Editorials

Critical care trials: sample size, power and interpretation

Two recent contributions to this journal have highlighted particular aspects of significant clinical trials in Critical Care:[1,2] the French Pulmonary Artery Catheter Study Group's trial of pulmonary artery catheters (PAC) in septic shock and ARDS,[3] and the SAFE trial of saline and albumin resuscitation in ICU patients.[4] An apposite comment was made regarding the SAFE trial that "Clinician's interpretation of this … [trial] …will undoubtedly be influenced by previous convictions".[1]

Such "convictions" should properly include expectations regarding trial conduct; in particular, appropriate sample size and the correct interpretation of results when 'sufficient' trial sample size is not achieved. It is therefore instructive to consider, from these perspectives, the above two studies[3,4] and a third, Transfusion Requirements in Critical Care,[5] which has been repeatedly cited in follow-up transfusion studies, such as the recently reported CRIT study of anaemia and blood transfusions in the critically ill.[6]

In the French Pulmonary Artery Catheter Study Group trial,[3] 676 patients were randomised to receive a PAC (n = 335) or not (n = 341), with a primary end point of 28 day mortality. Initial sample size (with one interim analysis at 500 enrolled patients) was estimated at 1100 ($\alpha = 0.05$, $\beta = 0.1$), based upon an anticipated 10% mortality difference (35% vs. 45%, for a global mortality of 40%, with "balanced group mortalities of 35% and 45%"). The treatment estimate was a risk ratio (RR) = 0.97 with 95% CI 0.86 - 1.10 and P = 0.67. At study conclusion (cessation was at 30 months by the Data Safety & Monitoring Board due to slow recruitment), the power to detect a 10% mortality difference was 78% and, as the authors noted, underpowered also to detect the postulated 5% absolute mortality difference (≡ odds ratio of 1.24) of the previous Connors *et al* observational study of the mortality effect of PAC.[7] The observed RR (0.97) in the French Pulmonary Artery Catheter Study Group trial equated to an absolute risk difference (RD) of -1.6% (95% CI: -9% to 5.8%). The appropriate question is: what are we to make of this effect estimate when the final sample size and power is reduced?

One stratagem is to assess the treatment effect with respect to the estimated post-hoc power; with a total sample size of 680 and a single interim analysis (at 45% of total sample size), the power to detect 5% and 10% mortality difference was 26% and 76% respectively (by our calculations, using the S+SeqTrial2 module running under S-Plus® V 6.2 software, with O,Brien-Fleming stopping boundaries).[8,9] However, inference from retrospective power has been properly criticised,[10-12] and an alternative approach, adopted in the French Pulmonary Artery Catheter Study Group trial, is to consider inference from the observed difference, as outlined by Hauck and Anderson.[13] The latter employed an equivalence testing approach[14] to quantify (with the generation of appropriate P values) "…what was actually determined from the study….a possible outcome of the equivalence testing approach is the conclusion at the 5 per cent level that two…proportions …do not differ by more than some specified amount".[13] Using both 90% confidence intervals and two simultaneous one-sided (*t*) tests (the TOST procedure) for the specified difference,[15] it can be shown (we use the Stata™ module "equipi"[16] and the "Analysis of proportions" module in NCSS, release 2004)[17] that "equivalence" is achieved for a threshold 28-day mortality difference between the two treatment groups of 7.8%, in agreement with the estimate reported by the French Pulmonary Artery Catheter Study Group authors. Therefore "we can conclude at an α risk of 5% that the absolute difference in mortality rate between the 2 groups is no more than 7.8%".[3]

The Transfusion Requirements in Critical Care trial, conducted by Hebert *et al*,[5] and published in 1999, has been pivotal in determining critical care physician attitudes to transfusion; in particular, that a restrictive red blood cell transfusion strategy (in this case, haemoglobin concentrations maintained at 7.0 to 9.0 g/dL) was "equivalent" to a liberal strategy (haemoglobin concentrations maintained at 10.0 to 12.0 g/dL). The primary outcome of the trial was "death from all causes in the 30 days after randomisation". The trial enrolled 838 patients with 418 randomised to restrictive and 420 to liberal transfusion strategies. The 30-day mortalities of 18.7% and 23.3% respectively, were described in the published report as "similar" (P = 0.11) and the conclusion was that a "…restrictive strategy of red-cell transfusions is at least as effective as and possibly superior to a liberal transfusion strategy in critically ill patients". As opposed to the French Pulmonary Artery Catheter Study Group trial, Hebert *et al*, as outlined in the published methods, conducted an "equivalency trial".[5] As previously shown in this journal,[14] the null hypotheses in superiority and equivalence trials are reversed: in a superiority trial, the null hypothesis (H_0) is that the treatments have equal effects and in an equivalence trial, H_0 is that there is a specified

difference (Δ). Retention of H_0 in a superiority trial does not establish equivalence (the null hypothesis of no difference is not "proved");[18] rejection of H_0 (and acceptance of the alternative hypothesis, H_a) in an equivalence trial establishes that the treatments do not differ by more than the specified Δ. Testing for equivalence (or non-inferiority) and superiority within the same trial is possible,[19] but trial methodology statements must pre-specify this and there must be initial demonstration of equivalence.

