Evaluation of cohort construction methods for MCI to AD progression from EHR data. Discussion and Recommendations

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 Dimentia. 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 implications 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.

Despite these challenges, 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.

For these tasks, 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.

Objective of Analysis

The three examples above highlight the varied methods used in crafting data-based definitions for clinical diagnoses of AD and MCI from EHR data. However, as the availability of EHR data increases and interest grows in developing models within these systems for monitoring healthcare-covered populations, it’s crucial to anticipate potential disparities when applying published definitions and data collection workflows to EHRs different from those for which the protocols were developed. Moreover, understanding the data’s sensitivity to varying cohort definitions applied to the same pooled data from the same EHR is essential.

Our study aims to delineate diverse cohorts mirroring common criteria observed in recent publications. We seek to assess the implications of restructuring criteria on sample size, overall statistics, and the degree to which the represented sample from a given population aligns with commonly accepted statistics regarding the progression to AD from MCI. Additionally, we aim to provide recommendations regarding the validity of each cohort and the type of statistical learning task for which each cohort is most applicable.

Methods

Data Collection

Data for the study were obtained from the Fairview Health System Data Warehouse, comprising records of visits to any Fairview Facility with ICD diagnoses identifying the presence of MCI, AD, or Unspecified Mental Disorder. Patient records spanned between 2004 and 2020, capturing all visits regardless of their relevance to MCI/AD. Including maximum possible observation times benefits the analysis, particularly for those not progressing to AD from MCI, yielding more accurate estimates of time until censoring events and reducing bias.

The key step of data engineering involves defining what constitutes a ‘visit.’ A visit where a patient is diagnosed with MCI or AD is one unique date when at least one of the ICD codes for MCI, Unspecified Mental Disorder, or AD occurs. A list of all diagnoses codes is given in Table 3. We use ICD codes capturing MCI, AD, and other Dementia diagnoses, along with Unspecified mental disorder codes, as primary markers of conditions of interest. F09 in the ICD-10-CM denotes an unspecified mental disorder caused by a known physiological condition, commonly used by the research community to capture symptoms cited for MCI and AD. This inclusion expands the potential subject pool but may introduce negative bias by including patients who would never progress to AD.

A notable aspect concerns the handling of Unspecified mental disorder, commonly used in ongoing research to identify MCI. Although not specifically labeled as “MCI,” it captures a condition resembling MCI but perhaps with lesser impact on daily life, influencing the progression from some mild and developing mental issues to AD.

As mentioned in the introduction, EHR data originates from a structured database, but translating clinical notes into structured data results in a dataset with a large patient pool for analysis, requiring additional filtering for credibility. Figure 5 illustrates the data collection flow from the Fairview Health System Data Warehouse. Initially, data on 20,121 patients were gathered. To study transition times from initial MCI occurrence to the first AD diagnosis, data cleaning continued, removing patients with only AD events and those with AD recorded before MCI.

In refining study population restrictions, we applied general filtering criteria to all cohorts. Firstly, we imposed further restrictions on AD diagnosis, ensuring eligibility only for ICD codes ’G30*’ and ‘331,’ thereby forming a group with a homogeneous outcome variable. Secondly, patients with only one Unspecified Mental Disorder diagnosis, often younger adults, were removed, as dementia and Alzheimer Disease are age-related. Including such patients likely results in lower progression rates and biased estimates.

Cohort Definitions

After reviewing the literature on work using EHR, Healthcare Claims, and other types of digital health records, we identified common criteria used to create cohorts or study populations for developing predictive and statistical models. 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 diagnosis. It's important to clarify the definition of "MCI" diagnosis in this context, which encompasses either MCI or Unspecified Mental Disorder Diagnosis to identify anchoring events initiating the count of time from the occurrence or suspected occurrence of MCI. At this point, we define three cohorts for evaluation, labeled as "Cohort 4", "Cohort 5", and "Cohort 6", names retained after data cleaning steps.

Cohort 4 is constructed by imposing a filtering criterion that all individuals at the date of their first MCI or Unspecified diagnosis are at least 50 years old. This age restriction, common in the literature, ensures relevance. Additionally, considering adults below 50 would likely include patients with issues unrelated to potential AD occurrence in the near future.

