Point of view

Mortality and other event rates: what do they tell us about performance?

The outcomes paradigm is now a dominant influence within medicine, as witness the controversy regarding the public enquiry into the performance of the paediatric cardiac surgical service at the Royal Bristol Infirmary. Any reservations about the sway of this new paradigm (one commentator suggested that decisions to publish “league tables” of death rates are on political, not scientific grounds), should be tempered by the remembrance that the “third revolution” in medical care has a long history, dating back to Florence Nightingale in the mid 19th century in the United Kingdom and Ernest Codman in the early 1900s in the USA. Critical care is no exception to this movement, as witness the recent publication of state-of-the-art reviews. However, in our rush to embrace this new wisdom we should be well advised to integrate firmly into our vision the notion of uncertainty or variability. The purpose at hand will to be review the nexus between outcome and quality as it applies to critical care, but, as no process occurs in isolation, illustrative reference will be frequently made to other medical disciplines.

The 1986 paper by Knaus et al., exploring the APACHE II risk-adjusted mortality of a cohort of 13 intensive care units (ICU), would appear to have established the notion of “institutional” or “provider” comparison within critical care. The standardised mortality ratio (i.e. observed to expected numbers of deaths, O/E = SMR) was introduced to the critical care literature and a discordant debate has subsequently ensued regarding the relationship between the SMR and ICU performance or quality. The analytic strategy to establish differences between ICUs in the Knaus et al. paper was two fold: 1) a comparison of observed and predicted mortality rates for each hospital, via the t-test, and 2) the inclusion of “institutional effects” as a categorical covariate within the logistic regression analysis. Both methods identified the same two “outlying” ICUs with high (1.58) and low (0.59) SMRs. Knaus et al., referred in their methodology section to a previous 1983 paper by Wolfe and co-workers, who had used the SMR to compare mortality rates of 11 specialist burn units and had also analysed “institutional effects” (as log odds ratio, log OR) within a logistic regression equation. In this study, a good correlation was reported between the SMR and the log OR, \( r = 0.95 \), although the concordance (\( \rho_c \)) between the two measures was quite low (\( \rho_c = 0.19 \)).

As pointed out by Hosmer and Lemeshow, the SMR and its variability is problematic and statistical tests of the equality of observed to expected mortality have been based upon the difference (\( O - E \)). Hosmer and Lemeshow described various estimates of the confidence intervals (CI) of the SMR:

1. If expected mortality probabilities are considered fixed (for instance, in the calculation of patient mortality probabilities only published logistic regression coefficients from the literature are available), CI can be computed (for what is in effect an indirectly standardised mortality ratio)\(^{16}\) using:
   - a) the normal approximation: \( \left[ \text{observed deaths} \pm z_{1-\alpha} \sigma/\text{(expected deaths)} \right] \)
   - b) bootstrapping, in particular the \( B_\alpha \) (bootstrap percentile) and \( BC_\alpha \) (bias-corrected and accelerated) intervals.

\( z_{1-\alpha} \) is the \((1-\alpha) \times 100^{th}\) percentile of the standard normal distribution.

ii) if estimated probabilities from a logistic regression equation are considered random (that is, published coefficients from the literature and an estimate of the covariance matrix are available), then CI can be computed using the so-called delta method\(^{20}\) (Taylor series approximations of log \( O/E \)).\(^{21}\)

Although the four methods (normal approximation, bootstrap \( B_\alpha \) and \( BC_\alpha \) and delta method) were not equivalent with respect to CI width \( (p = 0.002) \), the practicality of calculation (routine lack of publication of covariance matrices) recommended the methods of (i) above\(^{15}\) and the majority of reports in the critical care literature have used the normal approximation. However, the authors suggested that “a detailed simulation study is needed before we can recommend a definitive choice between the methods”. At this juncture a formal debate about these matters has not proceeded within the critical care literature, although in another domain (ie. health economics), a lively exchange has recently appeared about appropriate CI for cost-effectiveness ratios.\(^{22-25}\)

In summary, in the standard assessment of the risk-adjusted mortality outcomes of a cohort of ICUs, sampling variability in the observed outcomes \( (O) \) is acknowledged, but the expected outcomes or probabilities \( (E) \) are considered as fixed; that is, the
calculation of the SMR ignores the fact that the estimated coefficients ($\hat{\beta}$) derived from a logistic regression are a random vector. More importantly, unequal numbers of patients from different providers may lead to substantial bias in the normal based CI, and Taylor series approximations are to be preferred.27