Estimated sample size in the Transfusion Requirements in Critical Care trial showed progressive re-adjustments over time and the final statement was that the (recalculated) sample size of 1620 "…allowed us to rule out an absolute difference in the 30-day mortality rate of 5.5 percent…".[5] However, due to poor recruitment, the study was terminated at 838 patient enrolments. Neither the study authors nor commentators formally canvassed the consequences of this early termination in terms of trial end-points. Applying the same methods as above, for 30-day mortality at an α risk of 5%, the absolute difference in mortality rate between the 2 groups is calculated to be no more than 9.3%, and for hospital mortality 10.9%. Hospital mortality, albeit a secondary end point, was a focal-point of discussion in both the trial report[5] and the accompanying editorial,[20] the latter describing the higher in-hospital mortality associated with liberal transfusion practices as "striking". However, the "significance" of the in-hospital mortality difference was marginal (P = 0.05, rounded: on a battery of 8 tests provided by NCSS software,[17] P was always ≥ 0.051) and no adjustments were made for multiple testing. Combined testing for both equivalence and superiority yielded 'significant' results (for in-hospital mortality difference) only at Δ = ± 10.9%. Thus the trial goal of an equivalence margin of 5.5% between restrictive and liberal red blood cell transfusion regimens was *not* achieved and was demonstrated only at the 9% - 11% level.

The publishing of the results of the SAFE trial[4] was welcomed on a number of fronts,[1,21] none the least of which was the ability to conduct large trials[22] in the critically ill over a relatively short period of time without insurmountable enrolment difficulties. A trial sample size of 7000 (with two interim analyses at 2333 (33%) and 4666 (67%) patients, using Haybittle-Peto boundaries)[8] provided a 90% power to detect a 3% absolute mortality difference between the two treatment groups from an estimated baseline mortality rate of 15%. The hypothesis tested (H_0) was that "when 4 percent albumin is compared with 0.9 percent sodium chloride (normal saline) for intravascular-fluid resuscitation in patients in the ICU, there is no difference in the 28-day rate of death from any cause"; a superiority hypothesis, which was not rejected: RR = 0.99 (95% CI: 0.91 - 1.09) corresponding to an absolute RD of 0.07% (95% CI: -2% to 1.8%). The conclusion drawn was that there was "…evidence that albumin and saline should be considered clinically equivalent for intravascular volume resuscitation in a heterogeneous population of patients in the ICU". However, the accompanying editorial noted that the overall treatment effect "suggests equivalence, although proof of equivalence would require a different sample-size calculation".[21] This editorial claim is of some importance: first, because a negative (superiority) trial cannot assert the null hypothesis "proved" (see above); second, no formal demonstration of equivalence was proposed in the published trial report[4] nor in the earlier methods paper.[23] What are we to think?

Nominal fixed sample size for a 3% difference in mortality outcomes would vary between 5500 (mortality reduction 15% to 12%) to 6000 (global mortality of 15% based upon "balanced group mortalities" (see above) of 16.5% and 13.5%). The influence of the interim analysis with Haybittle-Peto stopping rules is to increment the nominal sample size by a function $R_{H\text{-}P}$ which is defined by the number of analyses or groups of observations (K), α and β, such that total N =($N_{\text{fixed sample}} \times R_{H\text{-}P}$). For 3 analyses with two-sided $\alpha = 0.05$ and $\beta = 0.1$, $R_{H\text{-}P} = 1.007$,[8,24] a seemingly small increment. For the classical Pocock and O'Brien-Fleming designs, R is 1.15 and 1.016 respectively; these constants refer to the maximum sample size, not the average sample size (ASN), which, given the possibility of stopping early for these latter two designs, is less than the fixed sample size.[8] As Jennison and Turnbull note: "Although Haybittle-Peto tests do not attain the maximum possible reductions in expected sample size, this is not always the key issue…"; and it may be that where large sample sizes are needed "…the investigator's objective is really to gather as much information as possible on all aspects of treatment, and there is little incentive for early stopping apart from the ethical need to cease randomising patients to a clearly inferior treatment".[24] This perspective is contrasted to that operative in the recently reported ALVEOLI trial (high versus low PEEP in ARDS), where a 10% mortality reduction (28% to 18%) was sought: "Asymmetric stopping boundaries (with a two-sided (alpha) = 0.05) were designed to allow early termination of the trial if the use of higher PEEP was found to reduce mortality or if there was a low probability that the trial could demonstrate a lower mortality rate in the higher-PEEP group than in the lower-PEEP group (futility stopping rule)".[25]

