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

EHRs have become a rich source of data for the analysis of phenomenon related to Alheimer's Disease and Dimetia. Many studies has been published recently attempting to identify AD or progression to AD using patients information stored in the EHR systems. Search for factors associated with AD, such as comorbidities or health outcomes associated with the presence of AD have been a common topic as well, although taking a back seat to a vastly diverse and large amount of work in the predicative analysis area. Researchers acknowledge key challenges in creating the data for analysis using 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 and the ability to generalize results of a given analysis to broader applications. In particular, the diagnoses codes used to identify onset of MCI and early stages of AD can be inconsistent or incomplete. Inconsistency mainly refers to the choice of ICD diagnosis used on the record of patient's visit. As we will see in the future sections, there are two major options for coding, for example, an MCI diagnosis. Two options pose major applications both for the definition that describes a population, and to the results of statistical analyses. Incompleteness refers to the fact that a given person might have a condition occur, but diagnosis is not recorded in the EHR, when the person does not return to a given health system and gets diagnosed elsewhere, or comes in for an unrelated reason, and the diagnosis is not recorded because a patient does not bring it up, or share symptoms that may be indicative of MCI presence.

Zhang, Simon, and Yu (2017) demonstrate that the utilization of big data is pivotal in advancing research on Alzheimer's disease (AD), primarily due to challenges in patient recruitment, retention, and the time and cost constraints associated with traditional clinical research methods. Analyzing 38 studies, they identified seven key research areas, including diagnosing AD or mild cognitive impairment (MCI), predicting MCI to AD progression, stratifying risks for AD, mining literature for knowledge discovery, predicting AD progression, describing clinical care for individuals with AD, and understanding the relationship between cognition and AD.

Datasets utilized in AD research encompass a variety of sources such as the Alzheimer's Disease Neuroimaging Initiative (ADNI), AddNeuroMed study, Mayo Clinic Study on Aging, ZARAgoza DEMentia DEPression (ZARADEMP) study, electronic health records (EHR), MEDLINE, and other research datasets. Data analytics methods employed span a broad spectrum, including data mining, machine learning, natural language processing (NLP), text mining, and statistical analysis.

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 always show steps of developing robust data pipelines relevant to MCI and AD progression, but different analyses have massively varying flows and intermediate steps. 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.

Obective of Analysis

The three examples above demonstrate that there is no one way to make data based definitions for clinical diagnoses of AD and MCI from the EHR data. But with the rising availability of EHR data, and interest to develop models within these systems to monitor populations covered by healthcare systems for clinical guidance, decision making, and forecast and event anticipation, we need to have an idea of what can potentially happen when we apply published definitions and data collection workflows to the EHR that is different from the one that the flow was developed with. Additionally, we need to understand how sensitive the data could be when we apply varying cohort definitions to the same pooled data from the same EHR. The objective of our study is to construct varying cohorts, resembling those common criteria we saw in the recent publications. We wish to evaluate implications of the restructing criteria on the sample size, overall statistics, and if the represented sample from some population is similar to common accepted statistics when it comes to common knowledge surrounding progression to AD from MCI. We also hope to make recommendation about validity of each cohort and what type of statistical learning task type each cohort is applicable for.

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 by giving a more accurate picture of time until censoring event, and therefore 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. In line with relevant literature reviews, we use ICD codes capturing that widely capture MCI, AD and other Dementia diagnoses, and Unspecified mental disorder codes as primary markers of conditions of interest. According to the ICD-10-CM, F09 is the diagnosis code for an unspecified mental disorder caused by a known physiological condition. This code is listed by the WHO under the range of mental, behavioral, and neurodevelopmental disorders. Organic mental disorders in older adults can cause a range of symptoms, including: a decline in memory, comprehension, learning capacity, language abilities, judgment, and severe dementia. We believe that this diagnosis is widely used by the research community because it captures symptoms that are commonly cited for MCI and AD. Using this diagnosis to identify patients with this condition allows us to include a much broader range of potential subjects for a study. As a trade off for a larger data set we potential incur a subset of patients who in fact do not have an MCI, nothing close to MCI, and therefore can induce negative bias towards progression rate estimates by including members who would never progress to AD.