In the comparison of means and 95% CI the overlap of such intervals does not necessarily indicate a lack of significant difference (at $\alpha = 0.05$).28 If the null hypothesis is that the mean is equal to a fixed scalar quantity (for example, an SMR = 1), then the probability that the sample mean lies in the upper or lower 2.5th percentile is, in fact, 5%. Such is not the case when comparing two means (say, $Q_1$ and $Q_2$); in this scenario, the overlap method is conservative and lacks power (with respect to a test of the null hypothesis that $Q_1 - Q_2 = 0$).29 a deficiency that is increased when (i) the corresponding standard errors (SE$_1$ and SE$_2$) are nearly equal and (ii) $Q_1$ and $Q_2$ are positively correlated. Under a normality assumption, the appropriate width of intervals to achieve 5% significance level (with equal SEs) is $\pm 1.39\sigma$ (or 83% CI).28,30 The importance of the above relates to the interpretation of the graphic of a “league table” (say, the SMRs and CIs of an ICU cohort), whereby the error bars need to be prescribed at a particular significance level ($\beta$), such that the non-overlap significance level, averaged over all possible comparison pairs, is equal to a prescribed $p$ value, usually 0.05 or 0.01. (Technically, $Z_{\beta} \neq 1.96$ and is calculated (using a search procedure) as the average of $z_{\sigma}/(\sigma_{\sigma} + \sigma_{\beta})$ over all $(i, j)$ pairs; a CI for the $i^{th}$ category is given as $m_i \pm z_{\alpha}\sigma_{\alpha}$.)30,31

One of the definitive expositions of the standardised difference, $(O - E)$, as it applied to critically ill (burn) patients, was provided by Flora in 1978,32 albeit this was for patient survival compared with calculated “baseline” survival curves.33,34

$$Z = \frac{S - \sum P_i}{\sqrt{\sum P_i Q_i}}$$

where $Z$ is the familiar standard normal deviate, $S$ is the total number of survivors among the $n$ patients, $P_i$ the probability of survival, estimated from the baseline survival curve and $Q_i = 1 - P_i$, the probability of death. Lemeshow et al.35 also endorsed such a “simple” calculation for the assessment of a “. . . particular hospital’s mortality experience ..(compared with) . . . that expected with the MPM system . . . .”. Power determinations (to detect a difference) can also be formulated; in this case: $Power = \Phi(-K\sqrt{\frac{E(P_i Q_i) + \sum(P - P_i)}{\sum P_i Q_i}}) + 1 - \Phi(-K\sqrt{\frac{E(P_i Q_i) + \sum(P - P_i)}{\sum P_i Q_i}})$

where $\Phi$ is the cumulative normal distribution function and $K_\alpha$ are the familiar critical values of the standard normal distribution (significance level ($\alpha$); $0.1 = \pm 1.65$, $0.05 = \pm 1.96$, $0.01 = \pm 2.58$ and $0.001 = \pm 3.29$).

Although seemingly formidable in its complexity, the equation reveals that the power of the Z statistic to detect a difference $(O - E)$ is dependent upon: (i) the prescribed $\alpha$ error rate and will also increase with an increase in the difference between the study population and baseline survival rates, (ii) a decrease in survival probabilities, (iii) size of study and (v) increase in heterogeneity of survival probabilities.36,37 Studies in the critical care literature have looked at the variability of the SMR with patient demographics,38-40 but the statistical “performance” of the SMR ratio has not been subjected to similar scrutiny as in the trauma and burn literature.

The above exposition has suggested a degree of uncertainty or variation in the estimation of performance profile of an ICU when directly assessed by risk-adjusted mortality and the corresponding SMR. The analytic strategy which informs risk-adjustment of binary health outcomes (for example: mortality, complications, wound infections) is of some interest to practitioners,41,42 and a frequent complaint of providers is that predictive algorithms do not sufficiently adjust for differences in case mix,43,44 which, by formal definition, excludes provider characteristics and processes. Case mix adjustment may occur by a number of means:45-47

i) direct standardisation; whereby a simple summary of, say complication rates, is reported for a provider unaffected by the mix of surgical and medical patients. An adjusted rate would be reported for an average fraction of surgical and medical patients across providers (assuming that cases are found in each stratum). Rixom has also argued for direct standardisation of performance (league) tables when considering United Kingdom NHS trust performance indicators.48

ii) indirect standardisation; whereby an expected incidence for providers is calculated using rates by stratum from a standard population and then comparing observed vs expected outcome rates. Fidler49 has suggested that the method of calculation of SMR using say, the APACHE II algorithm, whereby “mortality ratios are calculated by projecting the APACHE II score-specific mortalities of the total group on case mix . . . of individual ICUs” amounts to an indirect standardisation, which (quoting Yule and Rothman), “is not fully a method of standardisation at all”. Fidler’s recommendation flowing from this was to use direct standardisation by either:
a) “logistic regression … with separate intercepts for each ICU. The intercepts are simply the logits of (directly standardised) mortality rates and can be used for rankings. This approach assumes constant slopes for all ICUs… and can be tested”, or

b) model the differences between ICUs as random effects

c) regression modelling, using linear logistic regression, although probit analysis was used by early pioneers in the field. The common critical care outcome prediction algorithms eschew “institutional” factors, although, as seen above, these may be utilised in “snapshot” analyses.

d) more recently, multilevel or hierarchical modelling with a full or empirical Bayesian analytic framework has been introduced. Formal (model based) account of uncertainty is undertaken and institutional differences are modelled explicitly using random effects; in particular, the ICU effect is derived from a specific (usually normal) distribution. There may be particular advantage in the Bayesian approach, by virtue of being able to pose and answer questions such as the probability that the mortality rate of a provider will exceed a certain threshold percentage (for instance, estimation of the posterior probability that mortality in a hospital is a certain fraction of the median mortality of all hospitals); that is the ability to quantify uncertainty. Compared with a “fixed-effects” frequentist method (for instance, each provider considered as a separate level in a nominal categorical covariate), Bayesian mortality (point) estimates are “shrunken” towards the average mortality rates of providers. In the middle and upper ranges of patient risk outlier status becomes constrained and any propensity for outlier status is thus reduced, but at lower levels of risk, where event rates are sparse, Bayesian methods have high sensitivity. When applied to the same data set, agreement between the frequentist and Bayesian approaches in identifying outliers was noted to be poor. The Bayesian framework was the statistical basis for the recent public enquiry into the performance of the paediatric cardiac surgical unit at the Bristol Royal Infirmary.

Assuming that the risk-adjustment proceeds from an algorithm based upon a separate sample population (historically and or geographically), the question of the “generalisability” of the algorithm must be addressed. The methodology of algorithm construction and validation has been reviewed extensively in the literature, and it is not the purpose here to re-visit this paradigm; rather, the following are noted:

i) predictive algorithms perform less well in validation studies, and/or across populations, the usual pattern being that of good discrimination with variable calibration. With respect to the APACHE III algorithm, the hospital length of stay adjustment appears problematic in geographical comparisons as evidenced by the study of Sirio et al, (critical care services contrasted between Japan and the USA) and acknowledged by members of the APACHE group: “because this adjustment is unique to US hospitals it has not been possible to adjust for the impact of hospital length of stay on mortality in international studies using APACHE III”.

ii) models have a tendency to over-predict mortality likelihood in patients with less severity of illness and under-predict mortality likelihood with high severity of illness.

iii) the relative importance of discrimination and calibration depends upon the intended application, for example, when comparing observed and expected mortality proportions, calibration is paramount; for classification of patients by severity, discrimination. This being said, if the discrimination of the model deteriorates, “re-calibration” will not be corrective; with good discrimination, “re-calibration” will reconstitute reliability of the predictor instrument.

iv) methods of assessment of discrimination and calibration are well defined, as are comparison of receiver operator characteristic (ROC) curve areas, although Lemeshow and Le Gall were critical of graphical evaluation of observed versus expected mortality (calibration) as “very ineffective for the purpose…since, even for poor models, the proportion of observed deaths in each successive probability group will tend to go up”. Despite its ubiquitous application, the familiar Hosmer-Lemeshow deciles-of-risk “C” statistic (HL C) is a “small sample statistic” and will invariably yield p values << 0.1 (HL C >> 13.6) with large data bases, a point noted by some investigators. Recent studies have also suggested that it is not necessarily the optimal test for logistic regression goodness-of-fit, albeit the “.2 by 10 table of observed and estimated expected frequencies…provides a useful overall summary of model fit…”

v) multiple predictive algorithms give rise to competing performance claims. Numerous studies have suggested differential algorithm performance and consequent provider ranking changes; in: myocardial infarction, general medical conditions (stroke, lung cancer, pneumonia, myocardial infarction, congestive heart failure and coronary artery surgery) and critical care. Less divergence
was noted with trauma algorithms and in coronary artery surgery when considered alone. The particular concerns above (i) through (v)) may be illustrated using a data set (1993-1997) of the ANZICS National Data Base, previously reported. A total of 73171 patients were available for comparison of APACHE II and SAPS II risk of death for hospital mortality outcome. ROC curve areas were 0.866 (SE: 0.002) for APACHE II and 0.868 (SE: 0.002) for SAPS II and, despite the large sample size, not significantly different (p = 0.065). The Hosmer-Lemeshow deciles-of-risk “C” statistic was 253.81 (APACHE II) and 186.87 for SAPS II, suggesting that overall, the SAPS II algorithm was preferred (“lower” Hosmer-Lemeshow statistic). However, different patterns of mortality prediction were evident across the deciles as seen in Table 1: APACHE II tended to over-predict deaths in the lower deciles whereas SAPS II over-predicted only in the last two upper deciles. Differential performance of the APACHE II algorithm also occurred across hospital classifications, as seen in Table 2.

