Leveraging Clinical Data and Knowledge Networks to Derive Insights Into Alzheimer's Disease

by

Alice Tang

DISSERTATION

Submitted in partial satisfaction of the requirements for degree of

DOCTOR OF PHILOSOPHY

in

Bioengineering

in the

GRADUATE DIVISION

of the

UNIVERSITY OF CALIFORNIA, SAN FRANCISCO

AND

UNIVERSITY OF CALIFORNIA, BERKELEY

Approved:

DocuSigned by:
Marina Sirota — Marina Sirota
925B61AB9C41499...
Chair

DocuSigned by:
Dena Dubal — Dena Dubal
D54082EE4174FD...

DocuSigned by:
Steven Conolly — Steven Conolly
B22ED1E848D494...

DocuSigned by:
Sergio Baranzini — Sergio Baranzini
07D1896A48A94E7...

DocuSigned by:
John A. Capra — John A. Capra
CBF388B27CEB4AC...

Committee Members

Acknowledgements

I would like to give a big thank you to my thesis advisor Professor Marina Sirota, as well as the smiles and support of her family and three children throughout my time in the lab. I would also like to acknowledge the support and contributions of my thesis committee mentors, Drs. Steven Conolly, Dena Dubal, Professor Sergio Baranzini, and Professor John (Tony) Capra. I would also like to acknowledge Professor Katherine Rankin as an extraordinary mentor for meeting with me weekly for intellectual and professional support. There are also many other remarkable individuals who supported my journey along the way, and I would like to acknowledge all of their contributions including: Zicheng Hu for introducing me to my advisor and for mentoring me in informatics, Billy Zeng for being my amazing big sib and stimulating my interest in clinical informatics and machine learning bias in medicine, Jonathan Golob for being an amazing scientist-engineer-physician role model and parental figure, Tomiko Oskotsky and Boris Oskotsky for endless emotional and intellectual support and forever my honorary research family. Also there have been so many others that contribute to my thesis and success including Valentina Pedoia, Atul Butte, Charlotte Nelson, Maria Glymour, Isabel Allen, Zachary Miller, and many other mentors throughout my PhD journey at UCSF.

I want to acknowledge the administrators that make my training and research work possible, including Geri Ehle, Tiffani Quan, Victoria Starrett, Rocio Sanchez, Evan Phelps, Hunter Mills, Albert Lee, Rhett Hillary, Grace Loll, Edna Rodas, and all the wonderful patients whose visit to the UC system contributed data for research. Furthermore, I'd like to acknowledge the rest of the research support teams at UCSF that made the data and computational environments available to researchers like me. I want to acknowledge my fellow lab members in the Sirota lab, and all my mentors and mentees. I want to acknowledge my friends in both medical and graduate

school along the way as well, particularly my parents, grandparents, the two wonderful Richard's in my life, and the rest of my friends and family that wish the best for me even if they do not understand the PhD journey.

Of course, most importantly, I want to acknowledge my partner who supported me so much along the entire pathway, Chi-Yo Tsai. I would not have gone so far without his support, encouragement, and immense aid both intellectually and emotionally.

Contributions

Several chapters contain previously published material or material that is currently in preparation for submission. They do not represent the final published forms and have been edited slightly.

Ch 1 includes edited combinations of the following publications: (1) Alice Tang, Sarah Woldemariam, Jacquelyn Roger, and Marina Sirota. Translational Bioinformatics to Enable Precision Medicine for All: Elevating Equity across Molecular, Clinical, and Digital Realms. Yearb Med Inform. 2022 Aug;31(1):106-115. doi: 10.1055/s-0042-1742513. PMID: 36463867; PMCID: PMC9719766. (2) Alice Tang, Tomiko Oskotsky, and Marina Sirota. Personalizing routine lab tests with machine learning. Nat Med 2^{5}7, 1514–1515 (2021). https://doi.org/10.1038/s41591-021-01486-4

Ch 2 of the thesis includes a reprint of a previous publication: Alice S. Tang, Tomiko Oskotsky, Shreyas Havaldar, William G. Mantyh, Mesude Bicak, Caroline Warly Solsberg, Sarah Woldemariam, Billy Zeng, Zicheng Hu, Boris Oskotsky, Dena Dubal, Isabel E. Allen, Benjamin S. Glicksberg, Marina Sirota. Deep phenotyping of Alzheimer's disease leveraging electronic medical records identifies sex-specific clinical associations. Nat Commun 13, 675 (2022). https://doi.org/10.1038/s41467-022-28273-0

Funding support is provided through Grant # NIA R01AG060393, R01AG057683, RF1AG068325, and Medical Scientist Training Program T32GM007618.

Author Contributions: A.T. and M.S. designed the question, experiments, and analytic plan. B.Z. and Z.H. helped with data acquisition, cleaning, and interpretation. A.T., C.W.S., and

B.O. helped with creation of Rshiny app. A.T., W.M., T.O., C.W.S. and D.D. interpreted results. J.H, M.B, and B.G helped acquire and analyze validation data. S.W. and I.A. aided in statistical methods. A.T. wrote the manuscript with editing from all the authors. All the authors edited and reviewed the manuscript.

Ch 3 contains material that is currently in preprint and currently in submission: Alice Tang, Katherine P. Rankin, Gabriel Cerono, Silvia Miramontes, Hunter Mills, Jacquelyn Roger, Billy Zeng, Charlotte Nelson, Karthik Soman, Sarah Woldemariam, Yaqiao Li, Albert Lee, Riley Bove, Maria Glymour, Nima Aghaeepour, Tomiko Oskotsky, Zachary Miller, Isabel Allen, Stephan J. Sanders, Sergio Baranzini, Marina Sirota. Leveraging Electronic Medical Records and Knowledge Networks to Predict Disease Onset and Gain Biological Insight Into Alzheimer's Disease. medRxiv 2023.03.14.23287224; doi: https://doi.org/10.1101/2023.03.14.23287224.

Support is provided through NIA R01AG060393, T32GM007618, F30AG079504-01, NSF GRFP 2038436, grant NIA P30-AG062422

Author Contributions: AT, KR, and MS developed and directed the entire project. AT, KR, JR, HM, TO, BZ, CN, KS, MG, IA, and MS aided in the study design approach regarding cohort selection, control selections, and time frame selection. AT and KR. acquired the data and executed selection approaches, and KR reviewed diagnostic codes for selection. AT, JR, HM, CN, SW, and AL aided in EMR data preprocessing. AT, KR, JR, HM, CN, KS, and SW aided in the design of predictive modeling and evaluation. AT executed all aspects of data acquisition, preprocessing, model implementation, model evaluation, and model feature importance analysis. AT, KR, GC, SM, TO, BZ, CN, SW, YL, MG, IA, ZM, and MS contributed to the interpretation of predictive models, with statistical insight from MG and IA, and clinical insight from KR and ZM. AT, KR,

JR, SM, SW, YL, MG, and IA aided in the design of the study for external EHR validation and survival analysis. GC, CN, KS, SEB aided in access and utilization of the SPOKE knowledge graph. AT, KR, GC, RB, CN, SEB, SS, And MS aided in approaches for knowledge network interpretation and genetic validation. GC and AT executed the genetic analysis, with input from CN, SS, SEB, and MS. AT generated the figures and tables with help from HM and GC. AT prepared and wrote the manuscript, with inputs from all the authors. All authors read and approved the final manuscript, with expertise from NA for machine learning expertise and manuscript revisions.

Ch 4 contains material currently in preparation for submission. Other authors include Sarah Woldemariam, Silvia Miramontes, Tomiko Oskotsky, and Marina Sirota, who provided edits and suggestions to the text.

Epigraph

"In the intricate web of life's complexity, we stand at the crossroads of exploratory wonder and hypothesis-driven clarity. The brilliance of a functional brain is our beacon, urging us to dispel the clouds of neurodegeneration. As dawn breaks on the future of medicine, AI (Artificial Intelligence) and ML (Machine Learning) light the way, crafting a precision in care that resonates with the unique rhythm of each soul."

ChatGPT-4

“Words do not express thoughts very well. They always become a little different immediately they are expressed, a little distorted, a little foolish. And yet it also pleases me and seems right that what is of value and wisdom to one man seems nonsense to another.”

Hermann Hesse (from the novel Siddhartha)

“It’s complicated”

Sirota Lab

Abstract

Leveraging Clinical Data and Knowledge Networks to Derive Insights into Alzheimer's Disease

Alice Tang

Alzheimer's Disease (AD) is a devastating neurodegenerative disorder that is difficult to study and treat despite decades of progress. This is due to disease heterogeneity, lack of precise phenotyping, and limited understanding of molecular mechanisms underlying clinical manifestations. Electronic medical records (EMR) are emerging as a real-world dataset with abundance of longitudinal human data across diagnoses, medications, and measurements with opportunity to derive insights without predefined selection criteria or limitations in scope. Recent developments of integrative heterogeneous graph databases that combine knowledge across omics relationships provide a means to further identify molecular hypotheses underlying complex clinical phenotypes. We performed deep phenotyping to characterize AD and sex differences in the EMR against a control cohort, and identified sex and AD associated comorbidities, medication use, and lab values. Extending this work to apply machine learning, we utilize clinical information to predict AD onset and identify prioritized genes via knowledge networks (e.g., APOE, ACTB, IL6) and genetic colocalization analysis (e.g., MS4A6A with osteoporosis). Our findings suggest that AD onset risk can be predicted based on clinical data and that there are sex-specific relationships in AD including musculoskeletal disorders among females with AD and neurological or sensory disorders among males with AD. Extensions to knowledge networks and molecular datasets further prioritize genes depending on an individual's comorbid conditions. By leveraging clinical data to identify hypotheses for complex disease, we can further make steps towards better understanding molecular mechanisms and advance personalized treatment approaches in AD.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

AD: Alzheimer's Disease

AI: Artificial Intelligence

ALT: alanine aminotransferase

ATC: Anatomical Therapeutic Chemical

AUPRC: Area under the precision recall curve

AUROC: Area under the receiver operating curve

EMR/EHR: Electronic Medical Records / Electronic Health Records

Glu: glucose

GWAS: genetic wide association study

HBMD : heel bone mineral density

Hgb: hemoglobin

H/L: Hispanic / Latino

HLD: Hyperlipidemia

HR / aHR: hazard ratio / adjusted hazard ratio

ICD: International Classification of Diseases

MAC: Memory and Aging Center

ML: Machine Learning

NHPI: Native Hawaiian or Pacific Islander

OMOP: Observational Medical Outcomes Partnership

PS: Propensity Score

Pt: Patient

QTL: Quantitative Trait Loci

RF: Random Forest

RWD: Real World Data

SD: standard deviation

SMD: standardized mean difference

SNOMED: Systemized Nomenclature of Medicine

SPOKE: Scalable Precision Medicine Open Knowledge Engine

UC: University of California

UCDDP: University of California Data Discovery Platform

UCSF: University of California, San Francisco

UMAP: Uniform Manifold Approximation and Projection

Chapter 1: Opportunities from Bioinformatics to Clinical Informatics for Understanding and Managing Disease

In the 21th century, the proliferation of electronics, digital tools, sequencing technologies, data storage capabilities, and methodological advancements paved the way for an abundance of diverse and large datasets that can be either utilized or repurposed for research or translational goals. These datasets range across the multi-omics spectrum: from genetics databases like dbGAP (database of Genotypes and Phenotypes, ncbi.nlm.nih.gov/gap) and The Cancer Genome Atlas (portal.gdc.cancer.gov), to realms of epigenetics, gene and protein expression, clinical records, sensor outputs, imaging, disease epidemiology, and more. Some repositories of datasets, like the ADNI (Alzheimer's Disease Neuroimaging Initiative, adni.loni.usc.edu) and NACC (National Alzheimer's Coordinating Center, naccdata.org) databases for neurodegeneration or March of Dimes (MOD) Preterm Birth Database (pretermbirthdb.org) for pregnancy outcomes, provide focal points for the investigation of specific research questions and topics. The digital transformation has amplified data collection in health maintenance and care delivery, as evidenced by extensive electronic medical records and novel utilization of data collection sources like smartwatches or cellphone apps.

Such wealth of data provide opportunity in exploring potential untapped research questions and development of tools for translational applications. In this chapter, a general review of translational informatics across multimodal data domains will be provided, with emphasis on data representation and equity. Then a commentary on machine learning based personalized lab result insights will be provided. These will serve as the basis that leads to EMR-based insights for

Alzheimer's Disease in following chapters, and finally an overview of methods and considerations for expanding EMR informatics into other disease areas.

1.1 Translational Informatics Across Multimodal Data Domains for Equitable Precision Medicine

1.1.1 Abstract

Challenges from the COVID-19 pandemic led to collaborative efforts among researchers in the realms of data sharing and algorithmic development efforts across molecular, clinical, and digital health domains for goals of prevention and treatment. We performed a literature assessment of trends and approaches in clinical and translational bioinformatics to characterize and describe recent advanced computational approaches since 2020. This includes applications of phenotyping, disease subtype characterization, predictive modelling, biomarker discovery, treatment selection, and artificial intelligence model utilization for advancement of human health. To pursue the goals of equity and inclusion in scientific advancements and translational applications, data representation and bias mitigation should be considered at every step including project design, data collection, model creation, clinical implementation, and evaluation. Data representativeness along with breakthroughs in big data and artificial intelligence will guide the future in precision medicine applications for health.

1.1.2 Introduction

With the increasing acquisition of multimodal data (e.g., measurements or records that span across multiple sources, such as genetic and imaging data), recent terms such as 'translational bioinformatics' are evolving to encompass the discipline of the use of computational approaches and tools across life sciences and clinical data for the purpose of advancing human health or medicine[1]. Bioinformatics computational approaches have aided in the advancement for scientific

understanding of biological and disease phenomenon, often from genetics, gene expression, and proteomics datasets. Recent increasing availability of molecular testing in the clinic and opportunity to combine these datasets with clinical data has enabled applications for translational applications for improved disease understanding or treatment.

Computational tools in bioinformatics have evolved along with technical advances in genomic and single-cell sequencing, microbiome sequencing, proteomics, imaging technologies, and other technologies to capture biological data at a cellular level[2-4]. Tools to analyze these large datasets span across traditional statistical analyses to machine learning and unsupervised clustering to better identify patterns and associations with minimal human intervention. Models that learn upon biological phenomenon, as well as integrate across clinically observed diseases and phenotypes, are increasingly being trained and applied for precision medicine approaches such as disease risk prediction, diagnostic reasoning and classification, and prognostic modelling. In the years following 2020, particularly due to challenges brought on by the COVID-19 pandemic, there has also been unprecedented efforts in collaboration and data or model sharing across academia and industry to tackle the challenges caused by the changing nature of SARS-CoV-2, the virus that caused COVID-19[5,6]. Systemic challenges due to shelter-in-place orders have also shifted paradigms in healthcare delivery to increasingly rely on the role of technology for remote patient-centric healthcare monitoring and delivery, leading to sustained efforts for convenient local health management. This includes utilization of nearby lab facilities, imaging facilities, and digital health devices to help with timely data collection and data sharing. Mobile clinics and healthcare teams are also increasingly able to visit patient homes to deliver relevant devices, sensors, and even medications. Telehealth visits are increasingly utilized for appointments that do not require extensive physical exams, including psychiatric appointments and follow-up appointments from a

procedure. These changes lead to efforts to overcome challenges in data sharing and data processing efforts across molecular omics, clinical data, and digital health for the advancement of algorithmic approaches in healthcare and precision medicine in the future[7].

Despite the rising opportunities available for leveraging multimodal datasets to understand and tackle disease, it is important to evaluate how inequities can arise across the computational pipeline. Inequities in the dataset and modelling approaches can lead to societally biased scientific insight or biased algorithms, which can be further propagated when applied to translation. These biases can arise from the beginning, with data collection and representation, to algorithmic bias in model design, behavioral biases in data availability, as well as bias propagation due to algorithm utilization despite data shifts or drift. Scientific advances should therefore be considered with a framework of equity and inclusion to prevent transmission of healthcare disparities in translational applications and ensure algorithmic advances can benefit all communities and populations.

For example, studies leveraging data from the All of Us database have explored differences in disease prevalence across diverse populations, such as eczema and cardiovascular disease[8,9], which provides a starting point to characterize disease epidemiology in a heterogenous patient sample and motivate further research into understanding and addressing causes. The All of Us population includes diverse racialized individuals, those over the age of 75, people with disabilities, people with lower income, and people with less formal education. The All of Us dataset has also been utilized to study disparities in family health history knowledge and the ability to afford medications for diseases such as glaucoma[10,11]. These types of studies are useful to inform current understanding of disease characteristics and phenotypes across society, with the benefit of inclusion of minority populations and allowing for informed changes across domains of health

policy, clinical decision making, and design of research studies to ensure proper representation in the study of disease biology[12].

In this section, a brief review will be provided on recent advances across the fields of bioinformatics, translational informatics, and clinical or medical informatics, and across omics domains including molecular, microbiome, clinical, and digital health.

1.1.3 Approach

A literature search was performed on PubMed, Google Scholar, and within specific journals for publications past 2019, including Nature, Nature Digital Medicine, Nature Bioengineering, Lancet, the Journal of American Medical Association, Journal of Medical Internet Research, the New England Journal of Medicine, and Bioinformatics. Keyword searches were also performed to identify relevant publications, with keywords chosen by both broad and specific translational informatics topics (**Table 1.1.1**). References were also acquired from citations in papers identified from reviewed journals and keyword searches. After surveying identified papers, chosen papers were determined by their breadth, novelty, impact, or relevance, with a particular focus on papers that touch upon equity or inclusivity in the informatics fields.

1.1.4 Survey of Translational Bioinformatics

Translational bioinformatics applications include various goals, such as disease phenotyping, disease characterization, predictive modelling, trajectory modelling, subphenotyping, and drug discovery. Among these applications and pathway, we mention how equity and inclusion which should be considered at every step of the process including population identification, data collection, methodology, and algorithmic applications (**Figure 1.1.1**).

Informatics with Molecular Data

There are many recent exciting advances in the utilization of omics data to gain insights into complex diseases, discover biomarkers, therapeutic targets, and perform drug discovery through computational approaches across machine learning disciplines. Molecular datasets include diverse data modalities measuring the genome and polymorphisms, cancer gene mutations, epigenetics, gene expression, proteomics, microbiome, and others. These measurements have also advanced to acquisition with high temporal and spatial resolution, including data at the single cell and/or single organism level. As technologies are becoming more advanced, not only has molecular measurements become more easily attainable in the clinic, but advances in both molecular measurements and algorithmic development have allowed for improved clinical care to include cancer phenotyping, infectious disease identification, and disease risk identification. As technologies become increasingly advanced, there is also the need to revisit goals of equitable representation and conclusions across molecular studies and clinical implementation.

The rise in available measurements across omics modalities across the same samples and patients have paved the way for research in understanding associations among the intricacies of disease and health. Microbial compositions and host transcriptome have been utilized for applications including understanding crosstalk that influencing disease risk in inflammatory bowel disease (IBD) and irritable bowel syndrome (IBS)[13,14], as well as prediction of preterm birth via a crowdsourced DREAM challenge[15]. Datasets provided in AMP-AD (Accelerating Medicines Partnership - Alzheimer's Disease) for Alzheimer's Disease also allow for investigations across both mice and human for understanding of shared disease mechanisms. Multi-omics datasets are also being utilized to help identify disease biomarkers. For example, the CCGA (Circulating Cell-free Genome Atlas) consortium employed an ML method to detect cancer and its origin by analyzing the DNA methylation patterns of participants as a means for potential usage in early

cancer detection and treatment[16]. Some studies have also employed similar methods in the identification of infectious disease based on cell-free DNA (cfDNA), particularly among infectious diseases such as COVID-19[17]. Furthermore, there exists great interest in identifying biomarkers for aging and AD, with many potentially identified markers in the blood and cerebrospinal fluid, and full-scale clinical integration still pending.

Beyond disease detection, there is notable progress in identifying new indications for existing FDA-approved medicines, called drug repurposing, which offers hope in treating a greater variety of disease[18,19]. Some initiatives, like the OCTAD (Open Cancer TheraApeutic Discovery) website, are dedicated to assisting researchers in cancer drug discovery by comparing compound-induced gene expression signatures with gene expression data from patients[20]. The urgency brought on by the COVID-19 crisis has also sparked efforts to find repurposed medicines, with some prior approaches identifying statins and antipsychotics, as well as remdesivir, as potential treatments in the early stages of the crisis[21,22]. Overall, repurposing existing drugs can revolutionize treatment methods for various diseases in the foreseeable future, and methods spanning across omics domains are evolving for that purpose.

Considering Equity: It is crucial that all advancements in multiomics informatics methods and applications cater to everyone equitably. The NHGRI (National Human Genome Research Institute) emphasizes a diverse workforce in genomics and inclusive research participant to advance understanding of diseases among diverse patient populations[23]. A current challenge is the overrepresentation of European individuals in genome studies (>80%), leading to potential health disparities due to decreased predictive power from polygenic risk scores in non-European individuals[24]. However, recent endeavors aim to address this imbalance. For example, the PAGE (Population Architecture using Genomics and Epidemiology) study incorporated a diverse non-

European participant pool and unearthed various novel findings and ancestry-specific polymorphisms[26]. The All of Us research initiative also emphasizes diverse participant involvement and combines various data types, aiming for a holistic understanding of human health by integrating data across molecular data, electronic health records, survey data, and other social determinants of health[26,27]. For the goal of equitable benefit of precision medicine to be applied in clinical translation, research inclusivity should be prioritized to ensure [28,29].

Informatics with Clinical Data

Clinical information includes data extracted from electronic health databases, clinical trials, imaging, and notes. Only recently has the potential of clinical data been utilized for bioinformatics studies. Within electronic medical records (EMRs), patient information across diagnoses, lab results, drug prescriptions, and outcomes can be linked and investigated, and potentially include information about a patient's economic and social backgrounds[30]. For example, a study in the UK on pediatric diabetes identified differences in treatment regimen due to racism exposures[31]. Clinical informatics have led to advancements in patient profiling, disease prediction, treatment decisions, and subtyping. Based on a patient's health profile, algorithms can be employed that span from basic association analyses and diagnostic groupings to machine learning models that intelligently embed and cluster patients. This has been applied to several diseases such as type 2 diabetes, Alzheimer's Disease, and depression[32-34]. These phenotyping approaches highlight the heterogenous nature of disease and provide opportunity for treatment, but this heterogeneity also highlights the necessity to include diverse cohorts for fair representation across these heterogenous presentations and potential variations in treatment approaches.

There have also been a surge in further applications of clinical data for forecasting and predicting clinical outcomes, triaging disease severity, and assessing treatment efficacy or adverse

effects[35-45]. Emphasis is particularly placed on interpretable models as opposed to 'black box' models, as well as models that can be validated across medical centers to account for heterogeneity in treatment and patient differences. Since clinical records also contain extensive records of patient treatment history and diagnoses, drug repurposing applications are investigated through virtual or in silico trials to understand treatment patterns and gain insight into drug associations and outcomes[46-49]. Ultimately, the aim of many clinical informatics studies is the integration into clinical routines and treatment decision making. Example successes of translational applications include stories of prediction of patients undergoing radiation therapy[52], identifying adults at risk for clinical deterioration[53], guiding ultrasound and procedures [54], and managing COVID-19 outbreaks[55].

Considering Equity: As clinical bioinformatics increasingly plays a role in medical care and insight, it is crucial that algorithms developed are beneficial to everyone, especially groups that have been historically marginalized or understudied. If models are trained on data that lack representation from diverse individuals, such as racial exposures and gender identity groups, machine learning algorithms will learn from the bias themselves[57-67]. Therefore, it is essential that research and algorithmic development takes into account equity along the pipeline of data collection and model development.

Informatics with Digital Data

In the past two years, the COVID-19 pandemic has created many challenges and opportunities in utilizing technology to aid in healthcare when direct face-to-face meetings are less feasible, including video visits (telehealth) and sensor use via phones or wearables[68]. This led to maturation in digital health, informatics, and machine learning as a way to combat the pandemic

from both a public health perspective on prevention and control, as well as with providing individualized healthcare[4].

Mobile phones and wearables help provide a source of data that can be analyzed for health outcomes. As an example, population level information has been utilized to help with contact tracing at the start of the pandemic[69], and modeling infectious spread throughout numerous countries[70,71]. Sensor data have also been utilized learning to detect COVID-19 infection[72] via tracking of vital signs, sleep, activity, and even speech[73,74]. These 'digital biomarkers' provide an alternative proxy to invasive blood tests or molecular biomarkers, with applications including screening for depression[75], prediction of Parkinson's disease severity[77], and evaluating frailty in older people[79]. These applications provide recommendations in consumer applications or are slowly integrating into medical care as evident in the use of digital biomarkers for onsite patient triage and evaluation.

There has also been much work from the translational perspective in applying modeling and analysis approaches to aid in the advancement of medical care. One application includes aiding in physician monitoring of disease progression to inform clinical decision-making and management for complex diseases. For example, there are efforts to improve inpatient and at-home monitoring of vital signs[80–82], and obtain non-invasive proxies for metrics such as glucose[83] and inflammation status[84]. There are also increasing efforts to utilize digital biomarkers for precision medicine applications, such as in cancer and autoimmune diseases[85,86], to optimize therapies that account for disease complexity and heterogeneity. Furthermore, computational approaches are in development to manage the large data complexity of information acquired to derive scientific or medical insights via phenotyping[87] and predicting clinical or behavioral states[88,89].

Some digital health applications explored include incorporation of interactivity and feedback through patient-facing mobile applications. Mobile applications can aid in patient-centric care via patient education and treatment support, which is of particular importance for healthcare affordability and access. There has been an increase in the availability of apps for a variety of diseases, such as for vital sign monitoring, glucose monitoring for diabetes, weight management[90], mental health[91], and even for managing postpartum maternal health[92]. Informatics and artificial intelligence techniques can also be used to guide patients in management of their own care[93], such as determining optimal drug dosage or timing[94,95], or predicting risk and providing recommendations from surveys and inputted data points[96,97]. These translational applications have great opportunities for improving equity and inclusion in disease care, such as in aiding health management for those with disabilities[98], complex diseases[99,100], or in under-resourced locations[101].

Considering Equity: With the impetus that comes from the COVID-19 pandemic, technology and digital health are expected to continually become integrated into clinical care and utilized for scientific and clinical research[68]. This spans a wide range of data types and applications, ranging from public health analysis of mobile phones, networks, the internet, and GPS to individualized applications from both the clinical perspective (EMR, telehealth, medical devices) and from the patient perspective (wearables, mobile applications). There is therefore not a better time than now to talk about opportunities in equity. These opportunities include access, affordability, decreased time in the hospital, as well as early detection and prevention for public health goals[98]. With the maturation of digital health approaches, beyond issues regarding privacy and regulations, applications should account for technological literacy[102]; accessibility for culturally diverse populations[103–105], older people[106,107], and people with disabilities[98,103];

adaptability to rural environments[108]; various levels of health literacy[109,110]; and even access to fundamental tools and technology[111-114]. With these considerations in place, digital health can become an essential way to bring informatics into accessible and equitable translational applications.

1.1.5 Discussion

This review of the field highlighted innovative research post COVID-19 pandemic, blending computational progress with equity. The growing literature focuses on equity throughout the bioinformatics process, from data collection to interpretation. Recognizing bioethicists like Sandra Soo-Jin Lee's viewpoint, we acknowledged the ethical responsibilities when using biomedical data, including nurturing trust with underrepresented communities and individuals[28,29]. In the current era, we can acquire vast amounts of data, driving the surge of informatics and machine learning. These methods advance scientific knowledge, pinpoint therapeutic targets, support medical decisions, and foster patient-centric care. With evolving technology, there are opportunities to improve accessibility and health literacy. However, challenges persist in translational informatics. Data representation is a hurdle; more needs to be done to ensure equity in data collection. Technological literacy is also a barrier for many, impacting the efficacy of translational tools. Access to institutions and technology is foundational for inclusivity in all areas, from data to healthcare delivery[115-129].

The upcoming decade demands a focus on equity in data collection, analysis, model implementation, and application development. In summary, with the vast amounts of genomic, clinical, and digital health data available, computational methods offer immense potential to advance human health. Machine learning's transformative power is evident, especially in

predictive modeling. Integrating diverse data and prioritizing equity throughout research can pave the way for universal precision medicine.

Table 1.1.1 Keywords in the Search for Publications or Related Publications in Translational Bioinformatics

These terms were utilized in identifying relevant publications within informatics and health translation.

Informatics terms	Broad AI or algorithmic terms	Disease or data related terms	Clinical or digital health relevant terms
• Precision medicine • Translational bioinformatics • Translational informatics • Bioinformatics • Bias informatics • Multiomics • Omics • Drug repurposing	• Machine learning • Machine learning bias • Artificial intelligence • Predictive modelling • Clustering • Disease subtyping • Subphenotyping	• Biomarker discovery • Phenotyping • Microbiome • All of us research program • Digital biomarkers • Disease trajectories	• Electronic medical records • Electronic health records • Clinical trials • Clinical informatics • Digital health • Mobile health • Telehealth

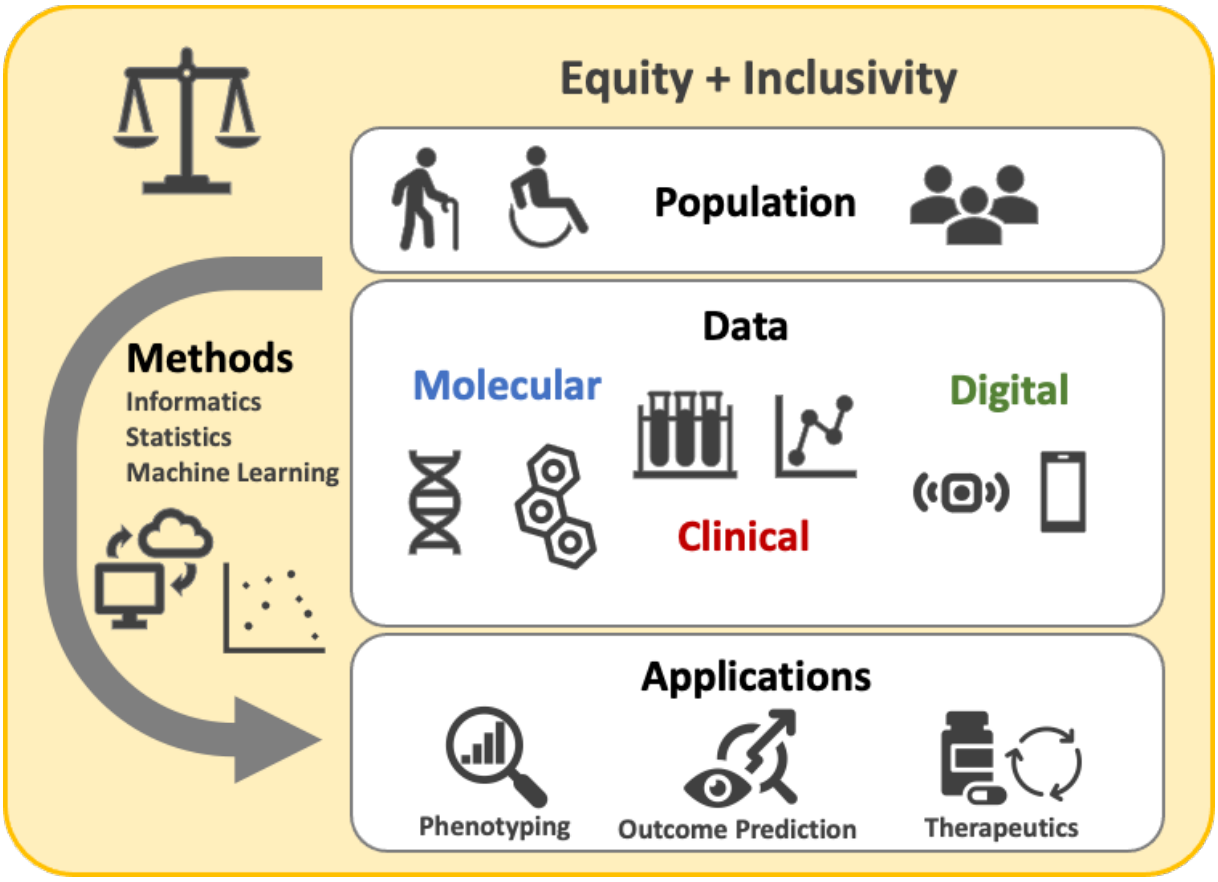


Figure 1.1.1 Translational Bioinformatics in the Era of Precision Medicine

Here we present recent translational bioinformatics approaches that leverage molecular, clinical, and digital data to advance precision medicine. We discuss specific applications such as phenotyping, outcome prediction, and therapeutics, as well as methods including informatics, statistics, and machine learning, all within the context of equity and inclusion.

1.1.6 References

1. Butte, A. J. Translational bioinformatics: coming of age. *J. Am. Med. Inform. Assoc. JAMIA* **15**, 709–714 (2008).
2. Nicholls, S. M. *et al.* CLIMB-COVID: continuous integration supporting decentralised sequencing for SARS-CoV-2 genomic surveillance. *Genome Biol.* **22**, 196, s13059-021-02395-y (2021).
3. Bennett, T. D. *et al.* Clinical Characterization and Prediction of Clinical Severity of SARS-CoV-2 Infection Among US Adults Using Data From the US National COVID Cohort Collaborative. *JAMA Netw. Open* **4**, e2116901 (2021).
4. Gunasekeran, D. V., Tseng, R. M. W. W., Tham, Y.-C. & Wong, T. Y. Applications of digital health for public health responses to COVID-19: a systematic scoping review of artificial intelligence, telehealth and related technologies. *Npj Digit. Med.* **4**, 40 (2021).
5. Maher, B. & Van Noorden, R. How the COVID pandemic is changing global science collaborations. *Nature* **594**, 316–319 (2021).
6. Luengo-Oroz, M. *et al.* Artificial intelligence cooperation to support the global response to COVID-19. *Nat. Mach. Intell.* **2**, 295–297 (2020).
7. Hartl, D. *et al.* Translational precision medicine: an industry perspective. *J. Transl. Med.* **19**, 245 (2021).
8. Leasure, A. C. & Cohen, J. M. Prevalence of eczema among adults in the United States: a cross-sectional study in the All of Us research program. *Arch. Dermatol. Res.* (2022) doi:10.1007/s00403-022-02328-0.
9. Acosta, J. N. *et al.* Cardiovascular Health Disparities in Racial and Other Underrepresented Groups: Initial Results From the All of Us Research Program. *J. Am. Heart Assoc.* **10**,

e021724 (2021).

10. Hull, L. E. & Natarajan, P. Self-rated family health history knowledge among All of Us program participants. *Genet. Med. Off. J. Am. Coll. Med. Genet.* S1098-3600(21)05465–4 (2022) doi:10.1016/j.gim.2021.12.006.
11. Delavar, A., Radha Saseendrakumar, B., Weinreb, R. N. & Baxter, S. L. Racial and Ethnic Disparities in Cost-Related Barriers to Medication Adherence Among Patients With Glaucoma Enrolled in the National Institutes of Health All of Us Research Program. *JAMA Ophthalmol.* (2022) doi:10.1001/jamaophthalmol.2022.0055.
12. Obermeyer, Z., Powers, B., Vogeli, C. & Mullainathan, S. Dissecting racial bias in an algorithm used to manage the health of populations. *Science* **366**, 447–453 (2019).
13. IBDMDB Investigators *et al.* Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases. *Nature* **569**, 655–662 (2019).
14. Mars, R. A. T. *et al.* Longitudinal Multi-omics Reveals Subset-Specific Mechanisms Underlying Irritable Bowel Syndrome. *Cell* **182**, 1460-1473.e17 (2020).
15. Tarca, A. L. *et al.* Crowdsourcing assessment of maternal blood multi-omics for predicting gestational age and preterm birth. *Cell Rep. Med.* **2**, 100323 (2021).
16. Liu, M. C. *et al.* Sensitive and specific multi-cancer detection and localization using methylation signatures in cell-free DNA. *Ann. Oncol.* **31**, 745–759 (2020).
17. Hogan, C. A. *et al.* Clinical Impact of Metagenomic Next-Generation Sequencing of Plasma Cell-Free DNA for the Diagnosis of Infectious Diseases: A Multicenter Retrospective Cohort Study. *Clin. Infect. Dis.* **72**, 239–245 (2021).
18. Taubes, A. *et al.* Experimental and real-world evidence supporting the computational repurposing of bumetanide for APOE4-related Alzheimer's disease. *Nat. Aging* **1**, 932–947

(2021).

19. Rodriguez, S. *et al.* Machine learning identifies candidates for drug repurposing in Alzheimer's disease. *Nat. Commun.* **12**, 1033 (2021).
20. Zeng, B. *et al.* OCTAD: an open workspace for *Nat. Protoc.* **16**, 728–753 (2021).
21. Beigel, J. H. *et al.* Remdesivir for the Treatment of Covid-19 — Final Report. *N. Engl. J. Med.* **383**, 1813–1826 (2020).
22. Morselli Gysi, D. *et al.* Network medicine framework for identifying drug-repurposing opportunities for COVID-19. *Proc. Natl. Acad. Sci.* **118**, e2025581118 (2021).
23. Green, E. D. *et al.* Strategic vision for improving human health at The Forefront of Genomics. *Nature* **586**, 683–692 (2020).
24. Martin, A. R. *et al.* Clinical use of current polygenic risk scores may exacerbate health disparities. *Nat. Genet.* **51**, 584–591 (2019).
25. Wojcik, G. L. *et al.* Genetic analyses of diverse populations improves discovery for complex traits. *Nature* **570**, 514–518 (2019).
26. The All of Us Research Program Investigators. The "All of Us" Research Program. *N. Engl. J. Med.* **381**, 668–676 (2019).
27. New Data Release: Filling Out the Pandemic Picture. (2021).
28. Fox, K. The Illusion of Inclusion — The "All of Us" Research Program and Indigenous Peoples' DNA. *N. Engl. J. Med.* **383**, 411–413 (2020).
29. Lee, S. S.-J. Obligations of the "Gift": Reciprocity and Responsibility in Precision Medicine. *Am. J. Bioeth.* **21**, 57–66 (2021).
30. Peterson, T. A., Fontil, V., Koliwad, S. K., Patel, A. & Butte, A. J. Quantifying Variation in Treatment Utilization for Type 2 Diabetes Across Five Major University of California Health

Systems. *Diabetes Care* **44**, 908–914 (2021).

31. Catherine, J. P., Russell, M. V. & Peter, C. H. The impact of race and socioeconomic factors on paediatric diabetes. *eClinicalMedicine* **42**, 101186 (2021).
32. Landi, I. *et al.* Deep representation learning of electronic health records to unlock patient stratification at scale. *Npj Digit. Med.* **3**, 96 (2020).
33. Alexander, N., Alexander, D. C., Barkhof, F. & Denaxas, S. Identifying and evaluating clinical subtypes of Alzheimer's disease in care electronic health records using unsupervised machine learning. *BMC Med. Inform. Decis. Mak.* **21**, 343 (2021).
34. Kung, B., Chiang, M., Perera, G., Pritchard, M. & Stewart, R. Identifying subtypes of depression in clinician-annotated text: a retrospective cohort study. *Sci. Rep.* **11**, 22426 (2021).
35. Pencina, M. J., Goldstein, B. A. & D'Agostino, R. B. Prediction Models — Development, Evaluation, and Clinical Application. *N. Engl. J. Med.* **382**, 1583–1586 (2020).
36. Rattsev, I., Flaks-Manov, N., Jelin, A. C., Bai, J. & Taylor, C. O. Recurrent preterm birth risk assessment for two delivery subtypes: A multivariable analysis. *J. Am. Med. Inform. Assoc.* ocab184 (2021) doi:10.1093/jamia/ocab184.
37. Feng, J., Lee, J., Vesoulis, Z. A. & Li, F. Predicting mortality risk for preterm infants using deep learning models with time-series vital sign data. *Npj Digit. Med.* **4**, 108 (2021).
38. Zhao, J. *et al.* Learning from Longitudinal Data in Electronic Health Record and Genetic Data to Improve Cardiovascular Event Prediction. *Sci. Rep.* **9**, 717 (2019).
39. Osborne, T. F., Veigulis, Z. P., Arreola, D. M., Röösli, E. & Curtin, C. M. Automated EHR score to predict COVID-19 outcomes at US Department of Veterans Affairs. *PLOS ONE* **15**, e0236554 (2020).

40. Lauritsen, S. M. *et al.* Explainable artificial intelligence model to predict acute critical illness from electronic health records. *Nat. Commun.* **11**, 3852 (2020).

41. Boland, M. R. *et al.* Harnessing electronic health records to study emerging environmental disasters: a proof of concept with perfluoroalkyl substances (PFAS). *Npj Digit. Med.* **4**, 122 (2021).

42. Song, X. *et al.* Cross-site transportability of an explainable artificial intelligence model for acute kidney injury prediction. *Nat. Commun.* **11**, 5668 (2020).

43. Tomašev, N. *et al.* Use of deep learning to develop continuous-risk models for adverse event prediction from electronic health records. *Nat. Protoc.* **16**, 2765–2787 (2021).

44. Ramachandran, A. *et al.* Predictive Analytics for Retention in Care in an Urban HIV Clinic. *Sci. Rep.* **10**, 6421 (2020).

45. Lee, C. K., Samad, M., Hofer, I., Cannesson, M. & Baldi, P. Development and validation of an interpretable neural network for prediction of postoperative in-hospital mortality. *Npj Digit. Med.* **4**, 8 (2021).

46. Wouters, O. J., McKee, M. & Luyten, J. Estimated Research and Development Investment Needed to Bring a New Medicine to Market, 2009-2018. *JAMA* **323**, 844 (2020).

47. Zhou, M., Zheng, C. & Xu, R. Combining phenome-driven drug-target interaction prediction with patients' electronic health records-based clinical corroboration toward drug discovery. *Bioinformatics* **36**, i436–i444 (2020).

