

UCLA

UCLA Electronic Theses and Dissertations

Title

Leveraging Genetic and Real-World Data for Precision Understanding of Alzheimer's Disease

Permalink

<https://escholarship.org/uc/item/09x613z6>

ISBN

9798297637023

Author

Fu, Joy

Publication Date

2025-10-13

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

Leveraging Genetic and Real-World Data for Precision Understanding of Alzheimer's
Disease

A dissertation submitted in partial satisfaction
of the requirements for the degree
Doctor of Philosophy in Medical Informatics

by

Mingzhou Fu

2025

© Copyright by
Mingzhou Fu
2025

ABSTRACT OF THE DISSERTATION

Leveraging Genetic and Real-World Data for Precision Understanding of Alzheimer’s
Disease

by

Mingzhou Fu

Doctor of Philosophy in Medical Informatics

University of California, Los Angeles, 2025

Professor Timothy Chang, Chair

Alzheimer’s disease (AD) represents one of the most pressing challenges in modern medicine, affecting over 6.7 million individuals in the United States with projections reaching 13 million by 2050. Despite significant advances in understanding individual risk factors, current approaches to AD research and clinical care face three fundamental limitations: the lack of robust methods for quantifying disease relationships, insufficient understanding of temporal disease progression patterns, and persistent health disparities in genetic risk assessment that exclude underrepresented populations from precision medicine benefits.

This dissertation presents an integrated framework for precision medicine in Alzheimer’s disease that addresses these limitations through three interconnected studies spanning computational methodology, temporal analysis, and health equity. The overarching goal is to develop predictive, preventive, personalized, and participatory approaches to AD research that can transform how we understand, predict, and prevent this devastating condition.

The first study established a foundational framework for quantifying disease relationships using SNOMED CT embeddings [1]. By integrating semantic, structural, comorbidity-based, and genetic correlation-based distance metrics, this work demonstrated that embedding-based approaches significantly outperform traditional methods in capturing clinically mean-

ingful disease relationships. This semantic framework provides the essential foundation for meaningful trajectory analysis and clinical decision support systems.

Building upon this foundation, the second study identified distinct temporal pathways leading to Alzheimer’s disease [2]. Four trajectory clusters were discovered: mental health, encephalopathy, mild cognitive impairment/neurodegenerative, and vascular cluster. These clusters exhibited significant differences in demographic characteristics, comorbidity patterns, and progression rates. Furthermore, the confirmation that multi-step trajectories confer higher risk than individual diagnoses underscores the importance of considering sequential disease progression in AD risk assessment.

The third study addressed critical health disparities in genetic research by developing machine learning approaches for dementia risk modeling in underrepresented populations [3]. Traditional polygenic risk scores showed poor performance in non-European populations, particularly African Americans where they performed worse than simple *APOE-ε4* counting. The novel Elastic Net SNP models incorporating functional genomic information achieved substantial improvements while identifying both shared and ancestry-specific risk factors. Validation in the All of Us Research Program confirmed the generalizability of this approach across diverse healthcare systems.

The integration of these three studies creates a comprehensive framework that spans from semantic disease understanding through temporal progression modeling to equitable genetic risk assessment. This framework enables trajectory-based risk stratification years before clinical symptoms appear, provides specific targets for personalized prevention strategies, and ensures that precision medicine benefits are accessible to all populations regardless of ancestry.

The clinical implications are immediate and transformative. The trajectory patterns enable clinicians to identify high-risk patients and implement targeted interventions based on specific progression pathways. The genetic risk models provide accurate assessment in diverse populations, supporting personalized prevention strategies. The semantic framework

supports clinical decision-making through quantified disease relationships and automated risk assessment systems.

Beyond Alzheimer’s disease, this work contributes methodological innovations applicable to other complex diseases with multi-step progression patterns. The semantic embedding approach provides a generalizable framework for disease relationship modeling, the trajectory analysis methodology can be applied to other chronic conditions, and the equity-focused genetic modeling addresses fundamental limitations in precision medicine more broadly.

This dissertation establishes foundations for a future where Alzheimer’s disease prevention is predictive (identifying risk years before symptoms), preventive (targeting specific progression pathways), personalized (accounting for individual genetic and clinical profiles), and participatory (including all populations in research and clinical benefits). The substantial lead times identified for intervention, combined with specific biological targets and equitable risk assessment approaches, provide unprecedented opportunities to transform the clinical course of this devastating disease while ensuring that precision medicine advances benefit all populations equitably.

The dissertation of Mingzhou Fu is approved.

Bogdan Pasaniuc

Keith Alan Vossel

Loes Marlein Olde Loohuis

Timothy Chang, Committee Chair

University of California, Los Angeles

2025

To God be the glory

TABLE OF CONTENTS

1	Introduction	1
1.1	The Grand Challenge of Alzheimer’s Disease	1
1.2	The Power and Challenge of Real-World Data	2
1.3	A Multi-Faceted Approach to Alzheimer’s Disease Research	4
1.4	Research Contributions and Dissertation Outline	5
2	Defining the Distance Between Diseases Using SNOMED CT Embeddings	8
2.1	Introduction	8
2.1.1	The Challenge of Quantifying Disease Relationships	8
2.1.2	Clinical Terminologies and SNOMED CT	9
2.1.3	Knowledge Graph Embeddings in Healthcare	9
2.1.4	Chapter Objectives and Contributions	10
2.2	Background and Related Work	10
2.2.1	Disease Similarity Measures	10
2.2.2	Clinical Terminologies and Embedding Techniques	11
2.2.3	Related Work in Medical Concept Embeddings	12
2.3	Methods	12
2.3.1	Data Sources and Preprocessing	12
2.3.2	SNOMED CT Embedding Generation	13
2.3.3	SNOMED CT Embedding-Based Distance Metric	13
2.3.4	ICD Tree-Based Distance Metric	14
2.3.5	Comorbidity-Based Distance Metric	15

2.3.6	Genetic Correlation-Based Distance Metric	16
2.3.7	Validation of Embedding Distance Metric	17
2.3.8	Comparisons of Embedding vs. Other Distance Metrics	18
2.4	Results	19
2.4.1	Development of Distance Metrics for Comparisons of Diseases	19
2.4.2	Validation of Embedding Distance Metric	20
2.4.3	Comparisons Across Different Distance Metrics	21
2.5	Discussion	24
2.6	Conclusion	26
2.7	Code and Data Availability	27
2.8	Acknowledgments	27
2.9	Figures	29
2.10	Supplementary materials	37
3	Identifying Common Disease Trajectories of Alzheimer’s Disease with Elec- tronic Health Records	40
3.1	Introduction	40
3.1.1	The Challenge of Understanding Disease Progression in Alzheimer’s Diseases	40
3.1.2	Disease Trajectories: A Temporal Perspective on Health and Disease	41
3.1.3	Electronic Health Records as a Window into Disease Progression . . .	41
3.1.4	Building on Disease Distance Metrics: From Static Relationships to Dynamic Progressions	41
3.1.5	Current Limitations in Disease Trajectory Analysis	42

3.2	Background and Related Work	42
3.3	Methods	43
3.3.1	Sample Selection in the UC Health Data Warehouse	43
3.3.2	Pairwise ICD Codes Associations in a Sampled Population	45
3.3.3	Patient Trajectories Preprocessing	46
3.3.4	Common AD Trajectory Identification	47
3.3.5	Evaluation of Common AD Trajectories	48
3.3.6	Causal Structural Learning	49
3.3.7	Sensitivity Analyses	49
3.3.8	Validation in the All of Us Research Program	51
3.3.9	Ethics	52
3.4	Results	52
3.4.1	Pairwise ICD Code Temporal Associations	52
3.4.2	Common AD Trajectory Identification	53
3.4.3	Evaluation of Common AD Trajectories	55
3.4.4	Risk Trajectories Test in Controls	57
3.4.5	Causal Structural Learning	57
3.4.6	Sensitivity Analyses	58
3.4.7	Validation in the All of Us	59
3.5	Discussion	61
3.6	Conclusions	64
3.7	Code and data availability	65
3.8	Acknowledgments	66

3.9	Figures	67
3.10	Supplementary materials	77
4	Improving Genetic Risk Modeling of Dementia in Underrepresented Populations	88
4.1	Introduction	88
4.1.1	The Promise and Peril of Precision Medicine in Dementia	88
4.1.2	The Burden of Dementia in Underrepresented Populations	88
4.1.3	Limitations of Current Genetic Risk Modeling Approaches	89
4.1.4	The Potential of Biobank-Linked Electronic Health Records and Machine Learning Approaches to Genetic Risk Modeling	90
4.2	Background and Related Work	91
4.2.1	The Evolution of Genetic Risk Assessment in Dementia	91
4.2.2	Polygenic Risk Scores: Promise and Limitations	91
4.2.3	Functional Genomics and Causal Variant Identification	91
4.2.4	Multi-Disease Approaches to Genetic Risk Modeling	92
4.3	Methods	93
4.3.1	Study Design and Participants	93
4.3.2	Patient Genetic Data Preprocessing	93
4.3.3	Genetic Predictors	94
4.3.4	Dementia Case Definition and Control Selection	95
4.3.5	Statistical Modeling and Machine Learning Approaches	96
4.3.6	Model Evaluation and Comparison	97
4.3.7	Gene Mapping and Pathway Analysis	98

4.3.8	Validation in the All of Us Research Program	98
4.3.9	Statistical Analysis and Software	99
4.4	Results	99
4.4.1	Sample Description	99
4.4.2	Performance Comparison for Dementia Phenotype Prediction Task	100
4.4.3	Featured Risk Variants and Mapped Genes	102
4.4.4	Validations in the All of Us Sample	104
4.5	Discussion	104
4.6	Conclusions	109
4.7	Code and data availability	111
4.8	Acknowledgments	111
4.9	Figures	112
4.10	Supplementary materials	118
5	Summary and Future Directions	129
5.1	Summary of Contributions	129
5.1.1	Defining Disease Relationships Through Semantic Embeddings	129
5.1.2	Identifying Disease Trajectories Leading to Alzheimer’s Disease	130
5.1.3	Advancing Health Equity in Genetic Risk Modeling	131
5.2	Methodological Integration	132
5.3	Clinical Implications	132
5.4	Limitations	133
5.5	Future Research Directions	134
5.5.1	Methodological Advances	134

5.5.2	Clinical Translation	134
5.5.3	Health Equity Applications	135
5.6	Conclusion	135

LIST OF FIGURES

2.1	SNOMED CT embedding-based distance metric	30
2.2	Distributions and summary statistics of distance metrics	31
2.3	Validation of embedding distance metric - Heatmaps of embedding distance . . .	32
2.4	Validation of embedding distance metric - Correspondence of ICD clinical chapter and predicted modules (hierarchical clustering with dynamic cut tree model) . .	33
2.5	Associations of embedding distances vs. genetic (A) and comorbidity (B) distances.	34
2.6	Comparison of embedding-based and ICD-tree distance (A).	35
2.7	Comparison of embedding-based and ICD-tree distance (B).	36
3.1	Study design	68
3.2	Risk factors of Alzheimer’s disease (AD) identified by the Fine-Gray proportional subdistribution hazards model, patients with AD (cases) and 1:1 matched controls (N = 48,946)	69
3.3	Summary of clusters of the Alzheimer’s disease trajectory	70
3.4	Summary of clusters of the Alzheimer’s disease trajectory	71
3.5	Venn diagram of Alzheimer’s disease patients in trajectory clusters	72
3.6	Alzheimer’s disease (AD) patient characteristics by trajectory clusters, patients with unique trajectory cluster only (N = 4,078)	73
3.7	Kaplan–Meier survival curve in Alzheimer’s disease (AD) patients by trajectory cluster	74
3.8	Distributions of time (in years) from central nodes to first Alzheimer’s disease diagnosis by trajectory cluster	75

3.9	Characteristics of patients with Alzheimer’s Disease (AD) that have clustered in the mild cognitive impairment cluster, single cluster vs. multi clusters	76
3.10	Associations between risk trajectories and incidence of Alzheimer’s disease	76
4.1	Sample selection steps and modeling steps description	113
4.2	Dementia patient characteristics by genetic inferred ancestry groups, UCLA ATLAS sample	113
4.3	Descriptive statistics of demographic and electronic health record features by case/control groups, UCLA ATLAS sample, stratified by genetic inferred ancestry group	114
4.4	Model performance of <i>APOE</i> -e4 count, polygenic risk score, and SNP models in dementia genetic prediction, UCLA ATLAS sample, stratified by genetic inferred ancestry	115
4.5	Top 10 featured risk SNPs from the best-performing Elastic Net SNP model, UCLA ATLAS sample, stratified by genetic ancestry	116
4.6	Descriptive statistics of demographic and electronic health record features by case/control groups, UCLA ATLAS sample, stratified by genetic inferred ancestry group	117
4.7	Overall model performance of <i>APOE</i> -e4 count, polygenic risk score, and Elastic Net SNP models in dementia genetic prediction in validation of All of Us sample, stratified by genetic inferred ancestry	117

ACKNOWLEDGMENTS

No PhD journey is traveled alone, and I am deeply grateful to the many people who have made this work possible.

To my advisory committee - thank you for your guidance, wisdom, and patience. Dr. Timothy Chang, you believed in my vision when it was just a collection of ambitious ideas and helped me transform them into rigorous science. Dr. Loohuis, you expanded my computational horizons and showed me the power of methodological innovation. Dr. Pasaniuc, you challenged me to think deeper about health equity and the social implications of our work. Dr. Vossel, you kept me grounded in clinical reality and reminded me that behind every data point is a human being.

To my research collaborators - this work would not exist without you. The UCLA Health System and the UC Data Warehouse provided not just data, but the clinical context that made this research meaningful. The All of Us Research Program gave us the opportunity to validate our findings in the diverse populations that precision medicine must serve. My laboratory colleagues shared countless hours of discussion, debugging, and discovery that shaped every aspect of this work.

To my family - you are my foundation. My parents instilled in me the values of service, perseverance, and compassion that drive this research. My siblings keep me grounded and remind me what truly matters beyond academic achievement. My extended family has provided prayers, encouragement, and unwavering support through every challenge and celebration.

To my friends - you are my community. My graduate school cohort shared this journey of discovery and growth, celebrating victories and supporting each other through setbacks. My church family provided spiritual support and constantly reminded me of my calling to serve others. My lifelong friends celebrated every milestone and provided comfort during the most challenging moments.

To my mentors and teachers - you opened my mind and heart to possibilities I never imagined. My former mentors Dr. Kelly Bakulski and Dr. Erin Ware sparked my passion for research and critical thinking. Clinical mentors showed me the human face of the diseases we study in our algorithms. Senior researchers opened doors and shared wisdom from their own journeys of discovery.

Most importantly, to the patients - every individual whose health journey is represented in our datasets made this research possible. The families affected by Alzheimer's disease inspire the urgency and purpose of this work. And the future patients who will benefit from more equitable and precise medicine give meaning to every late night and every line of code.

As Philippians reminds us, 'I can do all things through Christ who strengthens me.' This work is ultimately not about my achievements, but about how we can serve others through the gifts we've been given.

Thank you all for being part of this journey.

VITA

2012–2017	BS. in <i>Clinical Medicine</i> , School of Medicine, Shanghai Jiao Tong University, Shanghai, China.
2017–2020	MPH and MHI, <i>Public Health and Health Informatics</i> , University of Michigan, Ann Arbor, USA.

PUBLICATIONS

(The list includes only those works where I am the lead author during my Ph.D. studies.)

Fu M, Yan Y, Olde Loohuis LM, Chang TS. Defining the distance between diseases using SNOMED CT embeddings. *Journal of Biomedical Informatics*. 2023 Mar 1;139:104307.

Fu M, Sankararaman S, Pasaniuc B, Vossel K, Chang TS. Identifying common disease trajectories of Alzheimer’s disease with electronic health records. *eBioMedicine*. 2025 Aug 1;118:105831.

Fu M, Valiente-Banuet L, Wadhwa SS, Pasaniuc B, Vossel K, Chang TS. Improving genetic risk modeling of dementia from real-world data in underrepresented populations. *Communications Biology*, 2024 Aug 25;7(1):1–12

CHAPTER 1

Introduction

1.1 The Grand Challenge of Alzheimer’s Disease

Alzheimer’s Disease (AD), a progressive neurodegenerative disorder, stands as one of the most significant health challenges of our time. It is the leading cause of dementia worldwide, insidiously eroding an individual’s memory, cognitive functions, and ultimately, their capacity for independent life [4]. The global burden of AD is immense and growing; it is estimated that over 55 million people live with dementia, a number projected to rise to 139 million by 2050 [4]. Beyond its devastating personal impact, AD places an enormous strain on families, healthcare systems, and economies, with the global cost of dementia estimated to be in the trillions of dollars [5].

Despite decades of intensive research, the underlying mechanisms of AD remain only partially understood [5]. The disease is characterized by a complex and heterogeneous etiology, involving a confluence of genetic predispositions, environmental factors, and age-related changes [5]. This complexity is reflected in its clinical presentation; patients exhibit significant variability in their age of onset, rate of cognitive decline, and the sequence in which different comorbidities appear [6]. This heterogeneity poses a major obstacle to the development of effective treatments, as a ”one-size-fits-all” approach is unlikely to succeed [5].

The complexity of AD is further compounded by significant health disparities that affect both our understanding of the disease and the equitable delivery of care. While AD affects individuals across all demographic groups, there are notable differences in disease prevalence,

progression, and outcomes across racial and ethnic populations [7]. African Americans and Hispanic Americans, for example, are at higher risk for developing AD compared to non-Hispanic whites, yet they remain significantly underrepresented in clinical research and genetic studies [8]. This underrepresentation has profound implications for the development of precision medicine approaches, as genetic risk prediction models developed primarily in populations of European ancestry may not accurately predict disease risk in other populations [9].

Therefore, addressing the challenge of AD requires a dual approach: first, developing sophisticated computational methods to understand the heterogeneous progression patterns of the disease, and second, ensuring that these advances are developed and validated in diverse populations to promote health equity. Understanding common "disease trajectories" is paramount for developing targeted therapies and enabling earlier diagnosis, but these insights must be coupled with inclusive research practices that ensure the benefits of precision medicine are accessible to all populations.

1.2 The Power and Challenge of Real-World Data

The digital transformation of healthcare has ushered in an era of "big data," presenting unprecedented opportunities to study complex diseases like Alzheimer's at a massive scale [10]. Real-world data (RWD), particularly information contained within electronic health records (EHRs), offers a longitudinal, high-resolution view of the patient experience [10, 11]. EHRs capture a vast array of clinical information—including diagnoses, medications, laboratory results, and physician's notes—creating a rich tapestry of data that reflects the patient's journey through the healthcare system [11]. This data is an invaluable resource for moving beyond the controlled, and often limited, environment of traditional clinical trials to understand how diseases manifest and progress in large, diverse populations [10].

The advent of large-scale biobanks that link genetic information to de-identified EHR

data has created a particularly powerful resource for investigating the interplay between genetic risk and clinical phenotype [12]. The Electronic Medical Records and Genomics (eMERGE) Network, for example, represents a consortium of institutions that have successfully integrated genomic data with longitudinal clinical records from diverse patient populations [12]. These biobank-linked EHR systems enable researchers to study genetic associations across different ancestral backgrounds and to develop more inclusive approaches to precision medicine [13].

Importantly, many of these biobank initiatives have made deliberate efforts to include participants from underrepresented populations, creating opportunities to address historical disparities in genetic research [14]. For instance, the All of Us Research Program and similar initiatives have prioritized the recruitment of diverse participants, recognizing that genetic discoveries made in one population may not translate effectively to others [15]. This represents a critical shift toward more equitable research practices that can benefit all populations.

However, the sheer volume and complexity of RWD also present significant analytical challenges [11, 16]. The data is often unstructured, noisy, and contains missing values, requiring sophisticated computational methods to extract meaningful insights [16, 17]. Furthermore, to build truly comprehensive models of disease that can serve diverse populations, clinical data from EHRs must be integrated with genomic data in ways that account for population structure and genetic ancestry [9].

Applying advanced machine learning and artificial intelligence (AI) methodologies to these integrated datasets is essential to unlock their full potential, allowing researchers to identify patterns and relationships that would be impossible to discern through manual analysis [18]. However, these computational approaches must be developed and validated with careful attention to issues of bias, fairness, and generalizability across different populations [19]. This dissertation develops and applies such a computational framework to address both the methodological challenges and equity concerns in understanding Alzheimer’s Disease.

1.3 A Multi-Faceted Approach to Alzheimer’s Disease Research

This dissertation addresses the complexity of Alzheimer’s Disease through two complementary but distinct research directions, each tackling fundamental challenges in the field. The first direction focuses on developing novel computational methods to understand disease progression patterns, while the second addresses critical issues of health equity in dementia research.

Computational Framework for Disease Trajectory Analysis

To deconstruct the heterogeneity of Alzheimer’s Disease progression, this work proposes and applies a novel computational framework that operates on two fundamental principles: first, that the relationships between diseases can be quantitatively measured from clinical data, and second, that these relationships can reveal common underlying progression pathways. Traditional approaches often study diseases in isolation or rely on predefined comorbidity indices, which may not capture the nuanced, temporal relationships between different clinical events as they unfold over a patient’s lifetime [20]. To overcome this limitation, a new approach is needed to represent clinical concepts in a way that preserves their rich semantic meaning.

This work introduces a method to create dense, vector-based representations (embeddings) of medical concepts from the Systematized Nomenclature of Medicine Clinical Terms (SNOMED CT) ontology [21]. By representing diseases as vectors in a high-dimensional space, the ”distance” and ”relatedness” between them can be calculated mathematically. This provides a data-driven, objective measure of the relationships between tens of thousands of clinical concepts, forming the foundational layer of our analytical framework. With this foundation, the framework is then applied to the central challenge of identifying common disease trajectories in AD. By analyzing the sequence of diagnoses in the EHRs of a large cohort of AD patients, and leveraging the disease distance metric, we can map the pathways patients travel on their journey with the disease. This allows for the identifica-

tion of distinct patient subgroups who share common patterns of comorbidities and disease progression, moving beyond a single, monolithic view of AD.

Addressing Health Equity in Dementia Research

Parallel to the methodological advances in disease trajectory analysis, this dissertation also addresses a critical gap in dementia research: the underrepresentation of diverse populations in genetic studies. Most genetic research on Alzheimer’s Disease has been conducted primarily in populations of European ancestry, leading to significant disparities in the applicability of genetic risk models across different ethnic and racial groups [22]. This represents a fundamental issue of health equity, as genetic risk scores developed in one population may not accurately predict disease risk in others, potentially exacerbating existing healthcare disparities.

To address this challenge, this work develops and validates improved genetic risk models for dementia using real-world data from underrepresented populations. By leveraging large-scale biobank data that includes diverse patient populations, we can create more inclusive and equitable approaches to genetic risk prediction. This work demonstrates how computational methods can be applied not only to advance our scientific understanding of disease mechanisms, but also to promote health equity and ensure that the benefits of precision medicine are accessible to all populations.

1.4 Research Contributions and Dissertation Outline

This dissertation makes three primary contributions to the field of biomedical informatics and Alzheimer’s Disease research, addressing both methodological innovation and health equity concerns:

Contribution 1: Novel Disease Relationship Quantification

The first contribution develops a foundational methodology for quantifying the relation-

ships between diseases using SNOMED CT embeddings. This work provides a new tool for clinical data analysis that is both scalable and semantically rich, enabling a more nuanced understanding of disease comorbidities. Unlike traditional approaches that rely on predefined indices or simple co-occurrence statistics, this method captures the rich semantic relationships encoded in clinical terminologies.

Contribution 2: Disease Trajectory Identification in Alzheimer’s Disease

The second contribution applies this novel methodology to identify and characterize common, clinically-relevant disease trajectories in Alzheimer’s Disease from a large-scale EHR dataset. This is the central methodological contribution of the dissertation, providing new insights into the heterogeneity of AD and identifying distinct patient subgroups with different progression pathways. This work has significant implications for precision medicine, patient stratification in clinical trials, and early diagnosis by moving beyond a one-size-fits-all approach to AD understanding.

Contribution 3: Advancing Health Equity in Dementia Genetics

The third contribution addresses the critical issue of health disparities in dementia research by developing and validating improved genetic risk models for dementia in underrepresented populations. This work tackles a fundamental challenge in precision medicine: ensuring that genetic risk prediction tools are equitable and applicable across diverse populations. By leveraging real-world data from underrepresented groups, this study demonstrates how computational approaches can be used to promote health equity and reduce disparities in dementia care.

The remainder of this dissertation is structured as follows:

Chapter 2 details the development of the SNOMED CT embedding methodology for defining the distance between diseases, establishing the foundational tool for disease relationship quantification.

Chapter 3 presents the application of this methodology to identify common disease

trajectories in Alzheimer’s Disease and discusses the clinical implications of these findings for precision medicine and patient care.

Chapter 4 describes the parallel work on improving genetic risk modeling for dementia in underrepresented populations, highlighting the importance of health equity in precision medicine research.

Chapter 5 provides a synthesis of the key findings from both research directions, discusses the overall impact and limitations of the work, and proposes future research directions that can build upon both the methodological innovations and health equity advances presented in this dissertation.

CHAPTER 2

Defining the Distance Between Diseases Using SNOMED CT Embeddings

2.1 Introduction

2.1.1 The Challenge of Quantifying Disease Relationships

Studying disease relationships has been a growing interest in biomedical research [1]. Understanding disease relationships, including complex disease co-occurrences, causal relationships as well as temporal characteristics, is expected to provide in-depth insights into the human disease etiology [23, 24]. Disease distance, a quantitative metric of disease dissimilarity, is an essential disease relationship that serves as a cornerstone of various computational tasks that study relationships between diseases. For example, disease distance is needed when quantifying distances between patients’ disease trajectories (i.e., the temporal order of disease), which can be used in clustering patients to identify disease subtypes and common disease trajectories [25]. Identified patient clusters and common disease trajectories can then be used for the prediction of the disease onset and progression over time, and in return, can potentially facilitate early diagnosis and improve decision-making in clinical settings [26, 27].

Traditional approaches to measuring disease similarity have relied on predefined comorbidity indices, co-occurrence statistics, or hierarchical relationships within classification systems [24, 28]. However, these methods often fail to capture the nuanced, semantic relationships between diseases that are encoded in clinical terminologies and may not reflect the

complex, multifaceted nature of disease relationships observed in clinical practice.

2.1.2 Clinical Terminologies and SNOMED CT

The development of formal disease and research ontology databases, such as The Human Phenotype Ontology [29] and Systematized Nomenclature of Medicine, Clinical Terms (SNOMED CT) [21], have enabled the development of methods to define and evaluate disease distance metrics. SNOMED CT, in particular, represents the most comprehensive clinical terminology system available, providing a systematic collection of biomedical concepts and relations under a common ontological framework [21].

Given its polyhierarchical structure, SNOMED CT can be treated as a knowledge graph, where medical concepts are represented as nodes and their relationships as edges [30]. This rich semantic structure captures not only hierarchical relationships but also complex associative relationships between medical concepts, making it an ideal foundation for developing sophisticated disease similarity measures.

2.1.3 Knowledge Graph Embeddings in Healthcare

The distance metric we propose in this work relies on knowledge graph embeddings [30]. Knowledge graph embeddings represent a powerful approach for learning numerical representations that capture the semantic relationships encoded in structured knowledge bases. Specifically, knowledge graph embeddings can learn a low-dimensional numeric model of a knowledge graph that maps input relations to an output score indicating the likelihood of the relation being true, such that the high-dimensional graph representations of medical concepts are projected into a continuous vector space [31, 32].

These representations, known as embeddings, are more computationally manageable while preserving the inherent structure of the knowledge graph. Recent advances in knowledge graph embedding methods, such as RotatE [33], have shown promising capabilities in

learning biomedical knowledge representation and inferring phenotype relationships [30].

2.1.4 Chapter Objectives and Contributions

The growing volume of Electronic Health Records (EHRs), longitudinal clinical data collected during routine delivery of healthcare, has made the EHR a paramount source of phenotypic information. However, none of the previously discussed disease distance metrics have been broadly applied to International Classification of Diseases, version 10 (ICD-10) diagnostic codes, the most commonly used code in EHR systems [34]. Moreover, there is limited availability of well-powered genetic studies for the majority of ICD codes, limiting the use of genetically defined disease distances.

To address these barriers, this chapter presents a novel approach to defining disease distances using knowledge graph-based embeddings in the ICD-10 code format. The specific objectives of this work are to:

1. Develop a scalable methodology for computing disease distances based on SNOMED CT embeddings
2. Validate the semantic coherence and clinical relevance of the proposed distance metric
3. Compare the embedding-based approach with existing distance measures
4. Demonstrate the utility of this approach as a foundation for disease trajectory analysis

2.2 Background and Related Work

2.2.1 Disease Similarity Measures

Existing approaches to quantifying disease similarity can be broadly categorized into several types. Co-occurrence-based approaches rely on the frequency with which diseases appear together in clinical datasets, often using measures such as relative risk or odds ratios [24]. While intuitive, these methods are limited by the availability and quality of clinical data and

may not capture underlying biological relationships.

Ontology-based similarity measures leverage the hierarchical structure of disease classification systems, computing similarity based on the position of diseases within taxonomic trees [35]. The Human Disease Ontology [36], based on community-driven standards that focus on representing disease concepts captured across biomedical resources, was used to evaluate disease distances based on semantic similarity of disease-specific textual annotations [37]. However, these approaches are often limited by the depth and completeness of the ontological structure.

Network-based methods construct disease networks based on various types of relationships (genetic, proteomic, metabolomic) and use network topology measures to quantify similarity [38, 39]. While powerful, these approaches require extensive multi-omics data that may not be available for all diseases.

2.2.2 Clinical Terminologies and Embedding Techniques

SNOMED CT organizes disease concepts by grouping related terms into clinical chapters, with each subsequent character providing greater specificity [21]. The terminology includes concept identifiers, fully specified names, synonyms, and a rich set of relationships that capture various types of associations between medical concepts. This hierarchical and relational structure makes SNOMED CT particularly well-suited for embedding-based approaches.

The polyhierarchical nature of SNOMED CT means that concepts can have multiple parent concepts, reflecting the complex, multifaceted nature of medical knowledge. This structure allows for more nuanced representations of disease relationships compared to simple taxonomic hierarchies.

Knowledge graph embedding methods have emerged as powerful tools for learning representations of structured knowledge. These methods can be broadly categorized into translational models (such as TransE, TransH, TransR), semantic matching models (such as

RESCAL, DistMult), and neural network-based approaches [31].

RotatE, the embedding method used in this work, models relations as rotations in complex vector space, enabling it to handle various relation patterns including symmetry, anti-symmetry, inversion, and composition [33]. This flexibility makes it particularly suitable for modeling the complex relationships found in biomedical knowledge graphs.

2.2.3 Related Work in Medical Concept Embeddings

Prior research has explored the application of embedding techniques to medical concepts. Chang et al. [30] trained several knowledge graph embedding models on SNOMED CT and Human Phenotype Ontology knowledge graphs, demonstrating promising capabilities in learning biomedical knowledge representation. However, the application of these embeddings to define disease distances for ICD-10 codes, which are commonly used in EHR systems, remained unexplored.

Other approaches have used word embedding techniques on clinical text [40] or applied matrix factorization methods to patient-disease matrices [26]. However, these approaches do not leverage the rich semantic structure encoded in clinical terminologies like SNOMED CT.

2.3 Methods

2.3.1 Data Sources and Preprocessing

Our methodology relies on the SNOMED CT ontology, which provides a systematic collection of biomedical concepts and relations under a common ontological framework [21]. We first mapped three-digit ICD-10 codes with SNOMED CT concepts using Athena [41], an automated system for creating and maintaining the OMOP Standard Vocabularies [42].

The mapping process revealed that multiple SNOMED CT concepts can be mapped to a single ICD-10 code and vice versa, reflecting the different levels of granularity and

organizational principles between the two coding systems. A total of 1,619 three-digit level ICD-10 codes were successfully mapped to one to five SNOMED CT concepts. We limited our study to the first three digits of the ICD-10 codes (i.e., disease category) to: 1) represent EHR phenotypes at a general diagnosis level of granularity, 2) align with results found from other literature (e.g., genetic results from UK Biobank), and 3) improve computing efficiency.

2.3.2 SNOMED CT Embedding Generation

We retrieved pre-trained SNOMED CT embeddings provided by Chang et al. [30] to represent each SNOMED CT concept. These embeddings were generated using multiple knowledge graph embedding approaches, and we chose the embedding representations from their RotatE model based on its accurate and stable performance over multiple link prediction tasks.

The RotatE model treats SNOMED CT as a knowledge graph consisting of triples (head entity, relation, tail entity) that structurally describe the medical knowledge [33]. Through modeling processes, numerical representations capturing the medical knowledge of medical concepts can be learned from the knowledge graph. The RotatE model regards a relation as a rotation in the complex plane, enabling the modeling of symmetry/antisymmetry, inversion, and composition patterns in knowledge graph relations.

2.3.3 SNOMED CT Embedding-Based Distance Metric

Figure 2.1 illustrates the workflow of defining a SNOMED CT embedding-based distance metric and highlights an example of a pairwise comparison of two ICD-10 codes in our study. Our method relies on the SNOMED CT, an ontology that provides a systematic collection of biomedical concepts and relations under a common ontological framework [21]. Given its polyhierarchical structure, the SNOMED CT can also be treated as a knowledge graph [30].

We first mapped three-digit ICD-10 codes with SNOMED CT concepts using Athena [41], an automated system for creating and maintaining the OMOP Standard Vocabularies [42]. Multiple SNOMED CT concepts can be mapped to a single ICD-10 code and vice versa. A total of 1619 three-digit level ICD-10 codes were mapped to one to five SNOMED CT concepts. Next, we retrieved pre-trained SNOMED CT embeddings provided by Chang et al. [30] to represent each SNOMED CT concept. We chose the embedding representations from their RotatE model based on its accurate and stable performance over multiple link prediction tasks.

We represented each ICD-10 code as an $N \times 512$ matrix, where N denotes the number of SNOMED CT concepts mapped to the ICD-10 code, and 512 represents the dimensionality of the embedding vectors. For ICD-10 codes mapped to multiple SNOMED CT concepts, this matrix representation captures the semantic diversity within each diagnostic category.

We then calculated the pairwise matrix similarity between two matrices using the RV coefficient with Mayer’s modification, ranging from 0 to 1 [43]. Mayer’s RV analysis is based on the RV coefficient, a matrix correlation that can be interpreted as a generalization of the squared correlation from two single variables to two sets of variables. It has shown less bias in independent, high-dimensional data sets when evaluating mean square error and the power of the independence test [43].

Finally, we transformed the similarity to a distance using the formula: distance = 1 - similarity. The embedding distance between two ICD codes can range from 0 to 1, with 1 being the most distinct and 0 indicating identical or highly similar diseases.

2.3.4 ICD Tree-Based Distance Metric

We built a second distance matrix based on the tree-like hierarchical structure of ICD with the assumption that codes within a branch, i.e., a clinical chapter, should be highly correlated. Thus, we defined the distance between two codes within a clinical chapter by taking

the differences between their numeric parts (second and third digits), numerical representation of hierarchical structure ranging from 00 to 99) and multiplying by 0.01. For any two codes from different chapters, we assigned their distance as 1 since no formal relationship exists between the various chapters [34]. For example, the tree-based distance between C21 and C44 is equal to $(44 - 21) \times 0.01 = 0.23$, and the distance between A30 and C50 is equal to 1. The ICD tree-based distance is used as a naïve and simple baseline model to compare to our embedding distance. The ICD tree-based distance between two ICD codes ranges from 0 to 1, with 1 being the most different.

2.3.5 Comorbidity-Based Distance Metric

We built a distance metric based on disease comorbidity relationships using patient health records extracted from the University of California, Los Angeles (UCLA) Health System. A total of 1.1 million patients with more than one encounter were randomly selected from the de-identified EHR called UCLA Data Discovery Repository (DDR) (1.1 million patients and 23 million encounters total). Diagnoses for each patient in ICD-10 code format were ordered by time and truncated to a three-digit level, keeping only one record for repetitive diagnosis. A comorbidity relationship is defined as two diseases affecting the same individual relative to chance alone. With respect to pairwise disease distance, D_1 and D_2 , we used Jaccard distance as a measurement of dissimilarity between them [44], which was defined as:

$$J_{\text{distance}} = 1 - \frac{\# \text{patients with } D_1 \text{ and } D_2}{\# \text{patients with } D_1 \text{ or } D_2}$$

Thus, the comorbidity-based distance between two ICD codes also ranges from 0 to 1, with 1 being the most distinct.

2.3.6 Genetic Correlation-Based Distance Metric

We defined genetic correlation-based disease distance, which describes the genetic relationship between two traits and can be used to better understand the underlying biology of two traits. It is an informative metric to quantify the genetic similarity between complex traits [45]. Genetic correlation can be estimated in multiple ways, including traditional epidemiological approaches based on families [45], using single nucleotide polymorphism (SNP) data [46], and aggregating summary statistics from genome-wide association studies (GWAS).

Here, we built a genetic correlation-based distance matrix of the ICD-10 codes of common polygenic traits with the following two methods: 1) estimation of genetic correlations using linkage disequilibrium score regression (LDSC) [46] using GWAS summary statistics from the UK Biobank [47], and 2) mining genetic correlations reported in published literature based on family studies and case-control GWASs.

Method 1: LDSC Analysis

For method 1, we downloaded summary statistics for phenotypes reported in three-digit ICD-10 codes format from the UK Biobank website [47]. The LD Score regression analyses were performed using the software developed by Bulik-Sullivan et al. [46]. Some diseases in the UK Biobank data have small effective sample sizes, especially for rare diseases, which can result in low statistical power and thus, affect confidence in the stability of the LD Score regression results. To ensure the robustness of our results, we selected genetic correlations of pairs whose confidence of heritability was medium or high, as defined by Neale [48]. We only included significant pairwise genetic correlations with P-value < 0.01 after false discovery rate (FDR) adjustment [49].

Method 2: Literature Mining

We based method 2 on previously aggregated disease heritability and genetic correlations from 234 published studies [50]. We mapped diseases listed in their work to ICD-10 codes and extracted their estimated heritability and pairwise correlation information. We restricted our

analyses to SNP-based case/control and family/twin study-based study types only. Since correlation estimates vary by data source types and estimation method (i.e., study types - case/control or family/twin-study), we compared genetic correlations and corresponding standard deviations (SD) between two study types with Wilcoxon rank-sum test [51] and performed analyses on different study types separately.

Finally, we combined results from the two resources mentioned above (N pairs = 903) and defined the genetic distance as the genetic correlation multiplied by 1. Unlike other distance metrics, we allowed the genetic distance to be negative to retain the value of genetic correlations.

2.3.7 Validation of Embedding Distance Metric

Given that the hierarchical information of ICD was not transferred during the mapping from ICD to SNOMED CT, we used the ICD structural properties to validate our embedding distance metric. We first used the ICD-10 clinical chapters ($N = 22$, details see Supplementary Table 1). We assumed that the embedding distance could reflect the hierarchical structure of the ICD tree; that is, codes within a clinical chapter should be closer relative to codes across chapters. We tested our hypothesis by examining 1) distances between ICD and 10 codes within the same disease sub-chapter (first two characters, e.g., F30-F39); 2) average distances between disease sub-chapters of the same clinical chapter (e.g., F0-F9); and 3) average distances between different clinical chapters (i.e., Chapter I-XXII). Similar steps were also performed with the ICD tree-based distance for comparison. All results were visualized with heatmaps.

In addition, we clustered the embedding distances into 22 modules, the same number as the total number of ICD clinical chapters and compared the modules to the chapters. Multiple clustering methods, including K-medoids [52], hierarchical clustering with static cut trees [53] and dynamic cut trees [54], were used to ensure consistency and robustness of the clustering results. We used the Rand index [55] to measure the similarities between the

clustering results and the ICD-10 clinical chapters. We then tabulated the clustering results (with the highest Rand index) and the ICD clinical chapters to evaluate the alignment by chapter. Jaccard index [44] was used to evaluate the alignment performance, and results were visualized with a heatmap.

2.3.8 Comparisons of Embedding vs. Other Distance Metrics

We first examined the overall pairwise correlation among the different distance metrics. Since most of those distance metrics had moderate skew, we used the non-parametric Spearman correlation coefficient [56] to evaluate their monotonic relationships without assuming linearity [57].

Next, we investigated the relationships of the embedding-based metric with genetic and comorbidity metrics to examine common and unique characteristics in different disease categories. We first tested potential non-linear associations using a smooth function (equidistant knots = 30). A follow-up linear regression model was fit if the results showed no splines. Similar steps were also performed within clinical chapters as sensitivity analyses. We reported effect sizes from the final regression model along with their 95% confidence intervals (CI). The ICD tree-based distance is a straightforward metric that has been introduced as a common notion of distance between two diseases in previous research works [28]. Thus, we also compared our embedding-based metric to the ICD tree-based metric regarding their ability to capture the hierarchical structure of the ICD and the information from molecular and phenotypic levels.

We considered P-value < 0.05 for statistical significance if not specified. All analyses were performed in R statistical software (version 4.1.2) [58]. Code for all analyses in this manuscript are available online (https://github.com/TSChang-Lab/disease_distance_KGE).

Moreover, as a sensitivity analysis, we built another embedding distance metric that used

the TransE [59] model for training embeddings from SNOMED CT [30]. The pre-trained TransE model used a dimension of 1580×1580 compared to the pre-trained RotatE, which was 1619×1619 , as the earlier pre-trained embeddings from TransE had fewer SNOMED CT codes (269,486 TransE vs 273,860 RotatE). As a result, different embedding models are used. All statistical analyses mentioned above were also performed on the TransE embedding distance.

2.4 Results

2.4.1 Development of Distance Metrics for Comparisons of Diseases

Summary statistics and distributions for different distance metrics are shown in Figure 2.2. Except for the genetic correlation-based metric, which was limited due to the total number of well-powered genetic studies, all three other metrics had more than one million pairs of ICD-10 comparisons. Both ICD tree-based and comorbidity-based distance have a peak at one, representing codes from different clinical chapters (ICD tree-based) and no co-occurrence in the EHR (comorbidity-based), respectively. The distribution of embedding distance is moderately negatively-skewed (skewness coefficient = -2.46), with the mean (SD) at 0.82 (0.1). The distribution of genetic distance is approximately normal (skewness coefficient = -0.98), ranging from -1.41 to 0.42 and a mean (SD) of -0.27 (0.24). Further comparisons of genetic correlation-based distances by study type (N pairs = 61) suggest smaller reported genetic distances and smaller SDs in family/twin studies compared to SNP-based case/control studies (Supplementary Table 2). In order to minimize the effect of how genetic correlation was estimated, further analyses involving genetic distances were stratified by study type. All pairwise distance metrics are available online (https://github.com/TSChang-Lab/disease_distance_KGE).

2.4.2 Validation of Embedding Distance Metric

Our embedding distance retains the hierarchical structure of the ICD-10 classification system. Figure 2.3 shows an example heatmap of distance comparisons between ICD-10 codes in clinical sub-chapters (first two characters) and chapters (the first character). The average distances within clinical chapters are significantly lower (t-test p-value < 0.001) relative to distances across clinical chapters. Moreover, codes in chapters with non-specific diagnostic codes, such as Chapter XX (External causes of morbidity), XXI (Factors influencing health status and contact with health services), and XXII (Codes for special purposes), are further away from other precisely-defined chapters on average. A similar trend is also found in the comparisons within and across disease sub-chapters (t-test p-value < 0.001).

One limitation of the embedding metric, which occurs rarely, is that ICD codes that map to the same SNOMED CT have zero distance. For example, the distances between F34 (Persistent mood [affective] disorders), F38 (Other mood [affective] disorders), and F39 (Unspecified mood [affective] disorder) are all equal to zero since all three codes are mapped to the same SNOMED CT (46206005 - Mood disorder).

We clustered the ICD-10 codes into 22 modules based on their embedding distances in order to evaluate if the embedding distance could reflect the hierarchical structure of the ICD tree. Interestingly, the clustering results of the ICD-10 codes align well with ICD clinical chapters. The mean overall similarity (Rand index) between clinical chapters and modules predicted by the three clustering methods is 0.924 (K-medoids = 0.906, hierarchical clustering with static cut tree = 0.928, and dynamic cut tree = 0.937). According to the tabulated results shown in Figure 2.4, in most cases, ICD-10 codes of the same clinical chapter are clustered into one module, especially for disease chapters that are well-defined or belong to a specific body system, such as Chapter II (Neoplasms) and Chapter X (Disease of the respiratory system).

Some chapters and modules, however, do not align in a one-to-one manner. There are

clinical chapters that are represented by multiple modules (e.g., Chapter XVII - Congenital malformations, deformations, and chromosomal abnormalities), and similarly, there are modules that gather codes from multiple clinical chapters (e.g., module red). These discordances happened more often in chapters that contained broad definitions or conditions such as symptoms, signs, and health services. These discordant alignments result when chapters are rearranged by alternative concepts. For example, instead of grouping congenital malformations, deformations, and chromosomal abnormalities into one single chapter (Chapter XVII - Congenital malformations, deformations, and chromosomal abnormalities in the ICD coding system), our method clustered the "congenital malformations of eye" (Q10-Q15) together with Chapter VII (Diseases of the eye and adnexa). Taking the predicted module red as another example, several mental disorders, such as mental disorders due to physiological conditions and psychoactive substance use, are clustered together with some injuries from external causes, such as poisoning by drugs (Supplementary Table 3).