### Table 1. Deciles of risk comparison for APACHE II and SAPS II

<table>
<thead>
<tr>
<th>Deciles</th>
<th>Observed deaths</th>
<th>Expected AP II</th>
<th>Observed deaths*</th>
<th>Expected SAPS II</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>107</td>
<td>61</td>
<td>104</td>
<td>63</td>
</tr>
<tr>
<td>2</td>
<td>175</td>
<td>172</td>
<td>153</td>
<td>139</td>
</tr>
<tr>
<td>3</td>
<td>266</td>
<td>296</td>
<td>226</td>
<td>181</td>
</tr>
<tr>
<td>4</td>
<td>365</td>
<td>448</td>
<td>447</td>
<td>396</td>
</tr>
<tr>
<td>5</td>
<td>474</td>
<td>646</td>
<td>410</td>
<td>399</td>
</tr>
<tr>
<td>6</td>
<td>713</td>
<td>949</td>
<td>871</td>
<td>796</td>
</tr>
<tr>
<td>7</td>
<td>1141</td>
<td>1368</td>
<td>1088</td>
<td>973</td>
</tr>
<tr>
<td>8</td>
<td>1991</td>
<td>2074</td>
<td>2073</td>
<td>1876</td>
</tr>
<tr>
<td>9</td>
<td>3302</td>
<td>3407</td>
<td>3198</td>
<td>3238</td>
</tr>
<tr>
<td>10</td>
<td>5508</td>
<td>5682</td>
<td>5472</td>
<td>5778</td>
</tr>
</tbody>
</table>

Total 14042 15103 14042 13839

Observed deaths = observed deaths in each decile of APACHE II mortality probability, Observed deaths* = observed deaths in each decile of SAPS II mortality probability, Expected AP II = expected deaths with APACHE II risk of death, Expected SAPS II = expected deaths with SAPS II risk of death

### Table 2. Comparison of APACHE II performance across hospital classifications

<table>
<thead>
<tr>
<th>Deciles</th>
<th>Observed</th>
<th>Expected</th>
<th>Observed</th>
<th>Expected</th>
<th>Observed</th>
<th>Expected</th>
<th>Observed</th>
<th>Expected</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>I</td>
<td>II</td>
<td>I</td>
<td>II</td>
<td>III</td>
<td>III</td>
<td>IV</td>
<td>IV</td>
</tr>
<tr>
<td>1</td>
<td>7</td>
<td>9</td>
<td>21</td>
<td>13</td>
<td>105</td>
<td>50</td>
<td>8</td>
<td>12</td>
</tr>
<tr>
<td>2</td>
<td>12</td>
<td>33</td>
<td>50</td>
<td>45</td>
<td>161</td>
<td>129</td>
<td>15</td>
<td>27</td>
</tr>
<tr>
<td>3</td>
<td>28</td>
<td>52</td>
<td>69</td>
<td>77</td>
<td>217</td>
<td>207</td>
<td>18</td>
<td>35</td>
</tr>
<tr>
<td>4</td>
<td>53</td>
<td>87</td>
<td>99</td>
<td>118</td>
<td>321</td>
<td>319</td>
<td>21</td>
<td>50</td>
</tr>
<tr>
<td>5</td>
<td>84</td>
<td>140</td>
<td>125</td>
<td>183</td>
<td>415</td>
<td>473</td>
<td>28</td>
<td>69</td>
</tr>
<tr>
<td>6</td>
<td>93</td>
<td>178</td>
<td>145</td>
<td>228</td>
<td>706</td>
<td>706</td>
<td>40</td>
<td>94</td>
</tr>
<tr>
<td>7</td>
<td>148</td>
<td>257</td>
<td>261</td>
<td>345</td>
<td>1081</td>
<td>1069</td>
<td>56</td>
<td>127</td>
</tr>
<tr>
<td>8</td>
<td>265</td>
<td>378</td>
<td>410</td>
<td>512</td>
<td>1625</td>
<td>1636</td>
<td>111</td>
<td>181</td>
</tr>
<tr>
<td>9</td>
<td>572</td>
<td>626</td>
<td>715</td>
<td>814</td>
<td>2433</td>
<td>2531</td>
<td>167</td>
<td>267</td>
</tr>
<tr>
<td>10</td>
<td>1181</td>
<td>1209</td>
<td>1255</td>
<td>1376</td>
<td>3715</td>
<td>3861</td>
<td>510</td>
<td>622</td>
</tr>
</tbody>
</table>

