Peer and Statistical Review of the Report: 'Interim Findings On The Feasibility Of Using Predictive Risk Modeling To Identify New-Born Children Who Are At High Risk Of Future Maltreatment' and the 'Companion Technical Report'

April 30, 2013

John D. Fluke, Ph.D.
Kempe Center
University of Colorado School of Medicine

Lijun Chen, Ph.D.
Fred Wulczyn, Ph.D.
Chapin Hall
University of Chicago

Overview and Purpose

The purpose of this review is to provide feedback to New Zealand Ministry of Social Development (MSD) regarding the interim report by its contractors to develop a Predictive Risk Model (PRM), to identify newborn children who are at risk of maltreatment within a few years of birth. In order to complete this review the authors have read the interim report and the technical report, and we will refer to those reports as interim and technical reports throughout this document. Where appropriate we have also consulted the related literature.

Our intention is to provide comments regarding the overall approach based on the our research and evaluation experience with similar populations of children who are subject of child protection actions in other jurisdictions, including the Australia, Canada, the UK, and the US, as well as other high income countries. As a part of this review the authors do offer some specific comments regarding encouragement of additional analyses planned by the authors of the interim report, possible modifications to the methodology going forward, some discussion regarding ethical concerns, and some overall observations about the approach.

Summary of the PRM Approach

The contractors developing the PRM organized data from multiple administrative data sources including care and protection (CYF) data, benefit data, corrections, and birth data beginning with administrative data available as early as 1993. These data were linked through matching procedures to prepare a set of variables that are believed to statistically relate to children who are ultimately substantiated by CYF at some later point. Annual birth cohorts beginning with 2007 through 2010 were assembled based on birth

records with up to two years follow-up with the CYF subsequent substantiation event data.

Review of the Statistical Approach And Execution

1. Matching and Linking of Different Administrative Data Systems

First, different administrative data from several government agencies were linked together in order to obtain the information about the infants and their parents/caregivers needed to build the predictive model. These data systems include the benefit and care and protection data from the Ministry of Social Development, the birth notification and registration data, and the corrections and sentencing data. Because no single common unique identifier for the new-born child exists among the different data systems, data matching criteria and algorithms based on demographic information like first and last names, and date of birth were developed. The report claims that "conservative" matching criteria are adopted because only exact date of births and highly similar first and last names between different data systems are used to identify a child.

We agree that the procedures and steps for the linkage of data are basically sound and clearly specified with one major exception. The report does not mention how the matching algorithm addresses the issue of multiple matches of children with the same first and last names and date of births in the same data file and across data systems. From our experience in matching and linkage of different administrative data for the US children, the issue of multiple matches of children with same names and DOBs is not a trivial issue. As shown in the report, there are about 63,000 children in each birth cohort from 2007 to 2010. There must be many children with the same names and DOBs. The

report does not specify what further identifying information, such as addresses and parents' information, are used to differentiate these children.

To our knowledge, there are a few public domain data linking software available with probabilistic and deterministic record linkage algorithms, including Registry Plus developed by the National Program of Cancer Registry of CDC and the Link King.[1] Authors of the report may wish to consult the documentation for these programs and consider if the use of these procedures might improve the quality of their matched data.

2. Variable Selection

Among the many limitations of using administrative data is that the researcher is limited to the available data. Access to linked data clearly improves upon the range of such data that can be leveraged for analysis. Based on the linked data, the predictive risk model for the risk of maltreatment was developed using stepwise logistic regression models. The outcome variable is at least one substantiated finding of maltreatment by age two for each population. The predictor variables include child demographics, other siblings involved in care and protection system, parents and caregivers' age, benefit type and duration, corrections history, and childhood maltreatment experience, and other. Previous studies have found these variables to be strongly associated with the risk of child maltreatment. That said we have some specific comments regarding the nature of the variables selected for the analysis.

[1] Relevant information of Registry Plus can be found at: http://www.cdc.gov/cancer/npcr/tools/registryplus/lp.htm, and information on the Link King can be found at: http://the-link-king.com/screen_shots.html .

The first of these concerns the use of substantiated maltreatment as the primary dependent variable for the PRM. We recognize that it is tempting to consider substantiations as a clear demarcation of the existence of maltreatment or at least a conservative assessment of maltreatment status, and recognize that the authors identify substantiation as a potential issue (paragraph 74). However, in most child protection systems we are familiar with, the clarity around the substantiation decision is subject to considerable debate and discussion within the field (Fluke, 2009; Drake, 1996). A common practice in evaluating risk for child protection involvement entails looking at a combination of CPS response (investigation regardless of substantiation status), as well as, substantiation. The authors of the interim report may want to consider looking at both criteria and perhaps even notifications to child protection. However, the use of a single standard (substantiation or investigation) may also depend upon the type of intervention intended if PRM is implemented. This point is addressed a bit more fully in the general peer review.

