Evaluating the performance of Australian and New Zealand intensive care units in 2009 and 2010

J. Kasza, a*† J. L. Moran, b P. J. Solomon a for the ANZICS Centre for Outcome and Resource Evaluation (CORE) of the Australian and New Zealand Intensive Care Society (ANZICS)

The Australian and New Zealand Intensive Care Society Adult Patient Database (ANZICS APD) is one of the largest databases of its kind in the world and collects individual admissions’ data from intensive care units (ICUs) around Australia and New Zealand. Use of this database for monitoring and comparing the performance of ICUs, quantified by the standardised mortality ratio, poses several theoretical and computational challenges, which are addressed in this paper. In particular, the expected number of deaths must be appropriately estimated, the ICU casemix adjustment must be adequate, statistical variation must be fully accounted for, and appropriate adjustment for multiple comparisons must be made. Typically, one or more of these issues have been neglected in ICU comparison studies. Our approach to the analysis proceeds by fitting a random coefficient hierarchical logistic regression model for the inhospital death of each patient, with patients clustered within ICUs. We anticipate the majority of ICUs will be estimated as performing ‘usually’ after adjusting for important clinical covariates. We take as a starting point the ideas in Ohlssen et al and estimate an appropriate null model that we expect these ICUs to follow, taking a frequentist rather than a Bayesian approach. This methodology allows us to rigorously account for the aforementioned statistical issues and to determine if there are any ICUs contributing to the Australian and New Zealand Intensive Care Society database that have comparatively unusual performance. In addition to investigating the yearly performance of the ICUs, we also estimate changes in individual ICU performance between 2009 and 2010 by adjusting for regression-to-the-mean. Copyright © 2013 John Wiley & Sons, Ltd.

Keywords: hierarchical models; institutional comparisons; intensive care; mortality rates; regression-to-the-mean

1. Introduction

The need for monitoring and comparing hospital performance in Australia has recently been emphasised, with the Australian Government committing to publicly report hospital outcomes, [1, 2]. Comparing intensive care units (ICUs) is of particular importance, given the expense associated with intensive care and the severity of the cases treated in these units.

We compare the performance of ICUs contributing to the Australian and New Zealand Intensive Care Society Adult Patient Database (ANZICS APD), [3, 4], in 2009 and 2010, and use risk-adjusted mortality as an indicator of performance. We take as our starting point the method proposed by Ohlssen et al. for the identification of providers with divergent performance [5]. By estimating a null model that the performance indicators of usually performing ICUs are expected to follow, and estimating performance indicators and their variances directly from this fitted model, we are able to identify any ICUs that are characterised by unusual mortality. In addition, by generalising the method of Jones & Spiegelhalter for

*School of Mathematical Sciences, University of Adelaide, Adelaide, SA 5005, Australia
bDepartment of Intensive Care Medicine, The Queen Elizabeth Hospital, Woodville, SA 5011, Australia
†Correspondence to: Jessica Kasza, School of Mathematical Sciences, University of Adelaide, Adelaide, South Australia 5005, Australia.
E-mail: jessica.kasza@adelaide.edu.au

Copyright © 2013 John Wiley & Sons, Ltd.
monitoring changes in provider performance while controlling for regression-to-the-mean, [6], we are able to determine if any ICUs have undergone significant changes in mortality between 2009 and 2010.

Typically, in identifying ICUs with unusual performance, standardised mortality ratios (SMRs) are used as performance indicators. The expected number of deaths for each ICU is estimated by summing patient mortality probabilities that have been adjusted for the severity of a patient’s condition, and although some algorithms do account for other covariates, such as hospital location and length of time in hospital before admission to the ICU, there are clinically important covariates that are unaccounted for. To determine which ICUs have unusual SMRs, 95% or 99% confidence intervals are calculated, or funnel plots, [7], with 95% or 99% prediction limits are constructed, for example, [8–12]. The expected number of deaths is usually assumed to be fixed when calculating the variance of the SMRs. Those ICUs with confidence intervals that do not contain one, or with SMRs that fall outside of the prediction limits on the funnel plots, are then identified as unusual performers.

Even though the use of hierarchical models in the comparison of healthcare providers has become standard, for example, [13–15], such models, with patients clustered within ICUs, are not always used in the comparison of ICU performance. When hierarchical models are used, methods that estimate a null model describing the mortality experience in usually performing ICUs, to determine which ICUs have comparatively unusual performance, have not been implemented.

As discussed by [5, 14], when a hierarchical model is used in the identification of unusually performing providers, there are two approaches to determining which providers have unusual values of performance indicators. The first approach fits a random effects distribution that encompasses all of the variation between providers, and identifies those providers that have extreme random effects with respect to this model. The second approach involves fitting a random effects distribution that the random effects of the majority of providers are expected to follow, and identifies any providers that have outlying random effects with respect to this model, using hypothesis testing. We anticipate that the majority, if not all ICUs, contributing to the ANZICS APD in 2009 and 2010 will have usual, although not identical, performance, whereas some of the contributing ICUs may be characterised by unusually poor or good performance. The interest is in identifying those ICUs that are characterised by unusual performance, and so we estimate a random coefficient hierarchical logistic regression model for the in-hospital death of each patient, clustering patients within ICUs, with a random effects distribution that describes the mortality experience in an ICU with usual performance.

This model must be adequately adjusted for casemix, [16], so age, sex, patient severity score, patient diagnostic category, patient surgical status, patient ventilation status, source of admission to the ICU, geographical locality, ICU level, annual ICU volume, clinically meaningful interactions and a random intercept and random coefficient for the patient severity score are included in the model. The patient severity score we use is the third revision of the Acute Physiology and Chronic Health Evaluation (APACHE III) score [17], which is computed using the patient’s worst values during the first 24 h post admission to the ICU. To obtain confidence intervals with better coverage properties, following [18], we use logarithms of the SMRs (log-SMRs), instead of SMRs, as performance indicators. When calculating the variances of the log-SMRs, uncertainty in both the observed and expected numbers of deaths are accounted for. The multiplicity of hypothesis tests is accounted for by controlling the false discovery rate (FDR), [19]. We use funnel plots to display results, [7], and use the method of [20] to place thresholds controlling the FDR on the plots.

In addition to examining the yearly performance of ICUs, we also investigate and compare the changes in performance of ICUs between 2009 and 2010. An ICU not identified as having unusual performance in 2009 and again in 2010 may still have had a significant deterioration in performance, which should be investigated. Similarly, an ICU identified as a poor performer in two consecutive years may have significantly improved performance between those years, the merit of which should be recognised. We extend the method of [6], which accounts for regression-to-the-mean in the comparison of performance indicators estimated for two distinct periods when provider-level data is available, to the ANZICS APD patient-level data. In this way, we are able to determine if there were any ICUs with significant deteriorations or improvements in performance between 2009 and 2010.

The paper is organised as follows. In Section 2, we describe the ANZICS APD and the process used to find a good model for risk adjustment. In Section 3, after discussing previous approaches to the identification of ICUs with unusual performance, we describe the method we have used to identify ICUs with unusual performance, and assess the performance of ICUs. Changes in the performance of ICUs between 2009 and 2010 are compared in Section 4, and in Section 5, we present a discussion of our results.
2. The Australian and New Zealand Intensive Care Society Adult Patient Database and choice of risk-adjusted model

The ANZICS APD collects individual admissions data from ICUs in Australia and New Zealand and is one of the largest databases of its kind in the world. The database is voluntary, and data collection and quality procedures have been described in [4]. Our interest is in the current performance of ICUs, so we consider the most recently available data, from 2009 and 2010. The interrogation of the ANZICS APD to define an appropriate dataset has previously been described [3, 21, 22]. We follow these previous studies and exclude patients with unknown hospital mortality status and discharge date (4023 patients), patients with an ICU stay of ≤ 4 h (2110 patients) and patients < 16 years of age (3822 patients). Additionally, records for 10 170 separate hospital readmissions were excluded. ICUs with fewer than 150 patients in either 2009 or 2010 were excluded to ensure estimation stability. Of the 139 ICUs contributing to the APD in 2009, 24 were excluded, and of the 142 ICUs contributing in 2010, 27 were excluded.

