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: 1 Date: 06 Aug 2017

Author’s response to reviews:

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

Reviewer reports:

Associate editor: I have the following remarks:

1. In the introduction, I would simply focus on medical tests, as the difference between different types (e.g. screening, diagnostically, prognostic) can be rather ambiguous (and is not defined by whether or not they yield an absolute probability of disease presence or occurrence).

A1. We concur with the editor and the reviewers that our use of inconsistent terminology when referring to test types as well as the overbroad use of the concept of clinical utility caused unnecessary confusion and text inaccuracy. We would like to thank the editor and reviewers to highlight these terminology flaws and apologize for this.

To address this issue, we have carefully revised our wording to ensure consistency and in-line with these comments (editor comments: 1,2; Reviewer #1 comments: 2, general remark; Reviewer #3:points 4,5) and have removed references to “screening tests” and “diagnostics tests” throughout, and hence restricted the scope of the manuscript largely to discussing “prognostic tests”.

In line with this we also applied the following additional changes,

- we proposed a novel title which addresses the editors point 1 &2 by not referring to clinical utility. ROC is also in full in the title (Reviewer #1: point 1)
- we removed the background paragraphs referring to the Wilson and Jungner criteria, and their contemporary interpretation, for accepting novel screening tests.

- we moved language discussing the strengths and limitations of AUROC from the methods to the Background section

- we removed the paragraphs discussing AUROC in the context diagnostic tests

- we removed confusing language on differences between diagnostic and prognostic tests

In conjunction with the other changes made as per the editor and reviewers suggestions (see below), we hope the changes applied improve the readability of our manuscript and are acceptable to the editor.

2. Please refrain from using the term "utility" and "cost benefit" (which have for instance specific meaning in HTA, where actual costs become part of the equation). The authors are, in fact, referring to the predictive performance of the test (e.g. diagnostic accuracy), as measured by the AUROC in this case.

A2: Changes were made accordingly (see also A1); where appropriate we refer to prognostic test performance, relevance in a clinic context, or similar

3. As raised by the reviewers, sensitivity ans specificity of a test can actually vary by disease prevalence. It would be helpful to postulate the independence assumption (which is correct in theory), and come back to this in the discussion.

A3: Reference to patient spectrum effect is introduced in the background when introducing the strengths and limitations of AUROC; we also expanded the language on some of the other limitations a bit more. The role of patient spectrum and its relation to predictive values is also elaborated upon in the Discussion.

4. Please adjust the AUROC definition to better reflect its theoretical properties. I'm rather confused by the current description referring to `signal` and `noise`. Also, the essence of the AUROC relates to discrimination (i.e. ability to distinguish individuals who develop and do not develop the event), rather than identification (which requires some appropriate threshold as well).
A4: We added reference to discrimination and updated the definition for AUROC in accordance with the one employed by Usher-Smith et al in BMJ 2016;353:i3139; the sentence now reads: “The AUROC, essentially a measure of discrimination, corresponds to the probability that a classifier will correctly rank a randomly chosen person with the condition higher than a randomly chosen person without the condition”

5. In the methods, the authors raise the essence of calibration performance. This is indeed very important, and not merely relevant for tests used for prognostic purposes. Again, I would therefore refrain from splitting up the different types of medical tests. The authors could, however, indicate that sometimes (relatively often in diagnosis), calibration performance cannot be assessed because the test may only present a dichotomized value. This of course raises the discussion whether test manufacturers should continue to do so (especially when tests are not used in isolation, and their results do not directly affect decision making).

A5: We cannot formulate an advise on whether a test manufacturers should apply calibrations or not. We do observe that calibrated scores and predicted values have different usage (see statement below). The use of either method depends on the clinical context.

The calibration ensures that the test score reflects the likelihood of a patient to develop a condition (Steyerberg 2009). The predictive values gives the likelihood that a subset of selected patients develops (or not for the PPV) a condition.

6. In the discussion, please refer to other calibration performance measures (i.e. beyond PPV and NPV) or tools. For prediction models it is in fact recommended to provide at least calibration plots, which appear equally relevant when dealing with individual tests (based on the arguments raised in this article).