Cohort 5 is derived from Cohort 4 by restricting the timeline of MCI diagnoses before potential progression. For those not progressing to AD, meaning they have at least two MCI diagnoses, we require the time difference between the first and last MCI diagnoses to be at least 50 days apart. This adjustment, similar to Hane et al. (2020), extends the time span to 50 days, selecting patients seen on multiple occasions with confirmed or consistent MCI diagnoses and filtering out sporadic cases.

For patients with an AD diagnosis event, we do not apply this criterion to include as many people with AD diagnosis as possible for potential study and predictive modeling purposes, aiming to maximize data for individuals with a desired outcome variable and compare results to Cohorts 4 and 6.

Cohort 6 is obtained by applying the criterion of at least a 50-day time lapse between the first and last MCI diagnoses to all patients in the cohort. This restriction reduces sample size and the number of AD events but focuses on patients with confirmed, consistent MCI diagnoses preceding a possible progression to AD.

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

Statsitical Methods

To investigate the effects of age and initial diagnosis status on the progression from MCI to Alzheimer's Disease (AD), we employ the Cox Proportional Hazards Regression Model. This semi-parametric approach models the relationship between predictors and the hazard rate, which quantifies the instantaneous risk of experiencing the event of interest at a given time point.

The Cox model provides a robust framework for analyzing time-to-event data, such as the transition from MCI to AD diagnosis in our study. It allows us to estimate the hazard ratios associated with predictor variables while accounting for censored observations, a common occurrence in longitudinal studies where some individuals may not experience the event during the observation period. By applying the Cox Proportional Hazards model to the three cohorts (Cohorts 4, 5, and 6), we aim to examine how the same set of predictors, age and initial diagnosis status, relate to the hazard of progression across these distinct populations. Comparing the model estimates across cohorts enables us to assess whether inherent differences exist in the factors influencing the progression rate.

If substantial disparities in the effects of predictors are observed among the cohorts, it may indicate the presence of underlying cohort-specific characteristics that modulate the relationship between the predictors and the event of progression. Such findings could provide valuable insights into the unique factors or subpopulations represented by each cohort, potentially guiding the selection of appropriate cohorts for specific statistical tasks or modeling objectives. For instance, if a cohort exhibits substantially different predictor effects compared to others, it may suggest that this cohort represents a subpopulation with distinct progression dynamics or risk factors. Consequently, this cohort might be better suited for developing predictive models tailored to that subpopulation or for investigating the underlying mechanisms driving the observed differences. By leveraging the flexibility of the Cox Proportional Hazards model and the diversity of cohort definitions, our analysis aims to elucidate potential cohort-specific factors influencing the transition from MCI to AD. These findings could inform future research efforts by identifying cohorts most appropriate for specific analytical tasks, such as predictive modeling, risk stratification, or elucidating disease mechanisms.

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 elevated progression rate, but also a lengthier window following the final MCI diagnosis, permitting capture of additional progression events.

Figure 1 presents the distribution of age for the three cohorts at the time of initial MCI diagnosis. Imposing a criterion of having at least two MCI diagnoses over a minimum 50-day window significantly alters the population from which a study sample is derived. As discussed previously, MCI and Unspecified diagnoses differ in nature. Unspecified diagnosis can serve as a ‘catch-all’ diagnosis, leading to younger patients being more likely diagnosed with the ‘Unspecified’ ICD-10 code.

Commonly, Unspecified mental disorder, or unspecified mental disorder of unknown origin, is used to

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%)

identify individuals who potentially have MCI. While in our study we treat Unspecified diagnosis as MCI for the purpose of finding the anchor event—the earliest time we suspect MCI has occurred—it is evident from the initial data summaries that the two types of diagnoses capture different populations. We may consider these as varying degrees of MCI severity, such as Early and Late MCI diagnoses. Since they draw samples from different age groups, and consequently various other factors, we will explore the differences in progression to AD between those who begin with Unspecified and those who begin with MCI.

