Plan B Working Draft

Denis Ostroushko

Introduction

Background

Alzheimer's disease (AD) stands as a prevailing public health challenge, with an escalating impact on global societies. Characterized by progressive cognitive decline, AD places an immense burden on individuals, families, and healthcare systems. Mild Cognitive Impairment (MCI), often considered an intermediate stage between normal aging and AD, has become a focal point for researchers seeking to understand the early stages of neurodegeneration. The ability to accurately identify and track the progression from MCI to AD is pivotal for developing timely interventions and personalized treatment strategies.

In recent years, the advent of electronic health records (EHRs) has revolutionized the landscape of medical research, offering a wealth of longitudinal data that can illuminate patterns, risk factors, and potential predictors associated with the transition from MCI to AD. The integration of EHRs into scientific inquiry provides a unique opportunity to delve into the intricacies of disease progression, drawing upon real-world patient data in diverse clinical settings.

This paper aims to explore and critically examine the methodologies employed in studying the progression from MCI to AD using EHRs. By leveraging the wealth of information encapsulated in electronic health records, researchers gain access to comprehensive patient histories, enabling a nuanced understanding of the multifaceted factors contributing to the evolution of cognitive impairment. The utilization of EHRs not only offers a large-scale and diverse dataset but also facilitates the identification of potential biomarkers, risk factors, and temporal trends that may inform the development of predictive models for early AD detection.

Motivation

While predictive models utilizing regression, machine learning, and deep learning have become widespread in analyzing progression from MCI to AD, researchers acknowledge key challenges in applying these techniques to real-world electronic health records (EHRs). Although EHRs contain extensive longitudinal patient information, the data is inherently messy and heterogeneous across healthcare systems, posing obstacles for analysis. In particular, the diagnoses codes used to identify onset of MCI and early stages of AD can be inconsistent or incomplete. Methods to mitigate these issues

often require labor-intensive data cleaning and transformation to create usable datasets. Moreover, a set of decisions need to be made to create key identifying variables. As such, recent studies have focused efforts on developing robust data pipelines to harmonize variables from diverse EHR systems relevant to MCI and AD progression. This encompasses mapping diagnosis codes over time, extracting cognitive assessment scores, reconciling discrepancies in demographic information, and handling missing data. Additional work has centered on novel feature engineering and advanced natural language processing of clinical notes to augment more structured symptoms data available. By confronting the underlying complexity of real-world health data, predictive models applied to EHRs can become better equipped to uncover novel predictors and trajectories associated with progression from MCI to AD. However, barriers persist due to inherent variability across patients and care settings encapsulated in these records. Ongoing research on representation learning methods that can integrate heterogeneous data sources offers promise in this emerging area. Variation of all mentioned factors results in different cohort definitions.

For example, Aguilar et al. (2023) defined Alzheimer's disease (AD) and mild cognitive impairment (MCI) using clinical notes from the United States Veterans Affairs Healthcare System (VAHS) electronic health records (EHR). For MCI, the keywords "MCI" and "mild cognitive impairment" were selected, and for AD, the keyword "Alz∗" was used. Diagnostic codes related to cognitive impairment were also considered, such as ICD-9-CM '331.83' and ICD-10-CM 'G31.84' for MCI, and ICD-9-CM '331.0' and ICD-10-CM 'G30∗' for AD. Veterans had to be at least 50 years old to enter the study cohort.

Mattke et al. (2023) used Medicare Advantage population with an age cutoff of 65 years to construct their sample. Similar to the previous study, they used ICD9 and ICD10 diagnoses to get key identifying variables. However, authors identified the diagnosis based on ICD-10-CM code G31.84 (mild cognitive impairment of uncertain or unknown etiology) and the ICD-9-CM code 331.83 (mild cognitive impairment). Similarly, for ADRD they required two claims on separate days for the diagnosis of MCI. This approach poses less trust in the consistency of EHR records and introduces more criteria to make the sample more robust to false findings, in our opinion.

Xu et al. (2023) used longitudinal EHR records for identifying outcome-oriented progression pathways from MCI to AD. In their study, AD identification is based on ICD codes, specifically ICD 9 codes 331.0 and ICD 10 codes G30.*, excluding those with AD diagnosis before MCI. Similar to previous studies, MCI diagnosis is based on ICD codes, including ICD 9 codes 331.83 and 294.9, and ICD 10 codes G31.84 and F09. However, criteria include at least one year of data before and after the MCI onset, and a conversion time to AD of more than half a year. It appears that the authors of this study wanted to make sure that the an event of interest was observed relatively early to a potential time of censoring.

