Developing a Computational Model to Diagnose Patients with Dementia

Christina Vu Advisor: Jue Wang
Department of Mathematics, Union College

Background

Numerous individuals in the aging population do not find out that they are susceptible to neurodegenerative diseases until it is far too late. One of the more irreversible symptoms of these diseases is dementia. The time and cost involved in diagnosing a single patient may not be the most effective way to determine whether an individual is at risk. Dementias have costed the U.S. $305 billion in 2020 and that number could rise up to $1.1 trillion by 2050. However, could there be a simpler and faster method to diagnose patients using only basic clinical information? What is the best method for identifying and diagnosing? These were the questions I aimed to discover more about during my research.

Introduction

Clinical patient information is usually stored in the hospital's or office's electronic health records (EHR), a healthcare data depository. These databases are usually large and unsystematic as information filing has changed throughout a hospital's career. There has been numerous experimentation in using these large healthcare data depository to build predictive models that can predict and diagnose patients who are susceptible to dementia. These models utilize data that is routinely collected during primary care visits and therefore would be more efficient to implement and automate in calculating risk and diagnosing. In order to employ the databases, appropriate information has to be cleaned, extracted, and organized to ensure systematic performance for the models. Many of the prognostic and diagnostic models for dementia-related diseases, such as Alzheimer's, use machine learning, e.g. logistic regression or statistical methods, e.g. Cox regression. Due to the differences in databases and methodology, it is difficult to determine the best way to identify any one disease. My research project seeks to extract relevant information and explore the best approach in designing a model to diagnose patients who are susceptible to dementia.

Methodology

Study Population:

- Used MIMIC-III database for study population
- Contains de-identified clinical records stored as tables which include demographics, lab results, clinical diagnoses and procedures as defined by the International Classification of Diseases ver. 9 (ICD-9), and many other important factors
- 38,597 adult patients were admitted into the critical care unit of a single hospital between 2001 and 2012

ICD-9 Codes:

- The ICD-9 codes used to extract patients who had diseases with dementia as a major symptom
- The patients were categorized by ICD-9 codes: 290, 294, and 331.
- Examples of some of the diseases and their respective ICD-9 codes are:
  - Presenile dementia (290.1)
  - Vascular dementia (290.41)
  - Mixed dementia (294.2)
  - Alzheimer's disease (331.0 and 294.1)
  - Frontotemporal dementia (331.19 and 294.1)

Extracting Relevant Information:

- Information regarding the patients' demographics (ethnicity, sex, age), other medical conditions, vital signs, laboratory tests and medication were needed
- Used PostgreSQL, a relational database management system, to extract relevant information from the tables
- Organized the information by joining and exporting tables of the variables we wanted to extract
- Each patient possessed a unique subject identifier which was utilized to follow the patient's chart events and admissions
- The vital signs used were height, weight, and blood pressure
- The average height and average weight were used to calculate the body mass index (kg/m^2)
- Age was calculated from the patient's admission times and for analyzing, the minimum was used

Data Analysis:

- Will clean data by imputing missing values with study population mean
- Will split data 70% for training and 30% for testing

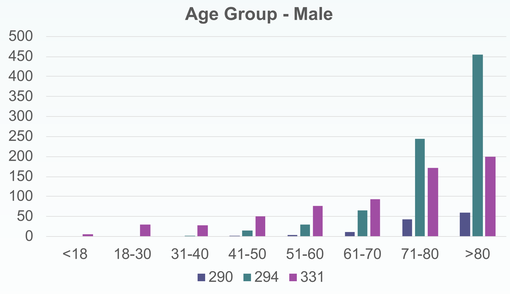
Results

Chi-Squared Test for Independence:

- There is a higher percentage of females overall.
- Compared to the other categories, the 290 Dementias group has a higher population of African Americans (17.5% vs 11.0% and 7.7%, $p<0.0001$).

Table 1 Example of Extracted Patient Information

Patients	Subject_id	10027
Admissions	Hadm_id	199395
Icustays	Icustay_id	286020
Patients	Gender	F
Admissions	Ethnicity_grouped	White
Patients/Admissions	First_admit_age	82
Chartevents	Height_cm	157.48
Chartevents	Weight_kg	46.8
	BMI	18.87
Diagnoses_icd	Diagnoses	41401, 4019, 25000, 4240, 2948, 5691, 79092, 4280, 3970, 4266, 42731, 5859
Chartevents	Baseline systolic blood pressure	131.2



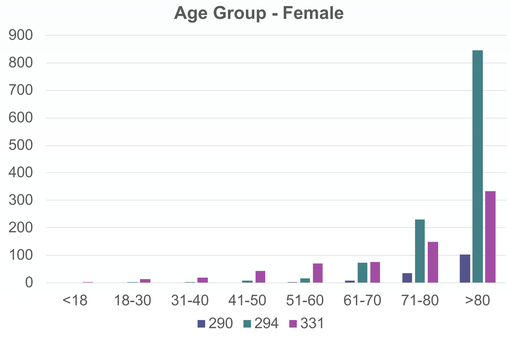


Table 2 Characteristics of the study population (n = 38, 597)

ICD9 Codes	290	294	331
	% (N)	% (N)	% (N)
Total	268	1986	1360
Gender			
Female	55.2% (148)	59.2% (1176)	51.9% (706)
Male	44.8% (120)	40.8% (810)	48.1% (654)
Ethnicity			
White	72.0 % (193)	75.3% (1495)	72.1% (981)
Black	17.5% (47)	11.0% (219)	7.7% (105)
Hispanic	3% (8)	2.1% (41)	4.7% (64)
Asian	3% (8)	2.4% (48)	3.4% (46)
Native	0	0.3% (5)	0.1% (2)
Other	4.4% (12)	9.0% (178)	11.9% (162)
	Median	Median	Median
Age	83	84	77

Future Works

- Build multivariable logistic regression models using different variables:
  - Only demographics and vital signs as variables
  - Using ICD-9 diagnoses (the top-20 variables) and demographics as variables
  - Laboratory tests and demographic variables
- Evaluate further with other machine learning algorithms such as a) decision tree and b) random forest.
- Compare models and evaluate using 10-fold cross-validation
- Assess the models' performance and discrimination using:
  - Sensitivity (true positive rate)
  - Specificity (true negative rate)
  - Area under the receiver operating characteristic

Acknowledgements

Union Undergraduate Research Fellowship and Davenport Research Fellowship

References

1. Chen Z, Bird VY, Ruchi R, Segal MS, Bian J, Khan SR, Elie M and Prosperi M. "Development of a personalized diagnostic model for kidney stone disease tailored to acute care by integrating large clinical, demographics and laboratory data: the diagnostic acute care algorithm - kidney stones (DACA-KS)" \\ BMC Medical Informatics and Decision Making, vol. 18, No. 72, 2018.
2. "Facts and Figures." *Alzheimer's Disease and Dementia*, www.alz.org/alzheimers-dementia/facts-figures#:~:text=More%20than%205%20million%20Americans%20of%20all%20ages%20have%20Alzheimer's,are%20age%2075%20or%20older.
3. Johnson A, Pollard T, Shen L. *et al.* MIMIC-III, a freely accessible critical care database. *Sci Data* 3, 160035 (2016). https://doi.org/10.1038/sdata.2016.35