A special note has to be made about handling of Unspecified mental disorder. Even though it is not specifically called "MCI", it is a common practice in relevant ongoing research to use Unspecified Mental Disorders diagnoses to identify MCI. Later in the analysis section we make a distribution between MCI and Unspecified diagnoses to make small inferences on the populations captured in each possible cohort. However, in the context of this paper, we commonly will say "progression from MCI to AD", which really means a progression from the first occurrence of Unspecified mental disorder to MCI until the event of AD. We deem this, and do many other research groups, an appropriate step considering that this diagnosis is given to capture a condition that resembles MCI, but perhaps does not have just as strong of symptoms and affect in the daily life of a patient.

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 data set with a large pool of patients available for analysis, but more filtering needs to be done before obtaining a credible data set. Figure 5 shows the flow of data collection from the Fairview Health System Data Warehouse. Initially, we were able to gather data on 20,121 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 only the event of AD. To keep the framework of analysis relevant with exciting research, we do not wish to study progression rates from the point of view of right-censored data. Additionally,events of AD recorded before MCI were removed from the analysis. It is a common occurrence to have cases in the EHR where MCI diagnoses show up after AD due to the subjective nature of diagnosis procedures that are inherent and specific to the family of these diagnoses.

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 adults. Since dementia and Alzheimer Disease are age-related, removing younger individuals made sense. 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.

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 populations. Using these general criteria, we identify 5,711 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. One note we would like to reiterate is what we mean by “MCI” diagnosis in this context. It really captures either of MCI or Unspecified Mental Disorder Diagnosis, as we determined that it is reasonable to make them equal to identify anchoring events that initiate the count of time form the occurrence or suspected occurrence of MCI. 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 is constructed by imposing 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 consider it. Additionally, it would not make sense to consider adults below the age of 50, even if they had more than two MCI diagnoses. It is likely that those who were filtered out in this step are in fact patients with mental, memory, and other relevant issues, but likely arising from other health problems, not related to a potential occurrence of AD within near future for those patients.

Cohort 5: is obtained from Cohort 4 by restricting the timeline of MCI diagnoses that take place before a potential progression. For those who do not end up progressing to AD, which means that they have at least two MCI diagnoses, we impose a rule that the time difference between the first and the last diagnoses of MCI must be at least 50 days apart. This approach is similar to Hane et al. (2020), but we increased 31 days time span to 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 to 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 4 and 6.

Cohort 6 is finally obtained by applying the criteria of at least 50 day time lapse between first and last diagnoses of MCI 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.

Data

Cohort Summaries

Table 1 compares key statistics across the three cohorts. Similarities exist for all summarized measures outside sample sizes and event rates. An extremely elderly patient population characterizes these cohorts on average. Age variances appear approximately equal between cohorts. Duration to progression and ages at progression also signify similarities. Furthermore, ages at progression imply relatively older individuals within these cohorts exhibit increased likelihoods of progressing.

As later cohort versions require increased time lags between initial and final MCI or Unspecified diagnoses, this likely explains observed increases in average diagnosis counts for later versions. Cohorts 4 and 6 differ in progression rates and observation windows post-final diagnosis. Cohort 6 demonstrates a slightly

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%)

elevated progression rate, but also a lengthier window following the final MCI diagnosis, permitting capture of additional progression 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 over at least 50 days window drastically changes the population from which a study sample is created. 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. While in out study we treat Unspecified diagnosis as MCI, for the purpose of finding the anchor event, the earliest time e suspect MCI has occurred, it is clear from the initial data summaries that the two types of diagnoses capture varying people. We can, perhaps, think of these as MCI diagnosis of varying severity degree, such as Early and Late MCI diagnoses. Since they capture samples from populations that are different in terms of their age, and therefore a whole variety of other factors, we will examine the disparities in terms of progression to AD between those who start with Unspecified and those who start with MCI.