Methods

Data Collection

Data for the study were obtained from the Fairview Health System Data Warehouse. Records with visits to any Fairview Facilities with ICD diagnoses identifying presence of MCI, AD, or Unspecified Mental Disorder. Patient records for visits due to these reasons spanned between 2004 and 2020. To obtain full available timeline of observations for these patients we collected data on all visits, relevant or irrelevant to MCI/AD visits. We are interested in progression times to AD, therefore, including maximum possible observation times is beneficial to the analysis, especially for those who do not progress to AD from MCI, as such variable for time until progression will reflect the reality more accurately and will result in estimates that should be less biased.

The key step of data engineering for this study involves the definition of what a 'visit' is. A visit where a patient is diagnosed with MCI or AD is one unique date when at least one of ICD codes for MCI, Unspecified Mental Disorder, or AD occurred. As mentioned in the introduction section, while EHR data comes from a structured data base, translation of clinical notes into the structured data type results in a messy dataset. We were able to gather data on XXX patients. To study transition times and rates from the initial occurrence of MCI in the EHR to the first occurrence of AD we continued to clean the data and removed patients who had events of AD recorded before MCI, and those who had only events of AD recorded. This resulted in a loss of YYY patients from the pool of patients that can create a cohort for the study of progression rates.

These steps provided a data set where all potential patients that one can use to form a study population had MCI or Unspecified Mental Disorder.

Cohort Definitions

After reviewing the literature on work using EHR, Healthcare Claims, and other types of digital health records, we identified common criteria that is used to create cohort or study populations to develop predictive model and statistical models for inference. In this study, we consider three common types of population. As mentioned previously, were were able to obtain data on approximately XXX patients. In the process of refining the restrictions and rules that help us construct this study population, we applied general filtering criteria to all cohorts. First, we further impose restrictions on what defines an AD diagnosis. We remove alcohol induced, Lewy body dementia, and other similar diagnosis groups from all cohorts, making it such that only ICD codes 'G30*' and '331' are eligible AD diagnoses. This step allows us to form a group of patients for the study who have a homogeneous outcome variable. Second, we remove those patients who have only 1 Unspecified Mental Disorder diagnosis. In the process of cleaning available EHR data, we found that those with one Unspecified diagnosis were much more likely to be younger adult. Since dementia and Alzheimer Disease are age-related, removing younger individuals made sense. Further research and documentation of this diagnosis code confirmed that this diagnosis

code is a 'catch-all' for cases where a patient exhibits cognitive dysfunction or mental impairment, and it is attributed to a known physiological condition, but the healthcare provider does not specify the exact nature of the mental disorder within the available coding options. We made a decision that including these patients into the study does not help form a population of patients who are likely to progress. Rather, such patients likely have other conditions that are further evaluated and documented in the EHR. Validation of this speculation was outside the scope of this study. Therefore, including such patients is more likely to results is lower rates of progression and biased estimates of progression times.

Using these general criteria, we identify YYY patients who have at least two diagnoses on separate dates, meaning that such patients either have at least two MCI diagnoses, or one MCI and one AD diagnoses. At this point, we define three cohorts that we will evaluate, and label them "Cohort 4", "Cohort 5", and "Cohort 6". These names remained after data cleaning steps.

Cohort 4: we impose a filtering criteria that all people at the date of their first MCI or Unspecified diagnosis are at least 50 years old. This age restriction is common in the literature, therefore we had to include it.

Cohort 5: this cohort restricts the timeline of MCI diagnoses that take place before a potential progression. For those who do not end up progressing to AD, we impose a rule that the time difference between the first and the last diagnoses of MCI or Unspecified diagnosis must be at least 50 days. This time lag allows to select patients that have been seen on multiple occasions over a prolonged period of time and had confirmed or consistent MCI diagnoses. This criteria helps tp filter out sporadic cases where a person might have a long inpatient admission and multiple Unspecified or MCI diagnoses were recorded and submitted to the system on separate days that were only a small period of time apart. For those patients who have an event of AD diagnosis, we do not apply this criteria. Our motivation is to include as many people with AD diagnosis as possible for a potential study. Additionally, for predictive modeling purposes, which is a common topic of research and publication, we want to include as much data as possible for people with a desired target, or outcome, variable. We also want to see how the results of estimated progression rates compare to cohort 6.

