Title: Using Machine Learning to Identify Patterns of Lifetime Health Problems in Decedents with Autism Spectrum Disorder

Authors: Lauren Bishop-Fitzpatrick, Arezoo Movaghar, Jan S. Greenberg, David Page, Leann Smith DaWalt, Murray H. Brilliant, Marsha R. Mailick

Introduction: As a large wave of individuals with autism spectrum disorder (ASD) diagnosed in the 1990s enters adulthood and middle age, knowledge about the patterning of lifetime health problems will become increasingly important for prevention efforts. However, although studies do suggest the presence of heightened morbidity and early mortality in ASD (Croen et al., 2015; Hirvikoski et al., 2016), we are aware of no studies conducted to date that have examined health problems throughout the life course of individuals with ASD using representative, population-level data. We sought to characterize lifetime health problems in a sample of decedents with ASD and matched controls.

Method: We retrospectively analyzed diagnostic codes associated with de-identified electronic health records (EHRs) from the Marshfield Clinic, a multi-specialty group practice with more than 700 physicians providing integrated, comprehensive care to over one million people across more than 50 locations in northern, central, and western Wisconsin. Previous research has validated EHR data from the Marshfield Clinic and found that patients are representative of northern, central, and western Wisconsin. Notably, 97% of the population in this region receives the majority of their care at the Clinic (Greenlee, 2003).

Our analysis included 91 decedents with ASD who were matched based on their sex and birth year (within 5 years) to a sample of 6,186 decedent community controls (a 1:68 ratio). Mean age of death for decedents with ASD was 67.3 (range=60-89). We used a machine learning algorithm (random forest) to classify participants into groups (ASD or control) based on their ICD-9 codes, V-codes, and E-codes using a 10-fold cross validation procedure. Information gain (IG) scores, which take into account information entropy for each class (ASD versus decedent community control) and feature (ICD-9 code, V-code, E-code), were used to measure the amount of information in each feature with respect to target class. Diagnoses related to developmental disabilities and mental health conditions (i.e., Chapter 5: Mental Disorders) were excluded from our algorithm to reduce overfitting.

Results: Diagnostic patterns distinguished decedents with ASD from matched decedent community controls with high sensitivity and specificity (AUC=0.88) and weighted average precision (98.2%) solely based on their lifetime ICD-9 codes, V-codes, and E-codes. Decedents with ASD had higher rates of epilepsy, choking, accidents, long-term medication use, cardiovascular screening, hypothyroidism, urinary tract dysfunction, motor abnormalities, skin conditions, respiratory problems, and digestive problems than decedent community controls. Decedents with ASD had lower rates of hypertension and cancer diagnosis and treatment than decedent community controls.

Discussion: This study is the first study to characterize health problems in decedents with ASD, the majority of whom were older adults, and the first to use a machine learning algorithm to differentiate decedents with ASD from decedent community controls with a high level of sensitivity, specificity, and precision based on lifetime ICD-9 codes, V-codes, and E-codes identified in EHRs. This analysis found distinctive lifetime profiles of health problems among decedents with ASD compared to decedent community controls. While preliminary, these findings have the potential to inform best practices for prevention and monitoring of health problems in ASD across diverse healthcare settings if replicated in additional population-level datasets.

References/Citations: