Derivation, Validation, and Testing of Novel Prediction Model to Identify Severe vs. Non-severe Epilepsy Patients

Chong H. Kim1, Ryan N. Hansen2, Jonathan D. Campbell1

*1 Department of Clinical Pharmacy, University of Colorado Anschutz Medical Campus*

*2 School of Pharmacy, University of Washington, Seattle, WA, USA.*

**ABSTRACT:**

*Objectives*: Dravet Syndrome (DS) is a rare and devastating epileptic encephalopathy marked by frequent and multiple seizure types. There are no standardized diagnosis codes assigned exclusively for DS or other severe forms of childhood epilepsy. We aimed to develop and validate a prediction model to differentiate severe and non-severe forms of epilepsy in medical claims data.

*Methods*: Our sample consisted of clinically confirmed DS patients (n=75), mild (i.e. childhood absence epilepsy, n=41) and other severe (i.e. infantile spasm and Lennox Gastaut, n=73) forms of epilepsy, which were split randomly to retain 1/3 for prediction modeling. The sample was drawn from Children’s Hospital Colorado requiring at least one year of follow-up. We collected data from medical claims (i.e. comorbidities, counts of procedures) and pharmacy claims (i.e. number of prescriptions filled). Continuous variables that were centered and scaled and variables that had near-zero variances were censored. Four different models (i.e. Classification and Regression Tree, Random Forest, Support Vector Machine, and Logistic Regression) were used to classify the patients as Severe (Dravet, infantile spasm, or Lennox Gastaut) vs. Mild (childhood absence epilepsy).

*Results*: The Random Forest and logistic regression algorithm yielded the highest AUC (0.856 and 0.814, respectively) compared to that of the other classification algorithms using the test cohort. The sensitivity for Random Forest and logistic regression algorithms were 0.643 and 0.714 while the specificity 0.9412 and 0.784, respectively. Using the Random Forest Algorithm, the top five variables with high variable importance scores were prescription count (100), laboratory count (57), number of chronic comorbidity conditions (54), frequency of clobazam prescribed (49) and insurance status (43).

*Conclusions*: The Random Forest algorithm made the most accurate predictions of epilepsy severity. However, logistic regression models may be more useful in predicting severe forms of childhood epilepsy in claims data.