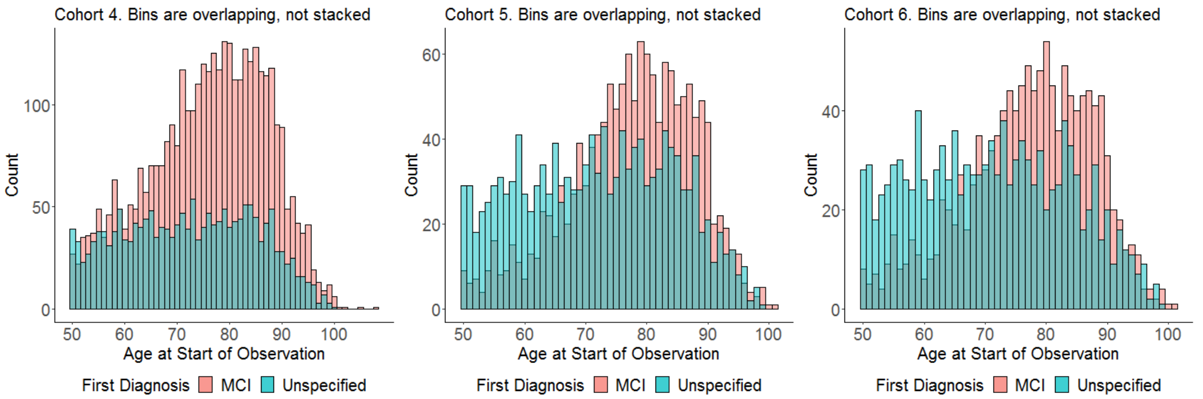


Figure 1: Breakdown of Age distribution

Kaplan Meir Curves

The next crucial preliminary data analysis step involves examining the Kaplan-Meier survival curves. Analogous to Figure 1, first diagnosis, either MCI or Unspecified, serves as a stratifying variable assessing time-to-event distribution differences between the two groups, as presented in Figure 3 for all three cohorts.

Immediately observable are patterns across cohorts similar to age distribution conclusions. Cohort 4 exhibited highly similar age distributions between MCI and Unspecified first diagnosis groups. Unsurprisingly, age manifests as a more powerful predictor of MCI to AD progression. Since age distributions appeared visually similar across the two groups, observed age distribution differences translate equivalently into time-to-AD progression estimate differences.

Notably, Cohorts 4 and 6 demonstrate quite similar overall progression rates in the samples, presumably due to similar overall progression rates in the data. However, the MCI/Unspecified subgroup difference increases. Within the first 7-9 years of follow-up, Cohorts 4 and 6 are indistinguishable overall, while the picture changes drastically when stratifying by the initial, or anchor, diagnosis. This key distinction, paired with largely differing age distributions, implies that while cohorts appear similar on paper, the two sub-populations comprising the overall cohort differ quite substantially. The unadjusted progression differences between groups with varying original diagnosis severity suggest imposing stricter criteria identifying EHR patients with more confirmatory MCI onset produces a population heavily stratified by inherent AD progression risk. While reducing sample size and potentially limiting statistical power, having such a sample where progression rates differ substantially based on initial diagnosis and age may aid statistical analyses with inference as the primary goal.

It is also essential to examine the time-to-AD progression distribution across age groups, since the three potential study cohorts differed in overall age distribution, as presented in Figure 2 and Figure 4. All cohorts reveal an expected picture: older age groups exhibit higher progression rates, a conclusion consistent across the three cohorts. Interestingly, there does not appear to be an additional age effect