H-L 216.53 193.75 116.23 265.02
ROCE 0.886(0.004) 0.855(0.004) 0.852(0.002) 0.856(0.007)
Total n 17465 17430 48174 11094
AP2 mean(SD) 12(9) 14(9) 15(9) 11(7)

Observed = observed deaths, Expected = expected deaths, I = regional hospitals, II = metropolitan hospitals, III = tertiary referral hospitals, IV = private hospitals, H-L = Hosmer-Lemeshow “C” statistic, ROC(SE) = ROC curve area and SE, Total n = total number of patients within hospital classification, AP2 = mean and SD of APACHE II score in hospital classification. Acceptable calibration: H-L < 13.6.
A plot of the two mortality probabilities (Figure 1) is similar to the “snow-storm” effect reported by Lemeshow et al when comparing APACHE II and MPM II.  

vi) the question of when to re-calibrate or customise “will be back – again and again.” In the trauma, cardiac surgical and critical care literature, recalibration has been undertaken with varying results. Metnitz et al. found wide changes in ΔO/E ratios (Δ = difference, before and after customisation) of -3.6 to +25% in 13 ICUs and Ivanov et al. demonstrated a change in half the rankings of surgeon risk-adjusted operative mortalities with re-calibration, although the accompanying editorial suggested that these changes were “relatively unimpressive.” Similarly, Champion et al. from the trauma perspective complained about being “sentenced to perpetual tweaking by certain researchers....” In one of the better know applications of risk-adjusted quality control (the Greater Cleveland Quality Choice study) initial recalibration was undertaken before study initiation; as Teres and Lemeshow observed: “Once there is a good, up-to-date local model, ...severity systems can be used for quality of care comparisons, assuming high quality data collection... clarification of definitions and good data management.” This sentiment was also echoed by Glance and Szalodos in an editorial comment upon the above referenced Sirio et al. cross-cultural comparison of critical care delivery: “Regardless of the model adopted, it is critical that the model coefficients be periodically updated so that institutions can benchmark against a contemporary reference point.” Similarly, Rowan et al. considered the question as to “whether an APACHE II equation derived from British data would provide better case mix adjustment than the existing American equation.” 

vii) implications of the above have been formally explored by DeLong et al. using an adult cardiac surgical database (total n = 3654, for 28 providers). Models were assessed using receiver operator area (ROC) area and deviance ($-2 \log$-likelihood; lower
for better fitting models). The methods of adjustment were:

1. external standard; the SMR being calculated as \( O / E \) using a previously derived mortality model.\(^{99}\) Mortality probabilities were considered as fixed.

2. mortality prevalence correction, whereby a constant correction term is added to each patients risk score, where the risk score \( = \) the logit (back transformed from mortality probabilities). The correction term was:

\[
\log \left( \frac{p_1}{1 - p_1} \right) - \log \left( \frac{p_0}{1 - p_0} \right)
\]

where \( p_1 \) and \( p_0 \) are mortality prevalences in the evaluation and original benchmark populations respectively.

3. modelling the risk score; the risk scores are calibrated by incorporating them as a covariate into a secondary logistic model

4. providers are modelled as fixed effects; the \( \beta^p \) provider is compared with the overall weighted average (using a deviation from means contrast matrix).\(^{100}\)

5. providers are modelled as random effects; the observed provider effects are drawn from a specific distribution (normal) and estimates of the individual effects are computed as “posterior means” (empirical Bayes predictions) which are shrunken to the null (that is, towards an SMR of 1) compared with the “fixed effects”\(^{101,102}\).

6. internal standard, equivalent to 1, above but based upon the algorithm adjusted to the evaluation population itself

7. providers as fixed effects base upon the internal standard algorithm

8. providers as random effects base upon an internal standard algorithm

The performance of the models improved, not surprisingly, with incorporation of provider effects and use of the internal standard. Risk score coefficients showed little change across models 2-5, suggesting that the relationship between the external risk score and in-hospital mortality was not confounded by provider effects. Provider outlier status (high or low) varied between the models and was reduced overall (in terms of the absolute count of ICUs) with the internal standard. As methods 2-8 adjust the external risk algorithm (method 1) to the evaluation population, the providers are effectively compared with one another, rather than the external standard. The reasoned conclusion of the authors was that method 5 “provided a realistic assessment of provider performance when the external model fits the data well”. DeLong \textit{et al}, presumably fitted a random intercept model (they used the GLIMMIX SAS macro), but this may not be optimal as it does not capture the full variability of the data. A more comprehensive approach would be to allow the logistic model slope to vary across hospitals (random coefficient model\(^{103}\)).