In summary, the results above show that our embedding distances retain many of the phenotypic characteristics of diseases and align well with the ICD hierarchical structure. At the same time, it highlights the structure of diseases not captured by the ICD.

2.4.3 Comparisons Across Different Distance Metrics

To compare our proposed distance metric to alternatives, we first calculated the pairwise correlations between the four distance metrics (embedding-, ICD tree-, genetic correlation-, and comorbidity-based) with the Spearman correlation coefficient [56]. We note that the Spearman correlation coefficient cannot compute exact p-values with ties in rank, which is common in the ICD-tree and comorbidity distances (Figure 2.2). Thus, comparisons involving these two metrics may yield less interpretable conclusions. All distance metrics are mild-to-moderately correlated with each other. We found the strongest association ($\rho = 0.57$, $p\text{-value} < 0.001$) between genetic and comorbidity distances. This met our expectation since diseases with a shared genetic basis tend to share similar pathogenic mechanisms

and thus, will likely co-occur in the same person [38]. Our embedding distances are positively associated with genetic distances ($\rho = 0.30$, $p\text{-value} < 0.001$) and with comorbidity distances ($\rho = 0.21$, $p\text{-value} < 0.001$). In other words, diseases with a shared genetic basis or frequently co-occurring in the EHR also fall closer to each other in the embedded knowledge-graph space.

Figure 2.5A shows associations of the embedding distances with genetic distances. The fitted smooth line (red line in the left scatterplot) suggests that their relationship is approximately positive and linear. In the fitted linear regressions stratified by clinical chapters, we observed stronger associations within most chapters, except for a few chapters with non-significant results which may be due to small sample sizes (e.g., Chapter IV - Endocrine, nutritional, and metabolic diseases and VI - Diseases of the nervous system).

Most of the ICD pairs follow the trend of positive linear association of embedding and genetic distances, which is consistent with our hypothesis that phenotypically similar diseases may also have higher genetic correlations. For example, we observed high genetic correlations between ICD pairs within the Chapter V - Mental, Behavioral and Neurodevelopmental disorders. Prior studies have provided evidence for shared genetic risks across psychiatric disorders such as F90 (attention-deficit hyperactivity disorder) and F32 (major depressive disorder) [60]. Another, perhaps more surprising example is that F34 (Persistent mood [affective] disorders) and J45 (Asthma) were relatively close in both embedding and genetic distances. This can be explained by their pathophysiologic connection in inflammation and immune responses, which may involve genetic and early developmental factors [61].

On the other hand, there are pairs lying off the diagonal of the linear regression. A total of 22 pairs were found to be two standard deviations away from the regression line and were defined as "off-diagonal" pairs. For example, we observed a relatively small genetic distance but a large embedding distance (i.e., similar genetically but not phenotypically) between F40 (Phobic anxiety disorders) and G47 (Sleep disorders). This suggests that there could be pleiotropy, in which a genetic locus affects more than one trait, between these two

diseases. However, few studies have focused on the genetic predispositions of the two diseases to date. Altogether, these results highlight how our embedding distance, when coupled with other metrics, may shed light on pathology, aiding the understanding of shared/different mechanisms, and developing new treatments for similar diseases.

Figure 2.5B shows associations of the embedding distances with comorbidity distances. The fitted smooth line is roughly flat, indicating a small to null association between them. Similarly, in chapter-stratified linear regressions, effect sizes are also relatively small and consistent across chapters for associations between embedding and comorbidity distances.

Comparing embedding- and ICD tree-based metrics, the embedding metric has advantages compared to the ICD tree-based metric in multiple dimensions. First, the embedding distance metric is able to capture subclusters within a disease group. Taking malignant neoplasms as an example (Figure 2.6), compared to the ICD tree-based metric, the embedding metric can separate not only malignant neoplasms of ill-defined, secondary, and unspecified sites from primary, specified sites but also differentiate pathogenic sites from each other, such as male genital organs and urinary tract. Second, the embedding metric has stronger associations with genetic and comorbidity metrics than the ICD tree-based metric, as shown by a greater linear correlation in Figure 2.5A than Figure 2.7, indicating that our embedding metric can better incorporate the information from molecular and clinical levels. Together, these results suggest that the knowledge graph-based embedding metric can capture more genetic and clinical co-occurrence information, even though no genetic or clinical data is directly involved in building the metric.

Finally, we repeated our analyses using the TransE model for knowledge-graph embeddings by first building a pairwise ICD-10 code distance metric. There is a significant positive association between the TransE and RotatE metrics (Pearson’s correlation 0.44, $p < 0.001$). Validation results of the TransE distance metric are shown in Supplementary Figure 1. Generally, the TransE embedding distance metric showed similar trends in alignment with ICD hierarchical structures but performed less well compared our RotatE metric. For example,

in the heatmaps of TransE embedding distances, the differences between average distances within vs. across clinical chapters/groups are not as clear as the RotatE metric. Moreover, in the clustering task (i.e., using TransE embedding distances to cluster ICD-10 codes into 22 modules), the mean overall similarity (Rand index) between clinical chapters and modules predicted by the three clustering methods is 0.828 (compared to RotatE = 0.924).

2.5 Discussion

Disease distance can be determined from different perspectives, including ontological, phenotypic, and molecular perspectives. A useful representation of disease distance can be widely applied in establishing disease associations, aiding clinical diagnosis, and functional prediction of genes. Our current study introduced a new way of defining distances using embeddings from the SNOMED knowledge graphs that can be applied to all pairs of ICD-10 codes. We also built three other distance metrics based on the hierarchical structure of ICD, clinical comorbidity, and genetic correlations. We systematically evaluated the embedding distance compared to the other three. Results showed that the embedding-based distance captures known phenotypic, clinical, and molecular structures at a finer granularity than the other three metrics. A distance metric in the ICD-10 code format can benefit many research goals such as patient clustering and disease trajectories identification.

Many disease similarities revealed by our embedding distance metric were consistent with previous literature. For example, we found a potential association between viral hepatitis (B1X) and hemolytic anemias (D5X) as both were clustered into the same module (green). Gonzalez-Casas et al. pointed out that aplastic anemia may follow the development of hepatitis [62]. Gkamprela et al. [63] also reported iron deficiency anemia as a frequent complication of patients with chronic liver diseases such as hepatitis B or C. Conversely, these clustering results can also reflect some unknown correlations of diseases that induce new research directions, such as evaluating causal relationships and finding underlying mechanisms

of shared biological pathways.

It is widely recognized that multiple factors, such as genetic variation and environmental perturbations, are risk factors for complex diseases [64, 65]. Instead of directly utilizing hierarchical structure presented in ICD-10 codes, we took advantage of knowledge graphs built with SNOMED CT concepts, which contained finer expert-curated biomedical knowledge from multiple aspects. Information from molecular and phenotypic levels can still be reflected in the embeddings derived from the knowledge graph, though no direct genetic or clinical data was involved in the building of the metric. Results from the validation analyses provided evidence that our embedding distance can retain disease characteristics of body systems or conditions. Our embedding distances also showed significant associations with the genetic and comorbidity distances, suggesting that they reflect clinical associations more closely compared to ICD codes alone. Different phenotypes can result from a shared genetic basis, while diseases that appear homogeneous clinically can be associated with different pathogenic mechanisms [66]. Therefore, despite the usefulness of the embedding metric individually, we plan to develop distance metrics in future work that integrates multiple metrics built in this study. This integrated distance metric may further explain known disease relationships, or uncover novel disease relationships and their underlying mechanisms.

There are several key contributions of our embedding distance metric. We are the first to build a full landscape of disease distances in the ICD-10 code format with fine granularity. Our embedding metric overcomes many limitations of the other three metrics. Compared to the ICD-tree-based metric, it has a full distribution (rather than a majority of distances equaling 1) and the ability to capture subclusters within a disease group (Figure 2.3). Unlike the genetic-correlation-based metric, which is subject to limited genetic studies and selection bias [67], the embedding metric is able to provide all possible pairwise comparisons of ICD-10 codes. In addition, the embedding metric used the well-established public SNOMED database. It does not have issues with patient privacy, data quality, computing power, and sample selection when using the EHR to build the comorbidity distance metric. Moreover,

ideally, a distance metric is independent of EHR-based comorbidities, as one would like to use a distance to identify novel or unexpected comorbidities and analyze the course of illness. In sum, the embedding metric enables many large-scale downstream analyses such as disease trajectory prediction and patient clustering [28] by providing a comprehensive pairwise disease distance matrix independent of clinical overlap.

There are limitations to our current three-digit ICD-code embedding metric. For example, we are not able to differentiate F38 (Other mood [affective] disorders) and F39 (Unspecified mood [affective] disorder). A more fine-grained distance metric based on four or more digits may help address the limitation and could be extended for future work. Moreover, the embedding distances were exclusively based on SNOMED CT knowledge graphs and, consequently, largely depended on the accuracy of the information included in those knowledge graphs. With an increasing understanding of disease characteristics and associations, more comprehensive knowledge graphs can be built in the future and, in turn, improve the metric’s performance. Furthermore, a small fraction of disease associations were identical or similar in different distance metrics while there were also cases of unique patterns revealed from a single distance metric. The embedding distances alone may not be enough to represent disease relationships. Thus, to identify additional patterns between diseases, it may be beneficial to combine different distance metrics.

2.6 Conclusion

In summary, we proposed a novel method to define disease distance in the ICD-10 code format and computed three other distance metrics based on ICD-tree structure, comorbidity, and genetic correlation separately for comparisons. Results show our metric provides a representation of disease relationships that could be exploited to discover important disease patterns. All distance metrics developed in this study have been made publicly available for further examination and downstream analyses. Research in many different fields can benefit

from this quantitative metric. Together with EHRs and available machine learning tools, it can accelerate the effective characterization of diseases, discover the underlying etiology of disease associations, and enable personalized medical practice in the clinical community.

This work establishes the methodological foundation for the disease trajectory analysis presented in Chapter 3. The embedding-based distance metric provides the essential tool for quantifying relationships between diseases in temporal sequences, enabling the identification of common disease progression patterns and the development of predictive models for disease onset and progression.

2.7 Code and Data Availability

The code and data supporting the findings of this study are made available to promote reproducibility and further research in this area. All pairwise distance metrics computed in this study are available online at: https://github.com/TSChang-Lab/disease_distance_KGE. This repository includes the computed distance matrices, code for reproducing the analyses, and documentation for using the distance metrics in other research applications.

The pre-trained SNOMED CT embeddings used in this study were obtained from Chang et al. [30] and are available through their published resources. The ICD-10 to SNOMED CT mappings were generated using the Athena system [41], which provides standardized vocabulary mappings for the OMOP Common Data Model.

2.8 Acknowledgments

We thank the UCLA Health System for providing access to de-identified electronic health record data through the UCLA Data Discovery Repository. We acknowledge the UK Biobank for making GWAS summary statistics publicly available, and we thank all participants and researchers who contributed to the genetic studies used in this analysis.

We are grateful to Chang et al. for making their pre-trained SNOMED CT embeddings available, which formed the foundation of our embedding-based approach. We also thank the OHDSI Collaborative for maintaining the Athena vocabulary mapping system, which enabled the critical ICD-10 to SNOMED CT mappings used in this study.

This work was supported by institutional resources and computing infrastructure that enabled the large-scale distance calculations presented in this chapter.

2.9 Figures

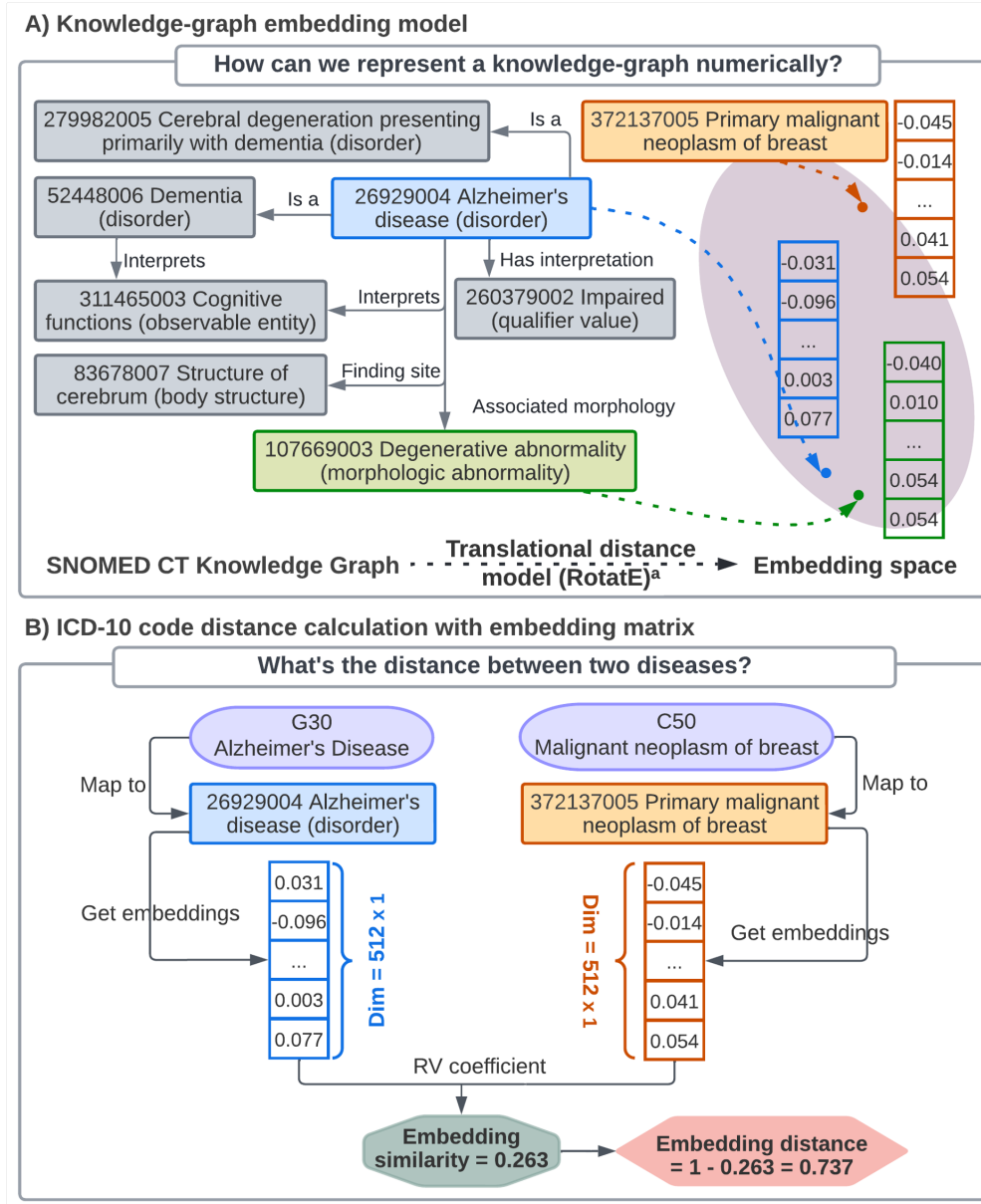


Figure 2.1: SNOMED CT embedding-based distance metric

A Knowledge-graph embedding model. Nodes and their relationships on the SNOMED CT knowledge graphs were input to RotatE model for relationship prediction tasks training (model details can be found in Chang et al.), from which we obtained an embedding vector (dimension = 1×512) as a simplified numerical representation for each node. Nodes with relationships are closer to each other when projecting their embedding vectors to the embedding space compared to nodes that are not related.

B Pairwise distance calculation between ICD and 10 codes. We first mapped ICD-10 codes with SNOMED CT concepts and obtained pre-trained SNOMED CT embeddings for each concept. Then we used the RV coefficient with Mayer's modification to calculate the matrix similarity, ranging from 0 to 1. Finally, the similarity was transformed to distance using distance = 1 - similarity.

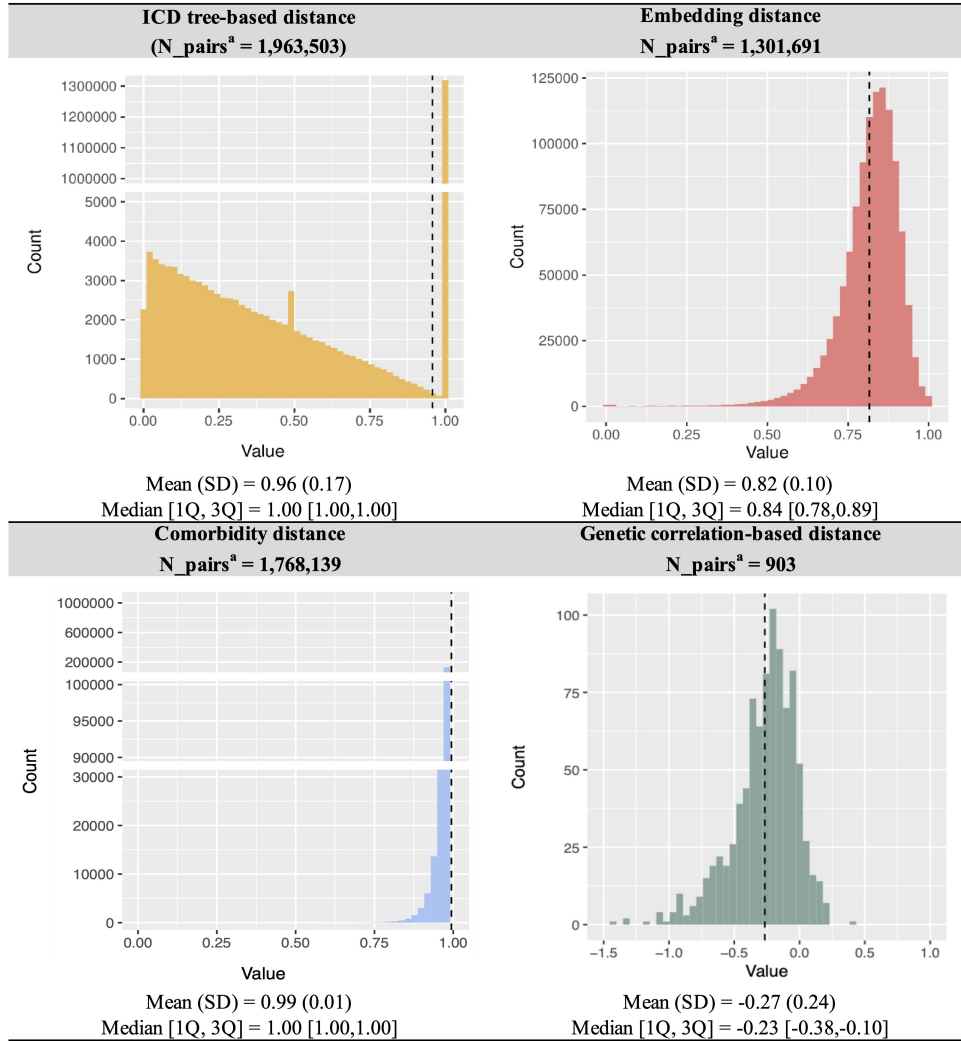


Figure 2.2: Distributions and summary statistics of distance metrics

Notes: [a] Self comparisons (e.g., A00 and A00) were excluded.

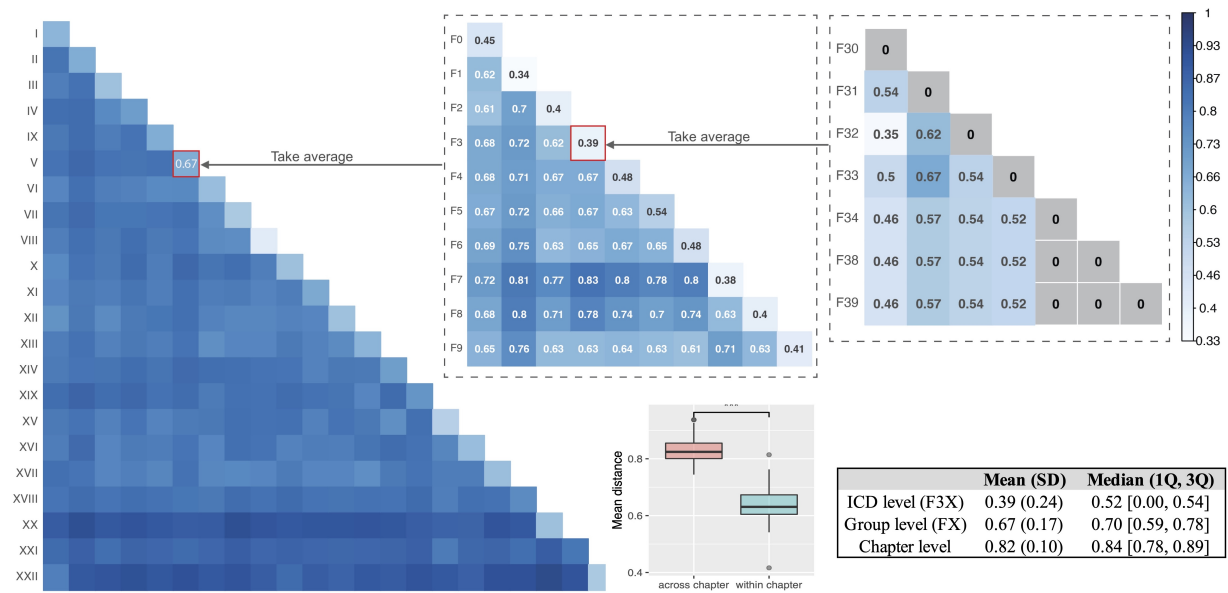


Figure 2.3: Validation of embedding distance metric - Heatmaps of embedding distance

Mean embedding distances between different clinical chapters (left); mean distances between disease groups in Chapter V - Mental and behavioral disorders (middle); and distances between ICD and 10 codes within F3 group - Mood [affective] disorders (right). A diagonal with lighter colors in the heatmap represents the average distances within clinical chapters are lower relative to distances across clinical chapters.

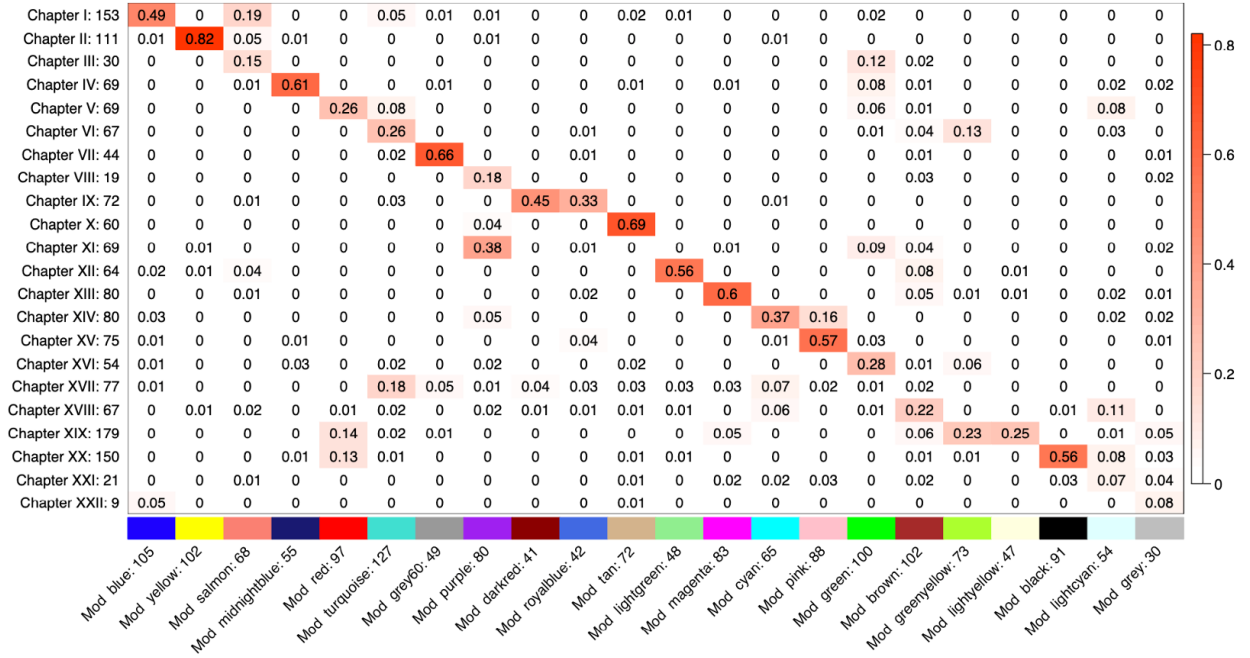


Figure 2.4: Validation of embedding distance metric - Correspondence of ICD clinical chapter and predicted modules (hierarchical clustering with dynamic cut tree model)

Y-axis represents ICD clinical chapters ($N = 22$) and number of codes in each chapter. X-axis represents modules predicted using the embedding distance and number of codes in each module. Numbers in each cell represents Jaccard index (range $[0,1]$, color-coded as larger values with darker colors) of the two disease classifications.

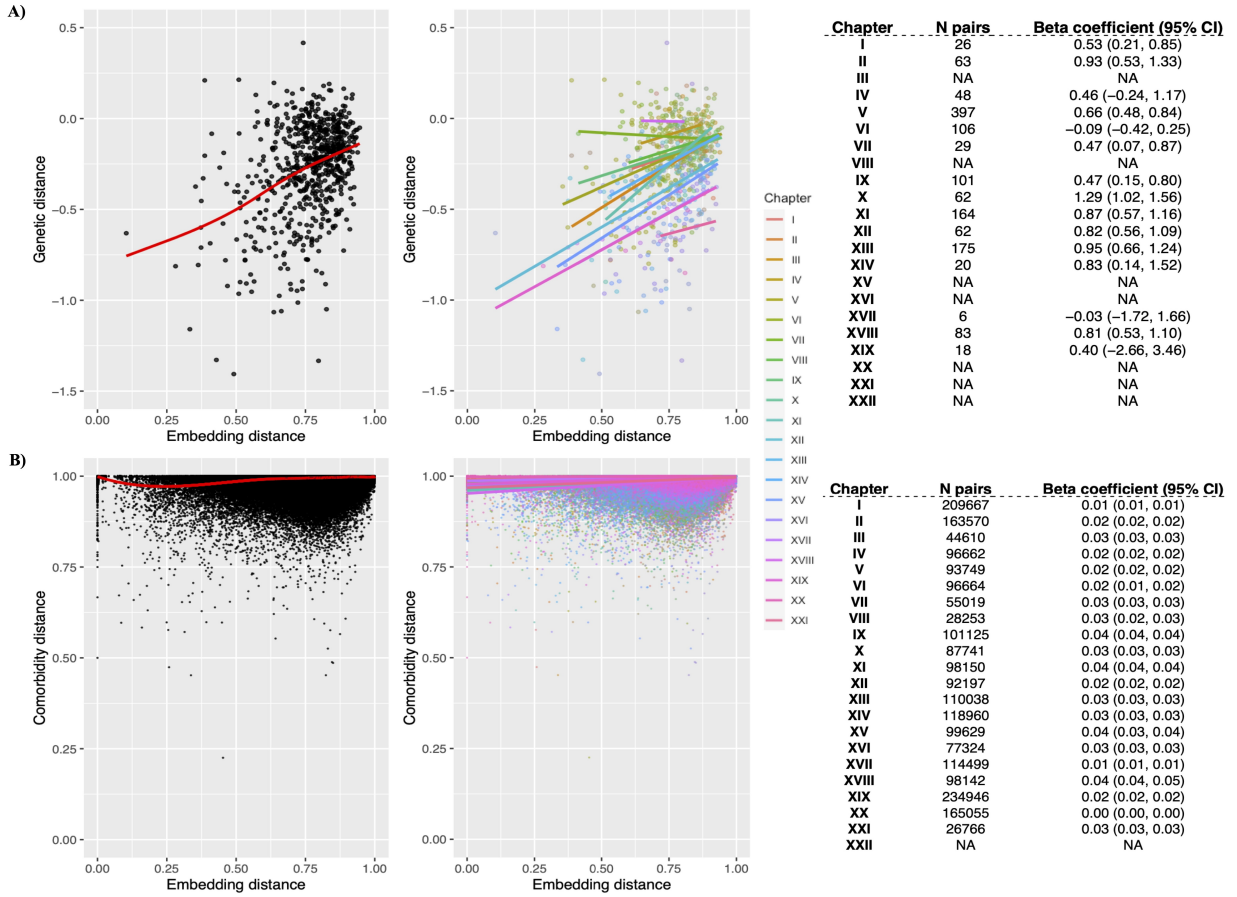


Figure 2.5: Associations of embedding distances vs. genetic (A) and comorbidity (B) distances.

Left scatterplots denote overall associations of distance metrics; a smooth line (red) was fit to check for non-linear relationships. Right scatterplots denote associations stratified by clinical chapters; lines in different colors were fitted from linear regression models of two distance metrics. Tables on the right show effect sizes of the associations by chapters. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

A) Example heatmap for C60-C80, embedding-based vs. ICD tree-based distance

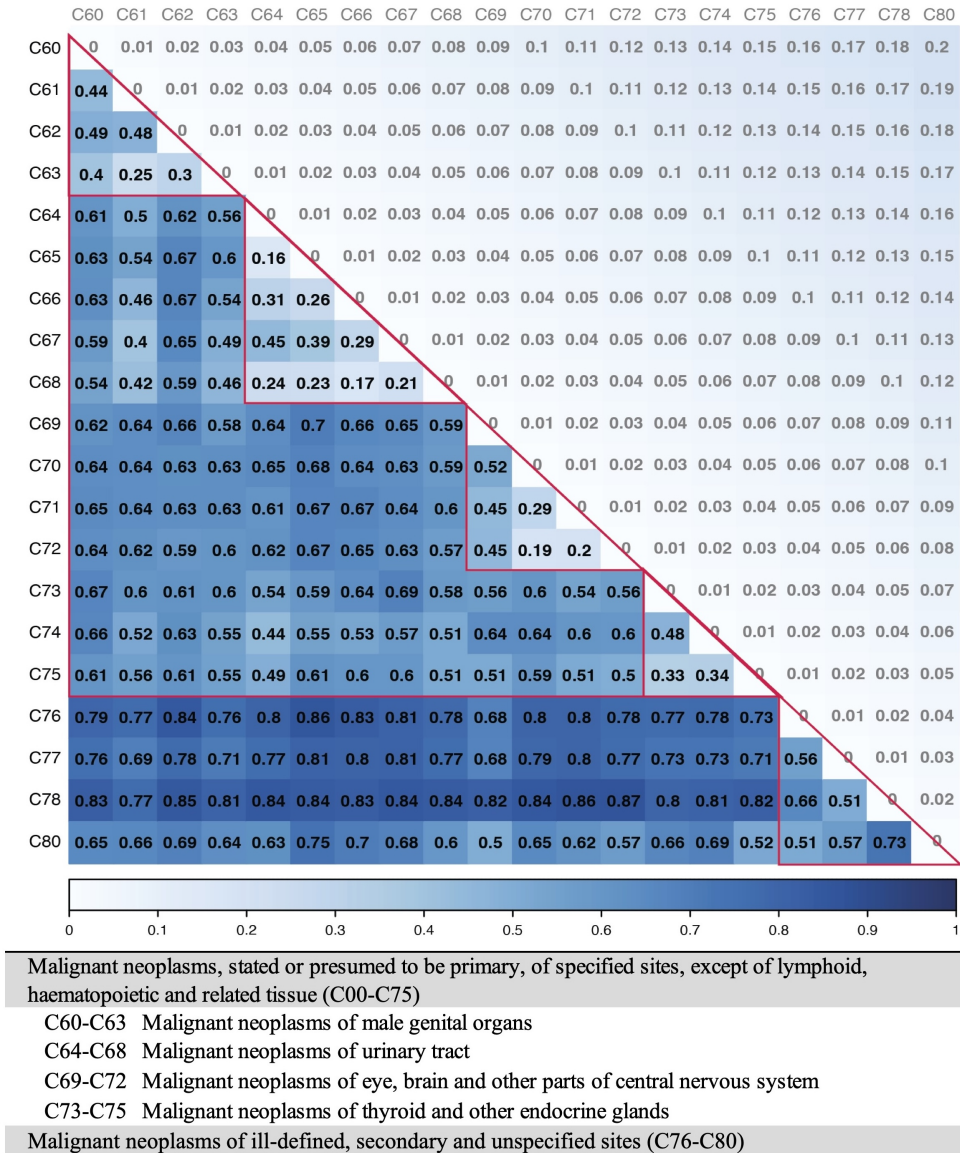


Figure 2.6: Comparison of embedding-based and ICD-tree distance (A).

Example heatmap for C60-C80. Colors and numbers in the heatmap represent pairwise distance between ICD and 10 codes calculated based on embeddings (bottom-left triangle) and ICD-tree (upper-right triangle).

B) Associations with genetic correlations (Spearman), embedding-based vs. ICD tree-based distance

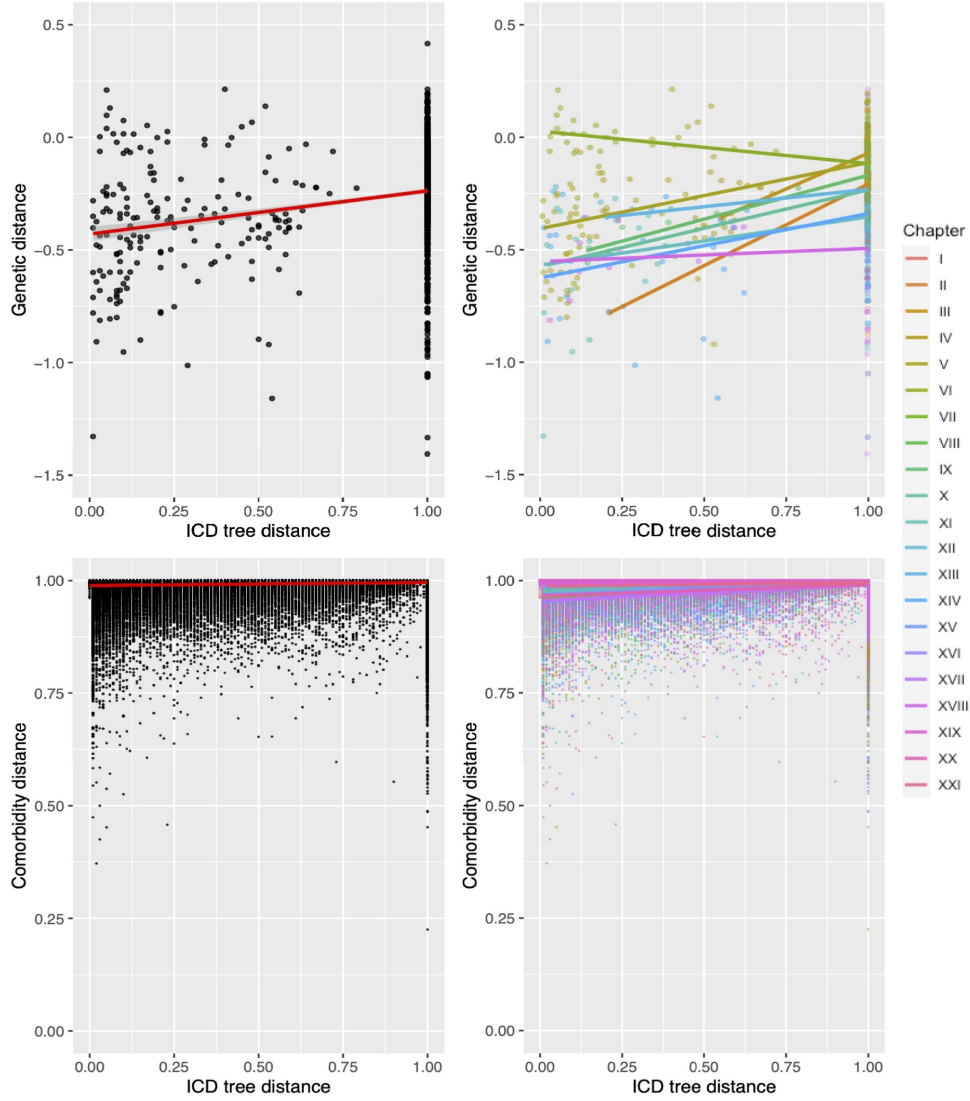
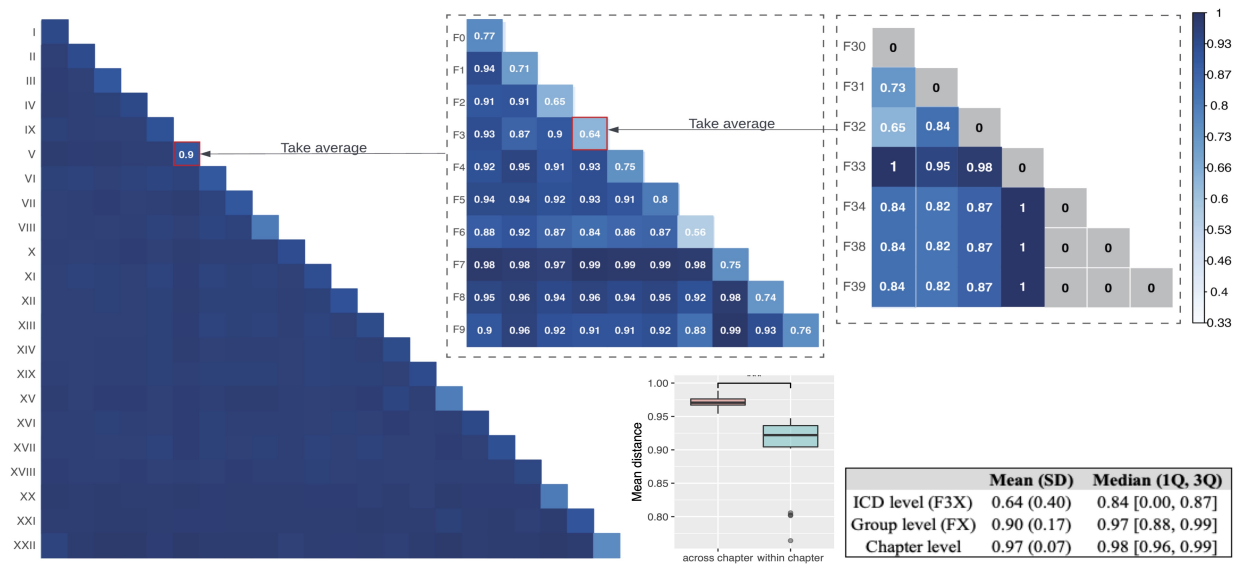


Figure 2.7: Comparison of embedding-based and ICD-tree distance (B).

Associations between ICD-tree distance and genetic (up) and comorbidity (down) distances. Left scatterplots denote overall associations of distance metrics; the red line was fit to check for linear relationships. Right scatterplots denote associations stratified by clinical chapters; lines in different colors were fitted from linear regression models of two distance metrics. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

2.10 Supplementary materials



Supplementary Figure 1. Validation of TransE embedding distance metric - Heatmaps of TransE embedding distances

Mean embedding distances between different clinical chapters (left); mean distances between disease groups in Chapter V - Mental and behavioral disorders (middle); and distances between ICD-10 codes within F3 group - Mood [affective] disorders (right). A diagonal with lighter colors in the heatmap represents the average distances within clinical chapters are lower relative to distances across clinical chapters.)

Supplementary Table 1. ICD-10-CM clinical chapters

Chapter	Code Range	Description
I	A00-B99	Certain infectious and parasitic diseases
II	C00-D49	Neoplasms
III	D50-D89	Diseases of the blood and blood-forming organs and certain disorders involving the immune mechanism
IV	E00-E89	Endocrine, nutritional and metabolic diseases
V	F01-F99	Mental, Behavioral and Neurodevelopmental disorders
VI	G00-G99	Diseases of the nervous system
VII	H00-H59	Diseases of the eye and adnexa
VIII	H60-H95	Diseases of the ear and mastoid process
IX	I00-I99	Diseases of the circulatory system
X	J00-J99	Diseases of the respiratory system
XI	K00-K95	Diseases of the digestive system
XII	L00-L99	Diseases of the skin and subcutaneous tissue
XIII	M00-M99	Diseases of the musculoskeletal system and connective tissue
XIV	N00-N99	Diseases of the genitourinary system
XV	O00-O9A	Pregnancy, childbirth and the puerperium
XVI	P00-P96	Certain conditions originating in the perinatal period
XVII	Q00-Q99	Congenital malformations, deformations and chromosomal abnormalities
XVIII	R00-R99	Symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified
XIX	S00-T88	Injury, poisoning and certain other consequences of external causes
XX	V00-Y99	External causes of morbidity
XXI	Z00-Z99	Factors influencing health status and contact with health services
XXII	U00-U85	Codes for special purposes

ICD-10-CM clinical chapters

Supplementary Table 2. Characteristics of genetic distances, stratified by study type (N pairs = 61)

Characteristic	Family/Twin-study ^a	Case/control-study ^a	p-value ^b
Genetic distance	-0.34 (-0.48, -0.22)	-0.13 (-0.26, -0.01)	<0.001*
Standard deviation	0.02 (0.02, 0.03)	0.06 (0.04, 0.09)	<0.001*

Notes:

[a] Summary statistics reported as Median (IQR)

[b] P-value calculated based on Wilcoxon rank sum test.

Characteristics of genetic distances, stratified by study type (N pairs = 61)

Supplementary Table 3. Details of ICD-10 disease groups in predicted module red

Group	Description
F0	Mental Disorders Due To Known Physiological Conditions
F1	Mental And Behavioral Disorders Due To Psychoactive Substance Use
F2	Schizophrenia, Schizotypal, Delusional, And Other Non-Mood Psychotic Disorders
F3	Mood [Affective] Disorders
F4	Anxiety, Dissociative, Stress-Related, Somatoform And Other Nonpsychotic Mental Disorders
F5	Behavioral Syndromes Associated With Physiological Disturbances And Physical Factors
F6	Disorders Of Adult Personality And Behavior
F9	Behavioral And Emotional Disorders With Onset Usually Occurring In Childhood And Adolescence
F9	Unspecified Mental Disorder
R4	Symptoms And Signs Involving Cognition, Perception, Emotional State And Behavior
T1	Injury Of Unspecified Body Region
T3	Poisoning By, Adverse Effects Of And Underdosing Of Drugs, Medicaments And Biological Substances
T4	Poisoning By, Adverse Effects Of And Underdosing Of Drugs, Medicaments And Biological Substances
T5	Poisoning By, Adverse Effects Of And Underdosing Of Drugs, Medicaments And Biological Substances
T5	Toxic Effects Of Substances Chiefly Nonmedicinal As To Source
T6	Toxic Effects Of Substances Chiefly Nonmedicinal As To Source
T6	Other And Unspecified Effects Of External Causes
T7	Other And Unspecified Effects Of External Causes
Y2	Event Of Undetermined Intent
Y3	Event Of Undetermined Intent
Y3	Legal Intervention, Operations Of War, Military Operations, And Terrorism

Details of ICD-10 disease groups in predicted module red

CHAPTER 3

Identifying Common Disease Trajectories of Alzheimer’s Disease with Electronic Health Records

3.1 Introduction

3.1.1 The Challenge of Understanding Disease Progression in Alzheimer’s Diseases

Alzheimer’s disease (AD) affects over 6.7 million individuals in the United States, with projections reaching 13 million by 2050 [4]. Unlike acute conditions, AD emerges through a gradual, multi-faceted process unfolding over years before clinical diagnosis [68]. While numerous individual risk factors have been identified—including cardiovascular conditions, depression, diabetes, and hearing loss [68, 69]—most research examines these factors in isolation rather than as interconnected elements of sequential disease progression.

Understanding not just what conditions precede AD, but when they occur and in what sequence, represents a paradigm shift from static risk factor identification to dynamic disease progression modeling. The pathway to AD is not merely the accumulation of independent risk factors, but rather the result of cascading pathophysiological processes that unfold over time [70].

3.1.2 Disease Trajectories: A Temporal Perspective on Health and Disease

Disease trajectories—sequences of health events occurring over time—offer a powerful framework for understanding complex disease progression [25]. Unlike cross-sectional analyses, trajectory analysis embraces the temporal nature of disease development, revealing patterns invisible to traditional epidemiological approaches.

This temporal perspective is particularly valuable for chronic conditions like AD, where the disease process may begin years before clinical symptoms appear [71]. Previous studies have demonstrated that certain sequences—such as progression from vascular risk factors to cerebrovascular disease to cognitive decline—represent common pathways informing both risk prediction and therapeutic intervention [72].

3.1.3 Electronic Health Records as a Window into Disease Progression

Electronic health records (EHRs) provide unprecedented opportunities to study disease trajectories at population scale [11]. EHRs capture longitudinal clinical data across diverse populations, documenting the real-world clinical course of disease as it unfolds in routine healthcare settings. The large scale of EHR databases provides statistical power to identify subtle patterns and rare progression pathways [73].

