Author’s response to reviews

Title: A novel method for interrogating Receiver Operating Characteristic curves for assessing prognostic tests

Authors:

Grégoire Thomas (gregoire.thomas@squ4re.com)
Louise Kenny (L.Kenny@ucc.ie)
Philip Baker (philip.baker@le.ac.uk)
Robin Tuytten (robin.tuytten@metabolomicdiagnostics.com; rtuytten@gmail.com)

Version: 2 Date: 10 Oct 2017

Author’s response to reviews:

Point by Point Answers to referee reports DAPR-D-17-00009R1

Re: Revision manuscript DAPR-D-17-00009_R1.

Dear Sir/Madam,

Please find the revision of our manuscript enclosed.

We would like to genuinely thank the editor and reviewers for the two instructive and constructive rounds of review of our manuscript. The reviews received allowed us to further build our own understanding of the proposed tool and reflect on its usability beyond the test development phase.

The suggestions for improvement as provided by Reviewer #1 following the review of the revised manuscript (ref. DAPR-D-17-00009_R1) have been taken into account to the best of our abilities. We also took the opportunity to correct minor grammar issues and a few minor typing errors.

As part of this revision, please find

- the point-by-point answers to all the points raised in the review – see below.
- a manuscript file on which all the changes made are highlighted (“track changes”); - a clean manuscript file on which all the changes made were accepted
We hope the changes applied improve the readability of our manuscript and render our manuscript acceptable for publication in Diagnostic and Prognostic Research.

We look forward to receiving your decision,

With sincere regards,

Robin Tuytten, PhD

Reviewer reports:

Reviewer #1:

The authors have addressed all my previous comments satisfactory or mentioned good reasons why they haven't. However, they also changed the structure and text of the whole manuscript tremendously. Therefore, I have peer-reviewed the manuscript as if it was a new one.

The authors highly appreciate the fact that Reviewer #1 committed time and effort to perform an additional in-depth review of our revised manuscript.

I still think the whole manuscript could be improved on a more conceptual level, as sensitivity and specificity on their own are seldom used or reported for genuine prognostic tests; and for diagnostic tests, taking prevalence into account can be done without the nice graphs the authors show here; and researchers should realize that one cannot just combine sensitivity and specificity with any prevalence, as prevalence is linked to patient spectrum and therefore sensitivity and specificity do vary when prevalence varies. But maybe that would also make things more complicated. So I have just picked up on some details.

1. Page 10, lines 1+2: "low prevalence disease, high sensitivity and specificity can still be associated with (very) low post-test probabilities.". It is the first time this term is introduced in the paper and not really clear. Besides, this would only be true for the post-test probability of being diseased, not of being non-diseased. Why not only state PPV here?

A1. The suggestion of the Reviewer is applied. The text is adopted accordingly.
2. The last paragraph of the results section seems to fit better in the Discussion section.

A2. The suggestion of the Reviewer is applied. The text is adopted accordingly.

3. In line with my general comment, the statement on page 17, lines 1 to 3 ("equi-predictive value lines ... Partially - account for variation in patient spectrum") is not true in my opinion. As the authors explain further on, it only holds in the same setting and therefore probably in the same spectrum of patients. One just cannot apply a sensitivity and specificity or AUROC or LR or whatever measure found in a population with, say, a prevalence of 50% to a situation with for example a prevalence of 10%, as both populations will likely have different patient characteristics (the Willis and Hyde paper contains references to several papers around this topic). So to put it straight, I think the methods presented here do only hold if they are applied in a situation where the ROC curve was derived in a setting where the prevalence more or less matches the prevalence used to derive the equi-predictive value lines. I would like to see that addressed in the Discussion section.

A3. The authors agree with the reviewer that the use of predictive value criterions is confined to gauging the impact of prevalence on defining relevant areas in the ROC space when considering the same intended-use populations and the same clinical care setting. Yet, within this framework, there might still be residual population heterogeneity, leading to differences in disease prevalence. In was in this context we used the term spectrum effect. However, to avoid unnecessary confusion about terminology, we removed the statement on page 17, lines 1 to 3 altogether.

Instead reference to patient spectrum was added at the end of the paragraph (page 17, lines 14-18), and the wrong word usage ("equi-prevalence lines” instead of "equi-PPV (NPV) lines") was corrected, as per below:

“Evidently, a significant change in application setting, e.g., from secondary care to primary care, will have a more profound impact on case mix, being the distribution of outcomes and predictive factors. [30] Such change of application setting and patient spectrum will be outside the utility scope of equi-prevalence PPV (NPV) lines for gauging test performance.”

We feel that the above section now meets the request from the reviewer to highlight that the utility of equi-PPV (NPV) lines is restricted to situations of defined clinical care setting.

4. I am not sure if I understand or agree with the statement made in lines 18 to 24. The authors state that use of NPV and PPV does not require calibration and that calibrated scores and predictive values have different use. But then the requirement for calibration has nothing to
do with using equi-PV lines, right? And then the statement that no calibration is required is not helpful. Calibration is still required, as it serves a different purpose.

A4. The authors agree with the reviewer and to ensure that the context of the statement is clearly understood, we have adjusted the sentence mentioned as follows:

“It is interesting to note that by itself the definition of a cutoff using PPV and NPV does not require prior calibration of the model.”

Note that to avoid confusion, the manuscript also states the purpose and usefulness of calibration:

“The calibration ensures that the test score reflects the likelihood of a test to predict a patient’s chance to develop a condition. [31]”


A5: The terminology used followed a suggestion by the editor as part of the first review (see below quote). For that reason, we kept the term multivariable test

“Terminology. Please use the term "multivariable model" (e.g. discussion) when referring to combination of prognostic tests. The term "multivariate" is reserved for situations where multiple outcomes are being predicted”

6. Page 19, lines 1 to 5: I do not understand these sentences. Particularly the link between being more sensitive to extreme disease and enhancing the transportability across different health care settings. Please elaborate and explain.

A6: The section has been reworded to clarify this point better. The text now reads:

“Finally, we consider it conceivable that multivariable tests which are developed to comply with either the rule-in or the rule-out test will also be more generalisable. Using the web-tool, it was found that models which comply with a (stringent) PPV criterion are characterised by a fraction of cases which are very well discriminated (following a tight risk score distribution in controls). Vice versa, models which comply with a (stringent) NPV criterion are characterised by a fraction of controls which are very well discriminated (following a tight risk score distribution in the cases). Provisionally this is not a result of mere overfitting or patient spectrum, it may well be the predictors constituting a good rule-in model are more directly associated with the pathophysiology of the condition (e.g., pre-eclampsia) or its severity. Likewise, a good rule-out model might constitute predictors which are strong determinants of non-disease (or health). If so
and arguably, the methods presented here may enhance the transportability of such models across different healthcare and demographic settings. Validation of dedicated rule-in or rule-out models will need to be done to confirm this hypothesis. Of note, we applied a rule-in criterion (PPV ≥ 0.20; Sn ≥ 0.50) to develop a biomarker based prognostic model for pre-eclampsia in low risk nulliparous once before; in that instance we were able to validate the model as developed in a cohort of New Zealand and Australian women in an European patient population.[1]”

Reviewer #2: The authors have dealt with my comments adequately