From the perspective of the current review, other important studies have used statistical modelling to investigate the relationship between provider mortality rates, outlier status and the quality of care in the (programmed) absence of case-mix variation.\(^{103-105}\) The predictive value of mortality rates to accurately classify outliers was poor: (a) sensitivity for detecting poor quality hospitals (“high” outliers) was 35% and positive predictive value was 52%\(^{103}\) and (b) of hospitals that delivered poor quality care, < 12% were identified as high outliers and > 60% of outliers were actually good quality hospitals.\(^{104}\) Glance and co-workers have also used simulation to test the performance of the APACHE II algorithm under a number of scenarios: calibration and discrimination appeared to be functions of case mix\(^{106}\) and the SMR was found to be a linear function of the simulated ICU mortality rate.\(^{107}\) Although the results of these studies have been the subject of a spirited repost from a member of the APACHE group,\(^{108}\) and were derived from a single unit’s data, they serve to underline the uncertainty aspect of mortality rates.

Further evidence for the poor performance of mortality rates as performance indicators was provided by Silber \textit{et al}\(^{109}\) who looked at 73173 admissions in 137 hospitals to measure the relative contribution of patient and hospital characteristics in three outcomes in surgical patients: death, adverse occurrence and death following adverse occurrence. Comprehensive risk-adjustment modelling was undertaken using 53 patient and 12 hospital variables and both hospital-hospital and hospital-patient interactions. The authors concluded that “most of the predictable variation in outcome rates among hospitals appears to be predicted by differing patient characteristics, rather than by differing hospital characteristics. That is, by who is treated rather than the resources available for treatment”.\(^{110}\)

Observation time. Critical care predictive algorithms invariably assess patient morality outcomes at hospital discharge, although such is not the case for the recent Veterans Affairs NSQIP study, which mandates 30 day outcomes.\(^{110}\) A key 1988 paper by Jencks \textit{et al}, demonstrated geographic mortality reversal when comparing inpatient versus 30 day mortality and concluded that “inpatient death rates depend on length-of-stay patterns and give a biased picture of mortality”.\(^{111}\) This study was cited in the original APACHE III publication,\(^{68}\) but in a subsequent review Knaus \textit{et al}, reported no significant change in ICU relative performance rankings using mortality rates “30 days after hospital discharge” compared with in hospital outcomes.\(^{52}\) A
number of other studies have looked at the potential bias of the definition of mortality, with varying results: little or no difference between provider SMRs for in-hospital versus 30-day mortality, although outlier status often differed; no difference for 30 versus 180 days follow up; and modest effects on mortality rates, rankings and outlier status. This being said, recent critical care commentaries have endorsed a move to 90 day mortality outcome.

ix) admission policies may also effect the SMR, by virtue of the propensity to admit or not the sicker patient; an example of selection bias to the extent that hospitalised patients are not randomly selected from the population at risk for hospitalisation. Miller et al, found a negative relationship between hospital SMR and higher relative risks of hospitalisation, such that as the relative risk of hospitalisation increased (and more “less sick” patients are admitted), the SMR decreased.

time change has been recently demonstrated for ICUs in the ANZICS national data base.

x) when considering the rankings or outlier status of providers, multiple comparisons are invariably undertaken, but few authors adjust appropriately for this and such failure has been the subject of a trenchant critique by Localio and co-workers. The actual rankings also demonstrate instability with wide confidence intervals and time variance. From a frequentist perspective, evidence has also been presented for wide variability of the 95% confidence intervals of the ranks for ICUs in the ANZICS national data base. From a frequentist perspective, evidence has also been presented for wide variability of the 95% confidence intervals of the ranks for ICUs in the ANZICS national data base.

xi) time change has been recently demonstrated for hospital mortality rates (consistent decreases over time noted) and indices of care (improvement confirmed); thus the impact of “quality of care” interventions using mortality outcome as a yardstick, needs careful assessment. The much publicised Cleveland Health Quality Choice and NewYork Cardiac Surgery programs yielded improved mortality outcomes over time, but their actual impact is questionable given that clear evidence of similar mortality improvements were occurring in geographically proximate areas where such (public) initiatives were not implemented.

xii) although not the principal focus of this presentation, the principles and processes pertaining to the assessment of morbidity rates, whether risk-adjusted or not, are subject to the same admonitions as above. Furthermore, it is important to realise that the correlations between performance and different outcomes (mortality, complications) and diagnoses appear quite variable.

xiii) the traditional manner of displaying rate indicators, especially over time, has also seen a recent revolution: statistical process control tools have been introduced, risk adjusted CUSUM and sequential probability test charts; cumulative risk-adjusted mortality (CRAM) charts, variable life-adjusted displays, funnel plots and time series monitors. Investigating the utility of these techniques for monitoring rate change over time in a national database is a question of some importance.