However, EHR analysis presents unique challenges including irregular timing of encounters, coding practice variations, and missing data [74]. The high dimensionality of EHR data requires sophisticated analytical approaches to extract clinically meaningful trajectories.

3.1.4 Building on Disease Distance Metrics: From Static Relationships to Dynamic Progressions

Chapter 2 established a framework for quantifying disease relationships using SNOMED CT embeddings [1]. While this distance metric excelled at identifying related diseases, it did not address the temporal dimension of disease progression over time.

The transition from static disease relationships to dynamic trajectories represents a natural evolution. The embedding-based distance metrics provide essential foundation for trajectory analysis by enabling precise quantification of relationships between sequential diagnoses, distinguishing meaningful clinical pathways from coincidental co-occurrences [24].

3.1.5 Current Limitations in Disease Trajectory Analysis

Current trajectory analysis methods suffer from several limitations: oversimplified disease progression representations, inability to handle temporal irregularities in clinical data, lack of robust causal inference methods, and limited generalizability across diverse populations [25, 74]. These limitations underscore the need for more sophisticated approaches that can capture the complex, multi-step progressions characterizing real-world disease development.

3.2 Background and Related Work

Healthcare temporal analysis has evolved from simple sequence mining to sophisticated trajectory analysis incorporating advanced techniques from time series analysis and machine learning [75, 76]. Dynamic Time Warping (DTW) has emerged as particularly powerful for healthcare trajectory analysis due to its ability to handle sequences of varying lengths and irregular timing [77, 78].

Network analysis provides systems-level perspectives on disease interactions and progression [23, 38]. These approaches have identified critical transition points and hub diseases in progression pathways, though traditional methods often treat relationships as static rather than dynamic [24, 79].

Machine learning methods, particularly clustering algorithms, have proven valuable for discovering complex patterns in clinical data [26, 80]. Unsupervised approaches like k-means and hierarchical clustering identify trajectory subtypes without prior knowledge, while deep learning approaches show promise for handling high-dimensional EHR data [81, 82].

Identifying causal relationships from observational data represents a major challenge in trajectory analysis [83]. Recent advances including directed acyclic graphs, structural equation modeling, and the Greedy Equivalence Search (GES) algorithm provide frameworks for discovering causal structures from observational data [84, 85].

Previous trajectory studies in dementia have identified important progression patterns, including cardiovascular-to-cognitive decline pathways and metabolic-neuropsychiatric sequences [68]. However, most studies have been limited by small sample sizes, homogeneous populations, or oversimplified analytical approaches, highlighting the need for more comprehensive trajectory analysis methods.

3.3 Methods

The general study design is illustrated in Figure 3.1. We first selected patients from the University of California Health Data Warehouse and pre-processed their longitudinal EHRs. We then retained diagnoses significantly associated with AD under a Fine–Gray survival model and constructed time-ordered trajectories. DTW was used to align and measure distances between these trajectories, followed by clustering to identify common patterns. Network analyses simplified each cluster into representative routes towards AD. We then assessed associations, evaluated risk in new cohorts, performed causal structural learning, and validated results in the All of Us Research Program.

3.3.1 Sample Selection in the UC Health Data Warehouse

Data Source and EHR Preprocessing

The University of California Health Data Warehouse is a secure, de-identified EHR repository aggregating data from six academic health centres: UC Davis, UC Irvine, UC Los Angeles, UC Riverside, UC San Diego, and UC San Francisco. Since 2012, it has captured inpatient and outpatient including primary-care encounters for 8.7 million unique patients.

For this study, we analysed records spanning from January 2012 to May 2024. Mortality ascertainment combines in-hospital deaths, family member reporting, and statewide death records from the California Department of Public Health. The data are de-identified to facilitate clinical research, with oversight from institutional review boards, privacy and compliance officers, and information security officers [86].

Patients were included if they had complete demographic information (age, gender, race and ethnicity), at least two encounters on different dates (one or more encounters per year), and were aged 65 to 90 at their last visit (the dataset censors patients older than 90) [87].

We extracted all International Classification of Diseases, 10th Revision (ICD-10) codes from multiple sources within the EHR, including primary and secondary diagnosis codes, problem list entries, medical history documentation, and billing codes. At the record level, three-digit ICD-10 codes were used after truncation to achieve a granularity appropriate for phenotyping [88]. We recorded the date of the first encounter bearing each three-digit ICD-10 code as the diagnosis date. ICD codes from Chapters XV (Pregnancy, childbirth, and the puerperium), XVI (Certain conditions originating in the perinatal period), XVII (Congenital malformations, deformations, and chromosomal abnormalities), and XX (External causes of morbidity and mortality) were excluded because they are less likely to be relevant to AD onset. Codes from Chapter XVIII (Symptoms, signs, and abnormal clinical and laboratory findings, not elsewhere classified) were excluded from the main trajectory-building process but retained for downstream analyses.

Case Definition

In this study, we define AD as dementia due to AD based on EHR diagnostic codes. This operational definition is distinct from both clinical diagnosis of AD, which relies on comprehensive neurological examination and cognitive testing, and biological AD, which requires biomarker confirmation of amyloid and tau pathology. Patients were classified as having AD if they had at least one encounter with the ICD-10 code G30. The AD diagnoses in our cohort were made across multiple departments, including neurology, geriatrics, and primary

care, following standard, real-world, clinical evaluation at health systems. These typically include cognitive assessments, neurological examination, and exclusion of alternative causes, though the specific diagnostic workup may vary by provider and department. Previous validation studies in EHR-based research have reported moderately high positive predictive values for AD codes (72–100%), suggesting that the misclassification of patients with AD is relatively limited [89]. In a UCLA Health internal validation, which was conducted within a subsample of the UC Health Data Warehouse cohort comprising 109 participants enrolled in the UCLA Alzheimer’s Disease Research Centre cohort, the positive predictive value was 0.77 for patients with ≥ 1 AD ICD code and 0.72 for those with ≥ 2 AD ICD codes, indicating that our diagnostic coding fidelity aligns with published benchmarks. The date of a patient’s first G30 diagnosis was defined as the AD index date, and only diagnoses on or before this date were kept for trajectory construction.

3.3.2 Pairwise ICD Codes Associations in a Sampled Population

We used a survival analysis framework to identify risk-associated diagnoses by examining pairwise ICD codes. This step aimed to reduce noise in patients’ disease trajectories by filtering out diagnoses that were not positively associated with subsequent AD.

Sample Set

We combined all patients with AD (cases) with non-AD (controls) into a discovery sample. Non-AD controls required ≥ 5 years of follow-up, ≥ 2 encounters, and ≥ 2 encounters per year to reduce the likelihood of undiagnosed AD. We performed 1:1 propensity score matching [90] on self-reported age, gender, race/ethnicity, and location sources (the six academic health centres within the UC Health system where patients received care). Although higher control-to-case ratios can enhance statistical power, 1:1 matching was chosen to keep computational complexity manageable. Subsequent tests of pairwise ICD code associations were then limited to patients with AD only, treating significant results as meaningful within the patients with AD subpopulation.

Fine–Gray Proportional Subdistribution Hazards Model

A Fine–Gray (FG) proportional subdistribution hazards model was employed to account for competing risks such as death [91]. We calculated follow-up time separately for each outcome, accommodating variations in event times and censoring. All three-digit ICD codes (excluding those from Chapters XV, XVI, XVII, XX, and certain symptomatic codes in Chapter XVIII) were evaluated as binary exposures indicating whether a condition occurred prior to the outcome. Positive risk patterns required a hazard ratio (HR) > 1 and an adjusted p-value ≤ 0.1 under false discovery rate (FDR) correction [92]. Only diagnoses meeting these criteria were retained in the patient with AD trajectories.

3.3.3 Patient Trajectories Preprocessing

Diagnoses recorded on the same date as AD were shifted one step before the AD date to preserve potentially relevant patterns. F02.8 codes overlapping with AD (often reflecting parallel mappings [93]) were removed if they occurred within 90 days before AD but were retained if earlier [89]. In the context of this study, a ‘step’ refers to a distinct diagnosis recorded in a patient’s EHR. A multi-step trajectory consists of a sequence of diagnoses occurring in a specific temporal order before the AD diagnosis. We split and concatenated all patient records that met the ≥ 3 -step criterion, building paths backward from AD using only significant diagnosis pairs from the FG model (Supplementary Figure 1). For patients with only two encounters, we applied a two-step expansion to reach three steps if the added diagnosis pairs were significant risk factors. The final trajectories ranged from three to 17 steps. To ensure alignment quality in the DTW analyses, we capped the maximum number of steps at nine, covering 99% of patients. Any patients with more than three distinct trajectories were reduced to three via clustering with DTW distances, ensuring minimal redundancy.

3.3.4 Common AD Trajectory Identification

Calculate Distances Between Trajectories

We used DTW [77] to align and calculate distances between patient trajectories. Unlike Euclidean distance, DTW accommodates sequences of varying lengths and timing distortions. The core component of DTW is the average accumulated distortion, which is the sum of distances minimised during alignment, defined as $d_{\Phi}(X, Y) = \frac{1}{M_{\Phi}} \sum_{k=1}^T d(\Phi_x(k), \Phi_y(k))m_{\Phi}(k)$ [77]. Local dissimilarity, $d(\Phi_x(k), \Phi_y(k))$, measures the distance between two ICD diagnoses in each time series, which was determined based on the Systemized Nomenclature of Medicine, Clinical Terms (SNOMED) embedding similarities established in our previous work [1]. We applied an asymmetric step pattern with slope weights ($m_{\Phi}(k)$) to each step (Supplementary Figure 2). The Open-Begin-End DTW method was used to allow for partial alignment with free endpoints, thereby skipping longer sequence segments if needed [94]. A normalisation constant (M_{Φ}) equal to the length of the query sequence ensured that accumulated distortions were comparable across paths of different lengths.

Trajectory Clustering and Network Analyses

We applied k-means [95], hierarchical [96], and partition around medoids [52] methods to cluster the resulting DTW distance matrix. We assessed clustering performance using the Calinski–Harabasz, Davies–Bouldin, and Silhouette indices [97]. We also calculated the adjusted rand index to compare the similarity of different cluster partitions [98].

For each cluster, we constructed a "raw" network of patient trajectories and retained edges present in at least 0.5% of patients. We then applied a backbone extraction based on modularity vitality [99]. This iterative process removed edges with minimal contribution, preserving primary routes to AD. The most frequently observed sequences in each cluster were designated as notable trajectories. Central nodes had the highest modularity contribution and often represented a key diagnosis or group of diagnoses. We also examined the decomposition of central nodes with more than three digits. For reporting, optional ICD

nodes are shown in brackets ("[]") to capture variations in disease progression that share a common core disease trajectory.

3.3.5 Evaluation of Common AD Trajectories

Patient Characteristics Across Clusters

After trajectory cleaning, each patient could have up to three unique trajectories, potentially placing them in multiple clusters. We first evaluated patient overlap across clusters via a Venn diagram [100]. We then compared demographic and EHR features across clusters after excluding patients assigned to multiple clusters, though all were retained for sensitivity checks. We examined Chapter XVIII symptoms, focusing on cognition, perception, emotional state, and behaviour at the three-digit ICD level. Symptoms from other systems were grouped at the two-digit level for clarity. We analysed these symptoms at any time before AD, three years before, and one year before and analysed the time between each cluster’s central node and AD.

We further characterised AD features across clusters, including age at onset, time from the first AD diagnosis to the most recent visit, and time to death (when applicable). Kaplan–Meier curves [101] evaluated the time from the first node of each cluster’s simplified network to the AD diagnosis, and from AD diagnosis to death. The log-rank test [102] determined statistical significance among survival curves. Finally, we examined alive vs. deceased patients within each cluster to identify diagnoses potentially linked to mortality risk.

Risk Trajectories Testing in AD and Controls

To validate the risk of multi-step disease trajectories for AD, we redefined a new control group comprised only of individuals never diagnosed with AD or AD-exclusionary phenotypes [89]. This redefinition was crucial to ensure the controls were true "controls" who were unlikely to have undiagnosed AD or related conditions that could confound our trajectory analyses. We used the FG model framework to compare those who followed a risk trajectory

to those who did not, employing similar 1:1 propensity score matching for baseline covariates. Two models were tested: 1) individuals who followed the multi-step trajectory vs. all others, and 2) individuals who followed the multi-step trajectory vs. those who had any diagnoses involved in that trajectory but did not fulfil the complete sequence.

3.3.6 Causal Structural Learning

To investigate potential causality, we used the Greedy Equivalence Search (GES) algorithm [85], a score-based procedure beginning with an empty, partially Directed Acyclic Graph (DAG). Edges were iteratively added or removed to optimise the Bayesian Information Criterion [103].

We applied GES to the same 1:1 matched AD-and-control sample used in FG modelling. A ten-fold cross-validation strategy was employed: in each iteration, we learned the DAG from the entire dataset excluding one-fold, repeating this for ten iterations. Only edges appearing in at least five of these iterations were deemed reliably causal. This approach offers computational efficiency while providing insights into possible causal structures underlying AD trajectories.

3.3.7 Sensitivity Analyses

Using a Different Case-to-Control Matching Ratio

To assess the stability of our findings, we conducted a sensitivity analysis using different case-to-control ratio (1:3 matching) in our propensity score matching procedure. The same matched sample was also used to test the causality with GES as described above. We then compared the resulting HRs and significance patterns across the different matching ratios to evaluate the robustness of our findings.

Trajectory Building with One-Month Buffer Period

To address potential reverse causality, we implemented a one-month buffer period be-

tween each diagnosis and the AD onset. This approach helps mitigate the risk that certain diagnoses might be part of the AD diagnostic workup rather than true risk factors. We repeated our trajectory construction and analysis with this buffer period and compared the results to our primary analysis. We only compared clusters that correspond by label (e.g., mental health cluster in the primary analysis vs. mental health cluster in the buffer-period analysis), ensuring a one-to-one alignment of patient-trajectory groups. To quantitatively assess differences between the original and buffer-period networks, we calculated Pearson correlation coefficients and cosine similarity metrics between standardised edge weights (divided by the total number of patients) for corresponding clusters. To maintain a consistent basis for correlation, we aligned the complete set of possible ICD-pair transitions observed in either network. Transitions present in one network but not the other were assigned a weight of zero, so the Pearson coefficient reflects both shared and differential edges.

AD Diagnosis Stability Analysis

To assess the stability of AD diagnoses in our cohort, we conducted an additional analysis examining potential changes in diagnosis after the initial AD identification. We reviewed all diagnoses recorded after the first AD onset date to identify other types of dementia (including progressive supranuclear palsy, frontotemporal dementia, vascular dementia, Lewy body dementia, cerebral dementia, and substance-induced dementia).

To strengthen the findings specifically for AD, we performed a sensitivity analysis focusing on diagnostic robustness. We selected patients with at least two AD diagnoses in their EHR and repeated the clustering analysis using the same methodology as our primary analysis. We then compared the consistency of clustering results with the adjusted rand index, which measures the similarity between two cluster assignments independent of permutation [98].

3.3.8 Validation in the All of Us Research Program

We validated our findings using data from the All of Us Research Program, a diverse cohort of over 500,000 participants from across the United States [15]. We used the Data Release v8, which includes longitudinal clinical records for 393,596 individuals, with data current through October 1, 2023. All analyses were conducted within the All of Us Researcher Workbench under approved data-use protocols.

We applied the same inclusion criteria and analytical pipeline to identify patients with AD (cases) and 1:10 matched controls in this dataset. We then assessed the replicability of our risk factor associations, trajectory clusters, and causal relationships to determine the generalisability of our findings beyond our primary cohort.

First, we replicated the FG modelling of all directed associations identified in the simplified networks for each trajectory cluster. For associations between specific diagnoses and AD, we utilized the full cohort of patients with AD (cases) and matched controls ($n = 8512$). For other pairwise ICD code associations, we restricted testing to the AD population ($n = 818$) to maintain consistency with our primary analysis framework. In total, 130 pairwise ICD code associations were tested, including 12 pairs of diagnoses directly linked to AD. We applied the same FDR correction methodology to adjust p-values for multiple comparisons.

Second, to validate the UCHWD AD disease trajectory clusters, we developed and implemented a procedure for assigning new patients to our previously identified clusters. For each cluster, we used all paths to AD in the simplified network as reference trajectories. We then calculated cluster-specific distance thresholds by determining the mean distance (using DTW) between each patient with AD in the UC Health (discovery sample) and all reference trajectories within their assigned cluster. Cutoff values were established at $Q3 + 1.5 \cdot IQR$ of all patients' mean distances within each cluster. For validation, we pre-processed the trajectories of All of Us patients with AD by extracting the first occurrence of each unique 3-digit ICD code and ordering them chronologically. We then calculated the mean distance

between each patient with AD in All of US and each UCHWD cluster’s reference trajectories, assigning patients to the cluster with the minimum mean distance, provided this distance fell below the predetermined cutoff for that cluster.

Third, we tested the risk trajectories identified in each cluster as potential risk factors in the All of Us cohort. We used the FG model framework to compare AD risk between individuals who followed a specific multi-step trajectory and those who did not, employing 1:10 propensity score matching for baseline covariates. We evaluated the same two comparative models as described above.

All analyses were conducted in R (version 3.6) [104]. Key packages used include: ‘survival’ [105] and ‘cmprsk’ [106] for Fine–Gray subdistribution hazard models; ‘MatchIt’ [107] for propensity score matching; ‘dtw’ [108] for dynamic time warping; ‘pcalg’ [109] for causal structural learning; ‘igraph’ [110] and ‘network’ [111] for network visualization and analysis; ‘cluster’ [112] and ‘factoextra’ [113] for clustering; and ‘ggplot2’ [114] and ‘survminer’ [115] for visualization.

3.3.9 Ethics

This study was considered human subjects exempt. All patient data used in this research was de-identified to protect patient privacy. We confirm that all authors have read the Journal’s position on ethical publication issues and affirm that this work is consistent with those guidelines.

3.4 Results

3.4.1 Pairwise ICD Code Temporal Associations

919,590 patients met our initial inclusion criteria (complete demographics, at least two encounters on different dates, and were aged 65 to 90) in the University of California Health

Data Warehouse. 24,473 (2.7%) met the inclusion criteria for AD. We acknowledge there is likely an underdiagnosis of AD in this population, but we used strict inclusion criteria to ensure higher precision of cases.

Using the FG subdistribution hazard model, we identified disease codes positively associated with AD incidence while accounting for the competing risk of death [91]. We compared all patients with AD with 1:1 matched non-AD controls, matched on age at last visit, gender, race/ethnicity, and location sources. All standardised differences were below 0.1, indicating balance between the AD and non-AD groups on the matched variables. The matched controls were slightly older and had longer EHR records, as expected by design (Supplementary Table 1).

In this AD-and-control cohort ($n = 48,946$), the FG model identified 16 ICD codes positively associated with AD incidence after false discovery rate adjustment [92] (Table 3.2). Many of these risk factors were in the ICD chapter on mental, behavioural and neurodevelopmental disorders (Chapter V), including depressive episode (F32), unspecified mood disorder (F39), personality and behavioural disorders due to known physiological condition (F07), alcohol-related disorders (F10), and unspecified psychosis not due to a substance or known physiological condition (F29). Other significant risk factors included other disorders of the brain (G93), other degenerative diseases of the nervous system (G31), other cerebrovascular diseases (I67), unspecified dementia (F03), and vascular dementia (F01). The hazard ratios ranged from 1.30 (95% CI: 1.20–1.41) for other cerebrovascular diseases to 6.04 (95% CI: 5.67–6.43) for unspecified dementia.

3.4.2 Common AD Trajectory Identification

After applying the ≥ 3 -step criterion and trajectory preprocessing, 5762 patients contributed 6794 unique AD progression trajectories. The DTW distance matrix was computed for all trajectory pairs, incorporating SNOMED CT embedding similarities to measure local dissimilarity between diagnoses. Clustering performance evaluation indicated an optimal

solution of four clusters based on the Calinski–Harabasz, Davies–Bouldin, and Silhouette indices.

The four trajectory clusters were characterised as follows (Figures 3.3 and 3.4):

1. Mental Health/Depression Cluster ($n = 1,448$, 1266 unique trajectories): Centred on depressive episode (F32), with 89% of patients having this diagnosis. Common precursors included disorders of lipoprotein metabolism, anxiety disorders, essential hypertension, type 2 diabetes, dorsalgia, and other functional intestinal disorders. Notable trajectories included "disorders of lipoprotein metabolism $\rightarrow$ depressive episode $\rightarrow$ AD" and "anxiety disorders $\rightarrow$ depressive episode $\rightarrow$ AD."

2. Encephalopathy Cluster ($n = 3,223$, 764 unique trajectories): Anchored by other disorders of the brain (G93), most frequently unspecified encephalopathy (G93.40, 45%). Common precursors included essential hypertension, other cerebrovascular diseases, other urinary system diseases, and benign prostatic hyperplasia. Examples included "essential hypertension $\rightarrow$ other disorders of the brain $\rightarrow$ [other cerebrovascular diseases] $\rightarrow$ AD" and "lipoprotein metabolism $\rightarrow$ other disorders of the brain $\rightarrow$ AD."

3. MCI/Neurodegenerative Cluster ($n = 1,502$, 1658 unique trajectories): Focused on other degenerative diseases of the nervous system (G31), with 70% of patients having mild cognitive impairment (G31.84). Common precursors were transient cerebral ischaemic attacks, menopausal disorders, male erectile dysfunction, and eyelid/lacrimal system diseases. Notable sequences included "other degenerative diseases of the nervous system $\rightarrow$ vascular dementia $\rightarrow$ AD" and "transient cerebral ischaemic attacks $\rightarrow$ other degenerative diseases of the nervous system $\rightarrow$ AD."

4. Vascular Cluster ($n = 1,446$, 3106 unique trajectories): Centred on other cerebrovascular diseases (I67). Common preceding diagnoses were joint disorders, soft tissue disorders, dorsalgia, and postprocedural states. Notable trajectories included "other cerebrovascular diseases $\rightarrow$ vascular dementia $\rightarrow$ AD" and "essential hypertension $\rightarrow$ other anaemias $\rightarrow$

unspecified dementia → AD.” In this group, 44% had unspecified cerebrovascular disease (I67.9).

Our methodology successfully identified both common two-step trajectories and longer multi-step pathways. Examples of longer trajectories include ”essential hypertension → lipoprotein metabolism → depressive episode → AD” from the mental health cluster and ”lipoprotein metabolism → essential hypertension → other joint disorder → other cerebrovascular diseases → AD” from the vascular cluster.

3.4.3 Evaluation of Common AD Trajectories

Patient Characteristics Across Clusters

A total of 4078 patients fell into only one cluster, 1511 appeared in two clusters, and 173 in three clusters (Figure 3.5). The MCI/neurodegenerative cluster showed the most overlap with others.

Focusing on the 4078 single-cluster patients, we observed significant demographic and EHR differences among clusters (Table 3.6). Patients in the encephalopathy cluster tended to be younger, while those in the depression cluster were more often female or Hispanic. The vascular cluster included more Asian individuals and exhibited higher comorbidity burdens. We observed significant variations in EHR record length across the four trajectory clusters. The encephalopathy cluster showed the shortest records (mean 2.7 years), followed by the depression cluster (mean 5.7 years) and MCI cluster (mean 6.0 years), while the vascular cluster had the longest records (mean 8.4 years). These variations reflect distinct patterns of healthcare utilisation and disease progression rather than bias in our cluster method, which did not use EHR record length as a feature. Including multi-cluster patients in the analysis did not substantially alter these findings (Supplementary Table 2).

When comparing Chapter XVIII symptoms (cognition, perception, emotional state, and behaviour) prior to AD diagnosis, vascular cluster patients were more likely to report memory

loss or dizziness/giddiness, particularly within one to three years before AD onset. Across all organ systems, the vascular cluster had higher non-cognitive symptoms than other clusters (Supplementary Table 3A and 3B).

Moreover, the encephalopathy cluster exhibited the shortest interval from AD diagnosis to the last recorded visit. Survival analyses based on Kaplan–Meier curves revealed that the shortest time from a cluster’s first node to AD diagnosis occurred in the encephalopathy cluster, followed by depression, MCI, and vascular (overall log-rank $p < 0.001$). The shorter time from encephalopathy cluster diagnoses to AD also aligns with shorter record length in the encephalopathy cluster. Among deceased patients, encephalopathy cluster members had a shorter interval from AD diagnosis to death than those in the vascular cluster ($p = 0.036$), whereas other comparisons were non-significant (Figure 3.7). The mean duration from the central node to AD was also shortest in the encephalopathy cluster (0.33 years vs. 1.14–1.41 in other clusters; Figure 3.8).

We investigated risk factors for mortality by comparing diagnoses between alive and deceased patients within each cluster (Supplementary Table 4). Certain diagnoses (e.g., other hypothyroidism, essential hypertension, or other urinary system diseases) were more frequent among deceased patients in the encephalopathy cluster, whereas atrial fibrillation, pneumonia, or acute kidney failure were frequent in deceased vascular cluster patients.

Single-Cluster vs. Multi-Cluster Patients

Because the MCI/neurodegenerative cluster showed the highest overlap, we compared single-cluster MCI patients to those belonging to multiple clusters. While no significant demographic differences emerged (Supplementary Table 5), patients in multiple clusters had longer, denser EHR histories and a higher number of diagnoses. They were also younger at AD diagnosis and had a shorter time from AD onset to death (e.g., three cluster 0.9–1.7 years vs. MCI only 2.4 years) (Table 3.9). Female and Hispanic individuals often overlapped with the depression cluster (e.g., MCI + depression Hispanic 13.8% vs. MCI only Hispanic 7.3%), while Black patients were more prevalent in two- or three-cluster combinations (e.g.,

three cluster Black 10.7–12.3% vs. MCI only Black 6.1%).

3.4.4 Risk Trajectories Test in Controls

To confirm that the identified trajectories confer elevated AD risk, we examined top trajectories from each cluster in a newly sampled control population devoid of AD or AD-exclusionary phenotypes. All representative trajectories (Table 3.10) were positively associated with AD risk. In most cases, multi-step trajectories posed a higher risk than having any single risk factor within the trajectory, suggesting that sequential disease progression amplifies the likelihood of AD.

3.4.5 Causal Structural Learning

We applied the GES algorithm to infer causal directions within each cluster’s simplified network. Overall, 26.2% of edges consistently showed direct, causal relationships. While this means that approximately one-quarter of edges were deemed causal, the remaining links did not reach the consistency threshold, indicating insufficient evidence or bidirectional relationships in the data. When stratified by cluster, directed edges ranged from 24.1% in the vascular cluster to 42.9% in encephalopathy (Supplementary Figure 3). Notably, the encephalopathy cluster demonstrated a higher proportion of directed edges, suggesting that this particular pathway may contain more definitive causal progressions than other clusters.

Among the AD risk factors identified in the FG model, six ICD codes, including unspecified dementia (F03), vascular dementia (F01), depressive episode (F32), unspecified mood disorder (F39), other disorders of the brain (G93), and other cerebrovascular diseases (I67), appeared causally linked with AD. Notably, three of these codes (F32, G93, and I67) were central nodes within their respective cluster. Representative multi-step pathways included "essential hypertension → depressive episode → AD," "acute kidney failure → other disorders of the brain → AD," and "other cerebrovascular diseases → vascular dementia → AD,"

suggesting that certain vascular, psychiatric, and systemic factors may act sequentially to increase AD susceptibility.

3.4.6 Sensitivity Analyses

Using a Different Case-to-Control Matching Ratio

The sensitivity analysis using the 1:3 matched sample demonstrated consistency with our primary findings. All 16 significant associations identified in our primary 1:1 matched analysis remained significant with similar effect sizes. The correlation between effect sizes (HRs) for all ICD codes associated with AD ($n = 387$ association tests) was 0.992 ($p < 0.001$), indicating highly consistent results regardless of the control-to-case ratio employed.

The causal structural learning analysis showed similar stability, with only minor differences in the identified causal links. Specifically, the F39 (Unspecified mood disorder) $\rightarrow$ AD link identified in the 1:1 matched sample was no longer classified as causal in the 1:3 matched sample, while a new causal link of F02 (Dementia in other diseases) $\rightarrow$ AD emerged. All other causal relationships remained consistent across the different matching ratios.

Trajectory Building with One-Month Buffer Period

The implementation of a 1-month buffer period resulted in 3306 patients with multi-step trajectories, compared to 5762 patients in our original analysis. The Calinski–Harabasz, Davies–Bouldin, and Silhouette analysis indicated an optimal solution of three clusters rather than the four identified in our primary analysis, with the most notable difference being the absence of a distinct "other degenerative diseases" (G31) cluster.

Quantitative comparison between the buffer-period networks and corresponding components of our original networks revealed varying degrees of similarity. The mental health cluster demonstrated remarkably high consistency (Pearson correlation = 0.97, cosine similarity = 0.97), while the brain/encephalopathy networks showed moderate correlation (Pearson correlation = 0.59, cosine similarity = 0.60). The cerebrovascular disease cluster exhibited

lower consistency (Pearson correlation = 0.43, cosine similarity = 0.44), indicating substantial reorganisation when the buffer period was applied.

AD Diagnosis Stability Analysis

Our analysis of diagnosis changes after initial AD identification revealed that among all patients with AD with follow-up data available ($n = 24,473$), 12.2% ($n = 2979$) had another dementia subtype diagnosis during follow-up. Within patients with multi-step trajectories ($n = 5762$), the proportion was slightly higher at 15.2% ($n = 875$). This finding reflects the clinical reality of mixed dementia and the evolving understanding of a patient’s condition over time.

In our sensitivity analysis examining diagnostic robustness, we identified 1822 patients with at least two AD diagnoses and multi-step trajectories after cleaning (compared to 5762 patients with at least one AD diagnosis in the original analysis). With the optimum clustering number of four as in our primary analysis, the adjusted rand index comparing these results to our original clustering was 0.94, indicating high consistency and demonstrating the robustness of our findings even when using a more stringent AD case definition.

3.4.7 Validation in the All of Us

Our validation analyses in the All of Us cohort yielded several important findings that largely corroborated our primary results. Compared to patients with AD in our discovery cohort, the patients with AD from the All of Us cohort had a younger age, fewer females, different race/ethnicity distribution, longer EHR records, and fewer visits per year. The AD features for these two AD populations were also different, with the All of Us patients being younger at AD diagnosis and having a shorter time from first AD diagnosis to death (Supplementary Table 6).

In the AD-and-control cohort ($n = 8512$), the Fine–Gray model validated 9 out of 12 ICD codes as being positively associated with AD after FDR adjustment (Supplementary Table

7). The three ICD codes that failed to replicate (F07: Personality and behavioural disorders due to known physiological condition, F10: Alcohol-related disorders, and F29: Unspecified psychosis not due to a substance or known physiological condition) had lower sample sizes in the validation cohort, though their hazard ratios trended in the same direction as in our primary analysis. Within the AD-only sample ($n = 818$), we were able to validate 3 out of 118 pairwise associations between non-AD diagnoses. The limited replication rate for these associations is likely attributable to the substantially smaller sample size available for validation ($n = 818$ compared to $n = 24,473$ in our primary analysis).

The cluster assignment validation demonstrated the robustness of our trajectory clustering approach. After minimum preprocessing, 785 patients with AD from the All of Us cohort had at least two distinct diagnoses documented before their first AD diagnosis and were therefore eligible for cluster assignment. Notably, 89.9% (706/785) of these patients were successfully assigned to at least one of our previously identified trajectory clusters using the predetermined distance cutoffs. The distribution of patients across clusters was comparable to our primary cohort, with the vascular disease (21.8%) and mental health (21.3%) clusters receiving the highest proportion of assignments. This high assignment rate confirms that the disease progression patterns we identified represent common pathways to AD that generalise beyond our initial study population.

Our analysis of multi-step trajectories as risk factors revealed that most of the associations identified in our primary analysis remained significant in the All of Us cohort. Specifically, 7 out of 9 tested representative trajectories were positively associated with AD risk after controlling for demographic factors and EHR features. Importantly, in 6 of these 7 validated trajectories, the multi-step progression conferred a significantly higher risk than having any single diagnosis within the trajectory, supporting our hypothesis that sequential disease progression amplifies the likelihood of developing AD (Supplementary Table 8). For example, the trajectory of disorders of lipoprotein metabolism $\rightarrow$ depressive episode $\rightarrow$ AD showed a hazard ratio of 1.63 ($p < 0.001$) compared to having either disorders of lipoprotein

metabolism or depression without the specific temporal sequence.

3.5 Discussion

Our study makes several important contributions to the field. First, we developed a framework for identifying and analysing disease trajectories in EHR data, combining DTW, clustering, network analysis, and causal inference. Second, we identified four distinct trajectory clusters leading to AD, each with characteristic demographic features, comorbidity patterns, and progression rates. Third, we demonstrated that multi-step trajectories confer higher AD risk than individual diagnoses, highlighting the importance of sequential disease progression. These multi-step trajectories provide a more comprehensive view of disease progression patterns directly observed in patient records leading to AD than previous approaches focusing on isolated risk factors or disease pairs. Fourth, we validated our findings in an independent cohort, confirming the generalisability of our approach.

Recent studies have applied various trajectory analysis methods to understand disease progression, including cross-sectional clustering [116], temporally ordered healthcare code pairs [117], and transitive sequential pattern mining [118]. In the field of AD and related dementia, two recent studies conducted by Li et al. [119] and You et al. [120] used disease trajectory mapping to study disease trajectory with UK Biobank data. While Li et al. identified three clusters initiated by circulatory, metabolic, and respiratory diseases 5–15 years before dementia, and You et al. characterised temporal evolution of clinical domains, these approaches typically concatenate directional pairs to form "pseudo" trajectories. Our method offers several advantages: we use a more diverse cohort (with higher percentage of non-white participants), capture multi-step disease trajectories directly observed in patients' records, employ DTW to accommodate temporal variations in disease progression, apply causal inference to determine directional relationships between diagnoses, and validate findings in an independent cohort. Despite methodological differences, our findings

on mental health and vascular trajectories align with these UK Biobank studies, reinforcing the importance of these pathways while providing unique insights into their sequential progression patterns. Our methodological approach enhances early detection strategies by identifying not just what conditions precede AD, but their specific ordering, interactions, and causal relationships.

Our study identified four distinct trajectory clusters leading to AD: mental health, encephalopathy, MCI/neurodegenerative, and vascular. These clusters exhibited significant differences in demographic characteristics, comorbidity patterns, and progression rates. The mental health cluster, centred on depression, predominantly affected women and Hispanic individuals, consistent with previous findings linking depression to AD risk [121, 122]. The encephalopathy cluster showed the most rapid progression to AD and subsequent death, suggesting a more aggressive disease course [123]. The MCI/neurodegenerative cluster followed the expected progression from mild cognitive impairment to AD, while the vascular cluster had the longest EHR histories and highest comorbidity burden, reflecting the chronic nature of cerebrovascular disease.

The presence of shared diagnostic precursors across multiple clusters, particularly vascular risk factors like hypertension, suggests that common pathophysiological mechanisms may underlie seemingly distinct progression pathways. From a clinical perspective, this highlights the potential importance of managing these shared risk factors for preventing progression along multiple pathways to AD. Future research should investigate whether targeted interventions for these shared diagnoses might prevent progression to one or more AD disease trajectories.

Our causal inference analyses provide valuable insights, they are based on observational data and should be interpreted with caution. The encephalopathy cluster demonstrated a higher proportion of directed edges, suggesting more definitive causal progressions compared to other clusters. The identification of six ICD codes causally linked to AD, including three that were central nodes in their respective clusters, further supports the validity of our

trajectory-based approach.

The validation of our findings in the All of Us Research Program demonstrates the generalisability of our approach across diverse populations and healthcare settings. The high assignment rate (89.9%) of All of Us patients to our previously identified clusters suggests that these trajectories represent common pathways to AD that extend beyond our initial study population. Furthermore, the confirmation that multi-step trajectories confer higher risk than individual diagnoses underscores the importance of considering sequential disease progression in AD risk assessment.

Our study has several limitations. First, our dataset automatically censored patients older than 90 years, which may limit the generalisability of our findings to the oldest-old population, a group with high AD prevalence. In addition, our discovery cohort was drawn from six academic health systems in California, which may limit representativeness relative to population-based samples. Individual records within the University of California Health Data Warehouse may be incomplete, and care delivered outside the UC system is not captured. Second, the potential misclassification of patients with AD in EHR data remains a concern. While previous validation studies have reported positive predictive values ranging from 72 to 100% for AD codes [89], we recognise that misclassification rates could impact our findings. The trajectories we identified should therefore be interpreted as pathways leading to clinically diagnosed AD rather than necessarily to biological AD confirmed by biomarkers. We acknowledge that ICD codes like 'unspecified dementia' may sometimes represent temporary diagnostic placeholders used by clinicians until more definitive assessments are completed, reflecting the reality of clinical coding practices. Third, our decision to focus specifically on AD rather than broadening to dementia represents a methodological trade-off between specificity and sensitivity. By focusing on AD codes, we aimed to identify trajectories most strongly associated with this specific dementia subtype, potentially missing cases with mixed pathology or those coded under more general dementia categories. Fourth, the varying EHR record lengths across trajectory clusters could potentially influence our find-

ings, though our DTW approach was specifically designed to accommodate such temporal variations. Finally, while our causal inference analyses provide valuable insights, they are based on observational data and should be interpreted with caution.

Despite these limitations, our study provides important insights into the complex, multi-step progressions that lead to AD. The identification of distinct trajectory clusters with characteristic features and progression patterns offers new opportunities for risk stratification, early detection, and targeted interventions. The validation of our findings in an independent, diverse cohort demonstrates the robustness and generalisability of our approach. Future research should focus on developing and testing targeted interventions for high-risk trajectories, investigating the biological mechanisms underlying different progression pathways, and extending this methodology to other complex diseases.

3.6 Conclusions

In summary, we developed a comprehensive framework for identifying and analysing disease trajectories leading to AD using longitudinal EHR data. Our approach successfully identified four distinct trajectory clusters—mental health, encephalopathy, MCI/neurodegenerative, and vascular—each with characteristic demographic features, progression patterns, and clinical implications. The demonstration that multi-step trajectories confer higher AD risk than individual diagnoses highlights the importance of considering sequential disease progression in clinical practice and research.

The validation of our findings in the All of Us Research Program confirms the generalisability of our approach across diverse populations and healthcare settings. The high assignment rate of validation cohort patients to our previously identified clusters demonstrates that these trajectories represent common pathways to AD that extend beyond our initial study population.

Our work establishes a methodological foundation for trajectory-based approaches to

understanding complex disease progression. The integration of advanced temporal alignment methods, sophisticated clustering approaches, network analysis, and causal inference provides a comprehensive framework that can be applied to other chronic conditions beyond AD. The identification of specific progression pathways offers new opportunities for risk stratification, early detection, and targeted interventions that could ultimately improve patient outcomes and reduce the burden of AD.

This trajectory-based perspective represents a paradigm shift from traditional risk factor identification to dynamic disease progression modeling. By revealing the temporal ordering and causal relationships between conditions that lead to AD, our approach provides insights that may inform the development of more effective prevention strategies and clinical interventions. Future research should focus on translating these findings into clinical practice through the development of trajectory-based risk prediction tools and targeted intervention strategies for high-risk progression pathways.

3.7 Code and data availability

All code used for trajectory analysis, clustering, and network construction is available at https://github.com/TSChang-Lab/AD_trajectory_analysis. The repository includes:

1. Complete analytical pipeline for trajectory identification and clustering
2. Dynamic time warping implementation with SNOMED CT embedding integration
3. Network analysis and backbone extraction algorithms
4. Causal inference implementation using the GES algorithm
5. Validation procedures for external cohort replication

The UC Health Data Warehouse data used in this study are available through appropriate data use agreements and institutional approvals. All of Us Research Program data are available through the All of Us Researcher Workbench under approved data-use protocols.

3.8 Acknowledgments

We gratefully acknowledge patients who have participated in the UCLA ATLAS Community Health Initiative and the All of Us Research Program. The UCLA ATLAS Community Health Initiative in collaboration with UCLA ATLAS Precision Health Biobank is a program of Institute for Precision Health (IPH), which directs and supports the biobanking and genotyping of biospecimen samples from participating UCLA patients in collaboration with the David Geffen School of Medicine, UCLA Clinical and Translational Science Institute and UCLA Health. We thank the All of Us Research Program participants and the research teams at the participating institutions for their contributions to this work.

This study was supported by the National Institutes of Health, National Institute on Aging (grants K08AG065519-01A1 to T.S.C. and M.F.), the National Science Foundation, the Hillblom and Fineberg Foundations, and the California Department of Public Health. S.S. was supported by the National Science Foundation (NSF CAREER Award DBI-1943147). B.P. was supported by the National Institutes of Health (R01HG009120, R01CA244670). K.V. was supported by the National Institutes of Health (R01AG064877, R01NS124882).

3.9 Figures

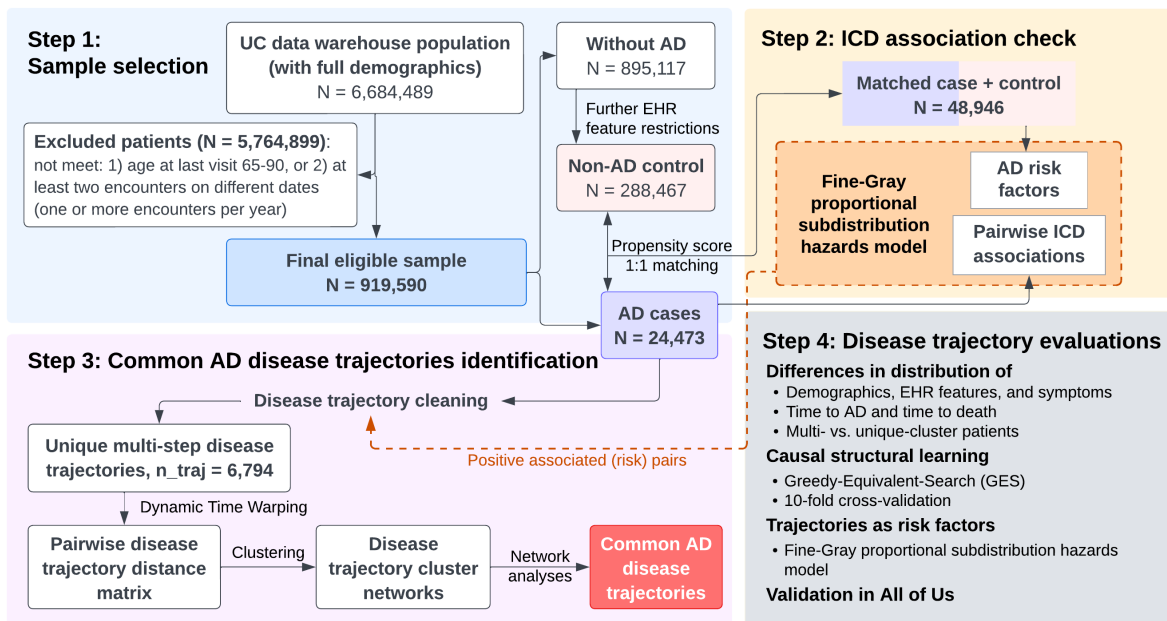


Figure 3.1: Study design

Table 1. Risk factors of Alzheimer's disease (AD) identified by the Fine-Gray proportional subdistribution hazards model, patients with AD (cases) and 1:1 matched controls (N = 48,946)

ICD	HR	Adjusted p-value*	Description	Chapter
F03	3.45	<0.001*	Unspecified dementia	Mental, Behavioural and Neurodevelopmental disorders
F09	2.39	<0.001*	Unspecified mental disorder due to known physiological condition	
F01	2.30	<0.001*	Vascular dementia, unspecified severity	
F07	2.29	<0.001*	Personality and behavioural disorders due to known physiological condition	
G31	2.27	<0.001*	Other degenerative diseases of nervous system, not elsewhere classified	Diseases of the nervous system
F29	1.91	<0.001*	Unspecified psychosis not due to a substance or known physiological condition	Mental, Behavioural and Neurodevelopmental disorders
F02	1.89	<0.001*	Dementia in other diseases classified elsewhere	
F22	1.71	<0.001*	Delusional disorders	
F06	1.53	<0.001*	Other mental disorders due to known physiological condition	
F31	1.50	<0.001*	Bipolar disorder	Diseases of the nervous system
G91	1.47	<0.001*	Hydrocephalus	
I67	1.30	<0.001*	Other cerebrovascular diseases	Diseases of the circulatory system
F39	1.18	0.003	Unspecified mood [affective] disorder	Mental, Behavioural and Neurodevelopmental disorders
F10	1.17	<0.001*	Alcohol related disorders	
F32	1.10	<0.001*	Depressive episode	Diseases of the nervous system
G93	1.10	<0.001*	Other disorders of brain	

Abbreviations: HR, Hazard Ratio.

Notes:

[a] Adjusted p-value using the False Discovery Rate for multiple testing.

* Statistical significance at p-value ≤ 0.1 level.