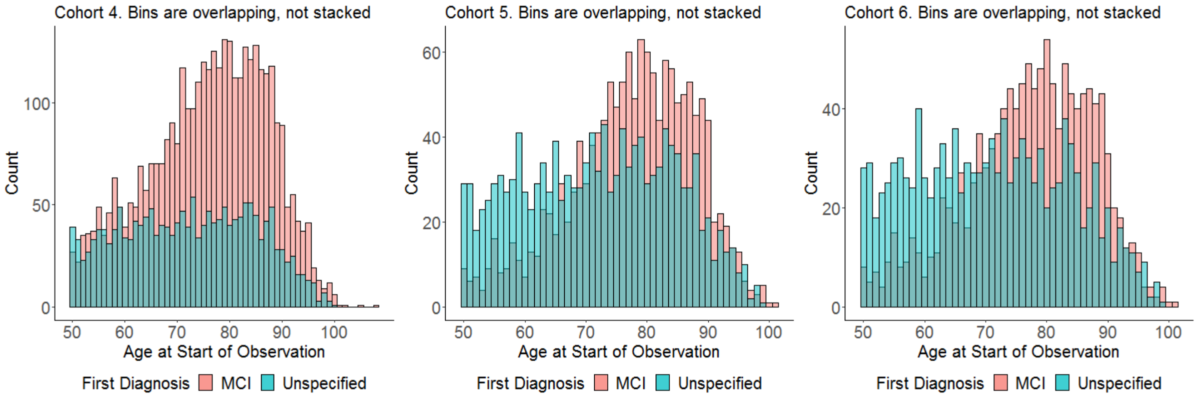


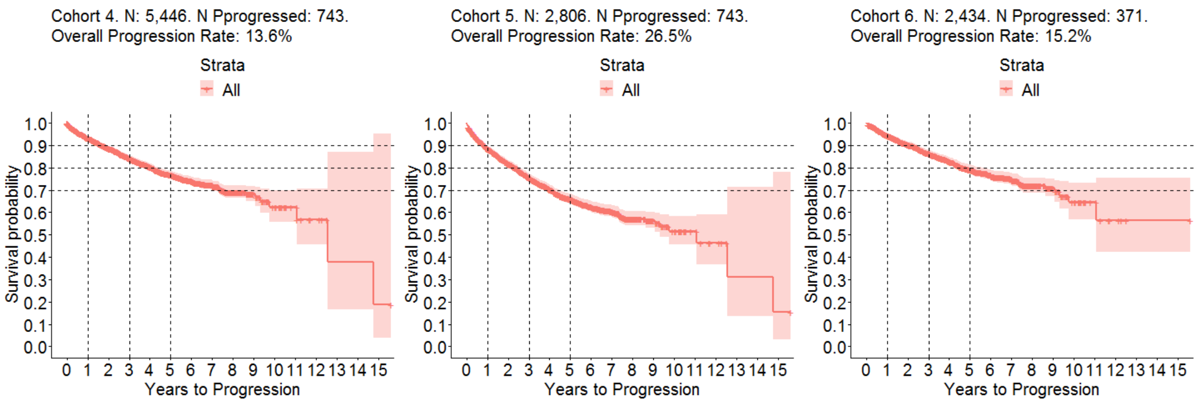
Figure 1: Breakdown of Age distribution

Kaplan Meir Curves

The subsequent essential step in preliminary data analysis involves examining the Kaplan-Meier survival curves. Similar to Figure 1, the first diagnosis, whether MCI or Unspecified, acts as a stratifying variable assessing differences in time-to-event distribution between the two groups, as depicted in Figure 3 for all three cohorts.

Immediately noticeable are patterns across cohorts akin to conclusions drawn from age distribution analysis. Cohort 4 demonstrates highly similar age distributions between MCI and Unspecified first diagnosis groups. It's expected that age emerges as a potent predictor of MCI to AD progression. Since age distributions visually appear similar across the two groups, observed differences in age distribution translate similarly into differences in time-to-AD progression estimates.