The oft quoted Knaus et al, 1986 study (see above) to evaluate outcome in intensive care located differences in mortality rates between ICUs in the “interaction and communication between physicians and nurses”. Although the data collection for the Knaus et al, study had actually occurred between 1979 and 1982 in self selected hospitals and, in some units, non-consecutive patients were recorded, the evidence for organisational determination of outcomes in ICU, even in the current environment of developed (albeit unequally) ICU services, seems persuasive if we are to believe the latest literature survey by Carmel and Rowan. A recent analysis attempted to quantify the physician organisational component of intensive care outcomes in a large retrospective review of outcome from abdominal aortic surgery. Overall hospital mortality rate was 7%, with 7% of the case load as emergency repair; no severity of illness score was available, a point belaboured by the authors. In the multivariate analysis, the only “ICU” characteristic that was predictive of hospital mortality was the categorical variable “No daily rounds by an ICU physician” (odds ratio 3.0, 95% CI, 1.9-4.9). At face value, this effect appears impressive; however, further question may be asked:

a) is this estimate reasonable, as it equates with a mortality reduction (absolute), all other variables being held constant, from 18% to 7%, the latter being the overall series mortality rate. Similar reservations about effect size in observational studies have been raised by Mant and Hicks, in particular, the report of large reductions in breast cancer mortality provided by specialist surgical services. Alternative explanations were offered by the commentators, relating to inadequate case-mix adjustment, better diagnostic facilities at specialist centres, the power of the study (provider sample size is critical) and the overall interpretation of non-randomised trials.

b) if the estimate is in fact believable, it may derive from a local uneven provision of ICU services rather than reflect a more general consequence.
c) with respect to treatment effect and adjustment for potential confounders, two further instances are most illustrative. Firstly, the belief that long term outcomes (5 year mortality) after trans-urethral prostate surgery for benign prostatic hyperplasia were better than open resection was demonstrated to be a function of inadequate accounting for severity of illness.\(^\text{149}\) Second and more pertinent to critical care concerns, the observational study by Connors et al.\(^\text{150}\) on the mortality impact of right heart catheterisation (RHC) used propensity score matching to adjust for pre-treatment observable differences between a group of treated (RHC) and a group of untreated patients. For the 1008 matched pairs, the 30 day mortality was increased with RHC (OR 1.24, 95% CI: 1.03-1.49) and a sensitivity analysis suggested that “a missing covariate would have to increase the risk of death 6-fold and the risk of RHC 6-fold for a true beneficial effect of RHC to be misrepresented as harmful”. However, as Lin et al., demonstrated:\(^\text{151}\) the metric of the sensitivity analysis was OR, not relative hazard and as RHC was a common procedure,\(^\text{152}\) misrepresenting the OR of RHC as a probability ratio resulted in the overestimation of the effects of an unmeasured confounder that would be required to misrepresent a neutral or beneficial RHC effect as harmful. Alternative specifications of the sensitivity analysis, provided by Lin et al., required only a 2-3 fold increase in unmeasured covariate effect (clinically, far more plausible than 6 fold); and overall, there was not “strong evidence” for either a harmful or beneficial effect of RHC.

d) of the 63 publications reviewed by Carmel and Rowan above,\(^\text{143}\) 28 demonstrated a null effect of the intervention or measurement, with extreme sample size variation (25 to 46,587). More importantly, only 3 were randomised and 22 were “before-after” studies, with the potential for confounding of effect by regression to the mean.\(^\text{153}\)

Coincident with the Knaus et al paper,\(^\text{10}\) Dubois and coworkers reported a study entitled “Adjusted hospital death rates: a potential screen for quality of medical care”.\(^\text{154}\) A second paper looked at quality of care components (at the sampled case record level) using both structured explicit and implicit review. Although clinicians’ subjective assessment identified differences between high and low mortality rate outliers, this was not confirmed for any condition where explicit structured process criteria were used. Since this seminal study, there have been other efforts, grounded in chart review, to locate a relationship between mortality and the process of “quality of care”: neither Gibbs et al.,\(^\text{155}\) in a surgical environment nor Best et al.,\(^\text{156}\) Thomas et al.,\(^\text{157}\) nor Park et al.,\(^\text{158}\) in a general medical setting, were able to establish such a relationship, although site-visit assessment of process and structure was able to distinguish differences between several dimensions of care in surgical units.\(^\text{159}\) However, a caution must apply to the accuracy of implicit review processes, especially when outcomes are known.\(^\text{160}\) Other studies looking at “prevalent care processes” and dialysis facility-specific mortality rates\(^\text{161}\) and physician profiles and cost and quality of care\(^\text{162}\) have not established a strong relationship.