Figure 3.2: Risk factors of Alzheimer's disease (AD) identified by the Fine-Gray proportional subdistribution hazards model, patients with AD (cases) and 1:1 matched controls (N = 48,946)

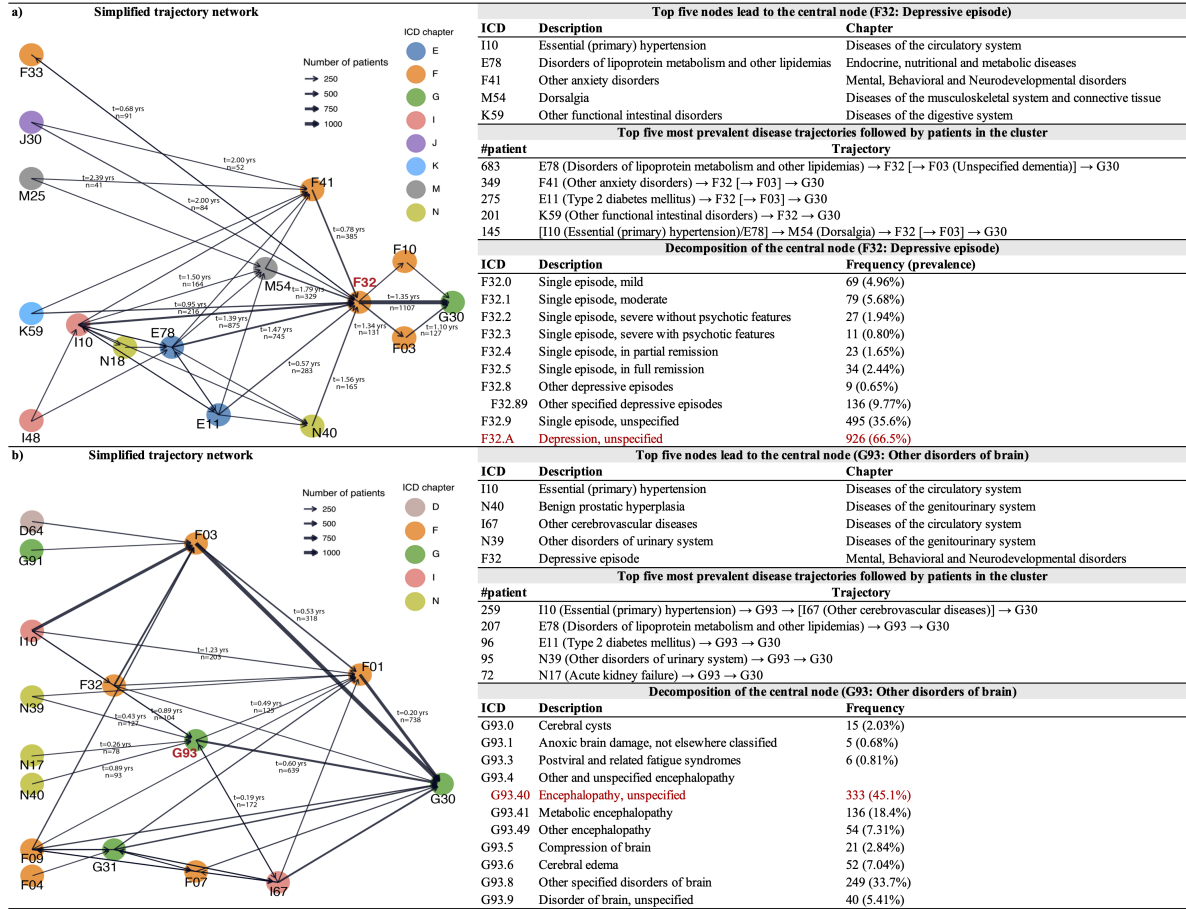


Figure 3.3: Summary of clusters of the Alzheimer's disease trajectory

a) Depression cluster (N patient = 1,448, N trajectory = 1266). b) Encephalopathy disease cluster (N patient = 3,223, N trajectory = 764). In the networks, “n” represents the number of patients followed the certain diagnosis switch and “t” represents the average time (in years) from the initial diagnosis to the ending diagnosis. ICD-10 codes enclosed in brackets (“[]”) indicate that the corresponding node is optional and may or may not be included in the complete trajectory.

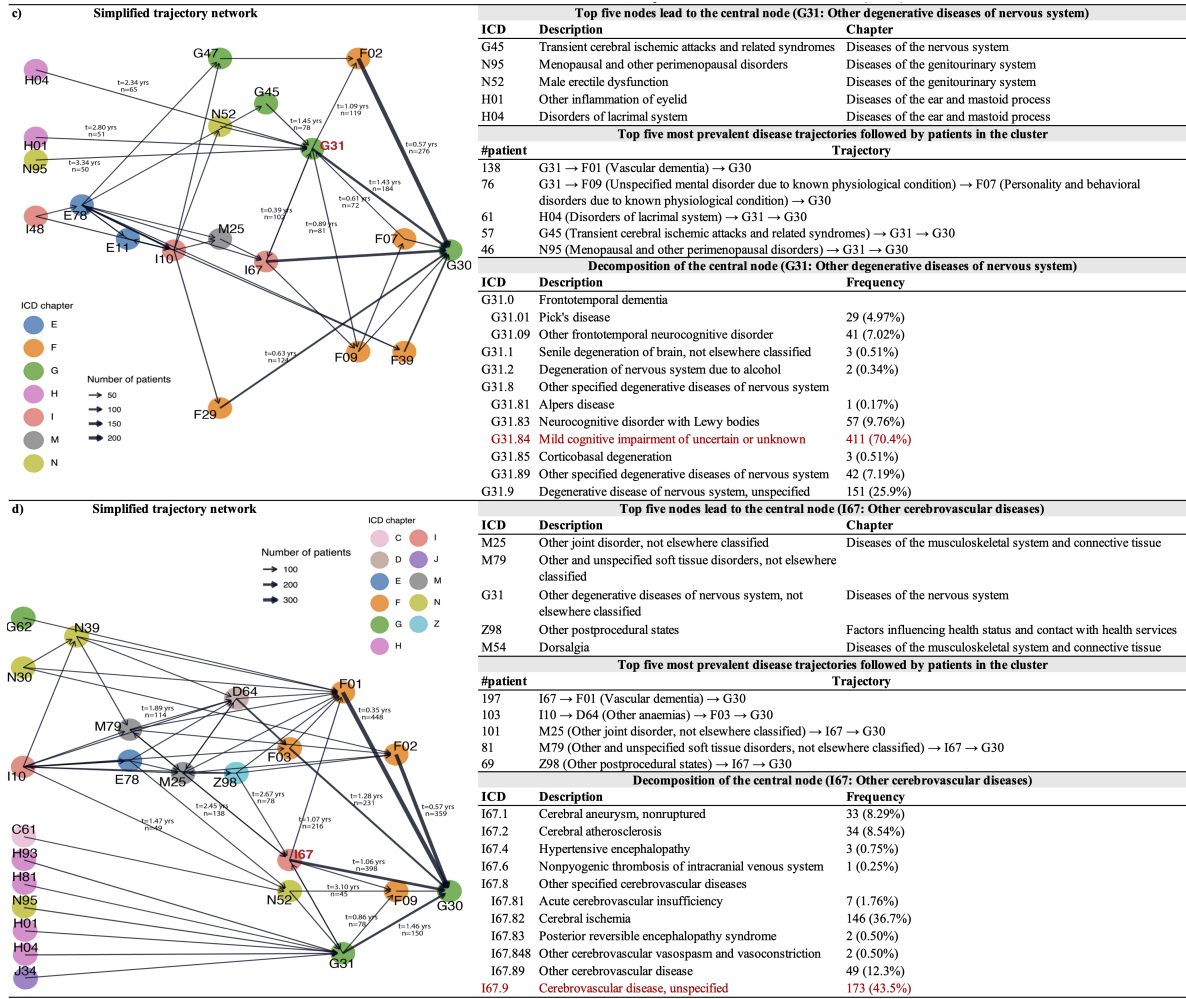


Figure 3.4: Summary of clusters of the Alzheimer's disease trajectory

c) Neurodegeneration cluster (N patient = 1,502, N trajectory = 1658). d) Vascular disease cluster (N patient = 1,446, N trajectory = 3106). In the networks, "n" represents the number of patients followed the certain diagnosis switch and "t" represents the average time (in years) from the initial diagnosis to the ending diagnosis. ICD-10 codes enclosed in brackets ("[]") indicate that the corresponding node is optional and may or may not be included in the complete trajectory.

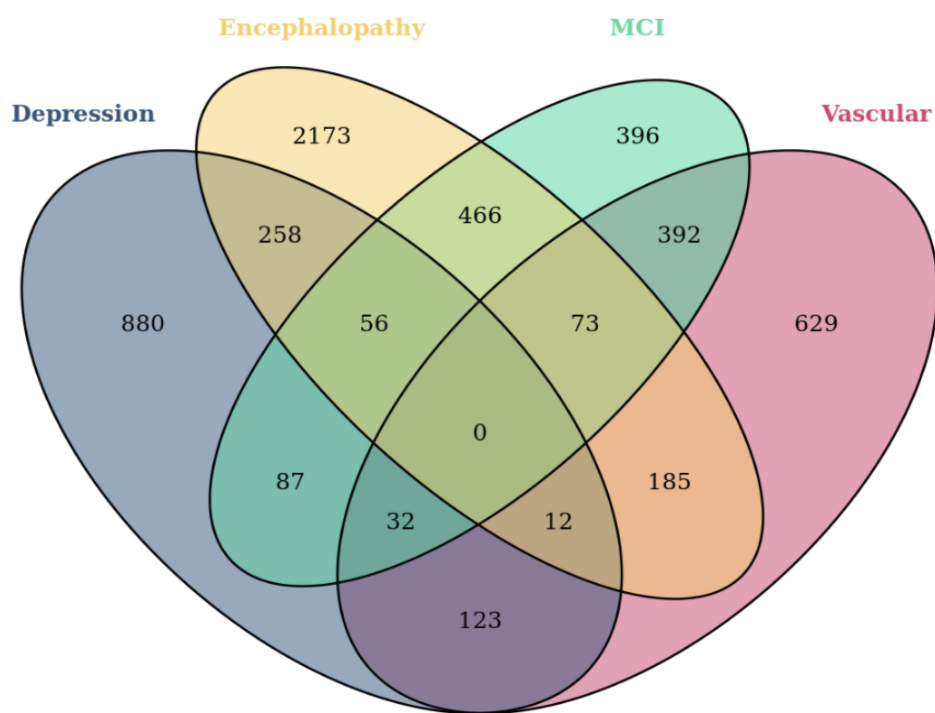


Figure 3.5: Venn diagram of Alzheimer's disease patients in trajectory clusters

Table 2. Alzheimer's disease (AD) patient characteristics by trajectory clusters, patients with unique trajectory cluster only (N = 4,078)^a

	Depression	Encephalopathy	MCI	Vascular	Adjusted p-value
Demographics					
N with unique cluster (% total)	880 (60.8%)	2173 (67.4%)	396 (26.4%)	629 (43.5%)	<0.001*
Age at last visit (in years)	82 (7)	81.3 (7.1)	83 (6.8)	83.2 (7.5)	<0.001*
Gender (female)	632 (71.8%)	1,305 (60.1%)	241 (60.9%)	369 (58.7%)	<0.001*
Race-ethnicity					<0.001*
American Indian	<10	<10	<10	<10	
Asian	84 (9.5%)	264 (12.1%)	51 (12.9%)	98 (15.6%)	
Black	43 (4.9%)	129 (5.9%)	24 (6.1%)	31 (4.9%)	
Hispanic	121 (13.8%)	219 (10.1%)	29 (7.3%)	63 (10.0%)	
Non-Hispanic White	575 (65.3%)	1,396 (64.2%)	274 (69.2%)	402 (63.9%)	
Others	54 (6.1%)	161 (7.4%)	16 (4.0%)	35 (5.6%)	
Death	425 (48.3%)	972 (44.7%)	162 (40.9%)	222 (35.3%)	<0.001*
EHR features					
EHR record length (in years)	5.7 (6.3)	2.7 (4.7)	6 (5.7)	8.4 (4.7)	<0.001*
Total number of encounters	42 (77)	12 (26)	36.5 (58.2)	72 (68)	<0.001*
EHR record density (per year)	10.4 (11.4)	7.4 (11.4)	8.3 (9.4)	9.8 (8.8)	<0.001*
Total number of diagnoses	47 (53.2)	21 (27)	39 (46)	65 (38)	<0.001*
Location sources					<0.001*
UC Site 1	140 (15.9%)	213 (9.8%)	36 (9.1%)	89 (14.1%)	
UC Site 2	105 (11.9%)	528 (24.3%)	60 (15.2%)	87 (13.8%)	
UC Site 3	297 (33.8%)	699 (32.2%)	151 (38.1%)	232 (36.9%)	
UC Site 4	<10	<10	<10	<10	
UC Site 5	223 (25.3%)	343 (15.8%)	74 (18.7%)	130 (20.7%)	
UC Site 6	114 (13.0%)	379 (17.5%)	75 (18.9%)	91 (14.5%)	
AD features					
Age at AD diagnosis (in years)	79.3 (5.7)	79.3 (7.1)	80.3 (6.3)	81 (6.9)	<0.001*
AD to last visit	1.8 (3.4)	1.2 (2.9)	1.6 (3.1)	1.5 (2.8)	<0.001*

Abbreviations: EHR, electronic health record; MCI, mild cognitive impairment.

Notes:

[a] Summary statistics are displayed as n (%) for categorical variables and mean (standard deviation) for continuous variables. Categorical variables are compared using chi-square tests. Continuous variables are compared using Kruskal-Wallis tests for overall comparisons across the four clusters. P-values represent comparisons across all trajectory clusters and have been adjusted for multiple comparisons using the False Discovery Rate method. Cells with counts less than 10 were masked as "<10" to comply with the requirements of the UC Health Data Warehouse.

* Statistical significance at p-value < 0.05 level.

Figure 3.6: Alzheimer's disease (AD) patient characteristics by trajectory clusters, patients with unique trajectory cluster only (N = 4,078)

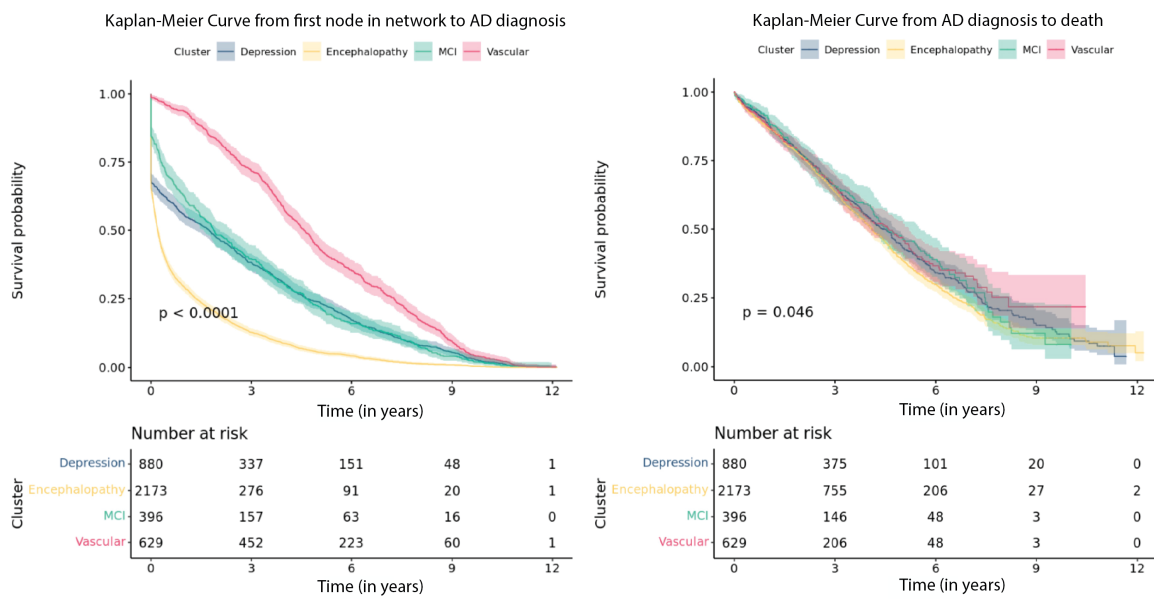


Figure 3.7: Kaplan-Meier survival curve in Alzheimer's disease (AD) patients by trajectory cluster

a) From first node in simplified trajectory (by cluster) to AD; b) From first AD diagnosis to death. Abbreviations: AD, Alzheimer's disease; MCI, mild cognitive impairment.

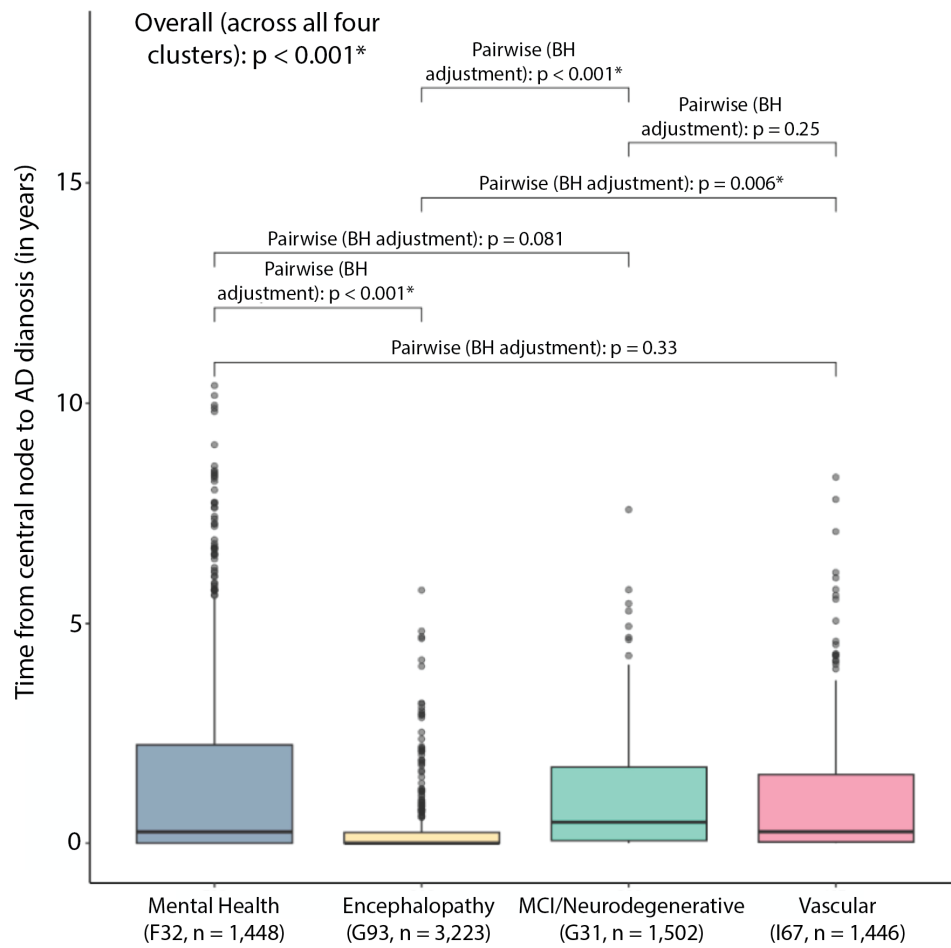


Figure 3.8: Distributions of time (in years) from central nodes to first Alzheimer's disease diagnosis by trajectory cluster

ICD-10 code descriptions: F32: Depressive episode; G93: Other disorders of brain; G31: Other degenerative diseases of nervous system; I67: Other cerebrovascular diseases.

Table 3. Characteristics of patients with Alzheimer's Disease (AD) that have clustered in the mild cognitive impairment cluster, single cluster vs. multi clusters^a

	Reference	Two clusters (MCI +)			Three clusters (MCI +)			p-value
	MCI only	Depression	Encephalopathy	Vascular	Depression + Encephalopathy	Depression + Vascular	Encephalopathy + Vascular	
Demographics								
N with unique cluster (% total)	396 (30.6%)	87 (6.7%)	466 (36.0%)	185 (14.3%)	56 (4.3%)	32 (2.5%)	73 (5.6%)	<0.001*
Age at last visit (in years)	83 (6.8)	81.1 (7.9)	81.7 (8.2)	82.7 (7.5)	79.2 (10)	82.4 (8.5)	84.4 (5.5)	0.02*
Gender (female)	241 (60.9%)	62 (71.3%)	254 (54.5%)	238 (60.7%)	37 (66.1%)	15 (46.9%)	40 (54.8%)	0.03*
Race-ethnicity								0.03*
American Indian	<10	<10	<10	<10	<10	<10	<10	
Asian	51 (12.9%)	<10	72 (15.5%)	57 (14.5%)	<10	<10	<10	
Black	24 (6.1%)	<10	28 (6.0%)	20 (5.1%)	<10	<10	<10	
Hispanic	29 (7.3%)	12 (13.8%)	58 (12.4%)	35 (8.9%)	10 (17.9%)	<10	<10	
Non-Hispanic White	274 (69.2%)	64 (73.6%)	276 (59.2%)	259 (66.1%)	33 (58.9%)	22 (68.8%)	50 (68.5%)	
Others	16 (4.0%)	<10	30 (6.4%)	20 (5.1%)	<10	<10	<10	
Death	162 (40.9%)	38 (43.7%)	208 (44.6%)	141 (36%)	31 (55.4%)	12 (37.5%)	29 (39.7%)	0.07*
EHR features								
EHR record length (in years)	6 (5.7)	4.9 (5.5)	3.8 (5.8)	7.1 (4.8)	3.9 (7.1)	8.4 (4)	6.9 (5.7)	<0.001*
Total number of encounters	36.5 (58.2)	38 (48)	20 (33.8)	56 (62.2)	30.5 (55.5)	60.5 (77.5)	41 (41)	<0.001*
EHR record density (per year)	8.3 (9.4)	8 (6.2)	8.1 (14.5)	9.5 (9.9)	11.7 (23.8)	9.3 (8.2)	7.7 (11.2)	<0.001*
Total number of diagnoses	39 (46)	50 (37)	30 (26.8)	56 (38)	37 (41.5)	55.5 (32.8)	47 (37)	<0.001*
Location sources								<0.001*
UC Site 1	36 (9.1%)	<10	38 (8.2%)	47 (12.0%)	<10	<10	<10	
UC Site 2	60 (15.2%)	17 (19.5%)	126 (27.0%)	69 (17.6%)	<10	<10	14 (19.2%)	
UC Site 3	151 (38.1%)	26 (29.9%)	136 (29.2%)	160 (40.8%)	18 (32.1%)	14 (43.8%)	26 (35.6%)	
UC Site 4	<10	<10	<10	<10	<10	<10	<10	
UC Site 5	74 (18.7%)	27 (31.0%)	99 (21.2%)	69 (17.6%)	24 (42.9%)	<10	15 (20.5%)	
UC Site 6	75 (18.9%)	11 (12.6%)	65 (13.9%)	47 (12.0%)	<10	<10	<10	
AD features								
Age at AD diagnosis (in years)	80.3 (6.3)	80.2 (8.5)	80.2 (7.3)	80.2 (7)	77.5 (9.1)	80.8 (10.3)	81.5 (5.6)	0.01*
AD to last visit	1.6 (3.1)	1.3 (2.9)	0.9 (2.6)	1.4 (2.6)	0.4 (3)	1.2 (2.7)	1.3 (2.8)	<0.001*
AD to death	2.4 (3.1)	2.3 (2.2)	2 (3.5)	2.1 (3)	0.9 (2.7)	1.3 (1.9)	1.7 (2.6)	0.03*

Abbreviations: EHR, electronic health record; MCI, mild cognitive impairment.

Notes:

[a] Summary statistics are displayed as n (%) for categorical variables and mean (standard deviation) for continuous variables. Categorical variables are compared using chi-square tests. Continuous variables are compared using Kruskal-Wallis tests for overall comparisons across clusters. P-values represent comparisons across trajectory clusters and have been adjusted for multiple comparisons using the False Discovery Rate method. Cells with counts less than 10 were masked as "<10" to comply with the requirements of the UC Health Data Warehouse.

* Statistical significance at p-value < 0.05 level.

Figure 3.9: Characteristics of patients with Alzheimer's Disease (AD) that have clustered in the mild cognitive impairment cluster, single cluster vs. multi clusters**Table 4.** Associations between risk trajectories and incidence of Alzheimer's disease

Cluster	Risk trajectory	Risk trajectory vs. no risk trajectory				Risk trajectory vs. single factor ^a			
		Case N	Control N	HR	P-value	Case N	Contr ol N	HR	p-value
Depression	Disorders of lipoprotein metabolism → F32	2910	40979	1.40	<0.001	2910	21701	1.61	<0.001
Central node - Depressive episode (F32)	Anxiety disorders → F32	1147	42742	1.57	<0.001	1147	7823	1.10	0.01
	Essential hypertension → F32	3295	40594	1.47	<0.001	3295	22962	1.67	<0.001
	Type 2 diabetes → F32	1217	42672	1.54	<0.001	1217	13285	1.39	<0.001
	Dorsalgia → F32	1910	41979	1.34	<0.001	1910	15105	1.50	<0.001
	Other functional intestinal disorders → F32	938	42951	1.49	<0.001	938	10687	1.24	<0.001
	Disorders of lipoprotein metabolism → G93	1762	42127	1.66	<0.001	1762	23526	1.87	<0.001
Encephalopathy	Disorders of lipoprotein metabolism → G93	1762	42127	1.66	<0.001	1762	23526	1.87	<0.001
Central node - Other disorders of the brain (G93)	Essential hypertension → G93	1999	41890	1.66	<0.001	1999	24837	1.82	<0.001
	Type 2 diabetes → G93	817	43072	1.74	<0.001	817	11212	1.71	<0.001
	Diseases of the urinary system → G93	914	42975	1.69	<0.001	914	10638	1.61	<0.001
	Benign prostatic hyperplasia → G93	521	43368	1.66	<0.001	521	8346	1.81	<0.001
	Acute kidney failure → G93	455	43434	1.67	<0.001	455	5726	1.33	<0.001
MCI	Transient cerebral ischemic attacks → G31	209	43680	2.30	<0.001	209	6628	0.87	0.002
Central node - Degenerative diseases of the nervous system (G31)	Menopausal disorders → G31	350	43539	2.39	<0.001	350	8237	1.60	<0.001
	Male erectile dysfunction → G31	252	43637	2.62	<0.001	252	7104	1.47	<0.001
	Other inflammation of eyelid → G31	254	43635	2.26	<0.001	254	7825	1.15	<0.001
	Disorders of the lacrimal system → G31	438	43451	2.20	<0.001	438	9984	1.60	<0.001
Vascular	Disorders of lipoprotein metabolism → I67	1438	42451	1.62	<0.001	1438	23682	1.82	<0.001
Central node - Other cerebrovascular diseases (I67)	Essential hypertension → I67	1556	42333	1.67	<0.001	1556	25207	1.80	<0.001
	Other joint disorders → I67	975	42914	1.58	<0.001	975	16542	1.96	<0.001
	Other degenerative diseases of the nervous system → I67	466	43423	2.82	<0.001	466	6233	1.08	0.04
	Conditions caused by postprocedural states → I67	568	43321	1.46	<0.001	568	12915	2.06	<0.001
	Unspecified soft tissue disorders → I67	822	43067	1.57	<0.001	822	14474	1.98	<0.001
	Other anaemias → I67	486	43403	1.47	<0.001	486	10361	1.46	<0.001

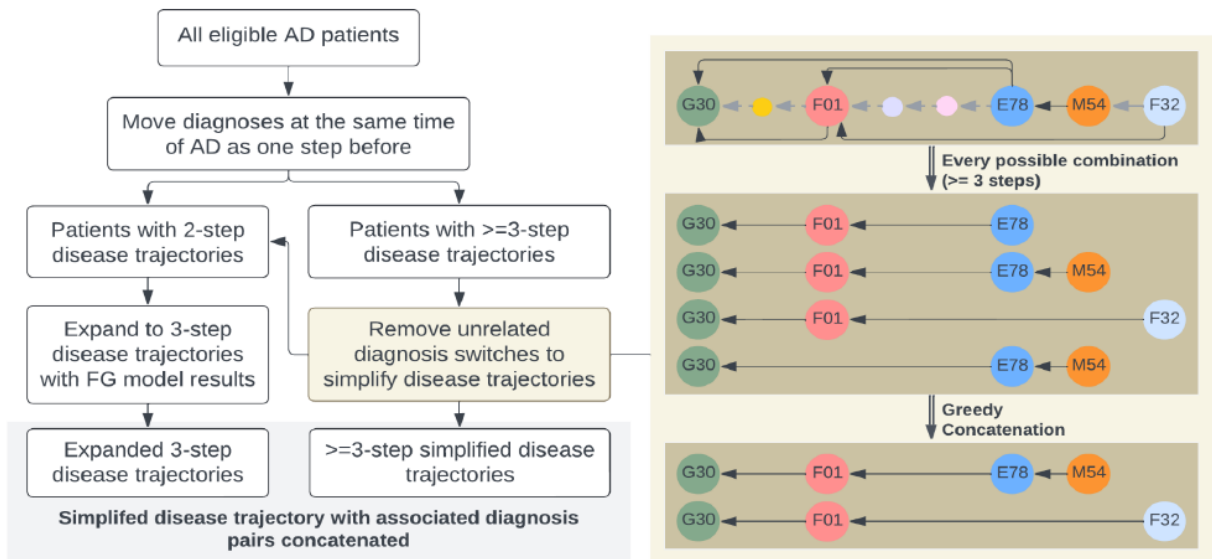
Abbreviations: MCI, mild cognitive impairment; HR, hazard ratio.

Notes:

[a] A single factor indicates that patients possess only one of the ICD codes within the risk trajectory.

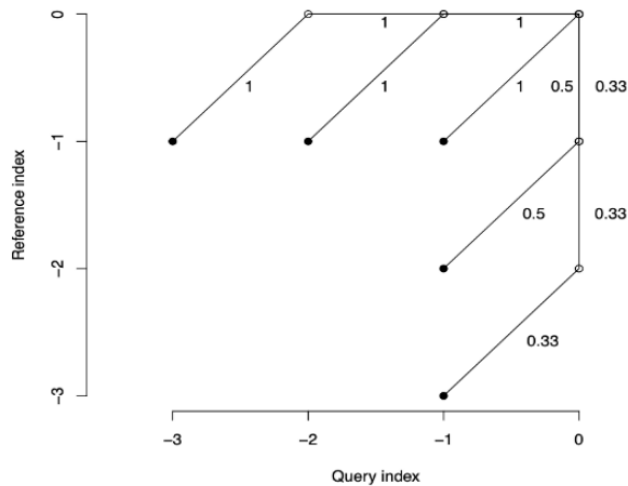
Figure 3.10: Associations between risk trajectories and incidence of Alzheimer's disease

3.10 Supplementary materials



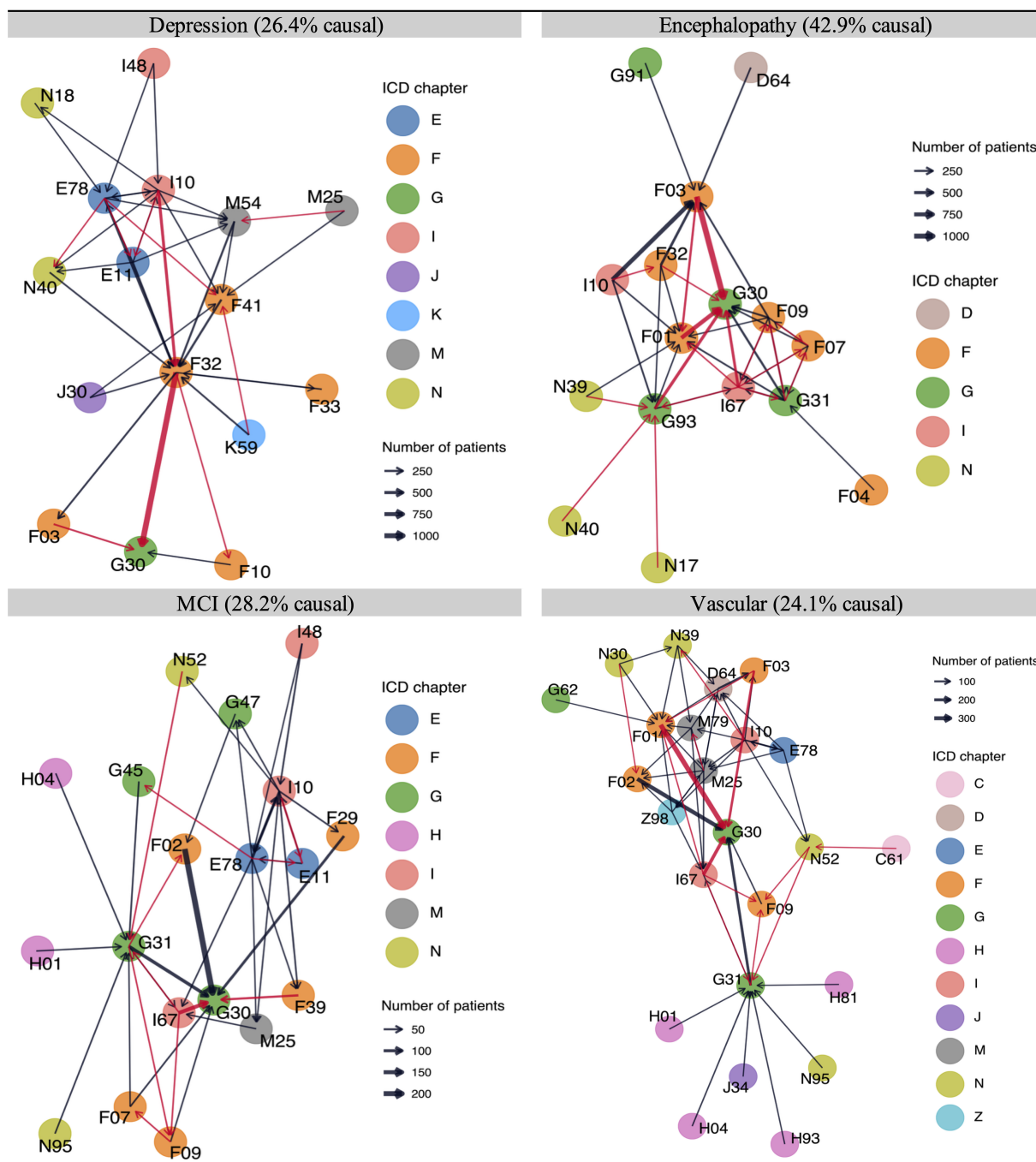
Supplementary Figure 1. Steps to simplify disease trajectories. Abbreviations: AD, Alzheimer's disease.

Steps to simplify disease trajectories. Abbreviations: AD, Alzheimer's disease.



Supplementary Figure 2. Asymmetric P05 slope weights

Asymmetric P05 slope weights



Supplementary Figure 3. Simplified networks by Alzheimer's Disease trajectory clusters.

Red arrows represent causal links detected by the Greedy Equivalent Search algorithm.

Supplementary Table 1. Descriptive statistics and matching information of AD cases vs. 1:1 propensity score-matched non-AD controls in UC data warehouse (N = 48,946)^a

Characteristic	AD N = 24,473	Non-AD N = 24,473	p-value
Age at last encounter visit	81.12 (5.4)	81.3 (5.6)	0.01*
Gender (female)	15164 (62.0%)	15018 (61.4%)	0.17
Race-ethnicity			<0.001*
American Indian	62 (0.3%)	51 (0.2%)	
Asian	2940 (12.0%)	2599 (10.6%)	
Black	1353 (5.5%)	1139 (4.7%)	
Hispanic	2775 (11.3%)	2507 (10.2%)	
NH-White	15891 (64.9%)	16647 (68.0%)	
Others	1452 (5.9%)	1530 (6.3%)	
Location resource			<0.001*
UC Site 1	3465 (14.2%)	3449 (14.1%)	
UC Site 2	3575 (14.6%)	3307 (13.5%)	
UC Site 3	8123 (33.2%)	8602 (35.2%)	
UC Site 4	71 (0.3%)	47 (0.2%)	
UC Site 5	5168 (21.1%)	4815 (19.7%)	
UC Site 6	4071 (16.6%)	4253 (17.4%)	
Follow-up time	2.89 (3.19)	8.53 (2.14)	<0.001*

Abbreviations: AD, Alzheimer's disease; NH, Non-Hispanic.

Notes:

[a] Summary statistics are displayed as n (%) for categorical variables, and median (IQR) for continuous variables.

Descriptive statistics and matching information of AD cases vs. 1:1 propensity score-matched non-AD controls in UC data warehouse (N = 48,946)

Supplementary Table 2. Patients with Alzheimer's disease (AD) characteristics by trajectory clusters, all patients in each cluster (N = 5,762)^a

Characteristic	Depression	Encephalopathy	MCI	Vascular	Adjusted p-value
Demographics					
N total	1448	3,223	1,502	1,446	<0.001*
Age at last visit	81.9 (7.5)	81.6 (7.5)	82.4 (8)	82.8 (7.8)	<0.001*
Gender (female)	992 (68.5%)	1,925 (59.7%)	887 (59.1%)	844 (58.4%)	<0.001*
Race-ethnicity					0.01*
American Indian	<10	<10	<10	<10	
Asian	150 (10.4%)	405 (12.6%)	203 (13.5%)	211 (14.6%)	
Black	74 (5.1%)	198 (6.1%)	93 (6.2%)	83 (5.7%)	
Hispanic	180 (12.4%)	339 (10.5%)	150 (10.0%)	140 (9.7%)	
Non-Hispanic White	959 (66.2%)	2,052 (63.7%)	978 (65.1%)	936 (64.7%)	
Others	81 (5.6%)	223 (6.9%)	73 (4.9%)	74 (5.1%)	
Death	676 (46.7%)	1,434 (44.5%)	621 (41.3%)	522 (36.1%)	<0.001*
EHR features					
EHR record length	5.4 (6.4)	3.2 (5.3)	5.7 (5.7)	7.8 (4.9)	<0.001*
Total number of encounters	40 (74)	16 (35)	35 (55)	64 (68)	<0.001*
EHR record density (per year)	10.4 (11.5)	7.9 (12.1)	8.6 (10.7)	9.7 (9.1)	<0.001*
Total number of diagnoses	47 (52)	26 (34)	42 (42)	62 (39)	<0.001*
Location sources					<0.001*
UC Site 1	200 (13.8%)	315 (9.8%)	145 (9.7%)	183 (12.7%)	
UC Site 2	174 (12.0%)	739 (22.9%)	295 (19.6%)	211 (14.6%)	
UC Site 3	512 (35.4%)	1,056 (32.8%)	531 (35.4%)	573 (39.6%)	
UC Site 4	<10	15 (0.5%)	<10	<10	
UC Site 5	383 (26.5%)	582 (18.1%)	314 (20.9%)	291 (20.1%)	
UC Site 6	176 (12.2%)	516 (16.0%)	215 (14.3%)	188 (13.0%)	
AD features					
Age at AD diagnosis	79.5 (6.5)	79.5 (7.3)	80.3 (7.2)	80.8 (7.2)	<0.001*
AD to last visit	1.5 (3.3)	1.2 (2.9)	1.3 (2.9)	1.4 (2.8)	<0.001*
AD to death	2.3 (3.1)	2.4 (3.2)	2.1 (3.2)	2 (2.9)	0.01*

Abbreviations: MCI, mild cognitive impairment.

Notes:

[a] Summary statistics are displayed as n (%) for categorical variables and mean (standard deviation) for continuous variables. Categorical variables are compared using chi-square tests. Continuous variables are compared using Kruskal-Wallis tests for overall comparisons across the four clusters. P-values represent comparisons across all cluster and have been adjusted for multiple comparisons using the False Discovery Rate method. Cells with counts less than 10 were masked as "<10" to comply with the requirements of the UC Health Data Warehouse.

* Statistical significance at p-value < 0.05 level.

Patients with Alzheimer's disease characteristics by trajectory clusters, all patients in each cluster (N = 5,762)

Supplementary Table 3A. Symptoms (ICD-10 Chapter XVIII) involving cognition, perception, emotional state and behaviour (selected significantly different in distributions) before Alzheimer's Disease (AD) diagnosis by disease trajectory clusters^a

Characteristics	All symptoms before AD				
	Depression	Encephalopathy	MCI	Vascular	Adj p-value
Somnolence, stupor and coma (R40)	37 (4.2%)	92 (4.2%)	18 (4.5%)	52 (8.3%)	<0.001*
Other symptoms and signs involving cognitive functions and awareness (R41)	513 (58.3%)	1,201 (55.3%)	274 (69.2%)	522 (83%)	<0.001*
Dizziness and giddiness (R42)	154 (17.5%)	117 (5.4%)	52 (13.1%)	171 (27.2%)	<0.001*
Disturbances of smell and taste (R43)	14 (1.6%)	<10	<10	<10	<0.001*
Other symptoms and signs involving general sensations and perceptions (R44)	30 (3.4%)	49 (2.3%)	<10	25 (4%)	0.04*
Symptoms and signs involving emotional state (R45)	82 (9.3%)	130 (6%)	29 (7.3%)	70 (11.1%)	<0.001*
Symptoms and signs involving appearance and behaviour (R46)	26 (3%)	80 (3.7%)	15 (3.8%)	37 (5.9%)	0.03*
Characteristics	Symptoms within 3 years before AD				
	Depression	Encephalopathy	MCI	Vascular	Adj p-value
Somnolence, stupor and coma (R40)	32 (3.6%)	86 (4%)	17 (4.3%)	39 (6.2%)	0.09
Other symptoms and signs involving cognitive functions and awareness (R41)	397 (45.1%)	1,087 (50%)	236 (59.6%)	384 (61%)	<0.001*
Dizziness and giddiness (R42)	82 (9.3%)	74 (3.4%)	38 (9.6%)	90 (14.3%)	<0.001*
Disturbances of smell and taste (R43)	<10	<10	<10	<10	0.32
Other symptoms and signs involving general sensations and perceptions (R44)	29 (3.3%)	48 (2.2%)	<10	22 (3.5%)	0.09
Symptoms and signs involving emotional state (R45)	66 (7.5%)	122 (5.6%)	27 (6.8%)	59 (9.4%)	0.01*
Symptoms and signs involving appearance and behaviour (R46)	23 (2.6%)	73 (3.4%)	15 (3.8%)	31 (4.9%)	0.12
Characteristics	Symptoms within 1 year before AD				
	Depression	Encephalopathy	MCI	Vascular	Adj p-value
Somnolence, stupor and coma (R40)	26 (3%)	78 (3.6%)	10 (2.5%)	23 (3.7%)	0.68
Other symptoms and signs involving cognitive functions and awareness (R41)	250 (28.4%)	867 (39.9%)	166 (41.9%)	221 (35.1%)	<0.001*
Dizziness and giddiness (R42)	32 (3.6%)	40 (1.8%)	25 (6.3%)	41 (6.5%)	<0.001*
Disturbances of smell and taste (R43)	<10	<10	<10	<10	0.33
Other symptoms and signs involving general sensations and perceptions (R44)	21 (2.4%)	42 (1.9%)	<10	16 (2.5%)	0.68
Symptoms and signs involving emotional state (R45)	51 (5.8%)	116 (5.3%)	21 (5.3%)	38 (6%)	0.89
Symptoms and signs involving appearance and behaviour (R46)	17 (1.9%)	62 (2.9%)	10 (2.5%)	23 (3.7%)	0.33

Abbreviations: MCI, mild cognitive impairment.

Notes:

[a] Summary statistics are displayed as n (%) for categorical variables and mean (standard deviation) for continuous variables. Categorical variables are compared using chi-square tests. Continuous variables are compared using Kruskal-Wallis tests for overall comparisons across the four clusters. P-values represent comparisons across all cluster and have been adjusted for multiple comparisons using the False Discovery Rate method. Cells with counts less than 10 were masked as "<10" to comply with the requirements of the UC Health Data Warehouse.

* Statistical significance at p-value < 0.05 level.

