Intensive Care Med (2010) 36:4–7
DOI 10.1007/s00134-009-1712-0

EDITORIAL

Shane M. Tibby

Does PELOD measure organ dysfunction…and is organ function a valid surrogate for death?

Received: 15 October 2009
Accepted: 18 October 2009
Published online: 10 November 2009

This editorial refers to the article available at:
doi:10.1007/s00134-009-1489-1.

S. M. Tibby (✉)
Paediatric Intensive Care Unit, Evelina Children's Hospital,
Guy's and St Thomas' NHS Foundation Trust,
Westminster Bridge Road, London SE1 7EH, UK
e-mail: shane.tibby@gstt.nhs.uk

The last 2 decades have seen considerable improvements in the delivery of care to critically ill children, such that paediatric intensive care unit (PICU) mortality rates are now typically below 10%, and indeed often closer to 5% [1, 2]. Paradoxically, this may hinder evaluation of new therapies aimed at further improving survival, because the number of patients required for an adequately powered clinical trial with mortality as the endpoint may now be unfeasibly large. Several strategies exist that may allow for a reduction in study size without loss of power. These include: (1) screening out extremely low- and high-risk patients, (where treatment effects may be less pronounced), (2) employing stratified randomisation based on mortality risk at trial entry [3], and (3) estimating treatment effects after covariate adjustment [4]. Identifying a suitable screening/stratification tool poses several challenges. It must be reliable, relatively simple, calculable within a short time frame, and should not reflect local treatment preferences, such as choice of monitoring tool or a particular ancillary therapy. At face value, contemporary mortality risk tools appear to fulfil these criteria, although valid criticisms of this approach have been voiced [5].

An alternative is to choose a relevant endpoint with a higher incidence than mortality. This could be a composite endpoint, such as mortality and morbidity combined, or indeed a different endpoint altogether, such as quality of life. However, both approaches have shortcomings. Composite endpoints have been criticised as being prone to subjectivity and providing misleading impressions of the impact of a given treatment [6]. Current tools for measuring paediatric quality of life are crude, and the optimal time to measure this post PICU discharge is unknown (1 month, 1 year, etc.) [7]. Recovery may take time to plateau; conversely, some deficits (e.g. post-traumatic stress) may not be apparent initially after discharge [7]. Prolonging the measurement period to accommodate this creates other problems; it increases the likelihood of confounding and may have major cost implications for a trial.

A third option involves the use of a surrogate endpoint for mortality. A surrogate is defined as "a variable that provides an indirect measurement of effect in situations where direct measurement of clinical effect is not feasible" [8]. Organ failure/dysfunction has been postulated as a surrogate for mortality, because it appears to meet the majority of criteria needed for validity [8]. First, there is biological plausibility of a causal link between the two. Second, epidemiological studies have shown prognostic value of organ failure for mortality [9–11]. The final major criterion is that there should be evidence from clinical trials that treatment effects on the surrogate produce similar effects on the main outcome, in other words, evidence that a treatment that ameliorates organ failure also reduces mortality. To date, this crucial piece of evidence is lacking.

Qualitative definitions for organ failure exist [9, 12]. However, a quantitative score, which also incorporates varying degrees of organ dysfunction, is likely to increase the sensitivity of this surrogate for identifying the likelihood of the endpoint (mortality). Three paediatric organ failure scores have been developed: PRISM III Acute Physiology Score [13], the paediatric logistic organ dysfunction score (PELOD) [14] and P-MODS [15]. All are clinical scores, and all have shown a correlation between degree of organ dysfunction and risk of mortality. Of the three, PELOD appears the most appealing. Unlike P-MODS, PELOD was developed in multiple centres and has been validated across several countries [14, 16]. And unlike PRISM III APS, the details of how to actually calculate the score are in the public domain.

Within the publication of the 2003 Lancet paper that validated PELOD externally, the authors stated the score was now fit for use as a surrogate in clinical trials [16], and indeed this has since become the case. However, in 2006, the same authors highlighted an error in their earlier paper, namely that the PELOD score did not, in fact, calibrate as originally stated [17]. Discrimination and calibration are two vital aspects of goodness of fit in a prognostic score. Discrimination refers to how well the score diagnoses or predicts the endpoint per se, while calibration refers to the accuracy of risk prediction within probability bands (for example, if the score assigns a 20% risk of mortality to 100 patients, we would expect 20 of them to die). An organ failure score that is poorly calibrated cannot be used to track changes in the degree of organ dysfunction, and thus loses a key criterion for validity as a surrogate.

