Using Deep Learning to Identify Cognitive Impairment in Electronic Health Records

Tanish Tyagi, Colin Magdamo, Ayush Noori, Mayuresh Deodhar, Zhuoqiao Hong, Dmitry Prokopenko, Rudy E. Tanzi, Deborah Blacker, Bradley T. Hyman, Shibani S. Mukerji, M. Brandon Westover, Sudeshna Das

Abstract

Background: Dementia is a neurodegenerative disorder that causes mental decline and affects more than 50 million people worldwide [1]. Dementia is under-diagnosed by healthcare professionals—only one in four people who suffer from dementia are diagnosed [2]. Often, a diagnosis is given by the time a patient has reached moderate dementia, and irreversible damage has already been done to the brain. A timely dementia diagnosis is crucial for patients to receive relevant treatment. Information relevant to cognitive impairment is often found within EHR records. Manual review of clinician notes by experts is both time consuming and often prone to errors (1 in 5 patients are misdiagnosed) [2]. Automated evaluation of these notes presents an opportunity to label patients with cognitive impairment who could benefit from an evaluation or from referral to specialist care, providing an incipient diagnosis and combating underdiagnosis.

Methods: We selected a cohort of patients from the Mass General Brigham (MGB) healthcare system who were older than 60 years (as of July 13, 2021), had at least one unique encounter with a match of a keyword pertinent to CI, and had *APOE* genotype data available from the BioBank (N=16,428). We extracted unstructured clinician notes, identified matches to dementia-related keywords, and constructed 800-character sequences from these matches. These sequences were then annotated by neurologists using a web-based annotation tool as 1) Yes, i.e., patient has CI; 2) No i.e., Patient does not have CI; and 3) Neither i.e., sequence has no information on patient's cognition. We performed TF-IDF (term frequency-inverse document frequency) vectorization on the annotated sequences and selected features based on a term's Pearson correlation coefficient with the outcome. Regularized logistic regression was applied with the annotated cognitive impairment labels. We used different correlation coefficients as thresholds to select features and iterated over different lambda values to determine the optimal lambda value and correlation coefficient threshold.

Results: The regularized logistic regression model was trained on a dataset comprised of 8,362 annotated sequences from (N=2,417) unique patients using 10-fold Cross Validation stratified across the three cognitive impairment labels. It achieved an area under the receiver operating characteristic curve (AUROC) of 0.94, accuracy of 0.90, sensitivity of 0.73 and specificity of 0.94 on the held-out test set (353 annotated sequences from N=77 unique patients) with a lambda value of 10 and a correlation coefficient threshold of 0.07. We then applied our model on the other 186,730 sequences from (N=13,941) unique patients that were not part of the training loop.

Conclusion: We developed a machine learning tool to identify potential patients with cognitive impairment. Our work can help combat the issue of underdiagnosis for dementia, as our model was able to accurately identify cognitive impairment for patients that would have normally been undiagnosed despite requiring medical attention. Our model is significantly more accurate and efficient than current clinical methods as it provides diagnoses in real time with 90% accuracy. Future plans include gathering more annotated sequences and using deep learning natural language processing (NLP) techniques (currently in development).

References:

1. https://www.alzint.org/about/dementia-facts-figures/dementia-statistics/
2. https://alzheimersnewstoday.com/alzheimers-disease-statistics/