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,[1] as witness the controversy regarding the public enquiry into the performance of the paediatric cardiac surgical service at the Royal Bristol Infirmary.[2] 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),[3] should be tempered by the remembrance that the "third revolution" in medical care[4] 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.[5,6] Critical care is no exception to this movement, as witness the recent publication of state-of-the-art reviews.[7,8] 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.[9] 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*,[10] exploring the APACHE II[11] 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*,[10] paper was two fold: 1) a comparison of observed and predicted mortality rates for each hospital, via the t-test,[12] 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,[13] 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)[14] between the two measures was quite low (rho_c = 0.19).

As pointed out by Hosmer and Lemeshow,[15] 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:

i) 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: ([observed deaths] ± $z_{1-\alpha}$ σ)/(expected deaths),

      where, $\sigma = \sqrt{\sum_{i=1}^{n} \hat{\pi}_i \left(1-\hat{\pi}_i\right)}$ and

      n = patient number, $\hat{\pi}_i$ is the predicted probability for the i^th^ patient and $z_{1-\alpha}$ is the (1-α) × 100th percentile of the standard normal distribution.[17]
   b) bootstrapping,[18] in particular the B_p (bootstrap percentile) and BC_a (bias-corrected and accelerated) intervals.[19]

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_p and BC_a) 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.[26] 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 (β), 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 $z_\alpha\sigma_{ij}/\left(\sigma_i+\sigma_j\right)$ over all (i, j) pairs; a CI for the i^{th} category is given as $m_i \pm z_\beta\sigma_i$).[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]

$$Z = \frac{S - \sum_{i=1}^{n} P_i}{\sqrt{\sum_{i=1}^{n} P_i Q_i}}$$

where Z is the familiar standard normal deviate, S is the total number of survivors among the n patients, Pi the probability of survival, estimated from the baseline survival curve and $Qi = 1-Pi$, 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 =

$$\Phi\left(\frac{-K_\alpha\sqrt{\Sigma P_i Q_i} + \Sigma(P_i - P'_i)}{\sqrt{\Sigma P'_i Q'_i}}\right) + \left[1 - \Phi\left(\frac{-K_\alpha\sqrt{\Sigma P_i Q_i} + \Sigma(P_i - P'_i)}{\sqrt{\Sigma P'_i Q'_i}}\right)\right]$$

where Φ is the cumulative normal distribution function and K_α are the familiar critical values of the standard normal distribution (significance level (α); 0.1 = ± 1.65, 0.05 = ± 1.96, 0.01 = ± 2.58 and 0.001 = ± 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 α 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 standarisation; whereby a single 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. Fidler[49] 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

iii) regression modelling, using linear logistic regression,[50] although probit analysis was used by early pioneers in the field.[34,51] The common critical care outcome prediction algorithms eschew "institutional" factors, although, as seen above, these may be utilised in "snapshot" analyses.[10,52]

iv) more recently, multilevel or hierarchical modelling with a full or empirical Bayesian analytic framework has been introduced.[31,53,54] Formal (model based) account of uncertainty is undertaken and institutionnal differences are modelled explicitly using random effects; in particular, the ICU effect is derived from a specific (usually normal) distribution.[55]

There may be particular advantage in the Bayesian approach,[56] 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 i^{th} hospital is a certain fraction of the median mortality of all hospitals[53]); 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.[57] 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.[2]

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.[58] The methodology of algorithm construction and validation has been reviewed extensively in the literature,[59-62] 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,[63] and/or across populations,[64,65] the usual pattern being that of good discrimination with variable calibration.[66,67] With respect to the APACHE III algorithm,[68] the hospital length of stay adjustment appears problematic in geographical comparisons as evidenced by the study of Sirio *et al*,[69] (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".[70]

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.[71,72]

iii) the relative importance of discrimination and calibration depends upon the intended application:[58] 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.[73]

iv) methods of assessment of discrimination[74] and calibration[50] are well defined, as are comparison of receiver operator characteristic (ROC) curve areas,[75] 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".[76] 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.[77,78] Recent studies have also suggested that it is not necessarily the optimal test for logistic regression goodness-of-fit,[79] albeit the "..2 by 10 table of observed and estimated expected frequencies….provides a useful overall summary of model fit…".[80]

v) multiple predictive algorithms give rise to competing performance claims.[81] Numerous studies have suggested differential algorithm performance and consequent provider ranking changes; in: myocardial infarction,[82] general medical conditions (stroke, lung cancer, pneumonia, myocardial infarction, congestive heart failure and coronary artery surgery)[83,84] and critical care.[85,86] Less divergence was noted with trauma algorithms[87] and in coronary artery surgery when considered alone.[88]

The particular concerns above ((i) through (v)) may be illustrated using a data set (1993-1997) of the ANZICS National Data Base, previously reported.[89] A total of 73171 patients were available for comparison of APACHE II[11] and SAPS II[90] 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

Deciles	*Observed deaths*$^{\$}$	*Expected_AP II*	*Observed deaths**	*Expected_SAPS II*
1	107	61	104	63
2	175	172	153	139
3	266	296	226	181
4	365	448	447	396
5	474	646	410	399
6	713	949	871	796
7	1141	1368	1088	973
8	1991	2074	2073	1876
9	3302	3407	3198	3238
10	5508	5682	5472	5778
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

Deciles	Observed I	Expected I	Observed II	Expected II	Observed III	Expected III	Observed IV	Expected IV
1	7	9	21	13	105	50	8	12
2	12	33	50	45	161	129	15	27
3	28	52	69	77	217	207	18	35
4	53	87	99	118	321	319	21	50
5	84	140	125	183	415	473	28	69
6	93	178	145	228	706	706	40	94
7	148	257	261	345	1081	1069	56	127
8	265	378	410	512	1625	1636	111	181
9	572	626	715	814	2433	2531	167	267
10	1181	1209	1255	1376	3715	3861	510	622
H-L	216.53		193.75		116.23		265.02	
ROC(SE)	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(8)		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.