The lack of calibration for PELOD has now been confirmed in a second paper, published in this issue of the Journal [18]. Garcia evaluated PELOD across 1,476 admissions from two PICUs in Brazil and Argentina [18], showing remarkably similar results to those from the original PELOD validation in 2003 [16]. The score continues to discriminate excellently (area under curve 0.93 for Garcia vs. 0.91 for Leteurte), but calibrates poorly, with both studies demonstrating under-prediction of mortality in lower risk groups and over-prediction in higher risk patients [16, 18]. The validity of Garcia's results is corroborated by their case mix and standardised mortality ratio, which are similar to current United Kingdom PICUs [1], and the fact that PIM2 is well calibrated within their study population.

Why is PELOD so poorly calibrated? A likely reason was highlighted previously by Garcia and colleagues: namely that PELOD does not assign risk on a continuous scale [19]. A risk score typically estimates probability of death using a ratio (continuous) scale ranging from 0 to 100%; this can be calculated via a number of mechanisms, most commonly using a logit transformation. Provided at least one of the variables within the model is continuous, an almost infinite number of covariate patterns are possible, limited only by the precision of the measuring tool (for example, base excess is usually expressed to one decimal place). PELOD is a composite score, with a single coefficient. However, its scale is ordinal, containing 33 ranks ranging from 0 to 71. This means that, for example, a patient cannot be assigned a risk between 3.1 and 16.3% nor between 39.2 and 79.6%. Thus, for example, a patient with a "true" risk of 4% would be calculated as 16.3% by PELOD.

The limitation of this approach can be highlighted by examining the original development paper for PELOD [14]. Continuous variables were first categorised using the Fisher algorithm, then further amalgamated into fewer and fewer (and indeed coarser) categories using a combination of cluster analysis and repeated logistic regressions. This approach loses considerable information and hence accuracy [20], and is further compounded when the dataset is small (594 patients, 51 deaths). The lowest score on PELOD (score 1) was formed by assigning the same weight (odds ratio 1.4, coefficient 0.33) to the lowest category of dysfunction seen in four different organ systems, which demonstrated odds ratios ranging from 1.4 to 3.6 (coefficients 0.32 to 1.27, table 6 of the paper). Similarly, a PELOD score of 10 (odds ratio 32.5) represented the amalgamation of four organ systems exhibiting moderate dysfunction showing considerably different odds ratios of 8.5 to 74.4. This resulted in over 80% of the variability in PELOD being explained by two organ systems: cardiovascular and neurologic. Conversely, pulmonary haematologic and hepatic dysfunction each accounted for less than 5%. This is in direct contrast to the adult MODS score, where variability (measured by partial correlation coefficients) is more evenly distributed among organ systems [21].

One implication of using PELOD as a surrogate endpoint is loss of power. To illustrate this, imagine a hypothetical treatment that produces a relative decrease in organ failure severity and hence mortality risk of 20% across a PICU population. The relativity assumption means that the absolute risk reduction is greater for higher risk patients (for example, the treatment may produce an absolute drop in risk of 16% for a patient with a baseline risk of 80%, but only a 2% drop if the baseline is 10%). I have illustrated this using a hypothetical, "true" organ failure score based upon the PIM2-derived risk profiles of 7,500 patients within my own PICU (Fig. 1). If we assume that only patients with a baseline risk greater than 10% are entered into this trial, the histogram on the left shows the effect of the treatment on risk. An adequately powered trial (80%) requires 120 patients per arm, giving a p value of 0.002. If PELOD is used, the "true" risk for each patient will be rounded up to the next PELOD category (e.g. a patient with risk of 4% will be classed by PELOD as 16.3%). The effect of using PELOD on this study population is shown in the right figure; indeed, the therapy is now classed as showing non-significant benefit, with a p value of 0.10.

Is PELOD broken beyond repair? Probably not, but it does need major resuscitation. Where possible, ordinal variables (e.g. heart rate ranges) should be re-expressed as continuous, and their relationship to risk (log odds of