48. Oskotsky, T. *et al.* Mortality Risk Among Patients With COVID-19 Prescribed Selective Serotonin Reuptake Inhibitor Antidepressants. *JAMA Netw. Open* **4**, e2133090 (2021).

49. MacLeod, A. R. *et al.* Personalised high tibial osteotomy has mechanical safety equivalent to generic device in a case–control in silico clinical trial. *Commun. Med.* **1**, 6 (2021).

50. Hong, J. C., Spiegel, D. Y., Havrilesky, L. J. & Chino, J. P. High-volume providers and brachytherapy practice: A Medicare provider utilization and payment analysis. *Brachytherapy* **17**, 906–911 (2018).

51. Braunlin, M., Belani, R., Buchanan, J., Wheeling, T. & Kim, C. Trends in the multiple myeloma treatment landscape and survival: a U.S. analysis using 2011–2019 oncology clinic electronic health record data. *Leuk. Lymphoma* **62**, 377–386 (2021).

52. Hong, J. C. *et al.* System for High-Intensity Evaluation During Radiation Therapy (SHIELD-RT): A Prospective Randomized Study of Machine Learning–Directed Clinical Evaluations During Radiation and Chemoradiation. *J. Clin. Oncol.* **38**, 3652–3661 (2020).

53. Escobar, G. J. *et al.* Automated Identification of Adults at Risk for In-Hospital Clinical Deterioration. *N. Engl. J. Med.* **383**, 1951–1960 (2020).

54. Muse, E. D. & Topol, E. J. Guiding ultrasound image capture with artificial intelligence. *The Lancet* **396**, 749 (2020).

55. Reeves, J. J. *et al.* Rapid response to COVID-19: health informatics support for outbreak management in an academic health system. *J. Am. Med. Inform. Assoc.* **27**, 853–859 (2020).

56. Adler-Milstein, J., Chen, J. H. & Dhaliwal, G. Next-Generation Artificial Intelligence for Diagnosis: From Predicting Diagnostic Labels to "Wayfinding". *JAMA* **326**, 2467 (2021).

57. Carcel, C. *et al.* Representation of Women in Stroke Clinical Trials: A Review of 281 Trials Involving More Than 500,000 Participants. *Neurology* **97**, e1768–e1774 (2021).

58. Unger, J. M. *et al.* Representativeness of Black Patients in Cancer Clinical Trials Sponsored by the National Cancer Institute Compared With Pharmaceutical Companies. *JNCI Cancer Spectr.* **4**, pkaa034 (2020).

59. Awad, E. *et al.* Minority participation in phase 1 gynecologic oncology clinical trials: Three

decades of inequity. *Gynecol. Oncol.* **157**, 729–732 (2020).

60. Trant, A. A. *et al.* Increasing accrual of minority patients in breast cancer clinical trials. *Breast Cancer Res. Treat.* **184**, 499–505 (2020).
61. Andrasik, M. P. *et al.* Increasing Black, Indigenous and People of Color participation in clinical trials through community engagement and recruitment goal establishment. *PLOS ONE* **16**, e0258858 (2021).
62. Jin, X. *et al.* Women's Participation in Cardiovascular Clinical Trials From 2010 to 2017. *Circulation* **141**, 540–548 (2020).
63. Mendis, S. R. *et al.* Female representation in clinical trials leading to FDA cancer drug approvals for gastrointestinal (GI) cancers between 2008 to 2018. *J. Clin. Oncol.* **38**, 809–809 (2020).
64. Martinkova, J. *et al.* Proportion of Women and Reporting of Outcomes by Sex in Clinical Trials for Alzheimer Disease: A Systematic Review and Meta-analysis. *JAMA Netw. Open* **4**, e2124124 (2021).
65. Baker, K. E., Streed, C. G. & Durso, L. E. Ensuring That LGBTQI+ People Count — Collecting Data on Sexual Orientation, Gender Identity, and Intersex Status. *N. Engl. J. Med.* **384**, 1184–1186 (2021).
66. Keuroghlian, A. S. Electronic health records as an equity tool for LGBTQIA+ people. *Nat. Med.* **27**, 2071–2073 (2021).
67. Vyas, D. A., Eisenstein, L. G. & Jones, D. S. Hidden in Plain Sight — Reconsidering the Use of Race Correction in Clinical Algorithms. *N. Engl. J. Med.* **383**, 874–882 (2020).
68. Ye, J. The Role of Health Technology and Informatics in a Global Public Health Emergency: Practices and Implications From the COVID-19 Pandemic. *JMIR Med. Inform.* **8**, e19866

(2020).

69. The Lancet Digital Health. Contact tracing: digital health on the frontline. *Lancet Digit. Health* **2**, e561 (2020).
70. Jewell, S. *et al. Npj Digit. Med.* **4**, 152 (2021).
71. Badr, H. S. *et al.* Association between mobility patterns and COVID-19 transmission in the USA: a mathematical modelling study. *Lancet Infect. Dis.* **20**, 1247–1254 (2020).
72. Mishra, T. *et al.* Pre-symptomatic detection of COVID-19 from smartwatch data. *Nat. Biomed. Eng.* **4**, 1208–1220 (2020).
73. Dash, T. K., Mishra, S., Panda, G. & Satapathy, S. C. Detection of COVID-19 from speech signal using bio-inspired based cepstral features. *Pattern Recognit.* **117**, 107999 (2021).
74. Verde, L., De Pietro, G. & Sannino, G. Artificial Intelligence Techniques for the Non-invasive Detection of COVID-19 Through the Analysis of Voice Signals. *Arab. J. Sci. Eng.* (2021) doi:10.1007/s13369-021-06041-4.
75. Rykov, Y., Thach, T.-Q., Bojic, I., Christopoulos, G. & Car, J. Digital Biomarkers for Depression Screening With Wearable Devices: Cross-sectional Study With Machine Learning Modeling. *JMIR MHealth UHealth* **9**, e24872 (2021).
76. Cavedoni, S., Chirico, A., Pedroli, E., Cipresso, P. & Riva, G. Digital Biomarkers for the Early Detection of Mild Cognitive Impairment: Artificial Intelligence Meets Virtual Reality. *Front. Hum. Neurosci.* **14**, 245 (2020).
77. the Parkinson's Disease Digital Biomarker Challenge Consortium *et al.* Crowdsourcing digital health measures to predict Parkinson's disease severity: the Parkinson's Disease Digital Biomarker DREAM Challenge. *Npj Digit. Med.* **4**, 53 (2021).
78. Robin, J. *et al.* Evaluation of Speech-Based Digital Biomarkers: Review and

Recommendations. *Digit. Biomark.* **4**, 99–108 (2020).

79. Park, C., Mishra, R., Golledge, J. & Najafi, B. Digital Biomarkers of Physical Frailty and Frailty Phenotypes Using Sensor-Based Physical Activity and Machine Learning. *Sensors* **21**, 5289 (2021).
80. Hamza, M. *et al.* Can vital signs recorded in patients' homes aid decision making in emergency care? A Scoping Review. *Resusc. Plus* **6**, 100116 (2021).
81. Mohammadzadeh, N., Gholamzadeh, M., Saeedi, S. & Rezayi, S. The application of wearable smart sensors for monitoring the vital signs of patients in epidemics: a systematic literature review. *J. Ambient Intell. Humaniz. Comput.* (2020) doi:10.1007/s12652-020-02656-x.
82. Leenen, J. P. L. *et al.* Current Evidence for Continuous Vital Signs Monitoring by Wearable Wireless Devices in Hospitalized Adults: Systematic Review. *J. Med. Internet Res.* **22**, e18636 (2020).
83. Wedlund, L. & Kvedar, J. Innovative new model predicts glucose levels without poking or prodding. *Npj Digit. Med.* **4**, 126, s41746-021-00501–9 (2021).
84. van den Brink, W. *et al.* Digital Resilience Biomarkers for Personalized Health Maintenance and Disease Prevention. *Front. Digit. Health* **2**, 614670 (2021).
85. Capobianco, E. & Meroni, P. L. Value of digital biomarkers in precision medicine: implications in cancer, autoimmune diseases, and COVID-19. *Expert Rev. Precis. Med. Drug Dev.* **6**, 235–238 (2021).
86. Solomon, D. H. & Rudin, R. S. Digital health technologies: opportunities and challenges in rheumatology. *Nat. Rev. Rheumatol.* **16**, 525–535 (2020).
87. Onnela, J.-P. Opportunities and challenges in the collection and analysis of digital phenotyping data. *Neuropsychopharmacology* **46**, 45–54 (2021).

88. Nahavandi, D., Alizadehsani, R., Khosravi, A. & Acharya, U. R. Application of artificial intelligence in wearable devices: Opportunities and challenges. *Comput. Methods Programs Biomed.* **213**, 106541 (2022).

89. Perez-Pozuelo, I., Spathis, D., Clifton, E. A. D. & Mascolo, C. Wearables, smartphones, and artificial intelligence for digital phenotyping and health. in *Digital Health* 33–54 (Elsevier, 2021). doi:10.1016/B978-0-12-820077-3.00003-1.

90. Lim, S. L. *et al.* Effect of a Smartphone App on Weight Change and Metabolic Outcomes in Asian Adults With Type 2 Diabetes: A Randomized Clinical Trial. *JAMA Netw. Open* **4**, e2112417 (2021).

91. Lau, N., O'Daffer, A., Yi-Frazier, J. P. & Rosenberg, A. R. Popular Evidence-Based Commercial Mental Health Apps: Analysis of Engagement, Functionality, Aesthetics, and Information Quality. *JMIR MHealth UHealth* **9**, e29689 (2021).

92. Tucker, L., Villagomez, A. C. & Krishnamurti, T. Comprehensively addressing postpartum maternal health: a content and image review of commercially available mobile health apps. *BMC Pregnancy Childbirth* **21**, 311 (2021).

93. Khoong, E. C. *et al.* Mobile health strategies for blood pressure self-management in urban populations with digital barriers: systematic review and meta-analyses. *Npj Digit. Med.* **4**, 114 (2021).

94. Sharma, R., Singh, D., Gaur, P. & Joshi, D. Intelligent automated drug administration and therapy: future of healthcare. *Drug Deliv. Transl. Res.* **11**, 1878–1902 (2021).

95. Eckardt, J.-N., Wendt, K., Bornhäuser, M. & Middeke, J. M. Reinforcement Learning for Precision Oncology. *Cancers* **13**, 4624 (2021).

96. Domin, A., Spruijt-Metz, D., Theisen, D., Ouzzahra, Y. & Vögele, C. Smartphone-Based

Interventions for Physical Activity Promotion: Scoping Review of the Evidence Over the Last 10 Years. *JMIR MHealth UHealth* **9**, e24308 (2021).

97. khan, Z. F. & Alotaibi, S. R. Applications of Artificial Intelligence and Big Data Analytics in m-Health: A Healthcare System Perspective. *J. Healthc. Eng.* **2020**, 1–15 (2020).
98. Jones, M., DeRuyter, F. & Morris, J. The Digital Health Revolution and People with Disabilities: Perspective from the United States. *Int. J. Environ. Res. Public. Health* **17**, 381 (2020).
99. Busse, M. *et al.* Web-based physical activity intervention for people with progressive multiple sclerosis: application of consensus-based intervention development guidance. *BMJ Open* **11**, e045378 (2021).
100. de Batlle, J. *et al.* Implementing Mobile Health–Enabled Integrated Care for Complex Chronic Patients: Intervention Effectiveness and Cost-Effectiveness Study. *JMIR MHealth UHealth* **9**, e22135 (2021).
101. Menictas, M., Rabbi, M., Klasnja, P. & Murphy, S. Artificial intelligence decision-making in mobile health. *The Biochemist* **41**, 20–24 (2019).
102. Smith, B. & Magnani, J. W. *Int. J. Cardiol.* **292**, 280–282 (2019).
103. Noel, K. & Ellison, B. Inclusive innovation in telehealth. *Npj Digit. Med.* **3**, 89 (2020).
104. Brewer, L. C. *et al.* Back to the Future: Achieving Health Equity Through Health Informatics and Digital Health. *JMIR MHealth UHealth* **8**, e14512 (2020).
105. Marwaha, J. S. & Kvedar, J. C. Cultural adaptation: a framework for addressing an often-overlooked dimension of digital health accessibility. *Npj Digit. Med.* **4**, 143, s41746-021-00516–2 (2021).
106. Yoon, H., Jang, Y., Vaughan, P. W. & Garcia, M. Older Adults' Internet Use for Health

Information: Digital Divide by Race/Ethnicity and Socioeconomic Status. *J. Appl. Gerontol. Off. J. South. Gerontol. Soc.* **39**, 105–110 (2020).

107. Liu, N., Yin, J., Tan, S. S.-L., Ngiam, K. Y. & Teo, H. H. Mobile health applications for older adults: a systematic review of interface and persuasive feature design. *J. Am. Med. Inform. Assoc.* **28**, 2483–2501 (2021).
108. Hilty, D. M. *et al.* Telehealth for rural diverse populations: telebehavioral and cultural competencies, clinical outcomes and administrative approaches. *mHealth* **6**, 20–20 (2020).
109. Knitza, J. *et al.* Mobile Health Usage, Preferences, Barriers, and eHealth Literacy in Rheumatology: Patient Survey Study. *JMIR MHealth UHealth* **8**, e19661 (2020).
110. Crawford, A. & Serhal, E. Digital Health Equity and COVID-19: The Innovation Curve Cannot Reinforce the Social Gradient of Health. *J. Med. Internet Res.* **22**, e19361 (2020).
111. Hoffman, D. A. Increasing access to care: telehealth during COVID-19. *J. Law Biosci.* **7**, lsaa043 (2020).
112. Yi, S. S. *et al.* With no data, *EClinicalMedicine* **41**, 101165 (2021).
113. Bakken, S. Replication studies and diversity, equity, and inclusion strategies are critical to advance the impact of biomedical and health informatics. *J. Am. Med. Inform. Assoc.* **28**, 1813–1814 (2021).
114. Sieck, C. J. *et al.* Digital inclusion as a social determinant of health. *Npj Digit. Med.* **4**, 52 (2021).
115. Chen, I. Y. *et al.* Ethical Machine Learning in Healthcare. *Annu. Rev. Biomed. Data Sci.* **4**, 123–144 (2021).
116. Mhasawade, V., Zhao, Y. & Chunara, R. Machine learning and algorithmic fairness in public and population health. *Nat. Mach. Intell.* **3**, 659–666 (2021).

117. Lett, E., Asabor, E., Beltrán, S., Michelle Cannon, A. & Arah, O. A. Conceptualizing, Contextualizing, and Operationalizing Race in Quantitative Health Sciences Research. *Ann. Fam. Med.* 2792 (2022) doi:10.1370/afm.2792.

118. Miron, M., Tolan, S., Gómez, E. & Castillo, C. Addressing multiple metrics of group fairness in data-driven decision making. *ArXiv200304794 Cs Stat* (2020).

119. Park, Y. *et al.* Comparison of Methods to Reduce Bias From Clinical Prediction Models of Postpartum Depression. *JAMA Netw. Open* **4**, e213909 (2021).

120. Wawira Gichoya, J., McCoy, L. G., Celi, L. A. & Ghassemi, M. Equity in essence: a call for operationalising fairness in machine learning for healthcare. *BMJ Health Care Inform.* **28**, e100289 (2021).

121. Leslie, D., Mazumder, A., Peppin, A., Wolters, M. K. & Hagerty, A. Does "AI" stand for augmenting inequality in the era of covid-19 healthcare? *BMJ* n304 (2021) doi:10.1136/bmj.n304.

122. Baxter, M. S., White, A., Lahti, M., Murto, T. & Evans, J. Machine learning in a time of COVID-19 - Can machine learning support Community Health Workers (CHWs) in low and middle income countries (LMICs) in the new normal? *J. Glob. Health* **11**, 03017 (2021).

123. Syeda, H. B. *et al.* Role of Machine Learning Techniques to Tackle the COVID-19 Crisis: Systematic Review. *JMIR Med. Inform.* **9**, e23811 (2021).

124. Ostaszewski, M. *et al.* COVID19 Disease Map, a computational knowledge repository of virus-host interaction mechanisms. *Mol. Syst. Biol.* **17**, e10387 (2021).

125. Dunn, P. & Hazzard, E. Technology approaches to digital health literacy. *Int. J. Cardiol.* **293**, 294–296 (2019).

126. Kemp, E. *et al.* Health literacy, digital health literacy and the implementation of digital health

technologies in cancer care: the need for a strategic approach. *Health Promot. J. Austr.* **32**, 104–114 (2021).

127. Kuek, A. & Hakkennes, S. Healthcare staff digital literacy levels and their attitudes towards information systems. *Health Informatics J.* **26**, 592–612 (2020).

128. Zhang, X. *et al.* Role of Health Information Technology in Addressing Health Disparities: Patient, Clinician, and System Perspectives. *Med. Care* **57**, S115–S120 (2019).

129. Triana, A. J., Gusdorf, R. E., Shah, K. P. & Horst, S. N. Technology Literacy as a Barrier to Telehealth During COVID-19. *Telemed. E-Health* **26**, 1118–1119 (2020).

1.2 Personalizing routine laboratory measurements from electronic health records with machine learning

1.2.1 Abstract

Machine learning applied to electronic medical records can be used to create personalized lab test reference ranges and to quantify disease risk, paving the way for precision medicine in clinical care.

1.2.2 News and Views

Precision Medicine is "an emerging integrative approach for disease prevention, early detection and treatment that takes into account individual variability in genes, environment, and lifestyle."[1] For instance, testing for genetic variants in a person's tumor sample is being increasingly performed as a part of diagnosing malignancies and determining therapeutic options, and is becoming the standard of care for some cancers[2,3]. This targeted approach to clinical care is enabled in part by basic research discoveries and is fueled by a growing volume of molecular, clinical, and epidemiological data. Electronic medical records (EMRs) provide an invaluable source of data for biomedical research and opportunity for precision medicine strategies, including the use of EMRs for personalized lab test modeling[4] (**Figure 1.2.1**).

EMRs capture clinical data on the population of patients, including demographics, diagnosis codes, medication orders and laboratory tests, which results in billions of data points on millions of patients. Even though EMR data are collected for individual patient-care purposes, there is an opportunity for de-identifying the data and, together with advanced computational approaches, leveraging it for clinical and translational research.

Integrative computational methods have become a valuable tool for turning various types of biomedical data into clinically actionable information. Machine learning is a branch of artificial intelligence based on the idea that systems can learn from data, identify patterns, and make decisions with minimal human intervention. These predictive modeling approaches have been successful in many fields, including computer vision and natural language processing, and have been extensively applied in the biomedical domain [5,6]. Bringing together rich clinical phenotyping in the EMR with advanced machine-learning techniques provides an incredible opportunity for advancing precision medicine in the clinic.

Currently, since clinicians rely on 'hard-coded' reference ranges for laboratory results, automated ways to use machine learning to identify personalized reference ranges are desired[4], especially since ranges can differ by age, sex, ethnicity, disease status and other relevant characteristics. A recent study has provided support for this rationale by demonstrating that the distributions of over 50% of laboratory tests with currently fixed reference intervals differ among healthy people, according to self-identified racial and ethnic groups[7]. Although computational analysis has been applied extensively to diagnosis codes[8,9] and medications in the EMR[10,11], relatively few studies have analyzed laboratory values. In Cohen et al., the authors made use of the rich EMR dataset from Israel's Clalit healthcare system to model 2.1 billion lab measurements in 2.8 million healthy adults and modeled the trajectories of 92 different lab tests[4]. They demonstrated the use of these models for personalized clinical applications, such as the creation of patient-specific reference lab-value ranges and showed impressive performance in the prediction of future test results and risks. This approach demonstrates the potential of using artificial intelligence on EMR data to personalize the interpretation of clinical results, to better quantify patient risk and to support clinical decision-making. Interpreting and acting upon lab data

has been an art form practiced by healthcare providers, but now precision-medicine approaches can help to increase the accuracy and reliability of that process.

To identify lab-test results that corresponded to healthy states, an unsupervised approach was utilized to filter out those associated with 131 chronic and acute conditions and 5,223 drug–test pairs that showed significant correlation. This generated a list of ~0.5 billion lab measurements from 2.8 million people and sufficient data for multivariate modeling across an age range of 20–90 years. These models were shown to better represent lab-test ranges for healthy people, performing better than the absolute normal ranges now used in clinical practice. They were able to consider variation due to sex-specific trends (e.g., changes in values for red-blood-cell distribution width due to menopause) or transient age-linked trends (e.g., peaks in alanine aminotransferase for males in their 30s) that absolute ranges fail to capture. These models can be applied in the clinic to generate a personalized quantitative reference range, similar to the utilization of the body-mass-index scale and percentile charts that pediatricians use to assess growth progress in children.

Looking at median lab-test values among healthy patients, age and sex were also contributors to less than 10% of within-norm variance in most tests, whereas personalized histories explained over 45% of variance in over half of the tests[4]. Predictive modeling was performed to account for this variance and these trained models could be used for personalized medicine, such as stratifying patients based on 2- to 3-year prediction of future lab abnormalities, mortality risk or disease risk. To demonstrate model utility, personalized risk stratification was performed on people who had normal lab-test values, with the goal of predicting future lab-test abnormalities or mortality. When used in a real-world setting, this type of modeling will allow healthcare systems to interpret tests and stratify patients by risk, years before a potential disease may manifest. In

another demonstration, temporal modeling of sparse patient histories allowed for quantification of the risk for developing diabetes, renal failure or colorectal cancer[4]. These models can be extended to any severe or insidious disease, which will allow more-precise early preventative measures in the clinic based on the risk of future disease.

Although there is considerable potential for EMR analysis, there are also limitations that need to be overcome, such as missing data, sampling bias, provider bias, reporting errors and unreported factors, including over-the-counter medication use. Nevertheless, approaches such as semi-automated strategy via unsupervised filtering has been utilized to model and replace the missing values, which has been shown to not negatively affect model performance[4]. Trained models could be extended in many potential future directions, some of which address the limitations mentioned above. Models should be further tested and applied to other EMR systems across a more diverse patient population to model 'normal' lab-test trajectories, which might also help alleviate the limitations related to bias. Inclusion of additional covariates can also help to further personalize the interpretation of lab values, even those within normal range. In addition, there is enormous potential for extension of this methodology beyond laboratory tests, to inclusion of medications, diagnoses, and longitudinal analysis. Finally, applications of this approach beyond the 'healthy' population will allow immense potential for interpreting laboratory ranges in diseased populations, which will potentially allow more-personalized interpretation and monitoring of patients' lab tests in this context.

Given the extensive clinical databases that exist across many countries and institutions, there is an incredible opportunity for paving the way forward in making best use of big health data. Applying modeling approaches to large-scale health data not only allows scientific progress through interpretability and observation, but also can bring actionable insights into the clinic to

improve personalized healthcare around the world. A stratified analysis according to personal covariates (e.g., age, sex, race, and other characteristics) can allow for better modeling of the complex extensive data within EMRs. Then, through the application of computational modeling and machine learning, results such as personalized lab test ranges, disease risk scores, and targeted therapeutics can be determined for each patient based on patient-specific covariates and medical history.

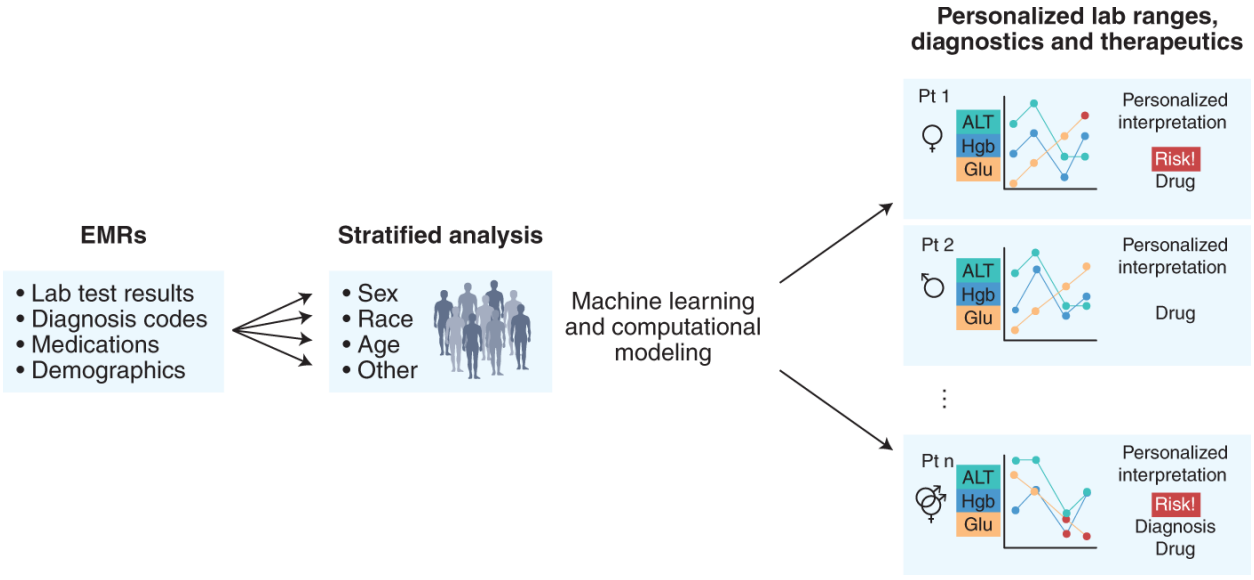


Figure 1.2.1 General workflow for modeling patient EMR data for personalized medicine

Information such as lab test results, diagnostic codes, medications, and other relevant demographic information or covariates can be extracted from electronic medical records. By stratifying lab results as outcomes to models and including demographic and relevant diagnostic or medication covariates in machine learning models, personalized lab ranges can be determined based on a patient's individual characteristics and health profile. These personal ranges can be utilized to identify potential diagnostic risks and therapeutic approaches.

1.2.3 References

1. MedlinePlus Genetics. *What is precision medicine?* https://medlineplus.gov/genetics/understanding/precisionmedicine/definition/.
2. Berger, M. F. & Mardis, E. R. The emerging clinical relevance of genomics in cancer medicine. *Nat. Rev. Clin. Oncol.* **15**, 353–365 (2018).
3. Denny, J. C. & Collins, F. S. Precision medicine in 2030—seven ways to transform healthcare. *Cell* **184**, 1415–1419 (2021).
4. Cohen, N. M. *et al.* Personalized lab test models to quantify disease potentials in healthy individuals. *Nat. Med.* **27**, 1582–1591 (2021).
5. Tarca, A. L., Carey, V. J., Chen, X., Romero, R. & Drăghici, S. Machine Learning and Its Applications to Biology. *PLoS Comput. Biol.* **3**, e116 (2007).
6. Wainberg, M., Merico, D., Delong, A. & Frey, B. J. Deep learning in biomedicine. *Nat. Biotechnol.* **36**, 829–838 (2018).
7. Rappoport, N. *et al.* Comparing Ethnicity-Specific Reference Intervals for Clinical Laboratory Tests from EHR Data. *J. Appl. Lab. Med.* **3**, 366–377 (2018
8. Li, L. *et al.* Identification of type 2 diabetes subgroups through topological analysis of patient similarity. *Sci. Transl. Med.* **7**, 311ra174 (2015).
9. Glicksberg, B. S. *et al.* Comparative analyses of population-scale phenomic data in electronic medical records reveal race-specific disease networks. *Bioinformatics* **32**, i101–i110 (2016).
10. Marić, I. *et al.* Decreased Mortality Rate Among COVID-19 Patients Prescribed Statins: Data From Electronic Health Records in the US. *Front. Med.* **8**, 639804 (2021).

11. Yee, S. W. *et al.* Drugs in COVID-19 Clinical Trials: Predicting Transporter-Mediated Drug-Drug Interactions Using In Vitro Assays and Real-World Data. *Clin. Pharmacol. Ther.* **110**, 108–122 (2021).

Chapter 2: Clinical Informatics Enables Deep Phenotyping and Discovery of Sex-Specific Differences in Alzheimer's Disease

2.1 Abstract

Alzheimer's Disease (AD) is a devastating disorder that is still not fully understood. Sex modifies AD vulnerability, but the reasons for this are largely unknown. We utilize two independent electronic medical record (EMR) systems across 44,288 patients to perform deep clinical phenotyping and network analysis to gain insight into clinical characteristics and sex-specific clinical associations in AD. Embeddings and network representation of patient diagnoses demonstrate greater comorbidity interactions in AD in comparison to matched controls. Enrichment analysis identified multiple known and new diagnostic, medication, and lab result associations across the whole cohort and in a sex-stratified analysis. With this data-driven method of phenotyping, we can represent AD complexity and generate hypotheses of clinical factors that can be followed-up for further diagnostic and predictive analyses, mechanistic understanding, or drug repurposing and therapeutic approaches.

2.2 Introduction

Alzheimer's Disease (AD) is the most common cause of dementia, making up 60-80% of cases, with a devastating and increasing burden on patients, caregivers, and society[1]. AD is characterized by brain atrophy and accumulation of beta-amyloid plaques and tau tangles seen on brain pathology after death. The disease erodes memory and cognitive functions, causing interference with daily activities and contributing to great emotional, social, and economic burden on patients and their families. AD is incurable and challenging to understand and diagnose. One reason AD is difficult to study is because it is a complex, heterogeneous, and multifactorial disease that takes many years to manifest[2]. This complexity, along with the slow insidious progression of the disease, makes it difficult to fully characterize disease phenotypes and associations.

Sex is one factor that has been shown to be important in AD, with higher prevalence in women afflicted by the disease at a 2:1 ratio compared to men[1]. While women have an increased estimated lifetime risk of AD, there is mixed evidence of risk between men and women of the same age[3,4]. Recent findings show that sex contributes to differing vulnerabilities or resilience to AD, as men with AD progress to death quicker[5,6] while women with this disease show higher cognitive resilience despite increased tau pathology[5,7,8]. How sex contributes to these differences in prevalence and vulnerability is a question of fervent interest among researchers in the AD field[9]. Recent studies in mice demonstrate that a second X chromosome may contribute to AD resilience[6]. Further sex-specific human studies in Alzheimer's Disease also show sex modification of AD risk[10], progression[11], and molecular phenotype[11–15]. As such, sex is a crucial factor to consider in studying and phenotyping AD.

While many efforts have evaluated the association of individual risk factors with AD, unbiased approaches to these associations are limited. Prior work, largely hypothesis-driven,

focused on select comorbidities associated with AD, such as hypertension[16], vascular disorders[17], diabetes[18], obesity[19], and others[20–22]. However, how sex modulates AD complexity and heterogeneity has still not been fully explored. Prior big data approaches to AD have examined genotype-phenotype associations[23,24] and molecular analyses[14,25–27] to characterize AD and sex differences[12,13]. Other work on phenotyping AD patients using clinical data has examined neuroimaging[28], neuropsychiatric phenotype[29], chart reviews[30], and billing records independently. Thus, an unbiased comprehensive approach to phenotype AD and identify sex associations using full clinical records is needed.

With the rise in electronic medical record (EMR) use over the past decade[31], there is abundant underutilized clinical data on patients covering comorbidities, medications, and lab values. This type of dataset provides a great opportunity to deeply investigate diseases and identify associations to facilitate understanding disease prevention and progression. Recently, EMR has been utilized for other diseases for creating comorbidity networks[32], identifying disease subtypes[33] and predicting disease outcomes[34,35] highlighting the potential of utilizing EMR data to extract insight and utility for complex and heterogeneous diseases[36], but a big data integrative analysis with EMR data has not yet been applied to characterize AD.

Deep phenotyping is a data-driven approach that has been used to provide more detailed stratification and representation of a disease in the era of precision medicine[37,38]. Here, we take an integrative approach through deep clinical phenotyping and network analysis to provide insight into AD clinical characteristics with a focus on sex differences. For the first time to our awareness, integrative phenotyping and association analysis is used to identify, in an unbiased manner, unique clinical features associated with AD itself - and reveals potential previously unknown sex-specific associations in the context of diagnoses, medications and lab test results.

2.3 Results

From the UCSF EMR database (~5 million patients), we identified 8,804 AD patients (5,558 female, 86.5 mean age (6.4 standard deviation)) and 17,608 propensity score (PS)-matched control patients (11,117 females, 86.5 mean age (6.4 standard deviation)). From the Mount Sinai EMR (~4 million patients), 5,958 AD patients (4,138 females, 88.3 mean age (8.7 standard deviation)) and 11,916 PS-matched controls (8,446 females, 88.7 mean age (11.4 standard deviation)) were identified. Post-matching analysis demonstrated adequate balance in covariates with standardized mean differences in age and categorical distributions below 0.1 (or below 0.2 between matched sex groups). Demographic characteristics of AD and matched control patients are shown in **Table 2.1** and **Supplementary Table 2.1**.

2.3.1 Embedding with Diagnosis Shows Separation Between AD and Controls

Due to the size of our cohort, we first performed low-dimensional visualizations using diagnoses as features to visualize patient separation. Low-dimensional UMAP visualizations of non-AD diagnoses (47,439 features, ICD-10-CM codes) show that distributions for AD and control patients are significantly different among the first two UMAP components (Mann-Whitney U-Test, p-value < 1e-5, **Figure 2.2A,B**) at both UCSF and Mount Sinai, with a progressive separation between groups. For the UCSF data, sex and death status show significant correlations with the first component, while age is significantly correlated with both components (Mann-Whitney U-Test p-value < 0.01, **Figure 2.2A**, **Supplementary Figure 2.1**). Sex, death status, and age are significantly correlated with both components at Mount Sinai (Mann-Whitney U-Test p-value < 0.01, **Figure 2.2B**, **Supplementary Figure 2.1**).

2.3.2 Alzheimer vs. Control Association Analysis Identifies Previously Known and Novel Associated Comorbidities in AD

Among each diagnostic hierarchical level (Level 2 categories, Level 3 categories, and full diagnosis names), AD disease networks contain more nodes and edges compared with control networks (**Supplementary Table 2.3**). In UCSF Level 3 diagnosis networks, more nodes and edges occur in AD vs control networks. As shown in Figure 3A, when thresholding Level 3 diagnosis categories by >5% of patients, there are 243 diagnosis pairs among AD patients compared to one pair in controls. When comparing node-level network metrics between groups, thresholded by >1% of patients within a group, AD and control networks are significantly different (Mann-Whitney U-Test, p-value < 0.01) when compared on average shortest path length, closeness centrality, neighborhood connectivity, and stress centrality, indicating a higher degree of connectivity among AD networks across all levels (**Figure 2.3C**). In Mount Sinai Level 3 diagnostic networks, more nodes and edges occur in AD networks compared to control networks, with significantly different distributions across AD and control networks on degree, eccentricity, neighborhood connectivity, and topological coefficient (Mann-Whitney U-Test, p-value < 0.01, **Supplementary Table 2.3**). Across the board, network metrics normalized by the metric are significantly correlated between UCSF and Mount Sinai (Spearman $\rho = 0.55$, p-value < 1e-4, **Figure 2.3E**)

Within Level 2 diagnosis categories, there were 166 significant diagnosis categories (Fisher's exact test, Bonferroni-corrected p-value < 0.05), with 120 diagnosis categories significantly enriched (OR > 2) uniquely in the AD group and no significantly enriched diagnosis categories uniquely in the control group (**Figure 2.4A top**). Within Level 3 diagnosis categories, there are 501 significant categories, with 391 and 4 categories significantly enriched in AD and

control groups, respectively (Fisher Exact test, Bonferroni-corrected p-value < 0.05, **Supplementary Data 2.1**). Within full diagnosis names, there are 1,627 significant diagnoses, with 1,491 and 7 diagnoses enriched uniquely in AD and control groups, respectively. Top significant diagnoses in AD include vascular dementia, hypertension, hyperlipidemia, urinary tract infection, syncope, hypothyroidism, and osteoporosis, while top significant diagnoses in controls include neoplasms of liver and brain (Fisher Exact test, Bonferroni-corrected p-value < 0.05, **Figure 2.4A bottom**, **Supplementary Data 2.1**). Top ICD diagnostic blocks in AD include mental health and behavioral diseases, genitourinary diseases, endocrine and metabolic diseases, and circulatory system diseases (**Figure 2.4B**). In the validation cohort, 1,495 of 1,627 significant UCSF diagnoses mapped to Mount Sinai EMR codes, of which 889 (60.13%) are significant (Fisher's exact test, Bonferroni p-value < 0.05). Overall comorbidity odds ratios at UCSF are significantly correlated with those of the validation cohort at Mount Sinai (Spearman $\rho = 0.65$, p-value < 1e-5, **Figure 2.4C**).

2.3.3 Sex-Stratified AD vs. Control Association Analysis Identifies Vascular and Musculoskeletal Disorders in Female AD & Behavioral/Neurological Disorders in Male AD

When stratifying diagnoses by sex (see **Methods**), AD disease networks are significantly different on metrics of betweenness centrality and neighborhood connectivity in both males and females compared to their respective controls among all diagnostic hierarchical levels (p-value < 0.001). Networks were significantly different in stress centrality among all diagnostic hierarchical levels when comparing AD males to AD females, but not when comparing control males to control females. Comparison of sex-specific network for diagnosis name show significantly greater closeness centrality, greater neighborhood connectivity, and lower eccentricity in female networks (Mann-Whitney U-Test, p-value<0.01 all three metrics, Figure 3D, **Supplementary Table 2.4**).

Within the validation cohort, similarly, female AD networks show significantly greater neighborhood connectivity compared to male AD networks (Mann-Whitney U-Test, p-value<0.01, **Supplementary Table 2.4**). When thresholding full diagnosis names by >5% of patients within a sex group, female AD patients have 45 shared co-diagnosis pairs compared to 14 in male AD patients (**Figure 2.3B**), and no pairs were identified for either control sex group.

For both males and females, there are 136, 338, and 714 shared significant diagnostic categories or diagnoses for Level 2, Level 3, and full diagnosis names, respectively. In a sex-stratified analysis, there were 29, 164, and 699 female-only significant hits and 5, 18, and 91 male-only significant hits for Level 2, Level 3, and full diagnosis names (Fisher Exact test, Bonferroni-corrected p-value < 0.05, **Figure 2.5A**, **Supplementary Data 2.1**). Compared to males among Level 2 diagnostic categories, females have a greater percent of significant diagnoses in blood-related disorders (e.g., nutritional anemia, coagulation defects) and congenital disorders and also have greater enrichment of pervasive and specific developmental disorders, musculoskeletal disorders (e.g. chondropathies, other osteopathies), injuries (e.g. injuries to the hip and thigh, injuries to the ankle and foot), infections with a predominantly sexual mode of transmission, and metabolic disorders (**Supplementary Data 2.1**). When comparing Level 2 categories in the validation cohort, among females, 153 out of 165 mapped with 60 (30.22%) significant, and among males, 133 out of 141 mapped with 64 (48.12%) significant (Fisher Exact test, Bonferroni-corrected p-value < 0.05 based on number of significant UCSF diagnoses). In general, Level 2 category sex-specific odds ratios are correlated between institutions (Females: Spearman $\rho = 0.77$, p-value < 1e-5; Males: Spearman $\rho = 0.83$, p-value < 1e-5). In the validation cohort, females have similar enrichment of blood-related disorders (e.g., nutritional anemia) and injuries (e.g., injuries to the hip and thigh), while males have enrichment of behavioral/emotional disorders.

Within full diagnosis names, unique significant diagnoses of female AD patients include asthma, atrial fibrillation, arthritis, fractures, and accidents while unique significant diagnoses of male AD patients include parkinsonism, sleep apnea, hypersomnia, neuropathy, irritability, and imbalance (Fisher Exact test, Bonferroni-corrected p-value < 0.05, Figure 5A, Figure 5B, **Supplementary Table 2.4**). Among full diagnosis names significant in both males and females, female AD patients have greater association in depression, hypertension, hyperlipidemia, urinary tract infections, upper respiratory infections, anemia, osteoporosis, and pneumonia, while male AD patients have greater effect size with behavioral phenotypes, hearing loss, and agitation (**Supplementary Data 2.1**). Among the full diagnosis names in the validation cohort, for females, 1,149 out of 1,383 significant diagnoses mapped, of which 240 (20.89%) were significant, and for males, 702 out of 805 significant diagnoses mapped, of which 216 (30.77%) were significant. In general, sex-specific diagnosis odds ratios were correlated for both females (Spearman ρ= 0.77, p-value < 1e-4) and males (Spearman ρ=0.83, p-value < 1e-4, **Figure 2.5C**). In the validation cohort, similarly, female AD patients have greater association in depression, hypertension, and osteoporosis while male AD patients have greater association in hearing loss and agitation (**Supplementary Data 2.1**).

2.3.4 Few Comorbidities Change with Sensitivity Analysis Taking Encounters Into Account

For our sensitivity analysis that included only patients with ≥ 10 encounters and records in EMR spanning > 1 year, there were 6,612 AD patients (2,382 males, 4,223 females) and 13,224 control patients (4,674 males, 8,539 females) identified by PS-matching on the number and timespan of encounters in addition to demographic characteristics and death status. A summary of the demographic characteristics of these cohorts are shown in **Supplementary Table 2.1.** We identified 100, 222, and 561 significant level 2, level 3, and full diagnosis names respectively

(Fisher Exact test, Bonferroni-corrected p-value threshold of 0.05), and an increase in odds ratio for chromosomal abnormalities and cerebrovascular disorders in AD patients (**Supplementary Table 5**). With sex-stratified enrichment analysis, encounter controlling increased enrichment of cerebrovascular disease in females, and increased significant enrichment of behavioral disorders, vision problems, and vascular dementia in males (**Supplementary Data 2.2**).

2.3.5 Visualization of Enriched Comorbidities via Rshiny App

An interactive visualization of **Figure 2.3** and **Figure 2.4** are made available in an RShiny app vizad.org.