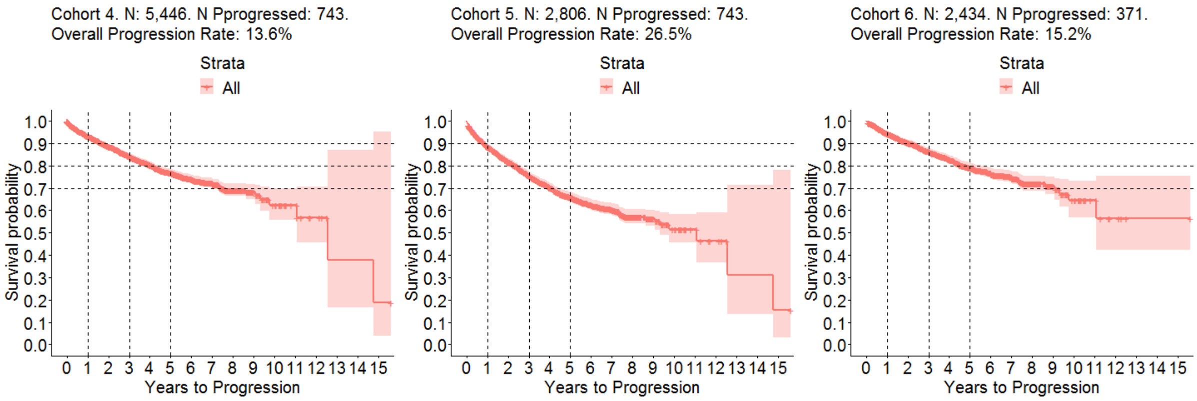


Figure 2: text

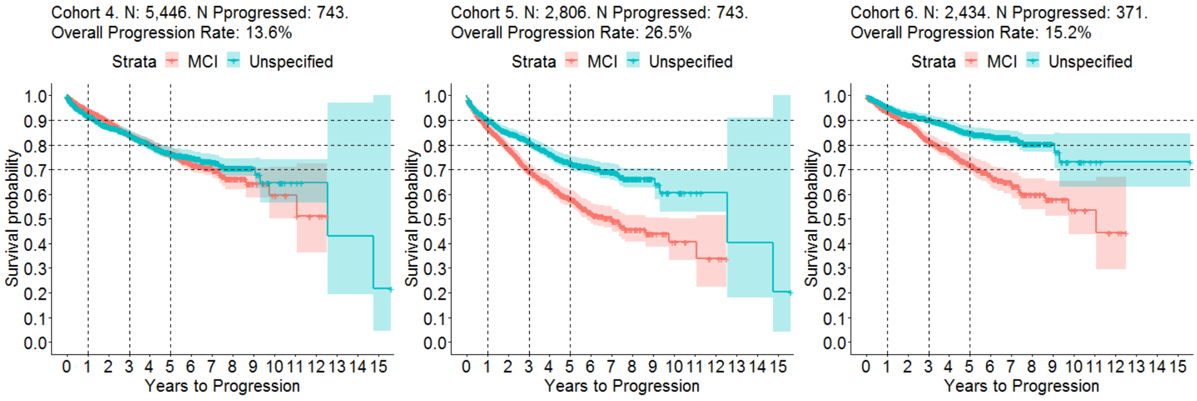
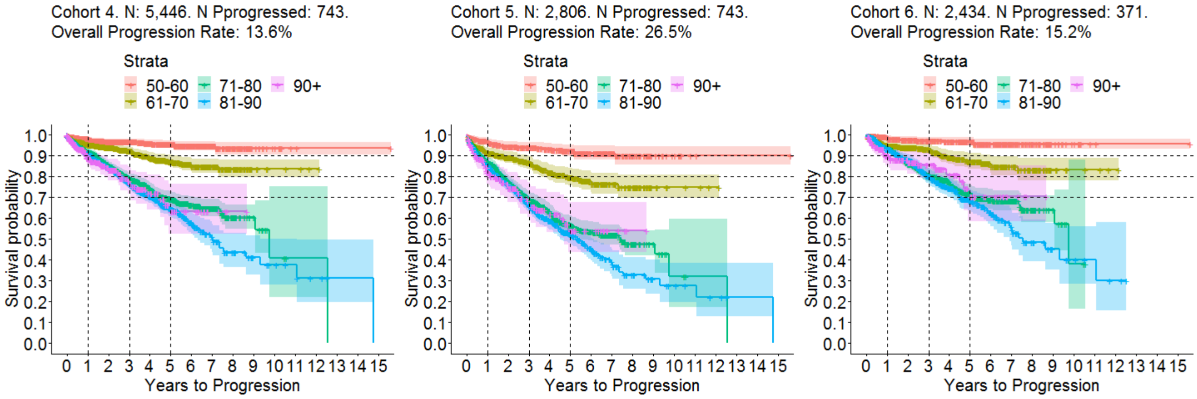


