Reviewer’s report

Title: Prediction of breast cancer by profiling of urinary RNA metabolites using SVM-based feature selection

Version: 1 Date: 21 January 2009

Reviewer: Marlin Friesen

Reviewer’s report:

The authors of this article show that urine levels of a selected set of nucleosides/ribosylated metabolites, isolated by cis-diol specific affinity chromatography and measured in an extremely semiquantitative way with HPLC-MS can be used to generate an algorithm which clearly separates a learning set of breast cancer case and control subjects. Application of this method to an independent cohort of case and control subjects will eventually determine if the method is valid.

Minor Essential Revisions

Title: Abbreviations like SVM-based should be avoided in article titles.

Page 3, line 4: The reference (http://who.int/healthinfo/statistics/bodqbdddeathdalyestimates.xls) does not lead the reader to the data cited. Please provide a suitable reference for this citation. To what year is this data referring?

Page 7, line 6: Figures 4, 5 and 6 are referred to in the text before figures 2 and 3. Figures should be renumbered so that their numbering is parallel with their appearance in the text.

Discretionary Revisions

Page 12: Avoidance of False Positives In this study, it appears that Reconstructed Mass Chromatograms extracted from HPLC-Full Scan Mass chromatograms used retention time as the sole criterion for identification of a peak. Why was selected Ion Monitoring of individual compounds not used to improve specificity and sensitivity? Even better, to greatly improve specificity, why was MS/MS not used?

In Yang et al., 2004, “Six reference peaks were firstly identified and used to match the individual peaks between the reference chromatogram and other chromatograms. This was done to avoid retention time fluctuation between HPLC samples. The chromatogram was then separated into seven zones, each zone bordering a reference peak. On the basis of these reference peaks, the peak matching was carried out based on retention factor (e.g. retention time, capacity factor k’, or lg k’) comparison between a sample chromatogram and the reference chromatogram.” Was any technique such as this used to avoid false
positives? If not, why not? To avoid misidentification, what tolerance in retention time was allowed for a positive identification?

Page 12, paragraph 2: Proof of Reproducibility Reproducibility was measured at one concentration only for each analyte. Was reproducibility as good if analytes were present at relatively high or low concentration compared to this test sample? Measurement of reproducibility at only one concentration is a poor measure of reproducibility over the range of concentrations is such a set of samples.

Page 12, paragraph 3: Proof of Linearity Calibration curve linearity is a poor measure of whether a method is accurately measuring concentration, especially at low concentrations where analyte levels can be inaccurate even though the standard curve is linear. Under Materials and Methods on Page 6, it is stated that many of the compounds chosen for analysis were available in the lab. Why were calibration curves for these compounds not done at least for these compounds in a conventional way using mixtures of internal standard and serial dilutions of analytes?

Page 14: Discussion Is it not conventional to apply the method developed using a learning set of cases and controls on another independent set of cases and controls and verifying that the method gives a good classification before publishing it as valid?

**Level of interest:** An article whose findings are important to those with closely related research interests

**Quality of written English:** Acceptable

**Statistical review:** Yes, but I do not feel adequately qualified to assess the statistics.

**Declaration of competing interests:**

I declare that I have no competing interests