Reviewer's report

Title: Fitness to Practice sanctions in UK doctors are predicted by poor performance at MRCGP and MRCP(UK) assessments: data linkage study

Version: 0 Date: 03 Aug 2018

Reviewer: Paul Tiffin

Reviewer's report:

This is a well written and interesting paper which provides evidence for the validity of postgraduate medical exams. In general my comments are fairly minor recommendations to improve the manuscript. The statistical approaches were generally as robust as possible, given the limitations of the data. I particularly liked the approach to trying to establish directionality of the association between the exam performances and fitness to practice risk.

More substantial points:

1. The paper repeatedly makes case that the relationship between performance in the exams and the risk of fitness to practice issues later is linear. However, unless I am missing something fundamental here, my reading of the results and the graphs in figure 1, suggest that the relationship is actually log-linear. That is, the graphs in figure 1 have a logarithmic scale. Therefore if the examination performance is plotted against the odds ratios related to the risk of fitness to practice, rather than the log-odds, presumably the graphs would be curvilinear rather than linear. This suggests that the risk of fitness practice does not increase linearly. Furthermore, a technical point would be that the examination performance, once concerted into Z scores, is not strictly an interval metric, though for these purposes it can properly be treated crudely as such, though strictly speaking scores would have to be treated as unidimensional and Rasch calibrated in order to produce units in a common metric. It would be interesting to see what the relationship between the examination performance was like plotted against the odds ratios, rather than log odds, as this may reveal certain threshold effects in this exam scores.

2. The paper also states that "This study is the first to demonstrate that clinical assessments, as opposed to knowledge assessments, are particular important for predicting FTP assessments...". This is not an accurate statement as our previous study, indeed cited in this report, compared the relative impact of the knowledge-based component of the PLAB part I (knowledge test), and the skills-based part two of the PLAB test. Indeed we state in the discussion section of our paper (Tiffin et al. 2016): "Performance at PLAB part 1 was also a relatively weak predictor of progression from referral to censure; whilst multiple sittings of that part of the test were associated with an increased risk of progression, the actual score at first attempt was not. Rather, it was achievement at part 2
of the PLAB that was more closely associated with the risk of a doctor progressing from referral to censure, along with male sex. This implies that various aspects of performance can increase the risk of being caught up in the rather broad and non-specific 'trawler-net' of the FtP referral process but that, once within this process, relatively few of these variables predict progression to censure. The weak association with PLAB part 1 performance suggests that it is not largely lack of semantic medical knowledge that is associated with the risk of eventual censure for those investigated by the GMC. In contrast, as outlined earlier, PLAB part 2 evaluated procedural skills, which will include ratings of interactions with role played patients. Thus, it is likely to be the capturing of these softer skills by PLAB part 2 that explains this association. Similarly, it is male sex and PLAB part 2 performance that are two of the three statistically significant predictors retained the multivariable models built with censure as the outcome." Thus our findings closely mirrored those reported in the current paper.

More minor points:

3. There is a comma missing from the sentence on page 4 "If medical examinations are indeed worthwhile than, as with all medical systems,…"

4. Again, the authors claim that "this study therefore differs from previous work in emphasising postgraduate examinations, and comparing the value of knowledge and clinical assessments..." Please see my point 2 above.

5. I agree that conversion of the scores (relative to the pass mark) to Z scores is the most sensible way of dealing with these. However it should be acknowledged as a limitation that this does not necessarily imply that the scores can be fully equated (the authors do not imply this but it should still be highlighted). Rather Z scores represents the relative performance to that particular cohort specific to the exam sitting. Whilst I think this is probably the best way of approaching this issue, it is still a limitation, especially when comparing different examination types, and should be mentioned as such.

6. I think it was reasonable to control for the decades since qualification, and in UK practice. However, I did wonder about the granularity of using decades (although the coefficients recovered are more easily interpreted, being in a larger metric). In particular I've noticed some non-linearity the relationship between the risk of fitness practice even within the first 10 years of practice. Although the results tend to be similar alternative approach would be to use a survival analysis approach, controlling for exposure time. This is the approach we took in our previous study, though, as we acknowledge in the paper at the time, this rests on the proportional hazards assumption.
7. In presenting the results it would probably be more conventional to merely present the odds ratios and their 95% confidence intervals, rather than the log odds and standard errors, which are not easily interpretable. Some additional interpretation, for the less statistically erudite reader, would also seem appropriate in the results section.

