Author's response to reviews

Title: Evaluation of Prediction Models for the Staging of Prostate Cancer

Authors:

Susie Boyce (susieboyce1@gmail.com)
Yue Fan (yue.fan@ucd.ie)
R William G Watson (william.watson@ucd.ie)
T Brendan Murphy (brendan.murphy@ucd.ie)

Version: 2 Date: 22 September 2013

Author's response to reviews: see over
Dear Editors,

Thank you for reviewing the original research article entitled “Evaluation of prediction models for the staging of prostate cancer”. I would also like to sincerely thank the reviewers for their substantive, helpful and informative reviews. I have addressed the comments and queries of Reviewer 1 (pages 2-3) and Reviewer 2 (pages 3 -13) below and attach the revised manuscript, with revisions/edits illustrated by means of underlining, as part of my submission.

Yours Sincerely,

Susie Boyce, BSc
Reviewer 1

Major Compulsory Revisions

COMMENT 1:
The outcome groups discussed in the introduction are NOT mutually-exclusive. While ECE and SVI define pT3a and T3b respectively, the two together are worse than either alone, and LNI can co-exist with either, neither, or both. Multiple, well-validated postoperative risk instruments (e.g., Kattan and Stephenson nomograms, CAPRA-S score) have shown this in the past. This is a major problem with the methods and study design.

RESPONSE 1:
The introduction has been revised to exclude the use of the term ‘mutually exclusive’, see page 3 of the attached revised manuscript. The overall introduction has been revised to re-focus the clinical question and we hope that this addresses issues referenced in relation to the overall study design.

COMMENT 2:
The introduction is much too long – should be limited to 1-2 paragraphs to frame the specific question being asked. This is particularly relevant because this is a well-established question. The majority of the material in the intro can be shortened or moved to the discussion.

RESPONSE 2:
The introduction has been significantly revised with the aim of re-focusing the research question being addressed by this study, see pages 3-4.

COMMENT 3:
A “nomogram” is a graphical representation of a prediction model, not the model itself. So searching Pubmed for “nomograms” will miss important tools such as the Partin tables and CAPRA score, both of which predict pathologic stage without use of a nomogram per se.

RESPONSE 3:
I agree that a nomogram is not a prediction tool, and the terminology is often confused in this field. Following the revision of the introduction, we have since removed this entire sentence as it is not the focus of the paper.

COMMENT 4:
Extent of biopsy involvement (# of cores, % of cores, % of tissue etc.) is a key predictor of outcome in multiple models (e.g., Stephenson, CAPRA), so its absence here is quite problematic and may help explain the poor performance of these models.

RESPONSE 4:
Unfortunately we do not have full data regarding extent of biopsy involvement available. As well as this, we have since revised our introduction to focus on the fact that that we are using the predictive variables used by Partin et al in their development of the Partin tables, and they do not include extent of biopsy involvement as a predictive variable, please see page 4.

COMMENT 5:
There are dozens of candidate markers in development. If the authors wish to raise this issue in discussion, they need to be more thorough in their review rather than just citing their own marker set in development.

RESPONSE 5:
We have since removed the reference to our own work from the discussion and have instead included a more general statement regarding the emerging biomarkers with a number of reviews referenced, see page 15.
Reviewer 2

Introduction/Background

COMMENT 1:
This introduction is outdated and should be improved to address several key issues on the subject. Current cancer statistics data has recently been updated.
RESPONSE 1:
The first sentence of the introduction has since been revised and updated to include the current prostate cancer statistics, see page 3.

COMMENT 2:
The current decision to perform a biopsy is provided in the 2010 NCCN prostate cancer guidelines
RESPONSE 2:
The introduction has been extensively revised to refocus the research question. As part of this, we have since removed some of the extra information regarding the clinical variables (including biopsy) as we feel the variables themselves are not the focus of the paper - we have highlighted the statistical techniques used to model and evaluate the variables. See pages 3-4.

COMMENT 3:
One of the most current clinical dilemmas involves overdiagnosis and overtreatment though has been noted often as a subject of the active surveillance option for CaP (2-6) and how to manage these decisions
RESPONSE 3:
We have added a sentence in the first paragraph of the introduction to address this point, please see page 3.

