Author's response to reviews

Title: A method for measuring productivity and encouraging the practice of research in hospitals: development and feasibility

Authors:

Caterina Caminiti (ccaminiti@ao.pr.it)
Elisa Iezzi (eiezzi@ao.pr.it)
Caterina Ghetti (cghetti@ao.pr.it)
Gianluigi De' Angelis (gianluigi.deangelis@unipr.it)
Carlo Ferrari (ferraric@ao.pr.it)

Version: 3 Date: 4 August 2015

Author's response to reviews: see over
This is an interesting paper about the measurement of research productivity in Italy, and I hope it will be published. In my view neither Compulsory nor Essential Revisions are required, e.g. to correct factual errors. However, I think the paper would be improved if the authors were to consider a number of Discretionary Revisions, as follows.

1. The title of the paper states that it describes a method for "encouraging the practice of research in hospitals". However, the paper includes almost nothing about this. I agree that a method for measuring research productivity is in general terms a Good Thing. It might also have the effect of encouraging subsequent research activity, but that is entirely speculative and remains to be demonstrated.

   We agree, and have modified the title to avoid drawing conclusions that have yet to be demonstrated

2. The abstract states (line 33) that a literature analysis was conducted to determine how best to assess research performance. However, the Methods don’t really explain how the analysis was done.

   We have added details on the research strategy used to identify recent, relevant studies. We point out, however, that our work is not intended to be a review.

3. The abstract also states (line 38) that a set of 12 indicators was selected. Again, the Methods don’t really explain how these 12 indicators were chosen (i.e. on what basis), and what indicators were rejected.

   To clarify that the indicators were not identified using a selection process, we have replaced the verb "selected" with "used" in the abstract, and the expression "greatly influenced" with "determined" in the Results section. Moreover, we have listed the indicators for which information is retrievable from existing administrative sources, and those which are derived from the answers provided by HCPs in the survey. Finally, we have explained that activities self-reported by HCPs in the questionnaire had to be documented.

4. The abstract concludes (line 47) that "this work suggests that the proposed method may be useful to promote research ...". I think this is a reasonable hope, although a cynic might suggest that it will do nothing of the kind. However, in the absence of any evidence at all, it is perhaps stretching things to claim that the present work suggests that the method will promote research. Just to be clear, I am not criticising what the authors have done. But I think they should be careful not to over-interpret their results, lest they be accused of hyperbole. So in my opinion it would be better to water down the title a bit (point 1 above) and make the conclusion a bit less sweeping (point 4 above).

   We agree, and have "softened" the conclusions of the abstract.
5. At the end of the Methods, the aim of the work is stated: to measure ... research productivity. This is fine, but not all readers will have the same understanding of the term "research productivity" and that should probably be defined at this point.

*We have now explained, at the end of the Background section, what is meant by research productivity in this work, using the definition provided by the reviewer in his paper on the subject (n. 12 of our references).*

6. The Results describe the 12 indicators which were chosen for the present analysis, chosen partly because of their ready availability (which seems eminently sensible). The indicators chosen seem perfectly reasonable, but I wonder about their definition. For example, one indicator concerns teaching, the relevant question to the HCP being "Did you carry out teaching activities concerning research?" Perhaps Italian doctors are better disciplined than the ones I have worked with over the years, but I find it difficult to imagine a doctor in a teaching hospital who would answer anything except Yes to this, whatever the actual nature of their work. This difficulty is touched on in reference [12], which points out that there is a very large number of possible indicator variables, and that as their complexity increases, agreement about their definition becomes more and more important.

*We have improved the description of the indicator on teaching, specifying that it concerned teaching activities carried out by HCPs on their own research underway in the year in question. We have also clarified that the indicator on training concerned participation in training in research methodology. We believe the remaining indicators were clearly explained and thus have not made any changes to their description.*

7. The other aspect of the 12 chosen indicator variables is their relative value. These values (i.e. the points awarded per item) are listed in Table 1. For example, a grant of 24,000 Euro is worth 1 point; and a grant which was submitted, but not awarded, is worth 0.1 points. Being a member of a national guidelines committee is worth 0.1 points and acting as a peer-reviewer for a national grant application is also worth 0.1 points. Clearly there is endless scope for argument about this. The values chosen don't seem unreasonable, but I think that the authors could discuss how they arrived at them.

