Reviewer's report

Title: Can American College of Radiology In-training Examination Scores be used to Predict Canadian Radiology Licensing Examination Results? A retrospective study.

Version: 1 Date: 24 November 2012

Reviewer: Chris McManus

Reviewer's report:

This paper has a very narrow focus, asking a highly specific question, and it will be of interest to those with the same interests as the authors. With a little extra effort though it might be able to provide information on other issues which are of somewhat broader interest.

Major compulsory revisions

1. For readers who are not Canadian radiologists, or even radiologists or from North America, it might be useful to provide a clear timeline of where these various assessments fit within the careers of the Canadian radiologists.

2. Much of the problem with these data is not the fault of the authors but the fact that their only outcome measure is a binary pass-fail, with most doctors passing. That makes any calculation of correlations difficult. The authors do well to use instead ROC curves, which do address the problem. However it might be that they need to explain why they are not using continuous scores for outcome (and I presume the College of Radiology will not provide them?). Either way, something should be said.

3. The correlations provided with binary outcome will be heavily constrained. Strictly what is needed here is a proper biserial correlation (i.e. equivalent in some ways to a tetrachoric correlation, and specifically not the point-biserial correlation which is the Pearson correlation), where the assumption is that a bivariate distribution has been dichotomised along one dimension. If the authors could provide that then it would be useful. Calculation is straightforward (and can be found in Howell’s Statistical Methods for Psychology). I have done it for PGY2 to PGY5 and Average and get values of .640, .633, .562, .664 and .692. These are noticeably higher than the Pearson R values and can be interpreted as the correlations if the College of Anaesthetists had provided full marks, rather than merely pass and fail.

4. A key question underlying these data is the extent to which behaviour in these residents remains constant across time (in other words, does performance ‘track’, so that poor performance at one time is related to poor performance at another). For that it would be helpful to have the correlations of the (continuous) scores available for PGY2, PGY3, PGY4, and PGY5. A clear prediction would be that they will be similar to the biserial correlations described above. If so then
there is strong stability during the course.

Minor corrections

5. P.8. “decreasing the skew from one inconsistent ACR score”. I'm not really sure that it is skew here. Essentially the error terms are being averaged out whereas the signal stays constant, and hence, via the Central Limit theorem the average is both more accurate and more likely to be normally distributed.

6. Table 1. I had presumed that the values after the plus/minus are standard errors (or are they 95% CIs?). However they are remarkably similar when N is 50 or 8, suggesting either that the SDs in the groups are very different or that these are SDs. Either way, a) it should be stated, and b) Standard Deviations/SEMs would be useful as well.

Level of interest: An article of importance in its field

Quality of written English: Acceptable

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

Declaration of competing interests:

I declare that I have no competing interests