Another variable that the authors may wish to consider is the age of the mother at the birth of the first child, we assume that the existing caregiver age variable is caregiver age at the birth of the target child. We also suggest inclusion of relevant variables about the family process, such as domestic violence. Since exposure to domestic violence is major predictor of emotional and psychological abuse, it should be included as a predictor. This information is available as the police violence notifications in the care and protection data for a child. The report also mentions the possible addition of health information in the future modeling efforts. Substance abuse should be an important

variable to consider since parental substance abuse is a strong predictor of child maltreatment in North America (e.g. Walsh et al. 2003).

It would also be helpful if the authors could provide more description of the neighborhood variable. We assume that this is a specific data set that is generally available in New Zealand, and there is some discussion of a mesh block in the technical report appendix. However, from the reports it is not entirely clear where this variable originates, why the coding was selected, or if there is other information that might be useful.

Finally, we would definitely encourage the authors to consider how endogenous variables could be better controlled in the PRM modeling process (paragraph 100 and 119). While this issue is brought up specifically in reference to ethnic disparities, the issue is generally important. The primary issue is whether families who have a history with child protection are at higher risk, or whether this is primarily a surveillance influence (Fluke, et al., 2008). While the authors do describe model performance in the presence of the different types of data sets, the relative risk impact of specific variables is not discussed; specific positive predictive values, and overall ROC performance are not presented. In the absence of this information and based on our experience with similar data we expect that these variables add considerable accuracy to the overall performance of the model, prior involvement in CYF especially for other children is likely to be highly predictive. Furthermore, intervening in such cases may also be redundant with ongoing CYF involvement given the short follow-up time frames. To the extent that the authors

could offer some sort of "control" for possible surveillance effects or would be interested in exploring this issue further, this would undoubtedly be helpful.

3. Population wide PRM tool for new-born children

The base model is built from a sample of a cohort, which includes all children with substantiated maltreatment by age 2 and 4 times as many children from the rest who are randomly selected. The sample data is further divided into a modeling sample (70% of the sample records) to develop the model and a validation sample (30% of the sample records) to test the model. The report also provides detailed information about the predictive accuracy of the base models, both with the validation sample, and through applying the 2010 base model to the 2007 birth cohort.

As mentioned above, the predictive risk model using logistic regression was developed on a sample of each birth cohort instead of the whole cohort population. The rationale for this practice is based on King and Zeng (2001), who state that logistic regression can sharply underestimate the probability of rare events. One way to overcome this is to over-sample the minority class or under-sample the majority class, the latter of which is adopted in the current report. However, we have doubts about the validity and usefulness of this practice in modeling the risk of maltreatment in this case. First, with 2.6% to 3% of the newborns found to be maltreated by age 2 in each birth cohort from 2007 to 2010, the total number of maltreated children will be over 1,500 in each birth cohort. The event of maltreatment may be rare, but the absolute size of the maltreated population is NOT a small sample, which may cause "small sample bias" and modeling difficulties like failure to converge. According to Paul Allison, the issue of rare events in

logistic regression is a red herring: what matters is not 'rare events' but 'small sample size'.[2] Second, even in the case of a small sample or cell size, there are easier ways to solve the problem. The Firth method which uses penalized likelihood can be adopted in the logistic regression model to overcome the small sample bias (Heinze and Puhr 2010). Multilevel models, as discussed below, offer an even more robust approach for weighting the risk factor estimates.

Splitting the samples (and the over/under sampling) is an effort to manage the variance of the parameter estimates but this is more efficiently done with a multilevel model. In the multilevel model, the goal is to place a confidence interval around the parameter estimates. Splitting the population (70% vs. 30%) for use with the fixed effects models reduces the sample of parameter estimates to two; in a multilevel model, the population of parameter estimates is much larger, so the estimates are more reliable. It also provides a framework for pooling the cohorts as opposed to treating them in serial fashion. Basically, the multi-level model exploits all of the data rather than in bits and pieces.