Cohort 6 applies criteria of at least 50 day time lapse between first and last diagnoses to all patients in the cohort. This restriction takes away sample size, and the number of AD events. However, it also focuses on those patients that have confirmed consistent diagnosis of MCI leading up to a possible progression to AD.

All three cohorts are valid target populations to study progression rates from MCI to AD using EHR records, and have been commonly cited in the literature focusing on modeling progression pathways. In the next section we evaluate primary data summary statistics, and discuss differences between the cohorts.

Table 1: Comparison of statistics for the three cohrots

	Cohort 4	Cohort 5	Cohort 6
N Patients	5,711	2,807	2,435
N Progressed	743	743	371
% Progressed	13.01%	26.47%	15.24%
Avg. Unspecified Diags. (SD)	1.19(2.83)	2.07(3.73)	2.29(3.95)
Avg. MCI Diags.	1.66(2.75)	2.38(3.75)	2.65(3.95)
Avg. AD Diags. (SD)	0.68(2.97)	1.38(4.12)	0.76(3.02)
Avg. Age at Start (SD)	74.77(11.78)	74.86(11.6)	74.18(11.8)
Avg. Age at Progression (SD)	80.89(8.9)	80.89(8.9)	80.79(8.66)
Avg. Years to Progression (SD)	1.98(1.91)	1.98(1.91)	2.29(1.96)
Avg. Total Obs. Years	2.66(2.69)	3.67(2.93)	3.69(3.01)
Avg. Total Years. After Last MCI	1.91(2.34)	2.17(2.64)	1.95(2.61)
Avg. Total Years. After Last AD	0.85(1.32)	0.85(1.32)	1(1.44)
N with MCI as last ever Diag.	571 (10%)	228 (8.1%)	228 (9.4%)
N with AD as last ever Diag.	143 (2.5%)	143 (5.1%)	59 (2.4%)

Data

Cohort Summaries

Table 1 compares keys statistics of interest for the three cohorts. We can see that these cohorts are similar in terms of all key summary statistics outside of sample size and event rates. Om average, this is a very old population of patients. Variance of ages appears approximately equal between cohorts. Duration to progression and age at progression appear quite similar as well. Additionally, age at progression implies that within these cohorts those who are relatively older are more likely to progress. Recall that as we create later version of cohorts, we require more time lag between first and last MCI or Unspecified diagnosis. This is likely why we see that for later version of cohort the average number of diagnoses increases. Note the difference between Cohort 4 and 6 in terms of progression rate and observation time after the last diagnosis. Cohort 6 has a slightly higher progression rate, but also a higher window of time after the last MCI diagnosis, which allows us to capture these events.

Figure 1 displays the distribution of age for the three cohorts at the time of initial MCI diagnosis.

Imposing a criteria of having at least two MCI or Unspecified diagnosis over at least 50 days window drastically changes the population that creates a study sample. As we discussed previously, MCI and Unspecified diagnoses are quite different in their nature. Unspecified diagnosis can be used as a 'catch-all' diagnosis, and therefore younger patients can be much more likely to be diagnosed with 'Unspecified' ICD-10 code.

Commonly Unspecified mental disorder, or unspecified mental disorder of unknown origin, is used to identity members who potentially have MCI. Our data summary shows that such patients are more likely to be younger. It is likely the case that physicians are more reluctant to give an MCI diagnosis to those who are younger, and perhaps require a follow up for those who are initially given an unspecified diagnosis.