Figure 3: text

after participants enter the 71 and older group. It is likely the aging process had already largely taken effect, with patients in the 71+ age group being very similar in terms of physical and mental health. While event rates slow for the oldest study participants, this is likely due to censoring induced by death events, an unverifiable speculative claim from the EHR data.

It is appropriate to again discuss similarities between Cohorts 4 and 6. Although the two cohorts differed in age distribution with respect to the initial diagnosis and progression rates to AD when stratified by the initial diagnosis, the age effect on progression time does not appear to differ between the two groups when not accounting for other progression predictors. This lack of difference in progression time by age group implies a comforting similarity between the populations representing Cohorts 4 and 6.


Cohort 4. N: 5,446. N Pprogressed: 743. Overall Progression Rate: 13.6%

Cohort 5. N: 2,806. N Pprogressed: 743. Overall Progression Rate: 26.5%

Cohort 6. N: 2,434. N Pprogressed: 371. Overall Progression Rate: 15.2%

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 had '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.

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

Discussion

Throughout the exploratory analysis and inference using Cox Proportional Hazard Models, we observed that overall progression, progression by subgroups, and rates of progression within subgroups can heavily depend on the cohort inclusion criteria. Recall the rationale for including stricter criteria is ensuring patients entering the study have onset MCI with higher confidence, before potential AD progression, rather than an MCI or Unspecified diagnosis recorded in the EHR due to other reasons or faulty initial diagnosis. The latter is more applicable to the MCI ICD diagnosis, while Unspecified diagnoses carry a higher risk of relating not to potential MCI presence itself, but other mental issues patients experience. With a lack of transparency and clinical notes in the EHR, we must ensure selected patients represent the general population, allowing inferences about the whole population based on the observed patient subset.

Numerous studies utilize EHR data to forecast or predict AD presence after/within a certain period, or identify factors associated with AD progression. These studies' results are used as general clinical practice advice, with authors advocating for various statistical models' use in practice. However, as discussed early, visit records, notes, code usage, and other data demonstrate homogeneity within a health system and database to a degree, but these data structures differ between independent EHR databases. Our study reveals that even within the same EHR database, using reasonable steps to limit the available population to obtain overall progression rate, progression rate by age-subgroup, and initial diagnosis as stratifying variables produces drastically different effects. This implies the patient inclusion criteria has an immense effect on the population makeup and, therefore, a strong effect on the type and quality of knowledge inferred from such samples.