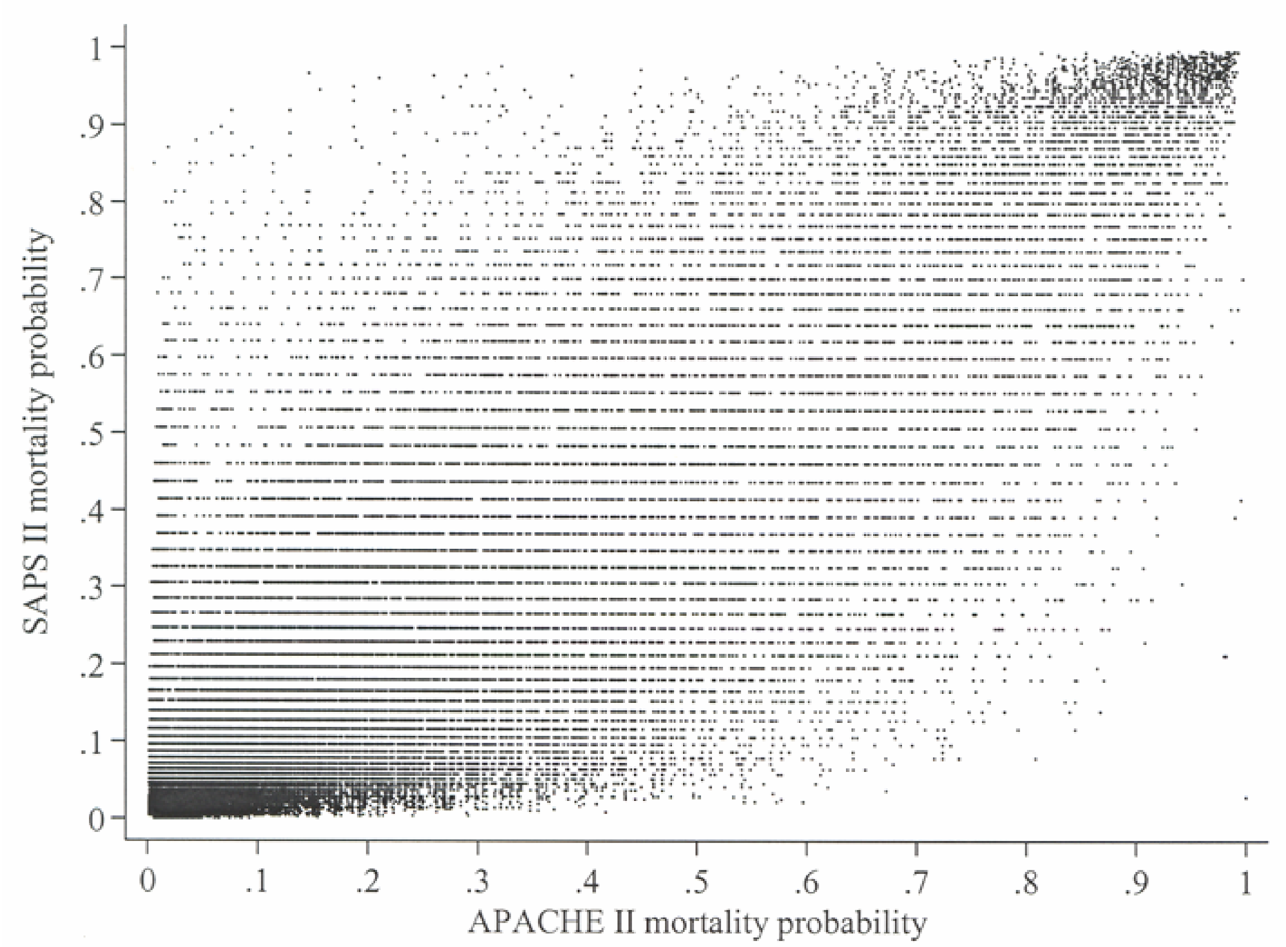

Figure 1. Comparison of APACHE II and SAPS II mortality probabilities

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_{24}.[91]

vi) the question of when to re-calibrate or customise "will be back – again and again".[92] In the trauma,[93] cardiac surgical[94] and critical care literature,[38,95] recalibration has been undertaken with varying results. Metnitz *et al*,[38] 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*,[94] 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".[71] Similarly, Champion *et al*, from the trauma perspective complained about being "sentenced to perpetual tweaking by certain researchers….".[92] 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".[96] This sentiment was also echoed by Glance and Szaldos in an editorial comment upon the above referenced Sirio *et al*,[69] 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".[97] Similarly, Rowan *et al*,[78] 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*,[98] 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 j^{th} 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 *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[53]).

