Author's response to reviews

Title: Collectives of diagnostic biomarkers identify high risk subpopulations of haematuria patients: exploiting heterogeneity in large-scale biomarker data

Authors:

Frank Emmert-Streib (f.emmert-streib@qub.ac.uk)
Funso Abogunrin (abogs5@yahoo.com)
Ricardo de Matos Simoes (r.dematossimoes@qub.ac.uk)
Brian Duggan (bduggan71@hotmail.com)
Mark W Ruddock (mark.ruddock@randox.com)
Cherith N Reid (cherith.reid@randox.com)
Owen Roddy (oroddy02@qub.ac.uk)
Lisa White (lwhite11@qub.ac.uk)
Hugh F O’Kane (hugh_o_kane@hotmail.com)
Declan O’Rourke (Declan.ORourke@belfasttrust.hscni.net)
Neil Anderson (neil.anderson@belfasttrust.hscni.net)
Nambi Rajan (nambi.rajan@belfasttrust.hscni.net)
Kate Williamson (k.williamson@qub.ac.uk)

Version: 3 Date: 29 September 2012

Author’s response to reviews: see over
Dr Sabina Alam
Editor
BMC Medicine

29 September 2012

Dear Dr Alam,

Collectives of diagnostic biomarkers identify high risk subpopulations of haematuria patients: exploiting heterogeneity in large-scale biomarker data
Frank Emmert-Streib, Funso Abogunrin, Ricardo de Matos Simoes, Brian Duggan, Mark W Ruddock, Cherith N Reid, Owen Roddy, Lisa White, Hugh F O’Kane, Neil H Anderson, Declan O’Rourke, Thiagarajan Nambirajan, Kate E Williamson.

Thank you for the referee reports for our manuscript 1271796137750173 and for giving us the opportunity to submit a revised manuscript. The reviewers raised some important points and in so doing have encouraged us to completely re-write the paper. We believe that our revised manuscript is easier to understand and that the data are now presented in a logical fashion.

We have uploaded a revised manuscript and a second highlighted version of the manuscript in which the text which is unchanged is black; the new text is blue underlined; and the text which was deleted and moved to a new position is underlined and highlighted in green. Deleted text is visible in the right margin.

With respect to the tables and figures we have summarised the changes in the table below

<table>
<thead>
<tr>
<th>Original submission</th>
<th>Revised</th>
<th>Replaced</th>
<th>New</th>
<th>Revised submission</th>
</tr>
</thead>
<tbody>
<tr>
<td>Table 1</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>Table 2</td>
</tr>
<tr>
<td>Table 2</td>
<td>N/A</td>
<td>yes</td>
<td>yes</td>
<td>Figure 2</td>
</tr>
<tr>
<td>Table 3</td>
<td>yes</td>
<td>no</td>
<td>no</td>
<td>Table 7</td>
</tr>
<tr>
<td>Table 4</td>
<td>yes</td>
<td>no</td>
<td>no</td>
<td>Table 1</td>
</tr>
<tr>
<td>Table 5</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>Table 4</td>
</tr>
<tr>
<td>Table 6</td>
<td>yes</td>
<td>no</td>
<td>no</td>
<td>Table 3</td>
</tr>
<tr>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>yes</td>
<td>Table 5</td>
</tr>
<tr>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>yes</td>
<td>Table 6</td>
</tr>
<tr>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>yes</td>
<td>Table 8</td>
</tr>
<tr>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>yes</td>
<td>Table 9</td>
</tr>
<tr>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>yes</td>
<td>Table 10</td>
</tr>
<tr>
<td>Figure 1</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>Figure 1</td>
</tr>
<tr>
<td>Figure 2</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>Figure 3</td>
</tr>
<tr>
<td>Figure 3</td>
<td>Yes</td>
<td>no</td>
<td>no</td>
<td>Figure 4</td>
</tr>
</tbody>
</table>
We detail below how we have addressed the comments from both reviewers:

**Referee 1**
This is an elegant study into an important area of research, by a well-respected group of researchers. I enjoyed reading it, but would suggest that two comments might be helpful.