We believe the issues discussed seem specific to how AD, especially MCI, are clinically diagnosed. For example, Lombardi et al. (2020) meta-analysis shows that MRI alone lacks accuracy in early diagnosing Alzheimer's disease dementia in individuals with MCI, with a high rate of misdiagnosis observed. They also discuss other challenges associated with the diagnoses of MCI and AD. Aslam et al. (2018) also alluded that certain tests display potential in detecting MCI and early dementia, issues such as small sample sizes, study replicability, and insufficient evidence hinder making clinical recommendations on their use for diagnosis, progression monitoring, and treatment response. Further research is essential to establish consistent cut-off points for automated computerized tests in diagnosing individuals with MCI or early dementia. While diabetes, hypertension, and cardiovascular conditions can be identified using established clinical measurements and biomarker levels, AD and MCI diagnosis is a more subjective process. Moreover, repeated measures need administration for the same patient to guarantee accurate data collection and allow averaging to address random variance associated with one observation for the three major comorbidity types. AD and MCI diagnosis subjectivity on a given day is an even greater issue. Currently, for most patients, Mini-Mental State Examination (MMSE) or Montreal Cognitive Assessment (MoCA) serve as screening tools. While highly developed and well-established, they remain prone to how a patient feels on a given day. Since MCI is a diagnosis for memory issues, there is no objective, consistent way to measure memory impairment degree. Moreover, like any clinical tool relying on a scale, a certain cutoff must be taken to justify further medical attention. If a patient falls just short, they might receive an 'Unspecified' diagnosis instead of further screening and MCI diagnosis, heavily affecting the clinical trajectory and observed EHR data. Additional AD diagnosis methods include imaging, scans, and other physician-evaluated methods, introducing more room for error. Currently, AD diagnosis accuracy is cited around 77%. Literature on MCI diagnosis accuracy and methods commonly shows inclusive results and lack of reproducibility, limiting clinical decision-making.

Bringing up repeated measurements for patients, it is common for suspected AD patients to be seen at three-to-six month intervals for re-evaluation. This seems a large enough time where patient follow-up can be lost. In the EHR data analysis context, we can lose patients to different EHR databases where they receive follow-up and repeated diagnosis data we do not observe. These issues require careful thoughtfulness about which patients are selected for the study and the population being studied.

Key Takeaway

Considering everything learned about Cohorts 4 and 6, with Cohort 5 aside for now, most important aggregate statistics in Table 1 appear very similar. The progression rate, defined as observed progressions over cohort patients, also looks quite similar. Examining the Kaplan-Meier survival curves, the two cohorts have almost identical curves, especially within the first seven years where most data is available. The two cohorts look similar in progression rates when stratifying by age groups, with every cohort showing increased age effect on progression rates initially, diminishing around 70 years old.

So why is the first diagnosis's impact so different when comparing Cohorts 4 and 6? Applying a highly restrictive criteria requiring at least two MCI (MCI or Unspecified) diagnoses creates a cohort where follow-up to the initial diagnosis is required. With such follow-up, the second diagnosis confirms presence of MCI or some unspecified mental condition. It is likely those initially diagnosed with an Unspecified condition are at less progression risk because they are just healthier, hence receiving that diagnosis initially. But why is the MCI first diagnosis impact so much greater compared to the Unspecified diagnosis in Cohort 6? It could be that observations from EHR are less credible, and lack the ability to accurately capture the development mental issues and captured medical history. In cohort 4, those who progress to AD are allowed to have one MCI diagnosis. It is likely that this diagnosis is captured 'by chance', or that there was an ordered follow up appointment, but the members either did not show up or MCI was not documented at that appointment. When requiring at least two total diagnoses, not at least two MCI-related as in Cohort 6, those progressing may have just one MCI-related diagnosis. Having just one MCI-related diagnosis allows more randomness and sporadic observations, diluting differences in progression between those with different initial, or anchor, diagnoses. While Cohorts 4 and 6 have the same overall progression rates aligning with accepted MCI to AD progression rates, they consist of different population makeups, potentially leading to different analysis results.

To reiterate, while Cohort 6 patients are more surely confirmed onset MCI cases, having an initial Unspecified diagnosis suggests doctors felt there might have been less concrete MCI evidence at initial evaluation, perhaps due to overall healthier patient status. This conjecture is confirmed by a Cox proportional hazards regression coefficient. Therefore, we may have constructed a study cohort that, by design, separates healthier, less likely to progress people from those in worse mental state at initial diagnosis, more likely to progress to AD from MCI/Unspecified mental diagnosis.

Recommendations

The first trade off to consider is sample size. Following Cohort 4's definition allows inclusion of more patients, potentially diversifying present patient types and providing data more representative of the entire population. An argument manifests that such an approach may be more applicable for machine learning tasks employing complex overparametrized models requiring large data samples to learn non-trivial relationships.

