NOT FOR CITATION WITHOUT PERMISSION

Data-Mining Approaches to Suicide and Suicidal Behavior

Final Report

March 2004

Submitted for Funding to Longer Life Foundation

Rumi Kato Price, Ph.D., M.P.E.[1]

Nathan K. Risk, M.A.[1]

Edward L. Spitznagel, Ph.D.[2]

[1] Department of Psychiatry, Washington University School of Medicine, St. Louis, Missouri, U.S.A.

[2] Department of Mathematics, Washington University, St. Louis, Missouri, U.S.A.

Executive Summary

We proposed to improve prediction of suicide and suicidal behavior by employing computer-intensive "data-mining" techniques focusing on the genetic algorithms (GA), the artificial neural networks (ANN), and the tree-based regression (TBR) techniques. Although suicide is now considered preventable, predicting who will commit or attempt suicide and when such an act will occur, nonetheless, remains in the realm of intuition among clinicians or family members, thus making targeted and economical prevention difficult. So-called data-mining techniques can overcome limitations inherent in more traditional parametric-oriented approaches. The three techniques proposed in this application are complementary to each other with their combined use maximizing the strengths of each. We proposed to use three data-mining techniques to: **a)** select the most predictive measures of suicide and suicidal behavior using the GA; **b)** examine the patterns of interaction among the most predictive measures chosen by GA; **c)** maximize the predictive power of the selected measures using ANN and compare the results with those by other methods; and **d)** to examine the structure of associations among the most predictive measures using the information stored in the trained ANNs.

We utilized two large datasets available in the public domain: The National Comorbidity Survey, 1990-1992 (NCS, N=8,098) contains detailed timing information on suicidal behavior and environmental and vulnerability factors. The National Mortality Followback Survey, 1993 (NMFS93; N=22,957) includes 1% of all 1993 U.S. deaths ascertained, and 86% of the sampled deceased members followed back with informant interviews. These datasets were chosen for three reasons: large general-population samples to allow for generalization of findings; availability of sufficient numbers of suicide attempters or those who committed suicide; and availability of sufficient predictive measures.

From the NCS datafile, an "initial" dataset containing 57 variables were developed, which contained mostly dichotomous and psychiatric diagnostic variables. An "expanded" dataset containing 77 variables was developed after preliminary analyses indicated the need to refine measures. The latter dataset included many ordinal variables such as symptom counts. After variable selection analyses were completed, all 77 of the variables and the best 15 variables chosen by the GA were separately input to RPART in S-Plus, a relatively widely used tree-based regression (TBR) method. The 15 most predictive measures from each dataset were input to the Multi-Layer Perceptron (MLP) estimation, the most commonly used method in the ANN field, to assess the maximum predictive power of the selected variables. The Receiver Operating Characteristic (ROC) analysis was used to evaluate model performance of the MLP in comparison to the quadratic discriminant analysis (QDA) and logistic regression, where all analyses were cross-validated. Males and females were separately analyzed to obtain gender specific results. The analysis dataset from the NMFS93 was developed to replicate results obtained from NCS. Completed suicide was compared with the category of "deaths due to accident."

Finding Summary:

a) Best predictors of past-year suicidal thought using the NCS: When applied to both the initial and expanded datasets derived from NCS, the GA-chosen "best" predictors included many variables that are non-significant, if the traditional statistical significance were required. However, those measures chosen by both the forward selection and GA yielded $p<.01$. Using the expanded dataset, the GA tended to pick more ordinal variables compared to the forward selection. The best predictors for males and females were very similar, although not identical.

b) Interaction among predictive measures using the NCS: When GA-QDA selected variables were input, major gender differences emerged. For males, fewer variables were sufficient in constructing the best generalizable tree: depression, impairment due to substance abuse, financial problems, and a loss of daily activities remained in the final model. For females, several more predictors intricately interacted with each other, including relational strengths with relatives. The final tree was more complex than that for males.