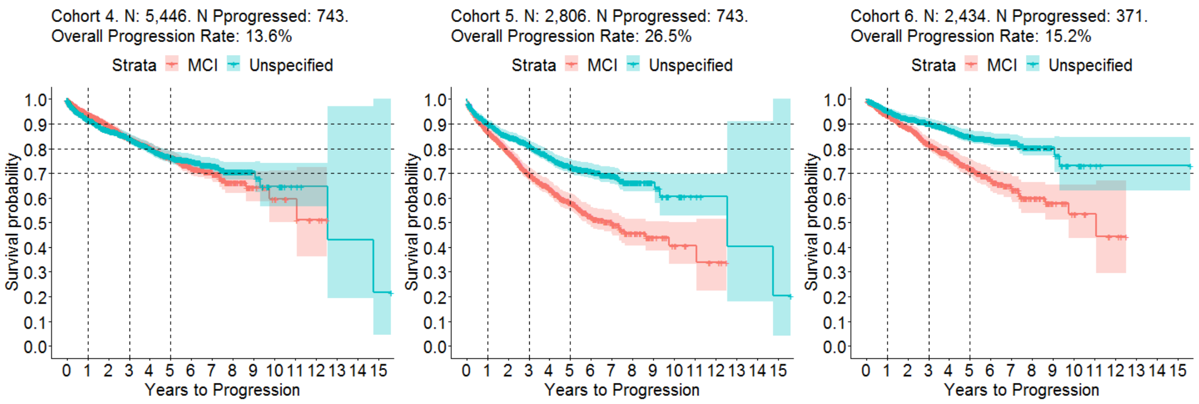
Significantly, Cohorts 4 and 6 exhibit quite comparable overall progression rates in the samples, likely due to similar overall progression rates in the data. However, the difference in the MCI/Unspecified subgroup becomes more apparent. Within the first 7-9 years of follow-up, Cohorts 4 and 6 appear overall indistinguishable, yet the scenario changes considerably when stratifying by the initial, or anchor, diagnosis. This crucial differentiation, coupled with largely disparate age distributions, suggests that while cohorts may seem similar on paper, the two sub-populations comprising the overall cohort differ significantly. The unadjusted progression disparities between groups with varying initial diagnosis severity imply that imposing stricter criteria to identify EHR patients with more confirmatory MCI onset yields 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 substantially differ based on initial diagnosis and age may facilitate statistical analyses with inference as the primary goal.


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 2: text


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 3: text

It is imperative to analyze the time-to-AD progression distribution across age groups since the three potential study cohorts varied in overall age distribution, as depicted in Figure 2 and Figure 4. All cohorts present an expected trend: older age groups demonstrate higher progression rates, a consistent conclusion across the three cohorts. Interestingly, there seems to be no additional age effect once participants enter the 71 and older group. It is plausible that the aging process had already predominantly occurred, with patients in the 71+ age group being quite homogeneous in terms of physical and mental health. While event rates decrease for the oldest study participants, this is likely attributed to censoring induced by death events, albeit an unverified speculative assumption from the EHR data.

It is pertinent to reiterate the similarities between Cohorts 4 and 6. Despite the differences in age distribution concerning the initial diagnosis and progression rates to AD when stratified by the initial diagnosis, the age effect on progression time does not seem to differ between the two groups when not considering other progression predictors. This absence of variation in progression time by age group suggests a reassuring resemblance between the populations representing Cohorts 4 and 6.