The final dataset used for analysis consisted of records for 163 795 patients from 115 ICUs. The mean age was 61.7 (standard deviation 18.2) years, mean APACHE III severity score was 51.3 (27.2); 10.2% of patients died in hospital, 58.1% were male, 42.1% were ventilated within the first 24 h of admission to the ICU, 7.7% of patients were transferred from another hospital, and 58.8% of patients were nonsurgical. Patients were categorised into eight diagnostic groups, a consolidation of the APACHE diagnostic categories, with the three largest categories being the cardiovascular category with 24.6% of patients, the gastrointestinal category with 17.5% of patients and the respiratory category with 15.3% of patients. Of the 12 610 patients transferred between hospitals, 1308 patients were transferred directly between ICUs. Given the lack of linkage of patient records between hospitals in the APD, it is unknown from which hospitals these patients were transferred. The 1308 patients transferred between ICUs are potentially evaluated twice, but given the small number of such patients, the impact on the results, overall and for particular ICUs, will be small.

The ICUs were classified as being situated in hospitals located in rural/regional areas, hospitals in metropolitan areas, tertiary hospitals (larger, usually academic, hospitals) or in privately funded hospitals. ICUs in tertiary hospitals were the largest contributors, with 43.1% of patients from such ICUs. Geographical locations of ICUs were recorded as New Zealand or as one of the eight states/territories of Australia. Most patients were from Australian ICUs, with 5.6% of patients from New Zealand ICUs. Of the states and territories of Australia, New South Wales was the largest contributor, with 31.2% of patients, followed by Victoria, with 25.3% of patients, and Queensland, with 22.8% of patients. Patient characteristics are further described in Table I, and the breakdown of patients by ICU hospital level and location is given in Table I of the web-based Supporting Information‡. The median number of patients in each ICU over the two-year period was 1194, with a minimum of 319 patients and a maximum of 3881 patients.

If mortality experience could be thought of as independent from year to year, it would be appropriate to fit separate models to the data from 2009 and the data from 2010. However, given the continuity of many hospital-level and ICU-level factors from year to year, many of which are outside the control of ICUs, we do not expect such independence, so we consider models incorporating data from both years, with patients from both years clustered within ICUs.

We use inhospital mortality as the outcome of interest. Patients from both 2009 and 2010 are nested within ICUs, and we let

\[ Y_{ij} = \begin{cases} 1 & \text{if patient } i \text{ in ICU } j \text{ dies in hospital} \\ 0 & \text{otherwise,} \end{cases} \quad i = 1, \ldots, n_j, \quad j = 1, \ldots, 115. \]

Although a more appropriate measure of ICU performance may be mortality up to 30 days after discharge from hospital, lack of linkage between the data in the ANZICS APD and the state death registers means that such measures are not currently available.

Letting \( X_{ij} \) represent the (centred) APACHE III severity score of patient \( i \) in ICU \( j \) and \( U_j \) the vector of covariates for patient \( i \) in ICU \( j \), the hierarchical logistic regression model is given by

\[
Y_{ij} | X_{ij}, U_j \sim \text{Bernoulli}(P_{ij}), \quad \log \left( \frac{P_{ij}}{1 - P_{ij}} \right) = \gamma' X_{ij} + U_j' (1, X_{1ij})', \quad U_j \sim N_2(0, \Sigma). \tag{1}
\]

‡Supporting information may be found in the online version of this article.

Table I. Characteristics of 163,795 patients used for analysis. Age and third revision of the Acute Physiology and Chronic Health Evaluation score are given as mean (standard deviation). Patient volume is total patient volume of intensive care units over 2009 and 2010, given as median (interquartile range).

<table>
<thead>
<tr>
<th>Characteristics</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age (years)</td>
<td>61.7 (18.2)</td>
</tr>
<tr>
<td>APACHE III score</td>
<td>51.3 (27.2)</td>
</tr>
<tr>
<td>ICU mortality (%)</td>
<td>6.5</td>
</tr>
<tr>
<td>Hospital mortality (%)</td>
<td>10.2</td>
</tr>
<tr>
<td>2009–2010 patient volume</td>
<td>1194 (1153)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Ventilation</th>
<th>n (%)</th>
<th>Hospital mortality (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Not ventilated</td>
<td>94,802 (57.9)</td>
<td>6.3</td>
</tr>
<tr>
<td>Ventilated</td>
<td>68,993 (42.1)</td>
<td>15.6</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Gender</th>
<th>n (%)</th>
<th>Hospital mortality (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>95,128 (58.1)</td>
<td>10.3</td>
</tr>
<tr>
<td>Female</td>
<td>68,667 (41.9)</td>
<td>10.1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Patient surgical status</th>
<th>n (%)</th>
<th>Hospital mortality (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nonsurgical</td>
<td>96,364 (58.8)</td>
<td>13.9</td>
</tr>
<tr>
<td>Elective surgical</td>
<td>47,847 (29.2)</td>
<td>2.4</td>
</tr>
<tr>
<td>Emergency surgical</td>
<td>19,584 (12.0)</td>
<td>11.5</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Patient diagnostic category</th>
<th>n (%)</th>
<th>Hospital mortality (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cardiovascular</td>
<td>40,230 (24.6)</td>
<td>15.8</td>
</tr>
<tr>
<td>Gastrointestinal</td>
<td>28,639 (17.5)</td>
<td>8.9</td>
</tr>
<tr>
<td>Metabolic</td>
<td>11,424 (7.0)</td>
<td>3.2</td>
</tr>
<tr>
<td>Neurologic</td>
<td>18,216 (11.1)</td>
<td>12.6</td>
</tr>
<tr>
<td>Respiratory</td>
<td>25,057 (15.3)</td>
<td>13.9</td>
</tr>
<tr>
<td>Trauma</td>
<td>90,30 (5.5)</td>
<td>8.3</td>
</tr>
<tr>
<td>Renal/Genitourinary</td>
<td>86,12 (5.3)</td>
<td>4.8</td>
</tr>
<tr>
<td>Hematological</td>
<td>22,587 (13.8)</td>
<td>2.2</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>ICU source</th>
<th>n (%)</th>
<th>Hospital mortality (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>No transfer</td>
<td>151,185 (92.3)</td>
<td>9.7</td>
</tr>
<tr>
<td>Hospital transfer</td>
<td>12,610 (7.7)</td>
<td>16.5</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>ICU hospital level</th>
<th>n (%)</th>
<th>Hospital mortality (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rural</td>
<td>21,348 (13.0)</td>
<td>10.1</td>
</tr>
<tr>
<td>Metropolitan</td>
<td>29,294 (17.9)</td>
<td>13.2</td>
</tr>
<tr>
<td>Tertiary</td>
<td>70,587 (43.1)</td>
<td>12.7</td>
</tr>
<tr>
<td>Private</td>
<td>42,566 (26.0)</td>
<td>4.1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>ICU location</th>
<th>n (%)</th>
<th>Hospital mortality (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Northern Territory</td>
<td>2,153 (1.3)</td>
<td>10.0</td>
</tr>
<tr>
<td>New South Wales</td>
<td>51,046 (31.2)</td>
<td>10.5</td>
</tr>
<tr>
<td>Australian Capital Territory</td>
<td>40,14 (2.5)</td>
<td>9.5</td>
</tr>
<tr>
<td>South Australia</td>
<td>12,772 (7.8)</td>
<td>13.7</td>
</tr>
<tr>
<td>Victoria</td>
<td>41,426 (25.3)</td>
<td>10.3</td>
</tr>
<tr>
<td>Western Australia</td>
<td>32,79 (2.0)</td>
<td>11.0</td>
</tr>
<tr>
<td>New Zealand</td>
<td>91,64 (5.6)</td>
<td>13.4</td>
</tr>
<tr>
<td>Queensland</td>
<td>37,337 (22.8)</td>
<td>7.6</td>
</tr>
<tr>
<td>Tasmania</td>
<td>2,604 (1.6)</td>
<td>11.6</td>
</tr>
</tbody>
</table>

The random effects formulation explicitly allows for some variation between the mortality experiences at each ICU: even patients with identical characteristics at different ICUs with usual performance are not expected to have the same mortality experience because of unknown ICU-level explanatory variables. Indeed, under the model in Equation (1), some degree of variation is considered usual.

