Mild Cognitive Impairment Prediction and Diagnosis Procedure Recommendation using Machine Learning

Sajjad Fouladvand, PhD1, Morteza Noshad, PhD1, Jonathan H Chen, MD, PhD1,2

1Stanford Center for Biomedical Informatics, Stanford University, Stanford, CA, USA;

2Division of Hospital Medicine, Stanford University, Stanford, CA, USA;

Abstract

# **Introduction**

[WHAT IS MCI AND WHY IT IS IMPORTANT TO PREDICT IT AND PREDICT DIAGNOSIS PROCEDURE]

Early detection of cognitive decline may be critical to the efforts to stop dementia progression, including Alzheimer’s disease (AD) and AD-related dementias (ADRD)(1). Dementia and mild cognitive impairment (MCI) are under-diagnosed even though dementia is one of the major causes of mortality and morbidity in older people worldwide. Early detection of cognitive decline may be critical to the efforts to stop dementia progression [https://grants.nih.gov/grants/guide/rfa-files/RFA-AG-20-051.html]. A recent research study based on the Medicare data showed that 85% of people first diagnosed with dementia were diagnosed by a non-dementia specialist physician, usually a primary care doctor, and an “unspecified dementia” diagnosis was common [<https://news.usc.edu/160355/dementia-diagnosis-usc-study-specialty-care/> ]. Among those diagnosed by a non-dementia specialist, 33% of patients were given a diagnosis that lacked a specific type of dementia, compared to 22% of patients diagnosed by a specialist. Diagnosis often occurs late in the illness trajectory. Inefficient identification of dementia type leads to low-value care including ineffective and burdensome treatments and procedures.

[WHY IT IS IMPORTANT TO PREDICT MCI AND DIAGNOSIS PROCEDURE WITH AI INSTEAD OF TRADITIONAL METHODS AND SCREENING]

[LITRATURE REVIEW ON MCI PREDICTION USING AI]

[WHAT DID WE DO IN THIS WORK]

# **Materials and Methods**

## Data

## [GENERAL INFORMATION ABOUT SHC\_CORE DATA]

## The cases in this study include patients with at least one ICD diagnosis of MCI (ICD10s=G31.84, F09 and ICD9s=331.83, 294.9) and two years of data availability. Controls are patients with no ICD diagnosis in their records and two years of data availability. Data availability is the time between patients first record date to an index date. The index date for cases is the first MCI diagnosis date and for the controls it’s one year before their last record in the data. Controls are matched with cases based on age and gender to create a balanced data set including 9,274 patients with equal number of cases and controls. Note, our models are trained using balanced training dataset to predict MCI one year prior to the index date for cases and controls; however, our testing experiments include both balanced and imbalanced scenarios and the MCI prediction models are tested on imbalanced test sets too.

## Cohort selection

The dataset includes n patients . For a patient their diagnosis, medication, procedure and demographic records. Records forare in a format where is the timestamp for and indicate medications, diagnosis or procedures codes for this patient at . We converted this longitudinal format to a stationary format for each patient by computing the frequencies of the features across all patient’s timestamps and created a stationary dataset where is a vector including frequencies of each medication, diagnosis and procedure features concatenated with patient demographic features, age, gender and race, and is the target variable. For the MCI prediction is a binary variable indicating if will develop MCI in a year or not and for the MCI diagnosis procedure recommendation system is a multi-hot vector indicating the recommended procedure for . Note, the prediction window for MCI prediction task is one year and for the procedure recommendation task is two months. Table 1 describes statistics of our final dataset. Demographic features in terms of age and gender were similar among cases and controls as we matched based on these variables. The average age of cases and controls were 69.61 years (25th and 75th percentiles = 62, 82) and 71.72 years (25th and 75th percentiles = 66, 83), respectively. The majority were female (54.43% among both cases and controls), cases had a higher number of patients with races as black, native American and white in cases than in controls and lower number of Asian patients in cases than controls. Prevalence of the top-5 medications, diagnoses and procedures selected by a random forest model are also presented in Table 1.