*We have specified in the Results section that the weighting system was constructed mainly considering the hypothesized effort, and for some indicators, also applying specific criteria.*

8. Having selected various indicators and assigned values to them, the authors have then calculated scores for 896 practitioners in six different departments. The results are summarised in Table 3. The authors comment (line 205) that in their sample, the highest score was for publications, followed by grant income and peer-reviewing. In other words, the measurement of research productivity here is heavily dependent on the publication element. This is perhaps easier to see in a pie chart-

What the authors don't discuss -- and I think they should -- is that the proportions in this pie chart are a simple consequence of the values awarded to the chosen indicators. Thus if different values are used for grant income (say 2 points per 24,000 Euro) or for peer reviewing activities (say 0.5 points per application), then a completely different pie chart results. Since there is no right answer here, the relative weights of the
12 chosen indicators are a matter for debate. As it stands, the present index of research productivity is 70% dependent on publication output, and 30% dependent on other factors, such as grant income, PhD supervision etc. This may make perfect sense in the context of Italian teaching hospitals, but other readers may question the dominance of publication output, and may feel that a more balanced scorecard would be preferable.

We have added a table showing raw data for each indicator. This has confirmed the observation made by the reviewer, whom we thank for pointing out the issue. This aspect has now been dealt with in the Discussion section.

9. In the Discussion, the authors comment that this is the first comprehensive approach to measuring research output which includes hidden research activities (line 230). This claim may be disputed by previous authors. For example, Mezrich et al (ref 8) used administrative and community service, and student ratings, in their scoring system. Iyengar et al (Acad Med 2009,84:1610–1616) used research space allocation in their scoring system.

We have now specified at the beginning of the Discussion section the hidden activities which distinguish our work from already published research. We thank the reviewer for this important observation, which has enabled us to better highlight some relevant aspects of our work.

10. Figures. It is interesting to see in Figs 1a-f that the research productivity for the six departments studied is skewed, as was found in the Norwegian study (ref 12). I'm not sure that more than one bar chart is necessary to make this point. It's not clear what is being plotted in Fig 2. Is this the frequency distribution for the 896 productivity scores? Do we know anything about the 3 high achievers identified in Fig 2? Can it be verified independently that these people are first-class researchers? That might strengthen the choice of scoring system.

We believe the inclusion of the 6 bar charts can be useful to depict the research productivity of the single Departments, highlight variability, and enable comparison with other institutions.

11. It would make the paper easier to read if the authors used a consistent terminology: research productivity, activity and output are used at various points. I believe these are all the same thing.

This is true. We have now clarified at the end of the Background section that the terms research productivity, research output and research performance are used interchangeably.
The authors have piloted an array of metrics to attempt to provide some quantification of research activity. They found that it seemed to have some discriminant validity (specialized medicine outperformed other areas). This is an increasingly important topic as funders (direct and indirect) seek value for money in those they support. The authors are to be commended for trying to integrate various metrics of time and output.

Major Compulsory Revisions

Essentially this is a paper describing the development of a measurement tool. It therefore falls into the domain of psychometrics – does the tool measure reliably some underlying construct that is important? I fully appreciate that with any new tool there is a balance between publishing early to let others see what is happening and use it but without extensive validation, and conversely producing a vast amount of robust internal and external validation but at the risk of never publishing anything. At the moment I feel the authors have erred slightly too far towards early publication. I would suggest that the authors take the principles of psychometrics / tool validation and apply (or at least discuss their application) to their tool. This might usefully be objective three.

The effect of this will be that the authors will include some (new) information about reliability, construct validity, face validity, discrimination etc. Some of this they already have but it is implied rather than up front. For instance, they have shown that their tool highlights specialized medicine as being research heavy – is this supported by other methods of assessment? (The question of whether this research bias is appropriate is for debate elsewhere).

We thank the reviewer for pointing out this relevant issue. In the Discussion section we had mentioned that scores had to be validated; we have now emphasized its importance and explained why it still has not been done.

The authors describe use of the JCR as a source for normalized impact factor. I have just been onto JCR and can’t find this; Scopus does something along these lines. So at present I am unable to replicate the study method.

This was our mistake. We have corrected Table 1 to clarify the misunderstanding, as impact factor values were derived from the JCR, while the Normalized Impact Factor was calculated according to the work by Owlia (ref. 17).

How much intra-individual variation is there? Grants and papers are episodic so any single year may be good or bad for any single individual.
We assume intra-individual variation is present. However, the aim of our work was to develop a method for measuring individual research productivity and to test its feasibility, by implementing it for one year, as specified in the Discussion section.

The authors describe a literature search. Can they please provide details.

This correct observation was also raised by reviewer 1. As stated above, we have added details on the research strategy used to identify recent, relevant studies, although this work is not intended to be a literature review.

Minor Essential Revisions

Line 189: ‘...indicators for each HCP and data analysis was about 9 weeks/person...’ There were 896 individuals – did it really take 8000 weeks to do this study?

We have modified the text to clarify.

The manuscript has been copyedited by Edanz Group Global Ltd