To properly adjust for risk, patient-level covariates, hospital level (tertiary, metropolitan, rural or private), volume (treated as a continuous variable) and location (New Zealand or state of Australia) were included in the model. Patient-level covariates included were age, age squared, APACHE III score, APACHE III score squared, sex, patient diagnostic category, patient surgical status, patient ventilation status and source of admission to the ICU. Patient-level by patient-level interactions included in the initial model were APACHE III by each of age, ICU source, patient category, surgical status and...
ventilation status; ventilation status by each of age, sex, patient category and surgical status; and surgical status by each of patient category and ICU source. Also included in the model were interactions between hospital level and each of patient surgical status and ICU source. More detail on the included covariates is provided in Appendix A.

To build an appropriate model for risk-adjustment, patients were split into an 80% training set and a 20% test set. Model parameters were estimated using the xtmelogit module in Stata 12, (StataCorp, College Station, TX, USA) [23], and for this model-building stage, the Laplace approximation was used. The first model considered (Model 1) included all covariates and two-way interactions as described previously, but excluded a random coefficient for APACHE III score. Inclusion of a random coefficient for APACHE III score (Model 2) led to a reduction in AIC and BIC. Interactions insignificant at the 5% level were ICU source by each of APACHE III score and surgical status, and ventilation status by each of age, sex and surgical status. Removing these interactions from the random coefficient model to get Model 3 led to a further decreases in AIC and BIC. No interactions were insignificant at the 5% level in Model 3, which contained 110 fixed effects parameters and three random effects parameters. Statistics for each model using the training and test sets are shown in Table II. For all considered models, discrimination was excellent as measured by the area under the ROC curve, and the Hosmer–Lemeshow goodness-of-fit statistics indicate a good fit of the model to the data [24]. We use Model 3 as our model for risk adjustment, and in Section 3.2.1, it is re-estimated using the entire dataset.

### 3. Identifying intensive care units with unusual performance

#### 3.1. Usual approaches

Usually, in studies identifying ICUs with unusual performance, patient mortality probabilities are estimated using one of two methods. The first method, discussed in [8, 25], estimates mortality probabilities using a scoring algorithm, such as the APACHE III algorithm. The ANZICS Centre for Outcome and Resource Evaluation (ANZICS CORE) take this approach in their evaluations of ICUs contributing to the ANZICS APD, using the APACHE III-J severity score, the tenth recalibration of the third revision of the APACHE score and other variables as in the algorithm, to estimate mortality probabilities [12, 26]. The second method fits a logistic regression model similar to that in Equation (1), where the log odds of inhospital deaths of patients are dependent upon a severity score [13, 27]. Such models have been extended to account for the hierarchical structure of patient-level data and may include additional covariates such as patient gender and age and a random coefficient for the severity score [9, 13].

Both of these methods suffer from statistical inadequacies. With respect to the first method, the scoring systems used fail to account for the clustering of patients within ICUs, [9], and these scoring systems may not be appropriately calibrated for the population under study [27–29]. Although the second method is an improvement on the first with respect to these problems, adjustment for risk tends to be inadequate, with many clinically important covariates and interactions between them often not included in the models. The most significant inadequacy of these typical methods is that the presence of unusual ICUs is not accounted for in the estimation of a model describing the mortality experience in ICUs with usual performance [5, 14]. A model estimated using data from both usually and unusually performing ICUs is likely to have inflated random effect variance estimates, which will lead to an overestimation of the usual levels of variation in SMRs.

### Table II. Model statistics for selecting a risk-adjustment model: AIC, BIC, area under the ROC curve and p-values for the Hosmer–Lemeshow statistic (H-L p-value). The numbers in brackets are the values for the 20% test set. Model 1: all covariates and two-way interactions, no random coefficient for the third revision of the Acute Physiology and Chronic Health Evaluation score. Model 2: Model 1 with a random coefficient for the third revision of the Acute Physiology and Chronic Health Evaluation score. Model 3: Model 2 with insignificant interactions removed.

<table>
<thead>
<tr>
<th>Model</th>
<th>AIC</th>
<th>BIC</th>
<th>ROC</th>
<th>H-L p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>56 174.85</td>
<td>56 918.37</td>
<td>0.8976 (0.8938)</td>
<td>0.1817 (0.5284)</td>
</tr>
<tr>
<td>2</td>
<td>56 131.73</td>
<td>56 894.82</td>
<td>0.8982 (0.8943)</td>
<td>0.1463 (0.4171)</td>
</tr>
<tr>
<td>3</td>
<td>56 123.25</td>
<td>56 817.86</td>
<td>0.8982 (0.8943)</td>
<td>0.1230 (0.4453)</td>
</tr>
</tbody>
</table>
Confidence intervals for SMRs are often displayed, despite the risk of misinterpretation of such plots [7,30]. Funnel plots are preferred and have been used in the identification of unusually performing ICUs [11,12]. To construct such plots, we require estimates of the variances of SMRs, but in the calculation of such variances, expected numbers of deaths are assumed to be fixed, as in [31,32]. Additionally, the need to account for multiple comparison issues, [19,33,34], has not been recognised in the ICU comparison literature [8–12].

### 3.2. Identifying unusual intensive care units contributing to the Australian and New Zealand Intensive Care Society Adult Patient Database in 2009 and 2010

In our approach to the identification of unusually performing ICUs contributing to the ANZICS APD in 2009 and 2010, we consider three summaries of ICU performance: the log-SMR for the two-year period over 2009 and 2010, and log-SMRs for each year (2009 and 2010 log-SMRs). The log-SMR of ICU \( j \) is given by the logarithm of number of observed deaths \( O_j \) divided by \( E_j \), the expected number of deaths.

\[
\text{log-SMR}_j = \log(\text{SMR}_j) = \log\left(\frac{O_j}{E_j}\right) = \log\left(\sum_{i=1}^{n_j} Y_{ij}\right) - \log\left(\sum_{i=1}^{n_j} P_{ij}\right).
\]

To estimate the log-SMR, we require estimates of the probabilities of inhospital deaths. We estimate an initial model on the basis of the data from all ICUs, then identify those ICUs with observed numbers of deaths that are not well-predicted by the model (‘potentially unusual ICUs’), and re-estimate the model accounting for the data from those ICUs. Our analysis is based on ideas of [5] but takes a classical, rather than Bayesian, approach. Additionally, we proceed with the identification of unusually performing ICUs by estimating performance indicators and their variances directly from the fitted model, rather than by simulating from the model.