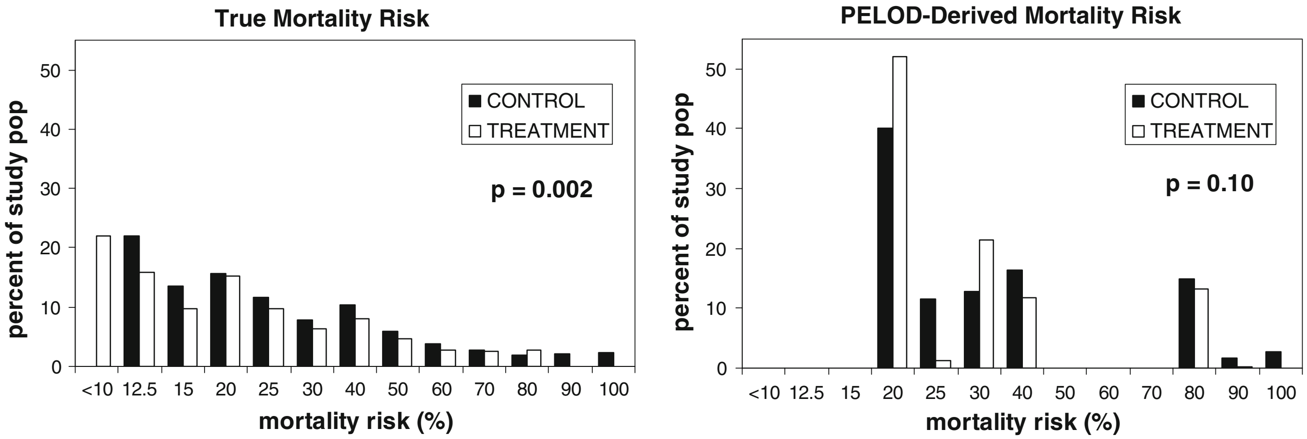

Fig. 1 Hypothetical trial for a treatment producing an average relative risk reduction of 20% as measured by true mortality risk (*left*) and PELOD (*right*). The reduction in risk is shown by a left shift in the histograms associated with treatment (*white bars*)

mortality) should be delineated using a technique that allows for non-linearity both in the univariate and multivariable modelling process (e.g. multivariable fractional polynomials or cubic splines). Alternative data reduction techniques should be considered (as per the cardiovascular variable in the MODS score) [21]. Therapies that are required to maintain a variable within the normal range may need to be incorporated within the score (e.g. inotrope dose, as in the SOFA score) [22]. These represent but a few suggestions; others can be found elsewhere [23, 24]. Lastly, this would require a much larger data set than previously [14].

Perhaps the most important point is that which goes beyond PELOD, or indeed any other organ failure score: we have not yet fully validated organ failure as a surrogate for death. Validation cannot rely purely on demonstrating a tight correlation between organ failure and death [25], nor upon fulfilling statistical criteria based on conditional distributions [26]. The demonstration of a treatment effect on the surrogate alone is insufficient; evidence must exist that this translates into a similar effect on mortality. A disconnect between the two has been demonstrated in numerous trials, where a beneficial effect on the surrogate translated into a harmful effect on the true endpoint [27]. This may also have relevance in the setting of organ failure, as it has been suggested that this entity may actually be protective, and hence attempting to reverse it may bring harm [28]. Regardless of this, the way forward requires (1) creation of a valid organ failure score, followed by (2) testing the score in the setting of a trial powered for a mortality endpoint. I think we may be waiting some time.

Conflict of interest statement The author has not disclosed any potential conflicts of interest.

References

1. Paediatric Intensive Care Audit Network 2008 Fifth PICANet National Report. Accessed Oct 05 2009 at http://www.picanet.org.uk/documentation.html
2. Slater A, Shann F, ANZICS Paediatric Study Group (2004) The suitability of the pediatric index of mortality (PIM), PIM2, the pediatric risk of mortality (PRISM), and PRISM III for monitoring the quality of pediatric intensive care in Australia and New Zealand. Pediatr Crit Care Med 5:447–454
3. Festa MS, Tibby SM, Taylor D, Durward A, Habibi P, Murdoch IA (2005) Early application of generic mortality risk scores in presumed meningococcal disease. Pediatr Crit Care Med 6:9–13
4. Roozenbeek B, Maas AI, Lingsma HF, Butcher I, Lu J, Marmarou A, McHugh GS, Weir J, Murray GD, Steyerberg EW, for the IMPACT Study Group (2009) Baseline characteristics and statistical power in randomized controlled trials: Selection, prognostic targeting, or covariate adjustment? Crit Care Med Aug 24 (Epub ahead of print)
5. Vincent JL, Opal SM, Marshall JC (2009) Ten reasons why we should NOT use severity scores as entry criteria for clinical trials or in our treatment decisions. Crit Care Med Aug 27 (Epub ahead of print)
6. Ferreira-González I, Busse JW, Heels-Ansdell D et al (2007) Problems with use of composite end points in cardiovascular trials: systematic review of randomised controlled trials. BMJ 334:786–793
7. Colville G (2008) The psychologic impact on children of admission to intensive care. Pediatr Clin North Am 55:605–616