COMMENT 4:
Not fully addressed in the introduction is the probability of missing significant cancer (high grade) during a systematic biopsy of 12 or more cores that include the lateral apex and central zone (4-5).
RESPONSE 4:
Please see RESPONSE 2 above.

**COMMENT 5:**
As to the subject of Gleason grading, besides the original work of Donald Gleason, this subject and its importance to the management of CaP should include the work of Dr. Jonathan Epstein and Peter Albertsen (8-11)

**RESPONSE 5:**
The introduction has been updated to reference to work of Epstein and Albertsen, see pages 3 and 5.

**COMMENT 6:**
Further, several other approaches to PCa staging, besides those mentioned, are not presented (12-15). Today the Partin Tables remain a standard of care for patients diagnosed with PCa among residents and practicing urologists.

**RESPONSE 6:**
Additional approaches to PCa staging have been referenced, please see page 3. As well as this, the introduction has been revised to focus on the Partin tables, as you correctly point out that the Partin tables are used as a standard of care for patients by clinicians, see pages 3-4.

**Methods**

*Study Population*

**COMMENT 1&2:**
1 The authors only used the Partin criteria of biopsy Gleason score, DRE status (Clinical stage) and total PSA value.
2. Why not introduce PSA Density (PSAD) as an additional variable?

**RESPONSE 1&2:**
We have since revised our introduction to focus on the fact that that we are using the predictive variables used by Partin et al in their development of the Partin tables. Partin et al do not include PSA Density (PSAD) as a predictive variable, please see page 4.

*Clinical and Pathological Assessment*

**COMMENT 1:**
Define organ-confined PCa clearly!

**RESPONSE 1:**
The definition of the PCa staging has been revised to use the T staging definitions, see page 5, reference [33].

Statistics

COMMENT 1:
For the multivariate logistic regression (MLR) modeling, what variable cut-off value was employed to construct the MLR models?
RESPONSE 1:
The logistic regression model consists of three clinical variables combined in one model – not only were the three variables modelled together, but the interactions between each was also allowed for in the model. Based on this, there is no single cut-off possible for three separate variables, rather a cut-off for a combination of the three variables combined, which have no direct interpretation.

COMMENT 2:
Did the authors apply forward or backward MLR modeling?
RESPONSE 2:
Neither backward nor forward feature selection logistic regression was used because no model selection was performed. The variables used in the model did not change, the exploratory variables were kept constant and therefore variable elimination was not required.

COMMENT 3:
Did the authors supply the final equations for their MLR models (i.e. model constant, Beta coefficients, etc.) to test their MLR models?
RESPONSE 3:
This was not supplied because this study explored three statistical classification methods – not just logistic regression. While it is possible to ascertain this information for a logistic regression model, the same cannot be said for k nearest neighbours or random forests. Therefore due to the fact this information would only be available for one of the three techniques examined we did not report this for reasons of consistency. However, we can supply the final equations for the LR model if the reviewer feels they are important.

COMMENT 4:
The equation #1 has poor resolution.
RESPONSE 4:
This has been changed to improve resolution, see page 8, equation (1).

COMMENT 5:
The last sentence in the paragraph uses different fonts.
RESPONSE 5:
The different fonts were used to illustrate the R packages used. This font is the same font used when reporting results from the R statistical software package.

Results
Table 1. Mathematical formulae underlying the classification models
COMMENT 1:
These are not mathematical formulae; rather these are the variables to input for the models and in its present form the Table is not useful.
RESPONSE 1:
We have since removed Table 1.

COMMENT 2:
Is the logistic Regression actually Multivariate Logistic Regression and is it backward or forward stepwise?
RESPONSE 2:
The logistic regression is not a multivariate logistic regression. Multivariate logistic regression is used when the outcome variable has more than two categories, which is not the case in this study. In relation to the second part of this question please see RESPONSE 2 under the Statistics section above.