Ohlssen et al. [5] discuss their methods as applied to models on the basis of both provider-level data and patient-level data. Their patient-level example is much smaller than ours: 6150 patients from eight providers, as compared with 163 795 patients from 115 providers. In their discussion of the methods, Jones & Spiegelhalter discuss the methods as applied to random effects models on the basis of provider-level data only [14].

#### 3.2.1. Stage 1: Identifying potentially unusual intensive care units.

We now briefly describe the three stages of the method, as applied to the ANZICS APD 2009 and 2010 data. In the first stage, ICUs with potentially divergent performance were identified. The hierarchical logistic regression model, given in Equation (1), was fitted to the entire dataset using the xtmelogit module in Stata 12, with seven point adaptive quadrature, [35], using Model 3 to adjust for risk. Parameter estimates are provided in Table II of the web-based Supporting Information.

Cross-validation is the preferred method for determining how well the data from each ICU is fitted by the stage 1 model, but because such an approach was not computationally feasible for this dataset, an approximate cross-validation approach was used. For each ICU \( j = 1, \ldots, 115 \), random effects were simulated from the fitted random effects distribution 5000 times, denoted by \( U_k^j \), \( k = 1, \ldots, 5000 \). Note that the parameters of the fitted random effects density are estimated given the data from all ICUs. The probability of death of each patient in ICU \( j \) given these simulated random effects was calculated:

\[
p_{ij}^k = \frac{\exp\left(\hat{\beta}'X_{ij} + (U_k^j)'(1, X_{1ij})'\right)}{1 + \exp\left(\hat{\beta}'X_{ij} + (U_k^j)'(1, X_{1ij})'\right)},
\]

and \( Y_{ij}^k \), the inhospital death of patient \( i \) in ICU \( j \) for iteration \( k \), was simulated from a Bernoulli distribution with probability of success \( p_{ij}^k \). The simulated number of deaths for ICU \( j \) was then calculated by summing the \( Y_{ij}^k \), for \( i = 1, \ldots, n_j \). Finally, for each ICU, the proportion of times the observed number of deaths exceeded the simulated number of deaths was obtained, giving an approximate \( p \)-value for each ICU, where we follow [5] in their use of the term ‘approximate \( p \)-value’. The nominal null hypothesis being tested here is that the SMR for each ICU is equal to one, and if the estimated model predicts the number of deaths for an ICU accurately, the simulated number of deaths would exceed the observed number of deaths in approximately 50% of the simulations. The approximate \( p \)-values measure how
Table III. Potentially unusual intensive care units (ICUs) in the 2009 and 2010 dataset. The final column gives the approximate \( p \)-values: an ICU with \( p \)-value < 0.05 is potentially over-performing, and an ICU with \( p \)-value > 0.95 is potentially under-performing.

<table>
<thead>
<tr>
<th>ICU identifier</th>
<th>Hospital level</th>
<th>( p )-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>Private</td>
<td>0.0166</td>
</tr>
<tr>
<td>57</td>
<td>Private</td>
<td>0.0182</td>
</tr>
<tr>
<td>48</td>
<td>Rural</td>
<td>0.0202</td>
</tr>
<tr>
<td>72</td>
<td>Rural</td>
<td>0.0220</td>
</tr>
<tr>
<td>108</td>
<td>Private</td>
<td>0.0258</td>
</tr>
<tr>
<td>49</td>
<td>Metropolitan</td>
<td>0.0290</td>
</tr>
<tr>
<td>19</td>
<td>Private</td>
<td>0.0422</td>
</tr>
<tr>
<td>45</td>
<td>Tertiary</td>
<td>0.0494</td>
</tr>
<tr>
<td>93</td>
<td>Private</td>
<td>0.9658</td>
</tr>
<tr>
<td>81</td>
<td>Private</td>
<td>0.9770</td>
</tr>
<tr>
<td>44</td>
<td>Private</td>
<td>0.9874</td>
</tr>
<tr>
<td>16</td>
<td>Private</td>
<td>0.9952</td>
</tr>
</tbody>
</table>

well the estimated model predicts the number of deaths for each ICU. ICUs with approximate \( p \)-values less than 0.05 were labelled as potential over performers, and ICUs with approximate \( p \)-values greater than 0.95 were labelled as potential under performers. Twelve ICUs were identified as being potentially over-performing or under-performing, and the approximate \( p \)-values of these ICUs and their hospital levels are given in Table III. The model was fitted again in Stage 2, accounting for the presence of these potentially unusual ICUs.

It is possible to expand the approximate \( p \)-value cut-offs for potentially unusual ICUs, and the choice of these cut-offs is dependent upon application. For this example, expanding the approximate \( p \)-value cut-offs to 0.1 and 0.9 led to collinearities in the Stage 2 model because of all of the ICUs in one Australian jurisdiction being labelled as potentially unusual.

Figure 1 shows the total patient volumes of the ICUs over 2009 and 2010 and indicates that the volumes of ICUs labelled as potentially unusual are spread throughout the volumes of all ICUs. Figure 2 displays the standardised empirical Bayes modal predictions, \([36]\), for the standardised random effects (the predicted random effects divided by their standard errors). Potentially unusual ICUs are identified by circles (poor performers) or diamonds (good performers). The pattern of potentially unusual ICUs in this plot indicates that the variance of the random intercept and random coefficient will decrease when the
Figure 2. The empirical Bayes modal predictions of the random effects, divided by standard errors. Potentially under-performing intensive care units are marked with circles, potentially over-performing intensive care units with diamonds.

data from these ICUs do not contribute to the estimation of the random effects density. Figure 2 shows the clear separation of the standardised random intercept predictions for potentially over-performing and under-performing ICUs; potentially under-performing ICUs have random intercept predictions above the average standardised intercept random effect prediction, whereas potentially over-performing units have predictions below the average.

3.2.2. Stage 2: Estimating a null model. In the second stage of the method, the presence of the potentially unusual ICUs identified at the first stage are accounted for in the estimation of the parameters for the random effects density. The aim is to estimate a null random effects density from which the random effects associated with usually performing ICUs are assumed to be drawn, thereby minimising the influence of those ICUs deemed potentially unusual at Stage 1. We estimate a model where only data from the supposedly in-control units contribute to the estimation of the random effect density, whereas data from all units contribute to the estimation of the fixed part of the model, as in [37], for example. This model may be written as

\[
Y_{ij} | X_{ij}, U_j \sim \text{Bernoulli}(P_{ij}), \quad U_j \sim N(0, \Sigma),
\]

\[
\log\left(\frac{P_{ij}}{1 - P_{ij}}\right) = (1 - b_j)\gamma_0 + b_j\gamma_{0j} + (1 - b_j)\gamma_1 X_{ij} + b_j\gamma_{1j} X_{ij}
\]

\[
+ \sum_{h=2}^{110} \gamma_h X_{hij} + (1 - b_j)U_{0j} + (1 - b_j)U_{1j} X_{ij},
\]

\[
b_j = \begin{cases} 
1 & \text{if ICU } j \text{ was identified as potentially unusual at Stage 1} \\
0 & \text{otherwise.}
\end{cases}
\]

Separate fixed intercepts and APACHE III coefficients were estimated for each of the potentially unusual ICUs listed in Table III. The estimates of the random effects variance matrix \( \Sigma \) for the Stage 1 and 2 models are given in Table IV, which shows that the variance of the random intercept is markedly reduced and the variance of the random coefficient for APACHE III is reduced by a small amount at Stage 2. Estimates of all parameters in the Stage 1 and 2 models are provided in Table 2 of the web-based Supporting Information.