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 underlie the uncertainty aspect of mortality rates.

Further evidence for the poor performance of mortality rates as performance indicators was provided by Silber *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".

viii)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 *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 *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;[112] no difference for 30 versus 180 days follow up;[113] and modest effects on mortality rates,[114] rankings and outlier status.[115] This being said, recent critical care commentaries have endorsed a move to 90 day mortality outcome.[96,97]

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.[116]

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.[117,118] The actual rankings also demonstrate instability with wide confidence intervals and time variance;[119] provider performance based upon such rankings is problematic.[31,120,121] 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.[122]

xi) time change has been recently demonstrated for hospital mortality rates (consistent decreases over time noted[123]) and indices of care (improvement confirmed[124]); thus the impact of "quality of care" interventions using mortality outcome as a yard-stick, needs careful assessment. The much publicised Cleveland Health Quality Choice[125] and NewYork Cardiac Surgery[126] 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.[127-129]

xii) although not the principal focus of this presentation, the principles and processes pertaining to the assessment of morbidity rates,[130,131] whether risk-adjusted[132,133] 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)[134,135] and diagnoses[136] 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[137] and sequential probability test charts;[138] cumulative risk-adjusted mortality (CRAM) charts,[139] variable life-adjusted displays,[140] funnel plots[141] and time series monitors.[142] 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*,[10] 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.[143] 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.[144] 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,[145] in particular, the report of large reductions (16%) in breast cancer mortality provided by specialist surgical services.[146] 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[147]) 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[148] 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".[149] Second and more pertinent to critical care concerns, the observational study by Connors *et al*,[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:[151] the metric of the sensitivity analysis was OR, not relative hazard and as RHC was a common procedure,[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,[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.[153]

Coincident with the Knaus *et al* paper,[10] Dubois and coworkers reported a study entitled "Adjusted hospital death rates: a potential screen for quality of medical care".[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*,[155] in a surgical environment nor Best *et al*,[156] Thomas *et al*,[157] nor Park *et al*,[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.[159] However, a caution must apply to the accuracy of implicit review processes, especially when outcomes are known.[160] Other studies looking at "prevalent care processes" and dialysis facility-specific mortality rates[161] and physician profiles and cost and quality of care[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;[163,164] critical care has been no exception.[165] For example, evidence exists for substantial variation in resource utilisation in care of sepsis patients across tertiary care centres.[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.[145,167] For a reduction of mortality from 25% to 20% with 90% power and two-sided α error of 0.05, approximately 3000 patients are required; 34 of the studies reported by Carmel and Rowan[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".[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".[6] Algorithmic scoring systems at best describe "elements" of performance.[169]