4. Sensitivity Testing

As mentioned above the PRM models appear to perform well, and in our view with considerably greater ROC AUC values than most similar child protection models we are familiar with. Even so, there are important considerations associated with the assessment of relative accuracy, many of which are addressed in the interim and technical

[2] See Paul Allison in http://www.statisticalhorizons.com/logistic-regression-for-rare-events .

reports. We will comment specifically on the issue of ethnicity and on the relative contributions in understanding sensitivity and specificity.

While it appears that the ethnic group was not a significant factor in the PRM model in terms of estimating model coefficients, it does appear from the discussion that the model will result in overall overrepresentation of Maori children and families if the threshold for intervention is 5% compared to the same threshold for non-Maori populations.

We would suggest that author's consider the possibility that even though average predictive accuracy may be consistent with the overall rates, AUC error variance and resulting confidence intervals for accuracy for the Maori population may be different as well. This is clearly a concern when setting cut points for potential interventions. Whether these potential differences in error variance are a function of ethnicity, or some other factor, if any differences are found, it may raise important ethical concerns including the possibility that use of the PRM could contribute to a higher average error rate for these populations over time, particularly given the inclusion of endogenous variables as discussed above. In other words the error may have cumulative impacts during implementation, rather than random impacts. The authors have developed separate models for this population and we would concur with the plans to do further modeling in preparation for the final report. For example, the suggested concept of setting cut points differently for Maori vs non- Maori has some possibilities in terms of addressing these concerns.

We also agree with the authors of the interim report that the idea of using the data set to develop a deeper understanding of issues associated with biases as suggested (paragraph 119) in their discussion of limitations would be beneficial. Additional discussion regarding ethnicity and ethical considerations are explored below.

The authors may want to consider a closer examination of predictive factors in the model that influence not only positive prediction and sensitivity, but also factors that are associated with improved model specificity. This issue is related to developing a better understanding of the population of children who might not be at risk of maltreatment. In particular some factors may be protective, for example, the receipt of benefits, or benefits of particular types. From the report there is no presentation of variable parameters making it difficult to assess potentially beneficial factors. In addition, variable specific analysis of ROC performance might be useful. Information of this sort may be suggestive of possible interventions that could be helpful in maltreatment prevention.

5. Person Level Risk

Our last methodological point has to do with the report's sole focus on person-level risks. There is strong evidence, across a range of disciplines and literatures, pointing to the importance of contextual risk factors (Baumann et al. 2011). The more salient point has to do with whether the risk factor effect sizes identified in the paper vary with place or some other unit of aggregation within the New Zealand context. The models used *assume* that a given risk is invariant across the sites where child protection decisions are being made. It is unlikely that that is strictly true, and in fact the endogenous nature of some of the prior child protection involvement variables is likely to

bear this out (see discussion above regarding surveillance effects). In practical terms, it means the predictive models as currently composed will represent what is true on average as opposed to what is true in a given context. In the long run information about the latter is more useful if risk varies with context, which is almost certainly the case, because most decisions are made in context.[3]

Conceptually what is needed is a unit of aggregation that meaningfully differentiates the social and/or administrative context in which decisions are made. These factors –attributes of the Decision Making Ecology (DME) – have to be added to the risk calculus via a multilevel model. The statistical methods, which are well established, solve a number of the aforementioned problems.

- Manages the non-independence of the observations
- Weights the parameter estimates for (relatively) rare events by the standard errors of the contributing population samples. This provides better protection against Type I and Type II errors than the method used.
- Incorporates contextual risks in a way that is substantively meaningful and surely strengthens the predictive model.

[3] It is true that the models used introduce neighborhood effects – what is called neighborhood deprivation in the paper. However, strictly speaking, the manner in which the neighborhood effects are exploited in the model would be more appropriate if one is concerned about how context affects risk. If the question is how the assessment of risk varies with context, then the model has to assess how risk-related effects vary with context – does context predict individual risk? For this specific formulation of the question, a random effects model is the model of choice.

Issues and Questions

A major issue identified by the authors is the ethical concerns that might be associated with PRM implementation. For one, the statistical framework of the PRM approach is most appropriate for understanding population level prediction, rather than individual prediction. As the authors of the interim report are clearly aware, checks and balances would need to be included in any implementation that involves individual prediction leading to intervention.

One consideration for implementation is the quality of the linked data at the individual level. Aside from its use statistically, it seems to us that verification, depending on the nature of the planned intervention, of whether a caregiver or child is actually correctly linked at the individual level would be a critical quality assurance need. For example, once a potential high-risk case is identified, it may be necessary to have a careful, independent capacity to review the records on which the risk rating was based and insure that the original linkage for the case is accurate. Tests of a quality assurance verification method for the linkages could be conducted on a sample basis to assess feasibility.