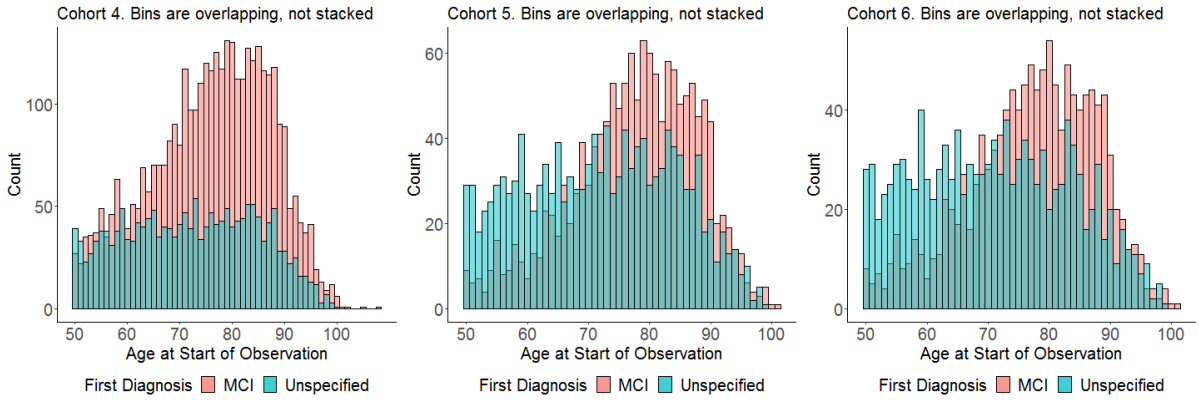

Figure 1: Breakdown of Age distribution

Kaplan Meir Curves

The next essential step of preliminary data analysis is the look at the Kaplan-Meier survival curves. Analogous to Figure 1, we use first diagnosis, either MCI or Unspecified, as a stratifying variable to assess the difference in the time-to-event distribution between the two groups. Figure 3 present the results for all three cohorts. We can immediately observe that the patterns across cohorts are similar to the age distribution conclusions. Cohort 4 had a very similar distribution of ages between those who had MCI and Unspecified as their first diagnosis. It should be of no surprise that age is a more powerful predictor of progression from MCI to AD. Since ages were distributed visually similarly across the two groups, observed difference in age distributions translates equivalently into the time-to-AD progression estimates.

It is of note that the overall progression rates for Cohort 4 and 6 are quite similar presumably to the overall similar progression rate observed in the data, but the difference between the MCI/Unspecified subgroups gets greater. Within the fist 7-9 years of follow-up cohorts 4 and 6 overall are indistinguishable, while the picture changes drastically when the start stratifying by the initial, or anchor, diagnosis. This key distinction, paired with the largely different distribution of the ages implies that while the cohorts look similar on paper, the make up of the two sub-populations that make up the overall cohort are quite

different.