Table 2. Prostate Cancer Research Consortium patient cohort characteristics
COMMENT 1:
There were 384 cases with no DRE reported and that will have a severe effect on the analysis and the outcomes since this variable is critical to any predictive solution.
RESPONSE 1:
We also carried out the analysis only using cases with entire data. We found the results to be very similar, despite including cases with missing data (those which were reported in the study). To illustrate this we include below a table of AUC values, the first row of AUC
values were calculated from the models developed using only cases with full data (N=379) and second row were calculated from models developed using the entire cohort (N=603). The fact that the AUC values created using N=603 and N=379 are similar is due to the fact that even a sample size of N=379 is more than sufficient for classification models developed using only 3 exploratory variables. We do not plan to include the below table in the revised manuscript.

Comparison of AUC values calculated excluding (N=379) and including (N=603) missing data

<table>
<thead>
<tr>
<th></th>
<th>Logistic Regression</th>
<th>Random Forests</th>
<th>K Nearest Neighbours</th>
<th>Biopsy Gleason Score</th>
<th>PSA</th>
<th>Clinical Stage (DRE)</th>
<th>Partin Table</th>
</tr>
</thead>
<tbody>
<tr>
<td>N=379</td>
<td>0.613</td>
<td>0.582</td>
<td>0.515</td>
<td>0.597</td>
<td>0.546</td>
<td>0.451</td>
<td>0.572</td>
</tr>
<tr>
<td>N=603</td>
<td>0.622</td>
<td>0.584</td>
<td>0.570</td>
<td>0.618</td>
<td>0.571</td>
<td>0.458</td>
<td>0.572</td>
</tr>
</tbody>
</table>

COMMENT 2:
Where are the data for the number of organ-confined and non-organ confined cases?
RESPONSE 2:
The data was included in the header section of the table, in the form of OC (N=427) and NOC (N=176), see Table 1 page 23.

COMMENT 3:
Where is the data for ECE, LN, SV, etc. involvement status?
RESPONSE 3:
This table has been altered to include the breakdown for ECE, SV and LN, see Table 1 page 23.

COMMENT 4:
If you have TRUS and serum PSA data, you should have PSAD data. Why not use it?
RESPONSE 4:
We have since revised our manuscript to focus on the fact that that we are using the predictive variables used by Partin et al in their development of the Partin tables. Partin et al do not include PSA Density (PSAD) as a predictive variable, please see page 4.

Table 3. Percentage of Gleason score upgrading or downgrading

**COMMENT 1:**
The upgrading error is quite significant but not necessarily unexpected. Was this error due to sampling errors contributed by changes in biopsy sampling methodology over the years?

**RESPONSE 1:**
While there have been changes in the biopsy methodology over the years we have reviewed the biopsy information for the patients carefully and have ruled out sampling errors contributed by changing methodology, for example, we have reviewed the number of cores taken during biopsy and found no significant difference in the number of cores taken across the years of the study (2006 – 2011).

**COMMENT 2:**
Once again the PSAD information could be of value to evaluating grading errors and also to possibly improve the models.

**RESPONSE 2:**
As mentioned previously, we have since revised our manuscript to focus on the fact that that we are using the predictive variables used by Partin et al in their development of the Partin tables. Partin et al do not include PSA Density (PSAD) as a predictive variable, please see page 4.

Table 4. Discrimination of prediction models and individual clinical variables

**COMMENT 1:**
The LR analysis of the data (Youden and Brier Indices) on pages 11- and 12 is quite useful. It clearly appears to me that Logistic Regression was the best performer and that is usually the case (i.e. Partin Tables) and other predictive algorithm publications.

**RESPONSE 1:**
We would agree with the reviewers comment.

**COMMENT 2:**
However, with 63.3% of the DRE status data missing, how much would the loss of so much data impact the predictive models?

RESPONSE 2:
Please see RESPONSE 1 above under Table 2. Prostate Cancer Research Consortium patient cohort characteristics section

COMMENT 3:
Should the authors have used only the cases with all variables available?

