Homework 4: Disease Prediction in Big Data

Julianne Ammirati

February 20, 2017

Prediction of Risk for psychosis

The North American Prodromal Longitudinal Study (NAPLS) Consortium is a group of researchers across American and Canadian Universities who standardized study methods across sites to acquire the largest dataset of adolescents deemed "clinical high-risk" for psychotic spectrum disorders. The purpose of this research is to identify the traits of the prodrome period of schizophrenia to allow clinicians to intervene earlier to prevent the deterioration of functioning that is a hallmark of the onset of psychosis.

Variables in the NAPLS2 Prediction Model

The authors made the decision a priori to include a maximum of eight variables in their model to ensure that each predictor had at least 10 prodromes who met that criteria. The eight variables identified were: (1) Severity of unusual thought content/delusions (P1) and persecutory thought (P2) items on the Structured Interview for Psychotic Risk Syndromes (SIPS), (2) Recent decline in social functioning as measured by the Global Functioning: Social scale, (3) Lower verbal functioning and memory as measured by the Hopkins Verbal Learning Test–Revised, (4) Slower speed of processing, per the BACS symbol coding test, (5) Age at baseline, (6) Presence of stressful life events, (7) Family history of psychosis, and (8) History of Traumas. Of these eight variables, only the first five were determined to be significant predictors of psychosis onset.

The authors adequately explain the decisions that led to the inclusion of these variables, but there are many potential predictors that were excluded in order to allow the prediction model to better translate into the clinical setting. Inflammatory markers, such as Interleukins or BDNF, have been theorized to have a role in psychosis development. Defecits in long term potentiation, as measured by a task based electroencephalograph, or changes in neurological structure and function that can be evaluated via functional MRI and diffusion tensor imaging, were also excluded from the model due to the cost of acquisition and inability for many clinics to replicate the NAPLS2 findings on these markers. Additionally, DNA and RNA analyses were completely excluded from the model.

Development of the model

The authors conducted their analyses in R, utilizing a multivariate proportional hazards model developed using the above predictors that were selected via published literature, not the NAPLS dataset itself. It seems as though this step was taken to prevent the study team from having to split the dataset for test/train purposes. The authors ruled out non-linearity of the relationship between the predictors and outcome variable using restricted cubic splines.

The predictive accuracy of the variables were tested and validated within this dataset using bootstrap resampling, and Harrell's concordance index for survival data (C-Index) was utilized to quantify the ability of the variables to discriminate the converters from the nonconverters. The authors asserted that the C-Index is analogous to an ROC curve. The model's performance was evaluated by comparing the probability of conversion from the model with the actual converter status of the participant.

While these methods of prediction are quite different from those discussed in this course, I am curious as to the authors decision to not derive predictor variables from the sample. I understand that 596 is not a huge sample size, and with converters only accounting for 15% of that group, a 90/10 or 80/20 data split would likely not provide any insight towards a meaningful model. While the easy solution would be to increase sample size, the expense of recruiting a large sample for a small proportion of converters is understandably a hurdle for this research. I would be curious as to what model for prediction can be derived from the entire NAPLS dataset (EEG, MRI, inflammatory cytokines, DNA, cortisol, etc.), which, while more difficult to replicate and utilize in a clinical setting, may provide more insight into the underlying mechanisms that motivate conversion to psychosis.

Markdown file used for word doc production is available at https://github.com/JulianneA/N741_Homework2

Article: Cannon, T. D., C. Yu, J. Addington, C. E. Bearden, K. S. Cadenhead, B. A. Cornblatt, R. Heinssen, C. D. Jeffries, D. H. Mathalon, T. H. McGlashan, D. O. Perkins, L. J. Seidman, M. T. Tsuang, E. F. Walker, S. W. Woods, and M. W. Kattan. 2016. 'An Individualized Risk Calculator for Research in Prodromal Psychosis', American Journal of Psychiatry, 173: 980-88.

Article is available from the Journal's Website using Emory proxy login:
http://ajp.psychiatryonline.org.proxy.library.emory.edu/doi/pdf/10.1176/appi.ajp.2016.15070890