For a sample size of 7000 and a simulation based (90%) two-sided CI approach for the difference (3%) in proportions based upon an equivalence hypothesis, the power at basal mortalities of 15% (projected mortality)

and 21% (the actual mortality of the SAFE study) is computed to be 93% and 87% respectively[26] (similar results were generated from PASS).[27] As noted, H_0 in the SAFE trial was not rejected and using the above method of "inference from the observed difference",[13] at an α risk of 5% the absolute difference in mortality rate between the 2 groups is no more than 1.7%. The 3% difference in mortality rates targeted in the SAFE trial was based on the "approximate minimal effect suggested by the lower confidence interval in the Cochrane Injury Group Albumin Reviewers Paper",[28] which used a fixed effect estimator to calculate the pooled difference, but this lower confidence interval was 1.6% for the random effect estimate. The choice of meta-analytic estimator is contentious, independent of the demonstration of heterogeneity.[29] An appropriate Δ has been defined as a fraction (0.2 - 0.5) of either the treatment effect, control drug versus placebo or of the lower 95% of this treatment effect, derived from a large trial or meta-analysis.[30] Another widely used therapeutic intervention, fibrinolysis for myocardial infarction, has an overall absolute 35-day mortality reduction of 1.84%(95% CI: -2.34 to -1.35)[31] and an absolute mortality difference of 1.5% has been suggested as a reasonable compromise for equivalence studies in this scenario, although lesser absolute levels (down to 0.40 %) have been implemented in cardiology equivalence trials.[8,32] Although these margins (which operationally may be defined as minimal clinically important differences) may seem small, from the perspective of a therapy applied to a "diverse population of critically ill patients" they translate into a noteworthy absolute number of (potential) deaths: for the 16 units in the trial enrolling 7000 patients over 20 months, 210 potential deaths at the 3% mortality treatment difference and 119 at a 1.7% difference. It would seem that a precise definition of these minimally clinically important differences will "…undoubtedly be influenced by previous convictions".[1]

What can be construed from the above. First, and obvious, that achieving planned sample size is critical for the assessment of trail reports.[33] Second, both the French Pulmonary Artery Catheter Study Group trial[3] and the Transfusion Requirements in Critical Care trial[5] failed to establish their primary end-points and treatment recommendations based on these end-points must be circumspect. Third, the "equivalence" of 0.9% saline and 4 percent albumin for intra-vascular-fluid resuscitation in patients in the ICU would appear to be located at the 1.7 % absolute risk-difference level. Fourth, minimal clinically important differences for critically-ill patient categories need to be established.

The interpretation of treatment effects in trial reports is never that simple; a consideration of "inference from the observed difference" is an aide to the perplexed clinician.

Dr. J. L. Moran
Intensive Care Unit,
The Queen Elizabeth Hospital, Woodville
SOUTH AUSTRALIA 5011

Associate Professor P. J. Solomon
School of Applied Mathematics,
University of Adelaide, Adelaide
SOUTH AUSTRALIA 5005

REFERENCES

1. Cooper J. Resuscitation fluid controversies - Australian trials offer new insights. Critical Care and Resuscitation 2004;6:83-84.
2. Morgan TJ. Life without the PA catheter. Critical Care and Resuscitation 2004;6:9-12.
3. Richard C, Warszawski, J, Anguel N, et al. Early use of the pulmonary artery catheter and outcomes in patients with shock and acute respiratory distress syndrome: a randomized controlled trial. JAMA 2003;290:2713-2720.
4. The SAFE Study Investigators. A comparison of albumin and saline for fluid resuscitation in the intensive care unit. N Engl J Med 2004;350:2247-2256.
5. Hebert PC, Wells G, Blajchman MA, et al. A multicenter, randomized, controlled clinical trial of transfusion requirements in critical care. Transfusion Requirements in Critical Care Investigators, Canadian Critical Care Trials Group. N Engl J Med 1999;340:409-417.
6. Corwin HL, Gettinger A, Pearl RG, et al. The CRIT study: Anemia and blood transfusion in the critically ill-Current clinical practice in the United States. Crit Care Med 2004;32:39-52.
7. 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.
8. Moran JL, Solomon PJ. Some aspects of the design and monitoring of clinical trials. Critical Care and Resuscitation 2003;5:137-146.
9. S+SEQTRIAL 2 User's Manual. Seattle, WA: Insightful Corporation; 2002.
10. Zumbo BD, Hubley AN. A note on misconceptions concerning prospective and retrospective power. The Statistician 1998;47:385-388.
11. Hoenig JM, Heisey DM. The abuse of power: The pervasive fallacy of power calculations for data analysis. The American Statistician 2001;55:19-24.
12. Goodman SN, Berlin JA. The use of predicted confidence intervals when planning experimants and the misuse of power when interpreting results. Ann Intern Med 1994;121:200-206.
13. Hauck WW, Anderson S. A proposal for interpreting and reporting negative studies. Stat Med 1986;5:203-209.