A6: There are indeed other methods to transform the score from a diagnostic test. We have added a paragraph mentioning calibration methods and a reference that reviews them. We do acknowledge the usefulness of calibration methods.

Of course this depends on knowing the prevalence and deciding upon an acceptable criteria. What the authors fail to do in their analysis is to take into account the uncertainty in the estimates for the prevalence which are unlikely to be known accurately as assumed here. This would undoubtedly widen the permissible regions. The paper would benefit from including such an analysis.
7. Notation. Please redefine Sn and Sp (e.g. using subscript) to avoid ambiguity with the parameters "n" and "p". Also, please be consistent (e.g. in the figures the term "Sens" and "Prev" are used instead of "Sn" and "p" respectively). Further, please move equation labels (A), (B), (C), ..., (G) to the right, as they now seem to be part of the equation.

A7: We implemented Sn and Sp throughout, resolved the ambiguity in the Figures and moved the equation labels.

8. 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.

A8: Thank you for highlighting this terminology ambiguity; in accordance with this we implemented “multivariable” throughout.

Reviewer #1: The authors present a novel and interesting method for assessing clinical utility of diagnostic tests. In itself, the method is not very difficult and the reasoning behind it makes sense. So if we address this study as a demonstration of a method, then the manuscript could either have been much shorter or some more application examples could have been provided.

On a different level, the manuscript raises a lot of questions about underlying mechanisms (do sensitivity, AUC and specificity indeed remain stable when the prevalence changes) and terminology (is a screening test the same as a prognostic test). I have tried to address all these concerns as well, but I am not sure if they should be treated as real concerns.

1. Title: I would have liked ROC in full in the title (Receiver Operating Characteristic curve) and probably also "of a prognostic test" at the end.

A9: The suggested changes to the title were made; see also A1.

2. Clinical utility is a term that is not well defined. Both CDC and Bossuyt et al (Clin Chem 2012) define it as the likelihood of or a measure for improved outcomes for patients when the test has been used. I do not think predictive values cover such a definition, so maybe another term (for example predictive value?) is more helpful. In any case, the authors should provide a definition of the term utility if they use it.
A10: Thank you for pointing out this terminology issue; all references to clinical utility were removed; see also A1.

3. The authors claim that this approach is novel. Which is surprising, given its relative simplicity. But after checking the literature, I couldn't find any similar articles either (which is not a guarantee of course). I did find a paper that looks remotely the same: Mandic et al, J Cardiopulm Rehab and Prev 2008; 28(6): 415-19. Maybe it is worthwhile to have a look at it and discuss it in this manuscript.

A11: At the time, we first established this method for plotting equi-predictive value criterions, we were also struck by its simplicity. Like the reviewer we searched for articles, but couldn’t find any reference to this method As the method helped us to get a better understanding on what type(s) of prognostic models we should develop (rather than “can develop”), we felt this might be helpful to other researchers also; Hence this manuscript. Thank you for pointing us to the Mandic paper, as it is a demonstration that the ROC curve, and the AUROC can be quite indifferent to underlying prevalence, it supports the notion that case-control study design is an acceptable approach when cases and controls are representative for the intended use population. We added a reference to the Mandic paper.

4. The authors start with the Jungner and Wilson criteria for screening tests. I am not sure if the claim really can be made that these criteria functioned as gatekeepers to the introduction of novel tests. Any test can be brought on the market without rigorous regulation. If I remember correctly (but I haven't checked), then these criteria are mainly useful to decide whether a screening test may indeed be useful to be used in screening programs. I think removing the part "and acted as gatekeepers to the introduction of novel tests" should suffice.

A12: To prevent terminology confusion with regards to screening vs prognosis, we removed the introductory section on Wilson and Jungner altogether; see also A1.