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

1. Davies HT, Crombie IK. Interpreting health outcomes. J Eval Clin Pract 1997;3:187-199.
2. Aylin P, Alves B, Best N, et al. Comparison of UK paediatric cardiac surgical performance by analysis of routinely collected data 1984-96: was Bristol an outlier? Lancet 2001;358:181-187.
3. McKee M, Hunter D. Mortality league tables: do they inform or mislead? Quality in Health Care 1995;4:5-12.
4. Relman AS. Assessment and accountability: the third revolution in medical care. N Engl J Med 1988;319:1220-1222.
5. Iezzoni LI. 100 apples divided by 15 red herrings: a cautionary tale from the mid-19th century on comparing hospital mortality rates. Ann Intern Med 1996;124:1079-1085.
6. Spiegelhalter DJ. Surgical audit: statistical lessons from Nightingale and Codman. Journal of the Royal Statistical Society A 1999;162:45-58.
7. Sibbald WJ, Bion J. Evaluating Critical Care: using health services research to improve quality. Berlin: Springer-Verlag; 2001.
8. Ridley S. Outcomes in Critical Care. Oxford: Butterworth-Heinemann; 2002.
9. Cox DR, Solomon PJ. Components of variance. Boca Raton: Chapman & Hall / CRC; 2003.
10. Knaus WA, Draper EA, Wagner DP, Zimmerman JE. An evaluation of outcome from intensive care in major medical centers. Ann Intern Med 1986;104:410-418.
11. Knaus WA, Draper EA, Wagner DP, Zimmerman JE. APACHE II: a severity of disease classification system. Crit Care Med 1985;13:818-829.
12. D'Agostino RB, Chase W, Belanger A. The appropriateness of some common procedures for testing the equality of two independent binomial proportions. The American Statistican 1988;42:198-202.
13. Wolfe RA, Roi LD, Flora JD, Feller I, Cornell RG. Mortality differences and speed of wound closure among specialized burn care facilities. JAMA 1983;250:763-766.
14. Steichen TJ, Cox NJ. A note on the concordance correlation coefficient. The Stata Journal 2002;2:183-189.
15. Hosmer DW, Lemeshow S. Confidence interval estimates of an index of quality performance based on logistic regression models. Stat Med 1995;14:2161-2172.
16. Zhou H, Romano PS. Confidence interval estimates of an index of quality performance based on logistic regression models. Stat Med 1997;16:1301-1303.
17. Rapoport J, Teres D, Lemeshow, S, Gehlbach, S. A method for assessing the clinical performance and cost-effectiveness of intensive care units: a multicenter inception cohort study. Crit Care Med 1994;22:1385-1391.
18. Efron B, Tibshirani R. An introduction to the bootstrap. New York: Chapman and Hall; 1993.
19. Carpenter J, Bithell J. Bootstrap confidence intervals: when, which, what? A practical guide for medical statisticians. Stat Med 2000;19:1141-1164.
20. Kirkwood BR, Sterne JA. Medical statistics. 2nd ed. Malden, MA: Blackwell Sciences Ltd; 2003.
21. Wolter KM. Taylor Series Methods. Introduction to Variance Estimation. New York: Springer-Verlag; 1985: 221-247.
22. Briggs AH, Mooney CZ, Wonderling DE. Constructing confidence intervals for cost-effectiveness ratios: an evaluation of parametric and non-parametric techniques using Monte Carlo simulation. Stat Med 1999;18:3245-3262.
23. Laska EM, Meisner M, Siegel C. Statistical inference for cost-effectiveness ratios. Health Econ 1997;6:229-242.
24. Polsky D, Glick HA, Willke R, Schulman K. Confidence intervals for cost-effectiveness ratios: a comparison of four methods. Health Econ 1997;6:243-252.
25. Wakker P, Klaassen MP. Confidence intervals for cost/effectiveness ratios. Health Econ 1995;4:373-381.
26. Willan AR, O'Brien BJ. Cost prediction models for the comparison of two groups. Health Econ 2001;10:363-366.
27. Faris PD, Ghali WA, Brant R. Bias in estimates of confidence intervals for health outcome report cards. J Clin Epidemiol 2003;56:553-558.
28. Austin PC, Hux JE. A brief note on overlapping confidence intervals. J Vasc Surg 2002;36:194-195.
29. Schenker N, Gentleman JF. On judging the significance of differences by examining the overlap between confidence intervals. The American Statistician 2001;55:182-186.
30. Goldstein H, Healy MJR. The graphical presentation of a collection of means. Journal of the Royal Statistical Society, A 1995;158:175-177.
31. Goldstein H, Spiegelhalter DJ. League tables and their limitations: Statistical issues in the comparisons of institutional performance. Journal of the Royal Statistical Society A 1996;159:385-443.
32. Flora JD Jr. A method for comparing survival of burn patients to a standard survival curve. J Trauma 1978;18:701-705.
33. Feller I, Flora JD Jr, Bawol R. Baseline results of therapy for burned patients. JAMA 1976;236:1943-1947.
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.
35. Lemeshow S, Teres D, Avrunin JS, Pastides H. A comparison of methods to predict mortality of intensive care unit patients. Crit Care Med 1987;15:715-722.
36. Taylor MS, Sacco WJ, Champion HR. On the power of a method for comparing survival of trauma patients to a standard survival curve. Comput Biol Med 1986;16:1-6.
37. Cottington EM, Shufflebarger CM, Townsend R. The power of the Z statistic: implications for trauma research and quality assurance review. J Trauma 1989;29:1500-1509.
38. Metnitz PG, Lang T, Vesely H, Valentin A, Le Gall JR. Ratios of observed to expected mortality are affected by differences in case mix and quality of care. Intensive Care Med 2000;26:1466-1472.
39. Murphy-Filkins R, Teres D, Lemeshow S, Hosmer DW. Effect of changing patient mix on the performance of an intensive care unit severity-of-illness model: how to distinguish a general from a specialty intensive care unit. Crit Care Med 1996;24:1968-1973.