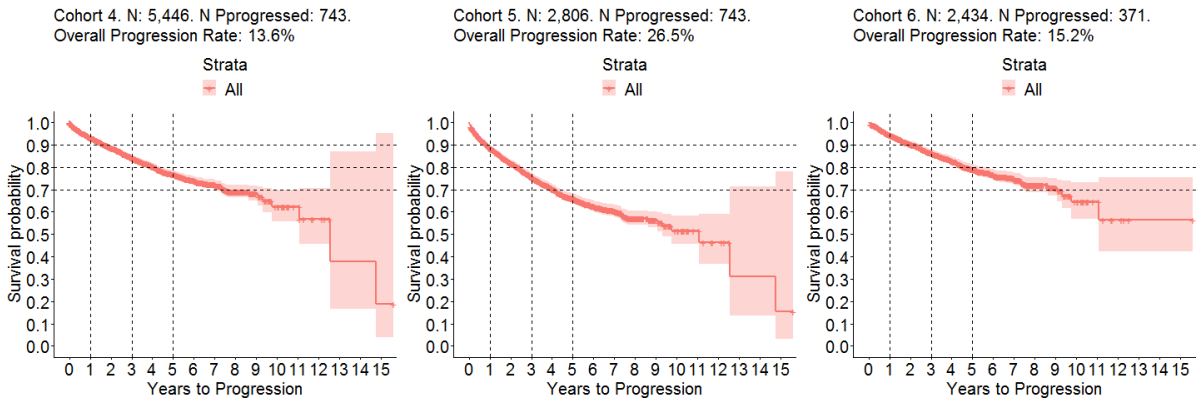

Figure 2: text

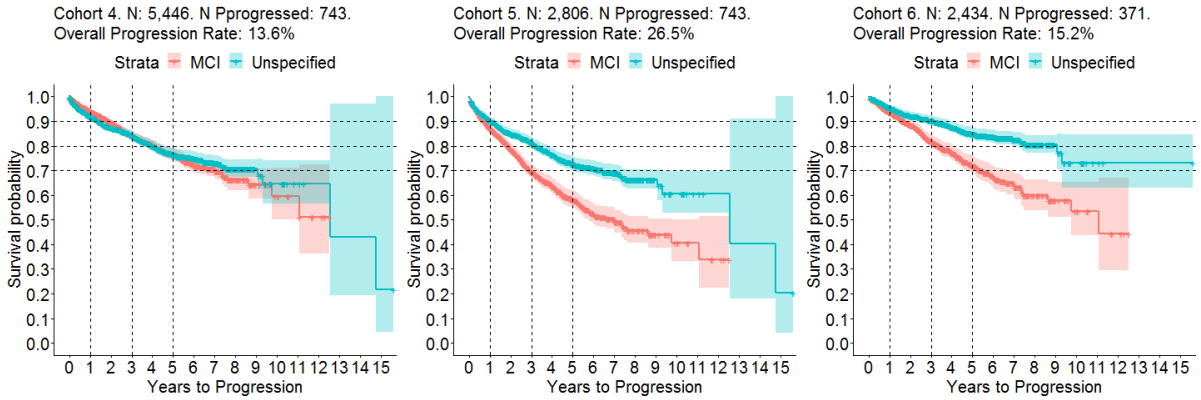

Figure 3: text

It is also essential to check the distribution of time-to-AD progression across age groups, since we observed that the three possible study cohorts differ in the overall age distribution. Figure 2 presents survival curves for each of the three cohorts. Figure 4 contrasts progression rates and times by major age groups. All cohorts reveal a picture that we shoudl expect to see: older age groups have higher progression rates. This conclusion is consistent across the three cohorts. What is of interest is that there does not seem to be an additional effect of age after participats entre the 71 and older age group. Is is likely that the process of aging already took effect for the most part, and patients in the 71+ age group are all very simialr in trems of their physical and mental health. We can also observe that while the rate of events slows down for the odlest people in the study, it is likely due to censoring which was induced by the events of deaths. This is a speculative claim that we are not able to verify from the EHR data.

It is appropriate to discuss the similarities between cohrots 4 and 6 once again. While the two cohorts were different in terms of age distbituion with respect to the intiail diagnosis and the progression rates to AD when stratified by the intiail diagnosis, the effect of age does not appear to differ betweeh the two groups, without accouting for some other predictors of progression time. Lack of difference in terms of progression time by age group implies a sort of comforting similarity between the two gorups of people

representing cohorts 4 and 6.

