicators.[37] In the context of hospital performance, possible continuous quality indicators include hospital length of stay or wait time for a given procedure (eg, time from hospital arrival to initiation of reperfusion therapy for patients arriving at hospital with a STEMI). To modify the multivariate response model, a logistic regression model for a binary indicator would be replaced with an appropriate linear model or

generalized linear model for the continuous outcome. While the focus was on binary indicators of healthcare quality, the proposed methods can be used to examine variation in any set of binary variables between hospitals or regions. Thus, given a set of comorbid conditions or disease risk factors, one could examine the within-hospital or within-region correlation between different comorbid conditions or risk factors. This would allow one to see whether the presence of comorbid conditions or risk factors tend to be clustered within certain hospitals or regions. In our case study we examined hospital performance on one patient outcome (survival to 30 days) and five measures of process of care. However, there is no reason that a specific application of this approach is limited to one patient outcome. One can consider any combination of patient outcomes and measures of process of care. Similarly, the latent variable approach can be modified to accommodate nonbinary outcomes.

An alternative approach to addressing multiple indicators is to explicitly create a composite indicator. While the latent variable model implicitly provides an estimate of a composite indicator (the latent variable denoting hospital quality), it is important to note that the multivariate response logistic regression model is not based on creating a composite indicator that pools information from a set of indicators. Compared to our multivariate approach, the explicit use of a composite indicator would likely result in a loss of information about variation in hospital performance. Furthermore, it is not clear what is the best method to pool multiple binary indicators to create a single composite indicator. Using our proposed approach, we are not weighting the different indicators, but examining the within-hospital correlation in the different indicators. Teixeira-Pinto and Normand[13] briefly discuss limitations with the creation of a composite hospital-level indicator of quality of care. One option is to calculate the proportion of patients receiving the given indicator at each hospital and then to compute the average proportion across indicators within each hospital (they refer to this as the raw average scores [RAS]). Alternatively, one can weight the indicator-specific proportions according to the number of patients eligible for the given indicator, to produce raw-weighted average scores (RWAS). They describe how paradoxical results can arise if some hospitals have no eligible patients for some of the indicators. Furthermore, they suggest that the computation of average scores is only meaningful when none of the indicators require risk-adjustment. Given that mortality requires risk-adjustment, we did not consider this approach in the current study. Furthermore, the use of RAS or RWAS is only feasible when all of the indicators are binary. In contrast to this, both methods used in our case study can be extended to include continuous outcomes such as hospital length of stay or procedural wait times.

Neither of the two methods described in this study require that each indicator be assessed on every subject. One of the indicators in the case study, reperfusion therapy, was only applicable to the subset of subjects with STEMI. This was achieved by setting the value of the indicator to missing for those subjects to whom the indicator did not pertain (ie, for those subjects who did not have a STEMI, the value of the reperfusion variable was set to missing). By doing so, those subjects to whom the indicator did not pertain provided no information on the performance of the hospital on that indicator. The ability of the method to incorporate indicators that apply to different subsets of the sample is important as many indicators apply only to subsets of the sample. For example, the outcome of readmission within 30 days of hospital discharge applies only to those subjects who are discharged alive from the index hospitalization episode.

In our case study we examined the probability that a hospital had below-average performance on all six indicators of quality of care. An advantage of MCMC methods is that researchers can create any summary measure of interest. For example, one could generate a warning flag indicator if a hospital was in the bottom quartile of performance on three or more indicators (without specifying what those three indicators were). One could then report the posterior probability of this warning indicator for each hospital and target quality improvement initiatives at those hospitals that had a high posterior probability for this warning indicator. Similarly, one could determine the posterior probability that a hospital has poor performance on at least one of the indicators. We examined each hospital's performance relative to that of an average hospital. One could also ask whether a hospital fell below externally set thresholds on each of the indicators. For instance, healthcare funders, regulators and caregivers could develop external thresholds for each indicator (eg, ASA use within 6 hours of admission in at least 80% of patients). The proposed methods can be easily modified to examine the probability that a hospital fell below these externally defined thresholds on a given number of quality of care indicators. The flexibility of the multivariate response Bayesian random effects logistic regression model is that the criteria for identifying hospitals as performance outliers can be modified by investigators to best address their objectives and criteria for classifying hospital performance.