5. In line with that, I think the authors mix up screening and prognosis. I know that these terms as well are used in different ways, so perhaps it would be good here as well to define what the authors exactly mean with screening test or prognostic test. To me, a screening test detects disease that is already there, but perhaps in an early stage. While a prognostic test to me predicts whether something will occur in the future. I know that some readers may think the same, but I also see these terms being used interchangeably. So the more reason to use clear terminology, to provide definitions and to be consistent in terminology.
A13: The reviewer is correct in the fundamental distinction between screening and prognosis and the manuscript has been adopted accordingly (see also A1).

6. Page 6, line 2: first write ROC in full.

A14: AUROC as well as ROC were written in full at its first appearance

7. On page 7, the authors state that a screening test is fundamentally different from a diagnostic test in that sense that screening tests can only estimate a probability. I disagree with this statement, as I think that a screening test is not necessarily different from a diagnostic test (but a real prognostic test may be different); and because I think that a diagnostic test in itself is rarely capable of confirmation presence or absence of disease.

A15: To focus the manuscript on the presentation of the novel method, the scope of the manuscript has been narrowed to prognostic tests; see also A1.

8. Page 7, lines 10+11: I do not understand the last part of this sentence ("these statistics therefore…"). Maybe it can be removed? It seems a bit out of place and irrelevant here.

A16: the paragraph on diagnostic tests is removed in its entirety, refer to A1.


A17: This sentence has been removed as it did not really add.

10. Equations A and B should be fairly well known by the audience of this journal. So this can be shortened or even removed. Same may be true for C and D, but I see that these are useful to understand the later equations E, F, G.

A18: The reviewer is correct that for the typical journal audience equations A,B and C are trivial. We felt it was however justified to keep the equations in the event this method would find traction in a larger audience (which we would hope)

11. Part of the Results section seems to be more appropriate under the Methods section, as it explains the methods they used to derive the data from their example.
A19: We reorganized the manuscript considerably. We hope the revised manuscript structure is clearer.

12. Page 13, lines 10 to 15: so the idea is to have two tests or testing algorithms; one with a high PPV and one with a high NPV. What would be the consequences if these two tests were used in the same patient and then both turn out to be positive? Or the one with the high PPV is negative and the one with the high NPV is positive? Maybe elaborate on this a bit more in the Discussion section (I know the authors explain that the two tests should be used in relation to a specific context, but a bit more explanation is required here).

A20: This is a pertinent observation; thank you for flagging this. We introduced some discussion on this particular limitation of using two independent tests for rule-out and rule-in in the discussion section.

13. Page 13, lines 16-22 and Figure 5. It is not clear to me how the sub-figures in Figure 5 are derived and what they add to the storyline. Could the authors please add some text to the Methods paragraph explaining all this? And are Figures 5A and 5B really necessary? I also read from the Figure caption that these figures are based on hypothetical data: this should have been made explicit in the Methods section.

A21: The reviewer is correct on the limited relevance of panels A and B in Figure 5. These were omitted from the revision. We indicated in the main manuscript the presented data is hypothetical.

14. Perhaps another real-life example, including practical considerations, would have been helpful.

A22: We expanded the discussion section considerably to better frame this tool in the development of novel prognostic tests. A reference to previous work, wherein the used a combined Sensitivity and PPV criterion to select and subsequently validate a biomarker based for preeclampsia risk prediction.

15. One more general comment: these methods imply that the sensitivity, specificity and AUROC of a test are the same, irrespective of whether it is used in a 5% prevalence situation or a 20% prevalence situation. Although I do understand that this is mathematically indeed the case, in practice the population with a prevalence of 5% will be a different population than that of 20%. And these different populations may cause differences in test
accuracy (sensitivity, specificity). Could the authors please address this point in their Discussion section?

A23: We amended the discussion section to address the impact of patient spectrum (within the same clinical setting). The limited value of equi-prevalence lines to assess test performance in different clinical sections is also highlighted.

16. Are there any limitations to this method or study that should be mentioned in the Discussion section?

A24: Limitations of the use of equi-prevalence lines as well as the development of 2 independent prognostic were addressed by earlier points A20 and A23.