14. Moran JL, Solomon PJ. The interpretation of lack of evidence of a difference in efficacy: equivalence trials and the treatment of fungal infections. Critical Care and Resuscitation 2003;5:216-223.
15. Rogers JL, Howard KI, Vessey JT. Using significance tests to evaluate equivalence between two experimantal groups. Psychol Bull 1993;113:553-565.
16. Goldstein R. sg21: Equivalence testing. Stata Technical Bulletin Reprints 1994;3:107-112.
17. NCSS 2004. Kaysville, Utah: Number Cruncher Statistical Systems; 2004.
18. Blackwelder WC. "Proving the null hypothesis" in clinical trials. Contol Clin Trials 1982;3:345-353.
19. Wiens BL. Something for nothing in noninferiority/superiority testing: A caution. Drug Inf J 2001;35:241-245.
20. Ely EW, Bernard GR. Transfusions in Critically Ill Patients. N Engl J Med 1999;340:467-468.
21. Cook D. Is Albumin Safe? N Engl J Med 2004;350:2294-2296.
22. Peto R, Collins R, Gray R. Large-scale randomized evidence: large, simple trials and overviews of trials. J Clin Epidemiol 1995;48:23-40.
23. Finfer S, Bellomo R, Myburgh J, Norton R. Efficacy of albumin in critically ill patients. BMJ 2003;326:559-560.
24. Jennison C, Turnbull BW. Group Sequential Methods with applications to clinical trials. Boca Raton: Chapman & Hall/CRC; 2000.
25. Higher versus Lower Positive End-Expiratory Pressures in Patients with the Acute Respiratory Distress Syndrome. N Engl J Med 2004;351:327-336.
26. Elashoff JD. Sample size Tables for Proportions. nQuery Advisor®Version 5.0 User's Guide. Cork, Ireland: Statistical Solutions Ltd; 2002: 15-1-15-46.
27. PASS 2002. Kaysville, Utah: Number Cruncher Statistical Systems; 2002.
28. ANZICS Clinical Trials Group and Institute for International Health SAFE Study Investigators. The Saline vs. Albumin Fluid Evaluation (SAFE) Study (ISRCTN76588266): Design and conduct of a multi-centre, blinded randomised controlled trial of intravenous fluid resuscitation in critically ill patients. B M J 2003;http://bmj.bmjjournals.com/cgi/content/full/326/7389/559/DC1.
29. Normand SL. Meta-analysis: formulating, evaluating, combining, and reporting. Stat Med 1999;18:321-359.
30. Blackwelder WC. Showing a treatment is good because it is not bad: when does "noninferiority" imply effectiveness? Control Clin Trials 2002;23:52-54.
31. Indications for fibrinolytic therapy in suspected acute myocardial infarction: collaborative overview of early mortality and major morbidity results from all randomised trials of more than 1000 patients. Fibrinolytic Therapy Trialists' (FTT) Collaborative Group. Lancet 1994;343:311-322.
32. Ware JH, Antman EM. Equivalence trials. N Engl J Med 1997;337:1159-1161.
33. Moore RA, Gavaghan D, Tramer MR, Collins SL, McQuay HJ. Size is everything--large amounts of information are needed to overcome random effects in estimating direction and magnitude of treatment effects. Pain 1998;78:209-216.

Strong ions – again

Traditionally, an acid-base defect is assessed by considering the arterial blood values of pH as a measure of the intensity of acidity or alkalinity, $PaCO_2$ as a measure of the respiratory component and HCO_3^- (with qualification) as a measure of the non-respiratory or metabolic component.[1] The pH and $PaCO_2$ are measured directly, whereas HCO_3^- concentration is calculated from the Henderson-Hasselbalch equation (i.e. $HCO_3^- = 0.0306 \times PaCO_2 \times 10^{(pH - pKa)}$). The equation illustrates the simple relationship between the variables (e.g. $pH \propto HCO_3^-/PaCO_2$) and how a change in ether $PaCO_2$ or HCO_3^- can alter the pH. The clinical defect is divided into a respiratory or metabolic acidosis or alkalosis (with or without compensation) and with the additional calculation of the anion gap one can determine whether there is an obvious anion or non-anion gap acidosis.[2] For the clinician, the disorder can be typified easily and managed appropriately.

Recently, an acid-base analysis based on the law of electroneutrality in aqueous solutions has provoked some interest.[3,4] The analysis contends that only independent variables of strong ions (e.g. sodium, potassium, calcium, magnesium, chloride and organic anions), PCO_2 and non-volatile weak acids (A_{TOT} which are predominantly the albuminate ions), can change acid-base status, as they change the dependent variables (predominantly HCO_3^-) to maintain electrical neutrality. The metabolic acid-base abnormality is characterised by calculating the strong ion difference (or $SID = [Na^+ + K^+ + Ca^{2+} + Mg^{2+}] - [Cl^- + lactate]$); a value which is essentially equal to the sum of the bicarbonate and albuminate ions[14] and similar to the buffer base component described by Singer and Hastings 50 years ago.[5-7] In this edition of the journal, Lloyd describes the use of a calculator to derive SID as well as A_{TOT} and a strong ion gap (as an "anion gap on steroids") using either measured or, in their absence, 'default' plasma values.[8]

The concept that weak acid anions, particularly bicarbonate, must change its concentration to conform to the space available that is altered by strong ions, reflects largely the effect in man of the bicarbonate buffer pair in an open system (i.e. CO_2 can be retained or can escape via the lungs altering the bicarbonate ion). Open systems, or systems that move substances in and out of the extracellular fluid space (the *raison d'être* for extracellular fluid) are not unique, and confusing the *in vivo* with *in vitro* events is not new in medicine. In the

instance of acid-base analysis it previously lead to the great trans-atlantic debate.[9]