While we touched on having one Unspecified diagnosis prior to AD progression being considered more of a sporadically occurring event, potentially 'recruiting' someone into the cohort unrepresentative of likely progressing patients, this conclusion is unclear from the data alone. However, the recommendation would be utilizing Cohort 4 when attempting predictive model development, as this cohort maximizes total sample size while preserving overall statistics such that the cohort, in aggregate, resembles a commonly cited population when evaluating MCI to AD progression.

On the other hand, Cohort 6 may be more appropriate for inference studies aimed at identifying factors associated with MCI to AD progression events. This argument arises as Cohort 6 is designed such that included individuals need at least two MCI diagnoses over some time frame (at least 50 days in our study). Having people more likely to actually experience an anchor MCI event provides increased credibility and confidence that Cohort 6's criteria identifies truly onset MCI patients. We can treat these observations as representative of the entire MCI population. Therefore, studying progression rates and factors associated with progression events/times using a Cohort 6-like sample seems more likely to produce marginal effect estimates covering true population effects.

Thus far, Cohort 5 was largely ignored due to summary statistics. In truth, inclusion occurred as an example of potentially bad practices or tweaking numbers in favor of the study conductor. By tightening non-AD event inclusion criteria but including all AD events even with one MCI diagnosis, one advantageous power increase manifests, with higher likelihood of discovering novel factors describing prognosis of AD progression events. However, this incurs a cost and decision potentially difficult to defend. On face value, Cohort 5 appears similar to 4 and 6, with age summary statistics, observation times, and specific diagnosis distributions not deviating largely and aligning with overall understandings of onset MCI and AD progression risk populations. However, due to the 'hybrid' criteria, the overall progression rate appears too high, as reality reflects much lower rates, with observing a 26% progression rate in a random sample when the true population rate is 15% being unlikely.

This cohort exemplifies how applying criteria to part of a sample can drastically change the population the sample is supposed to represent, with the sample potentially being unrepresentative of any population. Our regression model estimate supports this argument, as the coefficient for increased log-hazard associated with MCI being the first diagnosis falls 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.

Aslam, Rabeea W., Vicki Bates, Yenal Dundar, Juliet Hounsome, Marty Richardson, Anil Krishan, Rumona Dickson, et al. 2018. "A Systematic Review of the Diagnostic Accuracy of Automated Tests for Cognitive Impairment." *International Journal of Geriatric Psychiatry* 33 (4): 561–75. https://doi.org/10.1002/gps.4852.

Hane, Christopher A, Vijay S Nori, William H Crown, Darshak M Sanghavi, and Paul Bleicher. 2020. "Predicting Onset of Dementia Using Clinical Notes and Machine Learning: Case-Control Study." *JMIR Med Inform* 8 (6): e17819. https://doi.org/10.2196/17819.

Lombardi, Giovanni, Gabriele Crescioli, Enrica Cavedo, Ersilia Lucenteforte, Giovanni Casazza, Alberto Bellatorre, Chiara Lista, et al. 2020. "Structural Magnetic Resonance Imaging for the Early Diagnosis of Dementia Due to Alzheimer's Disease in People with Mild Cognitive Impairment." *Cochrane Database of Systematic Reviews* 2020: CD009628. https://doi.org/10.1002/14651858.CD009628.pub2.

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.

Zhang, Rui, Gyorgy Simon, and Fang Yu. 2017. "Advancing Alzheimer's Research: A Review of Big Data Promises." *International Journal of Medical Informatics* 106: 48–56. https://doi.org/https://doi.org/10.1016/j.ijmedinf.2017.07.002.