We suggest that the two approaches considered in this article be seen as complementary. The latent variable approach has at least two advantages. First, it provides a single numeric summary of each hospital's quality of care. Hospitals in the tails of this distribution (eg, bottom 10% or top 10%) can be classified as performance outliers. Second, it allows for identifying which indicators are most closely correlated with quality of care, and which indicators do allow for a meaningful discrimination between hospitals according to their quality of care. However, a drawback to this approach is that

it requires the assumption of an unmeasured variable denoting hospital quality and it may be difficult to communicate to physicians and hospital administrators how this variable is estimated. The use of multivariate response hierarchical models has at least two advantages. First, it allows for a formal quantification of the correlation in hospital performance on different indicators. Second, as discussed in the previous paragraph, one can create many different flags of poor (or excellent) hospital performance based on a hospital's performance on multiple indicators. This may result in classifications that are of greater relevance for physicians and hospital administrators. Furthermore, these actionable flags can make reference to normative standards or thresholds where these exist. The primary limitation to the multivariate response logistic regression model is that, unlike the latent variable approach, it is more difficult to create a single summary score reflecting hospital quality. Most of the metrics that we developed using this approach involved some synthesis of distance and the probability of poor performance on all six indicators.

The approach of measuring hospital quality using multiple indicators of quality of care shares some similarities with multiple informant analysis.[38-40] Multiple informant analysis is used in settings in which information on a given condition are provided by different sources (ie, the informants). For instance, a child's psychological state may be ascertained by interviewing parents, teachers, and clinicians. From a provider profiling perspective, the different indicators can be seen as different informants on the quality of care provided by a given hospital. However, multiple informant analysis differs from provider profiling in that in multiple informant analysis, the informants are seeking to answer the same question, whereas in profiling, the indicators truly are different aspects of hospital quality. Finally, multiple informant analysis differs in that it is generally interested in fitting regression models in which the outcomes are provided by different informants (but the regressors such as age or sex come from a single source) or in fitting regression models in which the regressors are provided by multiple informants (but the outcome variable comes from a single source). In contrast to this, provider profiling is typically trying to quantify variation in hospital performance and identify hospitals with outlying performance.

Directions for future research include creating methods that synthesize the multivariate response logistic regression model and the latent variable approach. One possible way to do so would be to have two latent variable models, one for those outcomes that require risk-adjustment (eg, death and hospital readmission) and one for process-of-care measures that do not require risk-adjustment. The latter model would implicitly assume that there is a perfect within-hospital correlation on the performance of the different process-of-care measures. This approach would suggest that hospital quality had two dimensions, rather than the single dimension assumed by the existing latent variable approaches.

In conclusion, we have developed multivariate response Bayesian random effects logistic regression models that can be used to compare the performance of healthcare providers on a set of binary indicators of healthcare quality. Use of this method allows one to formally quantify the magnitude of within-hospital correlation on the performance of different binary indicators. Furthermore, by using the Mahalanobis distance, one can quantify the distance of a given hospital from an average hospital.

ACKNOWLEDGEMENTS

This study was supported by ICES, which is funded by an annual grant from the Ontario Ministry of Health and Long-Term Care (MOHLTC). The opinions, results, and conclusions reported in this article are those of the authors and are independent from the funding sources. No endorsement by ICES or the Ontario MOHLTC is intended or should be inferred. This research was supported by an operating grant from the Canadian Institutes of Health Research (CIHR) (MOP 86508). Dr. Austin is supported by a Mid-Career Investigator award from the Heart and Stroke Foundation. Dr. Leckie is supported by a standard grant from the UK Economics and Social Research Council (ES/R010285/1). Dr. Lee is supported by a Mid-Career Investigator award from the Heart and Stroke Foundation and is the Ted Rogers Chair in Heart Function Outcomes. The EFFECT data used in the study were funded by a CIHR Team Grant in Cardiovascular Outcomes Research (Grant numbers CTP79847 and CRT43823).