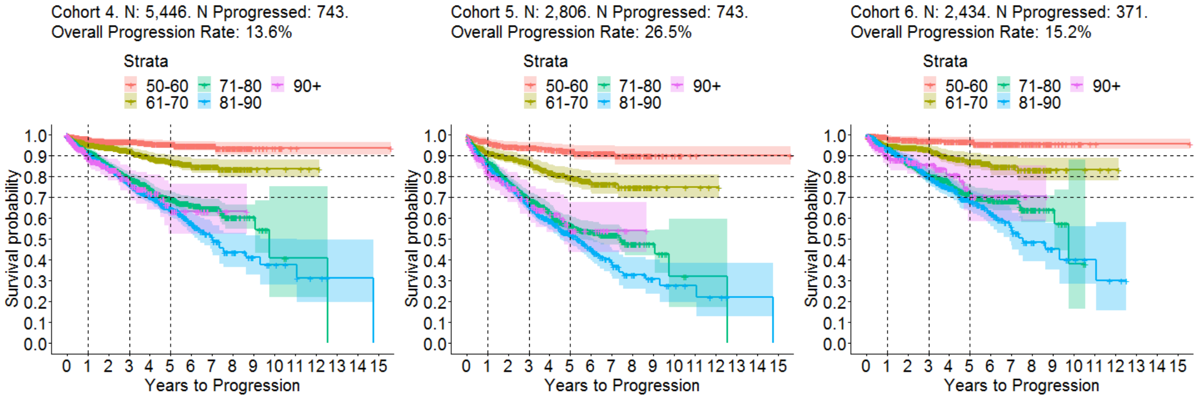


Figure 4: text

Results

Table 2 presents estimates from Cox Proportional Hazard regression models applied to the three cohorts utilizing the same set of predictors. Due to the large sample size and the straightforward nature of the model, all estimates are statistically significant. Nevertheless, the results of the three models exhibit significant variations.

The primary finding underscores the significance of the first diagnosis observation as a robust predictor of progression to AD. Across all cohorts, the occurrence of MCI as the first diagnosis increases the log-hazard (and hazard) of an AD event after adjusting for other factors. However, the magnitude of this effect varies considerably depending on the cohort. Cohorts with stricter inclusion criteria demonstrate a greater impact of the initial diagnosis on the log-hazard of AD events.

Table 2 also provides insights into the effect of age. For patients whose first diagnosis was 'Unspecified,' the effect of age remains consistent across all three cohorts. Each additional year of age at baseline is statistically associated with the progression to AD. However, in cohorts with stricter patient inclusion criteria, the effect of age at baseline on the log-hazard diminishes for those with MCI as the first diagnosis.

Discussion

Throughout the exploratory analysis and inference using Cox Proportional Hazard Models, we have observed that overall progression, progression by subgroups, and rates of progression within subgroups can heavily depend on the cohort inclusion criteria. The rationale for including stricter criteria is to ensure that 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. This latter scenario is more applicable to the MCI ICD diagnosis, while Unspecified diagnoses carry a higher risk of relating not to potential MCI presence itself, but to other mental issues patients experience. Given the lack of transparency and clinical notes in the EHR, it is imperative to ensure that selected patients represent the general population, allowing inferences about

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 entire population based on the observed patient subset.

Numerous studies utilize EHR data to forecast or predict AD presence after or within a certain period, or to 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 earlier, 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 that patient inclusion criteria have an immense effect on the population makeup and, therefore, a strong effect on the type and quality of knowledge inferred from such samples.

The issues discussed appear specific to how AD, especially MCI, is 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 alludes that certain tests display potential in detecting MCI and early dementia; however, 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 cutoff 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.

Regarding 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 the insights gained from analyzing Cohorts 4 and 6, excluding Cohort 5 momentarily, the primary aggregated metrics depicted in Table 1 exhibit remarkable similarity. Notably, the progression rate, defined as the observed progressions among cohort patients, appears strikingly alike across both cohorts. Upon examining the Kaplan-Meier survival curves, it becomes evident that the two cohorts boast nearly indistinguishable curves, particularly within the initial seven-year period where the bulk of the data lies. Additionally, both cohorts demonstrate similar trends in progression rates when stratified by age groups, with each cohort showing a heightened age effect on progression rates initially, tapering off around the age of 70.

However, a pertinent question arises: why does the impact of the first diagnosis differ significantly when comparing Cohorts 4 and 6? The application of highly restrictive criteria mandating at least two MCI (MCI or Unspecified) diagnoses results in a cohort necessitating follow-up after the initial diagnosis. Subsequent diagnoses confirm the presence of MCI or some unspecified mental condition. Individuals initially diagnosed with an Unspecified condition are likely at a lower risk of progression, presumably due to their better overall health status, hence receiving that diagnosis initially. Yet, why does the impact of the MCI first diagnosis appear much greater compared to the Unspecified diagnosis in Cohort 6? It is plausible that observations from the EHR lack credibility and fail to accurately capture the development of mental issues and the comprehensive medical history. In Cohort 4, individuals progressing to AD are permitted to have one MCI diagnosis. It is probable that this diagnosis is captured incidentally or that there was a scheduled follow-up appointment, but the individuals either did not attend or MCI was not documented during that appointment. By requiring at least two total diagnoses, not solely MCI-related as in Cohort 6, those progressing may only have one MCI-related diagnosis. This single diagnosis introduces more randomness and sporadic observations, thereby diluting the differences in progression between individuals with different initial diagnoses. Although Cohorts 4 and 6 exhibit similar overall progression rates consistent with accepted MCI to AD progression rates, they comprise distinct population compositions, potentially yielding different analytical outcomes.