40. Zhu BP, Lemeshow S, Hosmer DW, Klar J, Avrunin J, Teres D. Factors affecting the performance of the models in the Mortality Probability Model II system and strategies of customization: a simulation study. Crit Care Med 1996;24:57-63.
41. Blumberg MS. Risk adjusting health care outcomes: a methodologic review. Med Care Rev 1986;43:351-393.
42. Luft HS. The Baxter Allegiance Foundation Prize for Health Services Research Address. Statistics, tears, stories, and policy proposals: addressing the problems of risk adjustment. J Health Adm Educ 1998;16:339-351.
43. Buist M, Gould T, Hagley S, Webb R. An analysis of excess mortality not predicted to occur by APACHE III in an Australian level III intensive care unit. Anaesth Intensive Care 2000;28:171-177.
44. Goldhill DR, Withington PS. The effect of casemix adjustment on mortality as predicted by APACHE II. Intensive Care Med 1996;22:415-419.
45. Zaslavsky AM. Statistical issues in reporting quality data: small samples and casemix variation. Int J Qual Health Care 2001;13:481-488.
46. Wolfe RA. The standardized mortality ratio revisited: improvements, innovations, and limitations. Am J Kidney Dis 1994;24:290-297.
47. Clark DE. Comparing institutional trauma survival to a standard: current limitations and suggested alternatives. J Trauma 1999;47:Suppl-8.
48. Rixom A. Performance league tables. BMJ 2002;325:177-178.
49. Fidler V. The effect of case mix adjustment on mortality as predicted by APACHE II. Intensive Care Med 1997;23:711.
50. Hosmer DW, Lemeshow S. Applied Logistic Regression. Second ed. New York: John Wiley & Sons, Inc; 2000.
51. Waisbren BA, Stern M, Collentine GE. Methods of burn treatment: comparison by probit analysis. JAMA 1975;231:255-258.
52. Knaus WA, Wagner DP, Zimmerman JE, Draper EA. Variations in mortality and length of stay in intensive care units. Ann Intern Med 1993;118:753-761.
53. Normand S-LT, Glickman ME, Gatsonis CA. Statistical methods for profiling providers of medical care: issues and applications. Journal of the American Statistical Association 1997;92:803-814.
54. Thomas N, Longford NT, Rolph JE. Empirical Bayes methods for estimating hospital-specific mortality rates. Stat Med 1994;13:889-903.
55. Cox DR, Solomon PJ. Unbalanced situations. In: Cox DR, Solomon PJ. Components of Variance. Boca Raton: Chapman & Hall / CRC; 2003: 73-102.
56. Christiansen CL, Morris CN. Improving the statistical approach to health care provider profiling. Ann Intern Med 1997;127:1-8.
57. Austin PC, Naylor CD, Tu JV. A comparison of a Bayesian vs. a frequentist method for profiling hospital performance. J Eval Clin Pract 2001;7:35-45.
58. Justice AC, Covinsky KE, Berlin JA. Assessing the generalizability of prognostic information. Ann Intern Med 1999;130:515-524.
59. Braitman LE, Davidoff F. Predicting clinical states in individual patients. Ann Intern Med 1996;125:406-412.
60. Hadorn DC, Keeler EB, Rogers WH, Book RH. Assessing the performance of mortality prediction models. Santa Monica, CA: RAND; 1993.
61. Harrell FE Jr, Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. Stat Med 1996;15:361-387.
62. Laupacis A, Sekar N, Stiell IG. Clinical prediction rules. A review and suggested modifications of methodological standards. JAMA 1997;277:488-494.
63. Charlson ME, Ales KL, Simon R, MacKenzie CR. Why predictive indexes perform less well in validation studies. Is it magic or methods? Arch Intern Med 1987;147:2155-2161.
64. Dudley RA, Rennie DJ, Luft HS. Population choice and variable selection in the estimation and application of risk models. Inquiry 1999;36:200-211.
65. Sachdeva RC, Guntupalli KK. International comparisons of outcomes in intensive care units. Crit Care Med 1999;27:2032-2033.
66. Teres DM, Lemeshow SP. As American as apple pie and APACHE. Crit Care Med 1998;26:1297-1298.
67. Wood KE, Coursin DB, Grounds RM. Critical Care Outcomes in the United Kingdom: Sobering Wake-up Call or Stability of the Lamppost? Chest 1999;115:614-616.
68. Knaus WA, Wagner DP, Draper EA, et al. The APACHE III prognostic system. Risk prediction of hospital mortality for critically ill hospitalized adults. Chest 1991;100:1619-1636.
69. Sirio CAM, Tajimi KM, Taenaka NM, Ujike YM, Okamoto KM, Katsuya HM. A Cross-Cultural Comparison of Critical Care Delivery: Japan and the United States. Chest 2002;121:539-548.
70. Zimmerman JE, Draper EA, Wagner DP. Comparing ICU populations: Background and current methods. In: Sibbald WJ, Bion JF. Evaluating Critical Care: Using health services research to improve quality. Berlin: Spinger-Verlag; 2001: 121-139.
71. Krumholz HM. Mathematical models and the assessment of performance in cardiology. Circulation 1999;99:2067-2069.
72. Selker HP. Systems for comparing actual and predicted mortality rates: characteristics to promote cooperation in improving hospital care. Ann Intern Med 1993;118:820-822.
73. Harrell FE Jr, Lee KL, Califf RM, Pryor DB, Rosati RA. Regression modelling strategies for improved prognostic prediction. Stat Med 1984;3:143-152.
74. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 1982;143:29-36.
75. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics 1988;44:837-845.
76. Lemeshow S, Le Gall JR. Modeling the severity of illness of ICU patients. A systems update. JAMA 1994;272:1049-1055.
77. Castella X, Artigas A, Bion J, Kari A. A comparison of severity of illness scoring systems for intensive care unit patients: results of a multicenter, multinational study.

The European/North American Severity Study Group. Crit Care Med 1995;23:1327-1335.

78. Rowan KM, Kerr JH, Major E, McPherson K, Short A, Vessey MP. Intensive Care Society's APACHE II study in Britain and Ireland--I: Variations in case mix of adult admissions to general intensive care units and impact on outcome. BMJ 1993;307:972-977.

79. Hosmer DW, Hosmer T, Le Cessie S, Lemeshow S. A comparison of goodness-of-fit tests for the logistic regression model. Stat Med 1997;16:965-980.

80. Hosmer DW, Hjort NL. Goodness-of-fit processes for logistic regression: simulation results. Stat Med 2002;21:2723-2738.

81. Teres D, Lemeshow S. As American as apple pie and APACHE. Acute Physiology and Chronic Health Evaluation. Crit Care Med 1998;26:1297-1298.

82. Iezzoni LI, Ash AS, Shwartz M, Daley J, Hughes JS, Mackiernan YD. Predicting who dies depends on how severity is measured: implications for evaluating patient outcomes. Ann Intern Med 1995;123:763-770.

83. Iezzoni LI, Ash AS, Coffman GA, Moskowitz MA. Predicting in-hospital mortality. A comparison of severity measurement approaches. Med Care 1992;30:347-359.

84. Iezzoni LI. The risks of risk adjustment. JAMA 1997;278:1600-1607.

85. Glance LG, Osler TM, Dick A. Rating the quality of intensive care units: is it a function of the intensive care unit scoring system? Crit Care Med 2002;30:1976-1982.