Symptoms (ICD-10 Chapter XVIII) involving cognition, perception, emotional state and behaviour (selected significantly different in distributions) before Alzheimer's Disease diagnosis by disease trajectory clusters

Supplementary Table 3B. Symptoms (ICD-10 Chapter XVIII) (selected significantly different in distributions) before Alzheimer's Disease (AD) diagnosis by disease trajectory clusters^a

Characteristics	All symptoms before AD				
	Depression	Encephalopathy	MCI	Vascular	Adj p-value
Symptoms and signs involving the					
Circulatory and respiratory systems (R00-R09)	170 (19.3%)	260 (12%)	92 (23.2%)	183 (29.1%)	<0.001*
Digestive system and abdomen (R10-R19)	186 (21.1%)	250 (11.5%)	83 (21%)	185 (29.4%)	<0.001*
Skin and subcutaneous tissue (R20-R23)	121 (13.8%)	96 (4.4%)	47 (11.9%)	157 (25%)	<0.001*
Nervous and musculoskeletal systems (R25-R29)	219 (24.9%)	400 (18.4%)	102 (25.8%)	198 (31.5%)	<0.001*
Urinary system (R30-R39)	159 (18.1%)	201 (9.2%)	79 (19.9%)	175 (27.8%)	<0.001*
Cognition, perception, emotional state and behaviour (R40-R46)	358 (40.7%)	1,018 (46.8%)	219 (55.3%)	304 (48.3%)	<0.001*
Speech and voice (R47-R49)	64 (7.3%)	206 (9.5%)	26 (6.6%)	90 (14.3%)	<0.001*
General symptoms and signs (R50-R69)	199 (22.6%)	457 (21%)	98 (24.7%)	177 (28.1%)	<0.001*
Abnormal findings on					
Examination of blood, without diagnosis (R70-R79)	208 (23.6%)	250 (11.5%)	83 (21%)	215 (34.2%)	<0.001*
Examination of urine, without diagnosis (R80-R82)	57 (6.5%)	45 (2.1%)	23 (5.8%)	75 (11.9%)	<0.001*
Examination of other body fluids, substances and tissues, without diagnosis (R83-R89)	20 (2.3%)	10 (0.5%)	<10	16 (2.5%)	<0.001*
Diagnostic imaging and in function studies, without diagnosis (R90-R94)	147 (16.7%)	224 (10.3%)	65 (16.4%)	188 (29.9%)	<0.001*
Ill-defined and unknown causes of mortality (R95-R99)	24 (2.7%)	26 (1.2%)	12 (3%)	30 (4.8%)	<0.001*
Symptoms within 3 years before AD					
	Depression	Encephalopathy	MCI	Vascular	Adj p-value
Symptoms and signs involving the					
Circulatory and respiratory systems (R00-R09)	177 (20.1%)	254 (11.7%)	92 (23.2%)	187 (29.7%)	<0.001*
Digestive system and abdomen (R10-R19)	159 (18.1%)	237 (10.9%)	70 (17.7%)	150 (23.8%)	<0.001*
Skin and subcutaneous tissue (R20-R23)	85 (9.7%)	71 (3.3%)	39 (9.8%)	97 (15.4%)	<0.001*
Nervous and musculoskeletal systems (R25-R29)	198 (22.5%)	378 (17.4%)	96 (24.2%)	171 (27.2%)	<0.001*
Urinary system (R30-R39)	147 (16.7%)	183 (8.4%)	69 (17.4%)	140 (22.3%)	<0.001*
Cognition, perception, emotional state and behaviour (R40-R46)	357 (40.6%)	962 (44.3%)	209 (52.8%)	306 (48.6%)	<0.001*
Speech and voice (R47-R49)	50 (5.7%)	193 (8.9%)	20 (5.1%)	69 (11%)	0.01
General symptoms and signs (R50-R69)	231 (26.2%)	453 (20.8%)	97 (24.5%)	203 (32.3%)	<0.001*
Abnormal findings on					
Examination of blood, without diagnosis (R70-R79)	158 (18%)	219 (10.1%)	65 (16.4%)	167 (26.6%)	<0.001*
Examination of urine, without diagnosis (R80-R82)	38 (4.3%)	35 (1.6%)	19 (4.8%)	51 (8.1%)	<0.001*
Examination of other body fluids, substances and tissues, without diagnosis (R83-R89)	12 (1.4%)	<10	<10	<10	0.004
Diagnostic imaging and in function studies, without diagnosis (R90-R94)	123 (14%)	207 (9.5%)	61 (15.4%)	166 (26.4%)	<0.001*
Ill-defined and unknown causes of mortality (R95-R99)	11 (1.2%)	15 (0.7%)	<10	13 (2.1%)	0.02
Symptoms within 1 year before AD					
	Depression	Encephalopathy	MCI	Vascular	Adj p-value
Symptoms and signs involving the					
Circulatory and respiratory systems (R00-R09)	133 (15.1%)	197 (9.1%)	69 (17.4%)	128 (20.3%)	<0.001*
Digestive system and abdomen (R10-R19)	120 (13.6%)	201 (9.2%)	55 (13.9%)	89 (14.1%)	<0.001*
Skin and subcutaneous tissue (R20-R23)	39 (4.4%)	49 (2.3%)	21 (5.3%)	49 (7.8%)	<0.001*
Nervous and musculoskeletal systems (R25-R29)	137 (15.6%)	315 (14.5%)	72 (18.2%)	125 (19.9%)	0.002
Urinary system (R30-R39)	104 (11.8%)	153 (7%)	47 (11.9%)	87 (13.8%)	<0.001*
Cognition, perception, emotional state and behaviour (R40-R46)	273 (31%)	835 (38.4%)	165 (41.7%)	227 (36.1%)	<0.001*
Speech and voice (R47-R49)	33 (3.8%)	179 (8.2%)	15 (3.8%)	48 (7.6%)	<0.001*
General symptoms and signs (R50-R69)	206 (23.4%)	387 (17.8%)	75 (18.9%)	150 (23.8%)	<0.001*
Abnormal findings on					
Examination of blood, without diagnosis (R70-R79)	93 (10.6%)	168 (7.7%)	40 (10.1%)	87 (13.8%)	<0.001*
Examination of urine, without diagnosis (R80-R82)	15 (1.7%)	20 (0.9%)	<10	30 (4.8%)	<0.001*
Examination of other body fluids, substances and tissues, without diagnosis (R83-R89)	<10	<10	<10	<10	0.06
Diagnostic imaging and in function studies, without diagnosis (R90-R94)	84 (9.5%)	178 (8.2%)	43 (10.9%)	116 (18.4%)	<0.001*
Ill-defined and unknown causes of mortality (R95-R99)	<10	<10	<10	<10	0.44

Abbreviations: MCI, mild cognitive impairment.

Notes:

[a] Summary statistics are displayed as n (%) for categorical variables and mean (standard deviation) for continuous variables. Categorical variables are compared using chi-square tests. Continuous variables are compared using Kruskal-Wallis tests for overall comparisons across the four clusters. P-values represent comparisons across all cluster and have been adjusted for multiple comparisons using the False Discovery Rate method. Cells with counts less than 10 were masked as "<10" to comply with the requirements of the UC Health Data Warehouse.

* Statistical significance at p-value < 0.05 level.

Symptoms (ICD-10 Chapter XVIII) (selected significantly different in distributions) before Alzheimer's Disease diagnosis by disease trajectory clusters

Supplementary Table 4. Comparisons of diagnoses in alive vs. deceased Alzheimer's disease (AD) patients within disease trajectory cluster, significant diagnoses only^a

Depression (N = 880)				Encephalopathy (N = 2,173)				Vascular (N = 629)			
ICD	Alive	Deceased	Adj p-value	ICD	Alive	Deceased	Adj p-value	ICD	Alive	Deceased	Adj p-value
F33	82 (18%)	44 (10.4%)	0.01*	E03	122 (10.2%)	134 (13.8%)	0.04*	E11	130 (31.9%)	47 (21.2%)	0.07*
F41	195 (42.9%)	120 (28.2%)	<0.001*	F01	251 (20.9%)	245 (25.2%)	0.05*	F01	85 (20.9%)	73 (32.9%)	0.03*
G31	142 (31.2%)	73 (17.2%)	<0.001*	F02	220 (18.3%)	72 (7.4%)	<0.001*	F02	110 (27%)	27 (12.2%)	0.002*
G47	151 (33.2%)	106 (24.9%)	0.05*	F03	674 (56.1%)	608 (62.6%)	0.01*	F03	181 (44.5%)	128 (57.7%)	0.04*
G89	128 (28.1%)	67 (15.8%)	0.001*	F09	236 (19.7%)	154 (15.8%)	0.05*	I48	82 (20.1%)	66 (29.7%)	0.09*
H25	65 (14.3%)	37 (8.7%)	0.06*	G31	510 (42.5%)	286 (29.4%)	<0.001*	J18	34 (8.4%)	37 (16.7%)	0.04*
J06	57 (12.5%)	31 (7.3%)	0.06*	G47	171 (14.2%)	100 (10.3%)	0.02*	N17	40 (9.8%)	42 (18.9%)	0.04*
M47	59 (13%)	33 (7.8%)	0.07*	I10	497 (41.4%)	486 (50%)	<0.001*	R39	59 (14.5%)	16 (7.2%)	0.098*
M51	76 (16.7%)	35 (8.2%)	0.02*	I67	249 (20.7%)	162 (16.7%)	0.05*	R63	89 (21.9%)	72 (32.4%)	0.07*
M79	138 (30.3%)	91 (21.4%)	0.02*	N39	112 (9.3%)	129 (13.3%)	0.02*	Z01	182 (44.7%)	61 (27.5%)	0.002*
M85	81 (17.8%)	39 (9.2%)	0.02*	R26	163 (13.6%)	168 (17.3%)	0.05*	Z11	85 (20.9%)	25 (11.3%)	0.05*
R10	96 (21.1%)	64 (15.1%)	0.09*	R41	717 (59.7%)	484 (49.8%)	<0.001*	Z12	195 (47.9%)	68 (30.6%)	0.002*
R35	80 (17.6%)	51 (12%)	0.06*					Z13	174 (42.8%)	59 (26.6%)	0.003
R41	296 (64.1%)	217 (51.1%)	<0.001*					Z20	75 (18.4%)	22 (9.9%)	0.07*
R42	93 (20.4%)	61 (14.4%)	0.05*					Z66	29 (7.1%)	38 (17.1%)	0.005*
Z00	143 (31%)	134 (31.5%)	0.095*								
Z01	141 (31%)	82 (19.3%)	<0.001*	Mile Cognitive Impairment (N = 396)							
Z12	134 (29.5%)	74 (17.4%)	<0.001*	ICD	Alive	Deceased	Adj p-value				
Z13	123 (27%)	67 (15.8%)	<0.001*	F02	38 (16.2%)	<10	0.002*				
Z78	76 (16.7%)	45 (10.6%)	0.06*	F03	94 (40.2%)	100 (61.7%)	0.002*				
Z79	173 (38%)	125 (29.4%)	0.05*	G31	103 (44%)	48 (29.6%)	0.09*				
Z98	83 (18.2%)	50 (11.8%)	0.05*	Z11	36 (15.4%)	<10	0.002*				
				Z12	66 (28.2%)	26 (16%)	0.097*				

Abbreviations: MCI, mild cognitive impairment.

Notes:

[a] Summary statistics are displayed as n (%) for categorical variables and mean (standard deviation) for continuous variables. Categorical variables are compared using chi-square tests. Continuous variables are compared using Wilcoxon rank sum test. All p-values are adjusted using the False Discovery Rate for multiple testing. Cells with counts less than 10 were masked as "<10" to comply with the requirements of the UC Health Data Warehouse.

* Statistical significance at p-value ≤ 0.1 level.

Comparisons of diagnoses in alive vs. deceased Alzheimer's disease patients within disease trajectory cluster, significant diagnoses only

Supplementary Table 5. Alzheimer's Disease (AD) patient characteristics by disease trajectory clusters^a

	Uni-cluster	Multi-clusters (MCI +)	p-value
Demographics			
N with unique cluster (% total)	396 (26.4%)	1106 (73.6%)	-
Age at last visit	83 (6.8)	82.3 (7.9)	0.08
Gender (female)	241 (60.9%)	646 (58.4%)	0.43
Race-ethnicity			0.25
American Indian	<10	<10	
Asian	51 (12.9%)	152 (13.7%)	
Black	24 (6.1%)	69 (6.2%)	
Hispanic	29 (7.3%)	121 (10.9%)	
Non-Hispanic White	274 (69.2%)	704 (63.7%)	
Others	16 (4.0%)	57 (5.2%)	
Death	162 (40.9%)	459 (41.5%)	0.88
EHR features			
EHR record length	6 (5.7)	5.5 (5.9)	0.02*
Total number of encounters	36.5 (58.2)	35 (53.8)	0.25
EHR record density (per year)	8.3 (9.4)	8.8 (11.3)	0.01*
Total number of diagnoses	39 (46)	43 (40)	0.01*
Location sources			0.01*
UC Site 1	36 (9.1%)	109 (9.9%)	
UC Site 2	60 (15.2%)	235 (21.2%)	
UC Site 3	151 (38.1%)	380 (34.4%)	
UC Site 4	<10	<10	
UC Site 5	74 (18.7%)	240 (21.7%)	
UC Site 6	75 (18.9%)	140 (12.7%)	
AD features			
Age at AD diagnosis	80.3 (6.3)	80.3 (7.5)	0.55
AD to last visit	1.6 (3.1)	1.2 (2.8)	0.002*
AD to death	2.4 (3.1)	2 (3.2)	0.01*

Abbreviations: EHR, electronic health record; MCI, mild cognitive impairment.

Notes:

[a] Summary statistics are displayed as n (%) for categorical variables, and median (IQR) for continuous variables. P-values were calculated with Pearson's Chi-squared test or Wilcoxon rank sum test as appropriate. Cells with counts less than 10 were masked as "<10" to comply with the requirements of the UC Health Data Warehouse.

* Statistical significance at p-value < 0.05 level.

AD patient characteristics by disease trajectory clusters

Supplementary Table 6. Descriptive statistics of Alzheimer's disease (AD) patients by data sources, University of California Health Data Warehouse (UCHDW) vs. All of Us (AOU)^a

Characteristic	UCHDW	AOU	p-value
Demographics			
N total	5762	785	-
Age at last visit	82 (78, 85)	78 (73, 84)	<0.001*
Gender (female)	3548 (62%)	401 (51%)	<0.001*
Race-ethnicity			<0.001*
Amerian-Indian	13 (0.2%)	0 (0.0%)	
Asian	725 (13%)	12 (1.5%)	
Black	328 (5.7%)	88 (11%)	
Hispanic	611 (11%)	121 (15%)	
Non-Hispanic White	3730 (65%)	563 (72%)	
Others	355 (6.2%)	1 (0.1%)	
Death	2478 (43%)	68 (8.7%)	<0.001*
EHR features			
EHR record length	4.8 (1.8, 8.1)	5.7 (3.4, 7.9)	<0.001*
Total number of encounters	29 (10, 69)	22 (12, 36)	<0.001*
EHR record density (per year)	8.7 (4.7, 15.7)	4.3 (2.4, 7.3)	<0.001*
Total number of diagnoses	37 (19, 66)	44 (25, 67)	<0.001*
AD features			
Age at AD diagnosis	80 (76, 83)	76 (71, 81)	<0.001*
AD to last visit	1.36 (0.30, 3.28)	1.40 (0.36, 3.17)	0.97
AD to death	2.35 (1.00, 4.15)	1.44 (0.38, 3.66)	<0.001*

Notes:

[a] Summary statistics are displayed as n (%) for categorical variables, and median (Q1, Q3) for continuous variables. P-values were calculated with Pearson's Chi-squared test or Wilcoxon rank sum test as appropriate.

Descriptive statistics of Alzheimer's disease (AD) patients by data sources, University of California Health Data Warehouse (UCHDW) vs. All of Us (AOU)

Supplementary Table 7. Validation of pairwise ICD code associations using the Fine-Gray proportional subdistribution hazards model, All of Us sample

Exposure (Description)	Outcome	HR	Adjusted P ^a
F01 (Vascular dementia)	G30 (Alzheimer's disease)	5.51	<0.001*
F02 (Dementia in other diseases classified elsewhere)	G30 (Alzheimer's disease)	5.50	<0.001*
F03 (Unspecified dementia)	G30 (Alzheimer's disease)	8.15	<0.001*
F09 (Unsp mental disorder due to known physiological condition)	G30 (Alzheimer's disease)	3.86	<0.001*
F10 (Alcohol related disorders)	G30 (Alzheimer's disease)	1.33	0.106
F32 (Major depressive disorder, single episode)	G30 (Alzheimer's disease)	1.68	<0.001*
F39 (Unspecified mood [affective] disorder)	G30 (Alzheimer's disease)	1.51	0.028*
G31 (Oth degenerative diseases of nervous system, NEC)	G30 (Alzheimer's disease)	6.00	<0.001*
G93 (Other disorders of brain)	G30 (Alzheimer's disease)	2.23	<0.001*
I67 (Other cerebrovascular diseases)	G30 (Alzheimer's disease)	1.60	<0.001*
E78 (Disorders of lipoprotein metabolism and other lipidemias)	D64 (Other anemias)	0.91	0.627
I10 (Essential (primary) hypertension)	D64 (Other anemias)	1.14	0.562
M25 (Other joint disorder, not elsewhere classified)	D64 (Other anemias)	0.66	0.004*
M79 (Oth and unsp soft tissue disorders, not elsewhere classified)	D64 (Other anemias)	0.71	0.023*
N39 (Other disorders of urinary system)	D64 (Other anemias)	0.83	0.328
E78 (Disorders of lipoprotein metabolism and other lipidemias)	E11 (Type 2 diabetes mellitus)	0.41	<0.001*
I10 (Essential (primary) hypertension)	E11 (Type 2 diabetes mellitus)	0.62	<0.001*
I10 (Essential (primary) hypertension)	E78 (Disorders of lipoprotein metabolism and other lipidemias)	1.06	0.605
D64 (Other anemias)	F01 (Vascular dementia)	0.75	0.371
F03 (Unspecified dementia)	F01 (Vascular dementia)	0.84	0.592
F09 (Unsp mental disorder due to known physiological condition)	F01 (Vascular dementia)	0.68	0.556
F32 (Major depressive disorder, single episode)	F01 (Vascular dementia)	0.95	0.865
G31 (Oth degenerative diseases of nervous system, NEC)	F01 (Vascular dementia)	0.99	0.99
G62 (Other and unspecified polyneuropathies)	F01 (Vascular dementia)	1.35	0.441
G93 (Other disorders of brain)	F01 (Vascular dementia)	1.23	0.535
I10 (Essential (primary) hypertension)	F01 (Vascular dementia)	1.26	0.605
I67 (Other cerebrovascular diseases)	F01 (Vascular dementia)	1.02	0.969
M25 (Other joint disorder, not elsewhere classified)	F01 (Vascular dementia)	0.88	0.686
M79 (Oth and unsp soft tissue disorders, not elsewhere classified)	F01 (Vascular dementia)	1.00	0.99
N30 (Cystitis)	F01 (Vascular dementia)	0.88	0.813
N39 (Other disorders of urinary system)	F01 (Vascular dementia)	0.88	0.671
Z98 (Other postprocedural states)	F01 (Vascular dementia)	0.96	0.897
G31 (Oth degenerative diseases of nervous system, NEC)	F02 (Dementia in other diseases classified elsewhere)	1.15	0.667
G47 (Sleep disorders)	F02 (Dementia in other diseases classified elsewhere)	0.79	0.457
M25 (Other joint disorder, not elsewhere classified)	F02 (Dementia in other diseases classified elsewhere)	0.65	0.162
M79 (Oth and unsp soft tissue disorders, not elsewhere classified)	F02 (Dementia in other diseases classified elsewhere)	0.93	0.817
N30 (Cystitis)	F02 (Dementia in other diseases classified elsewhere)	0.59	0.449
Z98 (Other postprocedural states)	F02 (Dementia in other diseases classified elsewhere)	0.61	0.171
D64 (Other anemias)	F03 (Unspecified dementia)	0.77	0.09*
F32 (Major depressive disorder, single episode)	F03 (Unspecified dementia)	0.87	0.397
I10 (Essential (primary) hypertension)	F03 (Unspecified dementia)	0.90	0.605
G31 (Oth degenerative diseases of nervous system, NEC)	F07 (Personality & behavior disorders due to known physiologic cond)	1.23	0.809
G31 (Oth degenerative diseases of nervous system, NEC)	F09 (Unsp mental disorder due to known physiological condition)	0.51	0.028*
I67 (Other cerebrovascular diseases)	F09 (Unsp mental disorder due to known physiological condition)	0.31	0.021*
N52 (Male erectile dysfunction)	F09 (Unsp mental disorder due to known physiological condition)	0.66	0.433
F32 (Major depressive disorder, single episode)	F10 (Alcohol related disorders)	0.80	0.592
E11 (Type 2 diabetes mellitus)	F32 (Major depressive disorder, single episode)	1.09	0.605
E78 (Disorders of lipoprotein metabolism and other lipidemias)	F32 (Major depressive disorder, single episode)	0.61	<0.001*
F41 (Other anxiety disorders)	F32 (Major depressive disorder, single episode)	1.51	0.002*
I10 (Essential (primary) hypertension)	F32 (Major depressive disorder, single episode)	0.61	<0.001*
J30 (Vasomotor and allergic rhinitis)	F32 (Major depressive disorder, single episode)	0.57	0.005*
K59 (Other functional intestinal disorders)	F32 (Major depressive disorder, single episode)	0.46	<0.001*
M54 (Dorsalgia)	F32 (Major depressive disorder, single episode)	0.58	<0.001*
N40 (Enlarged prostate)	F32 (Major depressive disorder, single episode)	0.53	0.001*
F32 (Major depressive disorder, single episode)	F33 (Major depressive disorder, recurrent)	2.46	<0.001*
E78 (Disorders of lipoprotein metabolism and other lipidemias)	F39 (Unspecified mood [affective] disorder)	1.32	0.627
I10 (Essential (primary) hypertension)	F39 (Unspecified mood [affective] disorder)	0.88	0.815
E78 (Disorders of lipoprotein metabolism and other lipidemias)	F41 (Other anxiety disorders)	0.62	<0.001*
I10 (Essential (primary) hypertension)	F41 (Other anxiety disorders)	0.58	<0.001*
J30 (Vasomotor and allergic rhinitis)	F41 (Other anxiety disorders)	0.76	0.193
K59 (Other functional intestinal disorders)	F41 (Other anxiety disorders)	0.53	<0.001*
M25 (Other joint disorder, not elsewhere classified)	F41 (Other anxiety disorders)	0.44	<0.001*
M54 (Dorsalgia)	F41 (Other anxiety disorders)	0.52	<0.001*
G45 (Transient cerebral ischemic attacks and related syndromes)	G31 (Oth degenerative diseases of nervous system, NEC)	0.32	<0.001*
H04 (Disorders of lacrimal system)	G31 (Oth degenerative diseases of nervous system, NEC)	0.80	0.241
N52 (Male erectile dysfunction)	G31 (Oth degenerative diseases of nervous system, NEC)	0.71	0.106
E78 (Disorders of lipoprotein metabolism and other lipidemias)	G45 (Transient cerebral ischemic attacks and related syndromes)	0.74	0.415
E78 (Disorders of lipoprotein metabolism and other lipidemias)	G47 (Sleep disorders)	0.66	<0.001*
I10 (Essential (primary) hypertension)	G47 (Sleep disorders)	0.57	<0.001*
F32 (Major depressive disorder, single episode)	G93 (Other disorders of brain)	0.94	0.811
I10 (Essential (primary) hypertension)	G93 (Other disorders of brain)	0.76	0.271
I67 (Other cerebrovascular diseases)	G93 (Other disorders of brain)	1.30	0.328
N17 (Acute kidney failure)	G93 (Other disorders of brain)	1.56	0.028*

N39 (Other disorders of urinary system)	G93 (Other disorders of brain)	1.07	0.807
N40 (Enlarged prostate)	G93 (Other disorders of brain)	0.69	0.171
E78 (Disorders of lipoprotein metabolism and other lipidemias)	I67 (Other cerebrovascular diseases)	0.73	0.193
G31 (Oth degenerative diseases of nervous system, NEC)	I67 (Other cerebrovascular diseases)	0.61	0.023*
G93 (Other disorders of brain)	I67 (Other cerebrovascular diseases)	1.06	0.842
M25 (Other joint disorder, not elsewhere classified)	I67 (Other cerebrovascular diseases)	0.53	0.002*
M79 (Oth and unsp soft tissue disorders, not elsewhere classified)	I67 (Other cerebrovascular diseases)	0.82	0.406
Z98 (Other postprocedural states)	I67 (Other cerebrovascular diseases)	0.87	0.605
D64 (Other anemias)	M25 (Other joint disorder, not elsewhere classified)	0.52	<0.001*
E78 (Disorders of lipoprotein metabolism and other lipidemias)	M25 (Other joint disorder, not elsewhere classified)	0.60	<0.001*
I10 (Essential (primary) hypertension)	M25 (Other joint disorder, not elsewhere classified)	0.59	<0.001*
M79 (Oth and unsp soft tissue disorders, not elsewhere classified)	M25 (Other joint disorder, not elsewhere classified)	1.03	0.819
E11 (Type 2 diabetes mellitus)	M54 (Dorsalgia)	0.95	0.747
E78 (Disorders of lipoprotein metabolism and other lipidemias)	M54 (Dorsalgia)	0.54	<0.001*
I10 (Essential (primary) hypertension)	M54 (Dorsalgia)	0.58	<0.001*
M25 (Other joint disorder, not elsewhere classified)	M54 (Dorsalgia)	0.75	0.009*
I10 (Essential (primary) hypertension)	M79 (Oth and unsp soft tissue disorders, not elsewhere classified)	0.65	<0.001*
M25 (Other joint disorder, not elsewhere classified)	M79 (Oth and unsp soft tissue disorders, not elsewhere classified)	0.90	0.424
N39 (Other disorders of urinary system)	M79 (Oth and unsp soft tissue disorders, not elsewhere classified)	0.82	0.161
I10 (Essential (primary) hypertension)	N18 (Chronic kidney disease (CKD))	0.74	0.17
I10 (Essential (primary) hypertension)	N39 (Other disorders of urinary system)	0.61	0.001*
E11 (Type 2 diabetes mellitus)	N40 (Enlarged prostate)	0.88	0.581
I10 (Essential (primary) hypertension)	N40 (Enlarged prostate)	0.60	0.003*
E78 (Disorders of lipoprotein metabolism and other lipidemias)	N52 (Male erectile dysfunction)	0.46	0.002*
I10 (Essential (primary) hypertension)	N52 (Male erectile dysfunction)	0.37	<0.001*
I10 (Essential (primary) hypertension)	Z98 (Other postprocedural states)	0.50	<0.001*
M25 (Other joint disorder, not elsewhere classified)	Z98 (Other postprocedural states)	0.53	<0.001*

Abbreviations: HR, Hazard Ratio.

Notes:

[a] Adjusted p-value using the False Discovery Rate for multiple testing.

* Statistical significance at p-value ≤ 0.1 level.

Validation of pairwise ICD code associations using the Fine-Gray proportional subdistribution hazards model, All of Us sample (cont)

Supplementary Table 8. Associations between risk trajectories and incidence of Alzheimer's disease, validation in All of Us^a

Cluster	Risk trajectory	Risk trajectory vs. all others			Risk trajectory vs. single factor		
		Sample N	RR	P-value	Sample N	RR	P-value
Depression (Central node - F32)	Disorders of lipoprotein metabolism → F32	8512	1.42	<0.001	6341	1.63	<0.001
	Anxiety disorders → F32	8512	1.30	0.02	2457	0.98	0.85
	Essential hypertension → F32	8512	1.42	<0.001	6429	1.50	<0.001
	Type 2 diabetes → F32	8512	1.69	<0.001	3729	1.51	<0.001
	Dorsalgia → F32	8512	1.38	0.001	4478	1.50	<0.001
	Other functional intestinal disorders → F32	8512	1.08	0.63	2779	1.00	0.99
Encephalopathy (Central node - G93)	Disorders of lipoprotein metabolism → G93	8512	2.04	<0.001	6549	2.27	<0.001
	Essential hypertension → G93	8512	1.98	<0.001	6500	2.18	<0.001
	Type 2 diabetes → G93	8512	2.23	<0.001	3073	2.08	<0.001
	Diseases of the urinary system → G93	8512	1.83	<0.001	2678	1.84	<0.001
	Benign prostatic hyperplasia → G93	8512	1.62	0.01	2494	1.78	0.002
	Acute kidney failure → G93	8512	1.73	<0.001	1744	1.59	0.01
MCI (Central node - G31)	Transient cerebral ischemic attacks → G31	8512	0.64	0.1	1232	0.64	0.1
	Menopausal disorders → G31*	-	-	-	-	-	-
	Male erectile dysfunction → G31*	-	-	-	-	-	-
	Other inflammation of eyelid → G31*	-	-	-	-	-	-
	Disorders of the lacrimal system → G31	8512	1.81	<0.001	2180	1.81	<0.001
Vascular (Central node - I67)	Disorders of lipoprotein metabolism → I67	8512	1.43	<0.001	6496	1.59	<0.001
	Essential hypertension → I67	8512	1.37	0.0039	6463	1.45	0.001
	Other joint disorders → I67	8512	1.43	0.0035	5263	1.69	<0.001
	Other degenerative diseases of the nervous system → I67*	-	-	-	-	-	-
	Conditions caused by postprocedural states → I67	8512	1.38	0.07	2887	1.64	0.01
	Unspecified soft tissue disorders → I67	8512	1.50	0.001	4520	1.81	<0.001
	Other anemias → I67	8512	1.16	0.42	2705	1.20	0.35

Notes:

[a] A single factor indicates that patients possess only one of the ICD codes within the risk trajectory.

* Model experienced complete separation, where the Alzheimer's disease cases separate the predictor variable completely. Parameter estimates and p-values might be incorrect.

Associations between risk trajectories and incidence of Alzheimer's disease, validation in All of Us

CHAPTER 4

Improving Genetic Risk Modeling of Dementia in Underrepresented Populations

4.1 Introduction

4.1.1 The Promise and Peril of Precision Medicine in Dementia

Precision medicine offers unprecedented opportunities to tailor medical interventions to individual genetic profiles, particularly for complex neurodegenerative diseases like dementia [68, 124]. Genetic risk modeling through genome-wide association studies (GWAS) and polygenic risk scores (PRS) has emerged as a cornerstone approach, potentially identifying high-risk individuals years before clinical symptoms manifest [125].

However, the overwhelming majority of genetic research has been conducted in European populations, creating a profound "genomic gap" that undermines precision medicine principles and perpetuates health disparities [14, 126]. When European-derived genetic risk models are applied to non-European populations, their predictive accuracy often diminishes substantially due to differences in genetic architecture, linkage disequilibrium patterns, and allele frequencies [9, 127].

4.1.2 The Burden of Dementia in Underrepresented Populations

Dementia disproportionately affects underrepresented populations, with African Americans facing twice the risk and Hispanic Americans 1.5-fold increased risk compared to non-

Hispanic whites [128, 129]. Despite this burden, these populations remain systematically excluded from genetic research, with European ancestry individuals comprising over 78% of GWAS participants [130].

This exclusion perpetuates health disparities by limiting the clinical utility of genetic risk models in non-European populations and potentially exacerbating existing inequities in healthcare delivery [131]. Studying diverse populations could enhance disease understanding and identify novel therapeutic targets benefiting all populations [132].

4.1.3 Limitations of Current Genetic Risk Modeling Approaches

Traditional approaches to genetic risk modeling for dementia have relied heavily on polygenic risk scores (PRS), which aggregate the effects of multiple genetic variants identified through GWAS to quantify an individual’s genetic predisposition to disease [133]. While PRS have demonstrated considerable success in European populations, their performance in non-European ancestries is often suboptimal due to several fundamental limitations.

First, the transferability of genetic effect sizes across populations is constrained by differences in linkage disequilibrium (LD) patterns. Genetic variants identified as risk factors in European populations may not be the true causal variants but rather proxies in linkage disequilibrium with the actual causal mutations [134]. When these proxy variants are used to construct PRS for non-European populations, the different LD structures can result in poor predictive performance, as the proxy relationships may not hold across ancestral groups.

Second, allele frequencies vary substantially across populations, affecting both the statistical power to detect associations and the population-level impact of identified variants [135]. Variants that are common and influential in European populations may be rare or absent in other ancestries, while population-specific variants that contribute significantly to disease risk may remain undetected due to their absence or low frequency in European-focused studies.

Third, current PRS approaches often fail to account for population-specific genetic architecture and gene-environment interactions that may modulate disease risk differently across ancestral groups [67]. The assumption that genetic effects are uniform across populations ignores the complex evolutionary history and environmental adaptations that have shaped genetic variation in different human populations.

4.1.4 The Potential of Biobank-Linked Electronic Health Records and Machine Learning Approaches to Genetic Risk Modeling

Large-scale biobanks linked to electronic health records (EHRs) offer unprecedented opportunities to address traditional genetic research limitations [15]. These resources provide real-world clinical data, adequate statistical power for smaller population subgroups, and longitudinal disease progression data [136].

The UCLA ATLAS Community Health Initiative and All of Us Research Program exemplify this approach, systematically collecting genetic data from diverse populations and creating resources reflecting true population diversity [15, 137].

Machine learning methods offer promising avenues for addressing PRS limitations, particularly in diverse populations [138]. Elastic Net regression combines Lasso and Ridge regularization benefits, handling high-dimensional genetic data while performing automatic feature selection [139].

These approaches can incorporate functional genomic information to prioritize biologically relevant variants and integrate multi-disease information, potentially improving cross-population transferability and comprehensive risk assessment [140, 141].

4.2 Background and Related Work

4.2.1 The Evolution of Genetic Risk Assessment in Dementia

Dementia genetic risk assessment has evolved from single-gene analyses to sophisticated polygenic approaches. The *APOE* gene discovery in the 1990s provided the first evidence that common variants substantially influence disease risk [142, 143]. However, *APOE* alone has limited predictive utility [144].

GWAS revolutionized the field by identifying dozens of risk loci beyond *APOE*, implicating immune function, lipid metabolism, and tau pathology pathways [22, 145]. However, most studies focused on European populations, limiting generalizability to other ancestries [146].

4.2.2 Polygenic Risk Scores: Promise and Limitations

PRS represent major advances in genetic risk assessment, aggregating multiple variant effects into single disease predisposition measures [147]. AD-PRS demonstrate considerable success in European populations, with AUC values of 0.70-0.75 and effective risk stratification [125, 148].

However, European-derived PRS show consistently inferior performance in non-European populations, with dramatically reduced phenotypic variance explanation [9, 149]. This reduced transferability has important health equity implications, potentially excluding under-represented groups from precision medicine benefits [150].

4.2.3 Functional Genomics and Causal Variant Identification

Distinguishing causal variants from linkage disequilibrium proxies represents a fundamental challenge, particularly for cross-population studies where LD patterns vary [151]. Functional genomics approaches, including eQTL studies and chromatin interaction mapping,

offer strategies for prioritizing causal variants and improving biological interpretability [152, 153].

The FUMA platform provides comprehensive functional variant prioritization by integrating multiple genomic data types, potentially improving cross-population transferability [140].

4.2.4 Multi-Disease Approaches to Genetic Risk Modeling

Growing evidence suggests complex diseases share genetic risk factors and biological pathways, indicating multi-disease approaches may provide more comprehensive risk assessment [154]. In dementia, substantial overlap exists between different neurodegenerative diseases, reflecting shared pathophysiological mechanisms [141].

Machine learning methods can naturally incorporate multi-disease information and identify complex patterns missed by traditional approaches, though optimal combination strategies remain active research areas.

EHR-genetic data integration opens new possibilities for studying associations in real-world clinical settings [155]. EHR studies provide access to large, diverse populations and longitudinal disease progression data, though they present challenges including coding practice variations and missing data [136].

Large-scale initiatives like eMERGE Network, UK Biobank, and All of Us have demonstrated EHR-linked genetic study feasibility, paving the way for sophisticated analyses leveraging full EHR data richness [26, 156].

4.3 Methods

4.3.1 Study Design and Participants

We conducted a retrospective cohort study using electronic health record (EHR) data linked to genetic information from the UCLA ATLAS Community Health Initiative and validated our findings in the All of Us Research Program. The UCLA ATLAS Community Health Initiative collects biosamples during routine lab work at UCLA Health labs from a diverse population, which undergoes genotyping using a customized Illumina Global Screening Array [157]. Participants watch a short video explaining the initiative’s goals and record their consent decision. Detailed information regarding biobanking and consenting procedures is available in our previous publications [137, 158]. After the genotype quality control, there were 54,935 individuals with both genotype and UCLA EHR data. All ethical regulations relevant to human research participants were followed. As the genetic data and EHRs were de-identified, the study was exempt from human subject research regulations (UCLA IRB 21-000435).

We validated our models and findings using data from the All of Us Research Hub, one of the most diverse biomedical data resources in the United States. The All of Us Research Program serves as a centralized data repository, offering secure access to de-identified data from program participants [15]. For validation, we utilized data release version 7, encompassing 409,420 individuals, of which 245,400 have undergone whole genome sequencing.

4.3.2 Patient Genetic Data Preprocessing

Quality control was conducted using PLINK v1.9 [159], adhering to established guidelines [137]. Samples with a missingness rate exceeding 5% were removed. Low-quality SNPs with >5% missingness, monomorphic SNPs, and strand-ambiguous SNPs were excluded. Post-quality control, genotype imputation was performed via the Michigan Imputation Server

[160] to enhance the coverage of genetic variants and facilitate comparison across diverse genotyping platforms. SNPs with imputation $r^2 < 0.90$ or minor allele frequency $< 1\%$ were pruned. After these measures, 21,220,668 genotyped SNPs were retained across the 54,935 individuals. Finally, we restricted our analyses to SNPs that overlapped between UCLA ATLAS and All of Us, resulting in a total of 8,705,988 SNPs, ensuring consistency in genetic variables across datasets.

Genetic ancestry refers to the geographic origins of an individual’s genome, tracing back to their most recent biological ancestors [161]. GIA employs genetic data, a reference population, and inferential methodologies to categorize individuals within groups likely sharing common geographical ancestors [162]. In our UCLA ATLAS sample, we used the reference panel from the 1000 Genomes Project [135] and principal component analysis [163] to infer genetic ancestry. GIA groups included European American (EA), African American (AA), Hispanic Latino American (HLA), East Asian American (EAA), and South Asian American (SAA). For instance, individuals in the United States with recent biological ancestors inferred to be of Amerindian ancestry were designated as "HLA GIA" [164]. In addition, we calculated ancestry-specific principal components within each GIA group using principal component analysis.

4.3.3 Genetic Predictors

The initial step in our study involved identifying potential risk SNPs as candidate predictors for dementia using GWASs. A summary of the GWASs used and the steps to select candidate SNPs is provided in Supplementary Table 8 and Supplementary Figure 2.

We selected candidate SNPs from several neurodegenerative disease GWASs, including Alzheimer’s disease (AD) [22], Lewy body dementia (LBD) [165], Parkinson’s disease (PD) [166], progressive supranuclear palsy (PSP) [167], and stroke [168]. For AD, we used summary statistics from both European ancestry and African American ancestry GWASs, as well as a multi-ancestry meta-analysis. We prioritized independent genome-wide-significant

SNPs ($p < 5 \times 10^{-8}$) using a linkage disequilibrium (LD) r^2 threshold of 0.2 within a 250 kb window.

Additionally, we used the Functional Mapping and Annotation of Genome-Wide Association Studies (FUMA) tool [140] to prioritize independent gene-annotated SNPs. FUMA integrates positional, eQTL, and 3D chromatin mapping to identify SNPs that are likely to have functional consequences. We applied several filtering criteria to prioritize potentially causal variants:

1. Functional consequences: We prioritized SNPs with high Combined Annotation-Dependent Depletion (CADD) scores (≥ 12.37) [169], which predict the deleteriousness of genetic variants.
2. Expression effects: We included SNPs that were significant eQTLs ($p < 0.05$) in brain tissues from the GTEx database [152].
3. Chromatin interactions: We incorporated SNPs involved in chromatin interactions that could affect gene regulation in brain tissues.
4. Positional mapping: We included SNPs within gene boundaries or in regulatory regions based on chromatin state annotations.

4.3.4 Dementia Case Definition and Control Selection

We defined dementia cases as individuals with at least one encounter containing any of the following ICD-10 codes: G30 (Alzheimer’s disease), F03 (Unspecified dementia), F02 (Dementia in other diseases classified elsewhere), F01 (Vascular dementia), or G31 (Other degenerative diseases of nervous system, not elsewhere classified). This broad definition was chosen to capture the full spectrum of dementia presentations in real-world clinical settings, recognizing that mixed pathologies are common and that diagnostic practices may vary across providers [170].

Controls were defined as individuals without any dementia-related ICD codes and with

at least 5 years of follow-up in the EHR system to reduce the likelihood of undiagnosed dementia. We applied additional exclusion criteria to ensure control quality:

1. Age requirements: Controls must be at least 65 years old at their last visit to ensure adequate exposure time for dementia development.
2. Healthcare utilization: Controls must have at least 2 encounters per year on average to ensure adequate opportunity for dementia diagnosis if present.
3. Exclusionary diagnoses: We excluded individuals with diagnoses that might confound dementia assessment, such as severe psychiatric disorders or developmental disabilities.

4.3.5 Statistical Modeling and Machine Learning Approaches

We developed and compared four different types of genetic risk models for dementia prediction:

1. *APOE*- $\epsilon 4$ count model: A simple model using the number of *APOE* $\epsilon 4$ alleles (0, 1, or 2) as the sole genetic predictor.
2. Polygenic risk score (PRS) models: Traditional PRS models using summary statistics from various GWAS sources, including ancestry-specific and multi-ancestry studies.
3. Multiple PRS models: Models incorporating PRS from multiple neurodegenerative diseases simultaneously.
4. Elastic Net SNP models: Machine learning models using individual SNPs selected through Elastic Net regularization.

For all models, we applied offset corrections to account for demographic factors (age, sex) and genetic population structure (represented by principal components). The models were expressed as:

$$\hat{y}_i = g^{-1}(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \text{offset}_i)$$

where $\hat{y}_i$ represents the predicted dementia status, $g^{-1}(\cdot)$ is the inverse of the link function, and offset_i accounts for demographic and population structure covariates.

The Elastic Net regularization combines the benefits of both Lasso (L1) and Ridge (L2) regression methods to enhance model stability and handle multicollinearity [139]. This technique aids in variable selection by reducing the coefficients of less relevant variables to zero, simplifying the model and improving its ability to manage multicollinearity. The hyperparameter λ , which balances L1 and L2 regularization, was optimized using a grid search to maximize the penalized likelihood within each training set.

We employed a 5-fold cross-validation methodology across all models to evaluate performance, with final results reported on the combined hold-out testing sets (Figure 4.1). To enhance the robustness of our findings, we utilized bootstrapping to determine feature importance, determine confidence intervals (CIs), and establish statistical significance. Specifically, we repeated the modeling process 1000 times using random sampling with replacement of all subjects (cases and controls) within the analytical sample set of each GIA group.

4.3.6 Model Evaluation and Comparison

The primary assessment criterion was the area under the precision-recall curve (AUPRC), chosen for its suitability in scenarios involving imbalanced datasets where the number of cases is significantly outnumbered by controls [171]. Additionally, the area under the receiver operating characteristic curve (AUROC) was reported as a comprehensive metric for model evaluation.

To determine the optimal threshold, we selected the point that maximized the Matthews Correlation Coefficient (MCC) [172]. Subsequent performance metrics, such as the F1 score, accuracy, precision, recall, and specificity, were computed based on this threshold.

To compare models, we calculated DeLong test p-values [173], which are specifically tailored for comparing two AUROC values derived from identical observations. Given the

lack of an equivalent test for AUPRC comparisons, we employed the paired Wilcoxon signed-rank test [51] to compare AUPRC using the bootstrapping results.

4.3.7 Gene Mapping and Pathway Analysis

We further examined the features selected from the Elastic Net SNP models. During bootstrapping, each iteration potentially identified a subset of SNPs as important features contributing to the dementia prediction. SNPs consistently identified in at least 95% of the 1000 bootstrap iterations were retained. To facilitate biological interpretations, we employed FUMA’s positional, eQTL, and chromatin interaction mapping to associate dementia risk SNPs from the top-performing Elastic Net SNP models with specific genes [140].

These mapped genes were tested against gene sets procured from MSigDB, including positional gene sets and Gene Ontology (GO) gene sets, to assess the enrichment of biological functions through hypergeometric tests. The Benjamini-Hochberg adjustment was applied to correct for multiple testing [92].

4.3.8 Validation in the All of Us Research Program

We conducted a validation study using the All of Us cohort to assess the generalizability of our findings derived from the UCLA ATLAS sample. A comparable sample was selected, adhering to the same criteria and sampling scheme for the GIA groups as in the UCLA ATLAS sample. We employed the same methodologies to define dementia cases and controls, extracting the same genetic risk loci from the All of Us Whole Genome Sequencing data for PRS construction or those identified through Elastic Net models in the UCLA ATLAS sample.

Consistent methodologies were used to regress out demographic variables and genetic population structure (i.e., PCs) to derive offset corrections, ensuring statistical models accurately reflect intrinsic genetic associations without confounding from demographic or population

genetic structure.

In the All of Us sample, we compared three models: (1) the *APOE-ε4* model; (2) the best-performing PRS model; and (3) the best-performing Elastic Net SNP model. The same evaluation metrics were utilized for model comparisons.

4.3.9 Statistical Analysis and Software

All analyses were conducted using R version 4.0.0 [174] and Python 3.8. Key R packages included: 'glmnet' for Elastic Net regression [175], 'pROC' for ROC analysis [176], 'PRROC' for precision-recall analysis [177], and 'ggplot2' for visualization [114]. Python packages included 'scikit-learn' for machine learning implementations [178] and 'pandas' for data manipulation [179].

Statistical significance was set at $p < 0.05$ for all analyses. Confidence intervals were calculated using bootstrap methods with 1000 iterations. Multiple testing correction was applied using the Benjamini-Hochberg method where appropriate.

