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. However, there is no Internationally Classification of Disease version 9 (ICD-9) or ICD-10 code assigned exclusively for DS or other severe forms of childhood epilepsy. This study aimed to develop and validate a prediction model that can be used to differentiate severe and non-severe forms of epilepsy in medical claims data.

*Methods*: The study sample (n=189) consisted of clinically confirmed DS patients (n=75) along with mild (i.e. childhood absence epilepsy, n=41) and other severe (i.e. infantile spasm and Lennox Gastaut, n=73) forms of epilepsy. The sample was drawn from Children’s Hospital Colorado requiring at least one year of active retrospective follow-up in the electronic medical record. Measurements for modeling include medical claims (i.e. comorbidities, counts of procedures) and pharmacy claims (i.e. number of prescriptions filled) information. Data were pre-processed such that continuous variables that were included were centered and scaled, variables that had near-zero variances were removed, and Box-Cox transformation was applied. 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). The models were derived and validated using 124 subjects to generate the prediction models 65 subjects for model validation.

*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 identified and predicted epilepsy severity more accurately than other algorithms using medical and pharmacy claims data. However, logistic regression models may be easier to use for the purposes of predicting severe forms of childhood epilepsy within datasets where no identifying variables are available.