86. Glance LG, Osler TM, Dick AW. Identifying quality outliers in a large, multiple-institution database by using customized versions of the Simplified Acute Physiology Score II and the Mortality Probability Model II0. Crit Care Med 2002;30:1995-2002.

87. Meredith JW, Evans G, Kilgo PD, et al. A comparison of the abilities of nine scoring algorithms in predicting mortality. J Trauma 2002;53:621-628.

88. Landon B, Iezzoni LI, Ash AS, et al. Judging hospitals by severity-adjusted mortality rates: the case of CABG surgery. Inquiry 1996;33:155-166.

89. Moran JL, Leppard PI, Churches T, Hart G, Herkes D, McWilliam D. Performance of the acute physiology and chronic health evaluation (APACHE II) predicitve algorithm in a national database. Anaesth Intensive Care 1998;26:441.

90. Le Gall JR, Lemeshow S, Saulnier F. A new Simplified Acute Physiology Score (SAPS II) based on a European/North American multicenter study. JAMA 1993;270:2957-2963.

91. Lemeshow S, Klar J, Teres D. Outcome prediction for individual intensive care patients: useful, misused, or abused? Intensive Care Med 1995;21:770-776.

92. Champion HR, Sacco WJ, Copes WS. Injury severity scoring again. J Trauma 1995;38:94-95.

93. Jones JM, Redmond AD, Templeton J. Uses and abuses of statistical models for evaluating trauma care. J Trauma 1995;38:89-93.

94. Ivanov J, Tu JV, Naylor CD. Ready-made, recalibrated, or Remodeled? Issues in the use of risk indexes for assessing mortality after coronary artery bypass graft surgery. Circulation 1999;99:2098-2104.

95. Moreno R, Apolone G. Impact of different customization strategies in the performance of a general severity score. Crit Care Med 1997;25:2001-2008.

96. Teres D, Lemeshow S. When to customize a severity model. Intensive Care Med 1999;25:140-142.

97. Glance LG, Szalados JE. Benchmarking in Critical Care: The Road Ahead. Chest 2002;121:326-328.

98. DeLong ER, Peterson ED, DeLong DM, Muhlbaier LH, Hackett S, Mark DB. Comparing risk-adjustment methods for provider profiling. Stat Med 1997;16:2645-2664.

99. Hannan EL, Kilburn H Jr, Racz M, Shields E, Chassin MR. Improving the outcomes of coronary artery bypass surgery in New York State. JAMA 1994;271:761-766.

100. Hosmer DW, Lemeshow S. Interpretation of the fitted logistic regression model. Applied Logistic Regression. 2nd ed. New York: John Wiley & Sons, Inc; 2000: 47-88.

101. Ten Have TR, Localio AR. Empirical Bayes estimation of random effects parameters in mixed effects logistic regression models. Biometrics 1999;55:1022-1029.

102. Larsen K, Petersen JH, Budtz-Jorgensen E, Endahl L. Interpreting parameters in the logistic regression model with random effects. Biometrics 2000;56:909-914.

103. Hofer TP, Hayward RA. Identifying poor-quality hospitals. Can hospital mortality rates detect quality problems for medical diagnoses? Med Care 1996;34:737-753.

104. Thomas JW, Hofer TP. Accuracy of risk-adjusted mortality rate as a measure of hospital quality of care. Med Care 1999;37:83-92.

105. Zalkind DL, Eastaugh SR. Mortality rates as an indicator of hospital quality. Hosp Health Serv Adm 1997;42:3-15.

106. Glance LG, Osler TM, Papadakos P. Effect of mortality rate on the performance of the Acute Physiology and Chronic Health Evaluation II: a simulation study. Crit Care Med 2000;28:3424-3428.

107. Glance LG, Osler T, Shinozaki T. Effect of varying the case mix on the standardized mortality ratio and W statistic: A simulation study. Chest 2000;117:1112-1117.

108. Wagner DPP. Cannot Draw Generic Conclusions from a Single Study. Crit Care Med 2001;29:1095.

109. Silber JH, Rosenbaum PR, Ross R. Comparing the contributions of groups of predictors: which outcomes vary with hospital rather than patient characteristics. Journal of the American Statistical Association 1995;90:7-18.

110. Khuri SF, Daley J, Henderson W, et al. The Department of Veterans Affairs' NSQIP: the first national, validated, outcome-based, risk-adjusted, and peer-controlled program for the measurement and enhancement of the quality of surgical care. National VA Surgical Quality Improvement Program. Ann Surg 1998;228:491-507.

111. Jencks SF, Williams DK, Kay TL. Assessing hospital-associated deaths from discharge data. The role of length of stay and comorbidities. JAMA 1988;260:2240-2246.

