PEER REVIEW REPORT I

As part of its assignment for the New Zealand Ministry of Social Development, TCC Group reviewed the March 2013 research report on the feasibility of a population-wide predictive risk modeling tool based on linked social sector administrative data for predicting child maltreatment. TCC Group conducted this review by assessing the statistical analysis approach and its execution, the study findings and conclusions, and the quality and consistency of administrative data. An analysis of the information collected and the resulting models as it relates to the interests of the Ministry follows.

Statistical Approach and Execution

Variable Selection

Summary: It is generally a logical approach to recode predictor variables to binary or categorical variables; however, for some variables, recording appears to have resulted in masking the significance of the predictor. While data shows reasonable collinearity, many variables would seem to be highly correlated. It is suggested that the authors explore the possibility of reducing the predictors by utilizing an alternative, more systematic method.

Recoding: The 'parent/caregiver received benefits for a mental health disorder' or 'substance abuse disorder' variables included as predictors have been re-coded in the model into binary variables. As a result, the model no longer accounts for the severity of the conditions that resulted in the ultimate utilization of benefits. It is particularly important to be able to capture information about the degree of severity of the condition given the ethical implications associated with stigmatizing parents or children as "high risk". It is suggested that such variables be coded similarly to the way in which the 'corrections history' variable was coded (no history of sentencing', 'sentencing for non-violent crimes' or 'sentencing for violent crimes') so as to preserve this critical contextual information. The reviewers acknowledge that the authors may face constraints in the way of sample size and available data.

Collinearity: Variables such as 'multiple birth' and 'other children aged under 2' would seem to be 100% correlated. They would also appear to be closely related to the variable '3+ other children'. Similarly, 'single parent' and 'father not listed on birth registration' would likely have a high correlation; as well as 'benefit type' and 'benefit for substance abuse disorder' or 'benefit for mental health disorder', etc. While collinearity does not appear to be a concern statistically, an effort should be made to make the models more parsimonious.

Variable Derivation

Summary: The data sources seem reasonable, though there is an obvious need to improve the linkage capabilities.

Methodological Approach

Summary: There is the potential for many individuals to be dropped from the final dataset if they are not linked across data sources. Other statistical methods such as propensity scoring, weighting, and simple descriptive statistics may be used to make effective comparisons across groups about individual characteristics and missing data.

Matching: It is important to know if, for example, in the case of an individual whose birthday is not included in benefits data, the individual has characteristics that are distinct from those of other individuals with matched data. Consider propensity scoring using logistic regression to predict the likelihood of missing key linkage variables and comparing the prevalence of the key outcome (maltreatment by age 2) across groups included and not included due to missing linkage variables. Additionally, consider reporting information on the percentage removed in each cohort (e.g., 2007 vs. 2010).

Weighting: Similarly, weighting should be applied to the population estimates made from the sub-group of individuals with matched data to account for any differences between this sample and the population. (The authors note that SAS Enterprise Miner adjusts for non-random sampling design.)

Predictive accuracy:

Summary: While, from a methodological perspective, the predictive accuracy of the model is supported, it is impractical to subject children that are misclassified by the model as high-risk to child welfare system interventions. Additionally, from a cost-benefit standpoint, misclassification as a result of inappropriate referrals may result in an inefficient use of resources.

Thresholds: The tests for predictive accuracy are very thorough and many of the models fall well within thresholds for acceptable accuracy suggested by Leventhal (1988). In practice, however, the underlying logic supporting the thresholds used to assess the PPV may not be practical. Leventhal (1988) is suggesting that interventions should be provided to all of those identified as high risk even though only about 25% of children in those families will subsequently be maltreated. The reviewers question the justification that even those families with children who are misclassified as "high risk" will benefit from interventions aimed at parenting problems.

Comparative Advantage: The reviewers question the comparability of referrals identified by the PRM tool and those made by frontline professionals. They would seemingly need to be aligned if the PRM tool is expected to complement identifications made by frontline professionals. Moreover, the reviewers wonder whether it is possible to conduct a cost-benefit analysis of the PRM tool as compared to the referral method that is currently in place. Suggested topics for analysis include the net cost savings of early identification (indicators might include ‘number of hospitalizations’) and cost of missed linkages- whether as a result of weak linkages in administrative data or overly conservative practices.