4.4 Results

4.4.1 Sample Description

The primary dataset for model development was derived from EHR linked to the biobank of the UCLA Health System [137]. Figure 4.2 illustrates the finalized UCLA ATLAS samples, stratified by Genetic Inferred Ancestry (GIA) groups. The Hispanic Latino American (HLA) sample included 610 patients with 126 dementia cases, while the AA sample consisted of 440 patients with 84 dementia cases. The distribution of International Classification of Diseases, 10th Revision (ICD-10) diagnosis codes was relatively consistent across the two GIA samples, with Alzheimer's disease (G30) and unspecified dementia (F03) being the most prevalent diagnoses. However, the African American (AA) group exhibited a higher proportion of

vascular dementia (F01) diagnoses compared to the HLA group. The East Asian American (EAA) group, with a limited case count ($N = 75$), was excluded from primary analyses but included in sensitivity analyses.

Within each GIA group, eligible controls, due to the stringent inclusion criteria, had longer spans of records and more encounters. There were no significant differences in other EHR features between dementia cases and controls (Table 4.3).

4.4.2 Performance Comparison for Dementia Phenotype Prediction Task

We developed and evaluated machine learning models to predict the binary dementia phenotype within the UCLA ATLAS sample, stratified by GIA groups. After accounting for age, sex, and ancestry-specific genetic variations (represented by principal components (PCs)), we constructed genetic risk models for dementia, incorporating offset corrections within a linearized framework. The predictive capabilities of these models were assessed using four distinct sets of genetic markers: (1) *APOE*- $\epsilon 4$ counts, (2) AD PRS, (3) a composite of multiple PRSs, and (4) select SNPs refined through Elastic Net regularization [139]. For SNP selection, we utilized the Functional Mapping and Annotation of Genome-Wide Association Studies (FUMA) tool [140] to prioritize independent genome-wide-significant SNPs or independent gene-annotated SNPs.

Table 4.4 presents the overall performance of models for predicting dementia phenotypes. No discernible differences were observed among *APOE*- $\epsilon 4$ and all PRS models (AD-PRS and Multi-PRS), regardless of the SNP set used for PRS construction—whether derived from ancestry-specific GWASs, genome-wide-significant SNPs, or gene-annotated SNPs. Notably, the predictive performance in the AA GIA sample of all PRS models was inferior to *APOE*- $\epsilon 4$, particularly evident in the Area Under the Precision-Recall Curve (AUPRC).

Elastic Net SNP models demonstrated overall improvement in dementia prediction across both GIA groups. The model incorporating gene-annotated SNPs from AD and other

dementia-related disease GWASs (SNPs from AD + Neuro GWASs) emerged as the most effective, indicating a collective contribution from SNPs associated with various dementia-related diseases. Specifically, the leading Elastic Net SNP model for HLA GIA sample significantly enhanced the AUPRC by 31% (0.41 [95% CI: 0.27, 0.52] vs. 0.31 [95% CI: 0.21, 0.41], Wilcoxon signed-rank test p-value < 0.001), and the area under the receiver operating characteristic (AUROC) by 12% (0.73 [95% CI: 0.66, 0.79] vs. 0.65 [95% CI: 0.56, 0.72], DeLong test p-value = 0.01) compared to the best PRS model. Furthermore, this model outperformed the *APOE*- ϵ 4 count model, with an increase of 33% in AUPRC (0.41 vs. 0.31 [95% CI: 0.21, 0.41], Wilcoxon test p-value < 0.001) and 12% in AUROC (0.73 vs. 0.65 [95% CI: 0.56, 0.71], DeLong test p-value = 0.01).

This model's efficacy was even more pronounced within the AA GIA sample, with an increase in AUPRC by 84% (0.45 [95% CI: 0.31, 0.58] vs. 0.24 [95% CI: 0.11, 0.41], Wilcoxon test p-value < 0.001) and the AUROC by 16% (0.71 [95% CI: 0.63, 0.78] vs. 0.60 [95% CI: 0.46, 0.74], DeLong test p-value = 0.004) in comparison to the best PRS model. Relative to the *APOE*- ϵ 4 count model, the improvements were 65% in AUPRC (0.45 vs. 0.27 [95% CI: 0.16, 0.44], Wilcoxon test p-value < 0.001) and 17% in AUROC (0.71 vs. 0.61 [95% CI: 0.50, 0.74], DeLong test p-value = 0.01).

We also noted a substantial enhancement in the other performance metrics (based on the threshold that maximized the Matthews Correlation Coefficient (MCC)) of the Elastic Net SNPs models compared to other models across both GIA samples. This was evidenced by marked improvements in accuracy, precision, and the F1 score. In our sensitivity analysis, other non-linear models using gene-annotated SNPs from AD and other dementia-related disease GWASs, including Gradient Boosting Machine (GBM) and XGBoost, did not perform as well as the linear Elastic Net SNP models. Results from bootstrapping showed similar trends in model performances, as shown in Supplementary Figure 1. Applying a more stringent r^2 cut-off (< 0.1) for defining independent genome-wide-significant SNPs yielded results consistent with our initial findings, as detailed in Supplementary Table 1.

In summary, models leveraging SNPs as features identified through machine learning methods possess the potential to surpass those relying solely on summary scores such as PRSs in HLA and AA GIA. Furthermore, selecting SNPs mapped to genes using functional genomic data holds promise for further refining predictive performance.

4.4.3 Featured Risk Variants and Mapped Genes

In our analysis of the best-performing Elastic Net SNPs models, we examined the features selected by each model. According to results from bootstrapping (at least 95% of the 1000 iterations), the HLA and AA models identified 28 and 31 risk SNPs, respectively. The top 10 risk SNPs in variable importance selected by each model were shown in Table 4.5, with a detailed list, including related information, provided in Supplementary Table 2.

By assessing the feature importance of the SNPs chosen by the models, we found that for the HLA GIA group, the top three important predictors were rs429358 (chr19:44908684, nearest gene: *APOE*), rs2075650 (chr19:44892362, nearest gene: *TOMM40*), and rs483082 (chr19:44912921, nearest gene: *APOC1*), which together accounted for 25% of the total predictive importance. For the AA GIA group, the most influential predictors were rs2627641 (chr19:45205500, nearest gene: *BLOC1S3*), rs8073976 (chr17:44955857, nearest gene: *C1QL1*), and rs77283277 (chr7:143386852, nearest gene: *ZYX*).

Eight risk SNPs were identified by both GIA Elastic Net SNPs models, including two AD-associated SNPs (rs429358 and rs2075650) from the top 10 features in both GIA groups, though their relative importance varied slightly. Both models also identified several unique risk SNPs associated with Parkinson’s disease (PD), Progressive Supranuclear Palsy (PSP), and stroke as significant predictors of dementia. Notably, the AA GIA model highlighted the significance of a PSP-associated risk SNP, rs8073976, located on chromosome 17, underscoring the distinct genetic underpinnings influencing these different ancestry groups. These findings suggest that while there are common genetic markers associated with dementia across different ancestry groups, there are also unique genetic risk factors that could pro-

vide insights into the specific genetic architecture and risk profiles of dementia in diverse populations.

To better understand the biological functions and pathways associated with the identified risk variants, we mapped these risk SNPs to genes using FUMA, which integrates positional, eQTL, and 3D chromatin mapping [140]. Notably, 13 genes were identified by both non-European GIA models (Figure 4.6), suggesting shared biological pathways underlying dementia risk across different ancestry groups and suggests potential pathways for tailored interventions.

In the sensitivity analyses, we performed dementia risk modeling in the EAA GIA sample ($N = 673$). Similar to other GIA groups, the model incorporating gene-annotated SNPs from AD and other dementia-related disease GWASs performed the best compared to all other models. This model enhanced the AUPRC by 43% (0.34 [95% CI: 0.24, 0.43] vs. 0.24 [95% CI: 0.19, 0.29], Wilcoxon test p -value < 0.001) and the AUROC by 12% (0.80 [95% CI: 0.73, 0.86] vs. 0.71 [95% CI: 0.68, 0.74], DeLong test p -value = 0.001) compared to the best PRS model. It also outperformed the *APOE- $\epsilon 4$* count model, with increments of 42% in AUPRC (0.34 vs. 0.24 [95% CI: 0.19, 0.30], Wilcoxon test p -value < 0.001) and 11% in AUROC (0.80 vs. 0.71 [95% CI: 0.68, 0.73], DeLong test p -value = 0.004).

Among the featured 16 risk SNPs, rs429358 (chr19:44908684, nearest gene: *APOE*), rs66626994 (chr19:44924977, nearest gene: *APOC1P1*), and rs6857 (chr19:44888997, nearest gene: *NECTIN2*) were the most significant predictors for the EAA GIA group, collectively accounting for 45% of the overall predictive importance. After mapping these SNPs to gene, we identified the AD-risk gene cluster, *APOE-TOMM40-APOC1*, as well as the gene region near *chr17q21* (e.g., *FMNL1* and *SPPL2C*) (Supplementary Table 4A-D).

4.4.4 Validations in the All of Us Sample

We conducted a validation study using the All of Us cohort to evaluate the broad applicability of our findings obtained from the UCLA ATLAS sample. A comparable sample was selected from the All of Us Research Hub, employing the same selection scheme to their corresponding GIA groups in the UCLA ATLAS sample. However, due to the limited number of eligible dementia cases ($N_{\text{case}} = 8$) in the All of Us EAA GIA sample, we could only validate our models and findings in the HLA ($N_{\text{case}} = 68$) and AA ($N_{\text{case}} = 129$) GIA samples.

Table 4.7 presents the overall model performance of *APOE*- $\epsilon 4$ count, polygenic risk score, and Elastic Net SNP models in dementia genetic prediction in validation of All of Us sample, stratified by genetic inferred ancestry.

In particular, the Elastic Net SNP model demonstrated significant improvements over the other two models. In the HLA GIA sample, it outperformed both the *APOE*- $\epsilon 4$ and the best AD PRS model (AD AFR PRS.psig) by 17% and 4% in AUPRC (both Wilcoxon test $p\text{-value} < 0.001$), and by 2.7% and 4% in AUROC (DeLong test $p\text{-value} = 0.56$ and 0.25), respectively. Similarly, in the AA GIA sample, the Elastic Net SNP model showed a 35% and 13% enhancement in AUPRC (both Wilcoxon test $p\text{-value} < 0.001$), and 9% and 22% in AUROC (both DeLong test $p\text{-value} < 0.001$) over the *APOE*- $\epsilon 4$ and best AD PRS model, respectively.

4.5 Discussion

Traditional genetic risk models have faced limitations in effectively capturing causal disease risk variants and accurately assessing genetic risks across diverse populations. To address these challenges, our present study introduces a novel approach to predicting dementia risks by leveraging functional mapping of genetic data in conjunction with machine learning methods in the real-world EHR setting. Our proposed method shows remarkable improvements in prediction performance compared to well-known approaches like *APOE* gene and PRS

models. We successfully identified shared and ancestry-specific risk genes and biological pathways contributing to dementia risks for each non-European GIA group. Finally, we bolstered the reliability and generalizability of our findings by validating our models using a comparable EHR sample from the All of Us cohort.

Our study highlights the significance of prioritizing biologically meaningful SNPs in genetic prediction. GWASs often identify genomic regions with multiple correlated SNPs, which may encompass several closely located genes. However, not all of these genes are relevant to the disease [46]. Functional annotation of genetic variants enabled us to target potential causal SNPs by considering various factors, such as regional linkage disequilibrium (LD) patterns, functional consequences of variants, their impact on gene expression, and their involvement in chromatin interaction sites [140]. In our models developed on UCLA ATLAS samples, we achieved significant improvements in model performance by prioritizing biologically meaningful SNPs, ranging from 31–84% in AUPRC and 11–17% in AUROC across different GIA groups, compared to the *APOE*- ϵ 4 count and the best-performing PRS models. These results underscore the critical role of considering functional and biological information in enhancing the performance of genetic prediction models, especially in diverse populations.

It is worth highlighting that no discernible performance differences were observed between PRSs constructed using genome-wide-significant and gene-annotated SNPs. This can be attributed to the strong LD between genome-wide-significant and gene-annotated SNPs within the same genomic region. As a result, these SNPs tend to have similar effect estimates in the GWASs. Thus, it is expected that the PRSs built with these two sets of SNPs would exhibit a high correlation (Supplementary Table 7), which further supports the notion that the choice of genome-wide-significant or gene-annotated SNPs does not significantly impact the predictive performance of the PRSs in our study.

Moreover, our study emphasizes the significance of incorporating risk factors from multiple dementia-related diseases when developing predictive models for complex conditions like

dementia. Both ancestry-specific Elastic Net SNP models highlighted several PD and PSP risk variants as significant predictors of dementia. This finding aligns with the well-known complexity of dementia as a multifactorial disorder that shares common features with these related conditions [141]. However, it is worth noting that including PRSs of those diseases did not significantly improve the overall performance (Table 2). This result is consistent with research conducted by Clark et al. [180], in which they demonstrated that a combined genetic score, which incorporated risk variants for AD and 24 other traits, had an equivalent predictive power as the AD PRS on its own.

Our proposed Elastic Net SNPs models identified several shared risk factors across different ancestries. Notably, a substantial proportion of the identified shared genes were found near the *chr19q13* region, which is well-known for the AD risk gene cluster comprising *APOE-TOMM40-APOC1*. These findings align with previous research [143, 181, 182], further supporting the significance of this genomic region in contributing to the genetic risks associated with dementia.

At the same time, we have discovered compelling evidence supporting our hypothesis that risk SNPs associated with dementia, along with their corresponding weights, exhibit significant variations across diverse populations. Notably, our analysis of PRS models revealed that the performance of PRS built with the European population GWAS was worse when predicting a non-European GIA group. This is consistent with other previous studies. Using PRSs for 245 curated traits from the UK Biobank data, Privé et al. [149] revealed notable disparities in the phenotypic variance explained by PRSs across different populations. Specifically, compared to individuals of Northwestern European ancestry, the PRS-driven phenotypic variance is only 64.7% in South Asians, 48.6% in East Asians, and 18% in West Africans. Similarly, using a population from the Health and Retirement Study, Marden et al. demonstrated that the estimated effect of the AD PRS was notably smaller for non-Hispanic black compared to non-Hispanic white in both dementia probability score and memory score [183]. On the other hand, we also observed that the *APOE-ε4* count model performed better

than most PRS models in HLA and AA GIA samples. These finding further reinforces the limitations of standard PRS when applied to non-European populations, in which attempting to transfer GWAS effect size from one GIA to another GIA, or when using matched genetic ancestry GWAS with smaller sample size, as demonstrated in several AD and other phenotype studies [184–187].

In addition, we observed notable differences in the feature importance of various SNPs within the best-performing Elastic Net models across distinct GIA groups. Consequently, this led us to identify ancestry-specific genes and distinct biological pathways implicated in the genetic predisposition to dementia in diverse ancestral samples. These findings highlight the uniqueness of genetic risk factors and functional pathways in diverse population groups.

Finally, we validated our models using samples from separate EHR linked with genetic data (All of Us). Our proposed Elastic Net SNP model consistently outperformed the *APOE*- $\epsilon 4$ and the best PRS models. While the Elastic Net SNP model demonstrated improved performance in both HLA and AA populations, we observed a decrease in the general performance and significance (AUPRC and AUROC) in the All of Us sample compared to the UCLA ATLAS sample. One potential explanation for this discrepancy is the distinct population structure within each sample, as revealed by comparing patient characteristics (Supplementary Table 6). These findings underscore the influence of population-specific factors within GIA groups on the generalizability of genetic risk models, highlighting the critical need to account for population diversity in predictive models for complex diseases.

Our study boasts several notable strengths that contribute to its significance and impact. Firstly, we conducted our research with EHRs that mirror the practicalities of real-world community settings. Most current studies used longitudinal cohorts, which performed extensive testing and consensus criteria [188] applied by clinicians with expertise in dementias to diagnosis dementia. However, in real-world clinical care, the expertise in dementia may vary, and the criteria used for diagnosis may not always align with the stringent standards of research cohorts. Diagnoses documented in the EHRs capture these real-world data and, by routinely

capturing patient data over extended periods, form an expansive longitudinal cohort ideal for real-world research. Compared to traditional cohorts, EHR cohorts offer additional benefits, such as vast sample sizes, diverse phenotypes, and a more inclusive representation of often underrepresented groups, like minority groups and older adults [189]. Secondly, machine learning techniques applied in our study allowed us to infer crucial dementia risk factors for underrepresented populations, such as HLA and AA, with GWAS summary statistics from extensively studied populations like Europeans. This approach enabled a deeper understanding of the genetic landscape of dementia in underrepresented populations, particularly valuable given the current limitations in large-sample-size GWASs specific to these groups. Thirdly, we fortified the robustness and generalizability of our findings through the validation of our model on an independent dataset from the All of Us cohort. Furthermore, our innovative approach, which incorporated biologically relevant genetic markers and functional annotations, significantly enhanced the accuracy of disease prediction. This approach can be readily adapted to predict other complex diseases, extending the scope of its applications and enriching our understanding of diverse human populations' genetic traits.

However, we acknowledge certain limitations. Firstly, we observed variations in the composition of dementia subtypes among different GIA groups' case samples. Consequently, the distinct genes and biological pathways identified by different ancestry models should be interpreted with this consideration. Secondly, although our study identified potential risk SNPs and genes associated with dementia, additional experimentation is necessary to understand the precise mechanisms underlying the association of these factors. Thirdly, the limited number of dementia cases in our non-European GIA samples, after applying inclusion criteria, constrains the generalizability of our findings. Future studies should aim to replicate these findings in larger samples for each GIA to enhance their robustness. Finally, although detailed clinical guidelines for disease diagnoses exist [190, 191], clinical providers may adapt these criteria to fit specific research focuses or populations. This adaptation can lead to variations in diagnostic criteria across different studies or clinical practices. Consequently,

the precision of dementia diagnoses based on ICD-10 codes may vary compared to a gold standard of research criteria or autopsy findings.

In light of these limitations, further research with more extensive and diverse datasets, encompassing a broader range of dementia subtypes and GIA groups is imperative to strengthen the validity and applicability of our study’s outcomes. Such efforts will contribute to a more comprehensive understanding of the genetic complexities underlying dementia across diverse populations.

In conclusion, our study introduces a novel and robust approach to assessing individual genetic risks for dementia across diverse populations in a real-world setting. Our study demonstrates the importance of considering functional genomic information and machine learning approaches for developing more accurate and equitable genetic risk models. The identification of both shared and ancestry-specific risk factors provides insights into the biological mechanisms underlying dementia and highlights the need for population-specific approaches to precision medicine. Our findings have important implications for clinical practice and public health, particularly for addressing health disparities in dementia risk assessment and prevention.

4.6 Conclusions

In summary, we developed a novel machine learning approach for genetic risk modeling of dementia that significantly outperforms traditional methods across diverse populations. Our Elastic Net SNP models, which incorporate functional genomic information and variants from multiple neurodegenerative diseases, demonstrated substantial improvements in predictive performance compared to *APOE*-based and polygenic risk score models, with particularly pronounced benefits in underrepresented populations.

The identification of both shared and ancestry-specific genetic risk factors provides important insights into the biological mechanisms underlying dementia and highlights the com-

plex genetic architecture that varies across populations. The shared risk factors, particularly those in the *APOE-TOMM40-APOC1* gene cluster, reinforce the importance of established pathways while revealing additional functional variants that contribute to disease risk. The ancestry-specific findings underscore the critical need for population-specific approaches to genetic risk assessment and the limitations of applying European-derived models to diverse populations.

Our validation in the All of Us Research Program demonstrates the generalizability of our approach across different healthcare systems and populations, while also highlighting the importance of accounting for population-specific factors in genetic risk modeling. The consistent superior performance of our machine learning approach across multiple cohorts supports its potential for clinical implementation and its value for addressing health disparities in precision medicine.

This work establishes a methodological framework that can be readily adapted to other complex diseases and populations, contributing to the broader goal of developing equitable precision medicine approaches. The integration of functional genomics, multi-disease information, and machine learning provides a powerful paradigm for genetic risk modeling that addresses many of the limitations of current approaches.

Future research should focus on expanding these methods to larger and more diverse populations, investigating the biological mechanisms underlying the identified risk variants, and developing clinical decision support tools that can translate these findings into improved patient care. The ultimate goal is to ensure that the benefits of precision medicine are accessible to all populations, regardless of ancestry, and that genetic risk models contribute to reducing rather than exacerbating health disparities.

4.7 Code and data availability

Codes are publicly available on GitHub: <https://github.com/TSCchang-Lab/Dementia-prediction> (<https://doi.org/10.5281/zenodo.12754446>). Requests for additional information can be directed to the Lead Contact: Timothy S. Chang (timothychang@mednet.ucla.edu).

The Genome-Wide Association Study summary statistics data analyzed in this study are publicly available. Individual electronic health record data are not publicly available due to patient confidentiality and security concerns. Collaboration with the study authors who have been approved by UCLA Health for Institutional Review Board-qualified studies are possible and encouraged. Individual data from All of Us are publicly available for qualified researchers per the National Institutes of Health.

4.8 Acknowledgments

We gratefully acknowledge patients who have participated in the UCLA ATLAS Community Health Initiative and the All of Us Research Program. The UCLA ATLAS Community Health Initiative in collaboration with UCLA ATLAS Precision Health Biobank is a program of Institute for Precision Health (IPH), which directs and supports the biobanking and genotyping of biospecimen samples from participating UCLA patients in collaboration with the David Geffen School of Medicine, UCLA Clinical and Translational Science Institute and UCLA Health. We thank the All of Us Research Program participants and the research teams at the participating institutions for their contributions to this work.

This study was supported by the National Institutes of Health, National Institute on Aging (grants K08AG065519-01A1 to T.S.C. and M.F.), the National Science Foundation, the Hillblom and Fineberg Foundations, and the California Department of Public Health. B.P. was supported by the National Institutes of Health (R01HG009120, R01CA244670). K.V. was supported by the National Institutes of Health (R01AG064877, R01NS124882).

4.9 Figures

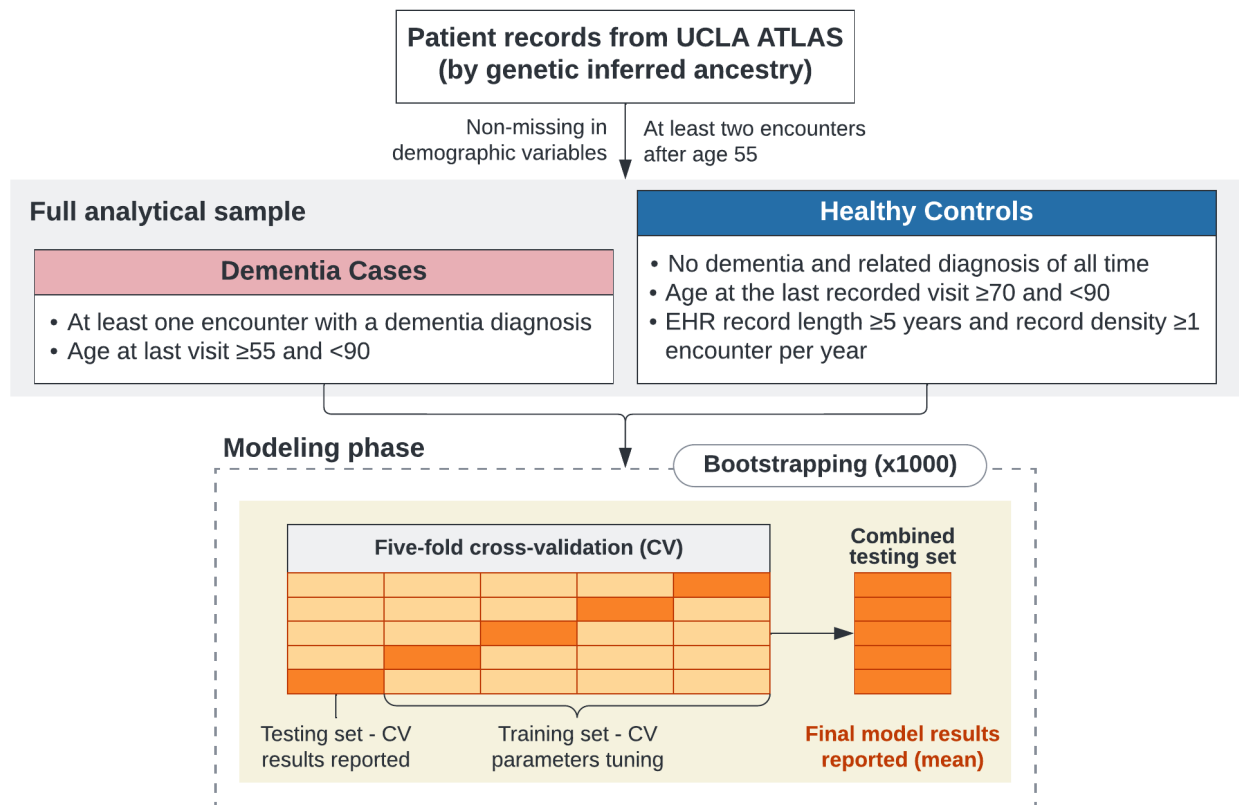


Figure 4.1: Sample selection steps and modeling steps description

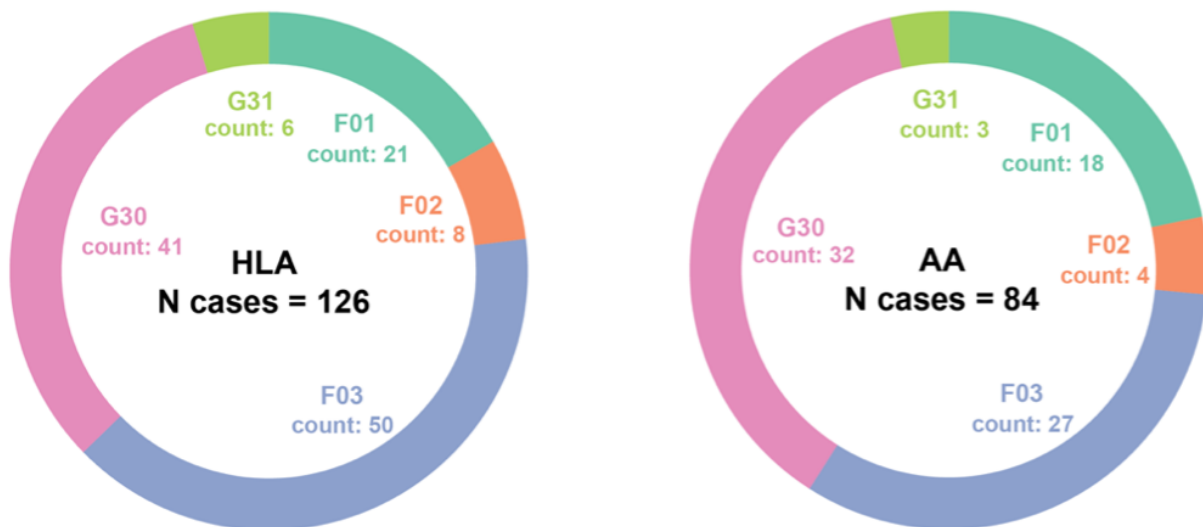


Figure 4.2: Dementia patient characteristics by genetic inferred ancestry groups, UCLA ATLAS sample

Distribution of diagnosis in ICD-10 codes by genetic inferred ancestry groups. Abbreviations: AA, African Americans; HLA: Hispanic Latino Americans. ICD-10 codes descriptions: G30, Alzheimer's disease; F03, Unspecified dementia; F02, Dementia in other diseases classified elsewhere; F01, Vascular dementia; G31, Other degenerative diseases of nervous system, not elsewhere classified.

Table 1. Descriptive statistics of demographic and electronic health record features by case/control groups, UCLA ATLAS sample, stratified by genetic inferred ancestry group

	Hispanic Latino Americans (N = 610)			African Americans (N = 440)		
	Cases	Controls	P value	Cases	Controls	P value
N	126	484	-	84	356	-
Age	78.4 (71.3, 81.7)	75.3 (72.6, 79.6)	0.2	78.0 (70.1, 82.6)	75.7 (72.7, 79.9)	0.7
Sex (Female)	72 (57%)	300 (62%)	0.30	46 (55%)	218 (61%)	0.30
Span of records (in yrs)	5.9 (2.8, 8.8)	9.6 (7.7, 10.9)	<0.001*	6.2 (3.1, 10.1)	9.9 (8.1, 11.4)	<0.001*
Encounters per year	16 (7, 25)	14 (8, 20)	0.05	14 (6, 28)	13 (9, 21)	0.60
Number of encounters	73 (26, 156)	124 (73, 205)	<0.001*	65 (28, 183)	140 (84, 210)	<0.001*
Number of unique diagnosis	68 (36, 113)	71 (47, 108)	0.40	61 (41, 99)	73 (47, 103)	0.20

Notes: Continuous variables were reported as median (IQR), and categorical variables were reported as n (%). P-values were calculated based on two-sided Wilcoxon rank sum test or Pearson's Chi-squared test as appropriate. * Statistically significant at level 0.05.

Figure 4.3: Descriptive statistics of demographic and electronic health record features by case/control groups, UCLA ATLAS sample, stratified by genetic inferred ancestry group

Table 2. Model performance of *APOE-e4* count, polygenic risk score, and SNP models in dementia genetic prediction, UCLA ATLAS sample, stratified by genetic inferred ancestry^a

		AUPRC	AUROC	F1 score	Accuracy	Precision	Recall	Specificity
Hispanic Latino Americans (N = 610)								
APOE	e4 count	0.308	0.652	0.424	0.707	0.357	0.524	0.754
AD-PRS models								
AD EUR PRS	P-significant	0.306	0.619	0.335	0.759	0.389	0.294	0.880
	Gene-annotated	0.288	0.615	0.387	0.397	0.245	0.921	0.260
AD AFR PRS	P-significant	0.312	0.644	0.409	0.692	0.339	0.516	0.738
	Gene-annotated	0.305	0.648	0.427	0.666	0.330	0.603	0.682
AD multi-ancestry PRS	P-significant	0.298	0.626	0.389	0.444	0.252	0.857	0.337
	Gene-annotated	0.298	0.640	0.401	0.448	0.259	0.897	0.331
Multi-PRS models								
PRSs using AD GWASs only ^b	P-significant	0.312	0.643	0.415	0.644	0.314	0.611	0.653
	Gene-annotated	0.302	0.646	0.404	0.670	0.322	0.540	0.705
PRSs using AD + Neuro GWASs ^c	P-significant	0.283	0.617	0.382	0.661	0.306	0.508	0.700
	Gene-annotated	0.309	0.643	0.411	0.662	0.321	0.571	0.686
Elastic Net SNPs models								
SNPs from AD GWASs only	P-significant	0.321	0.662	0.408	0.530	0.276	0.786	0.463
	Gene-annotated	0.351	0.679	0.436	0.602	0.308	0.746	0.564
SNPs from AD + Neuro GWASs	P-significant	0.359	0.715	0.472	0.633	0.336	0.794	0.591
	Gene-annotated	0.410	0.728	0.458	0.779	0.463	0.452	0.864
Non-linear SNPs models								
SNPs from AD + Neuro GWASs	GBM	0.304	0.634	0.381	0.707	0.337	0.437	0.777
Gene-annotated SNPs	XGBoost	0.298	0.642	0.375	0.710	0.338	0.421	0.785
African Americans (N = 440)								
APOE	e4 count	0.271	0.606	0.388	0.570	0.267	0.714	0.537
AD-PRS models								
AD EUR PRS	P-significant	0.221	0.592	0.369	0.432	0.234	0.869	0.329
	Gene-annotated	0.226	0.573	0.348	0.318	0.213	0.952	0.169
AD AFR PRS	P-significant	0.242	0.584	0.322	0.732	0.311	0.333	0.826
	Gene-annotated	0.241	0.581	0.344	0.584	0.246	0.571	0.587
AD multi-ancestry PRS	P-significant	0.234	0.592	0.360	0.386	0.225	0.905	0.264
	Gene-annotated	0.230	0.598	0.370	0.443	0.236	0.857	0.346
Multi-PRS models								
PRSs using AD GWASs only ^b	P-significant	0.238	0.589	0.358	0.527	0.242	0.690	0.489
	Gene-annotated	0.233	0.590	0.357	0.484	0.234	0.750	0.421
PRSs using AD + Neuro GWASs ^c	P-significant	0.187	0.516	0.311	0.195	0.186	0.952	0.017
	Gene-annotated	0.217	0.538	0.087	0.809	0.500	0.048	0.989
Elastic Net SNPs models								
SNPs from AD GWASs only	P-significant	0.356	0.669	0.356	0.802	0.471	0.286	0.924
	Gene-annotated	0.421	0.678	0.342	0.834	0.704	0.226	0.978
SNPs from AD + Neuro GWASs	P-significant	0.391	0.704	0.342	0.825	0.606	0.238	0.963
	Gene-annotated	0.446	0.710	0.365	0.834	0.677	0.250	0.972
Non-linear SNPs models								
SNPs from AD + Neuro GWASs	GBM	0.225	0.479	0.314	0.186	0.187	0.976	0.000
Gene-annotated SNPs	XGBoost	0.220	0.506	0.139	0.802	0.412	0.083	0.972

Abbreviations: AD, Alzheimer's Disease; APOE, apolipoprotein E; AUROC, Area Under the ROC Curve; AUPRC, Area Under the Precision-Recall Curve; EUR, European; GBM, Gradient Boosting Machine; GWAS, Genome-Wide Association Study; PRS, Polygenic Risk Score; SNP, Single-Nucleotide Polymorphism.

Notes:

[a] All models (if not other specified) have regressed out age, sex, and ancestry-specific principal components. Thresholds were determined by maximizing absolute Matthews correlation coefficient.

[b] All AD PRSs built with EUR, AFR, and multi-ancestry GWASs using P-significant/gene-annotated SNPs were included in the model at the same time.

[c] All AD PRSs built with EUR, AFR, and multi-ancestry GWASs, and neurodegenerative disease PRS (Parkinson's disease, progressive supranuclear palsy, Lewy body dementia, and stroke) using P-significant/gene-annotated SNPs were included in the model at the same time.

Figure 4.4: Model performance of *APOE-e4* count, polygenic risk score, and SNP models in dementia genetic prediction, UCLA ATLAS sample, stratified by genetic inferred ancestry

Table 3. Top 10 featured risk SNPs from the best-performing Elastic Net SNP model, UCLA ATLAS sample, stratified by genetic ancestry

rsID	CHR	POS	Variable Importance (95% CI)	Nearest Gene	AD EUR	AD AFR	AD multi	LBD	PD	PSP	Stroke
Hispanic Latino American ancestry (HLA)											
rs429358	19	44908684	0.089 (0.02, 0.11)	<i>APOE</i>		x					
rs2075650	19	44892362	0.089 (0.02, 0.11)	<i>TOMM40</i>		x	x	x			
rs483082	19	44912921	0.07 (0.017, 0.086)	<i>APOC1</i>		x	x				
rs157581	19	44892457	0.065 (0.014, 0.079)	<i>TOMM40</i>		x		x			
rs412776	19	44876259	0.057 (0.016, 0.066)	<i>NECTIN2</i>	x		x				
rs62120578	19	44713297	0.053 (0.022, 0.061)	<i>CTB-171A8.1</i>	x						
rs4803765	19	44855191	0.048 (0.015, 0.056)	<i>NECTIN2</i>	x						
rs80100206	4	705856	0.042 (0.02, 0.046)	<i>PCGF3</i>					x		
rs6857	19	44888997	0.038 (0.011, 0.042)	<i>NECTIN2</i>		x					
rs2276412	11	121590137	0.037 (0.018, 0.039)	<i>SORL1</i>	x						
African American ancestry (AA)											
rs2627641	19	45205500	0.096 (0.077, 0.099)	<i>BLOC1S3</i>	x						
rs8073976	17	44955857	0.079 (0.065, 0.082)	<i>CIQL1</i>						x	
rs77283277	7	143386852	0.076 (0.063, 0.079)	<i>ZYX</i>	x						
rs429358	19	44908684	0.071 (0.059, 0.074)	<i>APOE</i>		x					
rs2075650	19	44892362	0.068 (0.056, 0.071)	<i>TOMM40</i>		x	x	x			
rs73936967	19	44890485	0.064 (0.053, 0.066)	<i>TOMM40</i>		x					
rs13032148	2	127107524	0.063 (0.053, 0.065)	<i>BIN1</i>	x		x				
rs71352239	19	44926286	0.056 (0.047, 0.057)	<i>APOC1P1</i>	x		x	x			
rs435380	19	44903861	0.041 (0.037, 0.042)	<i>TOMM40</i>		x	x				
rs11223641	11	133950127	0.041 (0.037, 0.041)	<i>IGSF9B</i>					x		

Abbreviations: AD, Alzheimer's Disease; AFR, African American; CI, confidence interval; EUR, European; LBD, Lewy body dementia; PD, Parkinson's disease; PRS, Polygenic Risk Score; PSP, progressive supranuclear palsy; SNP, Single-Nucleotide Polymorphism. Note: SNPs marked in red are overlapped SNPs identified by both samples.

Figure 4.5: Top 10 featured risk SNPs from the best-performing Elastic Net SNP model, UCLA ATLAS sample, stratified by genetic ancestry

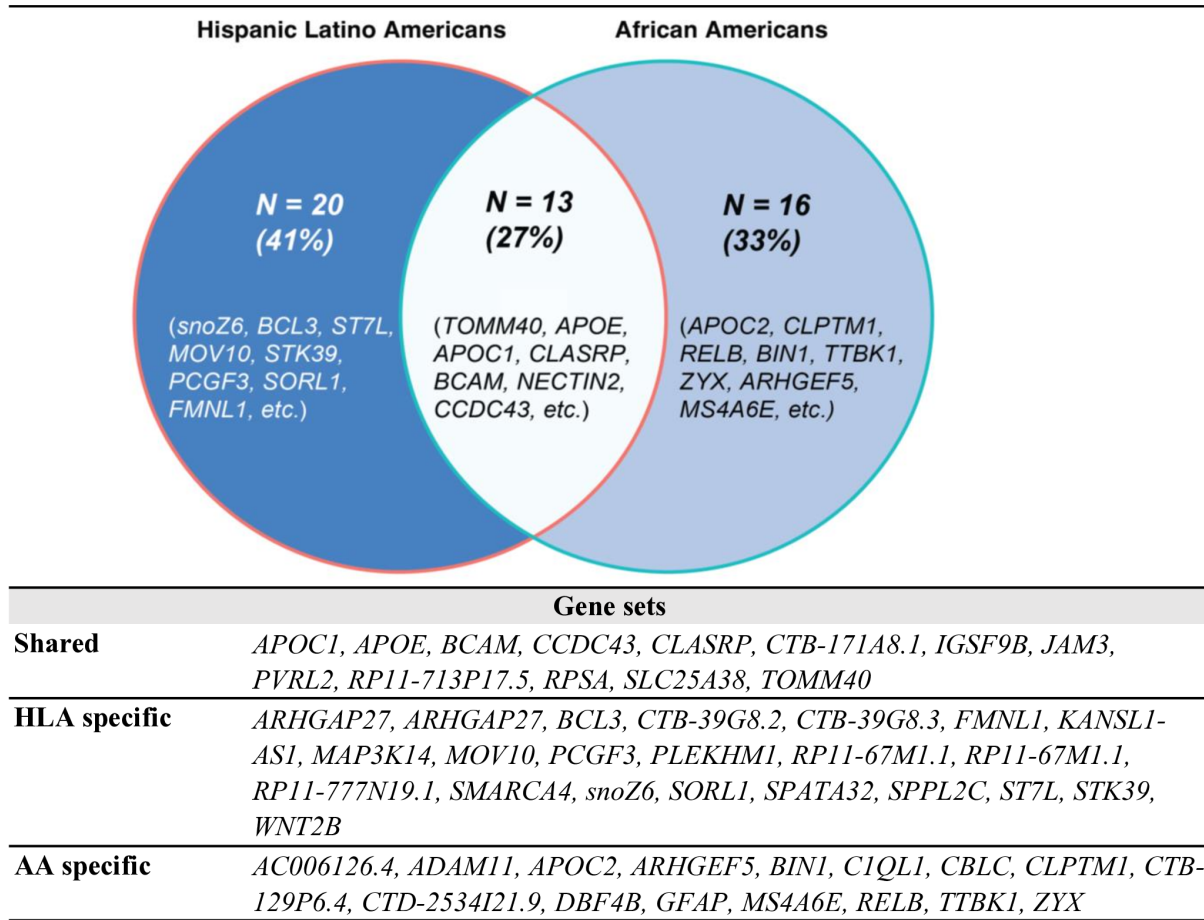


Figure 4.6: Descriptive statistics of demographic and electronic health record features by case/control groups, UCLA ATLAS sample, stratified by genetic inferred ancestry group

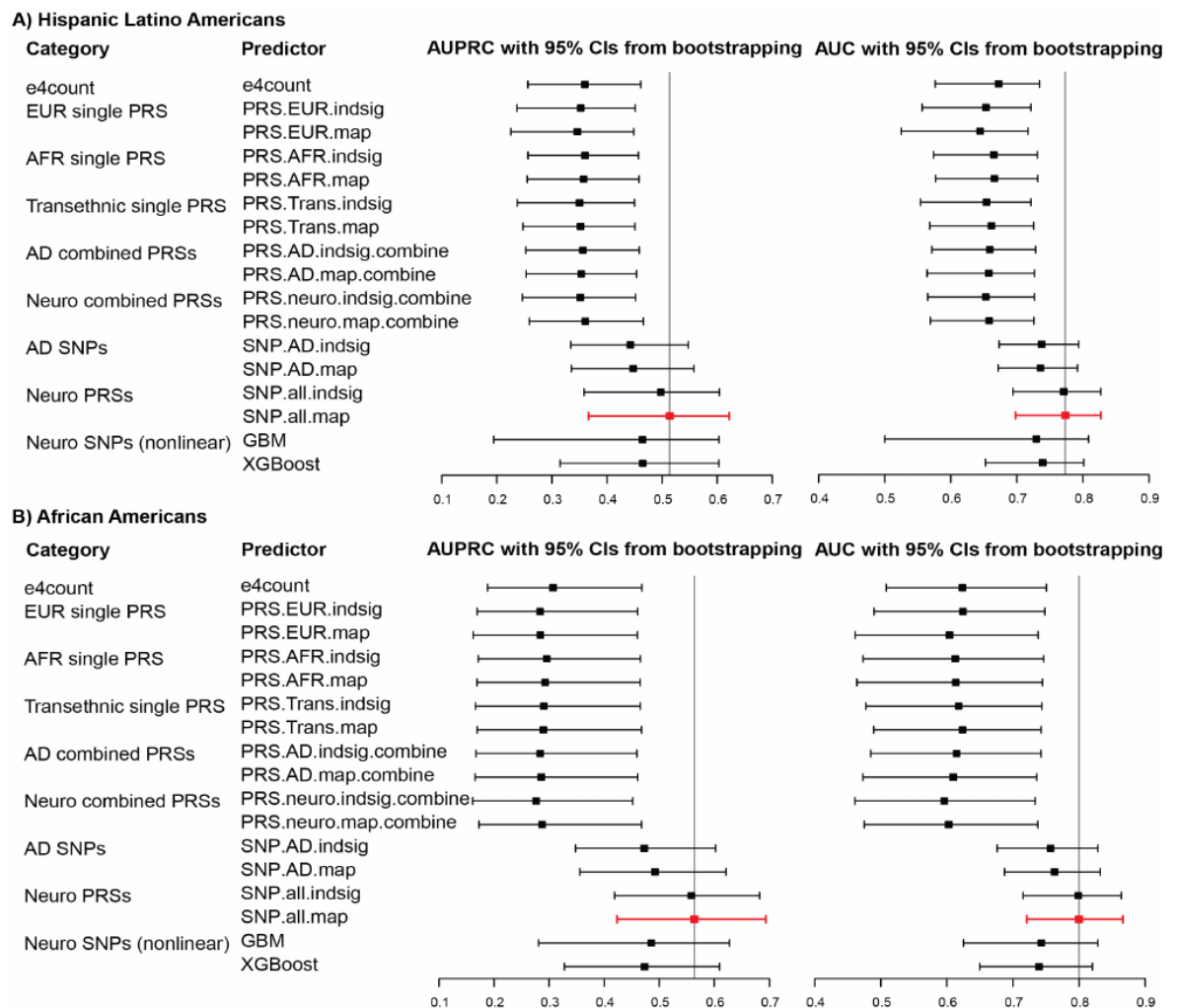
Table 4. Overall model performance of *APOE-e4* count, polygenic risk score, and Elastic Net SNP models in dementia genetic prediction in validation of All of Us sample, stratified by genetic inferred ancestry

N case		HLA (N = 458)		AA (N = 645)	
		Cases	Controls	Cases	Controls
Model	N	68	390	129	516
<i>APOE</i>	e4 count	AUPRC	AUROC	AUPRC	AUROC
Best single AD PRS	AFR P-significant	0.206	0.601	0.211	0.588
Best SNPs	Gene-annotated Neuro SNPs	0.231	0.593	0.253	0.524
		0.240	0.617	0.285	0.639

Abbreviations: AA, African Americans; AD, Alzheimer's Disease; AFR, African American; APOE, apolipoprotein E; AUROC, Area Under the ROC Curve; AUPRC, Area Under the Precision-Recall Curve; HLA: Hispanic Latino Americans; PRS, Polygenic Risk Score; SNP, Single-Nucleotide Polymorphism.