The estimated null model is given by

\[
\log\left(\frac{P_{ij}}{1 - P_{ij}}\right) = \hat{\gamma}_0 + \hat{\gamma}_1 X_{1ij} + \sum_{h=2}^{110} \hat{\gamma}_h X_{hij} + U_{0j} + U_{1j} X_{1ij} = \hat{\gamma}' X_{ij} + U_j'(1, X_{ij})', \quad U_j \sim N \left(0, \hat{\Sigma}\right).
\]
Table IV. Estimates and standard errors of the random effect density parameters for the Stage 1 and Stage 2 models.

<table>
<thead>
<tr>
<th></th>
<th>Stage 1</th>
<th>Estimate</th>
<th>SE</th>
</tr>
</thead>
<tbody>
<tr>
<td>APACHE III coefficient variance</td>
<td>0.0000318</td>
<td>7.74 x 10^{-6}</td>
<td></td>
</tr>
<tr>
<td>Intercept variance</td>
<td>0.0542223</td>
<td>0.0115764</td>
<td></td>
</tr>
<tr>
<td>Covariance of APACHE III and intercept</td>
<td>-0.0002500</td>
<td>0.0023700</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Stage 2</td>
<td>Estimate</td>
<td>SE</td>
</tr>
<tr>
<td>APACHE III coefficient variance</td>
<td>0.0000313</td>
<td>7.84 x 10^{-6}</td>
<td></td>
</tr>
<tr>
<td>Intercept variance</td>
<td>0.0271328</td>
<td>0.0073427</td>
<td></td>
</tr>
<tr>
<td>Covariance of APACHE III and Intercept</td>
<td>-0.0001876</td>
<td>0.0001879</td>
<td></td>
</tr>
</tbody>
</table>

SE, standard error; APACHE III, third revision of the Acute Physiology and Chronic Health Evaluation score.

Note that the separate intercepts $\gamma_{0j}$ and APACHE III coefficients $\gamma_{1j}$ for potentially unusual ICUs do not enter into this estimated null model.

For an ICU that was not deemed to be potentially unusual at Stage 1, the SMR was calculated using the estimated null model as

$$SMR_j = \frac{O_j}{E_j} = \frac{\sum_{i=1}^{n_j} Y_{ij}}{\sum_{i=1}^{n_j} \hat{P}_{ij}}, \quad \hat{P}_{ij} = \frac{\exp\left(\hat{\gamma}'X_{ij} + \hat{U}_j'(1, X_{1ij})\right)}{1 + \exp\left(\hat{\gamma}'X_{ij} + \hat{U}_j'(1, X_{1ij})\right)},$$

(2)

where $\hat{U}_j$ are the empirical Bayes modal predictions of the random effects. The potentially unusual ICUs were modelled without random effects, so for these ICUs, a usual ICU $k$ was randomly selected, and the random effect predictions from that ICU used to calculate the expected number of deaths for potentially unusual ICU $j$, substituting $\hat{O}_k$ for $\hat{O}_j$ in Equation (2).

3.2.3. Stage 3: Displaying results and accounting for multiple comparisons. To construct a funnel plot to display the log-SMRs, we assumed normality of the log-SMRs and used the method in [7, Appendix A.3.1] to construct the plot. In studies of this kind, when calculating the variance of performance indicators, it is usually assumed that the expected number of deaths are fixed. The variance of the log-SMRs plays an important role in determining which ICUs are deemed to have unusual performance, and is particularly important when monitoring changes in ICU performance over time. Figure 1 of the web-based Supporting Information shows that assuming a fixed number of expected deaths did result in underestimation of variances for this dataset. We have been unable to find a reference in the literature to our result on the variance estimator of the log-SMR, so we provide details in Appendix B.

To construct the funnel plots, we plotted each log-SMR against the effective sample size, which measures the variability of the log-SMR of each ICU relative to the total variability of all log-SMRs. In particular, the effective sample size of ICU $j$ is given by

$$n^{eff}_j = \frac{\sum_{i=1}^{115} n_i \hat{\sigma}_i^2 / 115}{\hat{\sigma}_j^2},$$

where $\hat{\sigma}_j^2$ is the variance of the estimated log-SMR of ICU $j$. Funnels are drawn controlling the FDR at 0.05, using the method of [20]. For comparison, funnels corresponding to the 95% prediction limits are also drawn.
Figure 3 displays the funnel plot for the log-SMRs for the 115 ICUs contributing to the ANZICS APD in 2009 and 2010, and the funnel plots for yearly performance are displayed in Figure 4. Log-SMRs for 2009 were calculated by restricting the sums in Equation (2) to those patients admitted to ICUs in 2009, and similarly for the 2010 log-SMRs. When the FDR is controlled at 0.05, ICUs 16, 44, 81 and 93 are identified as having unusually high log-SMRs over the two-year period, as is the case when 2009 admissions are considered. When the classical 95% prediction limits are used, in addition to the four ICUs with unusually high log-SMRs, ICUs 48, 49 and 100 are identified as having unusually low log-SMRs over

Figure 3. Funnel plot for the logarithms of the standardised mortality ratios (log-SMRs) for 2009 and 2010, calculated using the Stage 2 model. Intensive care unit 9 is marked with a triangle. The solid lines are the prediction limits when the false discovery rate is controlled at 0.05, and the dashed lines are the classical 95% prediction limits when multiple comparisons are not adjusted for.

Figure 4. Funnel plots for the 2009 logarithms of the standardised mortality ratios (log-SMRs; upper panel), and the 2010 log-SMRs (lower panel), using the Stage 2 model. The solid lines are the prediction limits when the false discovery rate is controlled at 0.05, and the dashed lines are the classical 95% prediction limits when multiple comparisons are not adjusted for. In both plots, intensive care unit 9 is marked with a triangle.
the two-year period and in 2009. In 2010, when the FDR is controlled at 0.05, no ICUs are identified as having unusual log-SMRs. When the classical limits are used, ICUs 9, 16, 44, 81 and 93 are identified as having unusually high log-SMRs, and ICU 48 is identified as having an unusually low log-SMR.

Both the log-SMRs and the effective sample sizes of ICUs 16, 44 and 93 are smaller in 2010 than in 2009, and the log-SMR and effective sample size of ICU 9 are larger in 2010 than in 2009. The changes in effective sample sizes indicate greater relative variability of the log-SMRs of ICUs 16, 44 and 93, and a reduced relative variability of the log-SMR of ICU 9 in 2010 from 2009. This helps explain why the results of the two-year analysis for ICUs 16, 44 and 93 are more similar to those of the 2009 analysis than those of the 2010 analysis, and why the two-year analysis for ICU 9 is more similar to the 2010 analysis than the 2009 analysis.

4. Identifying intensive care units with recent changes in performance

Consideration of the performance of ICUs at fixed points in time can sometimes result in inequitable classifications of unusual ICUs as good or poor performers. As discussed by [38], it is important to also consider the trajectory of the performance of providers and determine if the performance of each ICU is remaining steady, improving or deteriorating. We considered changes in the performance of ICUs, as indicated by the log-SMR, between 2009 and 2010.

Jones & Spiegelhalter noted that in assessing changes in recent institutional performance, investigators should account for regression-to-the-mean [6]. Instead of using a previous period’s performance indicator as a baseline from which to measure a change in performance, they calculate the expected value of the performance indicator in the most recent period given the performance indicator in the previous period as an adjusted baseline from which to measure the change. They show that the test based on this adjusted baseline is the more powerful of the two tests, although they do note that gains in power to detect changes in performance away from the mean come at the cost of a slight reduction in power to detect changes towards the mean. Their approach is based upon the use of aggregate provider-level data, which we generalise to the patient-level data of the ANZICS APD.

