NAME: Beth Malow, beth.malow@vumc.org
PI: Malow, Beth
First theme choice: Systems Neuroscience

Identifying Autism in Electronic Health Records- A Multimodal Approach

Authors: Beth A. Malow, Olivia J. Veatch, Xinnan Niu, Lea Davis

Introduction: Electronic health records (EHR) are rich sources of data for conducting studies in autism spectrum disorder (ASD). While diagnostic codes are useful for defining cases, they may not capture all patients or be applied by subspecialists (sleep, GI). We examined the value of natural language processing (NLP) to accurately identify ASD in a deidentified, EHR-derived cohort of patients.

Methods: Two cohorts, currently ages 20-25 years old, were retrieved from the Vanderbilt Synthetic Derivative (SD) implemented under the Observational Medical Outcomes Partnership model. All patients had at least four notes containing ASD key terms (autism, autistic, autism spectrum disorder, Asperger, pervasive developmental disorder, ASD, or PDD). In Cohort 1 only, patients also had at least one instance of a diagnostic code for ASD.
To validate and score notes within a patient's chart, NLP algorithms were applied. Data were pulled using SQL and validated using a UNIX bash shell script embedded with KnowledgeMap Concept Indexer. Notes were scored "1" for only positive, "0" for only negated ("does not have autism"), and "0.5" for only possible ("has possible autism"). Notes were scored "N/A" if no ASD key terms relevant to the patient were identified ("brother with autism"). If a combination of ASD key terms were present in a note, the average was determined (a note with a positive and a possible would receive a combined score of 0.75). To account for potential differences in the number of available notes per patient, a final score for each patient's chart was calculated by taking the mean of the scores for each individual note. Cases of ASD were defined based on a cutoff of ≥0.8. Sensitivity and specificity were calculated based on manual chart reviews performed on all notes using a standard rubric.

Results: Cohort 1 had 337 and Cohort 2 had 44 confirmed ASD cases by chart review. For Cohort 1, NLP algorithm sensitivity was 100% for high-, and 97% for both mid- and low-evidence patients. For Cohort 2, sensitivity was 100% for mid- and 94% for low-evidence patients. Specificity was 27% for Cohort 1 and 33% for Cohort 2; this was due to a combination of factors, including sentence structure, or evolution of ASD diagnosis over time.

Discussion: Our NLP methods identified ASD charts with high sensitivity, even in the absence of ICD codes. While specificity was low, we expect to achieve higher levels through NLP refinement. Had ICD codes been used as the sole criteria for ASD diagnosis, a sizable proportion of patients (11%) would have been missed. Our findings support the use of NLP in the identification of ASD.

Keywords:
Electronic health records, Language, ICD codes