112. Rosenthal GE, Kaboli PJ, Barnett 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.

113. Garnick DW, DeLong ER, Luft HS. Measuring hospital mortality rates: are 30-day data enough? Ischemic Heart Disease Patient Outcomes Research Team. Health Serv Res 1995;29:679-695.
114. Kaboli PJ, Barnett MJ, Fuehrer SM, Rosenthal GE. Length of stay as a source of bias in comparing performance in VA and private sector facilities: lessons learned from a regional evaluation of intensive care outcomes. Med Care 2001;39:1014-1024.
115. Johnson ML, Gordon HS, Petersen NJ, et al. Effect of definition of mortality on hospital profiles. Med Care 2002;40:7-16.
116. Miller MG, Miller LS, Fireman B, Black SB. Variation in practice for discretionary admissions. Impact on estimates of quality of hospital care. JAMA 1994;271:1493-1498.
117. Localio AR, Hamory BH, Sharp TJ, Weaver SL, TenHave TR, Landis JR. Comparing hospital mortality in adult patients with pneumonia. A case study of statistical methods in a managed care program. Ann Intern Med 1995;122:125-132.
118. Localio AR, Hamory BH, Fisher AC, TenHave TR. The public release of hospital and physician mortality data in Pennsylvania. A case study. Med Care 1997;35:272-286.
119. Marshall EC, Spiegelhalter DJ. Reliability of league tables of in vitro fertilisation clinics: retrospective analysis of live birth rates. BMJ 1998;316:1701-1704.
120. Marshall EC, Spiegelhalter DJ. Institutional performance. In: Leyland AH, Goldstein H. Multilevel modelling of health statistics. Chichester, West Sussex: John Wiley & Sons, Ltd; 2001: 127-142.
121. Spiegelhalter D. Ranking institutions. Journal of Thoracic & Cardiovascular Surgery 2003;125:1171-1173.
122. Moran JL, Newson R, Solomon P, ANZICS Adult Data Base. ICU ranking: appropriate confidnece intervals of the stanardised mortality raio and ICU ranks. Anaesth Intensive Care 2002;30:524.
123. Jarman B, Gault S, Alves B, et al. Explaining differences in English hospital death rates using routinely collected data. BMJ 1999;318:1515-1520.
124. Jencks SF, Huff ED, Cuerdon T. Change in the quality of care delivered to Medicare beneficiaries, 1998-1999 to 2000-2001. JAMA 2003;289:305-312.
125. Rosenthal GE, Quinn L, Harper DL. Declines in hospital mortality associated with a regional initiative to measure hospital performance. Am J Med Qual 1997;12:103-112.
126. Hannan EL, Kilburn H Jr, Racz M, Shields E, Chassin MR. Improving the outcomes of coronary artery bypass surgery in New York State. JAMA 1994;271:761-766.
127. Baker DW, Einstadter D, Thomas CL, Husak SS, Gordon NH, Cebul RD. Mortality trends during a program that publicly reported hospital performance. Med Care 2002;40:879-890.
128. Clough JD, Engler D, Snow R, Canuto PE. Lack of relationship between the Cleveland Health Quality Choice project and decreased inpatient mortality in Cleveland. Am J Med Qual 2002;17:47-55.
129. Ghali WA, Ash AS, Hall RE, Moskowitz MA. Statewide quality improvement initiatives and mortality after cardiac surgery. JAMA 1997;277:379-382.
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.
131. Romano PS, Schembri ME, Rainwater JA. Can administrative data be used to ascertain clinically significant postoperative complications? Am J Med Qual 2002;17:145-154.
132. Copeland GP. The POSSUM System of Surgical Audit. Arch Surg 2002;137:15-19.
133. Daley J, Khuri SF, Henderson W, et al. Risk adjustment of the postoperative morbidity rate for the comparative assessment of the quality of surgical care: results of the National Veterans Affairs Surgical Risk Study. Am J Coll Surg 1997;185:328-340.
134. Arozullah AM, Henderson WG, Khuri SF, Daley J. Postoperative mortality and pulmonary complication rankings: how well do they correlate at the hospital level? Med Care 2003;41:979-991.
135. Hartz AJ, Kuhn EM. Comparing hospitals that perform coronary artery bypass surgery: the effect of outcome measures and data sources. Am J Public Health 1994;84:1609-1614.
136. Rosenthal GE, Shah A, Way LE, Harper DL. Variations in standardized hospital mortality rates for six common medical diagnoses: implications for profiling hospital quality. Med Care 1998;36:955-964.
137. Cook DA, Steiner SH, Cook RJ, Farewell VT, Morton AP. Monitoring the evolutionary process of quality: risk-adjusted charting to track outcomes in intensive care. Crit Care Med 2003;31:1676-1682.
138. Grigg OA, Farewell VT, Spiegelhalter DJ. Use of risk-adjusted CUSUM and RSPRT charts for monitoring in medical contexts. Stat Methods Med Res 2003;12:147-170.
139. Sismanidis C, Bland M, Poloniecki J. Properties of the cumulative risk-adjusted mortality (CRAM) chart, including the number of deaths before a doubling of the death rate is detected. Med Decis Making 2003;23:242-251.
140. Lovegrove J, Valencia O, Treasure T, Sherlaw-Johnson C, Gallivan S. Monitoring the results of cardiac surgery by variable life-adjusted display. Lancet 1997;350:1128-1130.
141. Spiegelhalter D. Funnel plots for institutional comparison. Quality & Safety in Health Care 2002;11:390-391.
142. Marshall G, Shroyer AL, Grover FL, Hammermeister KE. Time series monitors of outcomes. A new dimension for measuring quality of care. Med Care 1998;36:348-356.
143. Carmel S, Rowan K. Variation in intensive care unit outcomes: a search for the evidence on organizational factors. Curr Opin Crit Care 2001;7:284-296.
144. Pronovost PJ, Jenckes MW, Dorman T, et al. Organizational characteristics of intensive care units related to outcomes of abdominal aortic surgery. JAMA 1999;281:1310-1317.
145. Mant J, Hicks NR. Assessing quality of care: what are the implications of the potential lack of sensitivity of outcome measures to differences in quality? J Eval Clin Pract 1996;2:243-248.