2.3.6 Medication Association Analysis Identifies Dexamethasone as Enriched in Controls

In addition to comorbidities, we performed medication enrichment analysis in order to phenotype patients and investigate drugs enriched in AD patients and controls. Medications found enriched (Fisher Exact test, Bonferroni-corrected p-value < 0.05, OR >2 or < .5) in AD patients include current treatments like donepezil and memantine, but also vitamin B12, antidepressants (escitalopram, citalopram, sertraline, mirtazapine, trazodone), antipsychotics (quetiapine, risperidone, olanzapine), carbidopa/levodopa, vitamin D3, and melatonin. Medications found enriched in control patients include dexamethasone, ondansetron, and alteplase. Significant medications in controls with lesser effect size (Fisher Exact test, Bonferroni-corrected p-value < 0.05, 0.5 < OR < 1) include midazolam, propofol, opioids (oxycodone, fentanyl citrate), and furosemide (**Figure 2.6A**). From the validation cohort, 116 out of 121 medications mapped, of which 66 (56.90%) were significant (Fisher Exact test, Bonferroni-corrected p-value < 0.05 based upon significant medications at UCSF). In general, odds ratios of medications are significantly correlated (Spearman $\rho = 0.85$, p-value < 1e-4, **Figure 2.6C**). Dexamethasone is significant among

controls in both institutions, and multiple medications including vitamin B12, antidepressants, and antipsychotics are significant in AD patients among both institutions.

In a sex-stratified analysis, medications enriched in AD males include Tdap vaccine, melatonin, and carbidopa/levodopa while methylprednisolone and phenylephrine are enriched in control males. Female AD patients have enrichments in diazepam, antipsychotics (risperidone, aripiprazole), buspirone, antidepressants (sertraline, mirtazapine, trazodone, bupropion), vitamin D2, and levothyroxine while control females are enriched in norepinephrine bitartrate and fentanyl citrate (**Figure 2.6B**). In the validation EMR, 18 of 23 (78.25%) significant medications found at UCSF are significant in females at Mount Sinai, and 13 of 16 (81.25%) in males (Fisher Exact test, Bonferroni-corrected p-value < 0.05 based upon significant medications at UCSF within a group). Overall, there is significant correlation of sex-specific medication odds ratios in females (Spearman $\rho = 0.7$, p-value = .001) and males (Spearman $\rho = 0.62$, p-value = .001, **Figure 2.6C**). Among both institutions, carbidopa/levodopa is significant in AD males only.

2.3.7 Comparing Labs Between Sex-Specific AD and Control Groups Identifies Clusters of Lab Value Differences

We also performed an unbiased analysis of laboratory test result differences between AD patients and controls to phenotype patient groups. Among significantly different median lab values in both UCSF and Mount Sinai, AD patients have higher levels of hematocrit, serum calcium, RBC count, serum albumin, and cholesterol and lower levels of glucose, activated partial thromboplastin time (aPTT), alanine transaminase (ALT), and aspartate transaminase (AST) compared to controls (Mann-Whitney U-test, Bonferroni-corrected p-value threshold of 0.05, **Figure 2.6D**, **Supplementary Figure 2.4A**).

Average significant median lab values across sex-stratified groups (AD females, AD males,

control females, control males) and across institutions were clustered into 7 significant clusters (Family-wise Error Rate (FWER) corrected p-value 0.05 cutoff, **Figure 2.6D**). Clusters 1, 4, and 7 show discordant results between UCSF and Mount Sinai. Clusters 2 represent groups of significant median lab values lowest in control males, and highest either in all AD patients (e.g. albumin, sodium and carbon dioxide) or highest in AD females (e.g. HDL cholesterol, lymphocytes, calcium). Cluster 3 represents significant labs with greater median values in females and in controls (e.g. Free T4, sedimentation rate). Cluster 5 represents labs with lower significant median values in AD patients than controls for either the whole group (e.g. B-Type Natriuretic Peptide, AST) or in a sex-specific way where significant median lab values for males are greater than for females (e.g. aPTT, ALT, ferritin). Cluster 6 shows labs greater in AD compared to controls in a sex-specific way where overall males have greater significant median lab values than females (e.g. hemoglobin, RBC count). Across the board, the normalized lab values are correlated between the institutions (Female control: Spearman $\rho = 0.45$, p-value < 0.001; Male control: 0.46, p-value < 0.001; Female AD: 0.59, p-value $<$ 1e-5; Male AD: 0.64, p-value $<$ 1e-5; **Supplementary Figure 2.4B**).

2.4 Discussion

In this work, we demonstrate the capability of utilizing data from EMRs in order to perform deep phenotyping of a complex and heterogeneous disease, Alzheimer's Disease, and derive insights into associations with AD in a combined and sex-stratified analysis.

First, we performed low-dimensional topographical embedding of patients using diagnoses as features in order to visualize patients spatially. We see that AD status is significantly correlated with the first two UMAP components at both institutions, suggesting that phenotypic representation of patients using diagnosis data can demonstrate separation of AD and control

patients. The UMAP representation demonstrates a progressive spectrum between controls and AD, as well as representing variance and heterogeneity at individual patient resolution. Furthermore, with the UMAP representation, we can visualize topographically the distribution of age, sex, and other variables among patients.

We then generated comorbidity networks between AD and controls which provide a phenotypic representation of disease interactions among patient groups and a difference in connectivity between diseases in AD and controls. AD networks contain a greater number of edges and network metrics that point to higher rates of comorbid conditions among AD patients at both institutions, particularly with stronger links of hypertension (HTN)-hyperlipidemia, HTN-urinary tract infection (UTI), and HTN-anemia. Indeed, other studies have found multimorbidities (such as neuropsychiatric and cardiovascular patterns) to increase risk for dementia[39], and to contribute to AD pathological heterogeneity[40,41] displaying the larger complexity and heterogeneous nature of AD.

With enrichment analysis, we applied an integrative, unbiased, big data approach to EMR and identified previously known associations and possible novel connections with AD. Some diagnoses found enriched in AD patients compared to control patients from our analysis at both institutions that have been previously identified as linked with AD include midlife hypertension[16,42], diabetes mellitus[18,43], anemia[44,45], vascular pathology[17,46], osteoporosis[47,48], and urinary tract infections[49]. Enrichment of hypertension and vascular risk factors supports many current hypotheses of potential vascular pathologies and inflammatory factors that may lead to AD[17,50–52] or 'unmask' the symptoms of AD by decreasing cognitive reserve by causing vascular brain disease. Enrichment of diabetes and dyslipidemia supports existing literature that found links with diabetes mellitus and dyslipidemia[53], with proposed hypotheses involving energy

metabolism[54–56], inflammation[57–59], or the integrity of the blood brain barrier[60–62]. Enrichment of degenerative diseases of age, such as osteoporosis, osteoarthritis, urinary issues, and sensory issues may align with theories of AD as being a disease linked with frailty[63–65]. This analysis therefore provides an unbiased integrative way to identify multifactorial associations with AD. Our enrichment analysis also identified neoplasms as enriched in controls at UCSF, especially cancer of brain and liver. While this is an associative finding, this supports ideas that cancer and AD co-occur less frequently than the general population[66,67]. Some theories propose that AD and cancer have similar mechanisms and molecular pathways, but are dysregulated in different directions[68,69].

Next, we generated sex-specific comorbidity networks to provide insight into sex differences in the complexity of the disease. In both EMRs, female AD networks contain more nodes with network metrics suggesting greater connectivity than female controls or male AD networks. This may support association with greater combined diagnoses and multimorbidity in female AD patients compared to males[70]. These associations would be consistent with theories of greater risk of dementia in females as a result of multiple diseases or the theory of greater cognitive and pathological resilience to AD in females due to taking on a greater burden of more comorbidities. Furthermore, sex-stratified networks show secondary interactions between comorbidities and AD, such as links of HTN-UTI and HTN-chest pain among female AD populations, but not in male AD patients. These findings give higher order comorbidity interactions associated with AD that have not been examined previously.

When performing enrichment analysis, we identify sex-specific enrichments that may be linked to AD that have not been previously explored in depth. Male AD patients show enrichment of neurological and sensory disorders (sleep disorder, parkinsonism, and irritability), and among diagnoses significant in both sexes, male AD patients have stronger effect size with behavioral

diagnoses, agitation, and hearing loss. These disorders are also mostly shown to be significant and associated with greater effect size compared to females in our validation cohort. Prior studies have found hearing loss to increase risk of dementia diagnosis[71,72] or cognitive decline[73,74] in men. The enrichment of behavioral and neurological disorders found in male AD patients may indicate lessened resilience or higher occurrence of co-pathology. Furthermore, this analysis found the psychiatric phenotype associated with AD to be related to behavioral phenotypes in males compared to females, which is consistent with prior studies[75,76].

Female AD patients have enrichment of unique significant diagnoses in musculoskeletal categories (arthritis, fractures), atrial fibrillation, and accidents, and among diagnoses significant in both sexes, female AD patients show stronger effect size with depression, hypertension, urinary tract infections, and osteoporosis. Some of these disorders are similarly significant and associated with greater effect sizes compared to males in our validation cohort. The diagnoses of hypertension and atrial fibrillation would be in line with the hypothesis of potential cardiovascular risk factors and pathology that may affect females more. Indeed, there is evidence supporting cardiovascular fitness to be protective or vascular risk factors to be harmful towards cognitive decline and dementia in women[42,77–79]. Furthermore, these diagnoses suggest a phenotype for female AD patients along with other degenerative diseases of aging and frailty. In particular, the increase in musculoskeletal and bone disorders in female AD patients, as well as high calcium and vitamin D deficiency, may point to a potential bone metabolism pathology or aberrant calcium metabolism in female AD patients. From a psychiatric standpoint, the female AD phenotype is more associated with depression compared to males as supported by studies that found depression associated with greater hippocampal volume loss in women[80], and is more likely to be a manifestation of mild cognitive impairment or AD in females[81,82].

We performed sensitivity analysis by taking the number of encounters for each group into account. In general, we see a decrease in statistical significance in our enrichment analysis consistently across all diagnoses. This is likely due to decreased power from a lower sample size, and a bias towards selection of patients with more severe disease due to encounter thresholding. Overall, enriched diagnoses are relatively similar, with an increase in cerebrovascular disorders observed in AD, and particularly female AD patients. Neuroimaging studies have identified differences in AD phenotypes and brain networks depending on presence of cerebrovascular disease[83,84], which may support cerebrovascular events as an associated phenotype for a different or severe phenotype of AD.

Medication enrichments show expected associations with AD, as the top medication hits are current therapies used to modify symptoms of AD (e.g. memantine, donepezil), or are associated with diagnoses found in comorbidity analysis (e.g. antidepressants for depression). These medications are also identified as AD-enriched in our validation cohort, although many of these medications are expected as they are associated with conditions of aging. Medications enriched in controls provide a more interesting story, as they not only suggest an 'opposite AD' phenotype, but control-enriched hits may provide a way to hypothesize potential targets for further exploration of protective drug effects or drug repurposing. From our medication analysis, we see control enrichments of opioids, sedatives, dexamethasone, and furosemide, with dexamethasone also found significant in our validation cohort. The negative association with opioids is inconsistent with prior studies that found associations between prescription opioid use and AD risk[85], although control enrichment of opioids could possibly be due in part to decreased ability to communicate pain and decreased opioid prescriptions after AD[86]. Nevertheless, studies have implicated the role of opioid system dysregulation in tau hyperphosphorylation and AD[87].

Dexamethasone is a corticosteroid that has been suggested to help reduce inflammation in AD[88,89], although the data on efficacy is still uncertain and may depend upon the need for combination therapy[90] or control of other factors that complicate the relationship between hormonal levels and the brain[91,92]. Furosemide is a diuretic drug used to treat hypertension and may confer a protective effect through the control of comorbid conditions that contribute to cardiovascular risk factors. Furosemide also reduces the production of CSF by inhibiting carbonic anhydrase, which may impact CSF dynamics and help decrease the risk of AD[93]. Prior studies have shown possible protective effects from diuretic drugs and AD[94–97], and one study identified furosemide as a potential probe molecule for reducing neuroinflammation[98].

Characterizing patients by lab values provides another way to phenotype patient groups. Through our analysis, greater calcium levels were identified, especially in AD females. A small observational study found calcium supplementation to increase risk of dementia in women with cerebrovascular disease[99]. Calcium dysregulation and homeostasis have been implicated in AD neuronal signaling pathology, and identified as a target for drug development[99,100]. Control-enriched labs may also be related to gastrointestinal cancers or liver/pancreatic dysfunction, as we observe increased AST, ALT, and glucose levels in controls and particularly among males. This result is not consistent with a study observing greater glucose levels to increase dementia risk[101], although one study did find low ALT[102] to be associated with AD, and some publications implicate altered glucose metabolism[103,104] and liver dysfunction in AD pathology[102,105,106]. Furthermore, since our control cohort has been matched on age and death status, control patients may encompass a population with terminal disease. Lab clusters also demonstrate phenotypes specific to a sex group. A lower clotting time (aPTT, PT) and greater platelet count, prealbumin, lymphocytes, and cholesterol levels in female AD patients may provide a multivariate way to identify potential AD

phenotype in females. Prior studies have shown high thrombin[107,108], abnormalities of hemostasis[109,110], and abnormal platelet activation[111–113] in AD patients that may contribute to a pro-thrombotic state in AD[114], leading to microinfarcts and cerebrovascular dysfunction[115,116], although sex-specific associations have not been studied previously. Furthermore, control sex phenotype may demonstrate protective labs or biomarkers that decrease risk of AD. We see lower free T3 in control males, and greater free T4 in control females. Indeed, studies on AD populations have shown high TSH and low free T4 to be associated with the disease[117–119], although sex-specific associations have not been explored in depth.

Some limitations do exist in our study. First, AD is an insidious and heterogeneous disorder, and is frequently misdiagnosed even in specialized dementia centers. Clinically, Alzheimer's Dementia is suspected when disease biomarker status is unknown, whereas Alzheimer's Disease is diagnosed when biomarker status is confirmed. Our current study did not rely on biomarker-positive cases of Alzheimer's Disease, and we did not exclude patients with other pathologies that can also impact brain health through different pathways, such as Parkinson's Disease. Nevertheless, Alzheimer's Disease often co-occurs with other dementias[120,121]. Second, EMRs, while a rich data source, is a very sparse dataset with a lot of missing data, such as sociological factors (e.g., income, education, etc). Nevertheless, the number of patients represented in the EMR is exceptionally large and provides robust opportunities for deriving meaningful insights or hypotheses. This limitation also applies to our validation EMR. Additionally, some associations may be different across the two systems due to differences in the underlying patient populations or standards of care. Therefore, it is possible that the UCSF EMR does not capture an association that may be more prevalent in a different population in New York, and vice versa. How other covariates including socioeconomic factors modify specific AD associations is a question

that can be followed-up in future work. Third, our definition of controls comes with limitations, as it is difficult to identify 'healthy' controls in the EMR. The institutions represented in our data includes both primary and tertiary care, which includes patients that seek hospital care for a variety of reasons. As such, there may be bias in the underlying patient population who chooses to seek medical care at a metropolitan medical center. Regardless, the power in utilizing EMR allows us to generate hypotheses with a large number of patients and versatility in choice of controls compared to many current AD studies. Lastly, our analysis only identifies associations with AD and does not take temporal factors into consideration, therefore causal relationships cannot be concluded. This will be the main focus of future work, as the temporal association can categorize an association as a risk/protective factor (if early in age), a diagnostic clue (if during AD diagnosis), or as a manifestation of AD progression or severity (if after AD diagnosis). Nevertheless, given AD is an insidious disorder, there can be brain perturbations a decade or more before a diagnosis is determined and documented in clinical records. While we made the assumption of independence in our statistical methods to identify significant associations, this method can be further extended to alternative statistical models that take covariates into account. Our current work allows the unbiased identification of associations and phenotyping, which can then be used to generate hypotheses for guiding follow-up studies.

Overall, our analyses leveraged an extensive clinical dataset to (1) phenotype and represent AD and (2) perform enrichment analysis to identify known or suggested novel associations with AD, as well as elicit sex-specific differences. We were therefore able to apply an integrative, unbiased, big data approach to identify associations with AD and provide phenotypic representations of an otherwise complex disease. With this approach, we can generate many new hypotheses to better motivate future work to understand AD complexity and develop diagnostic

strategies and therapeutic interventions. Future work will include temporal analysis in order to identify longitudinal relationships and predictive modeling for AD risk, diagnosis, or progression. More extensive analysis of medication and lab values, especially among opposite phenotypes in controls, may lead to better strategies for prevention or treatment of AD. Besides elucidating sex differences, next steps for phenotyping can include investigating race/ethnicity differences or differences based upon other covariates to better characterize Alzheimer's Disease heterogeneity. Furthermore, incorporation of molecular or genetic data with clinical data can help better elucidate potential mechanisms underlying identified associations.

2.5 Methods

We performed deep phenotyping and association analysis of AD and control patients. First, AD and control cohorts were identified from the UCSF EMR and topographically visualized via a low-dimensional projection of comorbidities. Comorbidity networks were created, and association and enrichment analysis were performed on all diagnoses, medications, and lab values. These analyses were further performed in a sex-stratified manner to identify sex-specific associations, and validation was performed on the Mount Sinai EMR. An overview of the workflow is shown in **Figure 2.1**.

2.5.1 Patient Cohort Identification

Patient cohorts were identified from over five million patients in the UCSF EMR database, which includes clinical data from 1982-2020. Due to the de-identification process, dates are shifted by at most a year (with relative dates preserved) and all birth dates before 1930 (= estimated age 90) are shifted to be no earlier than 1930. AD patients were identified by inclusion criteria of estimated age >64 years, and ICD-10-CM codes G30.1, G30.8, or G30.9, where estimated age is determined from the birth date. To identify a control group, we used propensity score (PS) matching method (*matchit* R package[115]) by a logistic regression model to match controls to AD patients. The control group was selected from patients >64 years old without AD diagnosis, matched on sex, estimated age, race, and death status at a 1:2 AD:control ratio using a nearest neighbors method. The demographic properties of the UCSF and Mount Sinai cohorts are shown in **Table 2.1**.

2.5.2 Dimensionality Reduction Patient Visualization

All identified patients were represented using one-hot encoding of diagnoses, excluding

encoding of diagnoses with Alzheimer's in the name (list in **Supplementary Table 2.2**, **Figure 2.2**). Patients were then visualized in a lower dimension using Uniform Manifold Approximation and Projection[122] (UMAP) with the umap-learn package from Python. Correlations between variables and UMAP coordinates were analyzed using Mann-Whitney U-Test for categorical variables, and Pearson's Correlation Coefficient for continuous variables.

2.5.3 AD vs. Control Enrichment Analysis of Comorbidities

To evaluate comorbidities, all diagnoses recorded from patient cohorts were identified with the earliest entry of every diagnosis. Comparisons were made at different ICD-10 hierarchical levels, specifically Level 2 categories (e.g. G30-G32: Other degenerative diseases of the nervous system), Level 3 categories (e.g. G30: Alzheimer's Disease), or full diagnosis names (e.g. G30.9 Alzheimer's Disease, unspecified). Level 2, Level 3, and full diagnosis names are also grouped by ICD-10 blocks (e.g. G00-G99: Diseases of the Nervous System. More information on ICD-10-CM codes can be found at the following website: www.cms.gov/Medicare/Coding/ICD10/ICD-10Resources.

Diagnosis networks were created based upon a diagnosis category or diagnosis shared by >1% patients in a group (node) or pair of diagnosis categories or diagnoses shared by >1% of patients in a group (edge). Network metrics were computed using Cytoscape app Network Analyzer[123]. Metrics were then compared between AD and control networks using Mann-Whitney U-test, with and without singleton nodes removed. Nodes and edges were thresholded by 5% of patients in a group for visualization purposes.

Enrichment analysis of diagnosis was compared between AD and control cohorts. For each diagnosis, the proportions of patients in each group were compared using Fisher Exact (if <5 patients in a category) or Chi-Squared test. Significant diagnoses were determined by a

Bonferroni-corrected threshold of p-value < 0.05, and directionality determined with Odds Ratio (OR). With inspiration from genetic and molecular approaches, the results were visualized using Manhattan plots by categorizing diagnoses in ICD-10 blocks.

2.5.4 Sex-Stratified AD vs. Control Enrichment Analysis of Comorbidities

Diagnostic networks were created for each sex, with diagnosis categories or diagnoses shared by >1% of patients in a group (node), and diagnosis category/diagnosis pair shared by >1% of patients in a group (edge). Network metrics were then computed using Cytoscape Network Analyzer app, and compared between sex-stratified AD patients and controls, and between males and females for both AD and controls separately with a Mann-Whitney U-test. Nodes and edges were thresholded by 5% of patients in a group for visualization.

Sex-specific enrichment analysis of diagnoses between AD and control cohorts were compared with a subset of equal numbers of AD and control patients for each sex. For each diagnosis, the proportions of patients in each group were compared using the Fisher Exact (if <5 patients in a category) or Chi-Squared test. Significance was determined by applying a threshold of 0.05 for Bonferroni-corrected p-values. Log-log plots were generated from odds ratios between Female and Male AD patients and controls, and Miami plots created by categorizing diagnoses in ICD-10 blocks.

2.5.5 Sensitivity Analysis Taking Encounters into Account

Sensitivity analysis of diagnosis enrichment analysis was performed with a subgroup of AD patients and a second control cohort to account for variability in the number of visits for each patient. AD cohorts were subgrouped by identifying patients with over 10 encounters in the EMR and records spanning over a year. The encounter-filtered control cohort was identified by

additionally matching on the number of encounters and years between the first and last record in the EMR. Diagnosis enrichment analysis was carried out as described above for general comorbidities and sex-specific analysis.

2.5.6 AD vs. Control Enrichment Analysis of Medications

All medications ordered for AD and control patients were extracted and grouped based upon the generic medication name, with route and dosage information removed. The proportions of AD and control patients prescribed each medication were compared using Fisher Exact (if <5 patients in a category) or Chi-Squared tests. Significantly enriched medications were identified by a Bonferroni-corrected threshold of p-value 0.05, and directionality was determined with an Odds Ratio. Sex-specific medication comparisons were also performed within a subset of equal numbers of AD and control patients for each sex and plotted with cutoffs based upon a Bonferroni-corrected p-value threshold of 0.05 and odds ratios (OR) threshold of <0.5 or >2.

2.5.7 AD vs. Control Comparisons of Lab Values

For laboratory values, median values for all numerical lab test results for each patient were identified. Lab tests missing data among 95% or more patients were removed. Lab value distributions were compared using Mann-Whitney U-test across three comparisons (AD vs. controls, Female AD vs. Female controls, and Male AD vs. Male controls) in order to identify significantly different lab values.

For clustering analysis, significant lab tests above a threshold of 0.05 for Bonferroni-corrected p-value were isolated, and mean values were then identified for each group (AD Females, AD Males, control Females, control Males) and normalized across groups as a Z-score. Clustering was then performed using the *sigclust2* R package[124] to determine significance of each cluster

break using permutations (Euclidean distance metric and average linkage).

2.5.8 Validation in External EMR

AD and PS-matched control patients were identified in the Mount Sinai EMR in the same fashion as described in [Patient Identification] in the UCSF EMR. All aforementioned analysis with dimensionality reduction, comorbidity networks, diagnosis/medication enrichments, sex-specific enrichments, and lab value comparisons were performed in the Mount Sinai dataset as they have been in the UCSF EMR dataset.

For network comparisons, network metrics were standard normalized across the 12 networks (6 at UCSF, 6 at Mount Sinai) by the metric and Spearman rank correlation coefficient and significance determined. For diagnosis comparison, Level 2, Level 3, and full diagnosis names were mapped and compared by the sub-chapter, three-digit codes, and full ICD-10-CM code of the ICD-10 hierarchy, respectively. Significant diagnosis in the validation cohort was determined by a Bonferroni-corrected threshold of 0.05 based upon the number of mapped UCSF-significant diagnoses. Correlations between odds ratios were determined by a Spearman rank correlation coefficient and significance. Medications were mapped based upon the generic name, and correlations between odds ratios determined with Spearman rank correlation coefficient.

For comparison of labs, the normalized lab values for each institution were combined, and clustering performed using Euclidean distance and average linkage to identify groups of labs with similar trends between AD/sex/institution stratified patient groups. The R package *sigclust2* was used to determine significant clusters of labs.

2.5.9 Data Visualization Using RShiny

An interactive visualization of comorbidity enrichments and networks between AD and

control groups and with sex stratification was implemented in an Rshiny[125] app: vizad.org.

2.6 Tables

Table 2.1 Patient Demographics

Summary table of sex, estimated age, death status, and first race among Alzheimer's and control cohorts at UCSF and Mount Sinai. Patients are propensity-score matched at a 1:2 Alzheimer to control ratio with the demographics shown in the table. SD: standard deviation. SMD: standardized mean difference. NHPI: Native Hawaiian/ Pacific Islander

	UCSF				Mount Sinai			
	Overall	**AD**	**Control**	**SMD**	**Overall**	**AD**	**Control**	**SMD**
n	26412	8804	17608		17874	5958	11916	
Sex, n (%)								
Female	16675 (63.1)	5558 (63.1)	11117 (63.1)	<0.001	12584 (70.4)	4138 (69.5)	8446 (70.9)	0.031
Male	9659 (36.6)	3220 (36.6)	6439 (36.6)		5290 (29.6)	1820 (30.5)	3470 (29.1)	
Unknown	78 (0.3)	26 (0.3)	52 (0.3)					
Estimated Age, mean (SD)	86.5 (6.4)	86.5 (6.4)	86.5 (6.4)	<0.001	88.6 (10.6)	88.3 (8.7)	88.7 (11.4)	-0.039
Race, n (%)								
American Native	27 (0.1)	9 (0.1)	18 (0.1)	<0.001	20 (0.1)	8 (0.1)	12 (0.1)	0.129
Asian	2638 (10.3)	879 (10.3)	1759 (10.3)		177 (1.0)	78 (1.3)	99 (0.8)	
Black/African American	1758 (6.9)	586 (6.9)	1172 (6.9)		3732 (20.9)	1214 (20.4)	2518 (21.1)	
Native Hawaiian/ Pacific Islander	1356 (5.3)	452 (5.3)	904 (5.3)		9 (0.1)	5 (0.1)	4 (0.0)	
Other	2230 (8.7)	743 (8.7)	1487 (8.7)		3922 (21.9)	1496 (25.1)	2426 (20.4)	
Unknown	2017 (7.6)	673 (7.6)	1344 (7.6)		786 (4.4)	253 (4.2)	533 (4.5)	
White/Caucasian	16386 (64.0)	5462 (64.0)	10924 (64.0)		9228 (51.6)	2904 (48.7)	6324 (53.1)	
Death Status, n (%)								
Alive	20146 (76.3)	6714 (76.3)	13432 (76.3)	0.001	9371 (52.4)	3264 (54.8)	6107 (51.3)	0.078
Deceased	6266 (23.7)	2090 (23.7)	4176 (23.7)		882 (4.9)	306 (5.1)	576 (4.8)	
Unknown					7621 (42.6)	2388 (40.1)	5233 (43.9)	

2.7 Figures

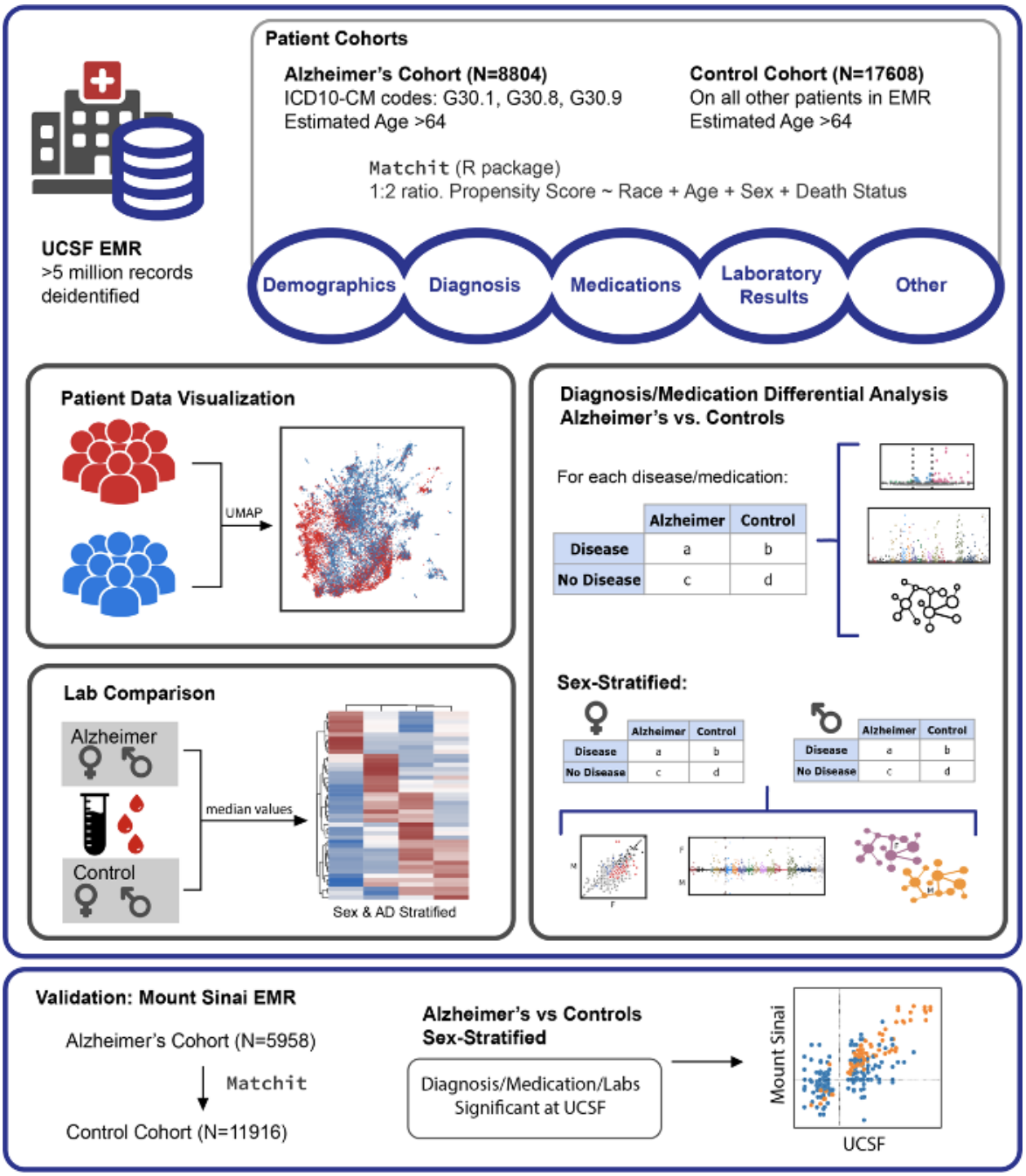


Figure 2.1 Overview of the workflow

Visualization of patient cohort identification from the UCSF EMR and methods for deep phenotyping and enrichment analysis. Validation analysis is done with Mount Sinai EMR to assess correlations.

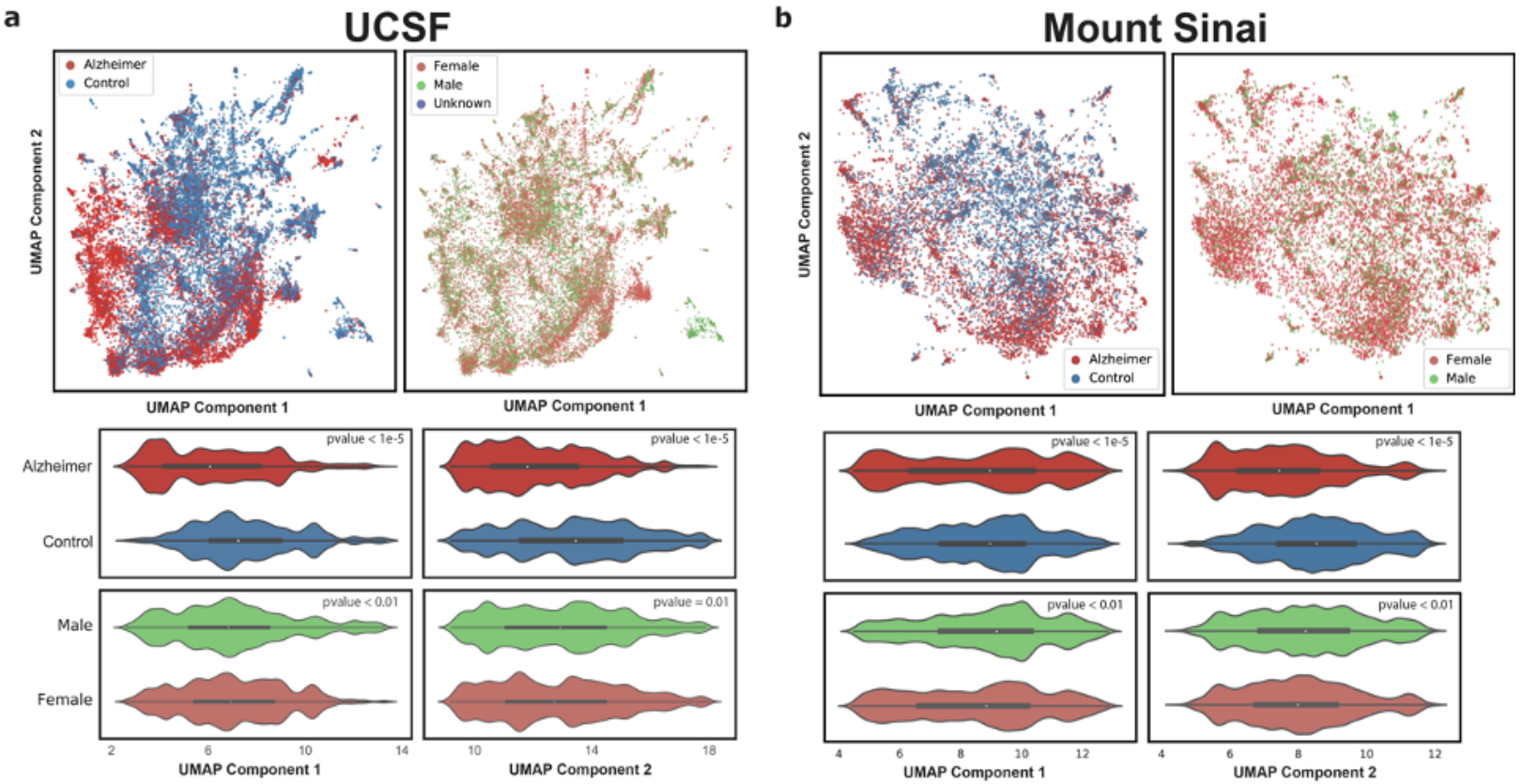


Figure 2.2 UMAPs using comorbidities as features provides a topographical view of the distribution of patients

Top row: UMAP of all patients (AD and controls), with each dot representing a patient, colored by AD status (**a** top left, **b** top left) or by sex (**a** top right, **b** top right).
Middle and Bottom rows: Violin plots show distribution of AD and control patients along the UMAP principal components for UCSF (**a**) and Mount Sinai (**b**), and p-values determined from comparing distributions with a Mann-Whitney U-Test.

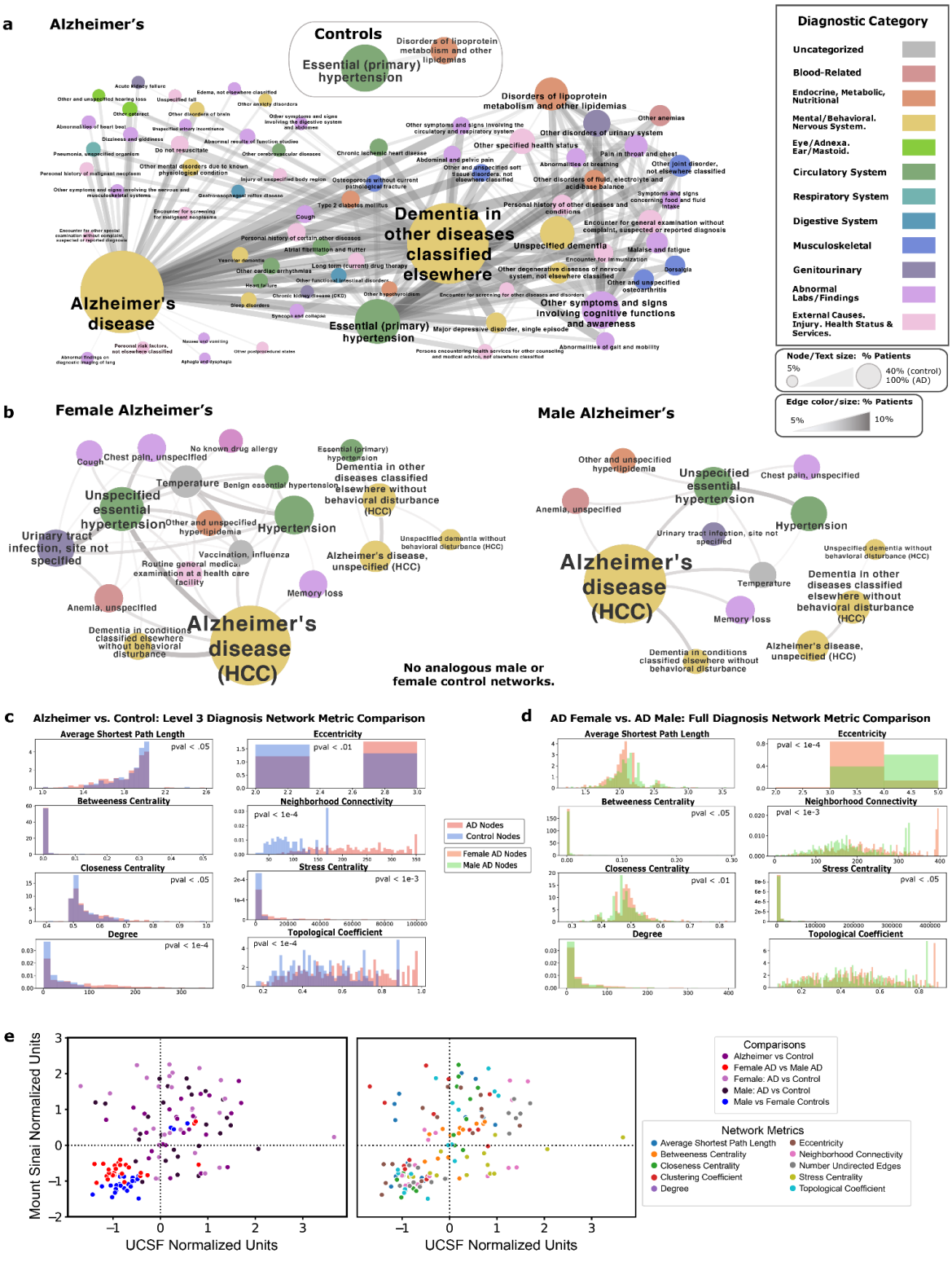


Figure 2.3 Comorbidity Networks Show Greater Co-Diagnosis in AD vs. Controls, and in Female AD vs Male AD patients
(Figure caption continued on the next page.)

(Figure caption continued from the previous page.)
a,b: Network Diagrams: For each network, the node size, text size, edge size, and edge color represent the number of patients sharing a diagnosis or diagnosis pair. Node colors are based on ICD-10 category. A threshold of 5% sharing was applied.
a. Network for Level 3 diagnosis categories in AD vs. control patients. Nodes and edges represent >5% of diagnosis or diagnosis pairs shared in each cohort, respectively.
b. Left: Female network of full diagnosis names. Each node and edge represent diagnosis or diagnosis pairs shared by >5% of AD females. No analogous comorbidity network was generated from control females.
Right: Male network of full diagnosis names. Each node and edge represent diagnosis or diagnosis pairs shared by >5% of AD males. No analogous comorbidity network was generated on control males.
c. Comparison of network metrics between AD and control Level 3 Diagnosis Category Networks. Statistical Tests are performed with Mann-Whitney U-Test.
d. Comparison of network metrics between Male and Female Alzheimer's Disease Full Diagnostic Name Networks. Statistical Tests are performed with Mann-Whitney U-Test.
e. Correlation of network metrics compared with validation EMR network metrics, normalized by the metric. Colors represent comparison type (left) or the specific network metric (right), Spearman $\rho = 0.55$, p-value < 1e-4.