Let \( S_{1j} \) be the estimated 2009 log-SMR of ICU \( j \) and \( S_{2j} \) be the estimated 2010 log-SMR. Assuming normality, the marginal distributions are given by

\[
S_{1j} \sim N \left( 0, \sigma_{1j}^2 \right), \quad S_{2j} \sim N \left( 0, \sigma_{2j}^2 \right).
\]

The joint distribution of \( S_{1j} \) and \( S_{2j} \) and the corresponding standard test statistic is given by

\[
\left( \begin{array}{c} S_{1j} \\ S_{2j} \end{array} \right) \sim N \left( \begin{array}{c} 0 \\ 0 \end{array} \middle| \begin{array}{cc} \sigma_{1j}^2 & \sigma_{12j} \\ \sigma_{12j} & \sigma_{2j}^2 \end{array} \right), \quad Z_j^* = \frac{S_{2j} - S_{1j}}{\sqrt{\text{var}(S_{2j} - S_{1j})}} \sim N(0, 1).
\]

The conditional distribution of \( S_{2j} \) given \( S_{1j} = s_{1j} \) and the corresponding adjusted test statistic is given by

\[
S_{2j} \mid S_{1j} = s_{1j} \sim N \left( s_{1j} \frac{\sigma_{12j}}{\sigma_{1j}^2}, \sigma_{2j}^2 - \frac{\sigma_{12j}^2}{\sigma_{1j}^2} \middle| s_{1j} \right), \quad Z_j = \frac{S_{2j} - E[S_{2j} \mid S_{1j}]}{\sqrt{\text{var}(S_{2j} \mid S_{1j})}} \sim N(0, 1).
\]

An estimator for \( \sigma_{12j} \) is derived in Appendix C.

In Figure 5, for each of the 115 ICUs that contributed to the ANZICS database in 2009 and 2010, the numerators of the standard test statistics (bullets) and adjusted test statistics (circles) are plotted against the inverse of the denominators of the test statistics. The standard and adjusted tests for each ICU are connected by line segments. As in the examples in [6], for each ICU, the adjusted test statistic has greater precision than the standard test statistic, and most of the adjusted test statistic numerators are closer to zero than the standard test statistic numerators.

When the FDR is controlled at 0.05, no ICUs are identified as having undergone significant changes in performance between 2009 and 2010. If the 95% prediction limits were used, five ICUs would be identified as having undergone significant deteriorations in performance, and seven would be identified as having significantly improved. The ICU with the smallest \( p \)-value for this test is ICU 9, marked with a triangle in Figures 4 and 5, and this ICU would be deemed to have deteriorated between 2009 and 2010 were the 95% limits used. When the 95% limits were used, ICU 9 was found to have an unusually high log-SMR in 2010 but not in 2009.
Figure 5. Funnel plot for testing for a significant change in logarithms of the standardised mortality ratio (log-SMR) between 2009 and 2010. The bullets represent the standard test results, which are connected to the adjusted test results for each intensive care unit (open circles). The result for intensive care unit 9 is marked with a triangle. The solid lines are the prediction limits when the false discovery rate is controlled at 0.05, and the dashed lines are the classical 95% prediction limits when multiple comparisons are not adjusted for.

Notably, ICUs 16, 44, 81 and 93, identified as poor performers in 2009 but not in 2010, are not identified as having significantly changed performance between the 2 years. This may be due to there actually being no significant change in the performance of these ICUs between 2009 and 2010 but could also be due to the increase in the variance of the log-SMRs between 2009 and 2010. The performance of ICUs contributing to the APD in the years 2000 to 2010 is currently being assessed, and the results of that analysis will help to elucidate this matter.

5. Discussion

In this study, we have introduced a number of key concepts and principled statistical methods for comparing intensive care units. By estimating a null model for the inhospital mortality of patients, which adjusts for important casemix factors, we have been able to rigorously monitor the performance of ICUs contributing to the ANZICS APD in 2009 and 2010. We have also formally compared changes in the performance of ICUs in 2009 and 2010 using regression-to-the-mean techniques. By properly accounting for statistical issues, which are often neglected in the comparison of ICUs (and indeed in the comparison of healthcare providers more generally), we have considerably reduced the risk of incorrectly identifying ICUs as unusual, and the risk of failing to identify ICUs performing unusually well or unusually poor.

In Section 2, we suggest that the random intercept will capture unknown ICU-level explanatory variables. It is possible that these unknown ICU-level variables are related to performance in some way, and it is for this reason that a null random-effects density that describes performance in usually performing ICUs is strongly recommended.

The proposed method for the identification of ICUs with unusual performance detects those ICUs with outlying log-SMRs. Any ICUs identified as having unusually good or poor performance should be investigated further, to determine if such results may be explained by data of poor quality or by other factors unrelated to performance. The current Outlier Management Policy of ANZICS CORE, [39], outlines the process by which an ICU with an unusual performance indicator is investigated, and emphasises scrutiny of data quality.

The ICUs 16, 44, 81 and 93, found to have unusually poor performance in 2009 and over the two-year period, are candidates for such further investigation. In 2010, when the FDR was controlled at 0.05, these ICUs were not found to have unusual performance, although these four ICUs and ICU 9 would have been identified as unusual, had the classical 95% limits been used. No ICUs were found to have significant changes in performance between 2009 and 2010 when the FDR was controlled at 0.05, which may indicate that ICUs 16, 44, 81 and 93 do not have systematically poor performance. Further investigation of data quality and data from additional years will help to determine if these observed results are in fact due to truly poor performance or simply due to chance or a run of bad luck.
In their comparison of the performance of ICUs in 2010, the ANZICS CORE identified two ICUs as having unusual performance indicators, [12], in contrast to the results we have obtained. There are differences between the dataset used in our study and the dataset used in the 2010 ANZICS Annual Report, namely, all contributing ICUs were considered in the ANZICS report, whereas we only considered those ICUs that contributed at least 150 patients in both 2009 and 2010. Despite this difference, it is still informative to compare the outcomes of the two analyses. In the ANZICS report, separate funnel plots for each ICU level were constructed, and ICUs outside of the 99% prediction limits, of which there were two, were deemed to have unusually poor performance. Upon further inspection of the two seemingly unusual ICUs, ANZICS CORE determined that one of these ICUs was unusual because of casemix, and the other is due to data quality issues. Given the possible implications of misidentification for the ICUs and the public, we recommend the rigorous statistical approach that we have used, including the development of a comprehensively calibrated Stage 1 model, for future ICU comparison studies.

The aim of the risk-adjustment model is to account for differences between observed deaths at different ICUs because of casemix and factors beyond the control of the ICU. To further adjust for risk, including random coefficients for interaction terms including APACHE III or for covariates such as age may be desirable; however, such additions to the model would come at a significant computational cost. Estimation of the Stage 2 model with random intercept and random coefficient for APACHE III took approximately 28 h, and estimation of the Stage 1 model took approximately 8 h on a 2×2.66 Dual-core Intel Xeon processor (Intel Corporation, Santa Clara, CA, USA) with 4 GB of RAM. Using the general formula for computation time given in the helpmelogit chapter in the Stata 12 manual, [23], were an additional random coefficient added to the model, estimation of the Stage 1 model would take about 7 times as long, estimation of the Stage 2 model would take about 8.5 times as long, and would possibly take longer in practice. Furthermore, including additional random coefficients could exacerbate concerns about inadvertently absorbing ICU differences through over adjusting for risk.

It is likely that different hospitals have different policies regarding end-of-life care, and as such, it may be desirable to use 30-day mortality or patient survival, rather than inhospital mortality, as the outcome for analysis. However, because of the present lack of linkage between databases in many Australian jurisdictions, information about patients following their discharge from hospital is unfortunately unavailable. Moreover, performing survival analysis on patients censored at hospital discharge is problematic and could lead to a biased analysis because of differences between the discharge policies of different hospitals. It is also the case that using a mixed effects survival analysis model would be prohibitively computationally intensive for a large dataset such as ours.