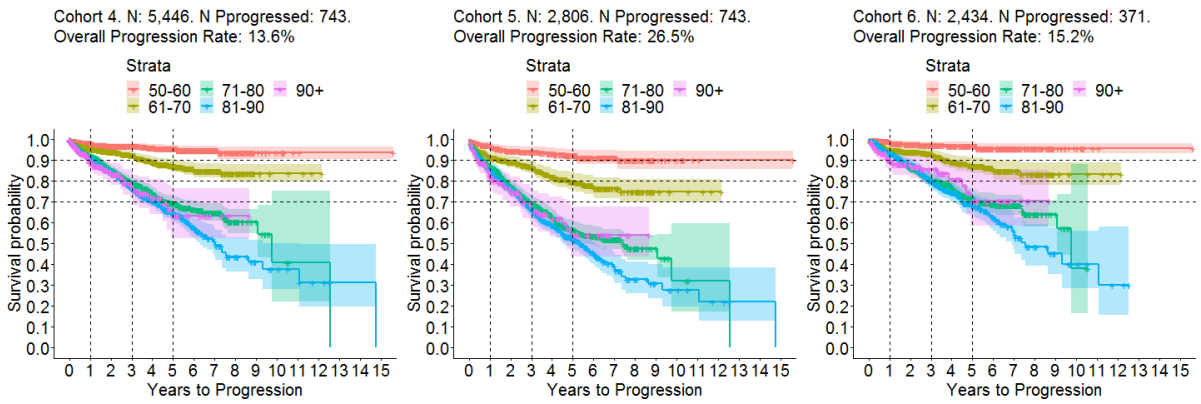

Figure 4: text

Results

Table 2 presents estimates of Cox Proportional Hazard regression models applied to the three cohorts with the same set of predictors. Due to the large sample size and non-complex nature of the model all estimates are statistically significant. However, the results of three models vary quite meaningfully. The biggest take away is that the first diagnosis observation is a strong predictor of progression to AD. In all cohorts the event of MCI being the first diagnosis increase log-hazard (and hazard) of an event of AD, after adjusting for other factors. However, the magnitude of this effect was highly dependent on the cohort. For each cohort with stricter inclusion of patients showed greater impact of initial diagnosis on the log-hazard of the event of AD. Table 2 also shows an interesting picture in terms of the effect of age. For those patients who has 'Unspecified' as the first diagnosis the effect of age was similar across the three cohorts. Each additional year of age at the baseline was statistically associated with the event of progression to AD. However, in cohorts with stricter patient inclusion criteria the effect of age at baseline on the log-hazard was diminished for those who has MCI as the first diagnosis.

Discussion

Through out the exploratory analysis and inference using Cox Propotional Hazard Models we saw that overall progression, progression by sub-groups, and the rates of progression within sub-groups can be higly dependent on the cohort inclusion criteria. Recall that the rationalle for including stricter criteria is to make sure that patients who are getting into the study in fact have on-set MCI before a possible progression to AD. With a lack fo transparency and clinical notes in the EHR we need to make sure that patients who are selected for the study are representative of the general population so that we can make inferences about the whole population based on the subset of patients who we get a chance to observe. There are numerous studies using EHR data to either forecast or predict presence of AD after or within a certain period of time, or identify factors associated with the event of progression to AD.

Table 2: Comparison of statistics for the three cohrots

	Cohort 4			Cohort 5			Cohort 6		
Predictor	Log-HR	95% C.I.	P-value	Log-HR	95% C.I.	P-value	Log-HR	95% C.I.	P-value
First Diag. is MCI	1.43	(0.34, 2.52)	0.01	2.94	(1.83, 4.06)	0	3.69	(2.14, 5.25)	0
Age at Start	0.06	(0.05, 0.07)	0.00	0.06	(0.05, 0.07)	0	0.07	(0.05, 0.08)	0
MCI * Age Interaction	-0.02	(-0.03, -0.01)	0.00	-0.03	(-0.05, -0.02)	0	-0.04	(-0.06, -0.02)	0

[a] HR = Hazard Ratio

[a] 'First Diag. MCI' is Compared to 'First Diag. is Unspecified' reference level

The results of these studies are used as a general clinical practice advise, and authors of these papers advocate that the results of predictive modeling methods advocate for the the use of various statistical models in practice. However, as we discussed in the beginning of the paper, records of visits, notes, use of codes and other data is homogenous within a health system and their data base to a certain degree, but these data structures are different between the independent EHR data bases.

Our study reveals that even within the same EHR database and using reasonable steps to limit population available for the study to obtain overall progression rate and progression rate by the age-subgroup and iniital diagnosis as stratifying variables produces drastically different effects.

We believe that the issues we dicsueed so far and will rouch on in this disucssion section seems to be specific to how AD, and especially MCI, are diagnosed clinically. While diabetes, hypertension, and cardiovascualr conditions, for example, can be identified using established clinical measurements and biomakrer levels, diagnosis of AD and MCI is more subjective process. Moreover, even for the three major types of comorbidities, repeated measures need to be administered for the same patient to guarantee accurate data collection and allow for averaging of measurements to address random variance accosiated with one observation. Subjectivity of AD and MCI diagnosis on a given day is an even greater deal. Currently, for most patients, Mini-Mental State Examination (MMSE) or Montreal Cognitive Assessment (MoCA) are administered as a screening tool. While these are highly developed and well established screening methods, they are still prone to how a patient might feel on a given day in a given moment. Since MCI is a diagnosis given for issues related to memory, there is no good way to objectively and consistently measure the degree of memory impairment. Moreover, like with any clinical tool relying on a scale, a certain cutoff needs to be taken in order to jsutify further medical attention to the patient. If a patient comes, for example, two points short, they might be given an 'Unspecified' diagnosis instead of further screening and finidng an MCI diagnosis, which can heavily affect the clinical trajectory of this

patient, and the data that we get to observe in the EHR. Additional methods of diagnosis of AD also inlude methods like imaging, scans, and other methods that need to be evaluated by a trained physician. This step then introduces more room for error, currently accuracy of AD diagnosis is cited at around 77%. Bringing up repeater measurements for the patients, it is common for patients who have suspected AD to be seen at the three-to-six months intervals in order to conduct a re-evaluation. This seems like a large enough time where you can lose a patient to follow up. In the context of EHR data analysis, we can lose this patient to a different EHR database where they will get their follow up and repeated diagnosis data, but will do not get to observe these data.