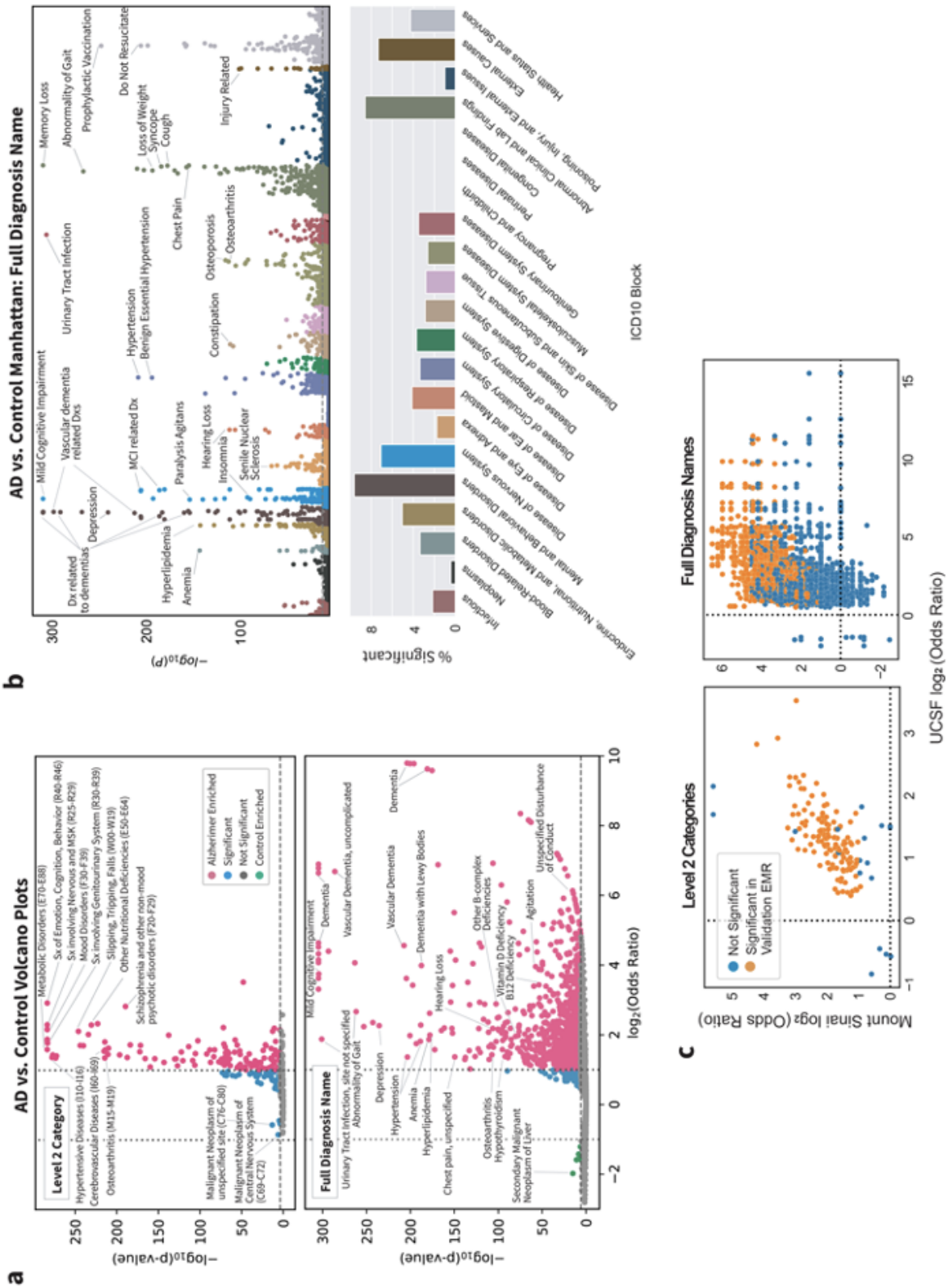


Figure 2.4 Comorbidity Enrichment Analysis identifies enriched diagnosis in AD vs. Controls
(Figure caption continued on the next page.)

(Figure caption continued from the previous page.)

a. Volcano plot for Level 2 categories (top) and full diagnosis names (bottom) compared between AD and control cohorts using Fisher Exact or Chi-Squared test. P-value cutoff is Bonferroni corrected (p-value < 2e-8 and 1e-6) with log 2 odds ratio cutoff of 1 for AD-enriched (pink) or log 2 odds ratio cutoff of -1 for control-enriched (green) and remaining significant diagnoses in blue. Some of the top significant diagnoses are labelled.
b. Top, a Manhattan plot with full diagnosis names colored by ICD-10 categories with Bonferroni-corrected p-value cutoff. Some of the top diagnoses in each category are labelled. Bottom, percentage of diagnosis in each ICD-10 category that is significant.
c. Diagnosis AD vs. Control odds ratio correlation plots between UCSF and Mount Sinai for Level 2 diagnosis categories and full diagnosis names that are significant at UCSF. Each dot represents a category or diagnosis, and dots in orange are significant at Mount Sinai with Bonferroni-corrected p-value threshold of 0.05.

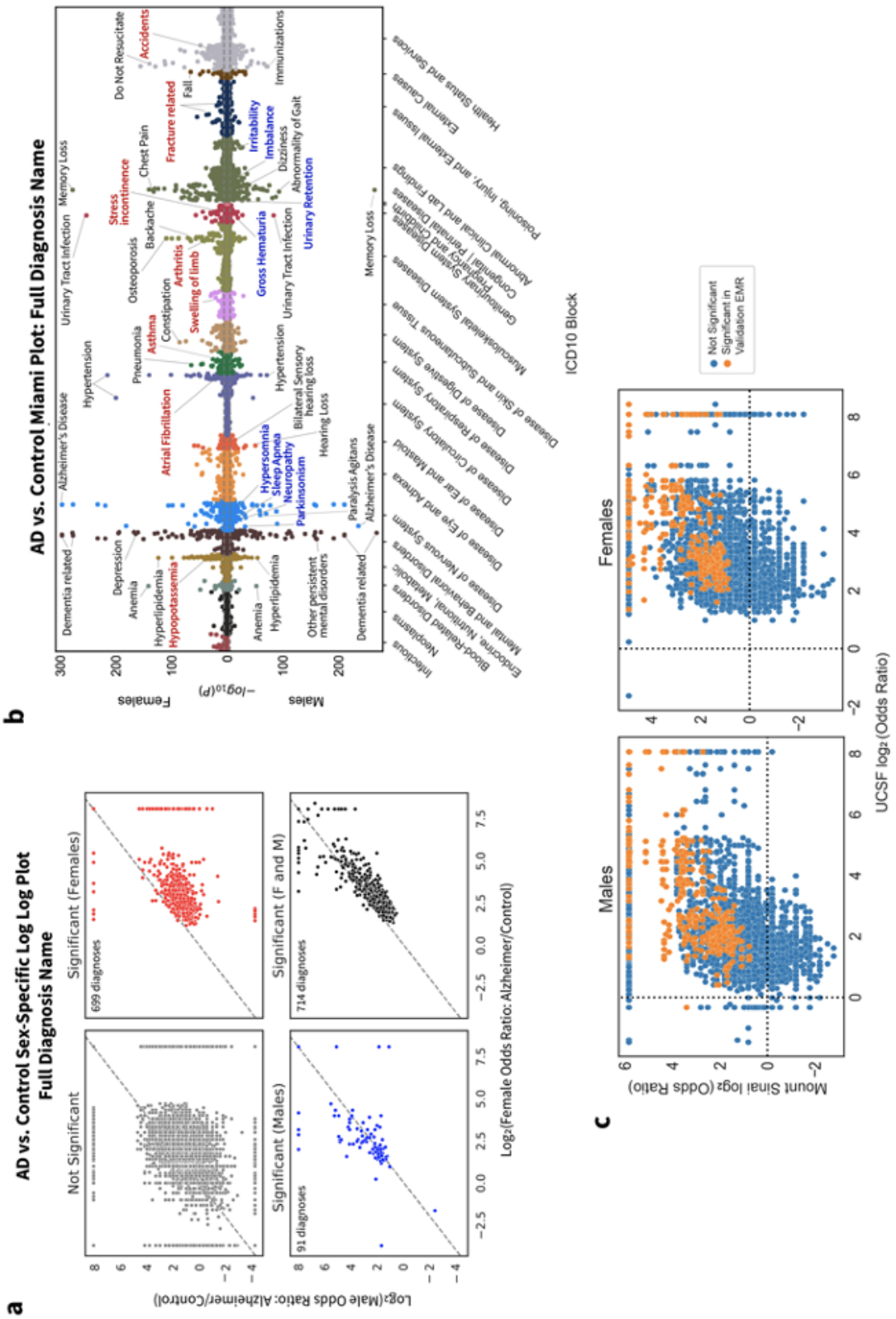


Figure 2.5 Comorbidity Enrichment Analysis identifies sex-specific enriched diagnoses in AD vs. Controls

(Figure caption continued on the next page.)

(Figure caption continued from the previous page.)

a. Full diagnosis names compared between AD and controls within each sex. The log2 of the odds ratio is plotted on the axis, and points are colored by significance (Bonferroni -corrected, p-val cutoff > 3e-6).
b. Miami plot of the diagnosis names grouped by sex and ICD-10 categories. Select top diagnoses are labelled, with diagnosis names colored by significance as female only (red), male only (blue), or significant in both sexes (black).
c. Correlation plots of AD vs. control odds ratios between UCSF and Mount Sinai for diagnoses that are significant at UCSF. Each dot represents a diagnosis, and dots in orange are significant at Mount Sinai with Bonferroni-corrected p-value threshold of 0.05.

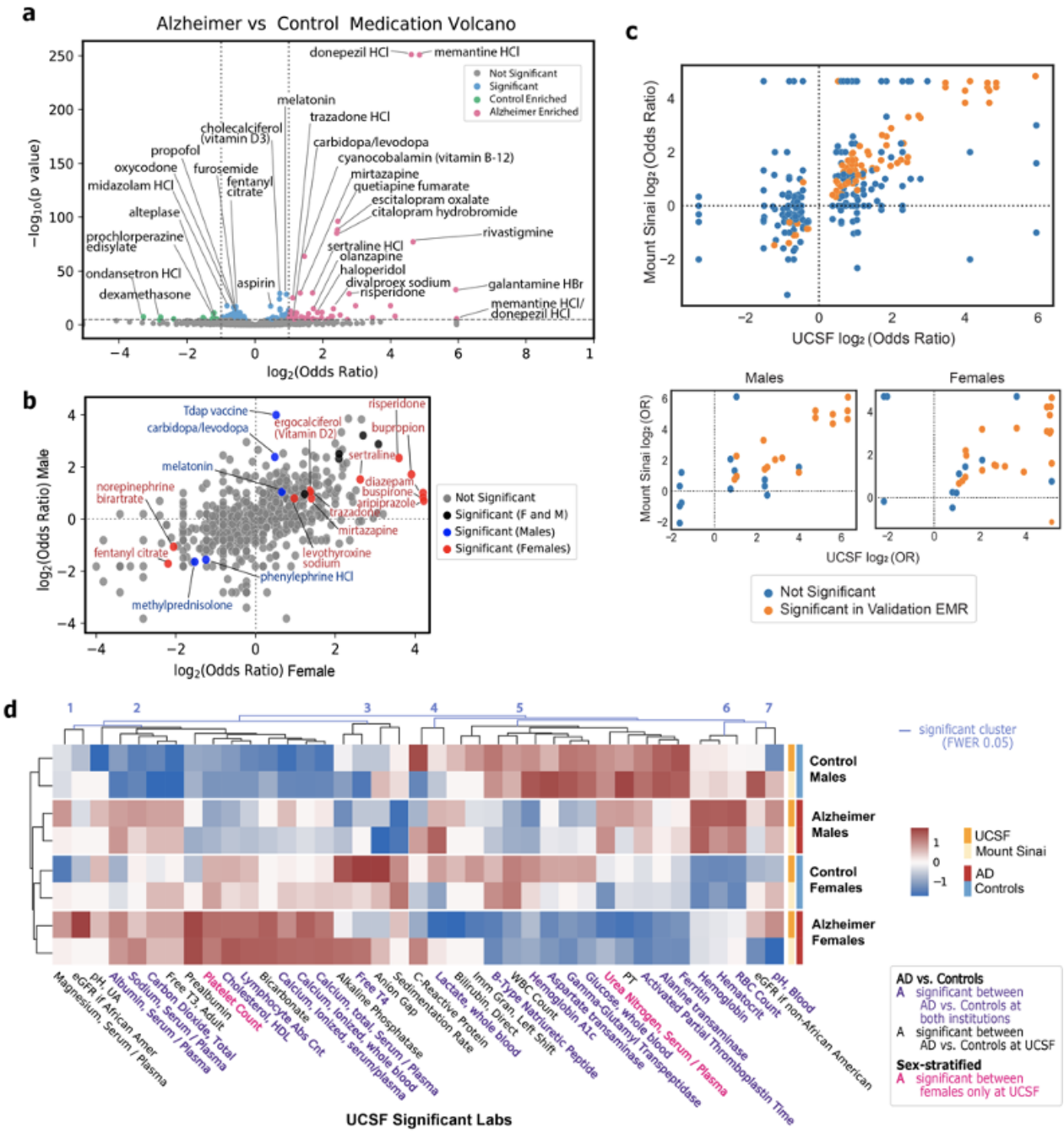


Figure 2.6 Medication and Lab Analysis shows Medication Enrichments and Median Lab Value Differences between AD and Controls

a. Volcano plot for generic medication names compared between AD and controls using Fisher Exact or Chi-Squared Test. P-value cutoff is Bonferroni corrected (p-value < 2e-5) with odds ratio cutoff at 2 for AD-enriched (pink) or 1/2 for control-enriched (green). Remaining significant diagnoses are in blue.

b. Log-log plot of generic medication names compared between AD and controls within each

(Figure caption continued on the next page.)

(Figure caption continued from the previous page.)

sex. The log of the odds ratio for each sex is plotted on the axis, with points colored by Bonferroni-corrected significance (p-value cutoff < 4e-5) if female only (red), male only (blue), or both (black).

c. AD vs control (top) and sex-specific (bottom) odds ratio correlation plots between UCSF and Mount Sinai for medications significant at UCSF. Each dot represents a medication, and dots in orange are significant at Mount Sinai with Bonferroni-corrected p-value threshold of 0.05.
d. Heatmap of lab values filtered on significance at UCSF in AD vs control comparison across sex-specific groups at UCSF and Mount Sinai. Labs are clustered with light blue lines representing significant cluster breaks (family-wise error rate (FWER) corrected p-value 0.05). Text color represents significant labs at both institutions (purple), significant among females only at UCSF (red), or significant between AD vs controls at UCSF only (black). Heatmap colors represent z-score of the average median value across the 4 groups at each institution.

2.8 Supplementary Tables

Supplementary Table 2.1 Patient Demographics with Encounter Thresholds and Controlling

Distribution of sex, estimated age, death status, and first race among Alzheimer's and control cohorts. These cohorts are thresholded on more than 10 encounters, and over a year representation in the EMR. Patients are matched at a 1:2 Alzheimer to control ratio with the demographics shown in the table. Estimated age shows mean and median (25%ile - 75%ile).

	Count	Age	Death Status	Race	
Alzheimer's Cohort	**6,612**	**86.4 90 (83-91)**	**Alive: 6714 (76.3%) Deceased: 2090 (23.7%)**	**White/Caucasian: 5462 (64.0%) Asian: 879 (10.3%) Black/African American: 586 (6.9%) Hawaiian/Pacific Islander: 452 (5.3%)**	**American Native: 9 (.1%) Other: 743 (8.7%) Unknown/Declined: 802 (4.7%)**
Males	2,382 (36.0%)	85.7 90 (84-91)	Alive: 1778 (74.6%) Deceased: 604 (25.4%)	White/Caucasian: 1570 (66.8%) Asian: 254 (10.8%) Black/African American: 128 (5.4%) Hawaiian/Pacific Islander: 117 (5.0%)	Other: 209 (8.9%) Unknown/Declined: 72 (3.1%)
Females	4,223 (63.9%)	86.8 90 (82-91)	Alive: 3084 (73%) Deceased: 1139 (27%)	White/Caucasian: 2525 (60.5%) Asian: 497 (11.9%) Black/African American: 393 (9.4%) Hawaiian/Pacific Islander: 217 (5.2%)	American Native: 8 (0.2%) Other: 404 (9.7%) Unknown/Declined: 130 (3.1%)
Other or Unknown	7 (0.10%)	90.7 91 (90.5-91)	Alive: 7 (100%)	White/Caucasian: 6 (85.7%) Unknown/Declined: 1 (14.3%)	
Control Cohort	**13,224**	**86.2 90 (83 – 91)**	**Alive: 13432 (76.3%) Deceased: 4176 (23.7%)**	**White/Caucasian: 10924 (64.0%) Asian: 1759 (10.3%) Black/African American: 1172 (6.9%) Hawaiian/Pacific Islander: 904 (5.3%)**	**American Native: 18 (.1%) Other: 1487 (8.7%) Unknown/Declined: 802 (4.7%)**
Males	4,674 (35.3%)	85.8 90 (82-91)	Alive: 3248 (69.5%) Deceased: 1426 (30.5%)	White/Caucasian: 3076 (66.8%) Asian: 490 (10.7%) Black/African American: 277 (6.0%) Hawaiian/Pacific Islander: 222 (4.8%)	American Native: 5 (.1%) Other: 384 (8.3%) Unknown/Declined: 247 (3.2%)
Females	8,539 (64.6%)	86.5 90 (84-91)	Alive: 6253 (73.2%) Deceased: 2286 (26.8%)	White/Caucasian: 5225 (61.9%) Asian: 1024 (12.1%) Black/African American: 768 (9.1%) Hawaiian/Pacific Islander: 387 (4.6%)	American Native: 11 (.1%) Other: 783 (9.3%) Unknown/Declined: 246 (2.9%)
Other or Unknown	11 (0.10%)	90.2 90 (90-91)	Alive: 10 (90.9%) Deceased: 1 (9.1%)	White/Caucasian: 1 (11.1%) Other: 3 (33.3%)	Unknown/Declined: 5 (55.6%)

Supplementary Table 2.2 UMAP Exclusion Terms

Table of diagnosis excluded in UMAP embedding. These terms contain the word ‘Alzheimer’.

AD (Alzheimer's disease) (HCC)
Alzheimer disease (HCC)
Alzheimer disease type 3 (HCC)
Alzheimer's dementia (HCC)
Alzheimer's dementia with behavioral disturbance (HCC)
Alzheimer's dementia with behavioral disturbance, unspecified timing of dementia onset (HCC)
Alzheimer's dementia without behavioral disturbance (HCC)
Alzheimer's dementia without behavioral disturbance, unspecified timing of dementia onset (HCC)
Alzheimer's dementia, late onset (HCC)
Alzheimer's dementia, late onset, with behavioral disturbance (HCC)
Alzheimer's disease (HCC)
Alzheimer's disease of other onset
Alzheimer's disease of other onset with behavioral disturbance (HCC)
Alzheimer's disease of other onset without behavioral disturbance (HCC)
Alzheimer's disease with delirium (HCC)
Alzheimer's disease with early onset (CODE) (HCC)
Alzheimer's disease with early onset (HCC)
Alzheimer's disease with late onset (CODE) (HCC)
Alzheimer's disease with late onset (HCC)
Alzheimer's disease with presenile onset (HCC)
Alzheimer's disease, early onset (HCC)
Alzheimer's disease, familial (HCC)
Alzheimer's disease, focal onset (HCC)
Alzheimer's disease, unspecified (CODE) (HCC)
Alzheimer's disease, unspecified (HCC)

Alzheimer's type dementia (HCC)
Alzheimer's type dementia with late onset with behavioral disturbance (HCC)
Alzheimer's type dementia with late onset without behavioral disturbance (HCC)
Alzheimers disease (HCC)
DAT (dementia Alzheimer type)
DAT (dementia of Alzheimer type) (HCC)
Dementia due to Alzheimer's disease (HCC)
Dementia in Alzheimer's disease (HCC)
Dementia in Alzheimer's disease with delusions (HCC)
Dementia in Alzheimer's disease with depression (HCC)
Dementia in Alzheimer's disease with early onset (HCC)
Dementia in Alzheimer's disease with early onset with behavioral disturbance (HCC)
Dementia in Alzheimer's disease with early onset without behavioral disturbance (HCC)
Dementia in Alzheimer's disease with early onset, with behavioral disturbance
Dementia in Alzheimer's disease with early onset, without behavioral disturbance
Dementia in Alzheimer's disease with late onset
Dementia of Alzheimer's type with behavioral disturbance (HCC)
Dementia of Alzheimer's type, with early onset, with depressed mood (HCC)
Dementia of the Alzheimer's type (HCC)
Dementia of the Alzheimer's type with early onset with behavioral disturbance (HCC)
Dementia of the Alzheimer's type with late onset without behavioral disturbance (HCC)
Dementia of the Alzheimer's type without behavioral disturbance (HCC)
Dementia of the Alzheimer's type, with late onset, uncomplicated (HCC)

Dementia of the Alzheimer's type, with late onset, with delirium (HCC)
Dementia of the Alzheimer's type, with late onset, with delusions (HCC)
Dementia of the Alzheimer's type, with late onset, with depressed mood (HCC)
Dementia of the Alzheimer's type, with late onset, with depressive mood (HCC)
Dementia, Alzheimer's, with behavior disturbance (HCC)
Early onset Alzheimer disease
Early onset Alzheimer's dementia without behavioral disturbance (HCC)
Early onset Alzheimer's disease with behavioral disturbance (HCC)
Family history of Alzheimer's disease
Focal Alzheimer's disease (HCC) 'Late onset Alzheimer disease (HCC)
Late onset Alzheimer's disease with behavioral disturbance (HCC)
Late onset Alzheimer's disease without behavioral disturbance (HCC)
Major neurocognitive disorder due to Alzheimer's disease (HCC)
Major neurocognitive disorder due to Alzheimer's disease, possible (HCC)
Major neurocognitive disorder due to Alzheimer's disease, probable, with behavioral disturbance (HCC)
Major neurocognitive disorder due to Alzheimer's disease, probable, without behavioral disturbance (HCC)
Major neurocognitive disorder due to Alzheimer's disease, with behavioral disturbance (HCC)
Major neurocognitive disorder due to possible Alzheimer's disease (HCC)
Major neurocognitive disorder, due to Alzheimer's disease, with behavioral disturbance, mild (HCC)
Major neurocognitive disorder, due to Alzheimer's disease, without behavioral disturbance, mild (HCC)

Major neurocognitive disorder, due to Alzheimer's disease, without behavioral disturbance, moderate (HCC)
Major neurocognitive disorder, due to Alzheimer's disease, without behavioral disturbance, severe (HCC)
Mild major neurocognitive disorder due to Alzheimer's disease with behavioral disturbance (HCC)
Mild major neurocognitive disorder due to Alzheimer's disease without behavioral disturbance (HCC)
Mild neurocognitive disorder due to Alzheimer's disease (HCC)
Mild possible major neurocognitive disorder due to Alzheimer's disease (HCC)
Mixed Alzheimer's and vascular dementia (HCC)
Mixed Alzheimer's and vascular dementia with behavior disturbances (HCC)
Moderate major neurocognitive disorder due to Alzheimer's disease without behavioral disturbance (HCC)
Moderate probable major neurocognitive disorder due to Alzheimer's disease with behavioral disturbance
Other Alzheimer's disease (HCC)
Possible major neurocognitive disorder due to Alzheimer's disease
Primary degenerative dementia of Alzheimer type (HCC)
Primary degenerative dementia of the Alzheimer type, senile onset (HCC)
Primary degenerative dementia of the Alzheimer type, senile onset, uncomplicated (HCC)
Primary degenerative dementia of the Alzheimer type, senile onset, with depression (HCC)
Probable major neurocognitive disorder due to Alzheimer's disease with behavioral disturbance
Probable major neurocognitive disorder due to Alzheimer's disease without behavioral disturbance
Progressive aphasia in Alzheimer's disease (HCC)

SDAT (senile dementia of Alzheimer's type) (HCC)	Senile dementia of Alzheimer's type (HCC)
	Sporadic Alzheimer's disease (HCC)

Supplementary Table 2.3 All Diagnosis Network Metrics

Diagnosis Networks are created with nodes representing a diagnostic category or diagnosis shared among >1% of patients in a group, and edges representing >1% of co-diagnosis in a group.

UCSF: Graph (>1%)	Number Nodes	Number Edges	Avg Number Neighbors	Network Diameter	Network Radius	Characteristic Path Length	Clustering Coefficient	Network density	Network heterogeneity	Network Centralization	Connected Components	Singletons
ADDiagnosisNameAll	1056	27504	62.15	4	2	2.043	0.830	0.070	1.626	0.763	171	169
ADDiagnosisNameFemale	962	25459	61.64	4	2	2.038	0.832	0.075	1.586	0.761	137	136
ADDiagnosisNameMale	924	20102	52.90	4	2	2.057	0.823	0.070	1.633	0.739	164	162
ADL3NameALL	483	23505	97.33	2	1	1.798	0.899	0.202	1.054	0.801	1	0
ADL3NameFemale	452	21958	97.16	2	1	1.785	0.899	0.215	1.021	0.788	1	0
ADL3NameMale	445	20099	90.33	2	1	1.797	0.899	0.203	1.046	0.800	1	0
ADL2NameALL	165	7960	96.48	2	1	1.412	0.896	0.588	0.479	0.417	1	0
ADL2NameFemale	160	7531	94.73	2	1	1.400	0.897	0.600	0.469	0.406	2	1
ADL2NameMale	158	7257	91.86	2	1	1.415	0.892	0.585	0.481	0.420	1	0
ConDiagnosisNameAll	421	2445	18.04	4	2	2.048	0.738	0.067	1.671	0.843	151	150
ConDiagnosisNameFemale	167	2109	25.72	3	2	1.892	0.848	0.158	1.156	0.797	4	3
ConDiagnosisNameMale	321	1417	13.43	4	2	2.078	0.681	0.064	1.717	0.815	111	110
ConL3NameALL	318	5772	43.89	2	1	1.832	0.760	0.168	1.135	0.839	56	55
ConL3NameFemale	190	5434	57.20	2	1	1.697	0.830	0.303	0.829	0.705	1	0
ConL3NameMale	282	4195	37.46	3	2	1.837	0.750	0.168	1.125	0.835	59	58
ConL2NameALL	150	3990	55.80	2	1	1.607	0.854	0.393	0.697	0.616	8	7
ConL2NameFemale	122	3760	61.64	2	1	1.491	0.866	0.509	0.552	0.499	1	0
ConL2NameMale	137	3200	48.48	2	1	1.630	0.862	0.370	0.726	0.640	6	5

Mount Sinai: Graph (>1%)	Number Nodes	Number Edges	Avg Number Neighbors	Network Diameter	Network Radius	Characteristic Path Length	Clustering Coefficient	Network density	Network heterogeneity	Network Centralization	Connected Components	Singletons
ADDiagnosisNameAll	483	1788	15.96	4	2	2.030	0.782	0.072	1.696	0.756	260	259
ADDiagnosisNameFemale	482	1753	15.72	4	2	2.035	0.769	0.071	1.700	0.751	260	259
ADDiagnosisNameMale	446	1034	12.16	4	2	2.084	0.722	0.072	1.674	0.700	277	276
ADL3NameALL	348	10434	59.97	2	1	1.827	0.910	0.173	1.152	0.832	1	0
ADL3NameFemale	352	10145	59.68	2	1	1.824	0.909	0.176	1.142	0.829	13	12
ADL3NameMale	332	8162	52.32	2	1	1.832	0.905	0.168	1.166	0.837	21	20
ADL2NameALL	141	4625	65.60	2	1	1.531	0.875	0.469	0.608	0.539	1	0
ADL2NameFemale	141	4480	64.93	2	1	1.526	0.876	0.474	0.602	0.534	4	3
ADL2NameMale	139	4037	59.37	2	1	1.560	0.878	0.440	0.642	0.569	4	3
ConDiagnosisNameAll	461	13	3.25	3	2	1.607	0.558	0.464	0.527	0.524	454	453
ConDiagnosisNameFemale	461	13	3.25	3	2	1.607	0.558	0.464	0.527	0.524	454	453
ConDiagnosisNameMale	423	13	3.25	3	2	1.607	0.558	0.464	0.527	0.524	416	415
ConL3NameALL	347	1038	17.16	2	1	1.857	0.788	0.143	1.267	0.871	227	226
ConL3NameFemale	351	1038	17.16	2	1	1.857	0.788	0.143	1.267	0.871	231	230
ConL3NameMale	331	980	16.47	3	2	1.863	0.780	0.140	1.280	0.867	213	212
ConL2NameALL	141	1323	32.67	3	2	1.594	0.877	0.408	0.677	0.568	61	60
ConL2NameFemale	141	1323	32.67	3	2	1.594	0.877	0.408	0.677	0.568	61	60
ConL2NameMale	139	1290	31.85	3	2	1.606	0.865	0.398	0.690	0.566	59	58

Supplementary Table 2.4 All Diagnosis Network Comparisons

Network metrics are computed for nodes in each network at UCSF and at Mount Sinai, and the distribution of metrics are compared between networks. Comparisons are performed with and without the removal of singletons (single nodes with no neighbors). A Mann-Whitney U-test is performed to compare the distribution of each network metric, with colors based upon p-value cutoff. The mean difference in metric between comparison groups is also shown.

Comparison	Nodes	Mann Whitney U Test	Avg Shortest Path Length	Betweeness Centrality	Closeness Centrality	Clustering Coefficient	Degree	Eccentricity	Neighbor-hood Connecti-vity	Number Undirected Edges	Stress Centrality	Topological Coefficient
UCSF AD vs Control: Diagnosis Name	All Nodes	Stat	2.68E+05	2.46E+05	2.63E+05	2.73E+05	3.07E+05	3.22E+05	3.73E+05	3.07E+05	2.56E+05	2.81E+05
		pval	3.75E-10	9.15E-05	2.31E-08	1.07E-12	1.02E-30	6.87E-48	1.58E-93	1.02E-30	2.48E-08	6.20E-16
		Δmean	3.96E-01	-1.52E-03	9.88E-02	2.20E-01	4.05E+01	1.01E+00	2.57E+02	4.05E+01	1.17E+04	1.35E-01
	Singletons Removed	Stat	1.23E+05	8.88E+03	1.18E+05	8.99E+04	1.61E+05	1.76E+05	2.28E+05	1.61E+05	1.89E+04	9.85E+04
		pval	6.05E-01	4.64E-11	6.05E-01	7.51E-01	1.22E-17	3.06E-42	4.30E-110	1.22E-17	1.31E-02	5.76E-02
		Δmean	-7.26E-03	-9.16E-03	2.03E-03	1.81E-02	4.40E+01	4.97E-01	2.76E+02	4.40E+01	3.19E+04	3.63E-02
UCSF AD vs Control: L3 Name	All Nodes	Stat	8.30E+04	8.19E+04	9.71E+04	1.02E+05	1.10E+05	9.02E+04	1.39E+05	1.10E+05	8.70E+04	1.13E+05
		pval	5.19E-02	7.44E-02	2.16E-10	3.00E-17	2.15E-25	6.25E-21	2.69E-85	2.15E-25	3.50E-04	4.21E-29
		Δmean	2.83E-01	-9.79E-04	1.08E-01	2.70E-01	6.10E+01	3.47E-01	1.76E+02	6.10E+01	6.20E+03	2.50E-01
	Singletons Removed	Stat	5.65E+04	1.03E+04	7.06E+04	5.77E+04	8.36E+04	6.36E+04	1.13E+05	8.36E+04	1.55E+04	6.79E+04
		pval	0.0121	0.0274	1.21E-02	1.03E-01	9.42E-13	6.64E-01	4.02E-69	9.42E-13	4.96E-05	2.53E-08
		Δmean	-3.44E-02	-3.70E-03	1.27E-02	1.80E-02	5.34E+01	1.73E-03	1.49E+02	5.34E+01	1.31E+04	1.00E-01
UCSF AD vs Control: L2 Name	All Nodes	Stat	8.40E+03	1.39E+04	1.75E+04	1.26E+04	1.88E+04	1.29E+04	2.06E+04	1.88E+04	1.50E+04	1.72E+04
		pval	8.40E-07	5.70E-02	2.06E-10	7.73E-01	1.51E-15	3.52E-02	3.65E-24	1.51E-15	9.45E-04	2.48E-09
		Δmean	-1.20E-01	-1.58E-03	1.23E-01	8.24E-02	4.33E+01	8.79E-02	3.31E+01	4.33E+01	7.29E+02	1.39E-01
	Singletons Removed	Stat	7.24E+03	3.93E+03	1.64E+04	1.08E+04	1.77E+04	1.17E+04	1.94E+04	1.77E+04	5.03E+03	1.54E+04
		pval	5.14E-09	0.11	5.14E-09	0.362	5.46E-14	6.51E-01	1.64E-22	5.46E-14	2.07E-01	3.12E-07
		Δmean	-1.95E-01	-4.17E-03	9.26E-02	1.80E-02	4.07E+01	-5.13E-03	2.85E+01	4.07E+01	4.30E+02	8.91E-02

Comparison	Nodes	Mann Whitney U Test	Avg Shortest Path Length	Betweeness Centrality	Closeness Centrality	Clustering Coefficient	Degree	Eccentricity	Neighbor-hood Connecti-vity	Number Undirected Edges	Stress Centrality	Topological Coefficient
UCSF Female AD vs Male AD: Diagnosis Name	All Nodes	Stat	4.57E+05	4.69E+05	4.88E+05	4.73E+05	4.85E+05	3.41E+05	5.22E+05	4.85E+05	4.72E+05	4.78E+05
		pval	5.15E-01	7.16E-01	6.35E-02	4.96E-01	9.92E-02	3.39E-31	2.96E-06	9.92E-02	4.90E-01	2.93E-01
		Δmean	1.98E-02	-8.89E-05	1.02E-02	1.69E-02	8.26E+00	-2.87E-01	4.19E+01	8.26E+00	2.75E+03	1.39E-02
	Singletons Removed	Stat	3.09E+05	5.33E+04	3.39E+05	2.68E+05	3.37E+05	1.92E+05	3.74E+05	3.37E+05	5.67E+04	2.73E+05
		pval	1.01E-01	4.56E-01	1.01E-01	9.79E-01	1.71E-01	1.10E-73	9.98E-08	1.71E-01	5.32E-01	5.68E-01
		Δmean	-1.50E-02	-3.69E-04	2.79E-03	1.27E-03	8.82E+00	-4.06E-01	4.32E+01	8.82E+00	6.85E+03	6.75E-03
UCSF Female AD vs Male AD: L3 Name	All Nodes	Stat	1.05E+05	1.04E+05	1.04E+05	1.05E+05	1.07E+05	1.04E+05	1.13E+05	1.07E+05	1.05E+05	1.06E+05
		pval	9.12E-01	9.13E-01	9.12E-01	9.61E-01	5.28E-01	9.74E-01	3.08E-02	5.28E-01	9.59E-01	7.47E-01
		Δmean	2.59E-03	-7.91E-05	-7.28E-04	-8.94E-04	3.28E+00	1.05E-04	1.36E+01	3.28E+00	9.47E+02	-2.65E-04
	Singletons Removed	Stat	1.05E+05	2.01E+04	1.04E+05	1.03E+05	1.07E+05	1.04E+05	1.13E+05	1.07E+05	2.07E+04	1.04E+05
		pval	0.912	0.461	9.12E-01	9.70E-01	5.28E-01	9.74E-01	3.08E-02	5.28E-01	8.03E-01	7.53E-01
		Δmean	2.59E-03	-2.17E-04	-7.28E-04	-1.29E-03	3.28E+00	1.05E-04	1.36E+01	3.28E+00	1.95E+03	-5.36E-04
UCSF Female AD vs Male AD: L2 Name	All Nodes	Stat	1.24E+04	1.27E+04	1.29E+04	1.27E+04	1.31E+04	1.26E+04	1.36E+04	1.31E+04	1.28E+04	1.29E+04
		pval	7.59E-01	9.08E-01	7.59E-01	9.74E-01	5.62E-01	9.92E-01	2.61E-01	5.62E-01	8.70E-01	7.35E-01
		Δmean	-7.12E-03	-7.87E-05	3.60E-03	5.89E-03	2.30E+00	1.58E-04	2.10E+00	2.30E+00	-6.22E+01	3.46E-03
	Singletons Removed	Stat	1.24E+04	5.31E+03	1.29E+04	1.27E+04	1.31E+04	1.26E+04	1.36E+04	1.31E+04	5.35E+03	1.29E+04
		pval	7.59E-01	0.429	7.59E-01	0.974	5.62E-01	0.992	2.61E-01	5.62E-01	0.48	7.35E-01
		Δmean	-7.12E-03	-3.26E-04	3.60E-03	5.89E-03	2.30E+00	1.58E-04	2.10E+00	2.30E+00	-2.47E+02	3.46E-03
UCSF Female: AD vs Control: Diagnosis Name	All Nodes	Stat	2.47E+05	2.23E+05	2.34E+05	2.47E+05	2.78E+05	2.49E+05	3.40E+05	2.78E+05	2.33E+05	2.55E+05
		pval	3.15E-11	1.16E-04	1.98E-06	6.28E-12	8.80E-29	3.89E-18	5.56E-91	8.80E-29	4.00E-08	3.53E-15
		Δmean	3.93E-01	-1.55E-03	9.58E-02	2.18E-01	3.99E+01	6.46E-01	2.45E+02	3.99E+01	1.14E+04	1.36E-01
	Singletons Removed	Stat	1.17E+05	8.17E+03	1.04E+05	8.21E+04	1.47E+05	1.19E+05	2.10E+05	1.47E+05	1.74E+04	9.03E+04
		pval	1.53E-01	1.32E-10	1.53E-01	7.09E-01	1.68E-16	2.64E-04	1.15E-107	1.68E-16	1.43E-02	5.98E-02
		Δmean	-1.64E-03	-9.37E-03	1.03E-04	1.76E-02	4.33E+01	8.01E-02	2.62E+02	4.33E+01	3.09E+04	3.67E-02
	All Nodes	Stat	7.79E+04	7.70E+04	9.01E+04	9.57E+04	1.02E+05	8.41E+04	1.30E+05	1.02E+05	8.16E+04	1.05E+05

Comparison	Nodes	Mann Whitney U Test	Avg Shortest Path Length	Betweeness Centrality	Closeness Centrality	Clustering Coefficient	Degree	Eccentricity	Neighbor-hood Connecti-vity	Number Undirected Edges	Stress Centrality	Topological Coefficient
UCSF Female: AD vs Control: L3 Name		pval	5.75E-02	7.26E-02	3.32E-09	3.11E-16	3.67E-23	1.91E-19	1.25E-80	3.67E-23	4.57E-04	2.62E-27
		Δmean	2.72E-01	-1.00E-03	1.03E-01	2.67E-01	5.77E+01	3.32E-01	1.67E+02	5.77E+01	5.77E+03	2.46E-01
	Singletons Removed	Stat	5.40E+04	9.75E+03	6.63E+04	5.42E+04	7.85E+04	6.02E+04	1.06E+05	7.85E+04	1.44E+04	6.36E+04
		pval	0.023	0.019	2.30E-02	1.23E-01	1.05E-11	6.68E-01	1.26E-65	1.05E-11	3.64E-04	9.73E-08
		Δmean	-3.09E-02	-3.80E-03	1.15E-02	1.81E-02	5.04E+01	1.75E-03	1.42E+02	5.04E+01	1.19E+04	9.75E-02
UCSF Female: AD vs Control: L2 Name	All Nodes	Stat	7.93E+03	1.29E+04	1.64E+04	1.22E+04	1.76E+04	1.21E+04	1.92E+04	1.76E+04	1.39E+04	1.64E+04
		pval	1.25E-06	1.10E-01	1.13E-09	5.29E-01	1.39E-14	6.37E-02	3.75E-22	1.39E-14	2.95E-03	8.95E-10
		Δmean	-1.25E-01	-1.58E-03	1.18E-01	8.53E-02	4.07E+01	7.65E-02	3.05E+01	4.07E+01	6.27E+02	1.42E-01
	Singletons Removed	Stat	6.97E+03	3.97E+03	1.54E+04	1.04E+04	1.67E+04	1.11E+04	1.82E+04	1.67E+04	5.00E+03	1.47E+04
		pval	1.70E-08	0.253	1.70E-08	0.573	2.92E-13	0.645	9.58E-21	2.92E-13	0.106	1.27E-07
		Δmean	-1.91E-01	-3.84E-03	9.11E-02	1.91E-02	3.84E+01	-5.36E-03	2.65E+01	3.84E+01	4.35E+02	9.01E-02
UCSF Males: AD vs Control: Diagnosis Name	All Nodes	Stat	1.76E+05	1.67E+05	1.81E+05	1.91E+05	2.11E+05	2.16E+05	2.52E+05	2.11E+05	1.74E+05	1.94E+05
		pval	1.84E-05	1.18E-03	2.82E-07	5.34E-13	3.46E-26	3.29E-34	1.28E-71	3.46E-26	9.63E-07	2.09E-14
		Δmean	3.44E-01	-2.21E-03	8.95E-02	2.38E-01	3.47E+01	9.30E-01	2.19E+02	3.47E+01	9.11E+03	1.42E-01
	Singletons Removed	Stat	7.94E+04	6.72E+03	8.42E+04	5.83E+04	1.15E+05	1.19E+05	1.56E+05	1.15E+05	1.43E+04	6.18E+04
		pval	5.14E-01	5.87E-07	5.14E-01	6.58E-01	2.12E-19	8.54E-33	8.81E-91	2.12E-19	4.09E-05	1.02E-01
		Δmean	-2.25E-02	-1.24E-02	5.42E-03	2.49E-02	3.92E+01	4.96E-01	2.43E+02	3.92E+01	2.60E+04	3.46E-02
UCSF Males: AD vs Control: L3 Name	All Nodes	Stat	6.90E+04	6.68E+04	8.31E+04	8.75E+04	9.24E+04	5.06E+04	1.18E+05	9.24E+04	7.20E+04	9.50E+04
		pval	3.34E-02	1.31E-01	5.52E-13	1.91E-20	4.43E-26	1.11E-09	8.58E-88	4.43E-26	3.40E-04	9.71E-31
		Δmean	3.35E-01	-1.19E-03	1.28E-01	3.03E-01	6.11E+01	4.82E-03	1.77E+02	6.11E+01	5.27E+03	2.74E-01
	Singletons Removed	Stat	4.31E+04	8.47E+03	5.72E+04	4.67E+04	6.65E+04	2.47E+04	9.23E+04	6.65E+04	1.36E+04	5.42E+04
		pval	0.00297	0.0189	2.97E-03	2.26E-02	5.72E-12	6.79E-61	1.56E-70	5.72E-12	1.35E-06	1.52E-08
		Δmean	-4.15E-02	-4.27E-03	1.46E-02	2.35E-02	5.34E+01	-5.09E-01	1.50E+02	5.34E+01	1.12E+04	1.11E-01
	All Nodes	Stat	6.83E+03	1.17E+04	1.56E+04	1.11E+04	1.69E+04	1.12E+04	1.88E+04	1.69E+04	1.29E+04	1.49E+04