RESPONSE 3:
Please see RESPONSE 1 above under Table 2. Prostate Cancer Research Consortium patient cohort characteristics section

COMMENT 4:
How well would have the Partin Tables worked with cases where all the variables were available? This can be done on line at the Brady Urological Research Institute website.

RESPONSE 4:
The study has been revised to now include the Partin tables as a comparison reference to the models developed in this study, please see page 8, page 11 and Figure 2.

Figures 2a-g: Calibration curves which illustrate predicted probabilities on the x-axis and the actual outcome on the y-axis. In this case the outcome is a binary variable hence Loess smoothing (red line) was used to estimate the actual outcomes for (a) logistic regression model, (b) random forests model, (c) k nearest neighbours model, (d) biopsy Gleason score, (e) age, (f) PSA and (e) clinical stage.

COMMENT 1&2:
1. Calibration curves support conclusions for the three models compared and Logistic Regression appears to be the best.
2. Also, the analysis of the three input variables seem acceptable.

RESPONSE 1&2:
We would agree with the reviewers comment.

Summary of the Results Section:

COMMENT:
There was a 30% error in staging in the patient cohort to start with. This is a serious defect in design since the Partin Tables and other algorithms have a very high degree of accuracy to predict OC.

RESPONSE:
30% of the patients included in this study had NOC disease because we needed a representative from both OC and NOC to be able to develop predictive tools to predict stage, this was essential for the study. The paper has been significantly revised and now includes the 2007 Partin table predictions, which we use to compare the models developed in this study to the current ‘gold standard’.

COMMENT:
The missing information for clinical staging (DRE) may have compromised the modeling. The weakest variable on calibration assessment.

RESPONSE:
Please see RESPONSE 1 above under Table 2. Prostate Cancer Research Consortium patient cohort characteristics section

COMMENT:
Perhaps the use of PSAD since the information was available may have been useful to include.

RESPONSE:
We have since revised our manuscript to focus on the fact that that we are using the predictive variables used by Partin et al in their development of the Partin tables. Partin et al do not include PSA Density (PSAD) as a predictive variable, please see page 4.

COMMENT:
The authors should have tested the Partin Nomogram and Tables to confirm their results.

RESPONSE:
The study has been revised to now include the Partin table, please see page 8, page 11 and Figure 2.

Discussion

COMMENT 1:
The authors clearly define weaknesses in their study as it relates to post-operative staging (30% incorrect) and the high degree of upgrading noted from biopsy to radical prostatectomy Gleason scores.

**RESPONSE 1:**
We agree with the reviewers comment.

**COMMENT 2:**
Adding a fourth variable, age, was of little consequence to the algorithm solutions.

**RESPONSE 2:**
The study has been revised and age has been excluded as an exploratory variable.

**COMMENT 3:**
No mention of other additional potential markers (i.e. PSAD, surface of area of cancer in the biopsy etc.).

**RESPONSE 3:**
The aim of this study was to evaluate statistical classification techniques for data of this nature, and not to evaluate new markers of prostate cancer stage. However, some select reviews around prostate cancer biomarker reviews have now been referenced in the discussion, please see page 15.

**COMMENT 4:**
What about new serum and urine biomarkers to make such predictions

**RESPONSE 4:**
While new serum and urine biomarkers are not the focus of this study, we have references to a number of studies looking at new potential biomarkers for prostate cancer, see page 15.

**COMMENT 5:**
Points not discussed were: race, new molecular (protein, genetic and epigenetic biomarkers), new PSA derivative etc.

**RESPONSE 5:**
In our Irish cohort of N=603 men all patients were Caucasian, therefore race was not discussed. We have not evaluated new markers in this study, but this is an area we are currently researching – using the methodological approach set put in this study we are currently evaluating 9 protein biomarkers.
COMMENT 6:
The solution is not more statistical models, rather better and more meaningful biomarkers to improve the prediction outcomes.

RESPONSE 6:
We would agree that more meaningful biomarkers are essential for better prediction of outcomes; however, we also propose that the accurate evaluation of these biomarkers is just as important. In this paper we have introduced an extensive marker integration and evaluation technique in the form of modelling using logistic regression and evaluating using discrimination, calibration and clinical relevance.