Comparison of the performance of ICUs over periods of more than 2 years is of interest and currently under study by the authors in the ANZICS context. To enable such comparisons, we require further extensions to the methods presented here. In particular, models for data for two or more years incorporating autocorrelation and seasonal effects are of interest and being developed at the risk-adjustment stage of the modelling process.

Appendix A. Models

The covariates included in the model (in the case of categorical covariates, reference level given in parentheses) are as follows:

- Mean-centred patient age, and age squared;
- Mean-centred APACHE III severity score, and APACHE III squared;
- Gender (male);
- Mechanical ventilation within the first 24 hours of admission to the ICU: ventilated, not ventilated (not ventilated);
- Descriptors of ICU admission primary organ system dysfunction (patient category): cardiovascular, gastrointestinal, metabolic, neurologic, respiratory, trauma, renal/genitourinary, hematologic (cardiovascular);
- Patient surgical status: postelective surgery, postemergency surgery, nonsurgical (nonsurgical);
- ICU source: no transfer, hospital transfer, (no transfer);
- ICU level: rural, metropolitan, tertiary, private (tertiary);
- States and territories of Australia or New Zealand (New South Wales);
- Mean-centred annual volume of the ICU (treated as a continuous covariate).
Two-way interactions are as follows:
- APACHE III by each of age, ICU source, patient category, surgical status, ventilation status;
- Ventilation status by each of age, sex, patient category, surgical status;
- Surgical status by each of patient category, ICU source;
- ICU level by each of surgical status and ICU source.

Insignificant interactions are the follows: ICU source by each of APACHE III score and surgical status, and ventilation status by each of age, sex and surgical status.

**Appendix B. Variance of the logarithms of standardised mortality ratio**

We assume a two-level model, with patients clustered within ICUs:

\[ Y_{ij} = \begin{cases} 
1 & \text{if patient } i \text{ in ICU } j \text{ dies} \\
0 & \text{otherwise}, 
\end{cases} \]

\[ Y_{ij} | X_{ij}, Z_{ij}, U_j \sim \text{Bernoulli}(P_{ij}), \quad P_{ij} = \frac{\exp\left(\gamma X_{ij} + U_j Z_{ij}\right)}{1 + \exp\left(\gamma X_{ij} + U_j Z_{ij}\right)}, \quad U_j \sim N(p_{1+i}(0, \Sigma)). \]

In this model, the vector \( X_{ij} \) contains the fixed-effect data and the vector \( Z_{ij} \) contains the data for the level 2 (ICU level) random effects for patient \( i \) in ICU \( j \). The SMR for ICU \( j \) is estimated by

\[ \hat{S}\hat{M}\hat{R}_j = \frac{O_j}{\bar{E}_j} = \frac{\sum_{i=1}^{n_j} Y_{ij}}{\sum_{i=1}^{n_j} \hat{P}_{ij}}, \quad \hat{P}_{ij} = \frac{\exp\left(\hat{\gamma}_j X_{ij} + \hat{U}_j Z_{ij}\right)}{1 + \exp\left(\hat{\gamma}_j X_{ij} + \hat{U}_j Z_{ij}\right)}, \]

where \( \hat{U}_j \) is the empirical Bayesian modal prediction of \( U_j \). We want a confidence interval for \( \hat{S}\hat{M}\hat{R}_j \), but [18] indicates that confidence intervals with better coverage properties are based on \( \log\left(\hat{S}\hat{M}\hat{R}_j\right) \).

We can then write

\[ \hat{\text{var}}\left[\log\left(\hat{S}\hat{M}\hat{R}_j\right)\right] = \hat{\text{var}}\left[\log\left(\frac{O_j}{\bar{E}_j}\right)\right] = \hat{\text{var}}\left[\log(O_j) - \log(\bar{E}_j)\right] = \hat{\text{var}}[\log(O_j)] + \hat{\text{var}}[\log(\bar{E}_j)] - 2 \hat{\text{cov}}[\log(O_j), \log(\bar{E}_j)]. \]

If each ICU has a small number of patients relative to the total number of patients, the covariance term can be assumed to be zero, and following [18], delta-method arguments give

\[ \hat{\text{var}}\left[\log\left(\hat{S}\hat{M}\hat{R}_j\right)\right] \approx \frac{\hat{\text{var}}(O_j)}{O_j^2} + \frac{\hat{\text{var}}(\bar{E}_j)}{\bar{E}_j^2}. \]

Expressions for \( \text{var}(O_j) \) and \( \text{var}(\bar{E}_j) \) are then required:

\[ \text{var}(O_j) = \text{var}\left(\sum_{i=1}^{n_j} Y_{ij}\right) = \sum_{i=1}^{n_j} \text{var}(Y_{ij}) + 2 \sum_{i<k} \text{cov}(Y_{ij}, Y_{kj}). \]

To find estimators for these variances and covariances, we use Goldstein’s linearisation, [40, p128]:

\[ Y_{ij} = (\gamma X_{ij} + U_j Z_{ij}) \frac{\exp(\gamma X_{ij})}{1 + \exp(\gamma X_{ij})^2} + e_{ij} \sqrt{\hat{P}_{ij}} \frac{1}{1 - \hat{P}_{ij}}, \]

\[ \hat{P}_{ij} = \frac{\exp(\gamma X_{ij})}{1 + \exp(\gamma X_{ij})}; \quad \text{var}(e_{ij}) = 1. \]

An expression for the variance of the outcome for patient \( i \) in ICU \( j \) is then

\[ \text{var}(Y_{ij}) = \hat{P}_{ij}^2 (1 - \hat{P}_{ij})^2 \text{var}(\gamma X_{ij} + U_j Z_{ij}) + \hat{P}_{ij} (1 - \hat{P}_{ij}) \]

\[ = \hat{P}_{ij}^2 (1 - \hat{P}_{ij})^2 Z_{ij}^T \text{var}(U) Z_{ij} + \hat{P}_{ij} (1 - \hat{P}_{ij}) \]

which can be estimated by

\[ \hat{\text{var}}(Y_{ij}) = \left(\hat{P}_{ij} - \hat{P}_{ij}^2\right) Z'_{ij} \hat{\Sigma} Z_{ij} + \hat{P}_{ij} (1 - \hat{P}_{ij}). \]

\[ \hat{P}_{ij} = \frac{\exp(\hat{\gamma} X_{ij})}{1 + \exp(\hat{\gamma} X_{ij})}. \]

Again, using Goldstein’s linearisation to obtain estimates of the covariance of the outcomes of two patients in ICU \( j \):

\[ \text{cov}(Y_{ij}, Y_{kj}) = E \left[ (Y_{ij} - E[Y_{ij}]) (Y_{kj} - E[Y_{kj}]) \right] \]

\[ = E \left[ \left\{ Z'_{ij} U_j \frac{\exp(X'_{ij} \gamma)}{1 + \exp(X'_{ij} \gamma)} + e_{ij} \sqrt{ \hat{P}_{ij} (1 - \hat{P}_{ij}) } \right\} \right. \]

\[ \times \left. \left\{ Z'_{kj} U_j \frac{\exp(X'_{kj} \gamma)}{1 + \exp(X'_{kj} \gamma)} + e_{kj} \sqrt{ \hat{P}_{kj} (1 - \hat{P}_{kj}) } \right\} \right] \]

\[ = E \left[ \frac{\hat{P}_{ij}}{1 + \exp(X'_{ij} \gamma)} \frac{\hat{P}_{kj}}{1 + \exp(X'_{kj} \gamma)} Z_{ij} U_j Z'_{kj} U_j \right] \]

\[ = \hat{P}_{ij} (1 - \hat{P}_{ij}) \hat{P}_{kj} (1 - \hat{P}_{kj}) Z_{ij} E \left[ U_j U'_j \right] Z_{kj}. \]

which can be estimated by

\[ \text{cov}(Y_{ij}, Y_{kj}) = \hat{P}_{ij} (1 - \hat{P}_{ij}) \hat{P}_{kj} (1 - \hat{P}_{kj}) Z_{ij} \hat{\Sigma} Z_{kj}. \]