8. I applaud the use of multiple imputation in this case, there would point out that it is usually conventional, in epidemiological research, to present the results on non-imputed data, treating imputed results as a sensitivity analysis for the 'missing at random' assumption for missing values. However, in this case it probably did make sense to present the results from the imputed datasets, though it would also probably be appropriate, possibly in a supplementary appendix, to briefly present the main results (in terms of odds ratios and associated confidence intervals) from the non-imputed data. Also, although is a minor point, I could not see any justification for the use of m=100 imputations. Usually I would tend to increase the number of imputations, in an exploratory fashion, until the recombined results stabilised. Traditionally relatively a small number of imputations (often m=5 to 10- though this may have been partly as a result of earlier limitations to computing power) have been used in epidemiological research although I find that in practice the recombined results usually stabilise at around 30 to 50 imputations. This may save the authors some computational time in the future.

9. For consistency results should probably be presented to 2 decimal places, rather than three.

10. It would be worth making some comment in the discussion section about the limitations of using the attempts at first sitting, as this is not usually the result that decisions by selectors are made on. However, in this case I agree with the authors that it made sense to use the first, rather and subsequent attempts at examinations.

11. Given that the relationship between the exam scores and the odds ratios for a fitness to practice issue are probably curvilinear, with this give rise to any particular thresholds for scores that should be used in selection and regulation policy? This is an issue that should be discussed in a revised discussion section.

12. In terms of tables 1, 2 and 3: they would probably be clearer if they merely presented the odds ratios and 95% confidence intervals. The p values, which are all the same, at less than .001 do not add information and should be omitted, although a footnote could state that all these coefficients were significant at the p less than .001 level.
13. In terms of figure 1, it would be useful to also see the graphs of the exponentiated odds ratios, to illustrate what I suspect is the non-linear nature of the relationship between the exam scores and the odds of a fitness to practice issue.

14. In terms of figure 2, whilst I generally find signal theory a very useful contribution in these contexts, the ROC plots themselves do not add much information beyond the information on the AUC values already given in the text and would be inclined to omit them.

I enjoyed reading this paper very much and hope that the authors find these comments helpful in improving the manuscript.

Are the methods appropriate and well described?
If not, please specify what is required in your comments to the authors.

Yes

Does the work include the necessary controls?
If not, please specify which controls are required in your comments to the authors.

Yes

Are the conclusions drawn adequately supported by the data shown?
If not, please explain in your comments to the authors.

No

Are you able to assess any statistics in the manuscript or would you recommend an additional statistical review?
If an additional statistical review is recommended, please specify what aspects require further assessment in your comments to the editors.

I am able to assess the statistics

Quality of written English
Please indicate the quality of language in the manuscript:

Acceptable
Declaration of competing interests
Please complete a declaration of competing interests, considering the following questions:

1. Have you in the past five years received reimbursements, fees, funding, or salary from an organisation that may in any way gain or lose financially from the publication of this manuscript, either now or in the future?

2. Do you hold any stocks or shares in an organisation that may in any way gain or lose financially from the publication of this manuscript, either now or in the future?

3. Do you hold or are you currently applying for any patents relating to the content of the manuscript?

4. Have you received reimbursements, fees, funding, or salary from an organization that holds or has applied for patents relating to the content of the manuscript?

5. Do you have any other financial competing interests?

6. Do you have any non-financial competing interests in relation to this paper?

If you can answer no to all of the above, write 'I declare that I have no competing interests' below. If your reply is yes to any, please give details below.

Researcher in similar area
Personal knowledge of several of the authors through my research
Lead author attended a short course I ran last year

Statement on potential review bias
Please complete a statement on potential review bias, considering the following questions:

1. Did you co-author any publication with an author of this manuscript in the last 5 years?

2. Are you currently or recently affiliated at the same institution as an author of this manuscript?

If you can answer no to all of the above, write 'I declare that I did not publish with these authors in the last 5 years and also meet the affiliation criteria”. If your reply is yes to any, please give details below.

1. No
2. No

I agree to the open peer review policy of the journal. I understand that my name will be included on my report to the authors and, if the manuscript is accepted for publication, my named report including any attachments I upload will be posted on the website along with the authors' responses. I agree for my report to be made available under an Open Access Creative Commons CC-BY license (http://creativecommons.org/licenses/by/4.0/). I understand that any comments
which I do not wish to be included in my named report can be included as confidential comments to the editors, which will not be published.

I agree to the open peer review policy of the journal