To recapitulate, while patients in Cohort 6 are more assuredly confirmed onset MCI cases, an initial Unspecified diagnosis implies that physicians may have perceived less concrete evidence of MCI during the initial evaluation, perhaps due to the overall healthier status of the patients. This speculation is corroborated by a Cox proportional hazards regression coefficient. Consequently, we may have assembled a study cohort that, by design, segregates individuals with better health, less likely to progress, from those with poorer mental states at the initial diagnosis, who are more inclined to progress to AD from MCI/Unspecified mental diagnoses.

Recommendations

The initial consideration revolves around sample size. Adhering to Cohort 4's criteria permits the inclusion of a larger number of patients, potentially diversifying the types of patients present and offering data more reflective of the entire populace. An argument can be made that such an approach might be more suitable for machine learning tasks utilizing complex ~~overparameterized~~ models that require ample data samples to discern non-trivial relationships.

Although we discussed the possibility of having a single Unspecified diagnosis prior to AD progression, signifying a sporadic occurrence, and potentially enrolling someone into the cohort who may not represent likely progressing patients, this conclusion remains ambiguous from the data alone. Nonetheless, the suggestion would be to employ Cohort 4 when undertaking predictive model development, as this cohort maximizes the total sample size while maintaining overall statistics resembling a commonly referenced population when assessing MCI to AD progression.

Conversely, Cohort 6 could be more suitable for inference studies aimed at identifying factors associated with MCI to AD progression events. This assertion arises because Cohort 6 is structured in such a way that included individuals require at least two MCI diagnoses over a certain time frame (at least 50 days in our study). Having individuals more likely to actually undergo an anchor MCI event enhances credibility and confidence that Cohort 6's criteria identifies genuinely onset MCI patients. These observations can be treated as representative of the entire MCI population. Therefore, analyzing progression rates and factors associated with progression events/times using a Cohort 6-like sample appears more likely to yield marginal effect estimates covering true population effects.

Up to this point, Cohort 5 has been largely disregarded due to summary statistics. In reality, inclusion occurred as an illustration of potentially poor practices or manipulation of numbers in favor of the study conductor. By tightening non-AD event inclusion criteria but encompassing all AD events even with a single MCI diagnosis, one advantageous power increase emerges, with a higher chance of discovering novel factors delineating the prognosis of AD progression events. However, this comes with a cost and a decision that may be challenging to justify. On the surface, Cohort 5 seems akin to Cohorts 4 and 6, with age summary statistics, observation times, and specific diagnosis distributions not deviating significantly and aligning with overall understandings of onset MCI and AD progression risk populations. However, due to the 'hybrid' criteria, the overall progression rate appears excessively high, as reality indicates much lower rates, making it improbable to observe a 26% progression rate in a random sample when the true population rate is 15%.

This cohort illustrates how applying criteria to a portion of a sample can drastically alter the population the sample is intended to represent, potentially rendering the sample unrepresentative of any population. Our regression model estimate supports this argument, as the coefficient for the increased log-hazard associated with MCI being the first diagnosis falls between Cohorts 4 and 6.

Further Questions

An intriguing aspect emerges from our analysis regarding the sequencing of initial diagnoses as a stratifying variable: why confine ourselves solely to the first diagnosis? Could we not leverage all diagnoses encountered along the patient journey, using them as the most recent accumulated information to gauge the risk of transitioning to AD at subsequent observations or within a specified timeframe? This proposition aligns with our earlier discussions throughout this paper.