Comment 1. Is the cohort of patients an all-comer, sequential series, or have patients with, for example, urinary tract infection or nephrotic syndrome been excluded by other investigations?

**Response to comment 1**

*Relevant insertions appear on pages 7 and 31 of the revised manuscript*

All patients with haematuria and either recent cystoscopy or planned cystoscopy were eligible for recruitment to the study. Therefore, we did not exclude patients with urinary tract infection or nephrotic syndrome. We have further addressed this point by including more detail about the patient recruitment, details of excluded patients and more details about the final diagnoses of all patients. This information is on page 7 of the revised manuscript in the first section of the Methods under a subheading “Patient information and samples”. Further information about the final diagnosis is included in Table 1 on page 31 of the revised manuscript.

Comment 2.
From the point of view of the clinician who would like something to help in the clinical setting (which this analysis sets out to do), is this really clinically applicable? when one sees a number of mathematical formulae, the reader may well feel that there is a separation between scientific endeavour and user-friendliness in the clinical setting.

**Response to comment 2.**

*The relevant section appears under the heading “Translation of risk and diagnostics classifiers from systems biology to the clinic” beginning on page 18 of the revised manuscript. Figure 4 which has been modified.*

To address this important point we have extensively modified the Results and discussion section to ensure that our findings are communicated logically and in a way that clearly demonstrates the clinical relevance. We have added headings to the Results and discussion section and one of these “Translation of risk and diagnostic classifiers from systems biology to the clinic” specifically addresses this comment. In addition, we have modified Figure 4 to better illustrate how we predict that risk classifiers could be applied in the primary care clinical setting to improve outcomes for a greater proportion of haematuric patients.

**Referee 2**
I have no competing interests. There may be some interesting ideas here, but the reporting and the analytical approach are both extremely confusing.

Comment 1.
The paper is poorly structured, including conclusions in the introduction section, and introduction in the results section. For example, the results section includes the statement “There were an estimated 150,200 deaths from bladder cancer worldwide in 2008”.

Response to comment 1.

*Inappropriate text which appeared in the Results and discussion section has been moved to page 4 and the top of page 5. In the highlighted version of the manuscript these sections of text appear as blue underlined text which is highlighted in green.*

We have edited our manuscript to ensure that all content is in the appropriate section.

Comment 2.
The research design seems circular in my view. It appears that the authors use biomarkers, as well as clinical data, to create different clusters and then set out to determine the association between the markers and the cluster. More specifically, the role of outcome in validation of the clusters is unclear. I wanted to know whether patients in some clusters had a higher cancer risk than those in other clusters, with an associated ROC curve. But this was skated over. In table 2, there does not seem to be any clear association between cluster and cancer incidence.

Responses to comment 2

(i) It appears that the authors use biomarkers, as well as clinical data, to create different clusters and then set out to determine the association between the markers and the cluster.

*Relevant sections appear on pages 9 and 10 of the revised manuscript*

We used only biomarkers to create first, patient clusters and second, biomarker clusters. This is detailed in the revised manuscript on pages 9 and 10 under the headings “Identification of patient clusters” and “Identification of biomarker clusters”, respectively.

(ii) More specifically, the role of outcome in validation of the clusters is unclear. I wanted to know whether patients in some clusters had a higher cancer risk than those in other clusters, with an associated ROC curve. But this was skated over.

*Relevant illustrative and text revisions appear in the following places: (1) Figure 2, (2) the section “Non-random distribution of cancer-risk characteristics across patient clusters” which begins on page 12 of the revised manuscript and (3) Table 3 which includes the ROC for the three largest patient clusters.*

The point raised about whether the patients in the patient clusters had a higher risk of cancer is now covered by the insertion of a new figure (Figure 2) into the revised manuscript and is also described in the section which appears under the heading “Non-random distribution of cancer-risk characteristics across patient clusters” which begins on page 12 of the revised manuscript. Table 3 which includes the ROC for the three largest patient clusters appears on page 33 of the revised manuscript and the appropriate row is labelled “Patient clusters”.