146. Gillis CR, Hole DJ. Survival outcome of care by specialist surgeons in breast cancer: a study of 3786 patients in the west of Scotland. BMJ 1996;312:145-148.
147. Hartz AJ, Kuhn EM, Krakauer H. The relationship of the value of outcome comparisons to the number of patients per provider. Int J Qual Health Care 1997;9:247-254.
148. Randolph AG, Pronovost P. Reorganizing the delivery of intensive care could improve efficiency and save lives. J Eval Clin Pract 2002;8:1-8.
149. Concato J, Horwitz RI, Feinstein AR, Elmore JG, Schiff SF. Problems of comorbidity in mortality after prostatectomy. JAMA 1992;267:1077-1082.
150. Connors AF Jr, Speroff T, Dawson NV, et al. The effectiveness of right heart catheterization in the initial care of critically ill patients. SUPPORT Investigators. JAMA 1996;276:889-897.
151. Lin DY, Psaty BM, Kronmal RA. Assessing the sensitivity of regression results to unmeasured confounders in observational studies. Biometrics 1998;54:948-963.
152. Zhang J, Yu K. What's the Relative Risk?: A Method of Correcting the Odds Ratio in Cohort Studies of Common Outcomes. [Article]. JAMA 1998;280:1690-1691.
153. Davis CE. The effect of regression to the mean in epidemiologic and clinical studies. Am J Epidemiol 1976;104:493-498.
154. Dubois RW, Brook RH, Rogers WH. Adjusted hospital death rates: a potential screen for quality of medical care. Am J Public Health 1987;77:1162-1166.
155. Gibbs J, Clark K, Khuri S, Henderson W, Hur K, Daley J. Validating risk-adjusted surgical outcomes: chart review of process of care. Int J Qual Health Care 2001;13:187-196.
156. Best WR, Cowper DC. The ratio of observed-to-expected mortality as a quality of care indicator in non-surgical VA patients. Med Care 1994;32:390-400.
157. Thomas JW, Holloway JJ, Guire KE. Validating risk-adjusted mortality as an indicator for quality of care. Inquiry 1993;30:6-22.
158. Park RE, Brook RH, Kosecoff J, et al. Explaining variations in hospital death rates. Randomness, severity of illness, quality of care. JAMA 1990;264:484-490.
159. Daley J, Forbes MG, Young GJ, et al. Validating risk-adjusted surgical outcomes: site visit assessment of process and structure. National VA Surgical Risk Study. J Am Coll Surg 1997;185:341-351.
160. Schroeder SA, Kabcenell AI. Do bad outcomes mean substandard care? JAMA 1991;265:1995.
161. Lowrie EG, Teng M, Lacson E, Lew N, Lazarus JM, Owen WF. Association between prevalent care process measures and facility-specific mortality rates. Kidney Int 2001;60:1917-1929.
162. Hofer TP, Hayward RA, Greenfield S, Wagner EH, Kaplan SH, Manning WG. The unreliability of individual physician "report cards" for assessing the costs and quality of care of a chronic disease. JAMA 1999;281:2098-2105.
163. Crombie IK, Davies HT. Beyond health outcomes: the advantages of measuring process. J Eval Clin Pract 1998;4:31-38.
164. Mant J. Process versus outcome indicators in the assessment of quality of health care. Int J Qual Health Care 2001;13:475-480.
165. Pronovost PJ, Miller MR, Dorman T, Berenholtz SM, Rubin H. Developing and implementing measures of quality of care in the intensive care unit. Curr Opin Crit Care 2001;7:297-303.
166. Yu DT, Black E, Sands KE, et al. Severe sepsis: variation in resource and therapeutic modality use among academic centers. Critical Care (London) 2003;7:R24-R34.
167. Mant J, Hicks N. Detecting differences in quality of care: the sensitivity of measures of process and outcome in treating acute myocardial infarction. BMJ 1995;311:793-796.
168. Allison JJ. Quality assessment tools: incremental advance or paradigm shift? Med Care 2003;41:575-578.
169. Linde-Zwirble WT, Angus DC. Can scoring systems assess ICU performance? J Intensive Care Med 1998;13:155-157.

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 multibillion dollar market, drug companies will go to almost any length in an attempt to improve shareholder equity.[3,4] 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.[2]

At a meeting of the International Committee of Medical Journal Editors, a statement supporting editorial freedom was prepared and promulgated.[5] 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

1. Editorial. The statin wars: why AstraZeneca must retreat. Lancet 2003;362:1341.
2. McKillop T. The statin wars. Lancet 2003;362:1498.
3. Bodenheimer T. Uneasy alliance: clinical investigators and the pharmaceutical industry. N Engl J Med 2000;342:1539-1544.
4. Thomas PS, Tan KS, Yates DH. Sponsorship, authorship, and accountability. Lancet 2002;359:351.
5. A Statement by the International Committee of Medical Journal Editors. Editorial freedom. Lancet 1988;ii:1089.