Comparison	Nodes	Mann Whitney U Test	Avg Shortest Path Length	Betweeness Centrality	Closeness Centrality	Clustering Coefficient	Degree	Eccentricity	Neighbor-hood Connecti-vity	Number Undirected Edges	Stress Centrality	Topological Coefficient
UCSF Males: AD vs Control: L2 Name		pval	4.73E-08	2.10E-01	6.05E-11	7.50E-01	1.41E-16	9.67E-02	1.55E-27	1.41E-16	3.31E-03	2.52E-08
		Δmean	-1.56E-01	-2.01E-03	1.24E-01	6.13E-02	4.51E+01	6.76E-02	3.51E+01	4.51E+01	8.48E+02	1.29E-01
	Singleton Removed	Stat	6.04E+03	3.88E+03	1.48E+04	9.94E+03	1.61E+04	1.04E+04	1.80E+04	1.61E+04	5.08E+03	1.38E+04
		pval	6.97E-10	0.808	6.97E-10	0.635	2.09E-15	0.674	2.50E-26	2.09E-15	0.00134	5.65E-07
		Δmean	-2.15E-01	-4.23E-03	1.01E-01	1.66E-02	4.34E+01	-5.08E-03	3.19E+01	4.34E+01	1.02E+03	9.47E-02
UCSF Male vs Female Controls: Diagnosis Name	All Nodes	Stat	6.07E+04	6.58E+04	7.05E+04	6.87E+04	6.88E+04	6.59E+04	7.18E+04	6.88E+04	6.63E+04	6.85E+04
		pval	6.60E-02	9.93E-01	8.92E-02	2.69E-01	2.80E-01	9.75E-01	3.14E-02	2.80E-01	8.26E-01	3.22E-01
		Δmean	-2.94E-02	-7.56E-04	3.84E-03	3.70E-02	3.02E+00	-2.01E-03	1.58E+01	3.02E+00	4.12E+02	1.96E-02
	Singletons Removed	Stat	2.30E+04	3.08E+03	3.28E+04	1.83E+04	3.11E+04	2.81E+04	3.40E+04	3.11E+04	3.54E+03	1.80E+04
		pval	9.78E-04	6.06E-01	9.78E-04	6.06E-01	3.14E-02	7.59E-01	3.46E-05	3.14E-02	2.82E-02	8.31E-01
		Δmean	-3.59E-02	-3.42E-03	8.11E-03	8.57E-03	4.71E+00	1.03E-02	2.48E+01	4.71E+00	1.99E+03	4.64E-03
UCSF Male vs Female Controls: L3 Name	All Nodes	Stat	4.40E+04	4.32E+04	4.66E+04	4.59E+04	4.67E+04	3.06E+04	5.20E+04	4.67E+04	4.39E+04	4.62E+04
		pval	8.32E-01	8.45E-01	1.48E-01	2.39E-01	1.30E-01	4.27E-13	4.51E-05	1.30E-01	8.74E-01	1.97E-01
		Δmean	6.56E-02	-2.72E-04	2.37E-02	3.49E-02	6.69E+00	-3.27E-01	2.38E+01	6.69E+00	4.43E+02	2.81E-02
	Singletons Removed	Stat	2.76E+04	5.40E+03	3.02E+04	2.17E+04	3.03E+04	1.42E+04	3.56E+04	3.03E+04	6.04E+03	2.20E+04
		pval	0.403	0.645	4.03E-01	4.86E-01	3.55E-01	6.24E-39	1.06E-05	3.55E-01	3.41E-01	3.75E-01
		Δmean	-8.03E-03	-6.85E-04	2.41E-03	4.14E-03	6.26E+00	-5.11E-01	2.23E+01	6.26E+00	1.27E+03	1.31E-02
UCSF Male vs Female Controls: L2 Name	All Nodes	Stat	9.32E+03	9.95E+03	1.06E+04	9.74E+03	1.09E+04	9.96E+03	1.16E+04	1.09E+04	1.02E+04	9.95E+03
		pval	3.26E-01	9.35E-01	3.96E-01	6.97E-01	1.73E-01	8.68E-01	2.06E-02	1.73E-01	8.08E-01	9.43E-01
		Δmean	-3.71E-02	-5.06E-04	9.92E-03	-1.81E-02	6.74E+00	-8.75E-03	6.73E+00	6.74E+00	1.59E+02	-9.02E-03
	Singleton Removed	Stat	8.61E+03	3.17E+03	9.87E+03	8.76E+03	1.02E+04	9.24E+03	1.09E+04	1.02E+04	3.38E+03	8.97E+03
		pval	3.31E-01	0.761	3.31E-01	0.976	1.30E-01	0.971	1.15E-02	1.30E-01	0.291	0.758
		Δmean	-3.09E-02	-7.11E-04	1.34E-02	3.35E-03	7.26E+00	4.33E-04	7.45E+00	7.26E+00	3.42E+02	8.03E-03

Comparison	Nodes	Mann Whitney U Test	Avg Shortest Path Length	Betweeness Centrality	Closeness Centrality	Clustering Coefficient	Degree	Eccentricity	Neighbor-hood Connect-ivity	Number Undirected Edges	Stress Centrality	Topological Coefficient
Mount Sinai AD vs Control: Diagnosis Name	All Nodes	Stat	1.62E+05	1.28E+05	1.60E+05	1.55E+05	1.61E+05	1.62E+05	1.62E+05	1.61E+05	1.29E+05	1.55E+05
		pval	6.92E-57	9.38E-17	2.49E-54	1.21E-48	1.91E-56	6.59E-58	1.48E-57	1.91E-56	3.27E-17	1.27E-47
		Δmean	9.13E-01	3.95E-04	2.20E-01	3.53E-01	7.35E+00	1.37E+00	5.33E+01	7.35E+00	4.23E+02	2.08E-01
	Singletons Removed	Stat	1.49E+03	1.20E+01	3.06E+02	7.93E+02	1.28E+03	1.47E+03	1.79E+03	1.28E+03	2.68E+02	3.96E+02
		pval	1.55E-03	1.98E-03	1.55E-03	1.04E-01	3.82E-02	3.72E-09	1.56E-06	3.82E-02	2.04E-02	1.74E-01
		Δmean	4.23E-01	-1.89E-01	-1.49E-01	1.49E-01	1.27E+01	6.79E-01	1.10E+02	1.27E+01	2.57E+03	-1.13E-01
Mount Sinai AD vs Control: L3 Name	All Nodes	Stat	9.83E+04	7.77E+04	1.01E+05	1.02E+05	1.09E+05	9.98E+04	1.19E+05	1.09E+05	7.97E+04	1.05E+05
		pval	4.40E-48	4.50E-17	2.23E-55	6.00E-62	4.78E-78	1.82E-74	1.90E-113	4.78E-78	9.06E-21	4.25E-66
		Δmean	1.18E+00	-1.21E-04	3.65E-01	6.35E-01	5.40E+01	1.30E+00	1.95E+02	5.40E+01	3.59E+03	4.59E-01
	Singletons Removed	Stat	1.96E+04	1.87E+03	2.25E+04	1.82E+04	3.04E+04	2.11E+04	4.06E+04	3.04E+04	3.82E+03	2.10E+04
		pval	0.263	0.000503	2.63E-01	8.41E-01	4.24E-13	7.67E-01	2.67E-52	4.24E-13	2.76E-03	3.08E-02
		Δmean	-2.98E-02	-1.61E-02	1.02E-02	1.04E-02	4.28E+01	2.52E-03	1.45E+02	4.28E+01	7.72E+03	5.50E-02
Mount Sinai AD vs Control: L2 Name	All Nodes	Stat	1.34E+04	1.20E+04	1.49E+04	1.41E+04	1.69E+04	1.35E+04	1.96E+04	1.69E+04	1.28E+04	1.49E+04
		pval	2.63E-07	6.47E-04	3.71E-13	3.72E-10	1.79E-24	3.02E-12	1.69E-45	1.79E-24	1.92E-06	4.61E-13
		Δmean	6.15E-01	-5.00E-04	3.05E-01	3.71E-01	4.68E+01	7.80E-01	7.09E+01	4.68E+01	1.43E+03	3.33E-01
	Singletons Removed	Stat	4.99E+03	1.27E+03	6.43E+03	5.63E+03	8.44E+03	5.07E+03	1.11E+04	8.44E+03	2.09E+03	6.41E+03
		pval	1.18E-01	0.15	1.18E-01	0.856	3.22E-09	0.00012	6.39E-32	3.22E-09	0.00077	0.128
		Δmean	-6.30E-02	-7.58E-03	2.93E-02	-1.83E-03	3.29E+01	-1.13E-01	4.71E+01	3.29E+01	2.31E+03	3.40E-02
Mount Sinai Female AD vs Male AD: Diagnosis Name	All Nodes	Stat	1.14E+05	1.09E+05	1.19E+05	1.17E+05	1.18E+05	1.15E+05	1.24E+05	1.18E+05	1.09E+05	1.16E+05
		pval	8.97E-02	5.42E-01	2.08E-03	5.16E-03	4.75E-03	3.29E-02	8.66E-06	4.75E-03	4.39E-01	1.02E-02
		Δmean	1.47E-01	-2.93E-04	4.45E-02	8.06E-02	2.64E+00	2.25E-01	2.12E+01	2.64E+00	1.52E+02	4.46E-02
	Singletons Removed	Stat	1.64E+04	1.99E+03	2.15E+04	1.44E+04	2.05E+04	1.78E+04	2.65E+04	2.05E+04	2.40E+03	1.37E+04
		pval	2.31E-02	6.72E-02	2.31E-02	2.50E-01	1.52E-01	4.96E-02	1.26E-11	1.52E-01	9.25E-01	8.51E-01
		Δmean	-4.90E-02	-3.85E-03	1.03E-02	2.26E-02	3.56E+00	-6.38E-02	3.12E+01	3.56E+00	7.35E+02	3.32E-03

Comparison	Nodes	Mann Whitney U Test	Avg Shortest Path Length	Betweeness Centrality	Closeness Centrality	Clustering Coefficient	Degree	Eccentricity	Neighbor-hood Connect-ivity	Number Undirected Edges	Stress Centrality	Topological Coefficient
Mount Sinai Female AD vs Male AD: L3 Name	All Nodes	Stat	5.94E+04	5.86E+04	6.06E+04	6.06E+04	6.26E+04	5.98E+04	6.77E+04	6.26E+04	5.93E+04	6.14E+04
		pval	7.17E-01	9.39E-01	4.12E-01	3.51E-01	1.03E-01	1.60E-01	3.56E-04	1.03E-01	7.02E-01	2.43E-01
		Δmean	4.04E-02	-1.67E-04	1.72E-02	2.79E-02	8.47E+00	4.96E-02	2.42E+01	8.47E+00	7.39E+02	2.51E-02
	Singletons Removed	Stat	5.24E+04	9.11E+03	5.36E+04	5.34E+04	5.57E+04	5.29E+04	6.07E+04	5.57E+04	9.81E+03	5.42E+04
		pval	8.06E-01	6.18E-01	8.06E-01	8.18E-01	2.63E-01	6.15E-01	1.36E-03	2.63E-01	5.75E-01	5.84E-01
		Δmean	-7.80E-03	-5.46E-04	2.77E-03	1.46E-03	7.36E+00	-2.68E-03	1.97E+01	7.36E+00	1.67E+03	6.90E-03
Mount Sinai Female AD vs Male AD: L2 Name	All Nodes	Stat	9.16E+03	9.68E+03	1.04E+04	9.94E+03	1.06E+04	9.80E+03	1.07E+04	1.06E+04	9.79E+03	1.03E+04
		pval	0.348	0.853	0.344	0.824	0.251	0.984	0.206	0.251	0.991	0.446
		Δmean	-0.033	-0.000305	0.016	-0.00158	5.46	0.000816	3.44	5.46	1.58E+02	0.0148
	Singletons Removed	Stat	8.74E+03	2.79E+03	1.00E+04	9.52E+03	1.02E+04	9.39E+03	1.02E+04	1.02E+04	2.90E+03	9.90E+03
		pval	0.33	0.518	0.33	0.821	0.238	0.991	0.193	0.238	0.278	0.434
		Δmean	-0.0342	-0.000165	0.0161	-0.00189	5.56	0.000213	3.49	5.56	4.69E+02	0.0149
Mount Sinai Female: AD vs Control: Diagnosis Name	All Nodes	Stat	1.61E+05	1.28E+05	1.60E+05	1.54E+05	1.61E+05	1.61E+05	1.61E+05	1.61E+05	1.28E+05	1.54E+05
		pval	1.05E-56	2.60E-16	4.01E-54	1.67E-47	3.26E-56	1.07E-57	2.35E-57	3.26E-56	9.41E-17	1.74E-46
		Δmean	9.14E-01	4.12E-04	2.19E-01	3.46E-01	7.22E+00	1.37E+00	5.27E+01	7.22E+00	4.13E+02	2.03E-01
	Singletons Removed	Stat	1.49E+03	1.20E+01	2.96E+02	7.81E+02	1.26E+03	1.46E+03	1.78E+03	1.26E+03	2.59E+02	3.80E+02
		pval	1.35E-03	2.04E-03	1.35E-03	9.68E-02	4.74E-02	3.99E-09	1.58E-06	4.74E-02	2.27E-02	1.57E-01
		Δmean	4.28E-01	-1.89E-01	-1.51E-01	1.49E-01	1.25E+01	6.79E-01	1.10E+02	1.25E+01	2.57E+03	-1.17E-01
Mount Sinai Female: AD vs Control: L3 Name	All Nodes	Stat	9.85E+04	7.88E+04	1.02E+05	1.03E+05	1.09E+05	1.00E+05	1.19E+05	1.09E+05	8.07E+04	1.05E+05
		pval	3.74E-44	2.28E-16	5.53E-52	1.98E-57	1.26E-72	7.58E-67	7.20E-105	1.26E-72	8.06E-20	6.21E-62
		Δmean	1.12E+00	-1.28E-04	3.50E-01	6.07E-01	5.17E+01	1.24E+00	1.83E+02	5.17E+01	3.40E+03	4.42E-01
	Singletons Removed	Stat	1.89E+04	1.85E+03	2.22E+04	1.78E+04	2.97E+04	2.06E+04	3.95E+04	2.97E+04	3.73E+03	2.06E+04
		pval	0.197	0.000661	1.97E-01	8.13E-01	3.61E-13	7.82E-01	2.38E-51	3.61E-13	3.23E-03	2.55E-02
		Δmean	-3.31E-02	-1.60E-02	1.14E-02	9.48E-03	4.25E+01	2.38E-03	1.41E+02	4.25E+01	7.53E+03	5.71E-02

Comparison	Nodes	Mann Whitney U Test	Avg Shortest Path Length	Betweeness Centrality	Closeness Centrality	Clustering Coefficient	Degree	Eccentricity	Neighbor-hood Connect-ivity	Number Undirected Edges	Stress Centrality	Topological Coefficient
Mount Sinai Female: AD vs Control: L2 Name	All Nodes	Stat	1.32E+04	1.19E+04	1.47E+04	1.39E+04	1.66E+04	1.33E+04	1.92E+04	1.66E+04	1.27E+04	1.47E+04
		pval	1.65E-06	1.52E-03	2.38E-12	1.99E-09	7.39E-23	8.21E-11	2.70E-42	7.39E-23	7.09E-06	3.44E-12
		Δmean	5.78E-01	-5.37E-04	2.94E-01	3.53E-01	4.48E+01	7.38E-01	6.69E+01	4.48E+01	1.34E+03	3.19E-01
	Singletons Removed	Stat	4.83E+03	1.26E+03	6.34E+03	5.54E+03	8.27E+03	4.96E+03	1.09E+04	8.27E+03	2.06E+03	6.31E+03
		pval	9.56E-02	0.239	9.56E-02	0.911	3.08E-09	0.000136	3.16E-31	3.08E-09	0.000274	0.112
		Δmean	-6.84E-02	-7.35E-03	3.19E-02	-1.29E-03	3.23E+01	-1.13E-01	4.52E+01	3.23E+01	2.26E+03	3.65E-02
Mount Sinai Male: AD vs Control: Diagnosis Name	All Nodes	Stat	1.29E+05	1.07E+05	1.28E+05	1.23E+05	1.29E+05	1.29E+05	1.29E+05	1.29E+05	1.07E+05	1.23E+05
		pval	2.85E-40	6.02E-13	3.40E-38	1.97E-32	1.05E-39	1.36E-40	9.84E-41	1.05E-39	2.27E-13	8.08E-32
		Δmean	7.64E-01	5.47E-04	1.74E-01	2.65E-01	4.58E+00	1.14E+00	3.15E+01	4.58E+00	2.61E+02	1.57E-01
	Singletons Removed	Stat	1.15E+03	1.00E+01	2.07E+02	5.52E+02	8.94E+02	1.13E+03	1.36E+03	8.94E+02	2.36E+02	2.91E+02
		pval	9.02E-04	2.24E-03	9.02E-04	1.70E-01	1.32E-01	1.93E-06	1.82E-06	1.32E-01	3.76E-03	1.98E-01
		Δmean	4.77E-01	-1.85E-01	-1.61E-01	1.26E-01	8.91E+00	7.43E-01	7.83E+01	8.91E+00	1.84E+03	-1.20E-01
Mount Sinai Male: AD vs Control: L3 Name	All Nodes	Stat	1.29E+05	1.07E+05	1.28E+05	1.23E+05	1.29E+05	1.29E+05	1.29E+05	1.29E+05	1.07E+05	1.23E+05
		pval	2.85E-40	6.02E-13	3.40E-38	1.97E-32	1.05E-39	1.36E-40	9.84E-41	1.05E-39	2.27E-13	8.08E-32
		Δmean	7.64E-01	5.47E-04	1.74E-01	2.65E-01	4.58E+00	1.14E+00	3.15E+01	4.58E+00	2.61E+02	1.57E-01
	Singletons Removed	Stat	1.15E+03	1.00E+01	2.07E+02	5.52E+02	8.94E+02	1.13E+03	1.36E+03	8.94E+02	2.36E+02	2.91E+02
		pval	0.000902	0.00224	9.02E-04	1.70E-01	1.32E-01	1.93E-06	1.82E-06	1.32E-01	3.76E-03	1.98E-01
		Δmean	4.77E-01	-1.85E-01	-1.61E-01	1.26E-01	8.91E+00	7.43E-01	7.83E+01	8.91E+00	1.84E+03	-1.20E-01
Mount Sinai Male: AD vs Control: L2 Name	All Nodes	Stat	1.30E+04	1.17E+04	1.40E+04	1.35E+04	1.58E+04	1.28E+04	1.85E+04	1.58E+04	1.24E+04	1.41E+04
		pval	6.41E-07	8.24E-04	8.73E-11	4.08E-09	4.30E-20	8.55E-10	5.46E-40	4.30E-20	4.95E-06	3.24E-11
		Δmean	5.91E-01	-3.78E-04	2.75E-01	3.55E-01	3.95E+01	7.12E-01	6.31E+01	3.95E+01	1.18E+03	3.08E-01
	Singleton Removed	Stat	5.00E+03	1.23E+03	6.01E+03	5.37E+03	7.81E+03	4.82E+03	1.05E+04	7.81E+03	1.96E+03	5.97E+03
		pval	2.61E-01	0.0909	2.61E-01	0.872	2.80E-07	5.81E-05	5.09E-29	2.80E-07	0.00993	0.232
		Δmean	-4.56E-02	-7.47E-03	2.05E-02	2.06E-03	2.75E+01	-1.26E-01	4.20E+01	2.75E+01	1.80E+03	2.81E-02

Comparison	Nodes	Mann Whitney U Test	Avg Shortest Path Length	Betweeness Centrality	Closeness Centrality	Clustering Coefficient	Degree	Eccentricity	Neighbor-hood Connect-ivity	Number Undirected Edges	Stress Centrality	Topological Coefficient
Mount Sinai Male vs Female Controls: Diagnosis Name	All Nodes	Stat	9.73E+04	9.74E+04	9.73E+04	9.74E+04	9.73E+04	9.73E+04	9.73E+04	9.73E+04	9.74E+04	9.74E+04
		pval	0.863	0.903	0.863	0.881	0.863	0.863	0.863	0.863	0.903	0.881
		Δmean	-0.00251	-0.000158	-0.00101	-0.00087	-0.00507	-0.0037	-0.00716	-0.00507	-0.00974	-0.000757
	Singletons Removed	Stat	3.20E+01	7.50E+00	3.20E+01	18	3.20E+01	3.20E+01	3.20E+01	3.20E+01	8.00E+00	18
		pval	0.957	1	0.957	0.933	0.957	0.95	0.958	0.957	0.876	0.935
		Δmean	0	-2.78E-17	0	0	0	0	0	0	0	0
Mount Sinai Male vs Female Controls: L3 Name	All Nodes	Stat	5.71E+04	5.80E+04	5.73E+04	5.75E+04	5.74E+04	5.59E+04	5.76E+04	5.74E+04	5.80E+04	5.77E+04
		pval	0.661	0.958	0.731	0.789	0.76	0.297	0.824	0.76	0.973	0.84
		Δmean	-0.0298	-0.00017	-0.00736	-0.00852	-0.00692	-0.0989	-0.551	-0.00692	3.69	-0.00383
	Singletons Removed	Stat	7.10E+03	7.34E+02	7.30E+03	5.43E+03	7.39E+03	5.82E+03	7.57E+03	7.39E+03	7.60E+02	5.57E+03
		pval	0.847	0.799	0.847	0.946	0.727	4.94E-07	0.492	0.727	0.996	0.802
		Δmean	-0.00638	-0.00132	0.00195	-0.00083	0.686	-0.193	1.6	0.686	40.6	0.00488
Male vs Female Controls: L2 Name	All Nodes	Stat	9.63E+03	9.77E+03	9.80E+03	9.81E+03	9.80E+03	9.68E+03	9.82E+03	9.80E+03	9.77E+03	9.91E+03
		pval	0.798	0.955	0.994	0.984	0.998	0.84	0.98	0.998	0.958	0.868
		Δmean	-0.0198	-0.000146	-0.00259	-9.77E-05	0.205	-0.0245	-0.274	0.205	-0.956	0.00301
	Singleton Removed	Stat	3.19E+03	8.50E+02	3.37E+03	3.26E+03	3.36E+03	3.24E+03	3.38E+03	3.36E+03	8.52E+02	3.36E+03
		pval	0.775	0.93	0.775	0.933	0.794	0.801	0.744	0.794	0.915	0.687
		Δmean	-0.0114	-0.000286	0.00475	0.00146	0.815	-0.0123	0.323	0.815	10	0.00653

Color Legend
pval < .05
pval < .01
pval < .001
pval < .0001
positive Δmean
negative Δmean

2.9 Supplementary Data

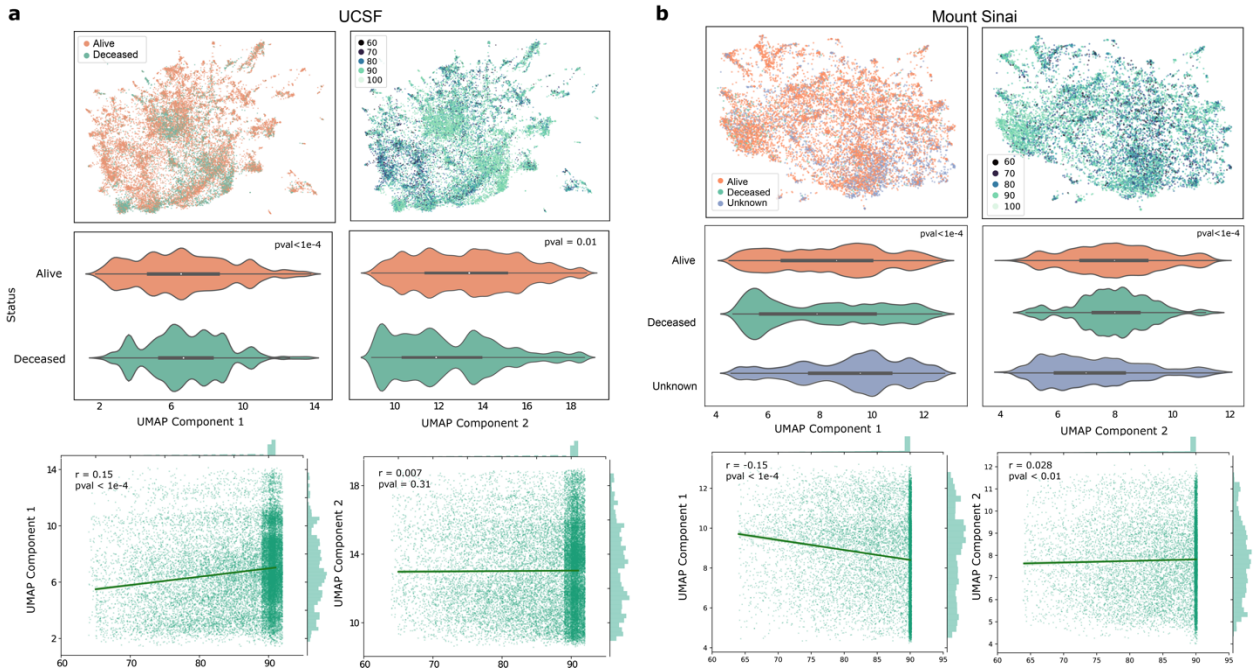
Supplementary Data 2.1 Thresholded Full Tables of Diagnosis Enrichment Analysis

An excel sheet with 3 levels of diagnosis categories and sex-specific analysis (6 tabs) for each institution (12 tabs total). Lists include diagnosis enriched between AD and control cohorts, and sex-specific enrichments. Diagnoses are thresholded to represent > 10 patients, with uncorrected p-values (from two-sided Fisher Exact or Chi Square test) and odds ratios. The data from UCSF can be visualized and explored in the Rshiny app: vizad.org.
Data can be downloaded at this link:
https://static-content.springer.com/esm/art%3A10.1038%2Fs41467-022-28273-0/MediaObjects/41467_2022_28273_MOESM4_ESM.xlsx

Supplementary Data 2.2 Encounter Controlled Diagnosis Enrichment Analysis

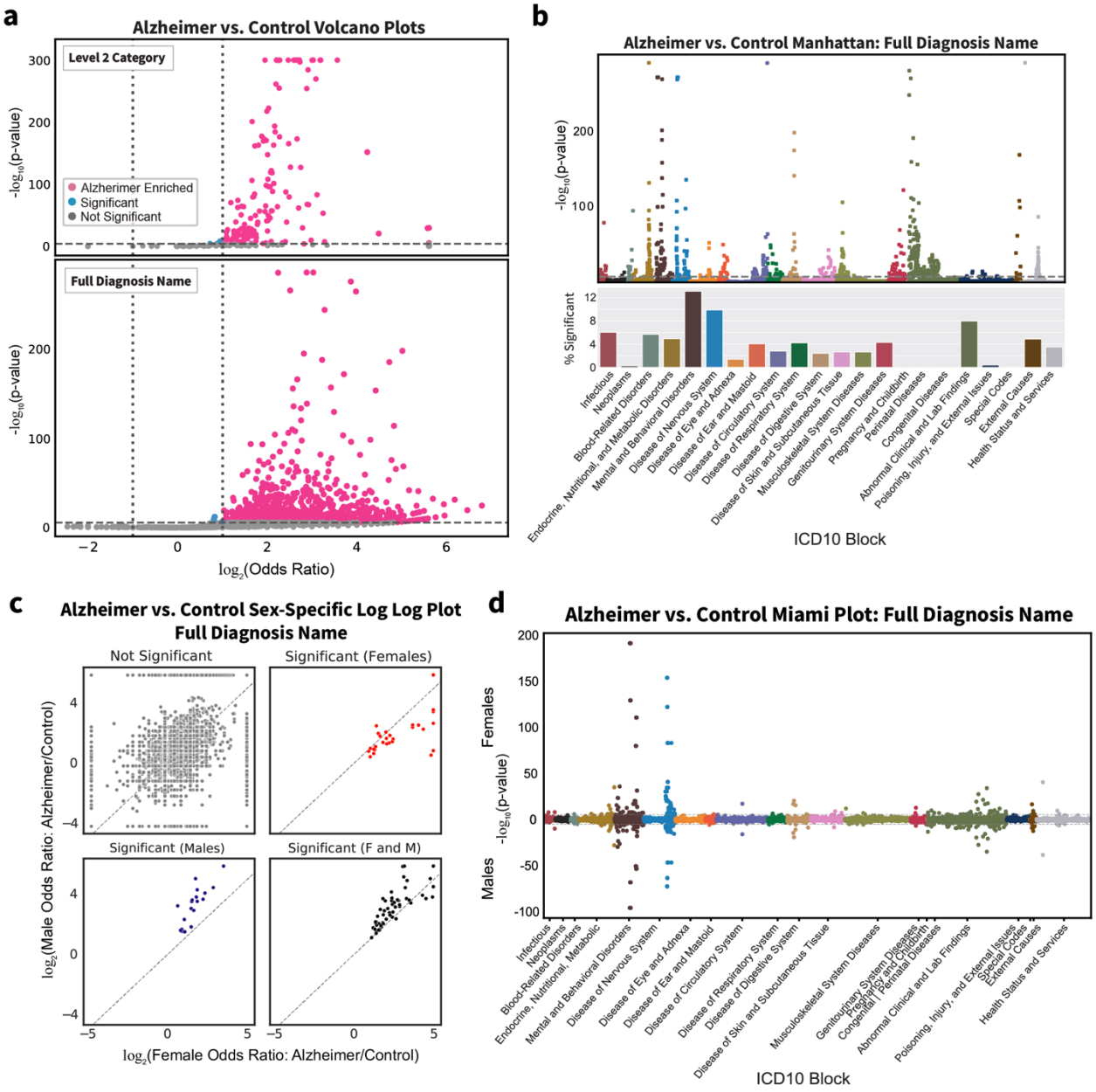
An excel sheets with 3 levels of diagnosis categories on encounter-controlled control cohorts (described in Methods) and sex-specific analysis at UCSF (6 tabs). Lists include diagnoses enriched between AD and control cohorts, and sex-specific enrichments. Diagnoses are thresholded to represent > 10 patients, with un-corrected p-values (from two-sided Fisher Exact or Chi Square test) and odds-ratios. The data can be visualized and explored in the Rshiny app: vizad.org.
Data can be downloaded at this link:
https://static-content.springer.com/esm/art%3A10.1038%2Fs41467-022-28273-0/MediaObjects/41467_2022_28273_MOESM5_ESM.xlsx

2.10 Supplementary Figures



Supplementary Figure 2.1 Demographic correlation across UMAP principal components

a. The top two graphs show the UMAP of AD and controls at UCSF, colored by deceased status (left) and estimated age (right). The middle graphs show distribution of deceased status among the two UMAP components, which are compared with a Mann-Whitney U-Test. The bottom graphs show estimated age across the two UMAP components, with marginal distributions shown on the sides. A regression line is plotted, and a Pearson's R correlation test is performed.
b. The same UMAP plots are shown as in **a**, but for Mount Sinai.



Supplementary Figure 2.2 Comorbidity Enrichment Analysis identifies diagnosis in AD vs. Controls and Sex-Specific Enrichments at Mount Sinai

a. Volcano plot for Level 2 categories (top) and full diagnosis names (bottom) compared between AD and control cohorts using Fisher Exact or Chi-Squared test. P-value cutoff is Bonferroni corrected at 0.05 with log2 odds ratio cutoff at 1 for AD enriched (pink) and remaining significant diagnoses in blue.

b. Above, a Manhattan plot with full diagnosis names colored by ICD-10 categories with Bonferroni-corrected p-value cutoff of 0.05. Bottom, percentage of diagnosis in each ICD-10 category that is significant.

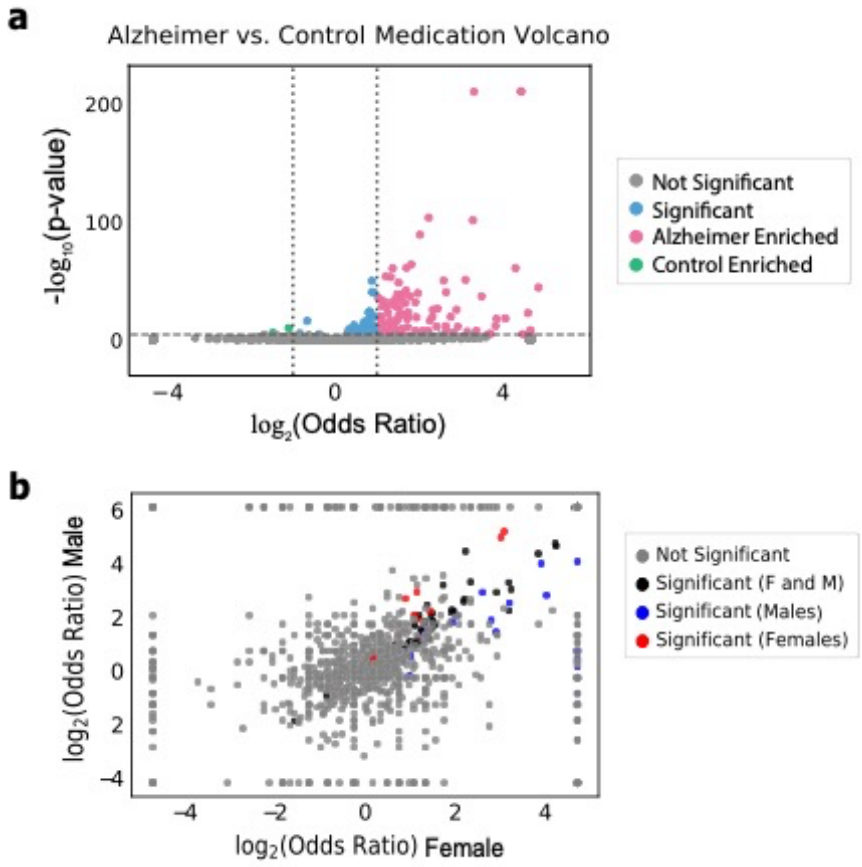
c. Full diagnosis names compared between AD and controls within each sex. The log of the

(Figure caption continued on the next page.)

(Figure caption continued from the previous page.)

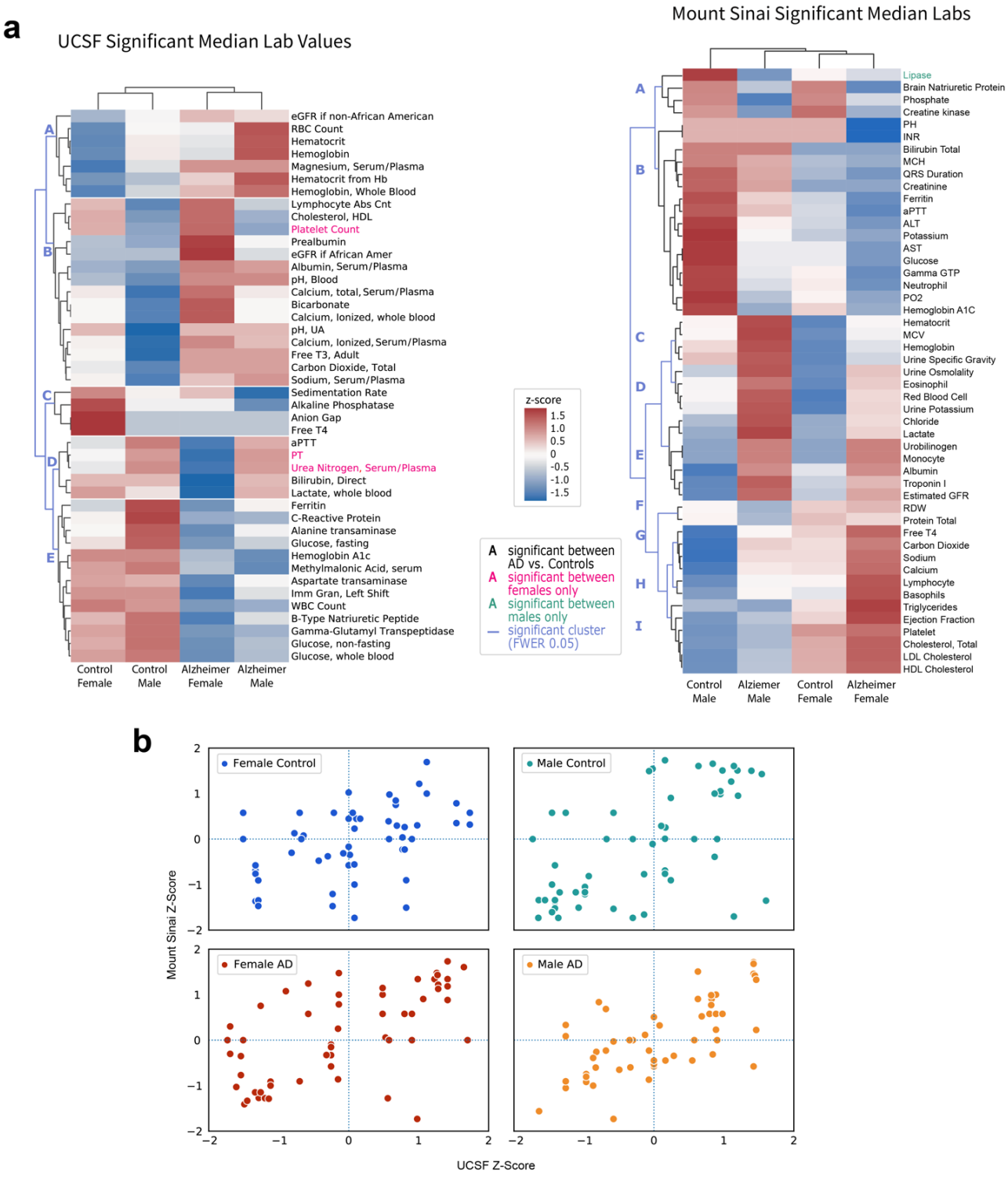
odds ratio is plotted on the axis, and points are colored by significance (Bonferroni-corrected, p-val cutoff > 3e-6).

d. Miami plot of the diagnosis names grouped by sex and ICD-10 categories.



Supplementary Figure 2.3 Medication Enrichment Analysis identifies Enriched Medications between AD and Controls

a. Volcano plot for generic medication names compared between AD and controls using Fisher Exact or Chi-Squared Test. P-value cutoff is Bonferroni-corrected at 0.05 with odds ratio cutoff at 2 for AD enriched (pink) or 1/2 for controlled enriched (green). Remaining significant diagnoses are in blue.
b. Log-log plot of generic medication names compared between AD and controls within each sex. The log of the odds ratio for each sex is plotted on the axis, with points colored by significance (Bonferroni-corrected p-value of 0.05) if female only (red), male only (blue), or both (black).



Supplementary Figure 2.4 Stratifying by AD status and sex allows identification of lab trends between groups

a. Heatmap of lab values filtered on significance at UCSF in AD vs control comparison across

(Figure caption continued on the next page.)

(Figure caption continued from the previous page.)

sex-specific groups. Labs are clustered with light blue lines representing significant cluster breaks (FWER corrected p-value 0.05). Text color represents significant labs among females only (pink), or significant between AD vs controls (black). Heatmap colors represent z-score of the average median value across the 4 groups.

b. Heatmap of lab values filtered on significance at Mount Sinai in AD vs control comparison across sex-specific groups. Labs are clustered with light blue lines representing significant cluster breaks (FWER corrected p-value 0.05). Text color represents significant labs among males only (green), or significant between AD vs controls (black). Heatmap colors represent z-score of the average median value across the 4 groups.
c. Comparison of z-scored lab values between UCSF and Mount Sinai showing significant correlations within each AD/sex-stratified groups. Female control: Spearman $\rho = 0.45$, p-value < 0.001; Male control: 0.46, p-value < 0.001; Female AD: 0.59, p-value $<$ 1e-5; Male AD: 0.64, p-value $<$ 1e-5.

2.11 References

1. 2022 Alzheimer's disease facts and figures. *Alzheimers Dement.* **18**, 700–789 (2022).
2. Rasmussen, J. & Langerman, H. Alzheimer's Disease – Why We Need Early Diagnosis. *Degener. Neurol. Neuromuscul. Dis.* **Volume 9**, 123–130 (2019).
3. Kivipelto, M. Midlife vascular risk factors and Alzheimer's disease in later life: longitudinal, population based study. *BMJ* **322**, 1447–1451 (2001).
4. Niculescu, A. B. *et al.* Blood biomarkers for memory: toward early detection of risk for Alzheimer disease, pharmacogenomics, and repurposed drugs. *Mol. Psychiatry* **25**, 1651–1672 (2020).
5. Alena V. Savonenko, Philip C. Wong, & Tong Li. Alzheimer diseases. (2023) doi:10.1016/b978-0-323-85654-6.00022-8.
6. Neugroschl, J. & Wang, S. Alzheimer's Disease: Diagnosis and Treatment Across the Spectrum of Disease Severity. *Mt. Sinai J. Med. N. Y.* **78**, 596–612 (2011).
7. Tang, A. S. *et al.* Deep phenotyping of Alzheimer's disease leveraging electronic medical records identifies sex-specific clinical associations. *Nat. Commun.* **13**, 675 (2022).
8. Taubes, A. *et al.* Experimental and real-world evidence supporting the computational repurposing of bumetanide for APOE4-related Alzheimer's disease. *Nat. Aging* **1**, 932–947 (2021).
9. Ben Miled, Z. *et al.* Predicting dementia with routine care EMR data. *Artif. Intell. Med.* **102**, 101771 (2020).
10. Tang, A., Woldemariam, S., Roger, J. & Sirota, M. Translational Bioinformatics to Enable Precision Medicine for All: Elevating Equity across Molecular, Clinical, and Digital Realms. *Yearb. Med. Inform.* **31**, 106–115 (2022).

11. Xu, J. *et al.* Data-driven discovery of probable Alzheimer's disease and related dementia subphenotypes using electronic health records. *Learn. Health Syst.* **4**, e10246 (2020).

12. Park, J. H. *et al.* Machine learning prediction of incidence of Alzheimer's disease using large-scale administrative health data. *Npj Digit. Med.* **3**, 46 (2020).