Consider the notion that an Unspecified diagnosis might be assigned when the presence of MCI isn't definitively clear, but indications of mental disturbances and memory issues are documented. In such cases, we can infer a lower confidence level regarding the likelihood of imminent progression to AD following an observation. However, what if a patient accrues multiple such diagnoses before transitioning to a diagnosis of MCI? It stands to reason that such individuals are at a heightened risk of AD progression, given the accumulation of systematically documented issues. Moreover, will it be reasonable to argue that a patient who incurs two Unspecified diagnoses followed by an MCI diagnosis is at a lower risk of progression compared to a person who incurs three MCI diagnoses in a shorter period of time?

The application of Markov chain models and G-computation approaches, commonly employed in causal inference, could provide valuable insights into analyzing such data and addressing the questions we pose in this section. These models and analysis frameworks facilitate the examination of observations over time, enabling the statistical evaluation of relationships between discrete time points and transitions from one state to another.

Of course, the successful implementation of these approaches necessitates not only a sizable dataset with a sufficient number of patients to discern complex patterns but also a robust set of predictors to enable accurate personalized predictions.

Summary

The present study aimed to explore and critically examine the methodologies employed in studying the progression from mild cognitive impairment (MCI) to Alzheimer's disease (AD) using electronic health records (EHRs). With the advent of EHRs revolutionizing the landscape of medical research by providing longitudinal patient data, we sought to delineate diverse cohorts mirroring common criteria observed in recent publications. Our goal was to assess the implications of restructuring cohort criteria on sample size, overall statistics, and the degree to which the represented sample aligns with commonly accepted statistics regarding MCI to AD progression.

We analyzed data from the Fairview Health System Data Warehouse, comprising records of patient visits with diagnoses of MCI, AD, or unspecified mental disorders. Three distinct cohorts were defined based on varying inclusion criteria, such as age restrictions, time between diagnoses, and the presence of confirmed MCI diagnoses across multiple visits.

Our analysis revealed that cohort inclusion criteria significantly impact the sample size, overall progression rates from MCI to AD, and the effect of key predictors like age and initial diagnosis type on disease progression. Notably, imposing stricter criteria to confirm MCI onset yielded a cohort heavily stratified by inherent AD progression risk, albeit with a reduced sample size.

We highlighted the subjectivity and challenges associated with MCI and AD diagnosis in clinical settings, which can introduce inconsistencies and incompleteness in EHR data capture. This underscores the importance of carefully considering patient inclusion criteria to ensure that the selected sample accurately represents the general population, thereby allowing for reliable inferences and recommendations.

Based on our findings, we provide recommendations for utilizing different cohort definitions based on the research objective. Cohort 4, with less stringent criteria, may be more suitable for machine learning tasks requiring larger sample sizes. In contrast, Cohort 6, with stricter criteria for confirming MCI onset, could be more appropriate for inference studies aimed at identifying factors associated with MCI to AD progression events.

In summary, this study underscores the significant impact of cohort selection criteria on the quality and generalizability of findings derived from EHR data when studying MCI and AD progression. Our work highlights the importance of carefully evaluating and reporting cohort definitions to ensure the validity and applicability of research outcomes in this field.