Figure 4.7: Overall model performance of *APOE-e4* count, polygenic risk score, and Elastic Net SNP models in dementia genetic prediction in validation of All of Us sample, stratified by genetic inferred ancestry

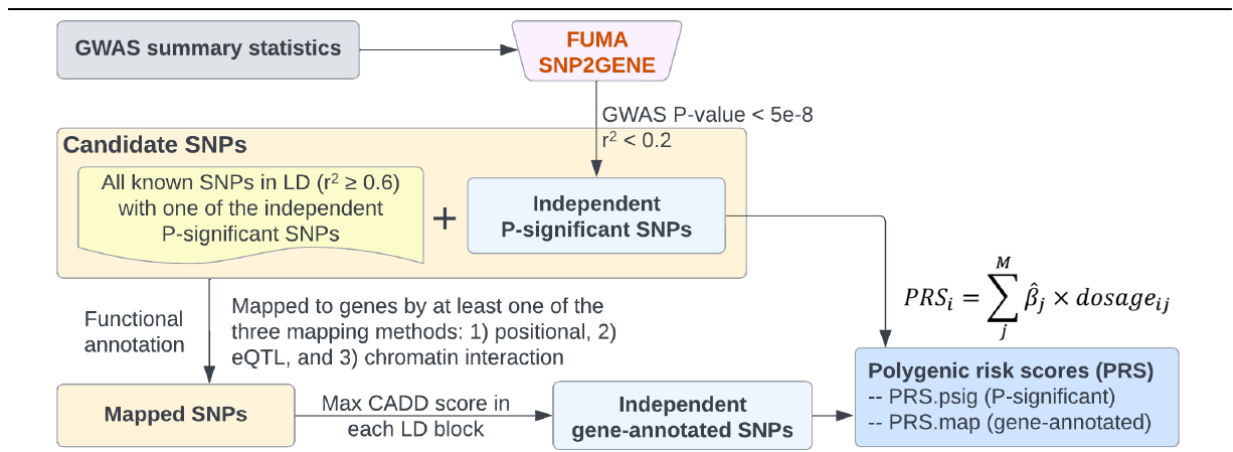
4.10 Supplementary materials



Supplementary Figure 1. Bootstrapping results on model performance of APOE-e4 count, polygenic risk score, and machine learning SNP models in dementia genetic prediction, UCLA ATLAS sample, stratified by genetic inferred ancestry group. All models (if not other specified) have regressed out age, sex, and ancestry-specific principal components. Abbreviations: AD, Alzheimer's Disease; AUROC, Area Under the ROC Curve; AUPRC, Area Under the Precision-Recall Curve; EUR, European; PRS, Polygenic Risk Score; SNP, Single-Nucleotide Polymorphism.

Supplementary Figure 1. Bootstrapping results on model performance of APOE-e4 count, polygenic risk score, and machine learning SNP models in dementia genetic prediction, UCLA ATLAS sample, stratified by genetic inferred ancestry group.

All models (if not other specified) have regressed out age, sex, and ancestry-specific principal components. Abbreviations: AD, Alzheimer's Disease; AUROC, Area Under the ROC Curve; AUPRC, Area Under the Precision-Recall Curve; EUR, European; PRS, Polygenic Risk Score; SP, Single-Nucleotide Polymorphism.)



Supplementary Figure 2. Workflow of candidate SNP selection. Functional annotation and prioritization of SNPs using FUMA. Two distinct sets of SNPs (independent genome-wide-significant SNPs and independent gene-annotated SNPs) were identified by FUMA and subsequently used in our polygenic risk scores building and modeling steps. Abbreviations: CADD, Combined Annotation Dependent Depletion; GWAS, genome-wide association study; LD, linkage disequilibrium; PRS, polygenic risk score; SNP, Single-Nucleotide Polymorphism.

Supplementary Figure 2. Workflow of candidate SNP selection.

Functional annotation and prioritization of SNPs using FUMA. Two distinct sets of SNPs (independent genome-wide-significant SNPs and independent gene-annotated SNPs) were identified by FUMA and subsequently used in our polygenic risk scores building and modeling steps. Abbreviations: CADD, Combined Annotation Dependent Depletion; GWAS, genome-wide association study; LD, linkage disequilibrium; PRS, polygenic risk score; SNP, Single-Nucleotide Polymorphism.)

Supplementary Table 1. Performance of models using lead SNPs (r^2 cut-off <0.1 for defining independent genome-wide-significant SNPs) in dementia genetic prediction, UCLA ATLAS sample, stratified by genetic

Hispanic Latino American ancestry sample (N = 610)							
	Overall performance		Threshold maximize absolute MCC				
	AUPRC	AUROC	F1 score	Accuracy	Precision	Recall	Specificity
AD EUR PRS	0.284	0.594	0.311	0.761	0.384	0.262	0.890
AD AFR PRS	0.314	0.651	0.418	0.716	0.363	0.492	0.775
AD multi-ancestry PRS	0.304	0.623	0.375	0.693	0.324	0.444	0.758
PRSs using AD GWASs only^b	0.308	0.640	0.419	0.705	0.353	0.516	0.754
PRSs using AD + Neuro GWASs^c	0.298	0.624	0.391	0.703	0.339	0.460	0.767
SNPs from AD GWASs only	0.343	0.684	0.431	0.602	0.306	0.730	0.568
SNPs from AD + Neuro GWASs	0.357	0.695	0.435	0.549	0.294	0.841	0.473
African American ancestry sample (N = 440)							
	Overall performance		Threshold maximize absolute MCC				
	AUPRC	AUROC	F1 score	Accuracy	Precision	Recall	Specificity
AD EUR PRS	0.206	0.549	0.322	0.627	0.247	0.464	0.666
AD AFR PRS	0.254	0.590	0.342	0.711	0.303	0.393	0.787
AD multi-ancestry PRS	0.264	0.604	0.368	0.516	0.245	0.738	0.463
PRSs using AD GWASs only^b	0.283	0.591	0.250	0.809	0.500	0.167	0.961
PRSs using AD + Neuro GWASs^c	0.240	0.572	0.343	0.295	0.209	0.964	0.138
SNPs from AD GWASs only	0.281	0.647	0.409	0.691	0.322	0.560	0.722
SNPs from AD + Neuro GWASs	0.327	0.683	0.432	0.755	0.387	0.488	0.817

Abbreviations: AD, Alzheimer's Disease; AFR, African American; AUROC, Area Under the ROC Curve; AUPRC, Area Under the Precision-Recall Curve; EUR, European; GWAS, Genome-Wide Association Study; MCC, Matthews Correlation Coefficient; PRS, Polygenic Risk Score; SNP, Single-Nucleotide Polymorphism.

Notes:

[a] All models (if not other specified) adjusted for age, sex, and first four ancestry-specific principal components.

[b] All AD PRSs built with EUR, AFR, and multi-ancestry GWASs using P-significant/gene-annotated SNPs were included in the model at the same time.

[c] All AD PRSs built with EUR, AFR, and multi-ancestry GWASs, and neurodegenerative disease PRS (Parkinson's disease, progressive supranuclear palsy, Lewy body dementia, and stroke) using P-significant/gene-annotated SNPs were included in the model at the same time.

Supplementary Table 1. Performance of models using lead SNPs (r cut-off ≥ 0.1 for defining independent genome-wide-significant SNPs) in dementia genetic prediction, UCLA ATLAS sample, stratified by genetic

Supplementary Table 2. Full featured risk SNPs from the best-performing Elastic Net SNP model, UCLA ATLAS sample, stratified by genetic ancestry

No.	rsID	CHR	POS	Variable Importance	Nearest Gene	Distance	Function	CADD	Position Map	eQTL Map	CI Map	AD EUR	AD AFR	AD trans	LBD	PD	PSP	STROKE
Hispanic Latino American ancestry (HLA)																		
1	rs429358	19	44908684	0.089 (0.02, 0.11)	<i>APOE</i>	0	exonic	12.64	1	0	1		x					
2	rs2075650	19	44892362	0.089 (0.02, 0.11)	<i>TOMM40</i>	0	intronic	7.974	1	0	1		x	x	x			
3	rs483082	19	44912921	0.07 (0.017, 0.086)	<i>APOC1</i>	1325	intergenic	11.01	1	0	1		x	x				
4	rs157581	19	44892457	0.065 (0.014, 0.079)	<i>TOMM40</i>	0	exonic	21.1	1	0	1		x		x			
5	rs412776	19	44876259	0.057 (0.016, 0.066)	<i>NECTIN2</i>	0	intronic	12.66	1	0	0	x		x				
6	rs62120578	19	44713297	0.053 (0.022, 0.061)	<i>CTB-171A8.1</i>	0	ncRNA_intronic	12.82	1	0	0	x						
7	rs4803765	19	44855191	0.048 (0.015, 0.056)	<i>NECTIN2</i>	0	intronic	1.401	1	0	0	x						
8	rs80100206	4	705856	0.042 (0.02, 0.046)	<i>PCGF3</i>	0	UTR5	19.73	1	1	0					x		
9	rs6857	19	44888997	0.038 (0.011, 0.042)	<i>NECTIN2</i>	0	ncRNA_intronic	5.597	1	0	1		x					
10	rs2276412	11	121590137	0.037 (0.018, 0.039)	<i>SORL1</i>	0	exonic	14.94	1	0	0	x						
11	rs2220427	4	110793733	0.033 (0.019, 0.035)	<i>RP11-777N19.1</i>	0:00	upstream	20.4	1	0	0						x	
12	rs117783785	19	45195929	0.03 (0.016, 0.032)	<i>BLOC1S3</i>	0	intronic	6.441	1	0	0	x						
13	rs7026	19	44821259	0.027 (0.013, 0.029)	<i>BCAM</i>	0	UTR3	11.62	1	0	0	x			x			
14	rs80335384	17	45321252	0.027 (0.017, 0.029)	<i>MAP3K14</i>	4203	intergenic	9.95	1	1	1							x
15	rs2627641	19	45205500	0.026 (0.014, 0.028)	<i>BLOC1S3</i>	0:00:00	intronic	11.65	1	0	0	x						
16	rs435380	19	44903861	0.025 (0.014, 0.028)	<i>TOMM40</i>	171	downstream	9.048	1	1	0		x	x				
17	rs13067212	3	39404095	0.025 (0.018, 0.028)	<i>RPSA</i>	2593	intergenic	13.8	1	1	0						x	
18	rs1551890	19	44829875	0.025 (0.01, 0.028)	<i>BCAM</i>	8458	intergenic	8.368	1	0	0	x		x				
19	rs10422350	19	44725238	0.023 (0.016, 0.026)	<i>snoZ6</i>	745	downstream	13.12	1	0	0	x		x				
20	rs7613	17	45394115	0.02 (0.016, 0.024)	<i>ARHGAP27</i>	0	UTR3	15.45	1	1	1							x
21	rs56870644	19	11065610	0.019 (0.016, 0.023)	<i>SMARCA4</i>	214	downstream	11.62	1	0	0					x		
22	rs76692773	19	44890954	0.019 (0.013, 0.023)	<i>TOMM40</i>	0:00	intronic	7.614	1	0	1	x		x	x			
23	rs35929607	2	168179226	0.012 (0.007, 0.019)	<i>STK39</i>	0	intronic	16.92	1	0	1					x		
24	rs17030651	1	112696760	0.012 (0.006, 0.018)	<i>MOV10</i>	0	exonic	19.29	1	1	0						x	
25	rs11223641	11	133950127	0.011 (0.004, 0.018)	<i>IGSF9B</i>	0	intronic	10.74	1	0	1					x		
26	rs10401176	19	44750234	0.009 (0.002, 0.018)	<i>BCL3</i>	0	intronic	13.68	1	0	0	x		x				
27	rs2619363	4	89837896	0.006 (0.003, 0.014)	<i>SNCA</i>	0	ncRNA_intronic	21.6	1	1	0				x	x		
28	rs28399637	19	44820881	0.004 (0, 0.012)	<i>BCAM</i>	0	intronic	0.114	1	0	0	x		x	x			
African American ancestry (AA)																		
1	rs2627641	19	45205500	0.096 (0.077, 0.099)	<i>BLOC1S3</i>	0:00	intronic	11.65	1	0	0	x						
2	rs8073976	17	44955857	0.079 (0.065, 0.082)	<i>C1QL1</i>	3835	intergenic	16.04	1	1	1						x	
3	rs77283277	7	143386852	0.076 (0.063, 0.079)	<i>ZFX</i>	0	intronic	5.616	1	1	1	x						
4	rs429358	19	44908684	0.071 (0.059, 0.074)	<i>APOE</i>	0	exonic	12.64	1	0	1		x					
5	rs2075650	19	44892362	0.068 (0.056, 0.071)	<i>TOMM40</i>	0	intronic	7.974	1	0	1		x		x			
6	rs73936967	19	44890485	0.064 (0.053, 0.066)	<i>TOMM40</i>	0	ncRNA_intronic	22	1	0	1		x				x	
7	rs13032148	2	127107524	0.063 (0.053, 0.065)	<i>BIN1</i>	168	upstream	7.853	1	0	0	x		x				
8	rs71352239	19	44926286	0.056 (0.047, 0.057)	<i>APOC1P1</i>	517	upstream	5.237	1	1	1	x		x	x			
9	rs435380	19	44903861	0.041 (0.037, 0.042)	<i>TOMM40</i>	171	downstream	9.048	1	1	0		x	x				
10	rs11223641	11	133950127	0.041 (0.037, 0.041)	<i>IGSF9B</i>	0	intronic	10.74	1	0	1					x		
11	rs2238682	19	44955336	0.036 (0.033, 0.036)	<i>CLPTM1</i>	0	UTR5	17.4	1	1	1	x		x				
12	rs73572039	19	44839373	0.034 (0.032, 0.034)	<i>NECTIN2</i>	6801	intergenic	4.628	1	0	0			x				
13	rs3800298	6	43284291	0.033 (0.032, 0.034)	<i>TTBK1</i>	0	exonic	15.75	1	0	0							x
14	rs28399635	19	44819986	0.028 (0.027, 0.028)	<i>BCAM</i>	0	UTR3	4.212	1	0	0	x		x				
15	rs4803791	19	45020325	0.023 (0.022, 0.024)	<i>RELB</i>	0	intronic	7.817	1	1	1	x		x				
16	rs41289510	19	44844379	0.022 (0.021, 0.023)	<i>NECTIN2</i>	1795	intergenic	3.223	1	0	0	x		x				
17	rs2889414	19	44794671	0.02 (0.018, 0.022)	<i>CBL3</i>	0:00	intronic	3.046	1	1	0	x		x				
18	rs75178253	19	44957750	0.02 (0.018, 0.022)	<i>CLPTM1</i>	0	intronic	0.32	1	0	1	x		x				
19	rs3810143	19	44846145	0.018 (0.016, 0.02)	<i>NECTIN2</i>	29	upstream	8.297	1	0	0	x		x				

Supplementary Table 2. Full featured risk SNPs from the best-performing Elastic Net SNP model, UCLA ATLAS sample, stratified by genetic ancestry

20	rs4263041	19	44935386	0.014 (0.013, 0.016)	<i>APOC1P1</i>	3999	intergenic	2.278	1	0	1	x						
21	rs76692773	19	44890954	0.012 (0.012, 0.015)	<i>TOMM40</i>	0	intronic	7.614	1	0	1	x			x			
22	rs166907	19	44883598	0.01 (0.009, 0.013)	<i>NECTIN2</i>	0	ncRNA_intronic	3.815	1	0	0		x	x				
23	rs62330386	4	154525388	0.009 (0.009, 0.014)	<i>RP11-158C21.2</i>	580	downstream	5.44	1	0	0						x	
24	rs12981508	19	44855243	0.009 (0.008, 0.013)	<i>NECTIN2</i>	0	intronic	1.322	1	0	0	x		x				
25	rs13067212	3	39404095	0.008 (0.007, 0.014)	<i>RPSA</i>	2593	intergenic	13.8	1	1	0						x	
26	rs71352248	19	44999245	0.008 (0.007, 0.013)	<i>RELB</i>	2184	intergenic	15.03	1	0	0	x						
27	rs76738189	19	45216235	0.007 (0.007, 0.013)	<i>EXOC3L2</i>	0	intronic	3.762	1	0	0	x						
28	rs157581	19	44892457	0.007 (0.006, 0.011)	<i>TOMM40</i>	0	exonic	21.1	1	0	1		x		x			
29	rs3852860	19	44879709	0.007 (0.006, 0.012)	<i>NECTIN2</i>	0	intronic	4.255	1	0	0	x			x			
30	rs73936968	19	44892559	0.004 (0.003, 0.011)	<i>TOMM40</i>	0	intronic	8.817	1	0	1	x						
31	rs2304933	11	60335034	0.002 (0.001, 0.009)	<i>MS4A6E</i>	0	exonic	17.97	1	0	0	x						

Abbreviations: AD, Alzheimer's Disease; AFR, African American; CI, confidence interval; EUR, European; LBD, Lewy body dementia; PD, Parkinson's disease; PRS, Polygenic Risk Score; PSP, progressive supranuclear palsy; SNP, Single-Nucleotide Polymorphism. Note: SNPs marked in red are overlapped SNPs identified by both samples.

Supplementary Table 2. Full featured risk SNPs from the best-performing Elastic Net SNP model, UCLA ATLAS sample, stratified by genetic ancestry (cont)

Supplementary Table 3. Mapped genes of selected risk SNPs from the best-performing Elastic Net SNP model, UCLA ATLAS sample, by genetic inferred ancestry

Hispanic Latino Americans (HLA)									
No	Gene	Symbol	CHR	Start	End	Type	posMap	eqtlMap (Direction)	ciMap
1	ENSG00000134245	WNT2B	1	113009163	113072787	protein_coding	No	Yes (+)	No
2	ENSG00000007341	ST7L	1	113066140	113163447	protein_coding	No	Yes (+)	No
3	ENSG00000155363	MOV10	1	113215763	113243368	protein_coding	Yes	No	No
4	ENSG00000198648	STK39	2	168810530	169104651	protein_coding	Yes	No	Yes
5	ENSG00000144659	SLC25A38	3	39424839	39438842	protein_coding	No	Yes (+)	No
6	ENSG00000168028	RPSA	3	39448180	39454033	protein_coding	Yes	Yes (+)	No
7	ENSG00000185619	PCGF3	4	699537	764428	protein_coding	Yes	Yes (-)	No
8	ENSG00000247775	RP11-67M1.1	4	90757559	90763129	antisense	No	Yes (-)	No
9	ENSG00000247775	RP11-67M1.1	4	90757559	90763129	antisense	No	Yes (+)	No
10	ENSG00000249519	RP11-777N19.1	4	111715559	111718500	lincRNA	Yes	No	No
11	ENSG00000137642	SORL1	11	121322912	121504402	protein_coding	Yes	No	No
12	ENSG00000080854	IGSF9B	11	133778459	133826880	protein_coding	Yes	No	No
13	ENSG00000255406	RP11-713P17.5	11	133896882	133898013	lincRNA	No	No	Yes
14	ENSG00000166086	JAM3	11	133938820	134021896	protein_coding	No	No	Yes
15	ENSG00000254481	PTP4A2P2	11	133993723	133994224	pseudogene	No	No	Yes
16	ENSG00000180329	CCDC43	17	42750437	42767147	protein_coding	No	No	Yes
17	ENSG00000184922	FMNL1	17	43298811	43324687	protein_coding	No	Yes (+)	No
18	ENSG00000184361	SPATA32	17	43331760	43339479	protein_coding	No	No	Yes
19	ENSG00000006062	MAP3K14	17	43340488	43394414	processed_transcript	Yes	No	Yes
20	ENSG00000267446	CTB-39G8.2	17	43448768	43449423	lincRNA	No	No	Yes
21	ENSG00000159314	ARHGAP27	17	43471275	43511787	protein_coding	Yes	Yes (+)	No
22	ENSG00000159314	ARHGAP27	17	43471275	43511787	protein_coding	Yes	Yes (-)	No
23	ENSG00000267344	CTB-39G8.3	17	43474298	43474843	antisense	No	No	Yes
24	ENSG00000225190	PLEKHM1	17	43513266	43568115	protein_coding	No	Yes (+)	No
25	ENSG00000214425	LRRC37A4P	17	43578685	43627701	pseudogene	No	Yes (+)	No
26	ENSG00000264070	DND1P1	17	43663237	43664295	pseudogene	No	Yes (-)	No
27	ENSG00000263503	RP11-707O23.5	17	43678235	43679706	pseudogene	No	Yes (-)	No
28	ENSG00000204650	CRHR1-IT1	17	43697694	43725582	pseudogene	No	Yes (-)	No
29	ENSG00000185294	SPPL2C	17	43922256	43924438	protein_coding	No	Yes (-)	No
30	ENSG00000214401	KANSL1-AS1	17	44270942	44274089	antisense	No	Yes (-)	No
31	ENSG00000262500	RP11-259G18.2	17	44320972	44322410	pseudogene	No	Yes (-)	No
32	ENSG00000262539	RP11-259G18.3	17	44336917	44337972	pseudogene	No	Yes (-)	No
33	ENSG00000261575	RP11-259G18.1	17	44344403	44346060	pseudogene	No	Yes (-)	Yes
34	ENSG00000127616	SMARCA4	19	11071598	11176071	protein_coding	Yes	No	No
35	ENSG00000266903	CTB-171A8.1	19	45135500	45222031	antisense	Yes	No	No
36	ENSG00000252200	snoZ6	19	45229248	45229322	snoRNA	Yes	No	No
37	ENSG00000069399	BCL3	19	45250962	45263301	protein_coding	Yes	No	No
38	ENSG00000187244	BCAM	19	45312328	45324673	protein_coding	Yes	No	No
39	ENSG00000130202	PVRL2	19	45349432	45392485	protein_coding	Yes	No	No
40	ENSG00000130204	TOMM40	19	45393826	45406946	protein_coding	Yes	No	No
41	ENSG00000130203	APOE	19	45409011	45412650	protein_coding	Yes	Yes (-)	No
42	ENSG00000130208	APOC1	19	45417504	45422606	protein_coding	Yes	No	No
43	ENSG00000214855	APOC1P1	19	45430061	45434643	pseudogene	No	No	Yes
44	ENSG00000104859	CLASRP	19	45542298	45574214	protein_coding	No	No	Yes
African Americans (AA)									
No	Gene	Symbol	CHR	Start	End	Type	posMap	eqtlMap (Direction)	ciMap
1	ENSG00000136717	BIN1	2	127805603	127864931	protein_coding	Yes	No	No
2	ENSG00000144659	SLC25A38	3	39424839	39438842	protein_coding	No	Yes (+)	No
3	ENSG00000168028	RPSA	3	39448180	39454033	protein_coding	Yes	Yes (+)	No
4	ENSG00000250609	RP11-158C21.2	4	155444639	155445959	pseudogene	Yes	No	No
5	ENSG00000146216	TTBK1	6	43211418	43255997	protein_coding	Yes	No	No
6	ENSG00000159840	ZYX	7	143078173	143088204	protein_coding	Yes	No	No
7	ENSG00000050327	ARHGEF5	7	144052381	144077725	protein_coding	No	Yes (+)	No
8	ENSG00000166926	MSA4E	11	60102304	60164069	protein_coding	Yes	No	No
9	ENSG00000080854	IGSF9B	11	133778459	133826880	protein_coding	Yes	No	No
10	ENSG00000255406	RP11-713P17.5	11	133896882	133898013	lincRNA	No	No	Yes
11	ENSG00000166086	JAM3	11	133938820	134021896	protein_coding	No	No	Yes
12	ENSG00000254481	PTP4A2P2	11	133993723	133994224	pseudogene	No	No	Yes
13	ENSG00000180329	CCDC43	17	42750437	42767147	protein_coding	No	No	Yes
14	ENSG00000161692	DBF4B	17	42785976	42829632	protein_coding	No	Yes (-)	No
15	ENSG00000073670	ADAM11	17	42836399	42859214	protein_coding	No	No	Yes
16	ENSG00000131095	GFAP	17	42982376	42994305	protein_coding	No	No	Yes
17	ENSG00000131094	C1QL1	17	43037061	43045439	protein_coding	Yes	Yes (+)	No
18	ENSG00000267788	CTD-2534I21.9	17	43059882	43060140	lincRNA	No	No	Yes
19	ENSG00000266903	CTB-171A8.1	19	45135500	45222031	antisense	No	Yes (+)	No
20	ENSG00000142273	CBLC	19	45281126	45303891	protein_coding	Yes	No	No
21	ENSG00000187244	BCAM	19	45312328	45324673	protein_coding	Yes	No	No
22	ENSG00000130202	PVRL2	19	45349432	45392485	protein_coding	Yes	No	No
23	ENSG00000267282	CTB-129P6.4	19	45385284	45394133	antisense	Yes	No	No
24	ENSG00000130204	TOMM40	19	45393826	45406946	protein_coding	Yes	No	No
25	ENSG00000130203	APOE	19	45409011	45412650	protein_coding	Yes	Yes (-)	No
26	ENSG00000130208	APOC1	19	45417504	45422606	protein_coding	No	No	Yes
27	ENSG00000214855	APOC1P1	19	45430061	45434643	pseudogene	Yes	Yes (+)	Yes
28	ENSG00000234906	APOC2	19	45449243	45452822	protein_coding	No	Yes (+)	No
29	ENSG00000104853	CLPTM1	19	45457842	45490599	protein_coding	Yes	No	No
30	ENSG00000104856	RELB	19	45504688	45541452	protein_coding	Yes	No	No
31	ENSG00000104859	CLASRP	19	45542298	45574214	protein_coding	No	No	Yes
32	ENSG00000267045	AC006126.4	19	45741890	45748628	lincRNA	No	No	Yes

Supplementary Table 3. Mapped genes of selected risk SNPs from the best-performing Elastic Net SNP model, UCLA ATLAS sample, by genetic inferred ancestry

Supplementary Table 4A. Descriptive statistics of demographic and electronic health record features by case/control groups, UCLA ATLAS sample (East Asian American ancestry group, N = 673)

	Cases	Controls	P value
N	75	598	-
Age	80.1 (73.2, 83.5)	76.3 (72.9, 80.7)	0.04*
Sex (Female)	48 (64%)	325 (54%)	0.11
Span of records (in yrs)	5.3 (2.7, 7.8)	9.8 (8.0, 12.0)	<0.001*
Encounters per year	17 (9, 28)	12 (7, 18)	0.002*
Number of encounters	73 (23, 137)	121 (68, 185)	<0.001*
Number of unique diagnosis	64 (35, 92)	60 (38, 90)	0.80

Abbreviations: EHR, electronic health record.

Notes: Continuous variables were reported as median (IQR), and categorical variables were reported as n (%). P-values were calculated based on Wilcoxon rank sum test or Pearson's Chi-squared test as appropriate. * Statistical significant at level 0.05.

Supplementary Table 4A. Descriptive statistics of demographic and electronic health record features by case/control groups, UCLA ATLAS sample (East Asian American ancestry group, N = 673)

Supplementary Table 4B. Model performance of *APOE-e4* count, polygenic risk score, and Elastic Net SNP models in dementia genetic prediction, UCLA ATLAS sample (East Asian American ancestry group, N = 673)^a

		AUPRC	AUROC	F1 score	Accuracy	Precision	Recall	Specificity
APOE	e4 count	0.238	0.712	0.347	0.782	0.260	0.520	0.814
AD-PRS models								
AD EUR PRS	P-significant	0.222	0.681	0.252	0.877	0.389	0.187	0.963
	Gene-annotated	0.215	0.676	0.262	0.866	0.340	0.213	0.948
AD AFR PRS	P-significant	0.234	0.706	0.336	0.765	0.245	0.533	0.794
	Gene-annotated	0.230	0.700	0.331	0.765	0.242	0.520	0.796
AD multi-ancestry PRS	P-significant	0.236	0.702	0.316	0.730	0.220	0.560	0.751
	Gene-annotated	0.232	0.696	0.318	0.796	0.254	0.427	0.843
Multi-PRS models								
PRSs using AD GWASs only^b	P-significant	0.234	0.699	0.327	0.731	0.227	0.587	0.749
	Gene-annotated	0.229	0.700	0.338	0.779	0.253	0.507	0.813
PRSs using AD + Neuro GWASs^c	P-significant	0.234	0.704	0.338	0.796	0.265	0.467	0.838
	Gene-annotated	0.224	0.706	0.346	0.770	0.253	0.547	0.798
Elastic Net SNPs models								
SNPs from AD GWASs only	P-significant	0.234	0.706	0.350	0.774	0.258	0.547	0.803
	Gene-annotated	0.256	0.743	0.338	0.685	0.220	0.720	0.681
SNPs from AD + Neuro GWASs	P-significant	0.271	0.748	0.366	0.835	0.320	0.427	0.886
	Gene-annotated	0.338	0.793	0.443	0.840	0.361	0.573	0.873
Non-linear SNPs models								
SNPs from AD + Neuro GWASs	GBM	0.113	0.511	0.219	0.416	0.129	0.733	0.376
Gene-annotated SNPs	XGBoost	0.206	0.673	0.300	0.660	0.194	0.653	0.661

Abbreviations: AD, Alzheimer's Disease; APOE, apolipoprotein E; AUROC, Area Under the ROC Curve; AUPRC, Area Under the Precision-Recall Curve; EUR, European; GBM, Gradient Boosting Machine; GWAS, Genome-Wide Association Study; PRS, Polygenic Risk Score; SNP, Single-Nucleotide Polymorphism.

Notes:

[a] All models (if not other specified) have regressed out age, sex, and ancestry-specific principal components. Thresholds were determined by maximizing absolute Matthews correlation coefficient.

[b] All AD PRSs built with EUR, AFR, and multi-ancestry GWASs using P-significant/gene-annotated SNPs were included in the model at the same time.

[c] All AD PRSs built with EUR, AFR, and multi-ancestry GWASs, and neurodegenerative disease PRS (Parkinson's disease, progressive supranuclear palsy, Lewy body dementia, and stroke) using P-significant/gene-annotated SNPs were included in the model at the same time.

Supplementary Table 4B. Model performance of *APOE-e4* count, polygenic risk score, and Elastic Net SNP models in dementia genetic prediction, UCLA ATLAS sample (East Asian American ancestry group, N = 673)

Supplementary Table 4C. Selected risk SNPs from the best-performing Elastic Net SNP model, UCLA ATLAS sample (East Asian American ancestry group, N = 673)

rsID	CHR	POS	Variable Importance (percentage, 95% CI)	Nearest Gene	AD EUR	AD AFR	AD multi	LBD	PD	PSP	Stroke
rs429358	19	44908684	0.205 (0.151, 0.214)	<i>APOE</i>		x					
rs66626994	19	44924977	0.133 (0.104, 0.137)	<i>APOC1P1</i>			x	x			
rs6857	19	44888997	0.118 (0.095, 0.122)	<i>NECTIN2</i>		x					
rs2075650	19	44892362	0.101 (0.084, 0.104)	<i>TOMM40</i>		x	x	x			
rs483082	19	44912921	0.092 (0.078, 0.094)	<i>APOC1</i>		x	x				
rs35106910	19	44781009	0.079 (0.07, 0.081)	<i>CBLC</i>	x						
rs1105568	17	45716061	0.05 (0.049, 0.051)	<i>CRHR1</i>						x	
rs76738189	19	45216235	0.049 (0.049, 0.051)	<i>EXOC3L2</i>	x						
rs34096562	8	16843772	0.047 (0.046, 0.049)	<i>RP11-13N12.1</i>					x		
rs7932740	11	60208638	0.029 (0.028, 0.038)	<i>MS4A4E</i>	x		x				
rs11724804	4	971991	0.021 (0.019, 0.032)	<i>DGKQ</i>				x	x		
rs13067212	3	39404095	0.019 (0.017, 0.031)	<i>RPSA</i>						x	
rs7613	17	45394115	0.018 (0.016, 0.03)	<i>ARHGAP27</i>						x	
chr1:207368589:D	1	207195244	0.015 (0.012, 0.028)	<i>RP11-164O23.7</i>	x						
rs59193782	17	45357199	0.01 (0.008, 0.025)	<i>CTB-39G8.2</i>						x	
rs10769263	11	47395632	0.007 (0.004, 0.023)	<i>RP11-750H9.5</i>	x						

Abbreviations: AD, Alzheimer's Disease; AFR, African American; CI, confidence interval; EUR, European; LBD, Lewy body dementia; PD, Parkinson's disease; PRS, Polygenic Risk Score; PSP, progressive supranuclear palsy; SNP, Single-Nucleotide Polymorphism.

Supplementary Table 4C. Selected risk SNPs from the best-performing Elastic Net SNP model, UCLA ATLAS sample (East Asian American ancestry group, N = 673)

Supplementary Table 4D. Mapped genes of selected risk SNPs from the best-performing Elastic Net SNP model, UCLA ATLAS sample (East Asian American ancestry group, N = 673)

No	Gene	Symbol	CHR	Start	End	Type	posMap	eqtlMap (Direction)	ciMap
1	ENSG00000203710	<i>CR1</i>	1	207669492	207813992	protein_coding	No	Yes (+)	No
2	ENSG00000127419	<i>TMEM175</i>	4	926175	952444	protein_coding	No	Yes (-)	No
3	ENSG00000145214	<i>DGKQ</i>	4	952675	980683	protein_coding	Yes	Yes (-)	No
4	ENSG00000145217	<i>SLC26A1</i>	4	972861	987228	protein_coding	No	Yes (+)	No
5	ENSG00000253496	<i>RP11-13N12.1</i>	8	16534414	16772553	lincRNA	Yes	No	No
6	ENSG00000175220	<i>ARHGAP1</i>	11	46698630	46722165	protein_coding	No	Yes (+)	No
7	ENSG00000255197	<i>RP11-750H9.5</i>	11	47404699	47430741	antisense	Yes	Yes (-)	No
8	ENSG00000165915	<i>SLC39A13</i>	11	47428683	47438047	protein_coding	No	Yes (-)	No
9	ENSG00000213619	<i>NDUFS3</i>	11	47586888	47606114	protein_coding	No	Yes (-)	No
10	ENSG00000196666	<i>FAM180B</i>	11	47608198	47610746	protein_coding	No	Yes (-)	No
11	ENSG00000184922	<i>FMNL1</i>	17	43298811	43324687	protein_coding	No	Yes (+)	No
12	ENSG00000233175	<i>CTD-2020K17.3</i>	17	43315395	43319101	antisense	No	Yes (+)	No
13	ENSG00000006062	<i>MAP3K14</i>	17	43340488	43394414	processed_transcript	No	No	Yes
14	ENSG00000199953	<i>RNA5SP443</i>	17	43404732	43404863	rRNA	Yes	No	No
15	ENSG00000267446	<i>CTB-39G8.2</i>	17	43448768	43449423	lincRNA	Yes	No	No
16	ENSG00000159314	<i>ARHGAP27</i>	17	43471275	43511787	protein_coding	Yes	Yes (-)	No
17	ENSG00000214425	<i>LRRC37A4P</i>	17	43578685	43627701	pseudogene	No	Yes (+)	No
18	ENSG00000263503	<i>RP11-707O23.5</i>	17	43678235	43679706	pseudogene	No	Yes (-)	No
19	ENSG00000204650	<i>CRHR1-IT1</i>	17	43697694	43725582	pseudogene	No	Yes (-)	No
20	ENSG00000185294	<i>SPPL2C</i>	17	43922256	43924438	protein_coding	No	Yes (-)	No
21	ENSG00000262500	<i>RP11-259G18.2</i>	17	44320972	44322410	pseudogene	No	Yes (-)	No
22	ENSG00000261575	<i>RP11-259G18.1</i>	17	44344403	44346060	pseudogene	No	Yes (-)	No
23	ENSG00000142273	<i>CBLC</i>	19	45281126	45303891	protein_coding	Yes	No	No
24	ENSG00000130204	<i>TOMM40</i>	19	45393826	45406946	protein_coding	Yes	No	No
25	ENSG00000130203	<i>APOE</i>	19	45409011	45412650	protein_coding	Yes	No	No
26	ENSG00000130208	<i>APOC1</i>	19	45417504	45422606	protein_coding	Yes	No	No
27	ENSG00000214855	<i>APOC1P1</i>	19	45430061	45434643	pseudogene	Yes	No	Yes
28	ENSG00000104859	<i>CLASRP</i>	19	45542298	45574214	protein_coding	No	No	Yes

Supplementary Table 4D. Mapped genes of selected risk SNPs from the best-performing Elastic Net SNP model, UCLA ATLAS sample (East Asian American ancestry group, N = 673)

Supplementary Table 5. Descriptive statistics of demographic and electronic health record features, by data sources (All of Us vs. UCLA ATLAS, stratified by the genetic inferred group)

Characteristic	Hispanic Latino Americans			African Americans		
	All of Us	UCLA ATLAS	p-value	All of Us	UCLA ATLAS	p-value
N	458	610	-	645	440	-
Age	75.0 (72.4, 79.0)	75.8 (72.5, 80.5)	0.1	70 (67, 75)	75.8 (72.5, 80.5)	<0.001*
Sex (Female)	288 (63%)	372 (61%)	0.5	437 (68%)	264 (60%)	0.01*
Span of records (in yrs)	6.5 (5.8, 7.8)	9.1 (6.8, 10.2)	<0.001*	6.5 (5.8, 6.8)	9.77 (7.27, 11.02)	<0.001*
Encounters per year	9 (5, 14)	14 (8, 22)	<0.001*	10 (5, 16)	13 (8, 22)	<0.001*
Number of encounters	59 (31, 105)	117 (64, 198)	<0.001*	60 (30, 108)	129 (71, 209)	<0.001*
Number of unique diagnosis	46 (28, 68)	71 (45, 109)	<0.001*	45 (28, 67)	71 (45, 103)	<0.001*
Dementia (Yes)	68 (15%)	126 (21%)	0.02*	129 (20%)	84 (19%)	0.7
<i>APOE-e4</i> count			<0.001*			<0.001*
0	425 (93%)	491 (80%)		517 (80%)	284 (65%)	
1	32 (7.0%)	117 (19%)		121 (19%)	137 (31%)	
2	1 (0.2%)	2 (0.3%)		7 (1.1%)	19 (4.3%)	

Notes: Continuous variables were reported as median (IQR), and categorical variables were reported as n (%). P-values were calculated based on Wilcoxon rank sum test or Pearson's Chi-squared test as appropriate. * Statistical significant at level 0.05.

Supplementary Table 5. Descriptive statistics of demographic and electronic health record features, by data sources (All of Us vs. UCLA ATLAS, stratified by the genetic inferred group)

Supplementary Table 6. Descriptive statistics of demographic and electronic health record features by case/control groups, All of Us sample, stratified by genetic ancestry

	Hispanic Latino Americans (N = 458)			African Americans (N = 645)		
	Cases	Controls	P value	Cases	Controls	P value
N	68	390	-	129	516	-
Age	76.1 (70.1, 80.6)	74.9 (72.8, 78.9)	0.5	70.6 (65.8, 77.1)	70.0 (67.7, 74.8)	0.5
Sex (Female)	42 (62%)	246 (63%)	0.8	81 (63%)	356 (69%)	0.20
Span of records (in yrs)	4.5 (2.8, 6.3)	6.6 (6.0, 8.1)	<0.001*	4.0 (2.1, 6.1)	6.6 (6.2, 7.2)	<0.001*
Encounters per year	10 (4, 19)	8 (5, 14)	0.2	15 (7, 23)	9 (5, 15)	<0.001*
Number of encounters	36 (15, 78)	63 (34, 108)	<0.001*	49 (20, 89)	65 (34, 110)	<0.001*
Number of unique diagnosis	52 (27, 69)	46 (28, 68)	0.9	53 (32, 74)	43 (27, 65)	0.01*

Notes: Continuous variables were reported as median (IQR), and categorical variables were reported as n (%). P-values were calculated based on Wilcoxon rank sum test or Pearson's Chi-squared test as appropriate. * Statistical significant at level 0.05.

Supplementary Table 6. Descriptive statistics of demographic and electronic health record features by case/control groups, All of Us sample, stratified by genetic ancestry

Supplementary Table 7. Pearson's correlation of polygenic risk scores that built with genome-wide-significant or gene-annotated SNPs, UCLA ATLAS sample, stratified by genetic inferred

Polygenic risk score GWAS	Pearson's Correlation	
	Hispanic Latino American	African American
Alzheimer's Disease (European ancestry)	0.885	0.837
Alzheimer's Disease (African ancestry)	0.953	0.963
Alzheimer's Disease (multi-ancestry)	0.953	0.919
Lewy Body Dementia	0.714	0.698
Parkinson's Disease	0.561	0.29
Progressive Supranuclear Palsy	0.813	0.657
Stroke	0.737	0.669

Abbreviations: GWAS, genome-wide association study, SNP, Single-Nucleotide Polymorphism.

Supplementary Table 7. Pearson's correlation of polygenic risk scores that built with genome-wide-significant or gene-annotated SNPs, UCLA ATLAS sample, stratified by genetic inferred

Supplementary Table 8. GWAS information and summary results of FUMA by phenotype

Phenotype	GWAS information			# SNPs (overlapped with UCLA and AOU data) ^b		
	Summary statistics	1000G reference ^a	N case/control	Candidate	Independent P-significant	Independent gene-annotated
Alzheimer's Disease	Kunkle et al. (2019)	EUR	21,982/41,944	1744	76	75
	Kunkle et al. (2021)	AFR	2,784/5,222	80	11	11
	Jun et al. (2017)	ALL	15,579/17,690	760	54	54
Parkinson's Disease	Nalls et al. (2019)	EUR	Meta-analysis	2640	27	27
Progressive Supranuclear Palsy	Chen et al. (2018)	EUR	1,646/10,662	732	21	21
Lewy Body Dementia	Chia et al. (2021)	EUR	2,981/4,391	356	9	9
Stroke	Malik et al. (2018)	ALL	67,162/454,450	903	21	20

Abbreviations: AFR, African American; EUR, European; GWAS, genome-wide association study; SNP, Single-Nucleotide Polymorphism.

Notes: [a] Reference panel population used for regional linkage disequilibrium patterns in identifying independent SNPs and lead P-significant SNPs. [b] Numbers of features used in PRSs building, feature selection for modeling, and phenotype prediction in the following steps.

Supplementary Table 8. GWAS information and summary results of FUMA by phenotype

Supplementary Table 9. ICD-10 codes used for dementia phenotype definition

ICD-10 code	Description
F01	Vascular dementia
F01.5	Multi-infarct dementia
F01.50	Multi-infarct dementia, unspecified
F01.51	Multi-infarct dementia with delirium
F01.511	Multi-infarct dementia with delirium in Alzheimer's disease
F01.518	Multi-infarct dementia with delirium, not elsewhere classified
F02.8*	Dementia in other diseases classified elsewhere
F02.80	Dementia in other diseases classified elsewhere, without behavioral disturbance
F02.81	Dementia in other diseases classified elsewhere, with behavioral disturbance
F02.811	Dementia in other diseases classified elsewhere, with behavioral disturbance in Alzheimer's disease
F02.818	Dementia in other diseases classified elsewhere, with behavioral disturbance, not elsewhere classified
F03	Unspecified dementia
F03.9	Unspecified dementia without behavioral disturbance
F03.90	Unspecified dementia with delirium
F03.91	Unspecified dementia with delirium in Alzheimer's disease
F03.911	Unspecified dementia with delirium in Alzheimer's disease, with behavioral disturbance
F03.918	Unspecified dementia with delirium in Alzheimer's disease, not elsewhere classified
G30	Alzheimer's disease
G30.0	Alzheimer's disease with early onset
G30.1	Alzheimer's disease with late onset
G30.8	Other Alzheimer's disease
G30.9	Alzheimer's disease, unspecified
G31	Other degenerative diseases of nervous system, not elsewhere classified
G31.0	Pick's disease
G31.01	Pick's disease with early onset
G31.09	Pick's disease with late onset
G31.1	Frontotemporal dementia
G31.83	Dementia with Lewy bodies
G31.85	Corticobasal degeneration
G23.1	Parkinson's disease with dementia

Supplementary Table 9. ICD-10 codes used for dementia phenotype definition

CHAPTER 5

Summary and Future Directions

5.1 Summary of Contributions

This dissertation presents a comprehensive approach to advancing precision medicine in Alzheimer’s disease research through three interconnected studies that address fundamental challenges in disease understanding, temporal progression modeling, and health equity. Each study contributes unique methodological innovations while collectively building toward a more inclusive and effective framework for dementia research and clinical care.

5.1.1 Defining Disease Relationships Through Semantic Embeddings

Chapter 2 established a foundational framework for quantifying disease relationships using SNOMED CT embeddings, addressing a critical gap in our ability to measure semantic similarity between clinical conditions. This work demonstrated that embedding-based distance metrics significantly outperform traditional approaches in capturing clinically meaningful disease relationships.