17. I have checked the website the authors mention at the end of the Conclusion section. It would have been helpful to mention this website a bit earlier and perhaps to use some more pictures from this website in the manuscript. It seems to be very helpful. Only point is: the bottom right figure states the AUC that belongs to the ROC, but it is depicted next to another graph. So it could be interpreted as the area under that bottom-right curve.

A25: We introduced the web-tool together with a brief description of its capabilities in the methods, and refer to it in the abstract.

18. Are there any benefits for the authors if readers access this website? In that case, it should be mentioned, I think.

A26: The authors will not benefit of interested people accessing the website. Yet, we did introduce an end-user license agreement to prevent others to commercially benefit from the software: an R-package to implement this tool is made available via the website. The R package is released under a GNU general public license. We thank the referee for raising the awareness regarding putative benefits from the web-tool

19. Figure 3 and 4: I have checked the calculations in Excel. These seem to be correct, but for Figure 3A they are only correct if you start with the sensitivity of 50%. But that is not how the methodology is described. In the Methods section, the authors describe that the PPV and specificity are known and that the sensitivity is calculated. However, if I use equation F to calculate the sensitivity for the numbers presented in figure 3A, I get three different sensitivities: 48%, 54% and 57%, while the figure shows three times sensitivity of 50%. So
apparently, the authors used the sensitivity of 50% to calculate the corresponding specificity and after rounding, this results in the same numbers as in the Figure. But this is not what was stated in the Methods section.

A27: The referee is correct that the data presented in Figure 3 correspond to a $Sn = 0.50$; the figure caption has been amended to clarify this.

With regards to using formula F to calculate back the value of Figure 3; the different values follow rounding of the Specificities in the figure. When using the non-rounded values 0.973684, 0.944444 and 0.875, the back-calculation works out.

Reviewer #2: Essentially the method proposed here relies upon treating the PPV and prevalence as mathematical parameters in order to define a straight line relationship between the sensitivity and specificity.

Thus for a given PPV and prevalence the locus of sensitivity of specificity pairs may be derived and follow a straight line. This idea is not new - Willis and Hyde do just this using the test positive rate and prevalence as the parameters to help define regions in ROC space for applicability (this should be acknowledged - J Clin Epidemiol 2014;67(5):538-46 and J Clin Epidemiol 2015;68(8):847-54 >> RT found 2nd paper; see enclosed ). Here, the authors use the straight line relationships to define regions in ROC space where either the PPV and specificity criteria are satisfied or the NPV and sensitivity criteria are satisfied. In the example, they derive a region in which both of these are satisfied. The potential of the method would lie with test development as acknowledged by the authors.

Of course this depends on knowing the prevalence and deciding upon an acceptable criteria. What the authors fail to do in their analysis is to take into account the uncertainty in the estimates for the prevalence which are unlikely to be known accurately as assumed here. This would undoubtedly widen the permissible regions. The paper would benefit from including such an analysis.

A28: We thank the referee for pointing us to the interesting papers of Willis and Hyde. We were not aware of these. We referred to them in the discussion.

This is indeed a very relevant remark. We have actually been working on extending the scope of the work. The uncertainty on the prevalence, but also on the predictive values are of high relevance. We are developing methods and an R package to cover this matter. This later work is more subject to discussion and will hopefully lead to further research. Presenting this work in the
present manuscript would make it overly complex and may distract it from its main focus. We have therefore chosen to present the work about uncertainty in another publication.

Other points

1. Page 7 Line 9

Although there is no strict mathematical relationship between the sensitivity and specificity and prevalence, it is a myth to think that they are independent. Spectrum bias/effects demonstrates this clearly. Indeed a change of prevalence is often a crude marker for a change of patient spectrum which leads to a change of sensitivity and specificity of the test. As such this could potentially affect the AUROC.

A29: We would like to thank the reviewer for flagging this matter; it triggered some instructive conversations amongst the authors. The importance of patient spectrum is now referred in the “Background”. A discussion on the interrelationship between lines of equi-predictive value and patient spectrum is also added to the “Discussion” (see also A3)

2. Page 7 line 15 & 16

Probably better to say ’correctly identifying the "noise" from the "signal plus noise" ’

A30: A more accurate definition was put in place; see also A4.