13. Qiu, S. *et al.* Multimodal deep learning for Alzheimer's disease dementia assessment. *Nat. Commun.* **13**, 3404 (2022).

14. Diogo, V. S., Ferreira, H. A., Prata, D., & for the Alzheimer's Disease Neuroimaging Initiative. Early diagnosis of Alzheimer's disease using machine learning: a multi-diagnostic, generalizable approach. *Alzheimers Res. Ther.* **14**, 107 (2022).

15. Ding, Y. *et al.* A Deep Learning Model to Predict a Diagnosis of Alzheimer Disease by Using ^{18}F-FDG PET of the Brain. *Radiology* **290**, 456–464 (2019).

16. Popuri, K., Ma, D., Wang, L. & Beg, M. F. Using machine learning to quantify structural MRI neurodegeneration patterns of Alzheimer's disease into dementia score: Independent validation on 8,834 images from ADNI, AIBL, OASIS, and MIRIAD databases. *Hum. Brain Mapp.* **41**, 4127–4147 (2020).

17. Chang, C.-H., Lin, C.-H. & Lane, H.-Y. Machine Learning and Novel Biomarkers for the Diagnosis of Alzheimer's Disease. *Int. J. Mol. Sci.* **22**, 2761 (2021).

18. Stamate, D. *et al.* A metabolite-based machine learning approach to diagnose Alzheimer-type dementia in blood: Results from the European Medical Information Framework for Alzheimer disease biomarker discovery cohort. *Alzheimers Dement. Transl. Res. Clin. Interv.* **5**, 933–938 (2019).

19. Dubal, D. B. Chapter 16 - Sex difference in Alzheimer's disease: An updated, balanced and emerging perspective on differing vulnerabilities. in *Handbook of Clinical Neurology* (eds. Lanzenberger, R., Kranz, G. S. & Savic, I.) vol. 175 261–273 (Elsevier, 2020).

20. Hampel, H. *et al.* Precision medicine and drug development in Alzheimer's disease: the importance of sexual dimorphism and patient stratification. *Front. Neuroendocrinol.* **50**, 31–51 (2018).

21. Nelson, C. A., Bove, R., Butte, A. J. & Baranzini, S. E. Embedding electronic health records onto a knowledge network recognizes prodromal features of multiple sclerosis and predicts diagnosis. *J. Am. Med. Inform. Assoc.* **29**, 424–434 (2022).

22. Belonwu, S. A. *et al.* Sex-Stratified Single-Cell RNA-Seq Analysis Identifies Sex-Specific and Cell Type-Specific Transcriptional Responses in Alzheimer's Disease Across Two Brain Regions. *Mol. Neurobiol.* (2021) doi:10.1007/s12035-021-02591-8.

23. Carlos A. Saura, Angel Deprada, Maria Dolores Capilla-López, & Arnaldo Parra-Damas. Revealing cell vulnerability in Alzheimer's disease by single-cell transcriptomics. *Semin. Cell Dev. Biol.* (2022) doi:10.1016/j.semcdb.2022.05.007.

24. Leonenko, G. *et al.* Polygenic risk and hazard scores for Alzheimer's disease prediction. *Ann. Clin. Transl. Neurol.* **6**, 456–465 (2019).

25. Alzheimer's Disease Neuroimaging Initiative *et al.* Multimodal Phenotyping of Alzheimer's Disease with Longitudinal Magnetic Resonance Imaging and Cognitive Function Data. *Sci. Rep.* **10**, 5527 (2020).

26. Himmelstein, D. S. *et al.* Systematic integration of biomedical knowledge prioritizes drugs for repurposing. *eLife* **6**, e26726 (2017).

27. Nelson, C. A., Butte, A. J. & Baranzini, S. E. Integrating biomedical research and electronic health records to create knowledge-based biologically meaningful machine-readable embeddings. *Nat. Commun.* **10**, 3045 (2019).

28. Morris, J. H. *et al.* The scalable precision medicine open knowledge engine (SPOKE): a massive knowledge graph of biomedical information. *Bioinformatics* **39**, btad080 (2023).

29. Bastarache, L. Using Phecodes for Research with the Electronic Health Record: From PheWAS to PheRS. *Annu. Rev. Biomed. Data Sci.* **4**, 1–19 (2021).

30. Schwartzentruber, J. *et al.* Genome-wide meta-analysis, fine-mapping and integrative prioritization implicate new Alzheimer's disease risk genes. *Nat. Genet.* **53**, 392–402 (2021).

31. Global Lipids Genetics Consortium. Discovery and refinement of loci associated with lipid levels. *Nat. Genet.* **45**, 1274–1283 (2013).

32. Morris, J. A. *et al.* An atlas of genetic influences on osteoporosis in humans and mice. *Nat. Genet.* **51**, 258–266 (2019).

33. Jansen, W. J. *et al.* Association of Cerebral Amyloid-beta Aggregation With Cognitive Functioning in Persons Without Dementia. *JAMA Psychiatry* **75**, 84 (2018).

34. Yagis, E. *et al.* Effect of data leakage in brain MRI classification using 2D convolutional neural networks. *Sci. Rep.* **11**, 22544 (2021).

35. You, J. *et al.* Development of a novel dementia risk prediction model in the general population: A large, longitudinal, population-based machine-learning study. *eClinicalMedicine* **53**, 101665 (2022).

36. Littlejohns, T. J. *et al.* Vitamin D and the risk of dementia and Alzheimer disease. *Neurology* **83**, 920–928 (2014).

37. Elbejjani, M. *et al.* Depression, depressive symptoms, and rate of hippocampal atrophy in a longitudinal cohort of older men and women. *Psychol. Med.* **45**, 1931–1944 (2015).

38. Goveas, J. S., Espeland, M. A., Woods, N. F., Wassertheil-Smoller, S. & Kotchen, J. M. Depressive Symptoms and Incidence of Mild Cognitive Impairment and Probable Dementia in Elderly Women: The Women's Health Initiative Memory Study: DEPRESSION AND INCIDENT MCI AND DEMENTIA. *J. Am. Geriatr. Soc.* **59**, 57–66 (2011).

39. Swerdlow, R. H. Is aging part of Alzheimer's disease, or is Alzheimer's disease part of aging? *Neurobiol. Aging* **28**, 1465–1480 (2007).

40. Kosyreva, A. M., Sentyabreva, A. V., Tsvetkov, I. S. & Makarova, O. V. Alzheimer's Disease and Inflammaging. *Brain Sci.* **12**, 1237 (2022).

41. Wallace, L. M. K. *et al.* Investigation of frailty as a moderator of the relationship between neuropathology and dementia in Alzheimer's disease: a cross-sectional analysis of data from the Rush Memory and Aging Project. *Lancet Neurol.* **18**, 177–184 (2019).

42. Kojima, G., Taniguchi, Y., Iliffe, S. & Walters, K. Frailty as a Predictor of Alzheimer Disease, Vascular Dementia, and All Dementia Among Community-Dwelling Older People: A Systematic Review and Meta-Analysis. *J. Am. Med. Dir. Assoc.* **17**, 881–888 (2016).

43. Wallace, L., Theou, O., Rockwood, K. & Andrew, M. K. Relationship between frailty and Alzheimer's disease biomarkers: A scoping review. *Alzheimers Dement. Diagn. Assess. Dis. Monit.* **10**, 394–401 (2018).

44. Barnes, L. L. *et al.* Sex Differences in the Clinical Manifestations of Alzheimer Disease Pathology. *Arch. Gen. Psychiatry* **62**, 685 (2005).

45. Davis, E. J. *et al.* Sex-Specific Association of the X Chromosome With Cognitive Change and Tau Pathology in Aging and Alzheimer Disease. *JAMA Neurol.* (2021) doi:10.1001/jamaneurol.2021.2806.

46. Campion, D. *et al.* Early-onset autosomal dominant Alzheimer disease: prevalence, genetic heterogeneity, and mutation spectrum. *Am. J. Hum. Genet.* **65**, 664–670 (1999).

47. Liew, T. M. Subjective cognitive decline, APOE e4 allele, and the risk of neurocognitive disorders: Age- and sex-stratified cohort study. *Aust. N. Z. J. Psychiatry* (2022) doi:10.1177/00048674221079217.

48. He, Z. *et al.* Genome-wide analysis of common and rare variants via multiple knockoffs at biobank scale, with an application to Alzheimer disease genetics. *Am. J. Hum. Genet.* (2021) doi:10.1016/j.ajhg.2021.10.009.

49. Nandar, W. & Connor, J. R. HFE Gene Variants Affect Iron in the Brain1–3. *J. Nutr.* **141**, S729–S739 (2011).

50. Wang, Z. *et al.* Deep post-GWAS analysis identifies potential risk genes and risk variants for Alzheimer's disease, providing new insights into its disease mechanisms. *Sci. Rep.* **11**, 20511 (2021).

51. Iivonen, S. *et al.* Heparan sulfate proteoglycan 2 polymorphism in Alzheimer's disease and correlation with neuropathology. *Neurosci. Lett.* **352**, 146–150 (2003).

52. Talwar, P. *et al.* Genomic convergence and network analysis approach to identify candidate genes in Alzheimer's disease. *BMC Genomics* **15**, 199 (2014).

53. Talwar, P. *et al.* Validating a Genomic Convergence and Network Analysis Approach Using Association Analysis of Identified Candidate Genes in Alzheimer's Disease. *Front. Genet.* **12**, 722221 (2021).

54. Zhu, M. *et al.* Mutations in the γ-Actin Gene (ACTG1) Are Associated with Dominant Progressive Deafness (DFNA20/26). *Am. J. Hum. Genet.* **73**, 1082–1091 (2003).

55. Vasilopoulos, Y., Gkretsi, V., Armaka, M., Aidinis, V. & Kollias, G. Actin cytoskeleton dynamics linked to synovial fibroblast activation as a novel pathogenic principle in TNF-driven arthritis. *Ann. Rheum. Dis.* **66**, iii23–iii28 (2007).

56. Lee, W.-C., Guntur, A. R., Long, F. & Rosen, C. J. Energy Metabolism of the Osteoblast: Implications for Osteoporosis. *Endocr. Rev.* **38**, 255–266 (2017).

57. Wang, F., Han, L. & Hu, D. Fasting insulin, insulin resistance and risk of hypertension in the general population: A meta-analysis. *Clin. Chim. Acta Int. J. Clin. Chem.* **464**, 57–63 (2017).

58. James, D. E., Stöckli, J. & Birnbaum, M. J. The aetiology and molecular landscape of insulin resistance. *Nat. Rev. Mol. Cell Biol.* **22**, 751–771 (2021).

59. Schrijvers, E. M. C. *et al.* Insulin metabolism and the risk of Alzheimer disease: The Rotterdam Study. *Neurology* **75**, 1982–1987 (2010).

60. Ferreira, L. S. S., Fernandes, C. S., Vieira, M. N. N. & De Felice, F. G. Insulin Resistance in Alzheimer's Disease. *Front. Neurosci.* **12**, 830 (2018).

61. Rahman, S. O. *et al.* Association between insulin and Nrf2 signalling pathway in Alzheimer's disease: A molecular landscape. *Life Sci.* **328**, 121899 (2023).

62. Ataie-Ashtiani, S. & Forbes, B. A Review of the Biosynthesis and Structural Implications of Insulin Gene Mutations Linked to Human Disease. *Cells* **12**, 1008 (2023).

63. Gillespie, M. *et al.* The reactome pathway knowledgebase 2022. *Nucleic Acids Res.* **50**, D687–D692 (2022).

64. Bowman, G. L., Kaye, J. A. & Quinn, J. F. Dyslipidemia and Blood-Brain Barrier Integrity in Alzheimer's Disease. *Curr. Gerontol. Geriatr. Res.* **2012**, 1–5 (2012).

65. Reitz, C. Dyslipidemia and the risk of Alzheimer's disease. *Curr. Atheroscler. Rep.* **15**, 307 (2013).

66. Goldstein, F. C. *et al.* Effects of hypertension and hypercholesterolemia on cognitive functioning in patients with alzheimer disease. *Alzheimer Dis. Assoc. Disord.* **22**, 336–342 (2008).

67. Sáiz-Vazquez, O., Puente-Martínez, A., Ubillos-Landa, S., Pacheco-Bonrostro, J. & Santabárbara, J. Cholesterol and Alzheimer's Disease Risk: A Meta-Meta-Analysis. *Brain Sci.* **10**, 386 (2020).

68. Bertram, L. & Tanzi, R. E. Genome-wide association studies in Alzheimer's disease. *Hum. Mol. Genet.* **18**, R137–R145 (2009).

69. Corder, E. H. *et al.* Gene Dose of Apolipoprotein E Type 4 Allele and the Risk of Alzheimer's Disease in Late Onset Families. *Science* **261**, 921–923 (1993).

70. Garcia, A. R. *et al.* APOE4 is associated with elevated blood lipids and lower levels of innate immune biomarkers in a tropical Amerindian subsistence population. *eLife* **10**, e68231 (2021).

71. Mahley, R. W. & Rall, S. C. Apolipoprotein E: Far More Than a Lipid Transport Protein. *Annu. Rev. Genomics Hum. Genet.* **1**, 507–537 (2000).

72. Kimura, R. *et al.* Albumin gene encoding free fatty acid and β-amyloid transporter is genetically associated with Alzheimer disease: Albumin gene and Alzheimer's disease. *Psychiatry Clin. Neurosci.* **60**, S34–S39 (2006).

73. Lv, X.-L. *et al.* Association between Osteoporosis, Bone Mineral Density Levels and Alzheimer's Disease: A Systematic Review and Meta-analysis. *Int. J. Gerontol.* **12**, 76–83 (2018).

74. Amouzougan, A. *et al.* High prevalence of dementia in women with osteoporosis. *Joint Bone Spine* **84**, 611–614 (2017).

75. Liu, Y., Jin, G., Wang, X., Dong, Y. & Ding, F. Identification of New Genes and Loci Associated With Bone Mineral Density Based on Mendelian Randomization. *Front. Genet.* **12**, 728563 (2021).

76. Fan, C. C. *et al.* Sex-dependent autosomal effects on clinical progression of Alzheimer's disease. *Brain* **143**, 2272–2280 (2020).

77. Deming, Y. *et al.* The MS4A gene cluster is a key modulator of soluble TREM2 and Alzheimer's disease risk. *Sci. Transl. Med.* **11**, eaau2291 (2019).

78. Chen, Y.-H. & Lo, R. Y. Alzheimer's disease and osteoporosis. *Ci Ji Yi Xue Za Zhi Tzu-Chi Med. J.* **29**, 138–142 (2017).

79. Li, S., Liu, B., Zhang, L. & Rong, L. Amyloid beta peptide is elevated in osteoporotic bone tissues and enhances osteoclast function. *Bone* **61**, 164–175 (2014).

80. Gale, S. A. *et al.* Preclinical Alzheimer Disease and the Electronic Health Record: Balancing Confidentiality and Care. *Neurology* **99**, 987–994 (2022).

81. Serrano-Pozo, A. *et al.* Mild to moderate Alzheimer dementia with insufficient neuropathological changes. *Ann. Neurol.* **75**, 597–601 (2014).

82. Nelson, P. T. *et al.* Alzheimer's disease is not 'brain aging': neuropathological, genetic, and epidemiological human studies. *Acta Neuropathol. (Berl.)* **121**, 571–587 (2011).

83. Jack, C. R. *et al.* NIA-AA Research Framework: Toward a biological definition of Alzheimer's disease. *Alzheimers Dement. J. Alzheimers Assoc.* **14**, 535–562 (2018).

84. Data Equity Taskforce sponsored by the Health Equity Council at UCSF Health. *UCSF Health's equity-related variables user's guide.* (2021).

85. Austin, P. C. An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies. *Multivar. Behav. Res.* **46**, 399–424 (2011).

86. Karlin, L. *et al.* Use of the Propensity Score Matching Method to Reduce Recruitment Bias in Observational Studies: Application to the Estimation of Survival Benefit of Non-Myeloablative Allogeneic Transplantation In Patients with Multiple Myeloma Relapsing after a First Autologous Transplantation. *Blood* **112**, 1133–1133 (2008).

87. Tipton, E. *et al.* Sample Selection in Randomized Experiments: A New Method Using Propensity Score Stratified Sampling. *J. Res. Educ. Eff.* **7**, 114–135 (2014).

88. Bingenheimer, J. B., Brennan, R. T. & Earls, F. J. Firearm violence exposure and serious violent behavior. *Science* **308**, 1323–1326 (2005).

89. Xia, Y. *et al.* Association between dietary patterns and metabolic syndrome in Chinese adults: a propensity score-matched case-control study. *Sci. Rep.* **6**, 34748 (2016).

90. Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* (2012) doi:10.48550/ARXIV.1201.0490.

91. scikit-learn developers. Scikit-Learn Documentation: Random Forest Parameters. https://scikit-learn.org/stable/modules/ensemble.html#random-forest-parameters.

92. Breiman, L. Random Forests. *Mach. Learn.* **45**, 5–32 (2001).

93. Azodi, C. B., Tang, J. & Shiu, S.-H. Opening the Black Box: Interpretable Machine Learning for Geneticists. *Trends Genet.* **36**, 442–455 (2020).

94. Wu, P. *et al.* Mapping ICD-10 and ICD-10-CM Codes to Phecodes: Workflow Development and Initial Evaluation. *JMIR Med. Inform.* **7**, e14325 (2019).

95. Assenov, Y., Ramírez, F., Schelhorn, S.-E., Lengauer, T. & Albrecht, M. Computing topological parameters of biological networks. *Bioinformatics* **24**, 282–284 (2008).

96. Ghoussaini, M. *et al.* Open Targets Genetics: systematic identification of trait-associated genes using large-scale genetics and functional genomics. *Nucleic Acids Res.* **49**, D1311–D1320 (2021).

97. Mountjoy, E. *et al.* An open approach to systematically prioritize causal variants and genes at all published human GWAS trait-associated loci. *Nat. Genet.* **53**, 1527–1533 (2021).

98. Sun, B. B. *et al.* Genomic atlas of the human plasma proteome. *Nature* **558**, 73–79 (2018).

99. Suhre, K. *et al.* Connecting genetic risk to disease end points through the human blood plasma proteome. *Nat. Commun.* **8**, 14357 (2017).

100. Neale Lab. UK Biobank GWAS Round 2. http://www.nealelab.is/uk-biobank/.

101. Giambartolomei, C. *et al.* Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. *PLoS Genet.* **10**, e1004383 (2014).

102. Võsa, U. *et al.* Large-scale cis- and trans-eQTL analyses identify thousands of genetic loci and polygenic scores that regulate blood gene expression. *Nat. Genet.* **53**, 1300–1310 (2021).

Chapter 3: Leveraging Electronic Medical Records and Knowledge Networks to Predict Disease Onset and Gain Biological Insight Into Alzheimer's Disease

3.1 Abstract

Early identification of Alzheimer's Disease (AD) risk can aid in interventions before disease progression. We demonstrate that electronic health records (EHRs) combined with heterogeneous knowledge networks (e.g., SPOKE) allow for (1) prediction of AD onset and (2) generation of biological hypotheses linking phenotypes with AD. We trained random forest models that predict AD onset with mean AUROC of 0.72 (-7 years) to .81 (-1 day). Top identified conditions from matched cohort trained models include phenotypes with importance across time, early in time, or closer to AD onset. SPOKE networks highlight shared genes between top predictors and AD (e.g., APOE, IL6, TNF, and INS). Survival analysis of top predictors (hyperlipidemia and osteoporosis) in external EHRs validates an increased risk of AD. Genetic colocalization confirms hyperlipidemia and AD association at the APOE locus, and AD with osteoporosis colocalize at a locus close to MS4A6A with a stronger female association.

3.2 Introduction

Neurodegenerative disorders are devastating, heterogeneous, and challenging to diagnose, and their burden in an aging population is expected to continue to grow[1]. Among these, Alzheimer's Disease (AD) is the most common form of dementia after age 65, and its hallmark memory loss and other cognitive symptoms are costly and onerous to both patients and caregivers. Approaches to curb this impact are moving increasingly to targeting interventions in at-risk individuals prior to the onset of irreversible decline[2–4]. To this end, advancements in AD biomarkers, diagnostic tests, and neuroimaging have improved the detection and classification of AD, and disease-modifying treatments have been approved, but there is still no cure and much remains unknown about its pathogenesis[5,6]. This is in part due to limited availability of longitudinal data or data linking molecular and clinical domains.

In the past few decades, electronic health records (EHRs) have become a source of rich longitudinal data that can be leveraged to understand and predict complex diseases, particularly AD. Prior applications of EHRs for studying AD include deep phenotyping of AD[7], identification of AD-related associations and hypotheses[8], and models classifying or predicting a dementia diagnosis from clinical data modalities[9]. Data available in clinical records can also better represent a clinician's knowledge of a patient's clinical history at a point in time prior to further diagnostic studies or imaging, allowing a prediction model to be low cost to implement as a first line application in primary care or for initial risk stratification[10]. While machine learning (ML) has been previously applied to EHRs for general dementia classification and prediction[11–13], these approaches are limited in their specificity for the AD phenotype, lack of biological interpretability, or rely on data modalities that may not be readily available in the EHR to facilitate early prediction (e.g. neuroimaging[14–16] or special biomarkers[17,18]). Sex as a biological variable is an important

covariate for AD heterogeneity with potential contributions to differing risks and resilience, but sex-specific contributions have often been omitted from prior AD machine learning models[19,20]. To our knowledge, there have not yet been approaches that utilize vast EHR data for predicting future risk of AD with consideration of applicability and explainability of models.

With recent advances in informatics and curation of multi-omics knowledge, there is increasing interest in integrative approaches to derive insights into disease. Heterogeneous biological knowledge networks bring in the ability to synthesize decades of research and combine human understanding of multi-level biological relationships across genes, pathways, drugs, and phenotypes, with vast potential for deriving biological meaning from clinical data[21]. There has been much AD-research leveraging specific data modalities or combining a few modalities (transcriptomics[22,23], genetics[24], neuroimaging[25]), but there is still a need for meaningful integration that allows for the understanding of the relationship between pathogenesis and clinical manifestations. Heterogeneous knowledge networks provide an opportunity to derive biological hypotheses from clinical data by synthesizing knowledge across multiple data modalities to explain potential relationships between many shared clinical associations[26,27].

Here, we utilize EHR data from the University of California, San Francisco (UCSF) medical center to develop ML models for AD onset prediction and generate hypotheses of high-level biological relationships between top predictors and AD. We carry out clinical model construction for prediction and proceed with interpretation of matched patient models, controlling for demographics and visit-related confounding, to identify biologically relevant clinical predictors. We further demonstrate interpretability using heterogeneous knowledge networks (SPOKE knowledge graph)[28] and validate predictors with supporting evidence in external EHR datasets and through genetic colocalization analysis. Our work not only has implications for

determining clinical risk of AD based on EHRs, but also can lead to further research in identifying hypothesized early phenotypes and pathways to help further the field of neurodegeneration.

3.3 Results

From the UCSF EHR database of over 5 million patients from 1980-2021, 2,996 AD patients who had undergone dementia evaluation at the Memory and Aging Center and thus had expert-level clinical diagnoses were identified and mapped to the UCSF Observational Medical Outcomes Partnership (OMOP) EHR database. From the remaining patients, 823,671 control patients were extracted with over a year of visits and no dementia diagnosis. After identifying an index time representing AD onset (mean onset age (SD) 74 (5.6), see Methods) and filtering for availability of at least 7 years of longitudinal data, 749 AD patients and 250,545 control patients were identified (demographics shown in **Table 3.1**). From that, 30% was held-out for model evaluation and 70% utilized for model training (**Figure 3.1B, Supplementary Figure 3.1**). For each time point and within sex strata, ML models were either trained for AD onset prediction or trained on the AD cohort and a subset of propensity-score matched controls for hypothesis generation, where balancing was performed on demographics (sex, race & ethnicity, birth year, age) and visit-related factors (years in EHR, first EHR visit age, number of visits, number of EHR concepts, and days since first EHR record, **Supplementary Table 3.4**, matched example in **Table 3.1**).

3.3.1 ML models based on clinical data can accurately predict Alzheimer's Disease onset up to 7 years in advance

Random forest (RF) models trained on only clinical features from time points between -7 years to -1 day to AD onset were evaluated on the held-out dataset with average bootstrapped Area Under the Receiver Operating Characteristic (AUROC) curve between 0.72 (median 0.75) for the -7 year time model to 0.81 (median 0.85) for the -1 day model. The RF models performed with

Area Under the Precision Recall Curve (AUPRC) greater than the reference held-out evaluation set AD prevalence of 0.003 (average/median of 0.05/0.01 for -7 year model and 0.10/0.06 for -1 day model, **Figure 3.1C**). With addition of demographics and visit-related features, RF model performance improved with average bootstrapped AUROC between 0.86 (median 0.89) to 0.90 (median 0.94) and AUPRC between mean 0.06 (median 0.04) and 0.27 (median 0.14) for the -7 year to -1 day model, respectively (**Figure 3.1C**).

Top decision features across each time point model (see **Methods**) included features across clinical data domains, including vaccines, abnormal feces content, hypertension, hyperlipidemia (HLD), and cataracts (**Supplementary Figure 3.2A**, **Supplementary Data 3.1**). Demographic and visit-related features became predictive for AD diagnosis when added to the model, which is not unexpected since these features may contribute to confounding that influence the identified features and predicted risk of AD diagnosis (**Supplementary Figure 3.2A**). EHR diagnoses mapped to phecode categories29 (see **Methods**) identified sense organs, circulatory, and musculoskeletal phecode categories for early models, and mental disorder category for late models (**Supplementary Figure 3.2B**). Among the clusters of top 50 ranked phecodes, one cluster identified phecode features that maintain high relative importance throughout the time models (HLD, hypertension, dizziness, abnormal stool contents), and other clusters contain features with relative importance at specific time points (**Supplementary Figure 3.2C**). While some of these features support prior identified AD risk factors, the lack of adjustment may lead to feature identification as proxies for age in risk determination but not directly relevant to disease pathogenesis. Therefore, we proceed to identify disease relevant features by training models on patients matched on demographics and hospital utilization for the goal of hypothesis generation.

3.3.2 Models trained on matched cohorts can identify hypotheses for biologically relevant AD predictors

To train models that are robust for AD prediction for identifying predictors without demographic and visit-related confounding, we train time point models on a matched set of participants at a 1:8 ratio between AD and controls. Sufficient balance was achieved on numerical covariates that were highly important in unmatched demographic models (**Supplementary Figure 3.3**, **Supplementary Table 3.3**).

RF models trained on only clinical features from -7 years to -1 day performed with average bootstrapped held-out evaluation set AUROC between .58 (median 0.57) for the -7 year time model to .77 (median 0.77) for the -1 day time model. The models performed with AUPRC greater than the held-out evaluation set AD prevalence of 0.003 with improvement closer to time 0 (mean/median of 0.02/0.008 for -7 year time model and 0.08/0.03 for -1 day model, **Figure 3.2A).** When demographics and visit-related information were added as features, the models performed with minimal improvement, with average bootstrapped test set AUROC between 0.61 (median 0.61) to 0.71 (median 0.72) and similar AUPRC (mean/median of 0.02/0.009 for -7 year time model and 0.05/0.03 for -1 day model, **Figure 3.2A**). For both the full and matched cohort models, the relative performances are consistent for balanced accuracy measures on the held-out evaluation, and an example permutation test demonstrates significance for the -1 day matched cohort model (**Supplementary Figure 3.7**).

Among top features sorted by average importance across time models, top features include amnesia and cognitive concerns, HLD, dizziness, cataract, congestive heart failure, osteoarthritis, and others (**Figure 3.2B**). These top features are consistently important even when demographics and visit information was added to the model, although demographic and visit features still had

minimal influence on prediction (**Figure 3.2B**). Compared to models trained on all patients, the models trained on matched cohorts have increased importance assigned to features like hyperlipidemia and amnesia, while decreasing importance of features like pain intensity rating scale and essential hypertension (**Supplementary Figure 3.6**).

Since matching allows for the control of the influence of visit and demographic-related on AD prediction, the remaining diagnoses features can be identified for hypothesis generation with greater specificity for AD predictive risk. Top phecode categories include mental disorders, sense organs, and endocrine/metabolic categories (**Figure 3.2C**). Among clusters of specific phecodes, one cluster included features with maintained predictive importance throughout time models (HLD and congestive heart failure), while other clusters include phecodes that are relatively predictive several years prior to AD onset (osteoarthritis, allergic rhinitis). A cluster of features emerges as important around -3 years (osteoporosis, dizziness, back pain, hemorrhoids, palpitations), and some features only emerge as important closer to the time of AD onset (memory loss, vitamin D deficiency, **Figure 3.2C**). Together, this shows that the model can identify a combination of conditions that can lead to AD risk identification for a patient of a given age and hospital utilization burden.

3.3.3 Stratification by sex allows identification of features that are predictive within a subgroup

Since sex plays a role in AD risk, models were trained within male or female-identified sex groups to perform sex-specific prediction and identify sex-specific predictive features, without and with matching on demographics and hospital utilization (demographics in **Supplementary Table 3.4**). Models trained on clinical features performed with average held-out evaluation set AUROC between 0.75 (median 0.76) and 0.71 (median 0.71) for -7 year female and male models to 0.84

(median 0.86) and 0.82 (0.89) for -1 day female and male models. For AUPRC, the models performed greater than the held-out evaluation set prevalence (0.0036 for females, 0.0023 for males) with performance of 0.056-0.11(median 0.022-0.061) and 0.041-0.15(median 0.015-0.056) for females and male -7 years to -1 day time models, respectively. With addition of demographics and visit-related features, AUROC/AUPRC improved considerably (**Supplementary Figure 3.4A**). Top features include sense organs and musculoskeletal phecode categories in female-only models, and circulatory system and digestive phecode categories as important among male-only models (**Supplementary Figure 3.4B**).

To identify sex-specific biologically relevant clinical predictors for hypothesis generation, models were also trained by matching on demographic and visit-related factors within each subgroup (matching results in **Supplementary Table 3.4**). Time point models trained only on clinical features performed with mean held-out evaluation set AUROC between 0.60-0.68 (median 0.58-0.74) and 0.41-0.75 (median 0.43-0.84) for female and male models respectively (**Figure 3.2D**). For AUPRC, models performed greater than held-out evaluation set prevalence with performance ranging from 0.031-0.095 (median 0.0076-0.046) and 0.0040-0.125 (0.0033-0.022) for female and male models, respectively. Slight improvement in performance was observed with the addition of demographics and visit-related information (**Figure 3.2D**).

Top phecode categories in the female models include respiratory/circulatory system features earlier on, to musculoskeletal features in the -5 year model, to sense organs and mental disorders in the -1 year and -1 day model. Top categories in male models include endocrine/ metabolic/circulatory disorders earlier, to digestive and genitourinary in -5 and -3 models, to mental disorders in -1 day model (**Supplementary Figure 3.4B**). When comparing specific phecodes, some are general across the subgroups such as HLD, congestive heart failure (early

models), and memory/cognitive symptoms (later models) (**Figure 3.2E**, **Supplementary Figure 3.4C**). Female-driven features across time models included osteoporosis, palpitations, allergic rhinitis, myocardial infarction, major depressive disorder, and abnormal stool contents. Male-driven features included chest pain, hypovolemia, sexual disorder, tobacco use disorder, and neoplasms (**Figure 3.2E**).

For all formulations of the prediction task, logistic regressions (LR) models performed comparably to random forests, and identify features with linear relationships with AD including some overlap with features identified from random forest models (**Supplementary Figure 3.5**). Nevertheless, for matched cohort models random forest performs better than logistic regression at the same time points (**Supplementary Table 3.5**) and can identify decision features with nonlinear relationships with AD (e.g., RF identifies osteoporosis). Balanced accuracy measures for all of the random forest models support trends in performance between models, including lower overall performance for matched cohort models, and improvement in model performance closer to onset of AD (**Supplementary Figure 3.7A**, **Supplementary Table 3.6**). As an example to evaluate the extent that clinical features meaningfully predict AD, random forest models were retrained on permutations of the ground truth label for the -1 day matched cohort (40 permutations) and the trained model distribution was significant compared to the null distribution ($p=0.024$, **Supplementary Figure 3.7B**).

3.3.4 Use of a knowledge graph allowed prioritization of known biological explanations underlying predictive features

Next, we utilized the SPOKE knowledge graph[28] in order to utilize existing knowledge to explain and prioritize biological relationships between groups of top clinical model features and

AD. We mapped biological features (genes, proteins, compounds, etc.) between top 25 clinical predictors (mapped to disease nodes) and AD node for each model (see **Methods**).

Genes that appear in shortest path networks among matched models across multiple time include APOE, AKT1, INS, ALB, IL1B, INF, ALB, IL6, SOD1, etc. and compounds include atorvastatin, simvastatin, ergocalciferol, progesterone, estrogen, cyanocobalamin, and folic acid (**Figure 3.3**). These genes and compounds also share relationships to multiple occurring model input nodes, particularly familial hyperlipidemia and osteoporosis among all time point models (**Figure 3.3**). Notable nodes that appear over at least 2 models include C9orf72, TREM2, APP, MAPT with relationships to input nodes of musculoskeletal and joint disorders, deafness, and depression (**Figure 3.3**).

3.3.5 Hyperlipidemia validates as a top predictor of AD in external EHRs and a genetic link confirmed in APOE locus

In order to further validate the utility of models to identify predictive disease associations, we followed up on HLD as a top feature that was a consistent predictor across all models. Utilizing a retrospective cohort study design in an EHR on five hospitals across the University of California system (University of California Data Discovery Platform (UCDDP)) with exclusion of UCSF, HLD-diagnosed patients (exposed group, n = 364,289) had a faster progression to AD event compared to matched unexposed patients (n = 364,289, matched demographics in **Supplementary Table 3.7**) (**Figure 3.4A**, **Supplementary Figure 3.8A**, log-rank test p-value<0.005). This was further confirmed with a Cox proportional hazards analysis (hazard ratio (HR) 1.52 (95% Confidence Interval (CI) 1.46-1.57), visit/demographic adjusted HR (aHR) 1.26 (1.21-1.31), p-value <0.005, **Supplementary Figure 3.8C**).

In order to investigate potential relationships between HLD and AD, the HLD-specific knowledge network demonstrated shared gene associations with LSS, APOE, INS, SMAD3, ALB, and GFPT1 (**Figure 3.4B**). Locus intersections between high LDL cholesterol and AD across two independent GWAS studies across 408,942 AD patients from Schwartzentruber et al.[30] and 94,595 LDL Cholesterol patients from Willer et al.[31] respectively identified multiple shared variants, including ch19:44,892,362(hg38):A>G (rs2075650) and ch19:44,905,579(hg38):T>G (rs405509) (https://genetics.opentargets.org/study-comparison/GCST002222?studyIds=GCST90012878). PheWAS for rs2075650 on the UK Biobank verified significant associations with cholesterol levels, HLD, AD, and family history of AD (**Figure 3.4C**). Colocalization H4 probability, a measure that determines the probability two traits are associated at a locus based on prior genetic studies, supports a causal link with locus variants for APOE protein QTL and both HLD traits and AD traits (**Figure 3.4D**).

3.3.6 Female-specific predictor of osteoporosis validates in an external EHR with potential explanations given in SPOKE and genetic colocalization analysis

Osteoporosis was identified as an important feature in the matched models as a female-specific clinical predictor of AD. In the UCDDP, osteoporosis-exposed patients (n=68,940) showed a quicker progression to AD compared to matched unexposed patients (n=68,940, matched demographics in **Supplementary Table 3.8**) (**Figure 3.5A**, **Supplementary Figure 3.8B**, log-rank test p-value<0.005). When stratified by sex, this progression is significant when comparing between female osteoporosis (n=57,486) vs female controls (n=58,636). Cox hazard analysis further supported osteoporosis as a general risk feature for AD (HR 1.81 (95% CI 1.70-1.92), aHR 1.59 (1.45-1.70), p<.005 **Supplementary Figure 3.8D**).

Osteoporosis-specific SPOKE network demonstrated shared gene associations with IL6, SMAD3, TNF, HSPG2, GATA1, GFPT1, HFE, INS, and ALB (**Figure 3.5B**). Based on previous GWAS studies across 472,868 AD patients from Schwartzentruber et al.[30] and 426,824 heel bone mineral density (HBMD) patients from Morris et al.[32], a shared risk locus was found in Chromosome 11 between HBMD and AD among the MS4A gene family (https://genetics.opentargets.org/study-comparison/GCST006979?studyIds=GCST90012877), with the closest gene as MS4A6A. A comparison of prior GWAS of up to 71,880 AD patients from Jansen et al.[33] and sex-stratified heel bone mineral density (HBMD) GWAS (111,152 Female, 166,988 Male) of UK Biobank patients from Neale Labs (www.nealelab.is/uk-biobank/) supports a female-specific association at the shared locus (**Figure 3.5C**). Colocalization analysis supports a link between MS4A6A and AD (H4 = 0.987), female-specific HBMD with AD, and phenotypes with MS4A6A expression (**Figure 3.5D**, AD vs Female HBMD H4 = 0.998, MS4A6A vs Female HBMD H4 = 0.997). This statistical significance is not replicated for male specific HBMD GWAS (**Figure 3.5D**, AD vs Male HBMD H4 = 0.00263, MS4A6A vs Male HBMD H4 = 0.00266). MS4A6A weighted associations with other phenotypes from Open Targets Genetics found locus associations with many inflammatory phenotypes including c-reactive protein, lymphocyte percentage, and neutrophil count (**Figure 3.5E**).

3.4 Discussion

While there is enormous potential in ML on clinical data, balancing clinical utility and biological interpretability can be challenging. To address this, we used thousands of EHR concepts to develop prediction models for expert-identified AD diagnosis, and selected an index time suggesting AD onset. Cohort selection and data preprocessing is a crucial first step to identify available clinical measures and optimal ground truth AD onset that is as close to biological AD

and avoid overly optimistic model performance due to nonspecific groundtruth or improper data preprocessing[34]. Our prediction model shows predictive power up to -7 years before the defined index time of AD onset with AUROC of 0.72 (and up to AUROC 0.86 with additional demographic and care utilization features), which is comparable with other models in literature that utilize clinical data to predict less specific dementia or AD diagnosis[11,35]. An application of the model trained on all patients includes determining early disease risk in primary care settings before time-consuming and costly detailed neuropsychological, biomarker, or neuroimaging assessments (after which imaging or biomarker classification models can be utilized[13]). The model may also identify at-risk patients for follow-up or inclusion in early intervention or clinical trials, with the -1 day model as suggesting possible AD onset to be considered at that visit to prevent underdiagnosis of AD. Furthermore, interpretable models, such as random forest models, can identify common decision point features and allow clinicians to understand what clinical features were used in determining prediction probability and assess the model output with greater trust compared to "black box" models.

In order to identify early clinical predictors that may be biologically relevant for AD diagnosis, we trained models on patients matched by pre-identified confounding variables such as demographics and visit-related features so that these features have less influence in AD prediction. Machine learning models still retain the ability to predict AD diagnosis with mean AUROC over .70 after the -3 year time model for random forests. Inclusion of demographic and visit-related features minimally improved model performance, which is expected since matching increased the specificity of the task to predict AD onset controlled on demographics and visit-related features. In terms of clinical utility, the models trained on matched patients provide predictive power for a given clinical scenario between two patients with similar pre-test probability of AD risk (e.g., same

age and disease burden), with application of this model as a tool for determining post-test probability of future AD risk. Furthermore, by balancing on pre-identified confounders such as demographics and visits, top features may be interpreted with more biological relevance for AD risk. For example, while we identified essential hypertension as an important feature in the models trained on the full cohort, this diagnosis became less important in the models trained on matched cohorts, suggesting hypertension may be nonspecific for AD and may instead be more related to aging or disease burden.

Our time models trained on matched cohorts identify or strengthen known or suggested hypotheses for early clinical predictors of AD, such as hyperlipidemia as a feature for all time point models. We also identify relative importance of features years in advance, such as allergic rhinitis and atrial fibrillation as early predictors, osteoporosis and major depressive disorder as non-neurological predictors, and cognitive impairment and vitamin D deficiency as late predictors of AD. Some of these prior predictors, such as depression and vitamin D deficiency, have been previously implicated in AD risk[36–38]. These findings potentially support hypotheses suggesting AD can be associated with general aging or frailty, which might present in non-neurologic body systems either prior to or concurrent with AD [39–43]. Furthermore, interpretation of these models allows for the identification of high order groups of predictors that may contribute to disease heterogeneity or together, contribute to AD risk. Nevertheless, while these models can identify hypotheses of predictive features, EHR data can still capture clinical biases or misdiagnoses, and further studies can investigate the influence of behavioral bias vs biological relevance.

We further trained models on sex-stratified subgroups (female vs male), with and without matching on demographics and visit-related covariates, in order to identify sex-specific drivers of clinical predictors. Given evidence that sex may influence different pathways to AD

diagnosis[22,44,45], it is important to consider how patient heterogeneity may impact the training, utility, and interpretation of a prediction model. From the matched cohort models, we identified clinical features in each subgroup that were consistent with the general models, such as hyperlipidemia as important in every model and memory loss as important in late models. Furthermore, we identified features that were sex-specific, such as osteoporosis, major depressive disorder, allergic rhinitis, and abnormal stool contents as predictors enriched among women, and chest pain, hypovolemia, prostate hyperplasia, and sensorineural hearing loss as predictive among men. Further work can seek to disentangle the biological meaning of these sex-specific predictive features: whether they reflect sex-specific non-neurological manifestation of prodromal states, contributing risk factors, or even sex biases in clinician evaluation and treatment (e.g., bone density evaluation may arise more often after a fall). These models also demonstrate that for a heterogeneous disorder like AD, subgroup composition, like sex ratio of a cohort, can influence the performance and the features that are identified as important. Differences in subgroup size and prevalence of AD contribute to greater predictive performance among female strata models, and differences observed in AUPRC are impacted by AD prevalence which can influence interpretation of the positive predictive value of models within each sex strata. In terms of identified features, the higher preponderance of females lead to sex-specific predictive factor, osteoporosis, being identified as a general predictive variable in the general group. This further indicates that both generalizable models and subgroup-specific models can provide valuable insight, both general and personalized, for a complex disease. Furthermore, in the context of ML fairness, the performance and identified features of general models may be influenced by the demographic make-up of the training population, just like how greater number and AD prevalence

among females influence greater female-strata performance and identification of osteoporosis in our general models.