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

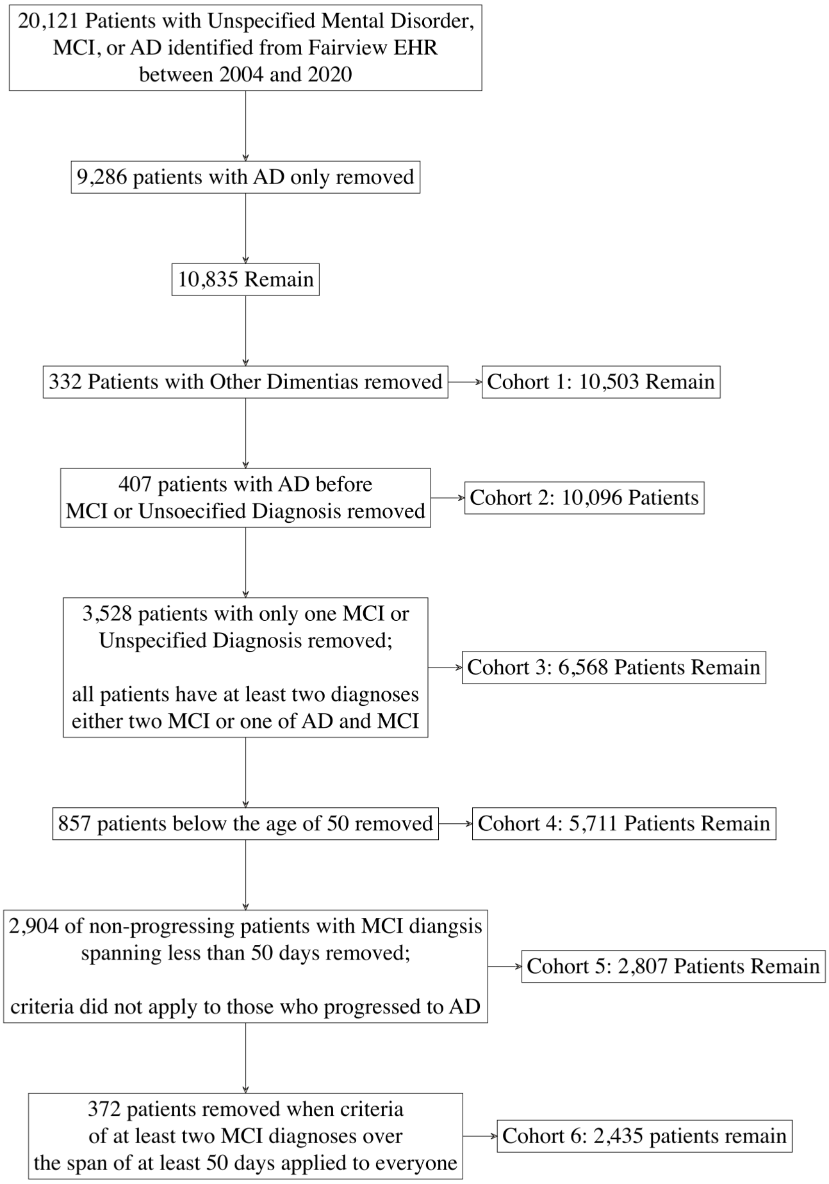

20,121 Patients with Unspecified Mental Disorder, MCI, or AD identified from Fairview EHR between 2004 and 2020
9,286 patients with AD only removed
10,835 Remain
332 Patients with Other Dimentias removed
Cohort 1: 10,503 Remain
407 patients with AD before MCI or Unsoecified Diagnosis removed
Cohort 2: 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

Table 3: Diagnoses from EHR

ICD Code	Version	Description
290.4	ICD9	VASCULAR DEMENTIA,UNCOMP
290.4	ICD9	Vascular dementia, uncomplicated
290.41	ICD9	VASC DEMENTIA W DELIRIUM
290.41	ICD9	Vascular dementia, with delirium
294.9	ICD9	MENTAL DISOR NOS OTH DIS
294.9	ICD9	Unspecified persistent mental disorders due to conditions classified elsewhere
331	ICD9	ALZHEIMER'S DISEASE
331	ICD9	Alzheimer's disease
331.19	ICD9	FRONTOTEMP DEMENTIA NEC
331.19	ICD9	Other frontotemporal dementia
331.82	ICD9	DEMENTIA W LEWY BODIES
331.82	ICD9	Dementia with lewy bodies
331.83	ICD9	MILD COGNITIVE IMPAIREMT
331.83	ICD9	Mild cognitive impairment, so stated
F01.50	ICD10	Vascular dementia without behavioral disturbance
F01.51	ICD10	Vascular dementia with behavioral disturbance
F09	ICD10	Unspecified mental disorder due to known physiological condition
G30.0	ICD10	Alzheimer's disease with early onset
G30.1	ICD10	Alzheimer's disease with late onset
G30.8	ICD10	Other Alzheimer's disease
G30.9	ICD10	Alzheimer's disease, unspecified
G31.01	ICD10	Pick's disease

G31.09	ICD10	Other frontotemporal dementia
G31.83	ICD10	Dementia with Lewy bodies
G31.84	ICD10	Mild cognitive impairment, so stated