8. No authors listed (1999) ICH Harmonised Tripartite Guideline. Statistical principles for clinical trials. International Conference on Harmonisation E9 Expert Working Group. Stat Med 18:1905–1942
9. Wilkinson JD, Pollack MM, Ruttimann UE, Glass NL, Yeh TS (1986) Outcome of pediatric patients with multiple organ system failure. Crit Care Med 14:271–274
10. Proulx F, Gauthier M, Nadeau D, Lacroix J, Farrell CA (1994) Timing and predictors of death in pediatric patients with multiple organ system failure. Crit Care Med 22:1025–1031
11. Tantaleán JA, León RJ, Santos AA, Sánchez E (2003) Multiple organ dysfunction syndrome in children. Pediatr Crit Care Med 4:181–185
12. Goldstein B, Giroir B, Randolph A, International Consensus Conference on Pediatric Sepsis (2005) International pediatric sepsis consensus conference: definitions for sepsis and organ dysfunction in pediatrics. Pediatr Crit Care Med 6:2–8
13. Pollack MM, Patel KM, Ruttimann UE (1997) The pediatric risk of mortality III—acute physiology score (PRISM III-APS): a method of assessing physiologic instability for pediatric intensive care unit patients. J Pediatr 131:575–581
14. Leteurtre S, Martinot A, Duhamel A, Gauvin F, Grandbastien B, Nam TV, Proulx F, Lacroix J, Leclerc F (1999) Development of a pediatric multiple organ dysfunction score: use of two strategies. Med Decis Making 19:399–410
15. Graciano AL, Balko JA, Rahn DS, Ahmad N, Giroir BP (2005) The pediatric multiple organ dysfunction score (P-MODS): development and validation of an objective scale to measure the severity of multiple organ dysfunction in critically ill children. Crit Care Med 33:1484–1491
16. Leteurtre S, Martinot A, Duhamel A, Proulx F, Grandbastien B, Cotting J, Gottesman R, Joffe A, Pfenninger J, Hubert P, Lacroix J, Leclerc F (2003) Validation of the paediatric logistic organ dysfunction (PELOD) score: prospective, observational, multicentre study. Lancet 362:192–197
17. Leteurtre S, Duhamel A, Grandbastien B, Lacroix J, Leclerc F (2006) Paediatric logistic organ dysfunction (PELOD) score. Lancet 367:897
18. Garcia PC, Eulmesekian P, Branco RG, Perez A, Sffogia A, Olivero L, Piva JP, Tasker RC (2009) External validation of the paediatric logistic organ dysfunction score. Intensive Care Med. doi:10.1007/s00134-009-1489-1
19. Garcia PC, Eulmesekian P, Sffogia A, Perez A, Branco RG, Piva JP, Tasker RC (2006) Limitation in paediatric logistic organ dysfunction score. Lancet 368:1151
20. Royston P, Altman DG, Sauerbrei W (2006) Dichotomizing continuous predictors in multiple regression: a bad idea. Stat Med 25:127–141
21. Marshall JC, Cook DJ, Christou NV, Bernard GR, Sprung CL, Sibbald WJ (1995) Multiple organ dysfunction score: a reliable descriptor of a complex clinical outcome. Crit Care Med 23:1638–1652
22. Vincent JL, Moreno R, Takala J et al (1996) The SOFA (sepsis-related organ failure assessment) score to describe organ dysfunction/failure. Intensive Care Med 22:707–710
23. Steyerberg E (2009) Clinical prediction models: a practical approach to development, validation, and updating. Springer, Germany
24. Royston P, Sauerbrei W (2008) Multivariable model-building. Wiley, Chichester, England
25. Baker SG, Kramer BS (2003) A perfect correlate does not a surrogate make. BMC Med Res Methodol 3:16–21
26. Berger VW (2004) Does the Prentice criterion validate surrogate endpoints? Stat Med 23:1571–1578
27. Shi Q, Sargent DJ (2009) Meta-analysis for the evaluation of surrogate endpoints in cancer clinical trials. Int J Clin Oncol 14:102–111
28. Singer M, De Santis V, Vitale D, Jeffcoate W (2004) Multiorgan failure is an adaptive, endocrine-mediated, metabolic response to overwhelming systemic inflammation. Lancet 364:545–548