We utilized a heterogeneous knowledge networks (SPOKE) to identify shared biological hypotheses underlying model-identified top clinical predictors and AD. By combining shortest paths in SPOKE between top predictors and AD, we can prioritize nodes (e.g., genes) that are consistently relevant for the high order combination of human data derived top clinical predictors and AD, and give novel insight via prioritization and combination of relationships. First of all, we were able to identify known genetic associations with dementia based upon top diagnoses, such as through identification of known autosomal dominant early AD genes such as APP and PSEN 1/2[46]. Other genes identified with known associations with AD include APOE, HFE, and HSPG2 variants that impact AD risk[47–51]. An example of novel insight gained through SPOKE integration includes ACTB relating to AD[52,53], sensorineural hearing loss[54], arthropathy, and arthritis[55]. The prediction model allows for the prioritization of ACTB for patients with the common comorbidities of sensorineural hearing loss and arthropathy/arthritis with risk of AD (where the connection through linking sensorineural hearing loss, arthropathy, arthritis, and AD all together through ACTB has not been previously implicated in literature).

The SPOKE network can also be leveraged to propose biological explanations based on common nodes and shared associations between clinical predictors identified from human data and AD. For example, ALB is identified through SPOKE as a shared genetic association between congestive heart failure, malnutrition, hyperlipidemia, and AD. While prior relationships have been identified between ALB and many individual diseases, each of those diseases also have many implicated genetic relationships. Leveraging human data through the predictive models allows for the prioritization of abundant gene connection with multiple disease predictors. Given ALB roles

in pathways such as heme biosynthesis (Reactome R-HSA-189445), HDL remodeling (Reactome R-HSA-8964058), and insulin-growth like factor regulation (Reactome R-HSA-8964058), prioritization of mechanistic hypotheses linking ALB related pathways with the pathophysiology of EHR-derived predictors (congestive heart failure, malnutrition, hyperlipidemia) can be explored in future studies. Another example insight includes INS as a shared association between osteoporosis[56], hypertension[57], hyperlipidemia[58], and AD[59,60]. Prior studies have identified potential mechanisms underlying the relationship between energy utilization, lipid levels, nutrition, and neurodegeneration (e.g., Reactome R-HSA-1266738, R-HSA-16368)[61–63], and this analysis allows for prioritization of mechanistic hypotheses to be further explored. While these associations are included in the SPOKE network due to evidence in literature, the association of these genes with specific early clinical predictors is less established, and thus this analysis allowed us to identify novel constellation of phenotypes and underlying genetic relationships observable in a clinical setting that, together, can lead a clinician to suspect future AD risk, prioritize molecular pathways for testing or personalized treatment, and guide biological hypotheses generation in AD pathogenesis for future studies.

To validate a few top clinical predictors, we utilized a hypothesis-driven approach to support the relationship between two identified features (hyperlipidemia and osteoporosis) and progress to AD diagnosis in an external database across the University of California EHR system. For both phenotypes, the UC-wide EHR database supports a potential increased AD diagnosis risk due to evidence of decreased time to AD and increased hazard of AD diagnosis in patients exposed to the predictor of interest. The association between hyperlipidemia and AD has been identified in prior clinical studies and systematic reviews[64–67]. In particular, APOE is a well-established associated genetic locus[68], and APOE polymorphism is known to modify AD risk, particularly in

individuals carrying the ε4 allele[69]. Many studies have also shown APOE association with elevated lipid levels and cardiovascular risk factors[70,71]. The validation of these well-known associations not only show that our ML models on clinical data can pick up hyperlipidemia as a risk factor, but also by utilizing the SPOKE network we can integrate known relationships in literature to potentially explain the association between hyperlipidemia and AD and identify the APOE locus as a potential shared causal mechanism as demonstrated in the colocalization results. Beyond the ability to identify known relationships, the SPOKE network also proposes biological explanations of higher-order shared associations between clinical predictors, such as ALB as a shared genetic association between congestive heart failure, malnutrition, hyperlipidemia, and AD, or INS as a shared association between osteoporosis, hypertension, hyperlipidemia, and AD. Prior studies have identified potential mechanisms underlying the relationship between energy utilization, lipid levels, nutrition, and neurodegeneration[59,60,72], although specific hypotheses of mechanistic relationships are an area for exploration in future studies.

The association between osteoporosis and AD is also validated to a lesser extent in clinical studies and meta-analysis[73,74], with unclear but possible sex-modification of this effect. Our study identifies osteoporosis as a predictor for AD among females prior to AD, but shows less of a *relative* predictive effect for males compared to other clinical features. Nevertheless, it is still possible that shared relationships between osteoporosis and AD exist in males. A bone mineral density GWAS analysis of female patients shows p-value association with AD GWAS around the MS4A family locus, and this is further supported by MS4A6A eQTL colocalization with both Alzheimer and female HBMD. These findings of osteoporosis as a potential sex-specific predictor of AD, with shared relationships through MS4A6A, is a potential new and unexpected results identified from single hypothesis-driven follow-up from our prediction models. Prior studies have

established the MS4A gene cluster as a risk for AD, with one study identifying the cluster based on mendelian randomization[75], and another that identifies a stronger female-specific effect size for MS4A6A[76]. Some studies investigating the role of the MS4A family suggest mechanisms that involve immune function, particularly among microglia[77]. While this gene may not have been identified in SPOKE, SPOKE did capture direct pathways through known markers of inflammation such as IL6 and TNF, and we also see MS4A6A as highly associated with measurements of immune cells in the blood. Further studies will be needed to validate the exact associative mechanism between osteoporosis and AD, although some prior hypotheses suggest the potential impact of genetic variants on osteoclast function, amyloid clearance, or oxidative stress response[78,79]. While we utilized knowledge networks to leverage knowledge to explain relationships between groups of predictors, we performed hypothesis-driven analysis on independent EHRs and genetics to further explore and validate a few chosen predictors (hyperlipidemia, osteoporosis) with AD. Hypothesis-driven approaches can be applied to any other selected predictor or phenotype identified by the models to understand their relationships with AD onset that may not yet be represented by the knowledge graphs.

This study has several limitations. First, EHR data complexity and quality can affect prediction models, and it is challenging to distinguish the influence of clinician/patient behavior, sociological factors, or underlying biology on identification of features. Matching can improve interpretability by removing influence of non-biological covariates, but follow-up validation of hypotheses across omics data types is needed. Due to changing patient demographics and societal factors, prediction models should be continuously trained, updated, and evaluated if implemented in the clinical setting to ensure effective utilization and account for biases that may have been learned from the data. Model utilization should investigate the impact of cohort selection biases

and matching methods on model generalizability, and model retraining and calibration should be a continual aspect of model application to account for possible data drifts and changing clinical practice approaches that would arise in the future. Second, clinical EHR data is sometimes sparse and provides a superficial interval snapshot of a patient's health, so the absence of a record may not necessarily reflect the absence of a condition and prior health information may not be available in the EHR. Therefore, the EHR provides a representation of an interval of a patient's health history and is more likely to pick up diagnosis of chronic or common conditions, as well as common drugs or measurements. Future work can investigate the impact of variations in data representation that can account for data sparsity, continuous lab result outcomes, and best temporal assignment of diagnosis onset beyond binary representation or considering drug prescriptions for assignment of diagnoses. Third, survival models have extensive right censorship and do not take into account competing risks. Fourth, since AD is heterogeneous and differential diagnosis is nuanced and subjective even in expert hands, predictive performance can be limited by label quality and the signal from clinical features can be noisy, limiting performance and generalizability. Future work investigating heterogeneity may identify subgroup-specific features where subgroups can be divided based on biotype, dementia syndromes, racialization, and so on. Future applications with hierarchical models, transfer learning, or fine-tuning on a subpopulation can increase personalization of models. Fifth, our sex-stratified analysis was restricted to patients that identified as female or male. Future studies could explore AD patterns among intersex individuals. Lastly, predictive features identified are relevant prior to AD onset, and future work is needed to identify diagnostic-relevant AD comorbidities, or conditions that can occur after AD progression. Since predictive features are identified as hypotheses, the direct mechanism and causal pathway relating

a phenotype to AD is not known. Future work can investigate causality with mendelian randomization or mechanistic studies.

In this study, we demonstrate how formulation of prediction models can influence utility for predictive application or biological interpretation. We show how models can be utilized to identify early predictors, and utilize SPOKE to explain relationships via shared biological associations. Lastly, we show that our models can pick up known associations with HLD through APOE, and identify a lesser known association with osteoporosis through MS4A6A that may be female-specific. This study contributes to the field of EHR integrative research that can inform future directions in both AD care and research.

3.5 Methods

3.5.1 Patient Identification

Alzheimer's Disease (AD) patients were identified based on UCSF Memory and Aging Center database containing over 9000 patients mapped to the UCSF Observational Medical Outcomes Partnership (OMOP)-format EHR. These patients have undergone dementia evaluation at the Memory and Aging Center and thus had expert-level clinical diagnoses. In clinical settings, since AD is often a syndromic diagnosis indicating general dementia for memory or cognitive concerns[80–82], we aimed to identify a highly accurate cohort diagnosed by neurodegeneration specialists to obtain AD diagnosis that is closer to the biological ground truth[83]. The remaining control patients were obtained from the rest of the UCSF EHR, with over 1 year of records and no existing records of dementia diagnosis among the G[123]* ICD-10 categories (**Supplementary Table 3.1**). These controls include patients seen at the UCSF Memory and Aging Center with EHR data, but without a dementia diagnosis given.

In order to best build models for prediction of AD onset, an index time was determined to identify input model features prior to first clinical indication of dementia. This was defined among the AD cohort as the first time of any AD diagnosis, dementia diagnosis, or prescription of cognitive drug (ATC codes N06D, **Supplementary Table 3.2**) to be the first time point of possible biological AD manifestation. This approach was utilized since AD patients may be prescribed an anticholinesterase inhibitor or given an alternative dementia diagnosis before a formal confirmation of an AD diagnosis. For controls, the index time was defined as 1 year before the last recorded her visit date, with no dementia diagnosis given within that year. In order to maintain a consistent patient population for training and evaluation of machine learning models, the final AD and control cohort was identified by filtering to patients who are at least 55 years of age at the

index time and have existing clinical visits and concepts 7 years prior to the index time. These patients were then split into 70% for model training and tuning, while the remaining 30% was held-out for model evaluation (**Supplementary Figure 3.1**).

3.5.2 Data Extraction and Preparation

Demographics (birth year, gender, race & ethnicity), clinical concepts (conditions, drug exposures, abnormal measures), and visit-related features (age at prediction, first visit age, years in UCSF EHR) were extracted before the index time for the AD and control cohort from the UCSF OMOP EHR database. Race & ethnicity is a single variable derived from an algorithm developed by the UCSF Data Equity Taskforce to codify aggregated sociopolitical categorizations based on EHR reported identifiers[84]. To train models in advance of the index time, clinical information was extracted for each patient including all clinical data up to a time point X before the index time, where X includes -7 years, -5 years, -3 years, -1 years, and -1 day. These time points represent the knowledge of a patient's clinical history leading up to time X before time. All existing clinical features (conditions, drug exposures, abnormal measurements) were one-hot encoded. Abnormal measures were extracted from the OMOP measurement table based on the numeric value falling either above range_high or below range_low. and abnormal measures were binary encoded based on abnormal flagging, following the approach from Nelson et al.[27]. If a clinical feature did not exist or if the clinical measure was within normal range, the encoding is represented as a 0 and therefore assumed to be normal. Since the UCSF database only captures an interval of a patient's interaction with the healthcare system, prior non-chronic conditions may not be captured within the EHR.

Demographic and visit-related features (prediction age, first visit age, years in UCSF EHR, log(number prior visits), log(number prior concepts), log(days since first clinical event)) were

scaled between 0-1 on the training data, where log indicates natural logarithm and feature scaling allows for multiple ML model approaches. Age at prediction is defined at the age of patient at which the model is applied (e.g., if a patient index time is at age 70, then the age of prediction for the -5 year model is 65). All features with no variance were removed for each model, with total number of features ranging from 5,211 features (-7 year model on matched cohorts) to 23,760 features (-1 day models on unmatched cohorts). Information about input features and specific OMOP concepts can be found in **Supplementary Data 3.1**. Some top feature prevalences are also included in **Supplementary Data**.

3.5.3 Machine Learning Preparation and Training

Binary classification time point models for AD were trained using the patient representation at each time point before the index time. We divided the data into two sets, 70% for model creation and 30% for evaluation. Training and optimal model selection (with hyperparameter tuning) was performed on the 70% split with cross-validation, and 30% was held-out for evaluation and not seen during model training and selection in any way. Final selected model evaluation was performed on the 30% held-out evaluation set as the common dataset to obtain and compare the performance of all final models (diagram in **Supplementary Figure 3.1**). Models were trained with clinical features only (clinical model) and with clinical features + demographics and visit-related information (clinical + demo/visits model). Models were also trained on samples matched by demographics and hospital utilization to account for biases and confounding in prediction. In these models, control patients were matched to AD patients at a 1:8 ratio on demographics (birth year, race & ethnicity, sex) and visit-related features (age, first visit age, years in EHR, log(# prior visits), log(# prior concepts), log(days since first clinical event)) utilizing propensity score matching[85] (propensity score estimated based upon a logistic regression

model, nearest neighbor matching without replacement). While propensity score is often utilized to balance treatment probabilities in cohort studies, it has also been utilized for sample selection[86,87], exposure likelihood[88], or for outcome-based case-control studies[7,89].

Random forest models were primarily utilized for both predictive performance and interpretability that takes into account the high collinearity between clinical variables. Random forests were trained using *scikit-learn* package[90], with balanced class weight parameter. Hyper-parameters were tuned (grid search) based on cross-validation performance (5 folds) of AUROC on the 70% model training set to determine parameters of *n_estimators* (n_features, n_features*2, n_features*3), *max_depth* (3, 5, 7, 9), and *max_features* (sqrt, log2). The number of estimators and max depth were tuned to balance between performance and overfitting, while a subset of features (max_features) was utilized per tree to help account for high correlation between features[91,92]. Models were evaluated on bootstrapped subsamples (50-200 iterations, 1000 samples) of the 30% held-out evaluation set to determine AUROC (area under the receiver operating curve) and AUPRC (area under the precision-recall curve) for model comparability. Balanced accuracy scores were also computed on the 30% held-out evaluation set. An elastic net logistic regression model was also trained on both the full and matched cohorts for comparison. We performed a permutation test on the -1 day matched cohort model to determine the significance of AUROC compared to a null distribution of AUROC scores of models trained from permuted ground truth labels (40 permutations) to determine to the extent clinical features can be predictive of AD.

Stratification: Both models for full patient cohorts and matched cohorts were re-performed in sex strata in the same fashion based upon sex reported the UCSF EHR to augment the OMOP database. Models were trained on two sex subgroups: female and male, due to lack of other subgroups labelled in the EHR. For each strata, AD patients were re-matched to controls within

each strata for the matched patient trained models. Models were evaluated similarly based on AUROC/AUPRC on the same bootstrapped held-out evaluation set, stratified by sex.

3.5.4 Top Feature Interpretation

Random forest models were investigated for feature interpretation due to the combined interpretable nature of the models (compared to neural networks) and the ability to capture nonlinear relationships (compared to logistic regression models)[83]. Average gini impurity decrease for each feature was utilized to evaluate the importance of each feature in the random forest models (feature importance). The average importance for each feature was taken across each time point models (-7yr, -5yr, -3yr, -1yr, -1day) to obtain an across-model importance for each model type, and normalized by the maximum importance value across all time point models within each model type (e.g., random forest) and group (e.g., female strata). Feature importances are then ranked within each model to obtain relative importance within each of the time points.

Since a patient's exposure to a medication or a laboratory test is often a result of a diagnosis, we pursued interpretability based on diagnostic features that have been mapped to phecodes, which is a semi-manual hierarchical aggregation of meaningful EHR phenotypes[29]. This allows for a lossy categorization of detailed OMOP features (OMOP IDs) to phecodes (OMOP ID → SNOMED → ICD10 → phecode) and phecode category. SNOMED IDs were mapped to ICD10 based upon recommended rule-based mappings from the National Library of Medicine (NLM) September 2022 release (www.nlm.nih.gov/healthit/snomedct/us_edition.html). ICD10 codes were then mapped to phecodes based on the release from Wu et al.[84] To obtain the importance within each phecode or phecode category, the average importance for the top 5 detailed OMOP features per phecode or phecode category was computed, and ranked between phecodes or

categories. For phecodes across all models and sex-stratified models, the ranking of importance of phecodes across each time model was hierarchically clustered with Ward linkage.

To compare top phecodes between sex-stratified models to identify sex-specific features, top random forest features over an average importance threshold of 1e-6 were identified per time model trained on matched participants. Upset plots were then generated for each time point based upon this overlap. Female-driven features are defined as features that exist in both the full model and female models, or only female models, and male-driven features defined analogously.

3.5.5 UC-wide validation analysis with hypothesis-driven retrospective cohort analysis

Two top clinical features were selected from the matched all patient model (hyperlipidemia) and matched sex-specific models (osteoporosis) and further followed up on an external EHR database to validate the feature as predictive and conferring risk for AD diagnosis. With these features defined as exposures, hypothesis-driven analysis was performed with a retrospective cohort study design on the University of California hospital EHR database (University of California Data Discovery Platform (UCDDP)) with exclusion of any patients seen at UCSF, so with included institutions consisting of UC Davis, UC Los Angeles, UC Riverside, UC San Diego, and UC Irvine. Exposed patients were identified with the exposure (hyperlipidemia or osteoporosis), which were identified by string-matching and mapping to all descendants or related concepts based on the OMOP relationship tables, and final SNOMED codes are shown in **Supplementary Table 3.6 and 3.7**. Controls were identified among the remaining patients. Recruitment age was defined as the age of exposure diagnosis (for exposed cohort) or the first visit age in the visit_occurrence table (for unexposed or control cohort), which was then matched to

represent the start of the cohort study timeline. All patients are then filtered to have at least 2 years of records in the EHR, and last visit age was utilized for right censorship.

The outcome of interest was AD diagnosis, which was identified based on SNOMED codes 26929004, 416780008, 416975007 (**Supplementary Table 3.5**). Exposed and control (unexposed) groups were then matched based on demographics (gender, race & ethnicity), birth year, and recruitment age (propensity score estimated based upon a logistic regression model, nearest neighbor matching without replacement). We utilized the gender_id column to identify sex, as the standard documentation intend for this column to represent biological sex (see www.ohdsi.org/web/wiki/doku.php?id=documentation:vocabulary:gender). Note that only two options exist (female concept_id=8532 and male concept_id=8507), and that accurate sex and gender information may be limited depending on the institution or EHR collection of sex information.

Analysis of time to AD diagnosis includes utilization Kaplan Meier survival curves fitted with 95% confidence interval and two-sided log-rank test to compare survival curves between groups. Sex-stratified curves were also fitted. Cox proportional hazard models were utilized to obtain unadjusted hazard ratios (HR) and adjusted hazard ratios by demographics and/or visit information (aHR), with and without stratification by recruitment age or birth year, and with 95% confidence intervals.

3.5.6 Heterogeneous Network Analysis

Heterogeneous knowledge networks, such as SPOKE, integrate known relationships across biological and phenotypic data realms in databases and literature. Such a network could provide hypotheses to explain relationships between groups of phenotypes that may not be immediately known[21,26]. We proceed with interpretation on the matched models, with the top 25 model features

taken per time point and mapped to SPOKE nodes based on Nelson et al.[27] Note that mappings may not be 1 to 1. All shortest paths were then computed from each input node to the Alzheimer's Disease node (DOID: 10652), and shortest paths were filtered to exclude certain node types (Anatomy, SideEffect, AnatomyCellType,Nutrient) and edges (CONTRAINDICATES_CcD, CAUSES_CcSE, LOCALIZES_DlA, ISA_AiA, PARTOF_ApA, RESEMBLES_DrD). Edges were also filtered based on the following criteria: TREATS_CtD at least phase 3 clinical trial, UPREGULATES_KGuG/ DOWNREGULATES_KGdG p-value at most 1E-4, PRESENTS DpS enrichment at least 5 and fisher p-value at most 1E-4.

If multiple detailed OMOP features map to the same node, the importance of the node was obtained by the average of OMOP feature importances. Networks for all time models were combined into a single network (union of nodes and edges), and total node importance was determined by the maximum across time. Network metrics were then computed with Cytoscape 'Network Analyzer' function[85]. The combined time model networks were then sorted by eccentricity metric on the x-axis (representing maximum distance to all other nodes, with lower number representing higher importance) and number of individual time model network occurrences in the y-axis (showing node importance persistence across time). With this layout, highly traversed nodes in the shortest paths between multiple EHR informed top model features and AD can be identified and prioritized for hypothesis generation and further investigation. Note that due to heterogeneous nature of edges and lack of edge weighting, distance in the figure is not meaningful.

To focus on two selected features for the full matched model (hyperlipidemia (HLD)) and the female-specific matched model (osteoporosis), the combined network was filtered based on first and second degree neighbors of the starting feature of interest. This allows for visualization

of associated genes and AD, as well as relationships with other top model features found from the clinical models.

3.5.7 Validation with Genetic Datasets

We further explored the association between clinical predictors and AD by identifying shared genetic loci between top model phenotypes and AD, based on colocalization probability and weighted evidence association scores computed from Open Targets Genetics[96,97] (genetics.opentargets.org). Colocalization analysis is a method that determines if two independent signals at a locus share a causal variant, which helps increase the evidence that the two traits (e.g., hyperlipidemia and AD, or protein expression and AD) also share a causal mechanism. It is a Bayesian method which, for two traits, integrates evidence over all variants at a single locus to evaluate the following hypothesis that two associated traits share a causal variant. This is the H4 probability.

We first identified shared loci between the selected phenotypes (HLD or osteoporosis) and AD by identifying the genetic intersection between AD and related phenotypes in Open Targets Genetics.

For HLD and AD, we utilized the Open Targets Genetics platform to identify overlapping variants and shared locus between LDL Cholesterol and Family History of AD or AD. PheWAS between a shared SNP and UK Biobank phenotypes were plotted and extracted from the Open Targets Genetics platform. Coloc analysis tables between the gene, molecular QTLs, and phenotypes were extracted, with protein QTLs for APOE specifically identified based on blood plasma data from Sun et al.[98] and Suhre et al.[99]

Similarly for osteoporosis and AD, we utilized the Open Genetics platform to identify shared locus between heel bone mineral density (proxy for osteoporosis) and Family History of

AD or AD. To further investigate the locus, we extracted GWAS summary statistics from Jansen et al.[48] for AD and sex-stratified GWAS summary statistics for heel bone mineral density (HBMD) from Neale's Lab GWAS round 2, Phenotype Code:3148, based on data from the UK Biobank (www.nealelab.is/uk-biobank/)[100]. We then conducted colocalization analysis using the coloc method described in Giambartolomei et al.[101], from R package coloc 5.1.0. Summary statistics for MS4A6A cis eQTLs in blood were extracted from eQTLGen[102], and colocalization analysis was performed between AD, sex-stratified HBMD, and MS4A6A eQTLs on the Locus Region 60050000-60200000 of Chromosome 11. To investigate further associations with the locus, MS4A6A associations with all other phenotypes was extracted from Open Targets Genetics platform with inclusion of a weighted literature evidence association scores.

3.5.8 Ethical Approval

This study was approved by the Institutional Review Board of University of California San Francisco (IRB #20-32422).

3.6 Code and Data Availability

EHR concepts and identification approaches are described in Methods, and concepts are provided in **Supplementary Tables 3.1 and 3.2**. Phecodes can be downloaded at phewascatalog.org/phecodes_icd10 or phewascatalog.org/phecodes, and mappings between ICD-10 codes and SNOMED can be accessed at www.nlm.nih.gov/healthit/snomedct/us_edition.html. Code for EHR prediction models, model feature interpretation, matching, external EHR survival analysis, and querying Open Targets API (genetics.opentargets.org/api) for P-P and eQTL plots can be found at github.com/al1563/ADprediction_code. Data for UK Biobank phenotype GWAS can be found at www.nealelab.is/uk-biobank/, and eQTL data can be downloaded from

www.eqtlgen.org/. Access to EHR databases are controlled due to the sensitive nature of the data. The UCSF EHR database can be accessed to UCSF-affiliated individuals by contacting UCSF Clinical and Translational Science Institute (ctsi@ucsf.edu) or UCSF's Information Commons team (Info.Commons@ucsf.edu). If the reader is unaffiliated with UCSF, they can set up an official collaboration with a UCSF-affiliated investigator by contacting the PI, Marina Sirota (marina.sirota@ucsf.edu). Requests should be processed within a couple of weeks. UCDDP is only available to UC researchers who have completed analyses in their respective UC first and have provided justification for scaling their analyses across UC health centers (more details at www.ucop.edu/uc-health/functions/center-for-data-driven-insights-and-innovations-cdi2.html or by contacting healthdata@ucop.edu). The SPOKE knowledge network can be accessed at spoke.rbvi.ucsf.edu/, and more details about the network can be found in Morris et al.[28] and mappings to EHR concepts can be found in Nelson et al.[27]

3.7 Tables

Table 3.1 Demographics of patients used in models, and an example matched cohort for the -1 year model

The top table shows characteristics of patients in the UCSF EHR with visits and concepts over 7 years prior to index time. Care utilization information can be found in Supplementary Table 3. The bottom table shows an example of training data where AD and controls are matched by the listed characteristics. Race & ethnicity (R&E) is a single variable derived from an algorithm developed by the UCSF Data Equity Taskforce[74]. NHPI: Native Hawaiian or Pacific Islander

All Filtered Patients (pre-test/train split)				
		Control	**AD**	
n		250545	749	
Age of AD onset (SD)			74.0 (5.6)	
Birth year, mean (SD)		1945.5 (10.2)	1933.9 (5.3)	
First visit age, mean (SD)		51.2 (11.4)	57.0 (10.4)	
Sex, n (%)	Female	139548 (55.7)	468 (62.5)	
	Male	110829 (44.2)	281 (37.5)	
	Nonbinary/Unknown	168 (0.1)		
R&E, n (%)	Asian/NHPI	32427 (12.9)	151 (20.2)	
	Black	17111 (6.8)	62 (8.3)	
	Latinx	15036 (6.0)	53 (7.1)	
	Other/Unknown	28177 (11.2)	45 (6.0)	
	White	157794 (63.0)	438 (58.5)	
Matched Train Patients for -1 year model				
		Control	**AD**	**SMD**
n		4184	523	
Birth year, mean (SD)		1934.2 (5.6)	1934.0 (5.3)	-0.042
First visit age, mean (SD)		57.2 (9.4)	56.9 (10.5)	-0.028
AD onset / index time age, mean (SD)		74.1 (5.8)	74.1 (5.8)	-0.002
Years in EHR, mean (SD)		15.9 (7.8)	15.9 (7.9)	-0.004
Log(n prev visits), mean (SD)		3.6 (1.5)	3.7 (1.6)	0.065
Log(n concepts), mean (SD)		3.1 (1.3)	3.3 (1.4)	0.108
Log(days since first event), mean (SD)		8.5 (0.4)	8.5 (0.4)	0.043
Sex, n (%)	Female	2343 (56.0)	317 (60.6)	0.094
	Male	1841 (44.0)	206 (39.4)	
R&E, n (%)	Asian/NHPI	705 (16.8)	112 (21.4)	0.219
	Black	520 (12.4)	35 (6.7)	
	Latinx	280 (6.7)	39 (7.5)	
	Other/Unknown	223 (5.3)	32 (6.1)	
	White	2456 (58.7)	305 (58.3)	

3.8 Figures

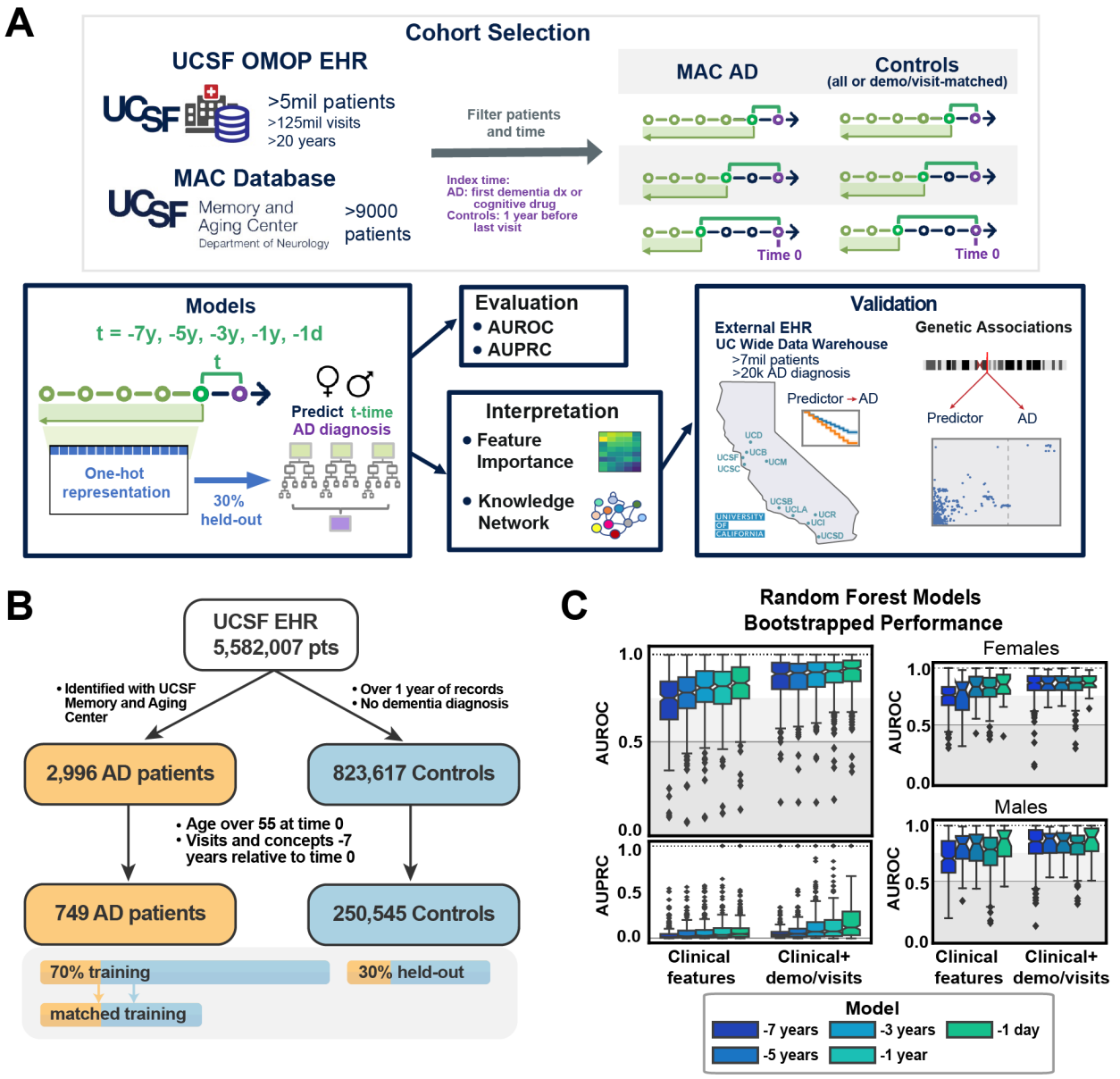


Figure 3.1 Overview of Patient Selection and Random Forest Model Performance

A. From the UCSF electronic health records and the UCSF Memory and Aging center database, patients and clinical information was extracted, filtered, and prepared for time points before the index time. All clinical features extracted were one hot encoded and trained on random forest models to predict future risk of Alzheimer's Disease diagnosis. Models were evaluated on a 30% held-out evaluation set to compute AUROC/AUPRC, and interpreted based on feature importances and using a heterogeneous knowledge network (SPOKE). Top features were then further validated in external databases.

(Figure caption continued on the next page.)

(Figure caption continued from the previous page.)

B. Filtering of a consistent set of AD and Control patients from the UCSF EHR for model training and testing. Filtered patient cohorts are shown in Table 1, and split with 30% held-out set for testing.
C. Bootstrapped performance of random forest models on the full held-out evaluation set (prevalence of AD on held-out set = 0.003). Bootstrapped AUROC performance for models trained and tested on female strata and male strata are also shown.

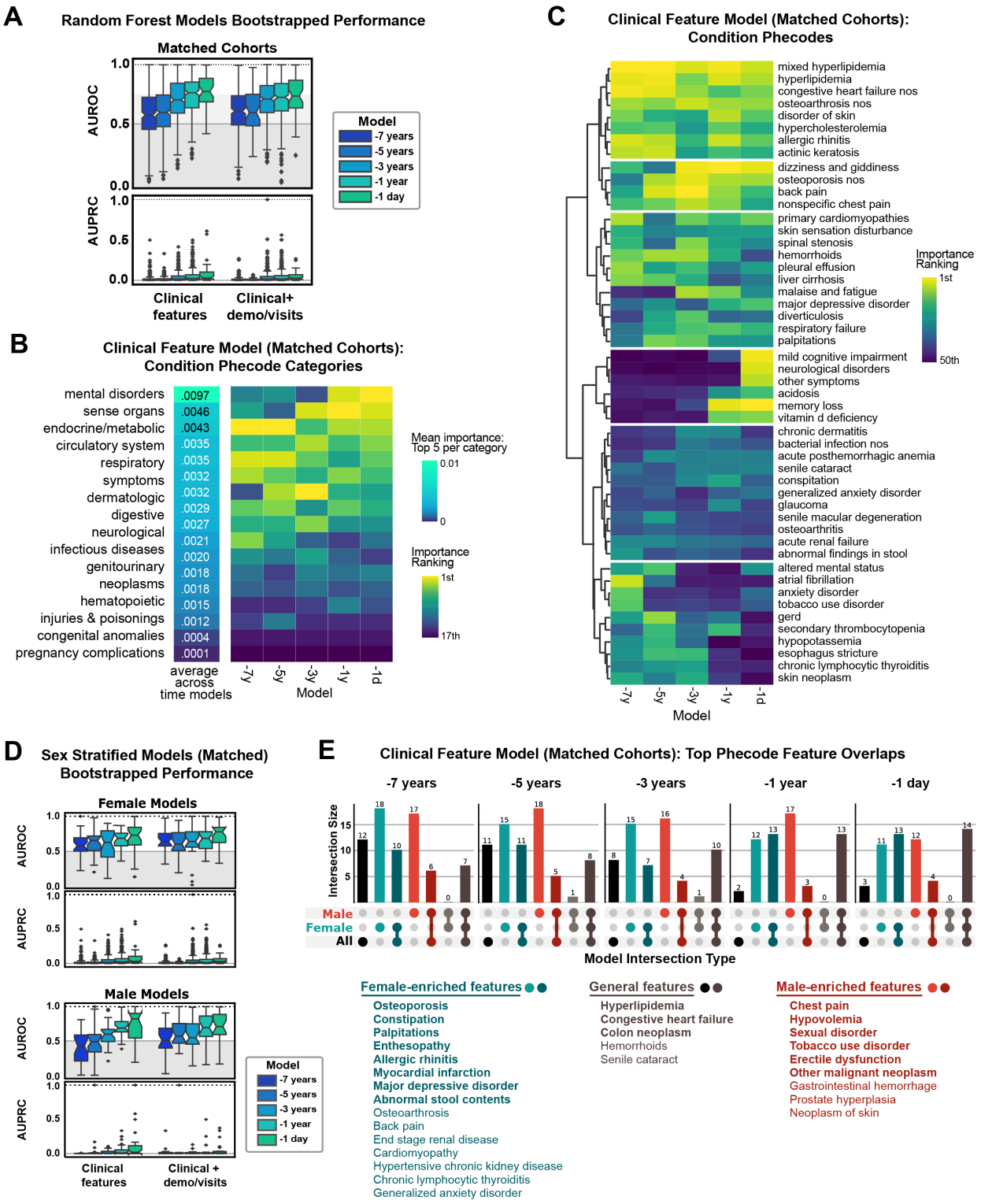


Figure 3.2 Models trained on matched cohorts allows for identification of hypotheses for AD predictors

(Figure caption continued on the next page.)

(Figure caption continued from the previous page.)

A. Bootstrapped performance of models trained on cohorts matched by demographics and visit-related factors on the full held-out evaluation set (prevalence of AD on held-out set = 0.003).
B. Top clinical phecode categories for matched models ranked by the average of the top 5 importance for each phecode category. Sorting is based on this average across time models.
C. Top 50 phecodes (detailed features) across time models, with features clustered based on ward distance of rankings.
D. Bootstrapped performances of sex-stratified matched models on the held-out evaluation set (reference AUPRC = .0036 female, .0022 male).
E. Overlap of top matched model features for models trained on all patients, female stratified patients, and male stratified patients, with model cutoff importance (RF average impurity decrease) greater than 1E-6. Specific features are listed, with bold features indicating top features across all 5 time models, and non-bolded features indicating top features across 4 time models.

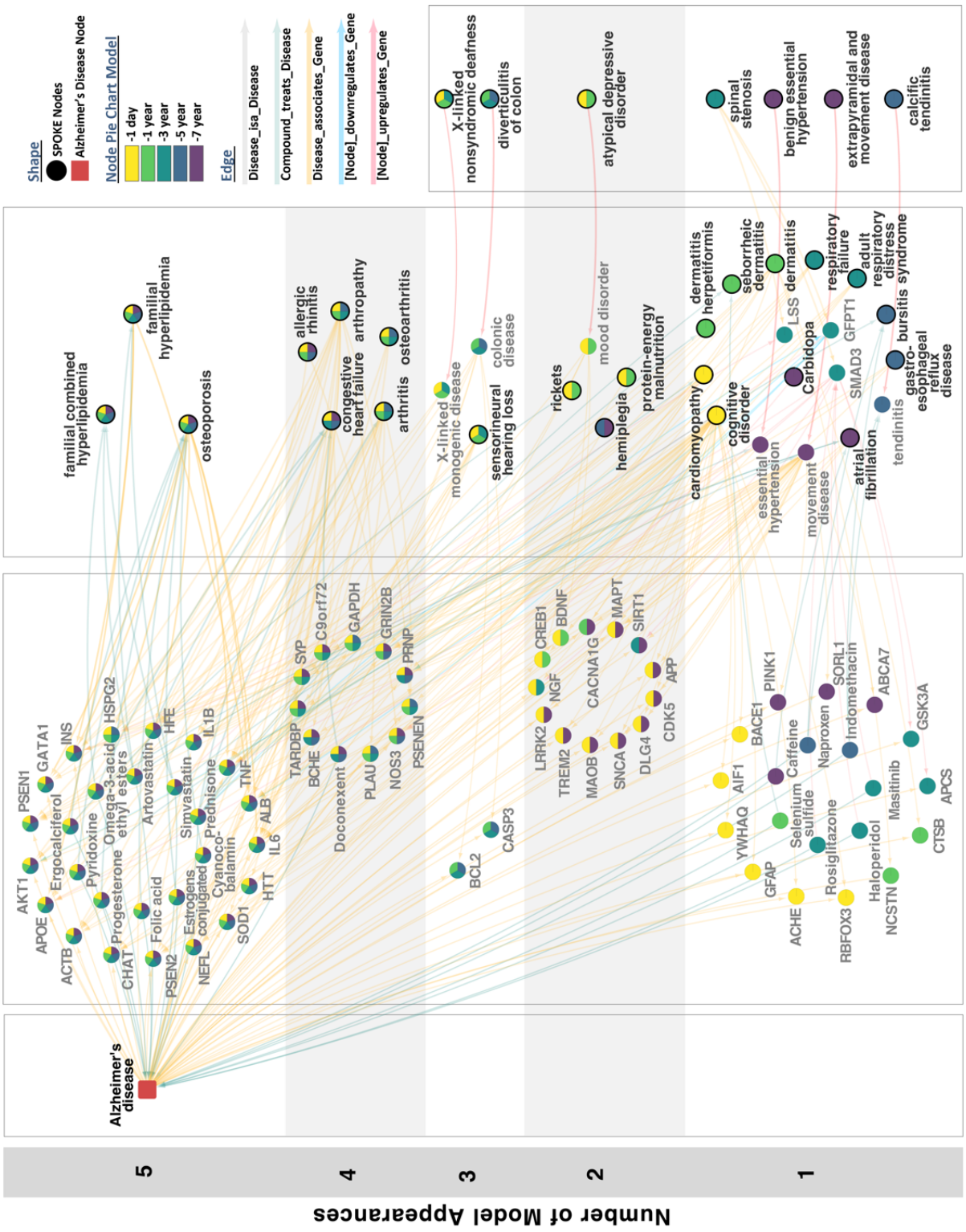


Figure 3.3 SPOKE prioritizes known biological hypotheses associated with shared clinical phenotypes
(Figure caption continued on the next page.)

(Figure caption continued from the previous page.)
Combined SPOKE network of all shortest paths to Alzheimer's Disease node (DOID:10652) for top 25 input features (bolded) from matched AD model at every time point. Network is organized based on the number of time point occurrences (y-axis) and eccentricity of a node in the subnetwork (x-axis). Specific time point occurrences are colored by the pie chart within each node.

