An EHR-based Cohort Discovery Tool for Identifying Probable AD

Donna Tjandra[1], Raymond Migrino[2], Bruno Giordani[3], and Jenna Wiens[1]
[1]Division of Computer Science and Engineering, University of Michigan, Ann Arbor, MI
[2] Phoenix Veterans Affairs Health Care System, Phoenix, AZ
[3]Neuropsychology Section, Department of Psychiatry, University of Michigan, Ann Arbor, MI

Background.
Electronic health records (EHRs) contain decades of longitudinal clinical data on hundreds of thousands of potentially at-risk individuals for Alzheimer's disease (AD). The ability to automatically identify probable AD patients within EHRs would facilitate downstream computational analyses on such large-scale datasets, by eliminating the need for labor intensive chart review. To this end, we developed and validated a cohort discovery tool that can be applied to EHR data for automatic classification of individuals with AD.

Methods.
We extracted EHR data from Michigan Medicine's Research Data Warehouse (RDW) pertaining to Michigan Alzheimer's Disease Center (MADC) participants with a consensus-based diagnosis ranging from cognitively normal to probable AD. We investigated the accuracy of different EHR-based rules for identifying patients with AD. Rules were based on combinations criteria pertaining to ICD diagnoses, medications, laboratory results and encounter types. Applied to data from the RDW, these rules were evaluated against MADC diagnoses (Figure 1), in terms of sensitivity, specificity, and positive predicted value (PPV). To optimize for the probability that patients identified by the rule have AD, we prioritized PPV when ranking different rules.

Results.
MADC and RDW records overlapped in 624 patients 65 years and older. Though a diagnostic code for AD alone resulted in relatively low specificity, in combination with an encounter involving medium/high complexity medical decision making it resulted in increased specificity and the highest PPV (Figure 2). This rule yielded a PPV, specificity, and sensitivity of 0.82 (95% confidence interval (CI) 0.75-0.87), 0.95 (95%CI 0.93-0.97), and 0.65 (95%CI 0.60-0.68) respectively. For true positives, the first RDW diagnosis of probable AD occurred on average three years before the first MADC diagnosis (95% CI 2.3-3.7 years). Applied to the entire RDW, the algorithm identified 4,152 patients with probable AD.

Conclusions.
EHR-based criteria can automatically and accurately identify patients with probable AD. Applied to large longitudinal EHR datasets, these labels can be used for downstream analyses, e.g., modeling patient trajectories of disease progression.

Figure 1: Comparing MADC and RDW encounters for a sample patient. Each row represents a timeline for the respective dataset, and encounters are indicated with squares. Shading along the MADC timeline indicates consensus-based diagnoses. A patient is considered to have probable AD six months prior to their first MADC encounter labeled as probable AD and anytime afterward. EHR-based criteria are applied to the RDW encounters. The encounters that occur on or after the first encounter that meets the criteria are labeled as probable AD. A true positive is counted if at least one predicted AD RDW encounter overlaps with the MADC defined probable AD window (e.g., the encounters in the orange circles).

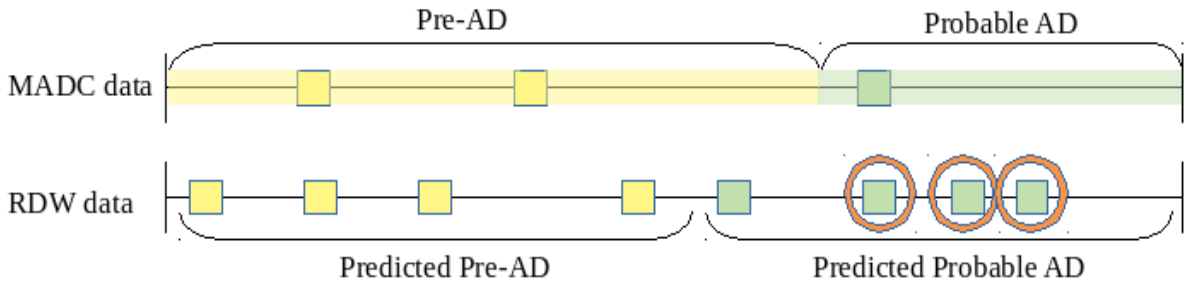

Figure 2: EHR-based rules with a PPV greater than 0.5 for identifying AD. Error bars correspond to 95% confidence intervals over 1,000 bootstrapped samples. Complexity in medical decisions is measured by the amount and variety of patient data examined by a physician, patient risk, and treatment options, as defined by CPT codes 99214 and 99215.