The key contributions include being the first to build a full landscape of disease distances in the ICD-10 code format with fine granularity. Our embedding metric overcomes many limitations of the other three metrics. Compared to the ICD-tree-based metric, it has a full distribution (rather than a majority of distances equaling 1) and the ability to capture subclusters within a disease group. Unlike the genetic-correlation-based metric, which is subject to limited genetic studies and selection bias, the embedding metric is able to provide

all possible pairwise comparisons of ICD-10 codes.

The embedding metric enables many large-scale downstream analyses such as disease trajectory prediction and patient clustering by providing a comprehensive pairwise disease distance matrix independent of clinical overlap. All distance metrics developed in this study have been made publicly available for further examination and downstream analyses. Research in many different fields can benefit from this quantitative metric. Together with EHRs and available machine learning tools, it can accelerate the effective characterization of diseases, discover the underlying etiology of disease associations, and enable personalized medical practice in the clinical community.

5.1.2 Identifying Disease Trajectories Leading to Alzheimer’s Disease

Chapter 3 built upon the disease distance framework to identify distinct temporal pathways leading to Alzheimer’s disease, representing a paradigm shift from static risk factor identification to dynamic disease progression modeling. We identified four distinct trajectory clusters leading to AD, each with characteristic demographic features, comorbidity patterns, and progression rates. We demonstrated that multi-step trajectories confer higher AD risk than individual diagnoses, highlighting the importance of sequential disease progression.

Our study identified four distinct trajectory clusters leading to AD: mental health, encephalopathy, MCI/neurodegenerative, and vascular. These clusters exhibited significant differences in demographic characteristics, comorbidity patterns, and progression rates. The mental health cluster, centred on depression, predominantly affected women and Hispanic individuals. The encephalopathy cluster showed the most rapid progression to AD and subsequent death, suggesting a more aggressive disease course. The MCI/neurodegenerative cluster followed the expected progression from mild cognitive impairment to AD, while the vascular cluster had the longest EHR histories and highest comorbidity burden, reflecting the chronic nature of cerebrovascular disease.

The validation of our findings in the All of Us Research Program demonstrates the generalisability of our approach across diverse populations and healthcare settings. The high assignment rate (89.9%) of All of Us patients to our previously identified clusters suggests that these trajectories represent common pathways to AD that extend beyond our initial study population. Furthermore, the confirmation that multi-step trajectories confer higher risk than individual diagnoses underscores the importance of considering sequential disease progression in AD risk assessment.

5.1.3 Advancing Health Equity in Genetic Risk Modeling

Chapter 4 addressed the critical challenge of health disparities in genetic research by developing machine learning approaches for dementia risk modeling in underrepresented populations. This work demonstrated that traditional polygenic risk scores perform poorly in non-European populations, with particularly pronounced limitations in African American populations where PRS showed inferior performance compared to simple APOE-4 counting.

The Elastic Net SNP models achieved substantial performance improvements across all populations studied. The functional genomics integration successfully prioritized meaningful variants over statistical proxies, revealing both shared and population-specific genetic architecture. External validation in the All of Us Research Program confirmed the robustness and generalizability of these improvements across different healthcare systems.

The work demonstrates that functional genomics integration and population-stratified modeling can preserve biological insights while accommodating genetic diversity. This provides a template for inclusive precision medicine approaches that can be applied to other diseases facing similar transferability challenges across populations.

5.2 Methodological Integration

The three studies integrate to create a comprehensive framework addressing multiple dimensions of precision medicine. Chapter 2 provides the semantic foundation necessary for meaningful disease relationship analysis. Chapter 3 builds on this foundation, using disease relationship understanding to enable trajectory analysis that can accommodate irregular timing and varying sequence lengths in real clinical data. Chapter 4 operates in parallel, ensuring that genetic precision medicine benefits reach all populations equitably.

Despite the usefulness of the embedding metric individually, we plan to develop distance metrics in future work that integrates multiple metrics built in this study. This integrated distance metric may further explain known disease relationships, or uncover novel disease relationships and their underlying mechanisms.

5.3 Clinical Implications

Our findings have several clinical implications. The identification of distinct trajectory clusters suggests that AD may develop through multiple pathways, each potentially requiring different prevention and treatment strategies. The recognition that certain diagnoses, such as depression or cerebrovascular disease, often precede AD by several years provides opportunities for early intervention. Furthermore, the finding that multi-step trajectories confer higher risk than individual diagnoses underscores the importance of viewing AD risk holistically rather than focusing solely on single risk factors.

For example, the sequence of acute kidney failure followed by encephalopathy may contribute to increased dementia risk, suggesting that systemic organ dysfunction can hasten neurodegenerative processes. Similarly, the vascular cluster’s pathways, such as hypertension followed by other cerebrovascular disease, mirror previous work on the links between vascular insufficiency and cognitive decline. These results suggest that clinicians should consider the

cumulative impact of sequential disease progression when assessing AD risk.

The genetic risk modeling provides equitable and accurate risk assessment across all populations. Population-specific models provide reliable risk estimates for underrepresented populations that have been historically excluded from precision medicine benefits. The functional genomics integration ensures that risk predictions are based on biologically meaningful variants rather than statistical proxies that may not transfer across populations.

5.4 Limitations

Several limitations must be acknowledged across the three studies. There are limitations to our current three-digit ICD-code embedding metric. For example, we are not able to differentiate F38 (Other mood [affective] disorders) and F39 (Unspecified mood [affective] disorder). A more fine-grained distance metric based on four or more digits may help address the limitation and could be extended for future work. Moreover, the embedding distances were exclusively based on SNOMED CT knowledge graphs and, consequently, largely depended on the accuracy of the information included in those knowledge graphs. With an increasing understanding of disease characteristics and associations, more comprehensive knowledge graphs can be built in the future and, in turn, improve the metric’s performance.

Our study has several limitations regarding the trajectory analysis. First, our dataset automatically censored patients older than 90 years, which may limit the generalizability of our findings to the oldest-old population, a group with high AD prevalence. In addition, our discovery cohort was drawn from six academic health systems in California, which may limit representativeness relative to population-based samples. Second, the potential misclassification of patients with AD in EHR data remains a concern. The trajectories we identified should therefore be interpreted as pathways leading to clinically diagnosed AD rather than necessarily to biological AD confirmed by biomarkers. Third, our decision to focus specifically on AD rather than broadening to dementia represents a methodological

trade-off between specificity and sensitivity.

For the genetic risk modeling, while substantially improved over traditional methods, the approach still requires larger and more diverse cohorts to fully capture the genetic architecture across all populations. The functional annotation approach depends on current understanding of variant function, which continues to evolve.

5.5 Future Research Directions

5.5.1 Methodological Advances

A small fraction of disease associations were identical or similar in different distance metrics while there were also cases of unique patterns revealed from a single distance metric. The embedding distances alone may not be enough to represent disease relationships. Thus, to identify additional patterns between diseases, it may be beneficial to combine different distance metrics [1].

Future studies could develop and validate predictive models that incorporate disease trajectories alongside biomarkers, genetics, and clinical notes. These models, when deployed within an EHR system, could flag patients with specific multi-step risk pathways for cognitive screening, clinical assessment, and early targeted treatments for AD [2].

5.5.2 Clinical Translation

Future research should focus on validating these trajectories in additional cohorts, investigating the biological mechanisms underlying these progression patterns, and developing targeted interventions for specific trajectory clusters [2].

The immediate priority is clinical implementation of the trajectory-based risk assessment framework. This requires development of electronic health record integration tools, clinical decision support systems, and provider training programs. Prospective validation studies are

needed to confirm that trajectory-based interventions improve clinical outcomes compared to standard care approaches.

5.5.3 Health Equity Applications

The health equity framework developed in Study 3 should be extended to other diseases and populations globally. This work provides a template for inclusive precision medicine approaches that can be applied to other diseases [3]. Priority should be given to diseases with significant health disparities and populations that have been historically underrepresented in genomic research.

5.6 Conclusion

In summary, we proposed a novel method to define disease distance in the ICD-10 code format and computed three other distance metrics based on ICD-tree structure, comorbidity, and genetic correlation separately for comparisons. Results show our metric provides a representation of disease relationships that could be exploited to discover important disease patterns [1].

Our study provides a comprehensive framework for identifying and analysing disease trajectories leading to AD using EHR data. By unveiling distinct and interconnected routes to AD, this approach offers insights that may improve risk assessment, timely diagnosis, and targeted interventions [2].

The genetic risk modeling work demonstrates that functional genomics integration and population-stratified modeling can preserve biological insights while accommodating genetic diversity, providing a template for inclusive precision medicine approaches [3].

Together, these three studies establish a foundation for precision medicine in Alzheimer’s disease that is both methodologically innovative and equitable across populations. The in-

tegrated framework enables semantic understanding of disease relationships, temporal modeling of disease progression, and equitable genetic risk assessment. This represents both a scientific advancement and a step toward ensuring that precision medicine benefits reach all populations.

The broader implications of this work extend beyond Alzheimer’s disease to the field of precision medicine more generally, providing methodological innovations that could be applied to other complex diseases and contributing to the goal of developing healthcare approaches that are truly predictive, preventive, personalized, and participatory.

The future directions identified in this work point toward exciting opportunities for continued innovation in precision medicine, from the integration of multi-modal data to the development of adaptive intervention strategies. The ethical considerations and implementation challenges highlighted here emphasize the importance of responsible development and deployment of precision medicine approaches that prioritize equity, privacy, and patient autonomy.

Ultimately, this dissertation contributes to a vision of precision medicine in Alzheimer’s disease that could transform how we understand, predict, prevent, and treat this devastating condition. By addressing methodological limitations, incorporating temporal dynamics, and prioritizing health equity, this work provides a foundation for developing approaches that could benefit all populations and contribute to reducing the global burden of dementia.

The journey toward precision medicine in Alzheimer’s disease is far from complete, but the work presented here provides important stepping stones toward a future where we can identify individuals at risk years before symptoms appear, implement targeted interventions that can delay or prevent disease onset, and ensure that the benefits of these advances are accessible to all populations regardless of ancestry or socioeconomic status. This vision of equitable, effective precision medicine represents both a scientific goal and a moral imperative that should guide future research and implementation efforts in this critical area of healthcare.

Bibliography

- [1] Mingzhou Fu et al. “Defining the distance between diseases using SNOMED CT embeddings”. In: *Journal of Biomedical Informatics* 139 (2023), p. 104307.
- [2] Mingzhou Fu and Timothy S Chang. “Identifying common disease trajectories of Alzheimer’s disease with electronic health records”. In: *eBioMedicine* 111 (2025), p. 105502.
- [3] Mingzhou Fu et al. “Improving genetic risk modeling of dementia from real-world data in underrepresented populations”. In: *Communications Biology* 7.1 (2024), p. 1142.
- [4] C Patterson. “World Alzheimer Report 2022: Life after diagnosis: Navigating treatment, care and support”. In: *Alzheimer’s Disease International* (2022).
- [5] Philip Scheltens et al. “Alzheimer’s disease”. In: *The Lancet* 397.10284 (2021), pp. 1577–1590.
- [6] Sukanya Jongsiriyanyong and Panita Limpawattana. “Comorbidity in Alzheimer’s Disease: A Review of the Literature”. In: *American Journal of Alzheimer’s Disease Other Dementias* 33.8 (2018), pp. 465–472.
- [7] Kala M Mehta and Gene W Yeo. “Systematic review of dementia prevalence and incidence in United States race/ethnic populations”. In: *Alzheimer’s & Dementia* 13.1 (2017), pp. 72–83.
- [8] Jennifer J Manly and Richard Mayeux. “Alzheimer’s disease in African Americans: The role of genetics, environment, and social determinants of health”. In: *Alzheimer’s & Dementia* 18.4 (2022), pp. 739–749.
- [9] Alicia R Martin et al. “Clinical use of current polygenic risk scores may exacerbate health disparities”. In: *Nature Genetics* 51.4 (2019), pp. 584–591.
- [10] Martin R Cowie et al. “Electronic health records to facilitate clinical research”. In: *Clinical research in cardiology* 106.1 (2017), pp. 1–9.

- [11] Peter B Jensen, Lars J Jensen, and Søren Brunak. “Mining electronic health records: towards better research and patient care”. In: *Nature reviews genetics* 13.6 (2012), pp. 395–405.
- [12] Omri Gottesman et al. “The electronic health record for translational research”. In: *Journal of the American Medical Informatics Association* 20.6 (2013), pp. 1053–1063.
- [13] William S Bush et al. “Genetic variation among 82 pharmacogenes: the PGRNseq data from the eMERGE network”. In: *Clinical Pharmacology & Therapeutics* 100.2 (2016), pp. 160–169.
- [14] Alice B Popejoy and Stephanie M Fullerton. “Genomics is failing on diversity”. In: *Nature* 538.7624 (2016), pp. 161–164.
- [15] Joshua C Denny et al. “The ”All of Us” research program”. In: *New England Journal of Medicine* 381.7 (2019), pp. 668–676.
- [16] Nicole G Weiskopf and Chunhua Weng. “Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research”. In: *Journal of the American Medical Informatics Association* 20.1 (2013), pp. 144–151.
- [17] Benjamin Shickel et al. “Deep EHR: a survey of recent advances in deep learning techniques for electronic health record (EHR) analysis”. In: *IEEE journal of biomedical and health informatics* 22.5 (2018), pp. 1589–1604.
- [18] Alvin Rajkomar, Jeffrey Dean, and Isaac Kohane. “Machine learning in medicine”. In: *New England Journal of Medicine* 380.14 (2019), pp. 1347–1358.
- [19] Ziad Obermeyer et al. “Dissecting racial bias in an algorithm used to manage the health of populations”. In: *Science* 366.6464 (2019), pp. 447–453.
- [20] Miguel J Divo, Carlos H Martinez, and David M Mannino. “A review of comorbidity indices and their applications in Alzheimer’s disease”. In: *Journal of Alzheimer’s Disease* 42.s3 (2014), S263–S274.

- [21] Kevin Donnelly. *SNOMED-CT: The advanced terminology and coding system for eHealth*. IOS Press, 2006.
- [22] Brian W Kunkle et al. “Genetic meta-analysis of diagnosed Alzheimer’s disease identifies new risk loci and implicates A β , tau, immunity and lipid processing”. In: *Nature Genetics* 51.3 (2019), pp. 414–430.
- [23] Albert-László Barabási, Natali Gulbahce, and Joseph Loscalzo. “Network medicine: a network-based approach to human disease”. In: *Nature Reviews Genetics* 12.1 (2011), pp. 56–68.
- [24] César A Hidalgo et al. “A dynamic network approach for the study of human phenotypes”. In: *PLoS Computational Biology* 5.4 (2009), e1000353.
- [25] Anders Boeck Jensen et al. “Temporal disease trajectories condensed from population-wide registry data covering 6.2 million patients”. In: *Nature Communications* 5.1 (2014), pp. 1–10.
- [26] Riccardo Miotto et al. “Deep patient: an unsupervised representation to predict the future of patients from the electronic health records”. In: *Scientific Reports* 6.1 (2016), pp. 1–10.
- [27] Alvin Rajkomar et al. “Scalable and accurate deep learning with electronic health records”. In: *npj Digital Medicine* 1.1 (2018), pp. 1–10.
- [28] Juyong Park and Mark EJ Newman. “The statistical mechanics of networks”. In: *Physical Review E* 70.6 (2004), p. 066117.
- [29] Sebastian Köhler et al. “The Human Phenotype Ontology in 2017”. In: *Nucleic Acids Research* 45.D1 (2017), pp. D865–D876.
- [30] David Chang et al. “Benchmark and Best Practices for Biomedical Knowledge Graph Embeddings”. In: *Proceedings of the 19th SIGBioMed Workshop on Biomedical Language Processing*. TLDR: This work trains several state-of-the-art knowledge graph embedding models on the SNOMED-CT knowledge graph, provides a benchmark with

- comparison to existing methods and in-depth discussion on best practices, and makes a case for the importance of leveraging the multi-relational nature of knowledge graphs for learning biomedical knowledge representation. Online: Association for Computational Linguistics, July 2020, pp. 167–176. DOI: [10.18653/v1/2020.bionlp-1.18](https://doi.org/10.18653/v1/2020.bionlp-1.18). URL: <https://aclanthology.org/2020.bionlp-1.18>.
- [31] Quan Wang et al. “Knowledge graph embedding: A survey of approaches and applications”. In: *IEEE Transactions on Knowledge and Data Engineering* 29.12 (2017), pp. 2724–2743.
 - [32] Shaoxiong Ji et al. “A survey on knowledge graphs: representation, acquisition, and applications”. In: *IEEE Transactions on Neural Networks and Learning Systems* 33.2 (2021), pp. 494–514.
 - [33] Zhiqing Sun et al. “RotatE: Knowledge graph embedding by relational rotation in complex space”. In: *arXiv preprint arXiv:1902.10197* (2019).
 - [34] World Health Organization. “International statistical classification of diseases and related health problems”. In: *World Health Organization* 1 (2016).
 - [35] Catia Pesquita et al. “Semantic similarity in biomedical ontologies”. In: *PLoS Computational Biology* 5.7 (2009), e1000443.
 - [36] Lynn Marie Schriml et al. “Disease Ontology: a backbone for disease semantic integration”. In: *Nucleic Acids Research* 40.D1 (2012), pp. D940–D946.
 - [37] Shreyasee Mathur and Deendayal Dinakarpanthian. “Finding disease similarity based on implicit semantic similarity”. In: *Journal of Biomedical Informatics* 45.2 (2012), pp. 363–371.
 - [38] Kwang-Il Goh et al. “The human disease network”. In: *Proceedings of the National Academy of Sciences* 104.21 (2007), pp. 8685–8690.
 - [39] Jörg Menche et al. “Uncovering disease-disease relationships through the incomplete interactome”. In: *Science* 347.6224 (2015), p. 1257601.

- [40] Edward Choi et al. “Multi-layer representation learning for medical concepts”. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (2016), pp. 1495–1504.
- [41] Odysseus Data Services, Inc. *ATHENA - OHDSI VOCABULARIES REPOSITORY*. <https://athena.ohdsi.org/search-terms/terms>. Accessed March 4, 2022. 2022.
- [42] OHDSI collaborative. *Observational Health Data Sciences and Informatics*. <https://www.ohdsi.org/web/wiki/doku.php?id=documentation:vocabulary:sidebar>. Accessed March 4, 2022. 2016.
- [43] C-D Mayer, J Lorent, and GW Horgan. “Exploratory analysis of multiple omics datasets using the adjusted RV coefficient”. In: *Statistical Applications in Genetics and Molecular Biology* 10.1 (2011), Article 14.
- [44] Paul Jaccard. “The distribution of the flora in the alpine zone”. In: *New Phytologist* 11.2 (1912), pp. 37–50.
- [45] Peter M Visscher et al. “Genetic correlations of polygenic disease traits: from theory to practice”. In: *Nature Reviews Genetics* 20.10 (2019), pp. 567–581.
- [46] Brendan K Bulik-Sullivan et al. “LD Score regression distinguishes confounding from polygenicity in genome-wide association studies”. In: *Nature Genetics* 47.3 (2015), pp. 291–295.
- [47] Cathie Sudlow et al. “UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age”. In: *PLoS Medicine* 12.3 (2015), e1001779.
- [48] Neale Lab. *Defining Confidence Levels for UKB Round 2 LDSR Analyses*. https://nealelab.github.io/UKBB_ldsc/confidence.html#summary_of_confidence_ratings. Accessed March 8, 2022. 2019.
- [49] Keegan Korthauer et al. “A practical guide to methods controlling false discoveries in computational biology”. In: *Genome Biology* 20.1 (2019), pp. 1–21.

- [50] Guiyan Jia et al. “Estimating heritability and genetic correlations from large health datasets in the absence of genetic data”. In: *Nature Communications* 10.1 (2019), pp. 1–10.
- [51] Frank Wilcoxon. “Individual comparisons by ranking methods”. In: *Biometrics Bulletin* 1.6 (1945), pp. 80–83.
- [52] Leonard Kaufman and Peter J Rousseeuw. “Finding groups in data: an introduction to cluster analysis”. In: *John Wiley & Sons* (1990).
- [53] Brian S Everitt et al. *Cluster analysis*. John Wiley & Sons, 2011.
- [54] Peter Langfelder, Bin Zhang, and Steve Horvath. “Defining clusters from a hierarchical cluster tree: the Dynamic Tree Cut package for R”. In: *Bioinformatics* 24.5 (2008), pp. 719–720.
- [55] William M Rand. “Objective criteria for the evaluation of clustering methods”. In: *Journal of the American Statistical association* 66.336 (1971), pp. 846–850.
- [56] Charles Spearman. “The proof and measurement of association between two things”. In: *The American journal of psychology* 15.1 (1904), pp. 72–101.
- [57] Jan Hauke and Tomasz Kossowski. “Comparison of values of Pearson’s and Spearman’s correlation coefficients on the same sets of data”. In: *Quaestiones geographicae* 30.2 (2011), pp. 87–93.
- [58] R Core Team. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria, 2021. URL: <https://www.R-project.org/>.
- [59] Antoine Bordes et al. “Translating embeddings for modeling multi-relational data”. In: *Advances in neural information processing systems*. 2013, pp. 2787–2795.
- [60] Cross-Disorder Group of the Psychiatric Genomics Consortium et al. “Genetic overlap between attention-deficit/hyperactivity disorder and major depressive disorder”. In: *Psychological medicine* 43.9 (2013), pp. 1875–1883.

- [61] Kate M Scott et al. “Association between mental disorders and subsequent medical conditions”. In: *New England Journal of Medicine* 369.18 (2013), pp. 1715–1721.
- [62] Raquel Gonzalez-Casas et al. “Aplastic anemia in a patient with autoimmune hepatitis”. In: *European journal of gastroenterology & hepatology* 19.12 (2007), pp. 1115–1118.
- [63] Eleni Gkamprela, Melanie Deutsch, and Dimitrios Pectasides. “Iron deficiency anemia in chronic liver disease: etiopathogenesis, diagnosis and treatment”. In: *Annals of gastroenterology* 30.4 (2017), pp. 405–413.
- [64] Greg Gibson. “Rare and common variants: twenty arguments”. In: *Nature Reviews Genetics* 13.2 (2012), pp. 135–145.
- [65] Evan A Boyle, Yang I Li, and Jonathan K Pritchard. “An expanded view of complex traits: from polygenic to omnigenic”. In: *Cell* 169.7 (2017), pp. 1177–1186.
- [66] Peter N Robinson et al. “The Human Phenotype Ontology: a tool for annotating and analyzing human hereditary disease”. In: *The American Journal of Human Genetics* 83.5 (2008), pp. 610–615.
- [67] Laramie Duncan et al. “Analysis of polygenic risk score usage and performance in diverse human populations”. In: *Nature communications* 10.1 (2019), p. 3328.
- [68] Gill Livingston et al. “Dementia prevention, intervention, and care: 2020 report of the Lancet Commission”. In: *The Lancet* 396.10248 (2020), pp. 413–446.
- [69] Deborah E Barnes and Kristine Yaffe. “The potential for primary prevention of Alzheimer’s disease: an analysis of population-based data”. In: *The Lancet Neurology* 10.9 (2011), pp. 788–794.
- [70] Rebecca F Gottesman et al. “Association between midlife vascular risk factors and estimated brain amyloid deposition”. In: *JAMA* 317.14 (2017), pp. 1443–1450.
- [71] Anselm L Strauss. “Chronic illness and the quality of life”. In: *Mosby* (1975).

- [72] Ronald C Petersen et al. “Mild cognitive impairment: clinical characterization and outcome”. In: *Archives of neurology* 58.3 (2001), pp. 303–308.
- [73] Nicole Gray Weiskopf and Chunhua Weng. “Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research”. In: *Journal of the American Medical Informatics Association* 20.1 (2013), pp. 144–151.
- [74] George Hripcsak et al. “Observational health data sciences and informatics (OHDSI): opportunities for observational researchers”. In: *Studies in health technology and informatics* 216 (2015), p. 574.
- [75] Rakesh Agrawal and Ramakrishnan Srikant. “Mining sequential patterns”. In: *Proceedings of the eleventh international conference on data engineering*. IEEE. 1995, pp. 3–14.
- [76] Carlo Combi, Elpida Keravnou-Papailiou, and Yuval Shahar. “Temporal reasoning and temporal data maintenance in medicine: issues and challenges”. In: *Computers in biology and medicine* 40.11-12 (2010), pp. 985–995.
- [77] Hiroaki Sakoe and Seibi Chiba. “Dynamic programming algorithm optimization for spoken word recognition”. In: *IEEE transactions on acoustics, speech, and signal processing* 26.1 (1978), pp. 43–49.
- [78] François Petitjean, Alain Ketterlin, and Pierre Gançarski. “A global averaging method for dynamic time warping, with applications to clustering”. In: *Pattern Recognition* 44.3 (2011), pp. 678–693.
- [79] Petter Holme and Jari Saramäki. “Temporal networks”. In: *Physics reports* 519.3 (2012), pp. 97–125.
- [80] Li Li et al. “Identification of type 2 diabetes subgroups through topological analysis of patient similarity”. In: *Science translational medicine* 7.311 (2015), 311ra174–311ra174.

- [81] Anil K Jain. “Data clustering: 50 years beyond K-means”. In: *Pattern recognition letters* 31.8 (2010), pp. 651–666.
- [82] Edward Choi et al. “RETAIN: An interpretable predictive model for healthcare using reverse time attention mechanism”. In: *Advances in neural information processing systems* 29 (2016).
- [83] Austin Bradford Hill. “The environment and disease: association or causation?”. In: *Proceedings of the Royal Society of Medicine* 58.5 (1965), pp. 295–300.
- [84] Judea Pearl. *Causality*. Cambridge university press, 2009.
- [85] David Maxwell Chickering. “Optimal structure identification with greedy search”. In: *Journal of machine learning research* 3.Nov (2002), pp. 507–554.
- [86] Abel N Kho et al. “Use of diverse electronic medical record systems to identify genetic risk for type 2 diabetes within a genome-wide association study”. In: *Journal of the American Medical Informatics Association* 22.2 (2015), pp. 472–480.
- [87] HIPAA. *Health Insurance Portability and Accountability Act of 1996*. 2003.
- [88] Wei-Qi Wei et al. “Evaluating phecodes, clinical classification software, and ICD-9-CM codes for phenome-wide association studies in the electronic health record”. In: *PLoS One* 14.7 (2019), e0175508.
- [89] Tom Wilkinson et al. “Identifying dementia outcomes in UK Biobank: a validation study of primary care, hospital admissions and mortality data”. In: *European journal of epidemiology* 34.6 (2019), pp. 557–565.
- [90] Paul R Rosenbaum and Donald B Rubin. “The central role of the propensity score in observational studies for causal effects”. In: *Biometrika* 70.1 (1983), pp. 41–55.
- [91] Jason P Fine and Robert J Gray. “A proportional hazards model for the subdistribution of a competing risk”. In: *Journal of the American statistical association* 94.446 (1999), pp. 496–509.

- [92] Yoav Benjamini and Yosef Hochberg. “Controlling the false discovery rate: a practical and powerful approach to multiple testing”. In: *Journal of the Royal Statistical Society: Series B (Methodological)* 57.1 (1995), pp. 289–300.
- [93] Oanh Kieu Phung et al. “Validation of a coding algorithm to identify patients with end-stage renal disease in a commercially insured population”. In: *Clinical Epidemiology* 10 (2018), pp. 613–621.
- [94] Paolo Tormene et al. “Matching incomplete time series with dynamic time warping: an algorithm and an application to post-stroke rehabilitation”. In: *Artificial intelligence in medicine* 45.1 (2009), pp. 11–34.
- [95] John A Hartigan and Manchek A Wong. “Algorithm AS 136: A k-means clustering algorithm”. In: *Journal of the royal statistical society. series c (applied statistics)* 28.1 (1979), pp. 100–108.
- [96] Joe H Ward Jr. “Hierarchical grouping to optimize an objective function”. In: *Journal of the American statistical association* 58.301 (1963), pp. 236–244.
- [97] Peter J Rousseeuw. “Silhouettes: a graphical aid to the interpretation and validation of cluster analysis”. In: *Journal of computational and applied mathematics* 20 (1987), pp. 53–65.
- [98] Lawrence Hubert and Phipps Arabie. “Comparing partitions”. In: *Journal of classification* 2.1 (1985), pp. 193–218.
- [99] Jean-Charles Delvenne, Sophia N Yaliraki, and Mauricio Barahona. “Stability of graph communities across time scales”. In: *proceedings of the national academy of sciences* 107.29 (2010), pp. 12755–12760.
- [100] Hanbo Chen and Paul C Boutros. *VennDiagram: Generate High-Resolution Venn and Euler Plots*. R package version 1.6.20. 2011.

- [101] Edward L Kaplan and Paul Meier. “Nonparametric estimation from incomplete observations”. In: *Journal of the American statistical association* 53.282 (1958), pp. 457–481.
- [102] Nathan Mantel. “Evaluation of survival data and two new rank order statistics arising in its consideration”. In: *Cancer chemotherapy reports* 50.3 (1966), pp. 163–170.
- [103] Gideon Schwarz. “Estimating the dimension of a model”. In: *The annals of statistics* 6.2 (1978), pp. 461–464.
- [104] R Core Team. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria, 2019. URL: <https://www.R-project.org/>.
- [105] Terry M Therneau and Patricia M Grambsch. “Modeling survival data: extending the Cox model”. In: *Springer Science & Business Media* (2000).
- [106] Bob Gray. *cmprsk: Subdistribution Analysis of Competing Risks*. R package version 2.2-9. 2019.
- [107] Daniel E Ho et al. “MatchIt: nonparametric preprocessing for parametric causal inference”. In: *Journal of statistical software* 42.8 (2011), pp. 1–28.
- [108] Toni Giorgino. “Computing and visualizing dynamic time warping alignments in R: the dtw package”. In: *Journal of statistical Software* 31.7 (2009), pp. 1–24.
- [109] Markus Kalisch et al. “Causal inference using graphical models with the R package pcalg”. In: *Journal of statistical software* 47.11 (2012), pp. 1–26.
- [110] Gabor Csardi and Tamas Nepusz. “The igraph software package for complex network research”. In: *InterJournal, complex systems* 1695.5 (2006), pp. 1–9.
- [111] Carter T Butts. “network: a package for managing relational data in R”. In: *Journal of statistical software* 24.2 (2008), pp. 1–36.

- [112] Martin Maechler et al. *cluster: Cluster Analysis Basics and Extensions*. R package version 2.1.0. 2019.
- [113] Alboukadel Kassambara and Fabian Mundt. *factoextra: Extract and Visualize the Results of Multivariate Data Analyses*. R package version 1.0.5. 2017.
- [114] Hadley Wickham. *ggplot2: elegant graphics for data analysis*. Springer, 2016.
- [115] Alboukadel Kassambara, Marcin Kosinski, and Przemyslaw Biecek. *survminer: Drawing Survival Curves using 'ggplot2'*. R package version 0.4.9. 2021.
- [116] Marianthi Alexander et al. “Patient-level clustering for disease subtyping”. In: *Nature medicine* 24.9 (2018), pp. 1284–1285.
- [117] Deirdre H Taylor et al. “Temporal patterns in diagnosis codes before first diagnosis of Alzheimer disease”. In: *Medical care* (2020), pp. 302–308.
- [118] Hossein Estiri, Jeffrey G Klann, and Shawn N Murphy. “Transitive sequential pattern mining for discrete clinical data”. In: *IEEE transactions on biomedical engineering* 67.8 (2019), pp. 2174–2182.
- [119] Shanquan Li et al. “Disease trajectories and mortality among individuals diagnosed with depression: a community-based cohort study in UK Biobank”. In: *Molecular psychiatry* 27.8 (2022), pp. 3304–3312.
- [120] Sheng Chieh You et al. “Temporal evolution of clinical domains in Alzheimer’s disease dementia”. In: *Alzheimer’s & Dementia* 18.12 (2022), pp. 2448–2456.
- [121] Raymond L Ownby et al. “Depression and risk for Alzheimer disease: systematic review, meta-analysis, and metaregression analysis”. In: *Archives of general psychiatry* 63.5 (2006), pp. 530–538.
- [122] Deborah E Barnes et al. “Midlife vs late-life depressive symptoms and risk of dementia: differential effects for Alzheimer disease and vascular dementia”. In: *Archives of general psychiatry* 69.5 (2012), pp. 493–498.

- [123] Tamara G Fong et al. “Delirium accelerates cognitive decline in Alzheimer disease”. In: *Neurology* 83.15 (2014), pp. 1346–1353.
- [124] Francis S Collins and Harold Varmus. “A new initiative on precision medicine”. In: *New England journal of medicine* 372.9 (2015), pp. 793–795.
- [125] Valentina Escott-Price et al. “Polygenic risk score analysis of pathologically confirmed Alzheimer disease”. In: *Annals of neurology* 82.2 (2017), pp. 311–314.
- [126] Giorgio Sirugo, Scott M Williams, and Sarah A Tishkoff. “The missing diversity in human genetic studies”. In: *Cell* 177.1 (2019), pp. 26–31.
- [127] Roseann E Peterson et al. “Genome-wide association studies in ancestrally diverse populations: opportunities, methods, pitfalls, and recommendations”. In: *Cell* 179.3 (2019), pp. 589–603.
- [128] Alzheimer’s Association. *2022 Alzheimer’s disease facts and figures*. 2022.
- [129] Elizabeth Rose Mayeda et al. “Inequalities in dementia incidence between six racial and ethnic groups over 14 years”. In: *Alzheimer’s & Dementia* 12.3 (2016), pp. 216–224.
- [130] Melinda C Mills and Charles Rahal. “The GWAS Diversity Monitor tracks diversity by disease in real time”. In: *Nature genetics* 52.3 (2020), pp. 242–243.
- [131] Amy R Bentley, Shawneequa Callier, and Charles N Rotimi. “Diversity and inclusion in genomic research: why the uneven progress?” In: *Journal of community genetics* 8.4 (2017), pp. 255–266.
- [132] Charles N Rotimi et al. “Research capacity. Enabling the genomic revolution in Africa”. In: *Science* 344.6190 (2014), pp. 1346–1348.
- [133] Shing Wan Choi, Timothy Shin-Heng Mak, and Paul F O’Reilly. “Tutorial: a guide to performing polygenic risk score analyses”. In: *Nature protocols* 15.9 (2020), pp. 2759–2772.

- [134] Noah A Rosenberg et al. “Genome-wide association studies in diverse populations”. In: *Nature Reviews Genetics* 11.5 (2010), pp. 356–366.
- [135] 1000 Genomes Project Consortium. “A global reference for human genetic variation”. In: *Nature* 526.7571 (2015), pp. 68–74.
- [136] Joan A Casey et al. “Using electronic health records for population health research: a review of methods and applications”. In: *Annual review of public health* 37 (2016), pp. 61–81.
- [137] Roch Johnson et al. “The UCLA ATLAS Community Health Initiative: promoting precision health research in a diverse biobank”. In: *Cell Genomics* 3.1 (2023), p. 100243.
- [138] Yann Li, Chia-Yen Chen, and Wyeth W Wasserman. “Machine learning in genetics and genomics”. In: *Nature Reviews Genetics* 21.12 (2020), pp. 747–762.
- [139] Hui Zou and Trevor Hastie. “Regularization and variable selection via the elastic net”. In: *Journal of the royal statistical society: series B (statistical methodology)* 67.2 (2005), pp. 301–320.
- [140] Kyoko Watanabe et al. “Functional mapping and annotation of genetic associations with FUMA”. In: *Nature communications* 8.1 (2017), p. 1826.
- [141] Jose A Santiago, Virginie Bottero, and Judith A Potashkin. “Transcriptomic and network analysis identifies shared and unique pathways across dementia spectrum disorders”. In: *International journal of molecular sciences* 21.6 (2020), p. 2050.
- [142] EH Corder et al. “Gene dose of apolipoprotein E type 4 allele and the risk of Alzheimer’s disease in late onset families”. In: *Science* 261.5123 (1993), pp. 921–923.
- [143] Lindsay A Farrer et al. “Effects of age, sex, and ethnicity on the association between apolipoprotein E genotype and Alzheimer disease: a meta-analysis”. In: *Jama* 278.16 (1997), pp. 1349–1356.

- [144] Emmanuelle Genin et al. “APOE and Alzheimer disease: a major gene with semi-dominant inheritance”. In: *Molecular psychiatry* 16.9 (2011), pp. 903–907.
- [145] John Hardy and Andrew Singleton. “Genome-wide association studies and human disease”. In: *New England Journal of Medicine* 360.17 (2009), pp. 1759–1768.
- [146] Brian W Kunkle et al. “Novel Alzheimer disease risk loci and pathways in African American individuals using the African Genome Resources Panel: a meta-analysis”. In: *JAMA neurology* 78.1 (2021), pp. 102–113.
- [147] Naomi R Wray, Michael E Goddard, and Peter M Visscher. “Prediction of individual genetic risk to disease from genome-wide association studies”. In: *Genome research* 17.10 (2007), pp. 1520–1528.
- [148] Itziar de Rojas et al. “Common variants in Alzheimer’s disease and risk stratification by polygenic risk scores”. In: *Nature communications* 12.1 (2021), p. 3417.
- [149] Florian Privé et al. “Portability of 245 polygenic scores when derived from the UK Biobank and applied to 9 ancestry groups from the same cohort”. In: *The American Journal of Human Genetics* 109.1 (2022), pp. 12–23.
- [150] Alicia R Martin et al. “Human demographic history impacts genetic risk prediction across diverse populations”. In: *The American Journal of Human Genetics* 100.4 (2017), pp. 635–649.
- [151] Peter M Visscher et al. “10 years of GWAS discovery: biology, function, and translation”. In: *The American Journal of Human Genetics* 101.1 (2017), pp. 5–22.
- [152] GTEx Consortium. “The GTEx Consortium atlas of genetic regulatory effects across human tissues”. In: *Science* 369.6509 (2020), pp. 1318–1330.
- [153] Stefan Schoenfelder et al. “Preferential associations between co-regulated genes reveal a transcriptional interactome in erythroid cells”. In: *Nature genetics* 42.1 (2010), pp. 53–61.

- [154] Nadia Solovieff et al. “Pleiotropy in complex traits: challenges and strategies”. In: *Nature Reviews Genetics* 14.7 (2013), pp. 483–495.
- [155] Joshua C Denny et al. “Systematic comparison of phenome-wide association study of electronic medical record data and genome-wide association study data”. In: *Nature biotechnology* 31.12 (2013), pp. 1102–1110.
- [156] Omri Gottesman et al. “The Electronic Medical Records and Genomics (eMERGE) Network: past, present, and future”. In: *Genetics in Medicine* 15.10 (2013), pp. 761–771.
- [157] Illumina. *Infinium Global Diversity Array-8 BeadChip — Array for Human Genotyping Screening*. 2024. URL: <https://sapac.illumina.com/products/by-type/microarray-kits/infinium-global-diversity.html>.
- [158] Clara M Lajonchere et al. “An integrated, scalable, electronic video consent process to power precision health research: large, population-based, implementation study”. In: *Journal of medical Internet research* 24.4 (2022), e30746.
- [159] Christopher C Chang et al. “Second-generation PLINK: rising to the challenge of larger and richer datasets”. In: *Gigascience* 4.1 (2015), s13742–015.
- [160] Sayantan Das et al. “Next-generation genotype imputation service and methods”. In: *Nature genetics* 48.10 (2016), pp. 1284–1287.
- [161] Iain Mathieson et al. “The genomic history of southeastern Europe”. In: *Nature* 555.7695 (2018), pp. 197–203.
- [162] Katarzyna Bryc et al. “The genetic ancestry of African Americans, Latinos, and European Americans across the United States”. In: *The American Journal of Human Genetics* 96.1 (2015), pp. 37–53.
- [163] Alkes L Price et al. “Principal components analysis corrects for stratification in genome-wide association studies”. In: *Nature genetics* 38.8 (2006), pp. 904–909.

- [164] Andrés Moreno-Estrada et al. “Reference populations and composite panels for next-generation sequencing-based characterizations of human genome variation”. In: *Nature genetics* 46.3 (2014), pp. 303–312.
- [165] Ruth Chia et al. “Genome sequencing analysis identifies new loci associated with Lewy body dementia and provides insights into its genetic architecture”. In: *Nature genetics* 53.3 (2021), pp. 294–303.
- [166] Mike A Nalls et al. “Identification of novel risk loci, causal insights, and heritable risk for Parkinson’s disease: a meta-analysis of genome-wide association studies”. In: *The Lancet Neurology* 18.12 (2019), pp. 1091–1102.
- [167] Günter U Höglinger et al. “Identification of common variants influencing risk of the tauopathy progressive supranuclear palsy”. In: *Nature genetics* 43.7 (2011), pp. 699–705.
- [168] Rainer Malik et al. “Multiancestry genome-wide association study of 520,000 subjects identifies 32 loci associated with stroke and stroke subtypes”. In: *Nature genetics* 50.4 (2018), pp. 524–537.
- [169] Martin Kircher et al. “A general framework for estimating the relative pathogenicity of human genetic variants”. In: *Nature genetics* 46.3 (2014), pp. 310–315.
- [170] Julie A Schneider et al. “Mixed brain pathologies account for most dementia cases in community-dwelling older persons”. In: *Neurology* 69.24 (2007), pp. 2197–2204.
- [171] Takaya Saito and Marc Rehmsmeier. “The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets”. In: *PloS one* 10.3 (2015), e0118432.
- [172] Davide Chicco and Giuseppe Jurman. “The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation”. In: *BMC genomics* 21.1 (2020), p. 6.

- [173] Elizabeth R DeLong, David M DeLong, and Daniel L Clarke-Pearson. “Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach”. In: *Biometrics* (1988), pp. 837–845.
- [174] R Core Team. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria, 2020. URL: <https://www.R-project.org/>.
- [175] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. “Regularization paths for generalized linear models via coordinate descent”. In: *Journal of statistical software* 33.1 (2010), p. 1.
- [176] Xavier Robin et al. “pROC: an open-source package for R and S+ to analyze and compare ROC curves”. In: *BMC bioinformatics* 12.1 (2011), p. 77.
- [177] Jan Grau, Ivo Grosse, and Jens Keilwagen. *PRROC: computing and visualizing precision-recall and receiver operating characteristic curves in R*. 2015.
- [178] Fabian Pedregosa et al. “Scikit-learn: Machine learning in Python”. In: *the Journal of machine Learning research* 12 (2011), pp. 2825–2830.
- [179] Wes McKinney et al. “Data structures for statistical computing in python”. In: *Proceedings of the 9th Python in Science Conference*. Vol. 445. Austin, TX. 2010, pp. 51–56.
- [180] Kristen Clark et al. “The prediction of Alzheimer’s disease through multi-trait genetic modeling”. In: *Frontiers in aging neuroscience* 15 (2023), p. 1168638.
- [181] Alexander M Kulminski et al. “Definitive roles of TOMM40-APOE-APOC1 variants in the Alzheimer’s risk”. In: *Neurobiology of aging* 110 (2022), pp. 122–131.
- [182] Michael E Belloy, Valerio Napolioni, and Michael D Greicius. “A quarter century of APOE and Alzheimer’s disease: progress to date and the path forward”. In: *Neuron* 101.5 (2019), pp. 820–838.

- [183] Jessica Rae Marden et al. “Validation of a polygenic risk score for dementia in black and white individuals”. In: *Brain and behavior* 4.5 (2014), pp. 687–697.
- [184] Ozan Dikilitas et al. “Use of polygenic risk scores for coronary heart disease in ancestrally diverse populations”. In: *Current cardiology reports* 24.9 (2022), pp. 1169–1177.
- [185] Sanjeev Sariya et al. “Polygenic risk score for Alzheimer’s disease in Caribbean Hispanics”. In: *Annals of neurology* 90.3 (2021), pp. 366–376.
- [186] Xiaoyu Ruan et al. “Application of European-specific polygenic risk scores for predicting prostate cancer risk in different ancestry populations”. In: *The Prostate* 83.1 (2023), pp. 30–38.
- [187] Sang-Hyuk Jung et al. “Transferability of Alzheimer disease polygenic risk score across populations and its association with Alzheimer disease-related phenotypes”. In: *JAMA network open* 5.7 (2022), e2247162–e2247162.
- [188] Guy M McKhann et al. “The diagnosis of dementia due to Alzheimer’s disease: recommendations from the National Institute on Aging-Alzheimer’s Association workgroups on diagnostic guidelines for Alzheimer’s disease”. In: *Alzheimer’s & dementia* 7.3 (2011), pp. 263–269.
- [189] Yuen-Shan Ho, Frank Hu, and Pui Lee. “The advantages and challenges of using real-world data for patient care”. In: *Clinical and translational science* 13.1 (2020), pp. 4–7.
- [190] Guy McKhann et al. “Clinical diagnosis of Alzheimer’s disease: Report of the NINCDS-ADRDA Work Group* under the auspices of Department of Health and Human Services Task Force on Alzheimer’s Disease”. In: *Neurology* 34.7 (1984), pp. 939–944.
- [191] Elias Engelhardt et al. “Vascular dementia: diagnostic criteria and supplementary exams. Recommendations of the Scientific Department of Cognitive Neurology and

Aging of the Brazilian Academy of Neurology. Part I". In: *Dementia & neuropsychologia* 5.4 (2011), pp. 251–263.