<u>Model performance:</u>

Summary: Using the area under the ROC as a test for goodness of fit, the models tested appear to perform reasonably well.

<u>Proportionality of predictions:</u>

Summary: The researchers transparently identified the problems they encountered with the proportionality of predictions. Developing separate models for Maori and non-Maori children is a sensible approach to improve the PRM’s performance. In addition, the researchers should consider building models for families with prior contact with social agencies, and families with no prior contact with social agencies. Some algorithms are able to develop rule sets and trees that can adapt to these different client typologies, and testing these algorithms is highly recommended.

<u>Statistical testing:</u>

Summary: While the statistical testing seems to have been conducted appropriately, in order to assess whether fewer variables might be as able to predict the target outcome variable as well as the PRM tool, further explanation of the stepwise selection process is recommended.

Stepwise Selection: In addition to the models detailed in Appendix 3, which show the combination of independent variables that are most able to predict the target outcome variable, the reader may be curious to know the log likelihood at step 0, as well as the comparative log likelihoods in the univariate logistic regression models. It is recommended that the authors first present the most important variables and then parse out those variables that are candidates for entry into the regression model before presenting the stepwise selection process.

<u>Sensitivity testing:</u>

Summary: Overall, the PPV, specificity and sensitivity rates appear to be markedly higher than similar models tested by others.

Sensitivity Thresholds: The authors presented clear tests of the sensitivity of the predictions and sensitivity thresholds were referenced from the literature (Leventhal, 1988). The models using risk scores in the top 5% fell within the suggested 40-60% sensitivity range.

Predictive Accuracy: The model shows better sensitivity in infancy due to the exclusion of newly emerging circumstances that threaten or protect the child as he/she ages. The reviewers question whether good sensitivity is achieved at the expense of predictive accuracy, as the tool does not incorporate new emerging circumstances that are threatening to the child as he/she ages.

Study Findings and Conclusions

Summary: The report makes it clear that there are still far too many inaccurate predictions for the current PRM model to be used as a sole mechanism for the identification of children at high risk. In addition, the research team states that frontline workers using the tool would need to be trained before it is used in the field to support decision-making.

Model utility: In practice the tool is identifying a group of very vulnerable families that require an array of services. While the PRM's focus was on identifying children at risk for maltreatment, the large number of false positive predictions does not yet make it useful for punitive child welfare services. The PRM model would be a useful tool for screening in vulnerable families that require additional attention from social services agencies.

Administrative Data

Summary: There are implications of using less conservative linkages such as "similar names" and "similar birthdates". Additionally, there are inconsistencies in the availability of administrative data that affect the ability to match individuals. Finally, some indicators are contextual, meaning that policy changes may result in over/under sampling a given population.

Data quality: Data linkages provide only a modest gain in terms of predictive accuracy (overdrawing solely based on benefit and care and protection data). While data linkages may help to address ethical concerns by widening the sampling frame, administrative data is imperfect. For example, care and protection history of caregivers in their own childhood is more complete for recent cohorts. In addition, there is no test to exclude stillborn children or those who have died from other children aged under two variable. Another example is that of residential addresses, which may be interpreted as different as a result of spelling errors. Finally, it is noteworthy that highest educational qualification was excluded due to many cases of missing information.

Data consistency: Policy changes have implications for the interpretation of administrative data and the ability to draw comparisons across time. For example, changes in police notifications for family violence saw increasing numbers of children with findings of emotional abuse, which had the affect of increasing rates of receipt of incapacity-related benefits. It is also noteworthy that some administrative data, for example notifications, are collected by third party individuals. Because their work is subject to less oversight, name and date of birth information may be captured inconsistently or in a manner that is vague or muddled.

Peer Review Conclusions and Recommendations

Summary: The research team, using Peters and Barlow (2004) as a gauge, has correctly pointed out that their model performs relatively well. The PRM trained on linked administrative data could be used to identify some of the new-born children at high risk of

maltreatment. In addition to integrating critical health information to improve the accuracy of the model, the research team should focus on the following issues:

- ***The interplay between PRM and agency workers***
- ***Rigorously testing and optimizing several algorithms***
- ***The relative understandability of the models for agency personnel/practitioners***