This does not mean that an analysis of acid-base based on the law of electroneutrality in aqueous solutions is wrong, it is just another way of taking a snapshot of plasma acid-base variables. So to realise that arterial pH is determined by the SID, $PaCO_2$ and A_{TOT} (where $A_{TOT} = HA + A^-$) is not a fundamental revelation, as arterial pH is also determined by the bicarbonate buffer pair (i.e. HCO_3^- and $PaCO_2$), the bicarbonate element of which may be written as (SID - A^-). In other words the Henderson-Hasselbalch equation may be written as pH $\propto$ (SID - A^-)/$PaCO_2$.[10]

What is wrong, then, with an analysis of acid-base based on the law of electroneutrality in aqueous solutions when managing patients with an acid base defect?

While Morgan highlighted the problems when a strong ion difference is calculated from plasma alone (when haemoglobin is the main contributor to A_{TOT} for whole blood) and when not taking into account the Gibbs-Donnan distributional effects between compartments in the ECF[11] (he also added an excellent review of the strong ion gap in comparison with other 'gaps' for unmeasured, or unsuspected, ions);[12] I believe it also leads the clinician to ignore the central role of bicarbonate in acid-base balance, causing erroneous conclusions to be drawn and inappropriate therapy to be considered. For example, beliefs such as:

a) *Sodium bicarbonate corrects an acidosis by altering the sodium concentration, not by increasing HCO_3.*[13]

The statement that "administration of HCl or NaOH causes an acidosis or alkalosis not due to OH^- or H^+ but by giving a strong anion in the case of HCl and strong cation in the case of NaOH",[8] provides us with the corollary of "administering HCl or NaOH causes hyperchloraemia or hypernatraemia not due to Cl^- or Na^+ but by giving a strong cation in the case of HCl and strong anion in the case of NaOH". The argument confuses the requirements of electrical neutrality with the determination of an acid or base (or sodium or chloride).

b) *Metabolic alkalosis caused by pyloric stenosis is due to a loss of chloride, not loss of H^+.*[14]

The concept that a metabolic alkalosis due to pyloric stenosis is due to loss of chloride not hydrogen ion was proposed by one author because "the amount of total body free H^+ is only about 1.6×10^{-7} mol. If physiology were just simple accounting, a patient with pyloric stenosis would rapidly run out of H^+",[14] a notion that gives no consideration to the vast movement of H^+ between buffers and one that leads to the erroneous disorder of 'hyperthermic acidosis'.[15]

It is also proposed that in a low anion gap acidosis, when the sodium concentration is elevated and therapy other than $NaHCO_3$ is being examined, one should consider "removal of $Cl^- > Na^+$ perhaps by use of renal replacement therapy (such as haemofiltration)".[14] This gives no consideration to the management of such a disorder with potassium citrate, acetate or lactate (with HCO_3^- generated by anion metabolism), which have been used for many years for these disturbances, particularly when hypokalaemia exists.

c) *Chloride is an 'acid'.*[16]

This issue was resolved more than 75 years ago by Brønsted[17] and Lowry,[18] who defined an acid as a proton (or H^+) donor, and a base as a proton (or H^+) acceptor. In the case of sodium and chloride; Cl^- has always been considered a base and Na^+ an acid.

d) *Alterations in plasma protein levels lead to conditions of hyperproteinaemiac acidosis and hypoproteinaemiac alkalosis.*[19]

This has been proposed despite there being no respiratory compensation (and thus no perceived acid-base abnormality by the human organism)[20] and no study to demonstrate a regulation of albumin metabolism for the purpose of pH regulation.

e) *The understanding of acid-base physiology has been hampered by using pH.*[8]

One of the reasons for differentiating pH from H^+ concentration is to separate 'intensive' and 'extensive' variables.[21] Confusing acidic intensity (as measured by pH, i.e. extremely small quantities of free H^+ or H_3O^+) with acidic capacity (as measured by titratable acidity, i.e. extremely large quantities of H^+) leads to the misunderstanding of the pH and acid load effects of intravenous solutions (e.g. 5% dextrose solution has a pH of 5.4 but a miniscule quantity of acid),[22] and an overestimation of the effects of 20 mmol of HCl in a 70 kg man,[23] when up to 360 mmol of HCl has been administered without the proposed theoretical effects being observed.[24] The statement that SI units were not established in medical practice "before Hasselbalch put the Hassle into the Henderson equation" is not the reason for the use of pH: pH is currently recognised in SI units as a legitimate measure of acidity.[25]

The Henderson-Hasselbalch approach focuses on the bicarbonate buffer pair whereas acid-base analysis based on the law of electroneutrality focuses on the antithesis of the HCO_3^- ion (i.e. 'strong ions' and A^-). In man, acid-base balance is regulated by the renal and respiratory system regulation of the bicarbonate pair, with all other body buffer systems adjusting to the alterations in this pair.[1,26] It does not regulate acid base balance by regulating 'strong ions' and A_{TOT}.