An expression for \( \text{var} \left( \hat{E}_j \right) \) is now required:

\[ \text{var} \left( \hat{E}_j \right) = \text{var} \left( \sum_{i=1}^{n_j} \hat{P}_{ij} \right) = \sum_{i=1}^{n_j} \text{var} \left( \hat{P}_{ij} \right) + 2 \sum_{i < k} \text{cov} \left( \hat{P}_{ij}, \hat{P}_{kj} \right). \]

To simplify the expression, let \( X^*_{ij} \hat{\gamma}^*_j = (X_{ij}, Z_{ij}) \left( \hat{\gamma}, \hat{U}_j \right) \) and \( \hat{P}_j = \left( \hat{P}_{ij}, \ldots, \hat{P}_{n_j} \right)^T \). Then, \( \text{var} \left( \hat{E}_j \right) \) is the sum of the elements of the covariance matrix of \( \hat{P}_j \). Once again, following [18] and using a delta-method argument,

\[ \text{var} \left( \hat{P}_j \right) = \left( \frac{\partial P_j}{\partial \gamma} \right)' \text{var} \left( \hat{\gamma}^* \right) \left( \frac{\partial P_j}{\partial \gamma} \right)_k = x^*_k \hat{P}_j (1 - \hat{P}_j), \]

where \( x^*_k \) is the \( k \)th element of \( X^*_{ij} \) and \( (\hat{\gamma}^*_j)_k \) is the \( k \)th element of \( \hat{\gamma}^*_j \). The variance of the log-SMR of ICU \( j \) can then be written as

\[ \text{var} \left[ \log \left( S \hat{M} R_j \right) \right] = \frac{1}{O_j^2} \left[ \sum_{i=1}^{n_j} \sum_{k=1}^{n_j} \hat{P}_{ij} (1 - \hat{P}_{ij}) \hat{P}_{kj} (1 - \hat{P}_{kj}) Z_{ij} \hat{\Sigma} Z_{kj} + \sum_{i=1}^{n_j} \hat{P}_{ij} (1 - \hat{P}_{ij}) \right] \]

\[ + \frac{1}{E_j^2} \sum_{i=1}^{n_j} \sum_{k=1}^{n_j} \hat{P}_{ij} (1 - \hat{P}_{ij}) \hat{P}_{kj} (1 - \hat{P}_{kj}) (X^*_{ij})' \text{var} \left( \hat{\gamma}^* \right) X^*_k. \]

Appendix C. Covariance of the 2009 logarithms of standardised mortality ratio and the 2010 logarithms of standardised mortality ratio

If \( O_{ij} \) is the number of deaths observed at ICU \( j \) in 2009, \( E_{ij} \) the expected number of deaths in ICU \( j \) in 2009, and \( n_{ij} \) the number of patients in ICU \( j \) in 2009, with \( O_{2j}, E_{2j} \) and \( n_{2j} \) are defined similarly for 2010; then, let

\[ S_{kj} = \log(O_{kj}) - \log \left( \hat{E}_k \right). \]
The covariance of $S_{1j}$ and $S_{2j}$ is given by

$$\text{cov}(S_{1j}, S_{2j}) = \frac{1}{2} \left[ \text{var}(S_{1j}) + \text{var}(S_{2j}) - \text{var}(S_{1j} - S_{2j}) \right]$$

$$= \frac{1}{2} \left[ \text{var} \left( \log(O_{1j}) - \log \left( \hat{E}_{1j} \right) \right) + \frac{1}{2} \text{var} \left( \log(O_{2j}) - \log \left( \hat{E}_{2j} \right) \right) \right]$$

$$- \frac{1}{2} \text{var} \left( \log(O_{1j}) - \log \left( \hat{E}_{1j} \right) - \log(O_{2j}) + \log \left( \hat{E}_{2j} \right) \right).$$

Assuming that $\text{cov} \left[ \log \left( O_{1j} \right), \log \left( \hat{E}_{1j} \right) \right] = 0$ and using delta-method arguments to obtain expressions for the variances of logarithms of random variables,

$$\text{cov}(S_{1j}, S_{2j}) = \frac{1}{2} \left\{ \frac{1}{O_{1j}^2} \text{var}(O_{1j}) + \frac{1}{\hat{E}_{1j}^2} \text{var}(O_{2j}) + \frac{1}{\hat{E}_{2j}^2} \text{var}(\hat{E}_{2j}) \right\}$$

$$- \frac{1}{2} \left\{ \frac{1}{O_{1j}^2} \text{var}(O_{1j}) + \frac{1}{O_{2j}^2} \text{var}(O_{2j}) + \frac{1}{\hat{E}_{1j}^2} \text{var}(\hat{E}_{1j}) + \frac{1}{\hat{E}_{2j}^2} \text{var}(\hat{E}_{2j}) \right\}$$

$$+ \frac{1}{O_{1j} O_{2j}} \text{cov}(O_{1j}, O_{2j}) + \frac{1}{\hat{E}_{1j} \hat{E}_{2j}} \text{cov}(\hat{E}_{1j}, \hat{E}_{2j})$$

$$= \frac{1}{O_{1j} O_{2j}} \text{cov}(O_{1j}, O_{2j}) + \frac{1}{\hat{E}_{1j} \hat{E}_{2j}} \text{cov}(\hat{E}_{1j}, \hat{E}_{2j}).$$

Letting $O_{kj} = \sum_{i=1}^{n_{kj}} Y_{ikj}$ and $\hat{E}_{kj} = \sum_{i=1}^{n_{kj}} \hat{P}_{ikj}$, an estimator for the covariance is obtained:

$$\hat{\sigma}_{12} = c \hat{\text{cov}}(S_{1j}, S_{2j}) = \frac{1}{\sum_{i=1}^{n_{1j}} Y_{1ij} \sum_{k=1}^{n_{2j}} Y_{k2j}} \sum_{i=1}^{n_{1j}} \sum_{k=1}^{n_{2j}} \text{cov}(Y_{1ij}, Y_{k2j}) + \sum_{i=1}^{n_{1j}} \sum_{k=1}^{n_{2j}} \text{cov}(\hat{P}_{1ij}, \hat{P}_{k2j})$$

$$= \frac{1}{\sum_{i=1}^{n_{1j}} Y_{1ij} \sum_{k=1}^{n_{2j}} Y_{k2j}} \sum_{i=1}^{n_{1j}} \sum_{k=1}^{n_{2j}} \hat{P}_{1ij} \left( 1 - \hat{P}_{1ij} \right) \hat{P}_{k2j} \left( 1 - \hat{P}_{k2j} \right) \left( X_{1ij}^* \right) \text{var} \left( \hat{y}^* \right) \left( X_{k2j}^* \right).$$

where $X_{1ij}^*$ is the matrix containing both the fixed and random effect data for patient $i$ in ICU $j$ in 2009, and $X_{2j}^*$ is the matrix containing both the fixed and random effect data for patient $i$ in ICU $j$ in 2010.

Acknowledgements

We gratefully acknowledge two anonymous reviewers for their helpful comments. This research was supported by Australian Research Council Discovery Project DP110102028.

References