When all 77 variables in the expanded dataset were input to RPART, the estimation yielded fewer variables after cross-validation.

c) Predictive power for past-year suicidal thought using the NCS: For both the initial and expanded datasets, predictive power was modestly improved with the GA-QDA, but the MLP further improved the predictive power by a considerable amount. Overall, MLP was able to improve the prediction by 8 to 18 % in the Area Under Curve (AUC) value when the GA-chosen variables were input. For example, the AUC=.98 was obtained for males using the NCS expanded dataset.

d) Replicating the results from the NCS using the NMFS93: With respect to the most predictive measures of suicide vs. accidental death, depression and depression-related variables were selected. Not unexpectedly, medically-related variables that may lead to reduced judgement, such as dementia symptom count, were selected to reflect their predictive power for accidental death. For this set of analyses, GA-selected variables did not vary greatly from the variables found with the forward selection. The MLP improved the prediction over QDA by a large margin (12% by the AUC value), and by a smaller margin (5%) over the logistic regression with forward selection. We found that most of the improvement from the logistic regression to MLP was due to MLP's estimation and not by GA variable selection.

e) Structure of ANN weights: Unlike regression coefficients of a parametric approach, weights on MLP paths do not have intuitive meanings. The ranking of the absolute values of weights obtained from the NCS ANN best models show that ranking of measures are more similar between males and females than would have been expected from results of TBR results. These rankings are not consistent with the importance of measures expected from logistic regression analyses run parallel to the GA variable selection runs.

Significance to the Mission of the Foundation:

Several concrete inferences can be drawn. The results of the GA analyses indicate that the current practice in medical research that emphasizes the use of diagnostic variables may undermine the ability to make accurate prediction. The patterns of interaction among predictive measures are rather different between males and females. The results suggest that clinical intervention should focus more on medication for depression and substance abuse intervention for males, but for females, more attention should be given to antisocial personality modification and improvement in social relations. Using a different numbers of predictors, we found that the TBR estimation yielded fewer variables to reach the minimum deviance, indicating that data-mining techniques such as TBR perform better when an initial effort is made to carefully select predictive measures. From the analyses attempting to show the predictive power of the epidemiologic measures, we ascertained high predictive power from our best measures (in particular, AUC=.98 for males in NCS). Such results indicate that existing detailed epidemiological measures contain sufficient information to predict past-year suicidal behavior, if appropriate variables are developed and more flexible estimation methods such as the GA and MLP are used with care.

Overall, the results from the NMFS data file were not as spectacular as those derived from the NCS. It is possible that the proxy measures by next of kin may contain a substantial amount of measurement errors. We also found that comparison of suicides with accidental death to be problematic for some analyses. A very large longitudinal dataset containing a sufficient number of suicides over many years would solve the problems we encountered with the NMFS93 dataset; however such a study would be prohibitively expensive and is very unlikely to be funded in the foreseeable future. The utility of ANN weight structure analysis is currently uncertain. Our analyses showed considerable inconsistencies with our results of linear models such as logistic regressions as well as non-linear modeling results such as the dissimilarities between the male and female TBR results.

Data-mining techniques are not a magic bullet. Researchers must work hard initially to select predictive measures, and create variables in such a way to be informative to particular techniques. Once this work is accomplished, the three methods together are likely to help improve predictive models of suicidal behavior and help us to better understand the patterns of interaction among predictive measures. Combined techniques would provide much improved prediction when sufficiently detailed predictive measures are available. With the arrival of a much faster processor, it may be possible to implement these data-mining techniques in a clinic and provide a rapid and up-to-date assessment of the patient's suicidal risk as new information is added on line to the patient chart.

Acknowledgment

Preparation of this final report was supported by the Independent Scientist Award (K02DA00221) and the Longer Life Foundation grant (LLF2001-003) to the first author. The authors thank Donald J Meyer, Tom Downey, William Shannon for technical assistance and Suzanne Culter and Angel Keller for clerical assistance.

Table of Contents

Note: Due to the length of this report, only the Executive Summary has been posted to the Longer Life Foundation Web Site. Anyone interested in obtaining more information on this study may contact Dr. Price at: price@wustl.edu.