Finally, citing the medical student turned philosopher, Arthur Schopenhauer (1788-1860), *"All truth*

passes through three stages: first, it is ridiculed; second, it is violently opposed; and third, it is accepted as self-evident"[8] may not have resonance with a probing scientist who believes that truth is never considered as self evident, as he, or she, continually challenges all theories (in this regard the strong ion theory is no different to the Henderson-Hasselbalch approach). To the scientist, eternal truth does not exist. History will decide whether acid-base analysis based on the law of electroneutrality in aqueous solutions takes a central rather than a peripheral role (or any role) in the clinical management of critically ill patients, because in reality a bad idea can't be started just as a good idea can't be stopped.

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

REFERENCES

1. Filley GF. Acid base and blood gas regulation. Lea & Febiger, Philadelphia. 1972
2. Worthley LIG. CH 45 Hydrogen ion metabolism and disorders. In Synopsis of Intensive Care Medicine. London: Churchill Livingstone, 1994: 451-466.
3. Stewart PA. Modern quantitative acid-base chemistry. Can J Physiol Pharmacol 1983;61:1444-1461.
4. Fencl V, Leith DE. Stewart's quantitative acid-base chemistry: applications in biology and medicine. Respir Physiol 1993;91:1-16.
5. Siggaard-Andersen O, Fogh-Andersen N. Base excess or buffer base (strong ion difference) as measure of a non-respiratory acid-base disturbance. Acta Anaesthesiol Scand Suppl 1995;107:123-128.
6. Singer RB, Hastings AB. An improved clinical method for the estimation of disturbances of the acid-base balance of human blood. Medicine (Baltimore) 1948;27:223-242.
7. Wooten EW. Analytic calculation of physiological acid-base parameters in plasma. J Appl Physiol 1999;86:326-334.
8. Lloyd P. Strong ion calculator – a practical bedside application of modern quantitative acid-base physiology. Critical Care and Resuscitation 2004;4:285-294.
9. Bunker JP. The great trans-atlantic acid-base debate. Anesthesiology 1965;25:591-594.
10. Worthley LIG. Strong-ion difference: a new paradigm or new clothes for the acid-base emperor. Critical Care and Resuscitation 1999;1:211-214.
11. Morgan TJ. Finding common ground in acid-base. Critical Care and Resuscitation 1999;1:123-126
12. Morgan TJ. What exactly is the strong ion gap, and does anybody care? Critical Care and Resuscitation 2004;6:155-159.
13. Dorje P, Adhikary G, McLaren ID. Dilutional acidosis or altered strong ion difference. Anesthesiology 1997;87:1011-1012.
14. Kellum JA. Metabolic acidosis in the critically ill: lessons from physical chemistry. Kidney Int 1998;53(Suppl 66):S81-S86.
15. Worthley LIG. Acid-base and strong ion difference. Critical Care and Resuscitation 1999;1:408-409
16. Story DA, Bellomo R, Kellum JA. Acid-base and strong ion difference. Critical Care and Resuscitation 1999;1:407-408
17. Brønsted JN. Einige Bemerkungen über den Begriff der Säuren und Basen. Rec Trav Chim Pays-Bas 1923;42:718-728.
18. Lowry TM. The electronic theory of valency. VI. The origin of acidity. Trans Faraday Soc 1924;20:13-15.
19. Jabor A, Kazda A. Modelling of acid-base equilibria. Acta Anaesth Scand 1995;39 (Suppl 107):119-122.
20. Wilkes P. Hypoproteinemia, strong-ion difference, and acid-base status in critically ill patients. J Appl Physiol 1998;84:1740-1748.
21. Clark WM. The determination of hydrogen ions. 3rd Ed. Baltimore, Williams & Wilkins Co., 1928.
22. Katz MA. pH vs acid load. N Engl J Med 1969;280.1480.
23. Leblanc M, Kellum JA. Section 5, Chapter 1. Biochemical and biophysical principles of hydrogen ion regulation. In, Critical care nephrology, Ed. Ronco C, Bellomo R, Kluwer Academic Publishers, Dordrecht, 1998.
24. Worthley LIG. The rational use of i.v. hydrochloric acid in the treatment of metabolic alkalosis. Br J Anaesth 1977;49:811-817.
25. Laposata M. SI conversion Guide. Boston: NEJM Books; 1992. p 6.
26. Pitts RF. Physiology of the kidney and body fluids. 2nd Ed. Year Book Medical Publishers. Chicago 1968.