The above failure to definitively relate outcome and quality of care process highlights the current debate over quality of care and its assessment by outcome or process measures;\(^\text{163,164}\) critical care has been no exception.\(^\text{165}\) For example, evidence exists for substantial variation in resource utilisation in care of sepsis patients across tertiary care centres.\(^\text{166}\) An argument for the increased sensitivity of process measures has been advanced because of the large sample sizes required to demonstrate small to modest changes in (mortality) outcome.\(^\text{145,167}\) For a reduction of mortality from 25% to 20% with 90% power and two-sided \(\alpha\) error of 0.05, approximately 3000 patients are required; 34 of the studies reported by Carmel and Rowan\(^\text{143}\) had less than 3000 patients. However, the felicity with which process may be measured is no guarantee that “measuring ...process and reporting performance will improve outcomes”.\(^\text{168}\)

Conclusions

Considerable uncertainty has been apportioned to the estimates of mortality as reflected in the SMR; in simulation studies where case-mix has been completely adjusted for, the ability of the SMR to identify outliers is sub-optimal. Furthermore, the translation of high SMR outlier status to identifiable process within providers by chart review has not been evidenced. Outlier status appears determined by random variation and/or flux over-time and needs to be more appropriately addressed. The above review would seem to have confirmed that “mortality is unlikely to be a sufficient statistic for quality.”.\(^\text{6}\) Algorithmic scoring systems at best describe “elements” of performance.

J. L. Moran
Department of Intensive Care Medicine, Queen Elizabeth Hospital, Woodville, South Australia

P. J. Solomon
School of Applied Mathematics, University of Adelaide, Adelaide, South Australia
REFERENCES


34. Stern M, Waisbren BA. A method by which burn units may compare their results with a base line curve. Surg Gynecol Obstet 1976;142:230-234.


Rosenthal GE, Kaboli PJ, Bartlett MJ. Differences in length of stay in Veterans Health Administration and other United States hospitals: is the gap closing? Med Care 2003;41:882-894.


130. Romano PS, Chan BK, Schembri ME, Rainwater JA. Can administrative data be used to compare postoperative complication rates across hospitals? Med Care 2002;40:856-867.


The Lancet is my hero; I shall not want

In what must be one of the most piercing medical editorials in recent times and with a crusading style reminiscent of Ralph Nader, The Lancet (October 25th 2003) condemned the CEO of AstraZeneca (Tom McKillop) for his marketing approach to rosuvastatin in attempting to muscle in on the billion dollar ‘statin’ market.1

What makes the editorial even more interesting is the fact that The Lancet has only two drug company advertisements in their October 25th 2003, issue, with AstraZeneca being one of the companies advertising Nexium® (esomeprazole) on the back page of the journal. This advertisement also appears on the back page of all previous issues of The Lancet for the year 2003, so it is (was?) quite a ‘money-spinner’ for the journal.

Following the publishing of the editorial, the initial conversation between Tom McKillop and Richard Horton (publisher and Editor of The Lancet) must have been robust. The formal reply to the editorial (November 1st 2003) only barely hides the pique felt by the AstraZeneca CEO when he declares his position in the ‘statin wars’ by saying “I deplore the fact that a respected scientific journal such as The Lancet should make such an outrageous critique of a serious, well studied, and important medicine”.2

This highlights once again the relationship between the pharmaceutical industry and the medical profession. Information for physicians should be completely independent and devoid of ‘spin’. However, in a multi-
billion dollar market, drug companies will go to almost any length in an attempt to improve shareholder equity. A position that many retirees may agree with when reviewing their superannuation portfolios, but perhaps not when considering their own health. The statement by the AstraZeneca CEO that “it is unthinkable that we should desist from our efforts to make this medicine [rosuvastatin] more widely available to physicians and patients” further indicates his strength of feeling when promoting his company’s product.

At a meeting of the International Committee of Medical Journal Editors, a statement supporting editorial freedom was prepared and promulgated. While editorial freedom for any medical journal may be a ‘given’, it may also come at a financial cost. Nevertheless, if a journal prefers not to compromise its fiscal position by confronting important scientific issues, it does so at the risk of becoming irrelevant.

Dr. L. I. G. Worthley
Department of Critical Care Medicine,
Flinders Medical Centre, Bedford Park
SOUTH AUSTRALIA 5042

REFERENCES