Appendix

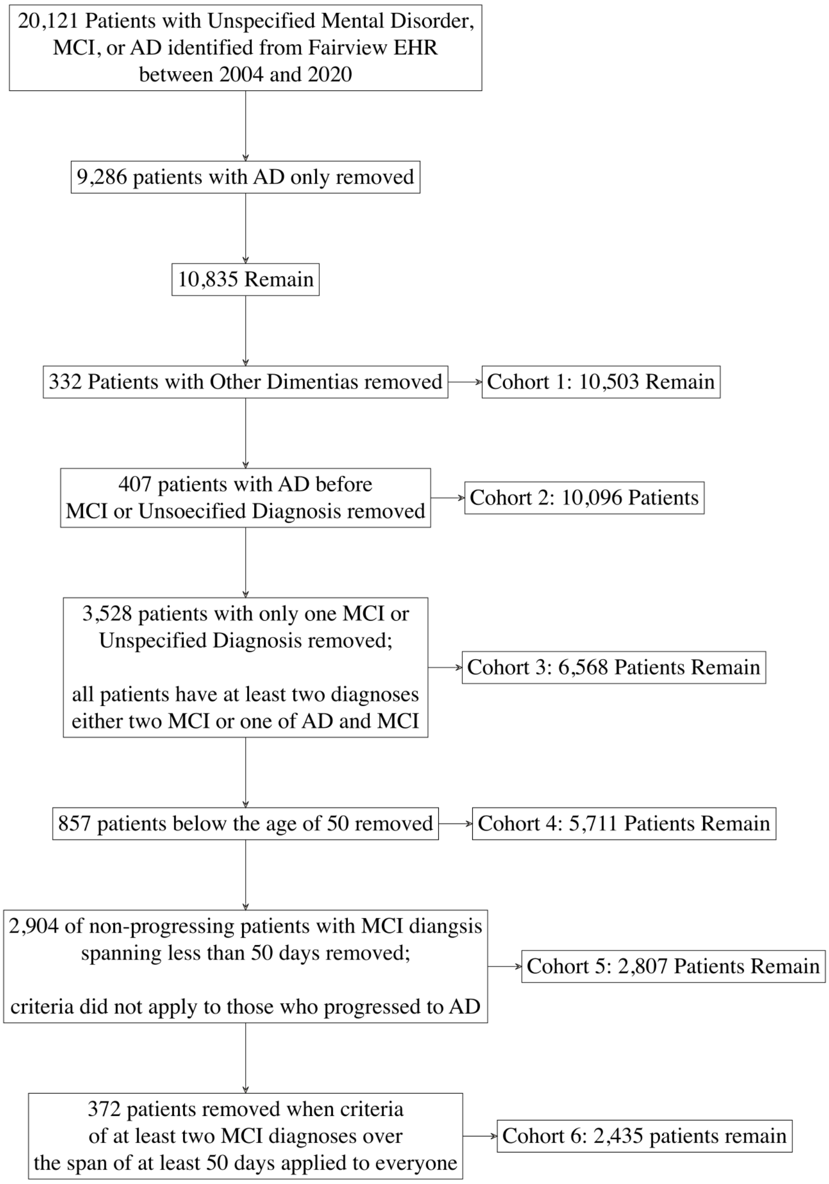

20,121 Patients with Unspecified Mental Disorder, MCI, or AD identified from Fairview EHR between 2004 and 2020
9,286 patients with AD only removed
10,835 Remain
332 Patients with Other Dimentias removed
Cohort 1: 10,503 Remain
407 patients with AD before MCI or Unsoecified Diagnosis removed
Cohort 2: 10,096 Patients
3,528 patients with only one MCI or Unspecified Diagnosis removed;
all patients have at least two diagnoses either two MCI or one of AD and MCI
Cohort 3: 6,568 Patients Remain
857 patients below the age of 50 removed
Cohort 4: 5,711 Patients Remain
2,904 of non-progressing patients with MCI diangsis spanning less than 50 days removed;
criteria did not apply to those who progressed to AD
Cohort 5: 2,807 Patients Remain
372 patients removed when criteria of at least two MCI diagnoses over the span of at least 50 days applied to everyone
Cohort 6: 2,435 patients remain

Figure 5: Flow of data collection