Critical care unit league tables – they're coming

The Australian Council for Safety and Quality in Health Care, at its inception, identified their top priorities to included the development of "*nationally driven standards/benchmarks against which performance can be measured"* and "*accurate and honest public reporting of performance"*.[1] Public reporting of hospital and practitioner performance has had earlier and greater momentum in the United Kingdom and the United States of America, being driven by events such as the public enquiry into the performance of the paediatric cardiac surgical service at the Royal Bristol Infirmary in the UK and financial accountability of the USA health plans. Following Britain's Department of Health's decision to publish hospital measures of clinical performance in 1997, it is not surprising that the ranking of hospital performance with respect to cardiac surgery, stroke and fractured hip patient mortality appeared in

the *Times* newspaper and the term "hospital league tables" was coined.[2,3] Prospective patients in the USA can peruse the "100 best hospitals" published annually in the *US News & World Report*.[4]

To date, public reporting of individual practitioner outcome has focused predominately on cardiac surgeons, with individual surgeon risk-adjusted mortality rates being available in some parts of the USA since the early 90's. In the UK, because of their view that a validated risk-adjustment method is currently not available, the published surgeon performance will be a three-star scale, either failing, meeting or exceeding the standards of the Society of Cardiothoracic Surgeons.[5] The Victorian Department of Human Services has led the way in Australia, publishing annual outcome reports for the six Victorian public hospital cardiac units.[6] The hospital comparative data is, however, de-identified. The actual and risk-adjusted mortality data compares favorably with the USA and UK data and while all hospitals fell within 3 standard deviations of the mean, mortality varied from 1% to over 3%. Currently public reporting of health outcomes tends to focus on surgery and surgeons and ignores the contribution of other medical and nursing practitioners, such as critical care to patient outcome.

Debate in this area centers on what constitutes an adequate performance indicator and whether public reporting is helpful. Nobody disputes the need to use appropriate performance indicators to drive change for quality improvement. However, for procedures such as elective coronary artery bypass grafting with such a low mortality rate, mortality may not be an appropriate measure of quality, and risk-adjustment for co morbidities, which is also crucial to outcome, is problematic. Manipulation of data is possible and is supported by the marked, three to four-fold increases in rates of chronic obstructive pulmonary disease and congestive cardiac failure reported associated with cardiac surgery and the introduction of public reporting of performance.[7] There have also been concerns and evidence that public performance reporting reduces willingness to operate on high risk cases.[8] However, public performance reporting also resulted in some surgeons, with low operating volumes and poor outcomes, stopping operating and mortality after cardiac surgery has improved.[9] However, over the same period the USA States that did not have public reporting saw a similar improvement in mortality.[10] Despite the availability of this data, it appeared to have minimal effect on cardiologist referral decisions and it was rarely discussed with the patients.[8] Public reporting of health outcome performance is in it's infancy and certainly has the power to drive change. Whether the Australian Council of Safety and Quality in Health Care will achieve it's aims of driving change to improve health care and restore public confidence in the health care system remains to be seen.

Public reporting of health outcomes performance is coming to Australia. How quickly, and in what form, will depend on many factors. However, the first crucial step is the development of an appropriate performance indicator where the practitioners, who are involved in the procedure or field of practice, have confidence that 'quality outcome' is truly being measured.

Where does critical care medicine stand in the reporting of performance? Well it is widely held that we have an appropriate validated performance indicator. At a recent forum to develop 'key' clinical performance indicators sponsored by the South Australian Department of Human Services and orchestrated by local epidemiologists, the view was consistent, "you lot are simple, you have APACHE, just give us the data". But how well does an APACHE derived standardised mortality ratio (SMR) stack up as a performance indicator? Not surprisingly the science of developing performance indicators is well established and the three criteria central to the integrity of the measurement are:

1) *Importance of what is being measured*

   Certainly mortality, especially excessive mortality in a critical care unit, has a significant impact on health and both health administrators and consumers would be concerned about this. Excessive mortality would also be a meaningful problem for a health care system to address. However a critical care unit's SMR does not necessarily tell us how a hospital cares for their critically ill patients. Critical care access block, covert therapy limitation or inappropriate palliation may all result in poor outcome of critically ill patients in a hospital that may have an acceptable critical care unit SMR. Therefore, it is crucial to compliment a critical care unit's SMR with some hospital wide performance indicator for critically ill patient outcome. Currently, in South Australia, individual hospital performance is being benchmarked using a hospital wide SMR, with the Charlson Index to risk-adjust case-mix variation.[11] A hospital wide SMR like this would compliment a critical care unit SMR in assessing quality outcome for critically ill patients regardless of access to a critical care unit. Another direction would involve developing an index of critical care accessibility, such as critical care episodes per hospital-wide death, to signal possible low rates of critical care admission for critically ill patients.

2) *Scientific soundness of the measure*

   Does an APACHE derived SMR as a critical care performance indicator measure critically ill patient safety and quality outcome? Does it also risk-adjust adequately for the variations in critical case-mix? These questions have been much debated, and I refer readers to an in-depth discussion and reference list

by Moran *et al*.[12] Despite these authors, and others, agreeing that SMR is a poor measure of quality and performance, many believe APACHE derived SMR has sufficient scientific validation to be used in this way.