3. Page 9 Lines 4-6

Depends on what the screening test is being used for, and how serious the disease is. It may be that deciding upon an acceptable false negative rate first and seeing where the ROC curve crosses the sensitivity criterion may be the most appropriate approach to take in some cases.

A31 We agree with the referee the clinical context is important in deciding which is the most appropriate way of establishing thresholds. We see the use of equi-prevalence criteria as an additional tool to assist researchers to develop prognostic tests.

4. Page 9 Line 9; See earlier point 1

A32: Please refer to A29 and A3.
5. Page 9 Line 14

This is just a rearrangement of Bayes' theorem and should be acknowledged as such. Further, both the specificity and sensitivity are fixed in Figure 2B, thus the likelihood ratio is fixed, and so it is illustrating PPV odds = LR x Prev odds (Bayes’ theorem) - namely for a fixed LR the PPV increases with the prevalence.

A33: The text has been modified accordingly; explicit reference to Bayes’ theorem is made. We established the relationships for Likelihood ratios also (Supplementary Information)

6. Page 10 Line 9

Again this a rearrangement of Bayes' theorem and should be stated

A34: The text has been modified accordingly; explicit reference to Bayes’ theorem is made.

7. Page 13 Lines 10-15

The 'rule-in' or 'rule-out' criteria really depends on the prevalence. The example shown in figure 5C, shows that setting the sensitivity >0.5 as the criterion is totally inadequate when the prevalence is 0.05. The corresponding PPV of 0.13 hardly rules in the condition. It is likely that the thresholds for sensitivity and specificity will depend on the prevalence or pre-test probability unless multiple testing is considered. This should be stated

A35: Albeit the proposed “enrichment” is modest, it would still have relevance. At this moment, clinicians have no means to stratify the 1st time pregnant women without overt risk factors (e.g. renal disease or diabetes) into a high-risk group warranting more vigilant care (and into a truly low risk group). The proposed rule-in PPV aligns the post-test probability to the risk in 2nd pregnant women who experienced pre-eclampsia before. This level of risk is currently sufficient to modify the care pathway. For that reason, we feel this modest enrichment is still appropriate.

Reviewer #3: In the methods section, sensitivity, specificity, NPV an PPV are clearly introduced.

I think the proposal that "...should mimic the pre-eclampsia risk information as available for a second-time pregnant woman" (p 12, line 11-12) is perhaps a little too ambitious, considering that it "remains to be determined by a stochastic process" and the authors find the absence of such a test unsurprising (p 13, line 1-2).
A36: We fully agree with the reviewer. It was upon this (sobering) realisation that the idea of developing separate rule-in and rule-out tests.

No explanation is given as to why "multivariate tests that comply with either the rule-in or the rule-out test will also be more robust to variance in prevalence" (p 13, line 20).

A37: This is a very valid point. Thus far, this is an hypothesis which needs to be proven by validation. We lifted this statement therefore from the results and put it in the discussion section, where we elaborated on this. Noteworthy, we applied this concept of a combined minimum PPV and minimum Sensitivity criterion once before, in the development of a protein biomarker based prediction test for pre-eclampsia. In this instance, we were able to validate this test in an independent patient population. We included a reference to this example in the discussion.

The central purpose of the equi-ppv and npv lines seems to be the visualization thereof. I find the visualization tool to be informative, and agree with the authors that the equi-PPV and equi-NPV lines may be valuable statistical tools. However, these tools can only be valuable to clinical researchers, if these are available to them. Therefore, I would like to suggest that the code be moved to a public repository or published as an R package.

A38: In addition to the web-based tool, the authors made a dedicated R-package available on the same website.

I'm having trouble interpreting Figure 5 A and B. So please add some more elaborate captions. E.g. in 5B, what do the blue and red indicate? Are these related to the rule-in and rule-out tests?

A39: Since these panels do not really add to the manuscript we removed them altogether.