Research Design: The authors suggest that Stepwise logistic regression was selected as the preferred algorithm due to its good predictive performance, the ease with which it can be explained to stakeholders, and its straightforwardness in terms of implementation (see page 12 of Technical Report); however, David W. Hosmer, Jr., and Stanley Lemeshow suggest in their book *Applied Logistic Regression* that this method is often used when the outcome being studied is new and the covariates are not known; or when their associations with the outcome are not well understood. Duncan Lindsey, Editor of Children and Youth Services Review, has repeatedly written about poverty as a key predictor of landing in the child welfare system. If the covariates are known, as Lindsey has shown, then Hosmer and Lemeshow might suggest utilizing an alternative research design. In addition, if the primary covariate with child welfare is poverty, in incorporating a multitude of poverty indicators (i.e. corrections data, benefit data, care and protection history, etc.), this tool serves merely to identify a large number of poor families in need rather than to justify a call to action that would involve referral to the child welfare system.

Efficient Use of Resources: The report suggests that Children's Teams are resource constrained, noting that the threshold for "high risk" is informed by the capacity of the Children's Teams. If the PRM tool correctly classifies only 25% of those identified as "high risk", not only is there a mental and emotional cost to having an unwarranted visit from/referral to child welfare services, but there is also a cost to society incurred by expending unnecessary resources on deploying Children's Teams and other resources.

Testing and Optimizing Algorithms: In the meeting with MSD administrators and researchers, it became clear that not all of the algorithms used in the study were optimized in the SAS data mining tool. Very often models will underperform if they are not optimized (e.g., with boosting, misclassification costs, etc.). In future iterations of this work it will be important to optimize the algorithms to avoid bias and ensure comparability.

Healthcare Data Inclusion: It is recommended that the MSD structure the healthcare data in a way that will support the modeling process and potentially increase the accuracy of the models. For example, the mental health of the mother should be advanced beyond a binary yes/no variable to ideally get at the type of mental health issue (diagnosis), the severity, and the frequency and duration of any mental health hospitalizations or intensive community placements. The questions could be ordered in the following way:

Does the mother have a mental health diagnosis?

- Yes
- No

If Yes, what are the mother's diagnoses?

- Anxiety
- Bipolar Disorder
- Depression
- Etc.

Macro vs. Micro Model Translation: In the book, *Clinical Data Mining: Integrating Practice and Research* (Epstein, 2009), Irwin Epstein suggests that PRM/data mining research should be conducted in partnership with the practitioners on the frontlines. It may be fruitful to build in some model translation component into the frontline social work training (micro level). Additionally, the administrators and supervisors (the macro level) in the agency should know how the model was constructed and tested. Showing the "scoring" is not as important as showing the key variables in the model, and how these variables behaved in the analysis. This issue will be very important in subsequent evaluations.

New Outcome Data: The addition of child abuse hospitalization data could be a very important addition to the dataset. Modeling substantiated abuse is essentially modeling the decision-making process that workers follow in a given jurisdiction. Predicting substantiated abuse is not predicting whether abuse actually occurred, rather it is predicting whether a worker will deem it abuse or not.

Available Resources/Supports: The function of the Children's Teams will be important to track. Given the large number of false positive predictions, it will be vital to have non-stigmatizing and non-punitive services/supports available to the flagged families. Ideally, if a family is flagged as high risk by the PRM tool, the Children's Teams will have a robust set of referral sources to help families meet a complex array of needs. Further quantitative analyses could be done to identify which referral sources are the most effective with regard to keeping children from entering the child welfare system.

Lacking Intervention Data in Administrative Dataset: In addition to the limitations of the administrative data that has been discussed above, another limitation the researchers face in the development of a PRM is that the data lacks important intervention information. For example, are there certain social programs in the community that seemed to mitigate child abuse risk for high risk children? The healthcare data supplement may get at some of this information, but it is important to think about this for PRM testing going forward.

Modeling Accuracy vs. Understandability: Depending on how a statistician/researcher is trained, certain types of models may be more or less understandable. Gauging the understandability of a PRM for the field is a different concern. The researchers should consider whether the output of their models is truly understandable to administrators, supervisors, and frontline staff. Decision tree generating algorithms could be a promising way to represent the PRM in a translatable way.

Enhancing Spatial Data: The researchers explained in their report that the spatial/geographic data available in the next iteration will be enhanced. This is critical because this information will allow the researchers to understand jurisdictional differences. For PRM purposes, child welfare/ social work units may behave quite differently depending on location.