All of these issues require us to be very careful and throughful about what patients are being seleced for the study, and what population is being studied in this case. COnsider everyhting we learned about cohorts 4 and 6, with cohort 5 being left aside for now. When lookign at cohorts in the aggregate through Table 1 most important statistics look very similar. Progression rate defined as the number of observed progressions over the number of patients in the cohort look quite similar as well. When looking at the Kaplan-Meier survival curves, the two cohorts have almost identical curves especially within the first seven years of observational time where the most data is available. The two cohorts look similar in terms of progression rates when stratifying by age groups, with every cohort showing increased effect of age at the beginning of observations on the progression rates, with a dimished effect when oatients reach the age of about 70 years old.

So why is the impact of the first diagnosis so different when comparing cohorts 4 and 6? When we apply a highly restrcive criteria and have patients with at least two MCI (MCI or Unspcified) diagnoses, we create a cohort where a follow up to the intial diagnosis is required. When such followup is required, the second diagnosis which confirms a presence of MCI or some unspecified mental condition. It is likely that those who are initially diagnosed with Unspecified condition are at less risk of progression beucase they are just healhier. That is why they are given this diagnosis in the first place.

But why is the impact of the MCI being the first diagnosis so much bigger when compared with the unspcieifed diagnosis in the cohort 6? Well, it can be the case that when we require members to have a follow up diangosis of MCI/Unsoecified, those who start with an Unsoecified diangosis and have a follow up MCI or Unspecified diagnosis are a healthier than those who started with MCI and have a confirmatory diagnosis at some point later in the window of at least 50 days. This conjecture is confirmed by a coeffieid from a Cox proportional hazard regression model.

Therefore, perhaps, we have found a way to construct a study cohort that, by design, separates healthier and less likely to progress people from those who are already at a worse mental state at the time fo the initial diagnosis and therefore are more likely to porgress to AD from MCI/Unspecifried mental diagnosis. At this time, we can return to cnotrast with cohort 4. When we require people to have at least two diangoses total, and not at least two MCI-related diagnosis like ln cohort 6, those who end up progressing may have jsut one MCI-related daingosis. Having just one MCI-related daignosis allows for mroe randomness and mroe sporradic obsevrations, therefore, which in turn dilutes the differences in

terms of progression betwene those with different initial, or anchor diangoses.

While cohorts 4 and cohorts 6 have same progression rates overall, and these rates align with the overall accepted rates of progression from MCI to AD, they consist of different make up of population, and therefore can potentially lead to different results of analyses.

First tradeoff to consider is the sample size. By following a defintion of cohort 4, we allow to include more patients, potentially diversifying type of patients present, and allow to have data that is much more representative of the entire population. An argument can be made that such approach can be more applicable for the machine learning tasks when we employ complex overparametrized models that need large samples of data to learn non-trivial relationships in the data. While we touched on the fact that having one unspecified diangosis before progression to AD may be considered a sporadically occuring event, and therefore 'recruit' a person into the cohort that may not be representative of those people who are likely to progress. The exact conclusion on this is obviously not clear from just looking at the data. However, my recommendation would be to use cohort 4 when attempting to develop a predictive model, since this cohort maximizes total sample size, while preserving overall statistics such that the cohort, at in the aggregate, looks like a population that is commonly cited when evalauting progression from MCI to AD.

On the other hand, cohort 6 may be more appropriate for the study that is aimed at the inference, and looking for factors that can be associated with the event of progression from MCI to AD. We make this argument because, as we discussed, cohrot 6 is designed in a way such that people who are included need to have at least two MCI diagnoses over some time frame (at least 50 days span in our study). By having people who are more likely to actually have an anchor event of MCI, and we have more credibility and confidence that such criteria in cohort 6 identifies peole who are truly on-set MCI patients, we can treat these obsevrations as representative of the entire population of people with MCI. THefore, studying their progressionr rates, and studyign factors associated with progression events and times using a sample similar to cohort 6, seems to be more likely to produce marginal effect estiamtes that will cover true effects in the population.