(iii) In table 2, there does not seem to be any clear association between cluster and cancer incidence.
Relevant insertions are included in the last paragraph on page 12 in the revised manuscript and in Figure 2. Figure 2 has replaced Table 2 as it appeared in the original submission.

There was a clear association between the patient clusters and cancer incidence as now stated in the last paragraph on page 12 and in illustrated in Figure 2 of the revised manuscript.

Comment 3.
The results are generally overinterpreted. For example, the authors claim that their findings “perhaps indicate the dominance of different biological pathways within the different subpopulations”. This is a pretty small study (fewer than 100 cases!) so breaking things apart by subgroup is a little silly. Table 6 for example, is trying to get too many conclusions from too few data. As another example, the authors state that their approach is “supported by the evidence that distinct gene linkages are involved in similar pathologies thus indicating the existence of distinct disease specific modules”. But this is marker research, which is about phenotype, not genotype.

Responses to comment 3

(i) The results are generally overinterpreted. For example, the authors claim that their findings “perhaps indicate the dominance of different biological pathways within the different subpopulations”. This is a pretty small study (fewer than 100 cases!) so breaking things apart by subgroup is a little silly.

Insertion of “suggest” on pages 3, 16, 18, 19, 30 and 35

Page 7

In our revised manuscript, we have inserted “suggest” on pages 3, 16, 18, 19, 30 and 35 to ensure that our findings are reported in a measured way.

The study was based on analyses of data from 157 patients, not 100 as the reviewer suggested. This is now clarified on page 7 of the revised manuscript.

We have deleted “perhaps indicate the dominance of different biological pathways within the different subpopulations” and “supported by the evidence that distinct gene linkages are involved in similar pathologies thus indicating the existence of distinct disease specific modules”.

(ii) Table 6 for example, is trying to get too many conclusions from too few data.

Relevant text starting from the bottom of page 10 of the revised manuscript

We justify our comparisons of contributory biomarkers at the bottom of page 10 of the revised manuscript.

Comment 4. The authors make a large number of questionable claims, that are poorly supported. For example, they say that the bootstrap “is similar to a cross-validation, but more precise”. The citations for this strange claim (does the bootstrap really reduce improve coverage probabilities?) is not to the statistical literature.

Response to comment 4
Relevant insertion on page 11 of the revised manuscript

In response to this query we have inserted the following on page 11 of the revised manuscript starting on the fifth line under the Random Forest Classification heading

“We estimated the AUROC by using out-of-bag samples, which means that the trees of a RFC are trained with bootstrap data which omit approximately one-third of the cases each time a tree is trained. These samples, called out-of-bag samples, are used as test data sets to estimate the classification errors (Breiman L. Bagging Predictors. Machine Learning 1996, 24(2):123–140).”

Comment 5. The paper is very poorly presented. For example, the authors discuss “principle components analysis” (presumably only for those who are principled?) and report numbers without any explanation as to the referent (e.g. number in parentheses after the AUROC in table 6).

Response to comment 5

Relevant insertion made on page 35 in the legend for Table 3

We have removed the comparison to Principle components analysis because it is not essential to the reporting of our findings.
We have added that the number in parentheses after AUROC in Table 6 (now Table 3 in the revised manuscript) is SD standard deviation.

Comment 6. There may be other problems in the paper, but after a while I do confess that I had to stop reading due to sloppy presentation and wooly logic.

Response to comment 6
We have extensively re-written the manuscript.

Please let me know if you have any queries or require any further information.

Yours sincerely,

Kate Williamson

Kate Williamson PhD
Senior Lecturer/PI Uro-oncology Group,
Centre for Cancer Research and Cell Biology,
Queen’s University Belfast,
97 Lisburn Road
Belfast BT9 7BL
Northern Ireland

T: +44(0)28 9097 2790
F: +44(0)28 9097 2776
Email: k.williamson@qub.ac.uk