3) *Feasibility of using the measure*

While APACHE data is widely collected, feasibility is an issue and data accuracy is the predominant problem. Concerns about accuracy and reliability of both diagnostic and physiological data were universal as APACHE II data collection was widely introduced into Australia in the late 80's. Data collector numbers, qualifications, training and experience could all have an adverse effect on data accuracy. We found that, despite training, there was a clinically significant lack of agreement between medical residents and nurses as data collectors when comparing predicted mortality for individual patients, even though overall agreement for total patient groups was good.[13] This lack of agreement was largely confined to patients with high predicted mortality. Medical residents more accurately assigned diagnostic category, chronic health evaluation and operative status. The commonest causes of data errors were choice between highest or lowest value as worst, Glasgow Coma Score error and data point outside first 24 hours. In this edition of *Critical Care and Resuscitation*, McHugh again reminds us that there can be significant lack of agreement with different data collectors.[14] In this study, while no measure of variability is provided, it can be inferred that there was even less agreement between junior and senior medical officers, with there being a significant difference in median APACHE II scores. There is no information provided on differences in diagnostic, co-morbidity and derived predicted mortality which has the potential to further skew predicted mortality. McHugh suggests the need for greater involvement by senior medical staff. This may be true for some aspects of data collection such as the assignment of diagnostic categories, however the major requirement is for dedicated data collectors. Our experience would suggest that 40 minutes per admission by an experienced data collector is required in addition to data entry time and data base management.

While dedicated data collectors and limited involvement by senior medical officers may solve data accuracy, the step to public reporting of critical care unit APACHE II derived SMRs requires that all stake holders have confidence in it as a performance indicator. Despite the many concerns discussed, the next important step, I believe, is to ensure that inter-hospital bias in data collection is minimised. The capacity to bias predicted mortality with diagnostic category selection and Glasgow Coma Score assignment for instance can be significant. Control of this potential bias requires inter-hospital observer reliability audits.

So how close are we to public reporting of critical care unit SMR performance? Well it's happening. The Victorian Government Department of Human Services again leads the way. In their Report to the Public; Intensive care for adults in Victorian public hospitals 2002, individual unit SMR is published in graphical form with 95 % confidence intervals for four tertiary, six metropolitan and 4 regional hospitals in a de-identified form.[15] To compliment this public reporting of critical care outcome, the Victorian Department of Human Services has recently announced funding for APACHE data collection in participating hospitals.

The pace of further developments in public reporting of health outcomes performance, to a large extent, will depend on the health consumers, i.e. our patients. However, in the field of critical care, most areas of Australia lag significantly behind Victoria. The Australian Council for Safety and Quality in Health Care have recently published their 10 Tips for Safer Health Care for prospective patients.[16] I wonder whether prospective Victorian patients, who face major surgery requiring post-operative critical care and work through their 10 tips, ask when they come to tip 7 "Is there more than one hospital to choose from?, If so, which has the best care and results for treating my condition?", and will they then ask for hospital G?

Critical care unit league tables are not coming, they're here.

Dr. A. W. Holt
Department of Critical Care
Flinders Medical Centre
SOUTH AUSTRALIA 5042

REFERENCES

1. A consumer vision for a safer healthcare system-May 2001. Australian Council for Safety and Quality in Health Care.
2. McKee M. Indicators of clinical performance: problematic, but poor standards of care must be tackled. BMJ 1997:315;142.
3. Vass A. Doctors urge caution in interpretation of league tables. BMJ 2001:323;1205.
4. Epstein AM. Public release of performance data: A progress report from the front. JAMA 2000:283;1884-1886.
5. Neil DA, Clark S, Oakley G. Public reporting of individual surgeon performance information: United Kingdom developments and Australian issues. Med J Aust 2004:181;266-268.
6. Reid CM, Rockell M, Skillington P, Shardey G, on behalf of the Australasian Society of Cardiac and

Thoracic Surgeons Victorian Database Steering Committee. Cardiac surgery in Victorian public hospitals: report to the public 2002. Victorian Department of Human Services.
7. Green J, Wintfeld N. Report cards on cardiac surgeons: assessing New York State's approach. N Engl J Med 1995:332;1229-1232.
8. Schneider EC, Epstein AM. Influence of cardiac surgery performance reports on referral practices and access to care: a survey of cardiovascular specialists. N Engl J Med 1996:335;251-256.
9. Hannan EL, Kilburn H, Racz M, Shields E, Chassin MR. Improving the outcomes of coronary artery bypass surgery in New York State. JAMA 1994:271;761-766.
10. Ghali WA, Ash AS, Hall RE, Moskowitz MA. Statewide quality improvement initiatives and mortality after cardiac surgery. JAMA 1997:277;379-382.
11. D'Hoore W, Sicotte C, Tilquin C. Risk adjustment in outcome assessment: the Charlson comorbid index. Methods Inf Med 1993:32;382-387.
12. Moran JL, Solomon PJ. Mortality and other event rates: what do they tell us about performance? Critical Care and Resuscitation 2003:5;292-304.
13. Holt AW, Bury LK, Bersten AD, Skowronski GA, Vedig AE. Crit Care Med 1992:20;1688-1691.
14. McHugh GJ. Quality and reliability of data collected in a regional hospital intensive care unit. Critical Care and Resuscitation 2004:6;258-260.
15. The Victorian Intensive Care Data Review Committee. Intensive care for adults in Victorian public hospitals 2002: Report to the public.www.health.vic.gov.au/ criticalcare/
16. Australian Council of Safety and Quality in Health Care. 10 tips for safer health care. www.safetyandquality.org/.