Thus far, cohort 5 was largely ingored due to the summary statistics. In truth, I wanted to include this cohort becuase it is an example of what can be consideered bad practices, or tweaking the numbers in favor of study conductor. By tightening the inclusion criteria on those who do not have an event of AD, but including everyone who has an event of AD, even with one MCI diangosis, we achieve one good advantage: increased power and thefore more likelihood of discovering some novel factors to describe and prognose the event of progression to AD. However, this comes at a cost, and with a decsion that can be hard to defent. At face value, the cohort looks similar to 4 and 6. Age summary statistics, obsevration times, distribution of specific diagnoses do not deviate largely from 4 adn 6, and align with overall understanding of populations that are at risk of on-set MCI and progression to AD. However, due to our 'hybrid' criteria, progression rate overall looks too high, in reality, these rates are much lower, and the chance of obsevroing 26% progression rate in a random sample when true popualtion progression

rate is 15% should be low.

This cohort presents an example of how applying some criteria to a part of sample can drastically change the population that this sample is suppoosed to represent, and in fact, such sample may not be representative of any popualtion, i.e. such population probably does not exist. This argument can be supported by the estimate from our regression model, since the coefficient for the increase in log-hazard associated with MCI being the first diagnosis is in the middle between cohorts 4 and 6.

Summary

References

Aguilar, Byron J., Donald Miller, Guneet Jasuja, Xuyang Li, Ekaterina Shishova, Maureen K. O'Connor, Andrew Nguyen, et al. 2023. "Rule-Based Identification of Individuals with Mild Cognitive Impairment or Alzheimer's Disease Using Clinical Notes from the United States Veterans Affairs Healthcare System." *Neurology and Therapy* 12 (6): 2067–78. https://doi.org/10.1007/s40120-023-00540-2.

Mattke, Soeren, Hankyung Jun, Emily Chen, Ying Liu, Andrew Becker, and Christopher Wallick. 2023. "Expected and Diagnosed Rates of Mild Cognitive Impairment and Dementia in the u.s. Medicare Population: Observational Analysis." *Alzheimer's Research & Therapy* 15 (1): 128. https://doi.org/10.1186/s13195-023-01272-z.

Xu, Jie, Rui Yin, Yu Huang, Hannah Gao, Yonghui Wu, Jingchuan Guo, Glenn E Smith, et al. 2023. "Identification of Outcome-Oriented Progression Subtypes from Mild Cognitive Impairment to Alzheimer's Disease Using Electronic Health Records." *medRxiv*. https://doi.org/10.1101/2023.07.27.23293270.

Appendix

Supplemental Figures

Cox PH models

Cohort 4 estimated model:

$$\begin{aligned} h(t) &= h_0(t)exp(\beta_1 I(First\ Diag\ is\ MCI) + \beta_2 Age + \beta_3 I(First\ Diag\ is\ MCI) * Age) \\ &= h_0(t)exp(1.43 I(First\ Diag\ is\ MCI) + 0.06 Age - 0.02 I(First\ Diag\ is\ MCI) * Age) \end{aligned}$$

Estimated Effect of I(MCI) = 1 and age = 65:

$$h(t) = h_0(t)exp(1.43 + 0.06 \times 65 - 0.02 \times 65) = h_0(t)exp(4.03)$$

Estimated Effect of I(MCI) = 0 and age = 65:

$$h(t) = h_0(t)exp(0.06 \times 65) = h_0(t)exp(3.9)$$

Ratio of effects: $exp(4.03)/exp(3.9) = 1.14$

Estimated Effect of I(MCI) = 1 and age = 85:

$$h(t) = h_0(t)exp(1.43 + 0.06 \times 85 - 0.02 \times 85) = h_0(t)exp(4.83)$$

Estimated Effect of I(MCI) = 0 and age = 85:

$$h(t) = h_0(t)exp(0.06 \times 85) = h_0(t)exp(5.1)$$

Ratio of effects: $exp(4.83)/exp(5.1) = 0.7633$

Cohort 6 estimated model:

$$\begin{aligned} h(t) &= h_0(t)exp(\beta_1 I(First\ Diag\ is\ MCI) + \beta_2 Age + \beta_3 I(First\ Diag\ is\ MCI) * Age) \\ &= h_0(t)exp(3.69 I(First\ Diag\ is\ MCI) + 0.07 Age - 0.04 I(First\ Diag\ is\ MCI) * Age) \end{aligned}$$