**Table 1**. Patient Characteristics among Cases and Controls.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Variable** | **Case** | **Control** | **Variable** | **Case** | **Control** |
| Age | a69.61  (62, 82) | 71.72  (66, 83) | Female | 2,524 (54.43%) | 2,524 (54.43%) |
| Race | | | Pacific Islander | 35 (0.75%) | 36 (0.78%) |
| Asian | b544 (11.73%) | 719 (15.50%) | White | 3,091 (66.66%) | 2,672 (57.62%) |
| Black | 262 (5.65%) | 167 (3.60%) | Unknown | 140 (3.02%) | 435 (9.38%) |
| Native American | 17 (0.37%) | 13 (0.28%) | Other | 548 (11.82%) | 595 (12.83%) |
| **Diagnoses** | | | Medical examination/evaluation | 2,233 (48.16%) | 993 (21.41%) |
| Nervous system signs and symptoms | 1,459 (31.46%) | 390 (8.41%) | Musculoskeletal pain, not low back pain | 1,944 (41.92%) | 833 (17.96%) |
| Exposure, encounters, screening or contact with infectious disease | 1,655 (35.69%) | 625 (13.43%) | Sleep wake disorders | 1,346 (29.02%) | 492 (10.61%) |
| **Medications** | | | SELECTIVE SEROTONIN REUPTAKE INHIBITOR (SSRIS) | 915 (19.73%) | 263 (5.67%) |
| ANTICONVULSANTS | 1,204 (25.96%) | 408 (8.80%) | OPIOID ANALGESIC AND NON-SALICYLATE ANALGESICS | 1,821 (39.27%) | 839 (18.09%) |
| ANTIHYPERLIPIDEMIC-HMGCOA REDUCTASE INHIB(STATINS) | 1,935 (41.73%) | 1,063 (22.92%) | ANTIEMETIC/ANTIVERTIGO AGENTS | 1,884 (40.63%) | 861 (18.57%) |
| **Procedure** | | | CBC WITH DIFFERENTIAL | 2,860 (61.68%) | 1351 (29.13%)) |
| METABOLIC PANEL, COMPREHENSIVE | 2,844 (61.33%) | 1,219 (26.29%) | External Lab Results | 42 (0.90%) | 572 (12.33%) |
| TSH | 2,196 (47.36%) | 756 (16.30%) | OTHER ORDER SCANNED REPORT | 2,123 (45.78%) | 925 (19.95%) |

a Numbers are in V(x, y) format, where V is the average and x and y are 25th and 75th percentile, respectively. b Numbers are in N (p%) format, where N is the number of patient and p% shows the percentage in the cohort.

Predictors in this study include medications, diagnosis, procedure and demographic features for each patient. Note, we grouped medications and diagnosis using pharmaceutical class and clinical classification software (CCS) codes, respectively. We included top-100 most frequent medications, diagnoses and procedures codes as well as the demographic features. Therefore, the feature set, for for example, includes 300 medication, diagnosis and procedure codes plus 3 demographic feature, age, gender and race. This feature set size then further reduced to 30 features using random forest algorithm. These predictors were used to train logistic regression, random forest and xgboost. These models were tuned on the train set including 6,491 patients’ data and a randomized parameter search with a 5-fold cross validation. The trained models were tested held out balanced and imbalanced test sets including 2,783 and 91,729 (with 1,391 cases and 90,338 controls) patients’ data. The optimum logistic regression has a sag as it’s solver with no penalty and . The optimum random forest model has 800 estimators with maximum depth of 110. The optimum values for the number of estimators, maximum depth, learning rate and gamma for the xgboost model used are 600, 8, 0.01, and 0.01. The optimized models were assessed using different methods including accuracy, precision, recall, F1-score and ROCAUC.

# **Experimental Results**

**Table 2.** Performance of MCI prediction using machine learning on unseen test sets. Balanced test set

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Model** | **Accuracy** | | **Precision** | | **Recall** | | **F1-score** | | **AUC** | |
| aBLN | bIMB | BLN | IMB | BLN | IMB | BLN | IMB | BLN | IMB |
| Logistic Regression | 0.72 | 0.63 | 0.77 | 0.03 | 0.64 | 0.64 | 0.70 | 0.05 | 0.79 | 0.70 |
| XGBoost | 0.77 | 0.53 | 0.79 | 0.02 | 0.74 | 0.74 | 0.77 | 0.05 | 0.85 | 0.71 |
| Random Forest | 0.77 | 0.48 | 0.78 | 0.02 | 0.77 | 0.77 | 0.78 | 0.04 | 0.86 | 0.70 |

a BLN shows the performance of the model on a balanced test set. b IMB shoes the model performance on imbalanced test set.

**Table 3.** Machine learning prediction of top-10 procedures for MCI patients. NOTE, THIS TABLE IS GOING TO BE UPDATED SOON.

|  |  |  |  |
| --- | --- | --- | --- |
| **Targets** | **precision** | **recall** | **f1-score** |
| 0 | 0.55 | 0.03630363 | 0.06811146 |
| 1 | 0.64527027 | 0.35834897 | 0.46079614 |
| 2 | 0.705882353 | 0.04411765 | 0.08304498 |
| 3 | 0 | 0 | 0 |
| 4 | 1 | 0.00283286 | 0.00564972 |
| 5 | 0 | 0 | 0 |
| 6 | 1 | 0.01310044 | 0.02586207 |
| 7 | 0.649805447 | 0.31568998 | 0.42493639 |
| 8 | 0.7 | 0.02734375 | 0.05263158 |
| 9 | 0.743589744 | 0.09477124 | 0.16811594 |
| micro avg | 0.65474339 | 0.12722877 | 0.21305668 |
| macro avg | 0.599454781 | 0.08925085 | 0.12891483 |
| weighted avg | 0.615008944 | 0.12722877 | 0.17723047 |
| samples avg | 0.133998134 | 0.06367978 | 0.07901215 |

**DISCUSSION**

# **Conclusion**

# [SUMMARY OF THIS WORK. LIMITATIONS. FUTURE WORK]

# **References**

1. Rasmussen J, Langerman H. Alzheimer’s Disease – Why We Need Early Diagnosis. Degener Neurol Neuromuscul Dis. 2019 Dec 24;9:123–30.