DATA AVAILABILITY STATEMENT

The data set from this study is held securely in coded form at ICES. While data sharing agreements prohibit ICES from making the data set publicly available, access may be granted to those who meet pre-specified criteria for confidential access, available at www.ices.on.ca/DAS.

ORCID

Peter C. Austin https://orcid.org/0000-0003-3337-233X

REFERENCES

1. New York State Department of Health. *Coronary Artery Bypass Graft Surgery in New York State 1989–1991*. Albany, NY: New York State Department of Health; 1992.
2. Jacobs FM. *Cardiac Surgery in New Jersey in 2002: A Consumer Report*. Trenton, NJ: Department of Health and Senior Services; 2005.
3. Massachusetts Data Analysis Center. *Adult Coronary Artery Bypass Graft Surgery in the Commonwealth of Massachusetts: Fiscal Year 2010 Report*. Boston, MA: Department of Health Care Policy, Harvard Medical School; 2012.
4. Naylor CD, Rothwell DM, Tu JV, Austin PC, The Cardiac Care Network Steering Committee. Outcomes of coronary artery bypass surgery in Ontario. In: Naylor CD, Slaughter PM, eds. *Cardiovascular Health and Services in Ontario: An ICES Atlas*. Toronto, NJ: Institute for Clinical Evaluative Sciences; 1999:189-198.
5. Pennsylvania Health Care Cost Containment Council. *Consumer Guide to Coronary Artery Bypass Graft Surgery*. Harrisburg, PA: Pennsylvania Health Care Cost Containment Council; 1995.
6. Luft HS, Romano PS, Remy LL, Rainwater J. *Annual Report of the California Hospital Outcomes Project*. Sacramento, CA: California Office of Statewide Health Planning and Development; 1993.
7. Pennsylvania Health Care Cost Containment Council. *Focus on Heart Attack in Pennsylvania: Research Methods and Results*. Harrisburg, PA: Pennsylvania Health Care Cost Containment Council; 1996.
8. Tu JV, Austin PC, Naylor CD, Iron K, Zhang H. Acute myocardial infarction outcomes in Ontario. In: Naylor CD, Slaughter PM, eds. *Cardiovascular Health and Services in Ontario: An ICES Atlas*. Toronto, Canada: Institute for Clinical Evaluative Sciences; 1999:83-110.
9. Tu JV, Donovan LR, Lee DS, et al. Effectiveness of public report cards for improving the quality of cardiac care: the EFFECT study: a randomized trial. *JAMA*. 2009;302(21):2330-2337.
10. Tu JV, Donovan LR, Lee DS. *Quality of Cardiac Care in Ontario*. Toronto, Ontario: Institute for Clinical Evaluative Sciences; 2004.
11. Landrum MB, Bronskill SE, Normand SL. Analytic methods for constructing cross-sectional profiles of health care providers. *Health Serv Outcomes Res Methodol*. 2000;1(1):23-47.
12. Landrum MB, Normand SL, Rosenheck RA. Selection of related multivariate means: monitoring psychiatric care in the Department of Veterans Affairs. *J Am Stat Assoc*. 2003;98(461):7-16.
13. Teixeira-Pinto A, Normand SL. Statistical methodology for classifying units on the basis of multiple-related measures. *Stat Med*. 2008;27(9):1329-1350.
14. Stone CA, Zhu X. *Bayesian Analysis of Item Response Theory Models Using SAS*. Cary, NC: SAS Institute Inc.; 2015.
15. Dunson DB. Bayesian latent variable models for clustered mixed outcomes. *J R Stat Soc B*. 2000;62(2):355-366.
16. O'Malley AJ, Normand SL, Kuntz RE. Application of models for multivariate mixed outcomes to medical device trials: coronary artery stenting. *Stat Med*. 2003;22(2):313-336.
17. *Markov Chain Monte Carlo in Practice*. London, UK: Chapman & Hall; 1996.
18. Eagle KA, Lim MJ, Dabbous OH, et al. A validated prediction model for all forms of acute coronary syndrome: estimating the risk of 6-month postdischarge death in an international registry. *JAMA*. 2004;291(22):2727-2733.
