Reviewer's report

Title: An evaluation of POSSUM and P-POSSUM scoring in predicting post-operative mortality in a level 1 critical care setting

Version: 1  Date: 30 July 2014

Reviewer: Alex Bottle

Reviewer's report:

Major comments

I was asked to assess the statistics.

This study compared these two common perioperative risk scores in level 1 ward patients and concluded that calibration was poor. This is hardly surprising given that the scores were not developed for that level of acuity. Also, as the Intro points out, “although POSSUM and P-POSSUM scoring systems have been validated for a number of surgical specialties, they are now 23 and 16 years old respectively and may not accurately reflect the risk faced by the today’s surgical patient.” Also, of course, if your unit has very good results in terms of mortality, then even a well calibrated model will look poor. The various limitations of a single-institution study should be pointed out.

Results: “Given this poor fit, logistic regression analysis was performed on the data set using IBM SPSS Statistics 19 (IBM, New York, USA). From this the following predictor equation was derived for mortality employing the originally recorded operative and physiological scores and urgency of surgery, and termed S-POSSUM.” You need to explain how this regression model was developed and where the coefficients come from – and present the coefficients in a table. Did you split into the usual training and testing data set portions, for instance? I suspect that a modest sample size (just 88 deaths) at a single institution will yield estimates with a fair amount of sampling error. You cannot propose the scores in your formula for use elsewhere. You make a lot of the “large” size of the data set. It may be large compared with some other studies, but in fact 88 deaths aren’t many, and it’s the number of deaths that really drives the power, not the denominator.

Minor comments

In the Intro, it would be good to say why people felt the need to develop P-POSSUM.

How many level 1 patients were excluded due to missing data? Might they differ in some important ways from those included?

P8: ROC curves don’t necessarily “compare the test against a gold standard” – in studies like yours they compare against the actual outcome (death).
Tables 2 and 3: I think columns 4 and 5 could go to save space. With Fig 2, you don’t really need these tables at all. Table 4 could probably go entirely – just comment in the text on the fact that the results are essentially the same as with the usual H-L bands.

Discussion: “In our analysis however the lowest bands of risk showed as good, if not better, model fit as higher risk bands, an important observation given that 92% of individuals within this study had P-POSSUM predicted mortalities of 20% or less.” This may be true, but I don’t think it’s at all important given that the whole model needs recalibrating anyway as it’s not currently fit for level 1 use.

Discussion: “Discrepancy between observed to expected mortality amongst individual studies is however large.” Not surprising if studies are often single-institution affairs.

Discussion: “Secondly, although data was collected prospectively by the admitting doctor it was analyzed retrospectively.” This isn’t a limitation. Analysis has to be “retrospective”.

Punctuation needs work throughout, esp around “however”.

Level of interest: An article whose findings are important to those with closely related research interests

Quality of written English: Needs some language corrections before being published

Statistical review: Yes, and I have assessed the statistics in my report.

Declaration of competing interests:

'I declare that I have no competing interests’