Going further with the issue of ethics and partly to address the fundamental concerns associated with the potential for disparate impacts on ethnic groups, we think the authors and MSD may want to reframe this issue and place some attention on what the nature of the actual intervention might be. From what is presented in the report it is implied, although very unclear, that one notion might be to involve CYF in a response not unlike what is currently performed when a notification is accepted for an

investigation. If so, and given the nature of the PRM this may not necessarily represent the most effective approach.

First, we think a more careful review of the potential options for programmatic response to the PRM identified children may be warranted at this stage. These responses could run the gamut from primary prevention such as providing educational information to more structured evidence based parenting training programs such as Nurse Family Partnership Home Visitation or Safe Care. Any of these would be far short of CYF child protection responses, and might have the advantage of providing some indication of what the likely cost effectiveness would be of the programmatic response in terms of potential avoidance of subsequent maltreatment related costs. As mentioned above regarding substantiation as the criteria, using a measure like notifications or investigations by CYF might be better depending on the intervention type.

Articulating the possible programmatic models for working with high-risk caregivers could also help to focus the ongoing PRM analysis toward a better understanding of the possible implementation options. For example, threshold levels could be lowered or raised depending on the nature and cost of the response. The PRM could be a source of information to forecast the potential need for responses of various types across a range of threshold values. Coupled with place specific analysis these analyses could also be used to target resources geographically as well.

Review of the Study Findings And Conclusions

Overall, the study is based on a rich, linked administrative database. The approach to constructing the data was thoughtful. There are ways to improve the match, but for a first pass, given the domain, the data has a lot of useful structure for understanding the policy and practice implications of contact with the child protection system early in life.

The analysis of the data requires more work. The twin risks of maltreatment and exposure to the decision-making processes that determine whether a child receives services from the child welfare system and for what reason are a function of person and context level factors. The current analysis appears to assume that risk is fixed across units of New Zealand's child welfare system, which is contrary to what is known about the nature of inefficiency in human service systems. To predict whether a given child will be reported for maltreatment, one has to know how context and place affects risk. As a source of meaningful variation, these effects are *not* represented in the model. If a multilevel framework is adopted, the predicted estimates of person-level risk will be significantly more realistic. This can be advanced in two ways: by taking advantage of the naturally nested structure of the data and by adding level two covariates to capture the influence of context on the level one parameter estimates (Wulczyn, Gibbons, Snowden, & Lery, 2013).

It seems to us that the potential for developing a fairly sound PRM model or set of models has been demonstrated based on the findings in the interim report. The precise nature of implementation of the PRM remains the weaker issue in terms of operational

feasibility, decision making context, and ethical concerns. These concerns are heightened since the analysis suggests the possibility that implementation could result in disparate impacts for Maori and potentially other ethnic groups as well. Clearly, as the work progresses it will become increasingly necessary to tie the modeling research more closely to the actual implications, both negative and positive for implementation. For this reason we think that an increasingly important aspect of the research on the model going forward needs to be directed toward addressing the range of implementation options and the degree to which the PRM model and methods would effectively support these.

References

Baumann, D.J., Dalgleish, L., Fluke, J., Kern, H. D. (2011). **The Decision Making Ecology**. American Humane Association: Washington, DC

Drake, B. (1996). Unraveling unsubstantiated. *Child Maltreatment, 1*(5), 261-271.

Fluke, J., Shusterman, G., Hollinshead, D., & Yuan, Y.T. (2008). Longitudinal analysis of repeated child abuse reporting and victimization: multistate analysis of associated factors. Child Maltreatment, 13 (1), 76 – 88.

Fluke, J. (2009). Allegory of the cave: On the theme of substantiation. *Child Maltreatment*, 14, 69-72.

Heinze, G., & Puhr, R. (2010). "Bias-reduced and separation-proof conditional logistic regression with small or sparse data sets." *Statistics in medicine, 29*(7-8), 770–777.

King, Gary, & Zeng, L. (2001) “Logistic Regression in Rare Events Data.” *Political Analysis* 9: 137-163.

Walsh C, MacMillan H.L., Jamieson E. (2003) “The relationship between parental substance abuse and child maltreatment: findings from the Ontario Health Supplement.” *Child Abuse & Neglect 27*(12): 1409-1425.

Wulczyn, F., Gibbons, R., Snowden, L., & Lery, B. (2013). Poverty, Social Disadvantage, and the Black/White Placement Gap. *Children and Youth Services Review*, *35*, 65–74.