Review of "Interim findings on the feasibility of using predictive risk modelling to identify new-born children who are at high risk of future maltreatment."

General Comments.

The report describes the application and testing of predictive risk models for future maltreatment of new-born children. The construction of the under-pinning database and the selection of statistical models for risk prediction are well-described and the choices made and methods used, seem generally sound. The record-linkage used to construct the database appears to have been carefully conducted. At first, I was surprised at the use of stepwise logistic regression, in view of the inherent instability of this method and the failure of most implementations to account for the uncertainty introduced through the model selection process. However, in the context of prediction these considerations appear less important than is the case for estimation problems. Moreover, the empirical results presented in the report regarding the predictive performance of the methods compared support the use of the stepwise logistic regression model for this problem.

The authors are appropriately cautious as to the conclusions that can be drawn for individual children based on their predicted maltreatment risk. The authors also recognise that a risk prediction tool should not be the only input into decision-making and that professionals will continue to need to exercise their professional judgement to refer children appropriately.

The report notes that not all of those determined by the model to be at high-risk were found to have been subject to mal-treatment by age 5. In fact, for all risk score cut-points considered, the positive predictive value was less than 50%. On the other hand, specificity and negative predictive values were high, indicating that the predictive modelling approach correctly classifies a very high proportion of those newborns not subsequently notified as subject to maltreatment and that the great majority of negative predictions proved to be accurate. The overall performance of the predictive modelling, as measured by the area under the ROC curve is good. Thus, the predictive modelling approach has some of the characteristics of a good screening test. Perhaps an analogy with medical screening is apt: Given a positive finding on a screening test the next step is a more-definitive investigation with intervention following only after the need is confirmed by the definitive investigation. Another similarity with the medical screening scenario is that, just as screening tests are not the only route to diagnosis and treatment, predictive risk modelling is not the only means by which at risk children and families are brought to the attention of the relevant agencies. Professionals in contact with such families should continue to refer them for follow-up and / or possible intervention, as appropriate, even if a risk scoring tool does not highlight the children concerned as high risk. This point is noted by the authors.

Although the authors' predictive risk modelling approach shows promise as the basis of a screening test it seems that additional work may be needed to optimise the threshold used to indicate high risk. This is a difficult task: With a rare outcome it is difficult to achieve high positive predictive value and the authors note the inherent tradeoff between positive predictive value and sensitivity. One useful input to this decision may be to reflect on the relative costs or, more abstractly, "consequences", of false positive and false negative classifications. In a setting where the risk score was a posterior probability from a Bayesian analysis, Austin & Brunner (2008) showed that if c_1 denotes the cost of a false positive (incorrectly predicting maltreatment) and c_2 the cost of a false negative (predicting as low risk a child subsequently subject to maltreatment) the expected cost is

minimised when the probability threshold is set at $p_{opt} = c_1/(c_1+c_2)$; thus if false negatives are regarded as three times as "costly" as false positives $p_{opt}=0.25$. Under this scenario, all those with a probability of the event of interest of 0.25 or more should be highlighted as at risk of the event. I am not sure whether this type of reasoning can be mapped directly to the current context but it might be of use informally in weighing up the reasonableness of the positive predictive values derived from alternative risk thresholds.

Some specific points.

Page 22. Table 3 (also Table 6, page 25). It would be helpful if the title included an indication of the meaning of the numerical entries in the body of the table. I assume they are an indication of the priority order of the variables, but it would be good to have this clarified.

Page 23, para 95, also page 26, paragraphs 99-102. The issue that some groups, particularly Maori, are over-represented in the group highlighted as at risk by the model compared to the observed distributions in those known to be maltreated is raised and discussed. The solution proposed is to stratify the database and construct separate models for separate groups. I have two comments. Firstly, the observation that the distribution of covariates between the group highlighted by the models as at risk of maltreatment and the group known to have experienced mal-treatment is just another way of saying the predictive model is not perfect. So some mismatch between the characteristics of the groups is inevitable. Secondly, I would be wary of going too far with the stratification and separate models for separate groups approach because of the potential for instability due to smaller sample sizes, keeping in mind that the outcome is rare. An alternative may be to experiment with adding some interaction terms to the logistic models or to look at alternatives such as tree based methods.

Page 25, para 98. The point that the data assembled for the feasibility study is a valuable resource for investigation of the over-representation of Maori in maltreatment statistics is very well made and I hope that the opportunity for additional research using the assembled data can be taken.

Companion Technical Report.

Page 11, Table 4. I understand the rationale and approach for under-sampling. However Table 4, confuses me, because number of maltreatments by age 2 in the constructed sample exceeds the total number in the full cohort. It looks to me as though the entries in the "Number" column in the "Constructed Sample" panel refer to the number *without* findings of mal-treatment, in contrast to the entries in the corresponding column of the "Population Studied" panel.

Reference.

Austin P.C., Brunner L.J. Optimal Bayesian probability levels for hospital report cards. Health Services and Outcomes Research Methodology, 2008, 8, 80-97.

Patrick Graham
Principal Statistician
Standards and Methods
Statistics New Zealand / Tatauranga Aotearoa