19. Geweke J. Evaluating the accuracy of sampling based approaches to the calculation of posterior moments. In: Bernardo JM, Berger JO, Dawid AP, Smith AFM, eds. *Bayesian Statistics*. Vol 4. Oxford, UK: Oxford University Press; 1992:169-193.
20. Snijders T, Bosker R. *Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling*. 2nd ed. London, UK: Sage Publications; 2012.
21. Goldstein H, Browne W, Rasbash J. Partitioning variation in generalised linear multilevel models. *Understanding Stat*. 2002;1:223-232.
22. Austin PC, Merlo J. Intermediate and advanced topics in multilevel logistic regression analysis. *Stat Med*. 2017;36(20):3257-3277.
23. Cohen J. *Statistical Power Analysis for the Behavioural Sciences*. 2nd ed. Hillsdale, NJ: Lawrence Erlbaum Associates; 1988.
24. Sheskin DJ. *Handbook of Parametric and Nonparametric Statistical Procedures*. 3rd ed. Boca Raton, FL: Chapman & Hall; 2004.
25. Everitt BS. *The Cambridge Dictionary of Statistics*. 2nd ed. Cambridge, UK: Cambridge University Press; 2002.
26. Christiansen CL, Morris CN. Improving the statistical approach to health care provider profiling. *Ann Intern Med*. 1997;127(8, pt 2):764-768.
27. Normand SLT, Glickman ME, Gatsonis CA. Statistical methods for profiling providers of medical care: issues and applications. *J Am Stat Assoc*. 1997;92(439):803-814.
28. Austin PC, Naylor CD, Tu JV. A comparison of a Bayesian vs. a frequentist method for profiling hospital performance. *J Eval Clin Pract*. 2001;7(1):35-45.
29. Austin PC. A comparison of Bayesian methods for profiling hospital performance. *Med Decis Making*. 2002;22(2):163-172.
30. Austin PC. Bayes rules for optimally using Bayesian hierarchical regression models in provider profiling to identify high-mortality hospitals. *BMC Med Res Methodol*. 2008;8:30.
31. Austin PC, Brunner LJ. Optimal Bayesian probability levels for hospital report cards. *Health Serv Outcomes Res Methodol*. 2008;8:80-97.
32. Austin PC. The reliability and validity of Bayesian methods for hospital profiling: a Monte Carlo assessment. *J Stat Plan Infer*. 2005;128:109-122.
33. Robinson JW, Zeger SL, Forrest CB. A hierarchical multivariate two-part model for profiling providers' effects on health care charges. *J Am Stat Assoc*. 2006;101(475):911-923.
34. Goldstein H. *Multilevel Statistical Models*. 4th ed. West Sussex, UK: John Wiley & Sons Ltd; 2011.
35. Leckie G. Avoiding bias when estimating the consistency and stability of value-added school effects using multilevel models. *J Educ Behav Stat*. 2018;43:440-468.

36. Goldstein H, Carpenter J, Kenward MG, Levin KA. Multilevel models with multivariate mixed response types. *Stat Model*. 2009;9(3):173-197.
37. Prior L, Goldstein H, Leckie G. School value-added models for multivariate academic and non-academic outcomes: a more rounded approach to using student data to inform school accountability. 2020. arXiv:2001.01996.
38. Lash TL, Thwin SS, Horton NJ, Guadagnoli E, Silliman RA. Multiple informants: a new method to assess breast cancer patients' comorbidity. *Am J Epidemiol*. 2003;157(3):249-257.
39. Horton NJ, Fitzmaurice GM. Regression analysis of multiple source and multiple informant data from complex survey samples. *Stat Med*. 2004;23(18):2911-2933.
40. Horton NJ, Laird N, Zahner GEP. Use of multiple informant data as a predictor in psychiatric epidemiology. *Int J Methods Psychiatr Res*. 1999;8(1):6-18.

How to cite this article: Austin PC, Lee DS, Leckie G. Comparing a multivariate response Bayesian random effects logistic regression model with a latent variable item response theory model for provider profiling on multiple binary indicators simultaneously. *Statistics in Medicine*. 2020;39:1390–1406. https://doi.org/10.1002/sim.8484