Reviewer’s report

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

Version: 1 Date: 23 July 2012

Reviewer: Andrew Vickers

Reviewer’s report:

BMC biomarkers and hematuria

There may be some interesting ideas here, but the reporting and the analytical approach are both extremely confusing.

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”.

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.

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.

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.

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).
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.

**Quality of written English:** Not suitable for publication unless extensively edited

**Statistical review:** Yes, and I have assessed the statistics in my report.

**Declaration of competing interests:**
I declare that I have no competing interests