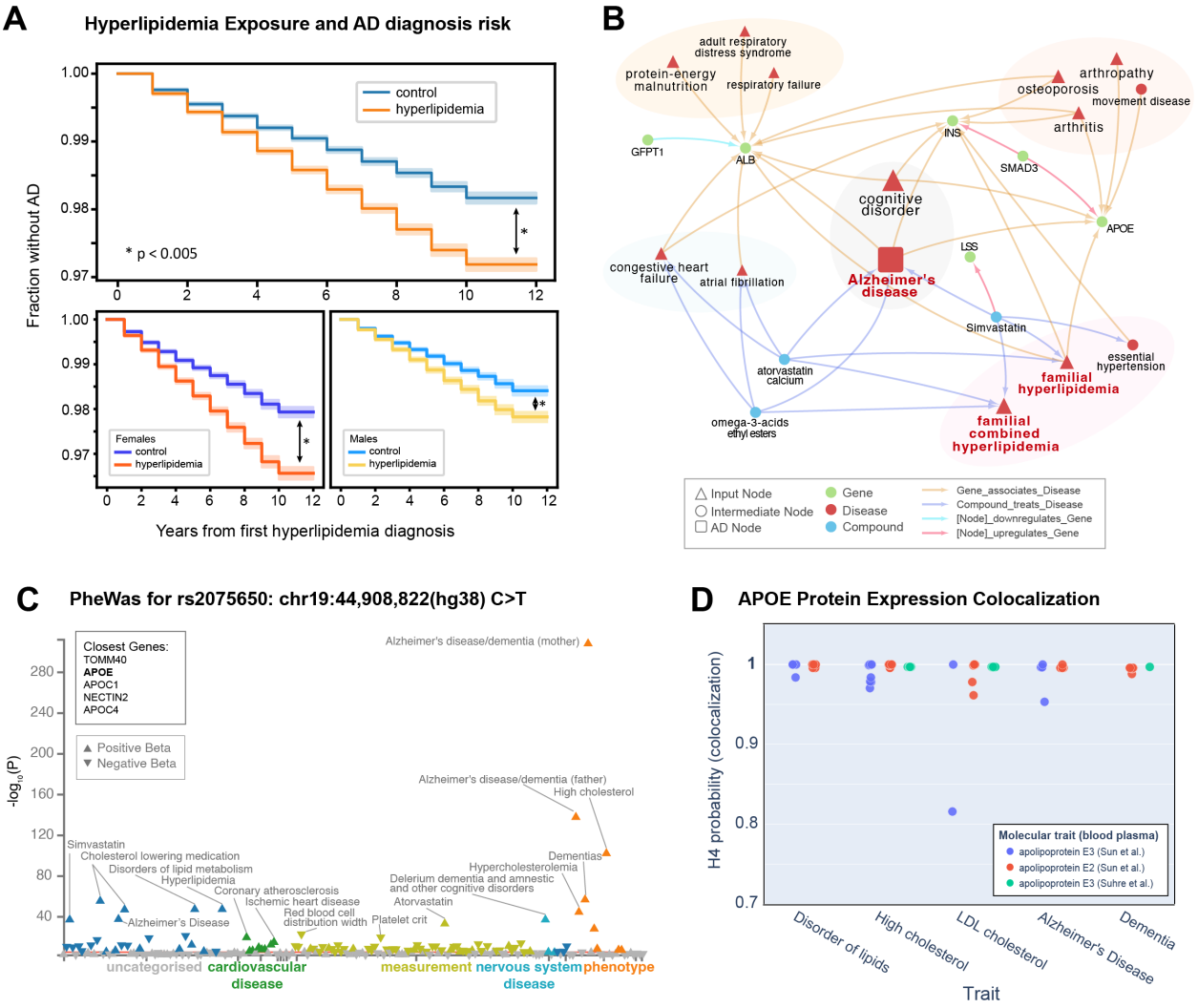


Figure 3.4 The hyperlipidemia and AD association is validated externally with APOE as a shared causal genetic link

A. Kaplan Meyer curve on UC-wide EHR for hyperlipidemia (HLD) as the exposure. Log rank test is significant for all HLD vs controls (p=2.36e-85), female HLD vs female controls (p=3.64e-69), and male HLD vs male controls (p=8.39e-22).
B. 1st and 2nd degree neighbor of hyperlipidemia on the full network representing all shortest paths from the top 25 features per time model.
C. PheWAS for variant rs2075650 on a shared loci associated with both hyperlipidemia and AD, plotted based on associations with phenotypes in the UK Biobank.
D. Plot of APOE protein expression colocalization with H4 (probability two associated traits share a causal variant) from Open Targets Genetics. Each dot represents a specific phenotype categorized based on trait (x-axis). Each color represents an APOE molecular trait measured from blood plasma from Sun et al. and Suhre et al.

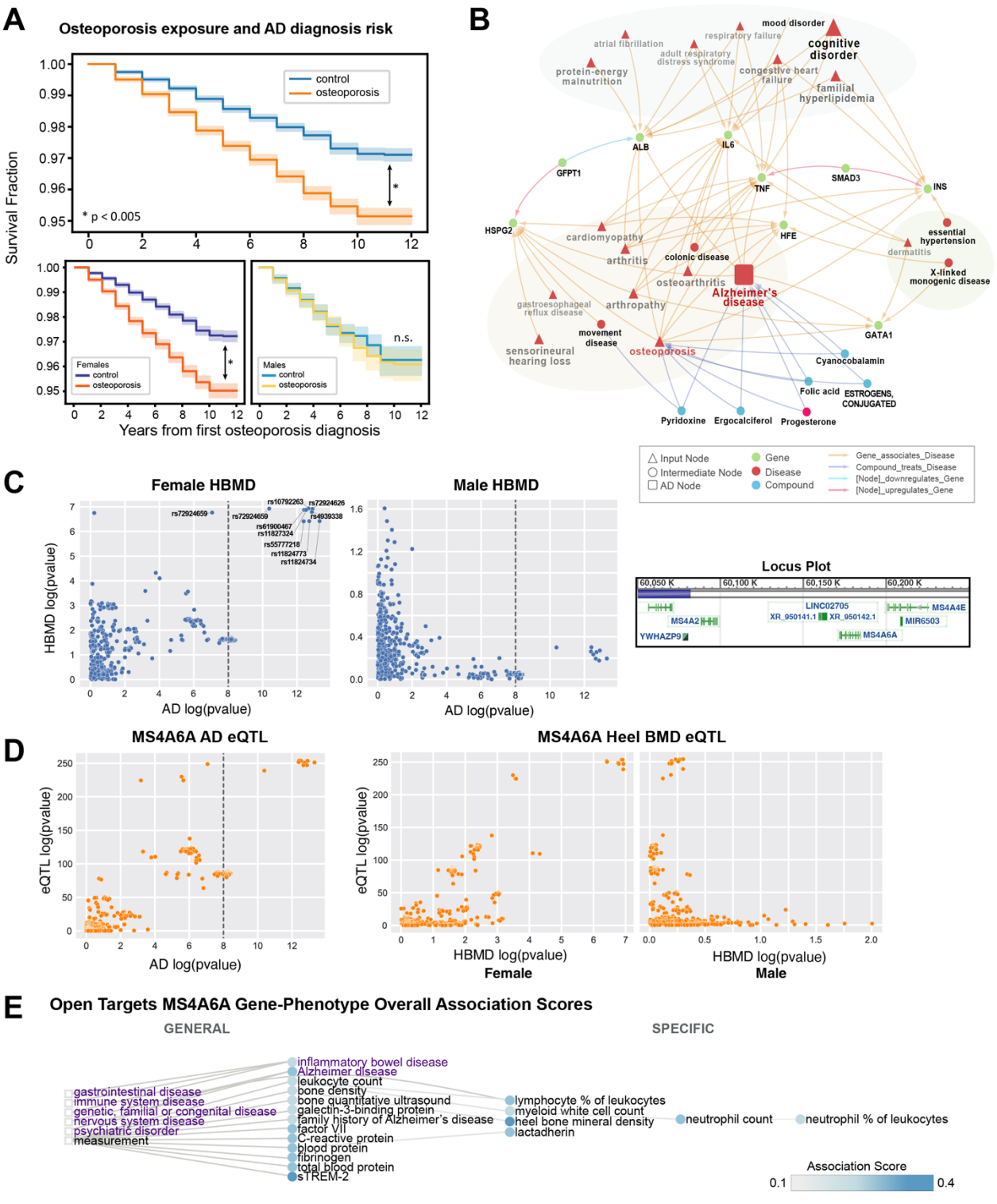


Figure 3.5 The association between osteoporosis and AD is validated externally with MS4A6A as a potential female-specific shared genetic link

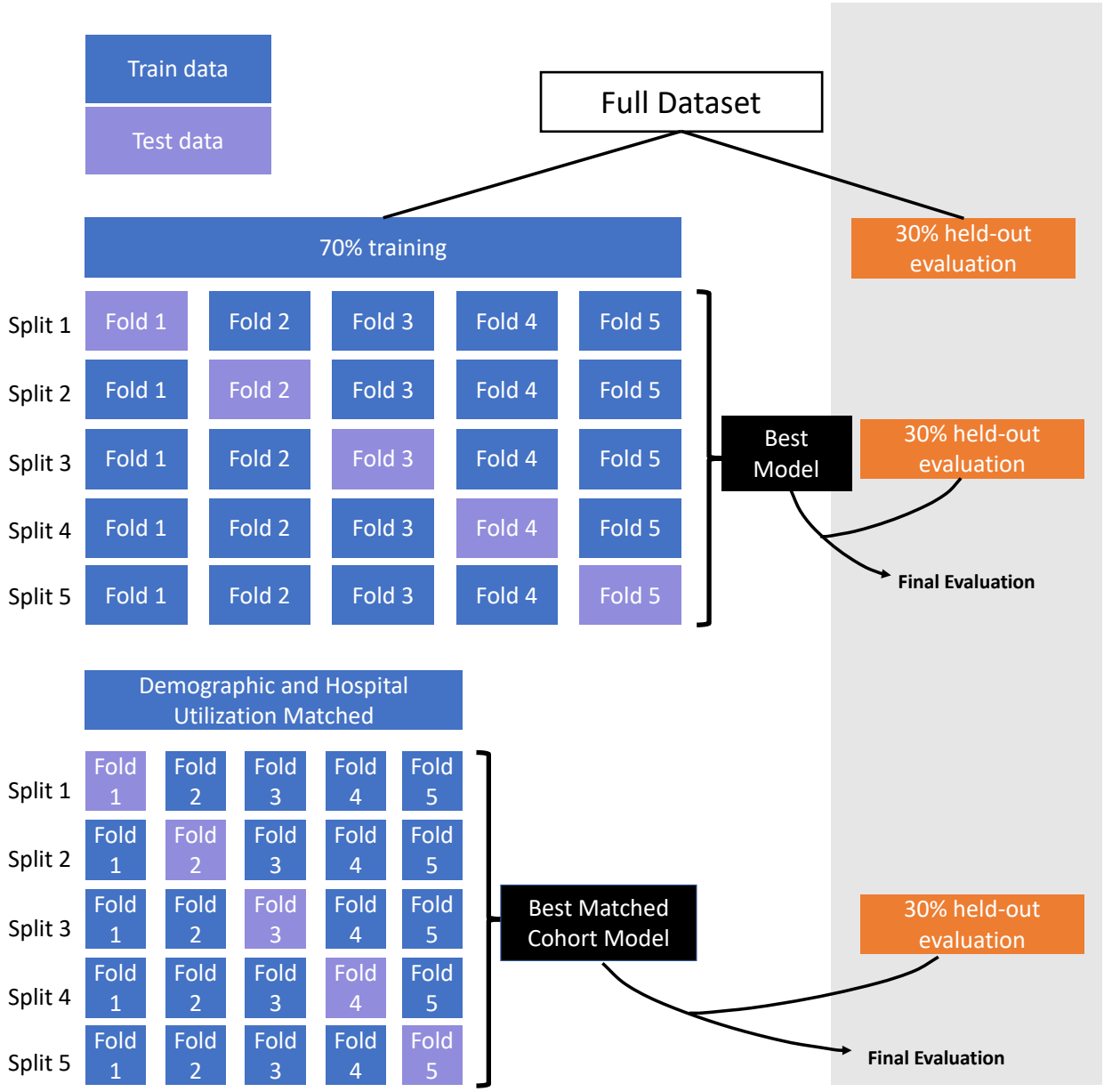
A. Kaplan Meyer curve on UC-wide EHR for osteoporosis as the exposure. Log rank test is

(Figure caption continued on the next page.)

(Figure caption continued from the previous page.)

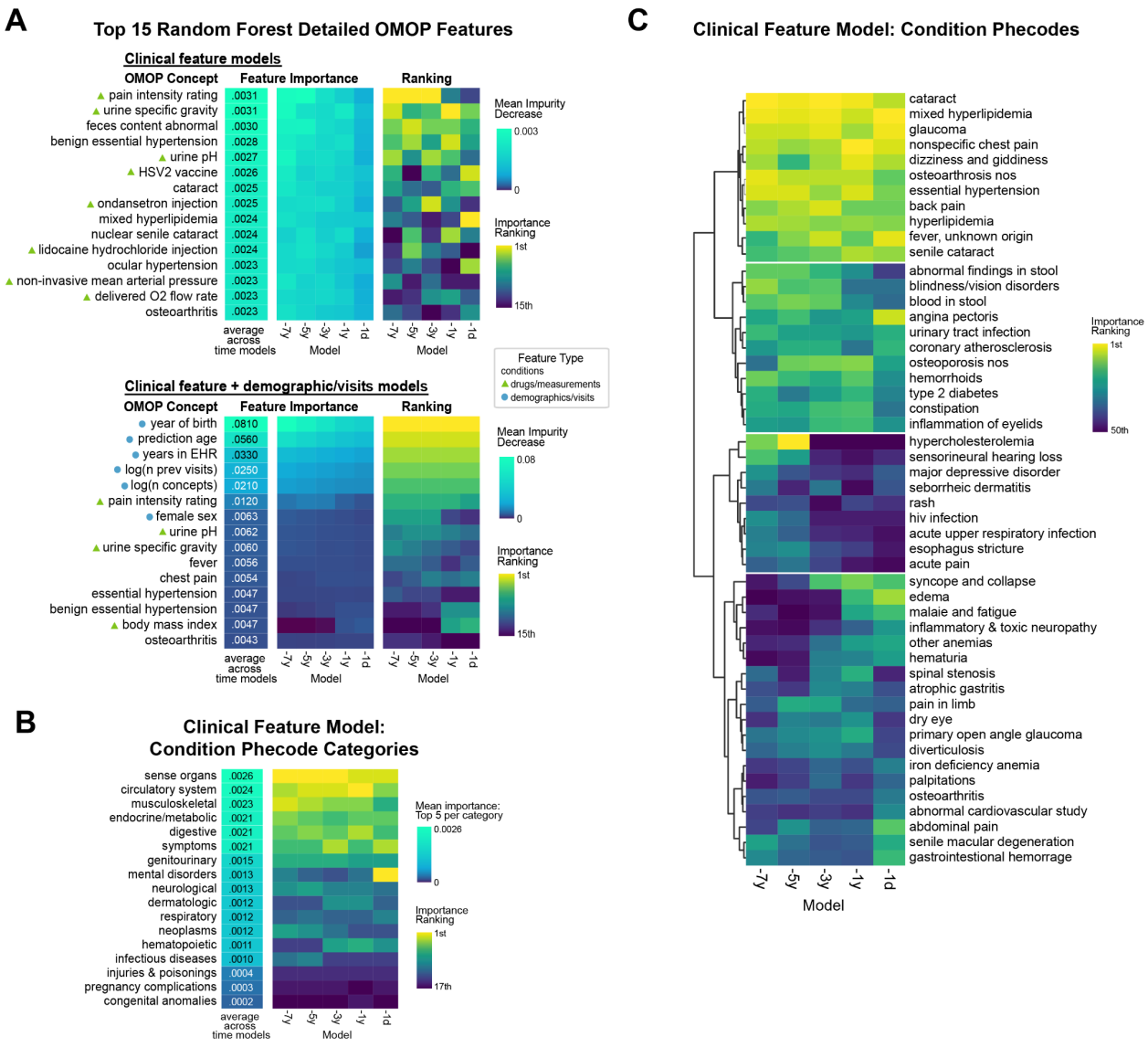
B. significant for all osteoporosis vs controls and similarly for female strata, * $p<0.005$.
C. 1st and 2nd degree neighbors of osteoporosis node on the network representing all shortest paths from top 25 feature per time model.
D. P-P plots between Alzheimer's Disease GWAS (Jensen et al. 2018, n= 455,258) and sex-stratified heel bone mineral density GWAS (Female n = 111,152, Male HBMD n = 166,988, UK Biobank / Neale's Lab GWAS) around the MS4A locus (left and middle plots) at region 60050000-60200000 of Chromosome 11 (locus plot on right).
E. MS4A6A cis eQTL association with AD, and association with sex-stratified heel bone mineral density, from eQTLGen.
F. Open Targets associated phenotype graph for MS4A6A with association score computed based on a weighted harmonic sum across evidence (described in platform-docs.opentargets.org/associations#association-scores). Purple words indicate diseases, while black words indicate measurements. Circles are phenotypes colored by the association score, and boxes represent the most general categories.

3.9 Supplementary Figures



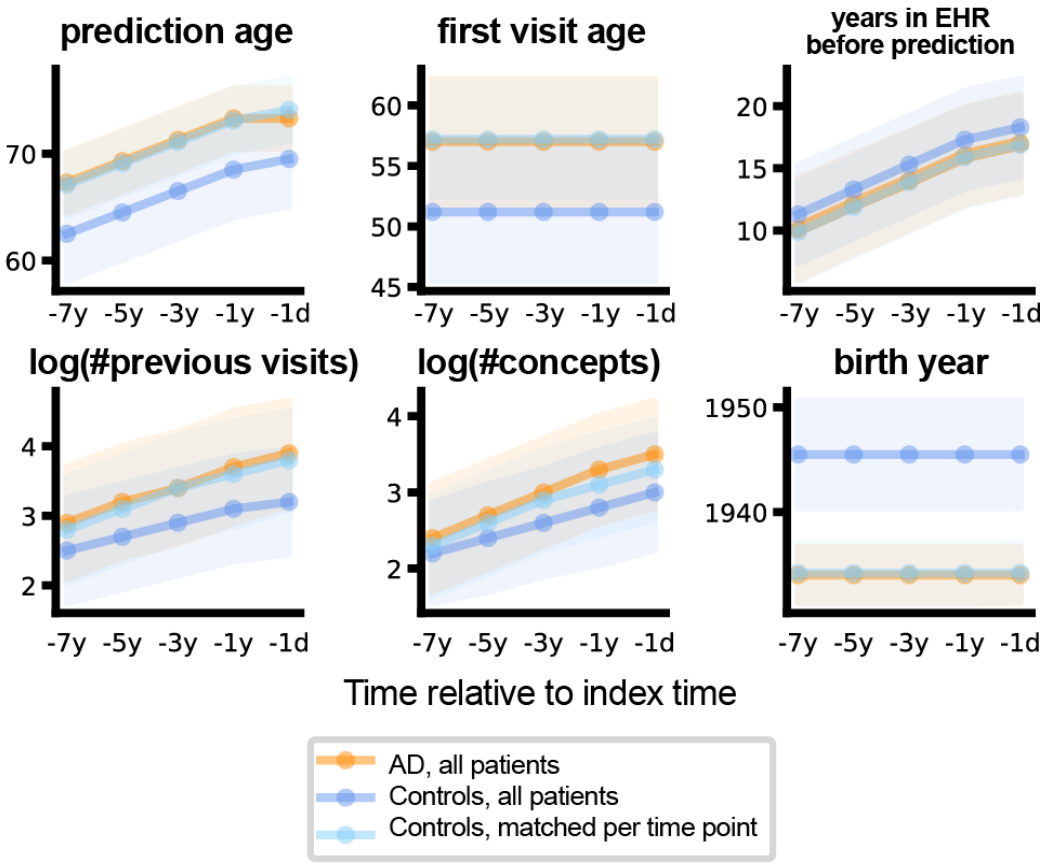
Supplementary Figure 3.1 Approach to Cross-Validation

The full dataset was split into 70% for training and chosing the best model, and 30% was set aside as the held-out evaluation set. Model selection and optimization was performed with cross-validation on the 70% training set. All final models are then evaluated on the 30% held-out evaluation set.



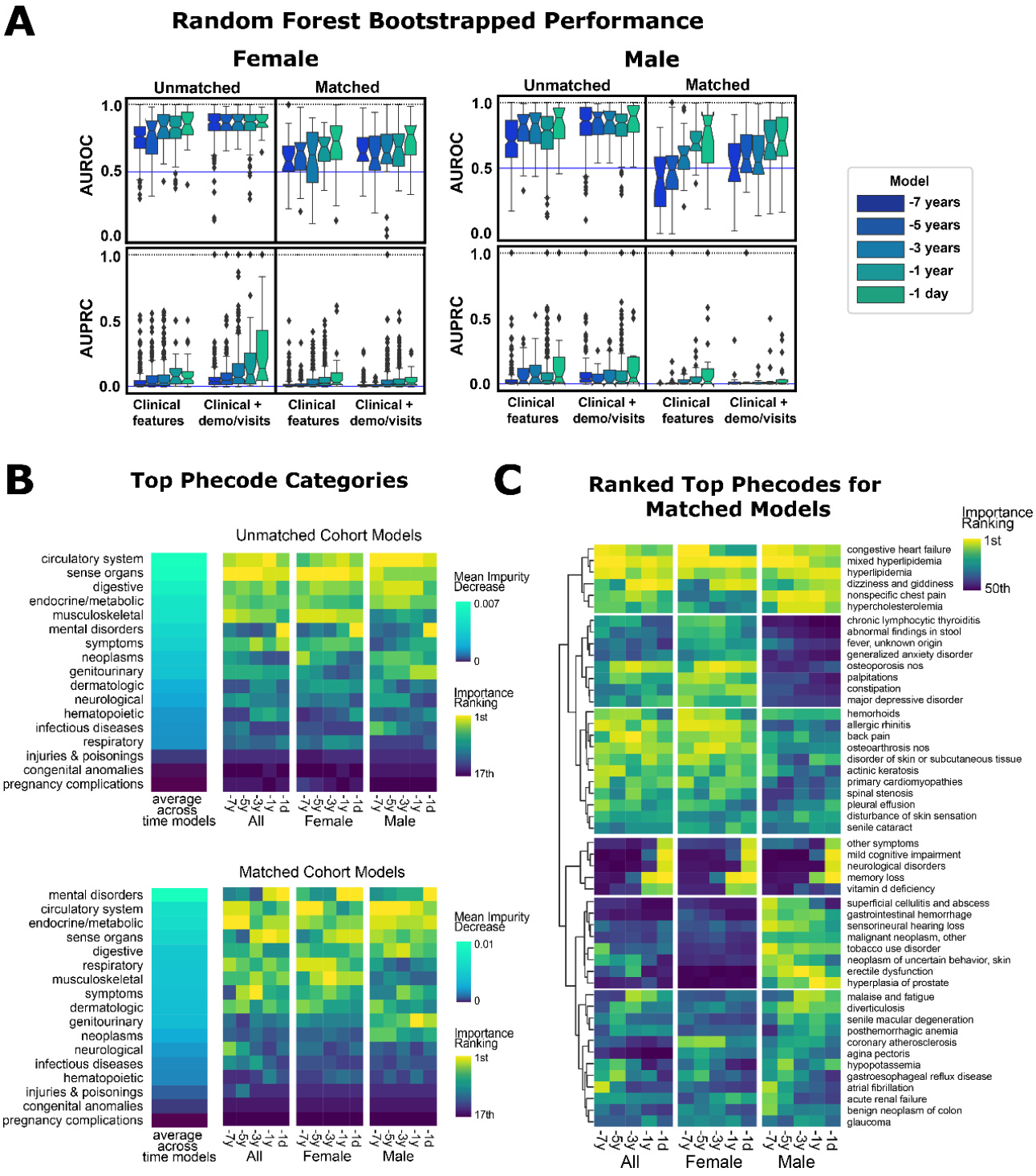
Supplementary Figure 3.2 Top detailed features and phecodes from the random forest model

A. Top detailed OMOP clinical features utilized in models for clinical feature only models (top), or clinical features + demographic + visit information models (bottom). Features within the drug/measurement categories are marked with a triangle, while demographic/visit features are marked with a circle.
B. Top phecode categories utilized in models, where importance is determined by the top 5 detailed features within each phecode mapping. The vertical order is based upon the average importance across time models.
C. Top 50 phecodes utilized in time models, clustered based on relative importance across time models.



Supplementary Figure 3.3 Comparison of age and visit-related factors between AD, controls, and matched controls

The plots demonstrate the distribution of continuous variables utilized in matching with standard deviation. Orange represents AD patients at each time point. Dark blue represents all controls, while light blue represents controls that have been matched at each time point.



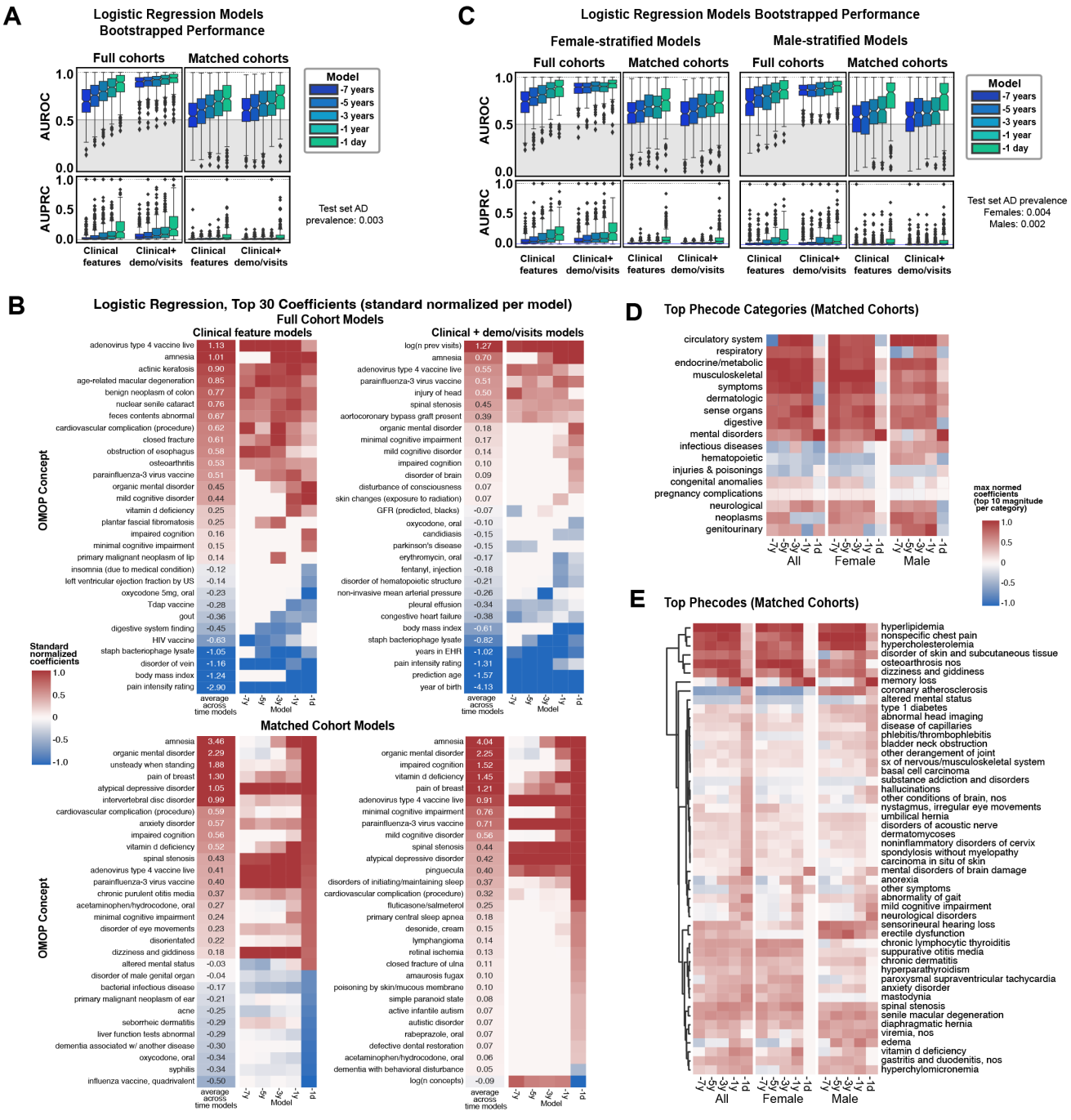
Supplementary Figure 3.4 Sex stratified models elucidate performance differences and sex predictive features that drive the total cohort models

A. The full performance of sex-stratified models are shown. The bootstrapped AUROC/AUPRC is determined by the male or female strata of the initial 30% held-out

(Figure caption continued on the next page.)

(Figure caption continued from the previous page.)

B. evaluation set. Horizontal lines represent median and quartiles for the bootstrapped performance.
C. Top phecode categories are listed by importance for all models, with inclusion of comparison with the general non-stratified model. Vertical ordering is determined by the average importance across time models.
D. Top 50 important phecodes clustered by relative importance across time models and across strata.



Supplementary Figure 3.5 Logistic regression models identifies some similar predictive features

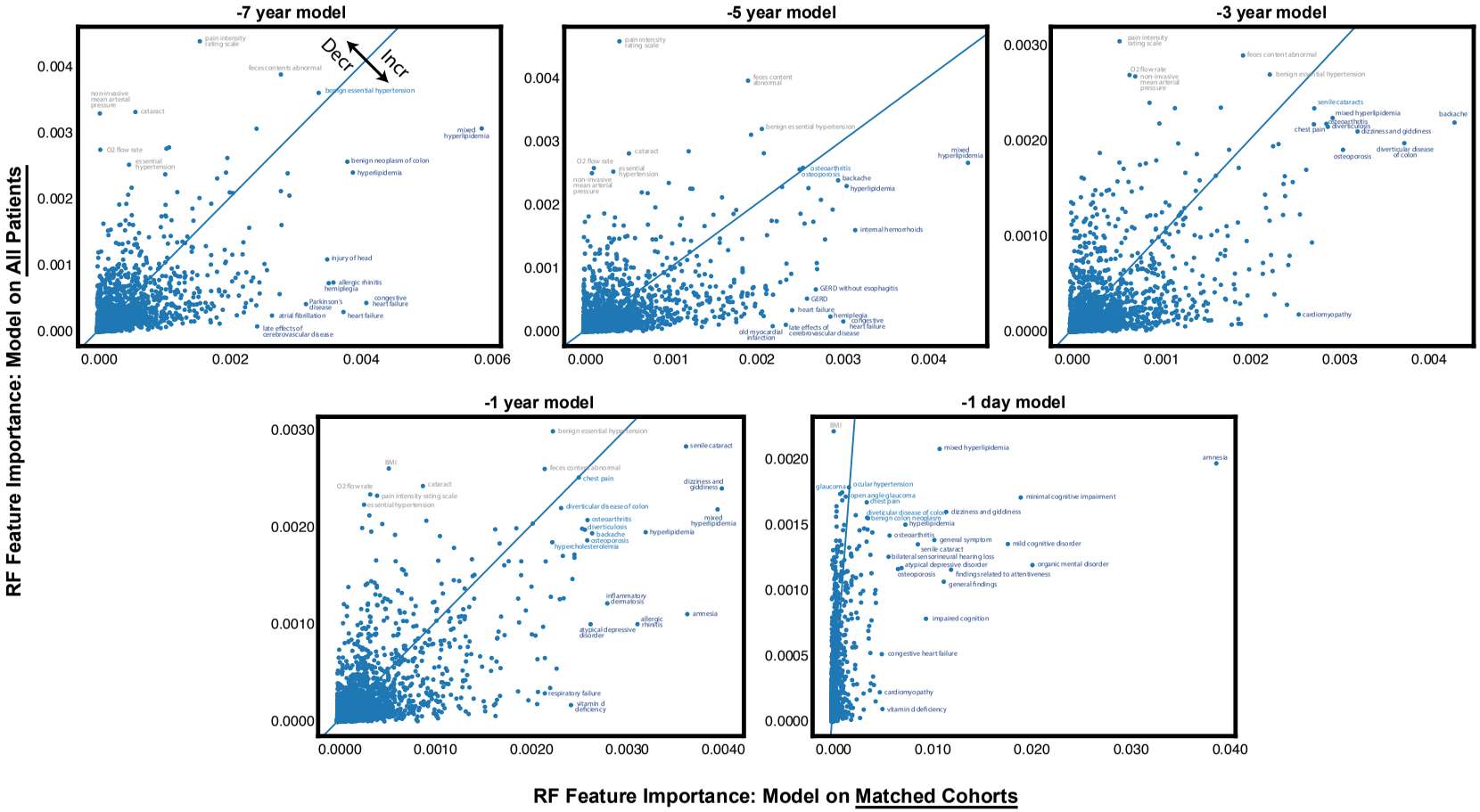
A. The full performance of logistic regression models. The bootstrapped AUROC/AUPRC is determined the 30% held-out evaluation set.
B. Top detailed OMOP feature logistic regression coefficients are listed by importance for all model formulations. Top row shows coefficients from the model trained on all

(Figure caption continued on the next page.)

(Figure caption continued from the previous page.)

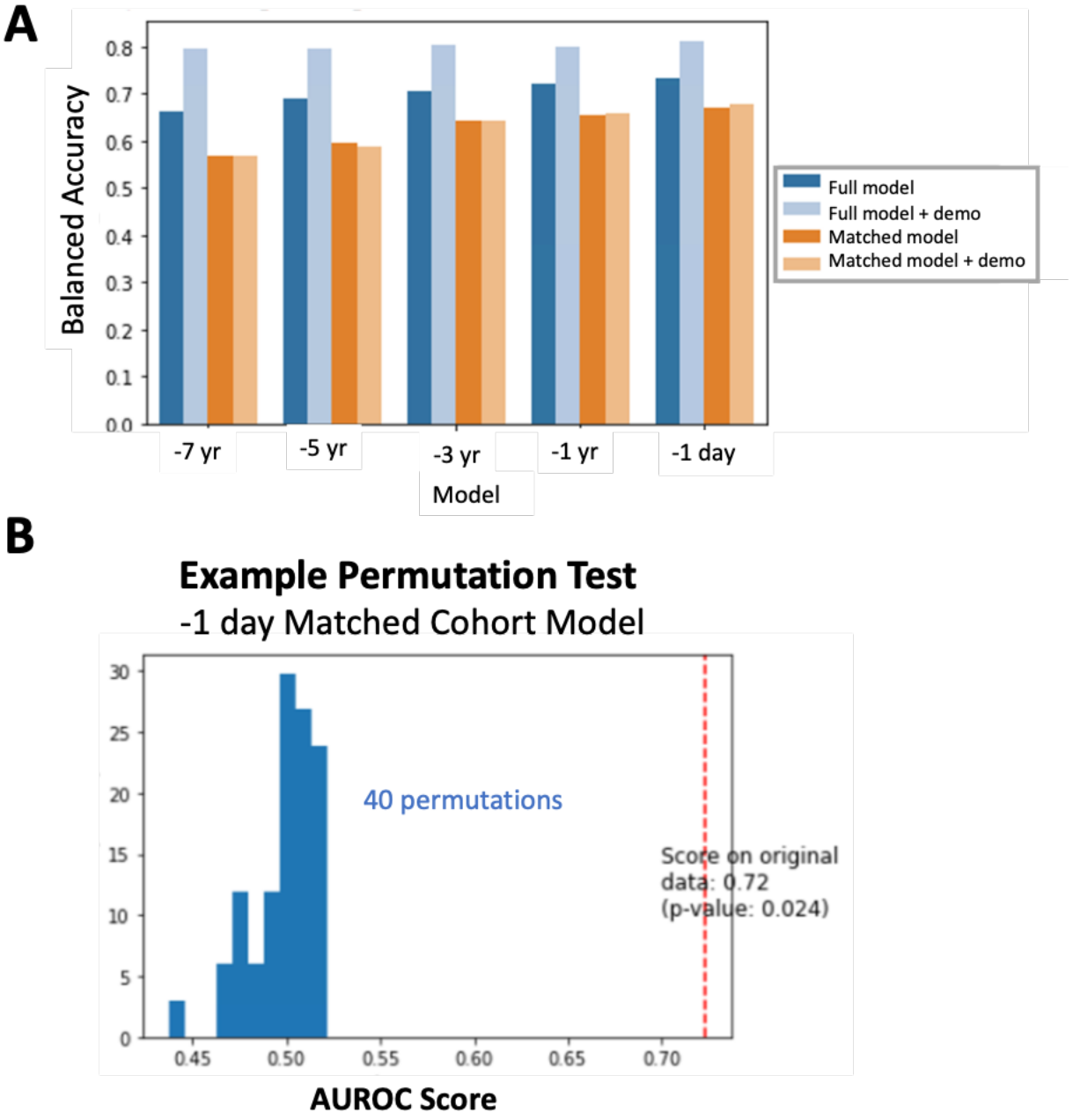
patients, while the bottom row shows coefficients from the model trained on matched cohorts.

C. The full performance of sex-stratified logistic regression models are shown. The bootstrapped AUROC/AUPRC is determined by the male or female strata of the initial 30% held-out evaluation set.
D. Top phecode categories across time models and across strata, determined by the top 10 logistic regression coefficient magnitudes within each category.
E. Top 50 important phecodes clustered by average logistic regression coefficient across time models and across strata, where the average logistic regression coefficient is determined by the top 10 logistic regression coefficient magnitudes within each category.



Supplementary Figure 3.6 Random Forest Feature Importance Changes Models

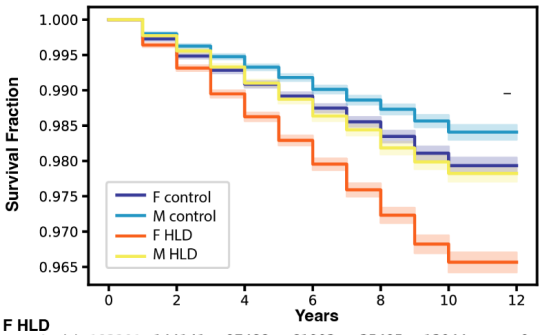
A comparison of the random forest model feature importances between the model trained on all patients (y-axis) and the model trained on demographics/care utilization matched cohorts (x-axis). The blue line represents no change in feature importance. Above the blue line represents a **decr**ease in feature importance in the model trained on the full cohort compared to matched cohorts, and below the line represents features with **incr**eased importance for the model trained on matched cohorts.



Supplementary Figure 3.7 Balanced Accuracy and Example Permutation Test

A. Balanced accuracy on the 30% held-out evaluation set was computed for all random forest models.
B. A null distribution for AUROC was computed based on retrained random forest models with permutations on the ground truth label (40 permutations)

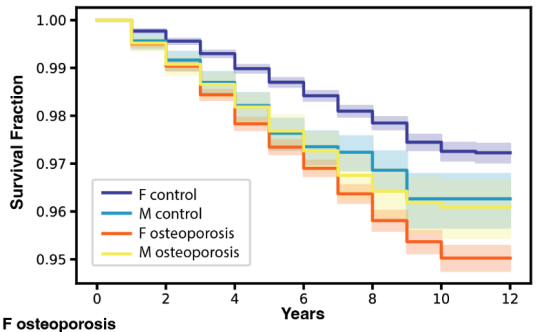
A Hyperlipidemia → AD survival curve



F HLD							
At risk	183201	144141	97488	61903	35495	13044	0
Censored	0	37863	83591	118579	144581	166818	179862
Events	0	1197	2122	2719	3125	3339	3339
M HLD							
At risk	178551	143424	97398	61518	33611	12010	0
Censored	0	34374	79793	115262	142925	164421	176431
Events	0	753	1360	1771	2015	2120	2120
F control							
At risk	185375	184420	112506	58973	28016	8729	0
Censored	0	0	71230	124432	155187	174375	183104
Events	0	955	1639	1970	2172	2271	2271
M control							
At risk	177265	176599	111455	60617	28098	9024	0
Censored	0	0	64658	115190	147561	166560	175584
Events	0	666	1152	1458	1606	1681	1681

Log Rank Test Comparison	test statistic	-log2(p)	pval
all: HLD vs control	383.32	281.13	2.36E-85
F: HLD vs control	308.98	227.35	3.64E-69
M: HLD vs control	92.06	70.01	8.39E-22
F vs M HLD	255.75	188.82	1.45E-57

B Osteoporosis → AD survival curve



F osteoporosis							
At risk	58737	47341	37081	25278	14093	4280	0
Censored	0	10854	20568	32042	42975	52690	56970
Events	0	542	1088	1417	1669	1767	1767
M osteoporosis							
At risk	9597	7282	5275	3208	1622	474	0
Censored	0	2231	4177	6199	7760	8903	9377
Events	0	84	145	190	215	220	220
F control							
At risk	60141	59876	59530	44003	15587	3199	0
Censored	0	0	0	15187	43391	55699	58897
Events	0	265	611	951	1163	1243	1244
M control							
At risk	8501	8430	8349	5079	1620	389	0
Censored	0	0	0	3197	6640	7861	8250
Events	0	71	152	225	241	251	251

Log Rank Test Comparison	test statistic	-log2(p)	pval
all: osteoporosis vs control	287.91	212.1	1.42E-64
F: osteoporosis vs control	321.39	236.33	7.22E-72
M: osteoporosis vs control	0.55	1.12	4.60E-01
F vs M osteoporosis	5.14	5.42	2.34E-02

C UCDDP: Hyperlipidemia Exposure, AD Diagnosis Outcome

	No Strata			Strata: recruitment age		
Model	**Hazard Ratio**	**95% CI**	**p-value**	**Hazard Ratio**	**95% CI**	**p-value**
Unadjusted	**1.53**	[1.47, 1.58]	2.18E-124	**1.49**	[1.44, 1.54]	9.79E-111
demographics adjusted	**1.43**	[1.38, 1.48]	1.12E-87	**1.47**	[1.42, 1.53]	1.43E-104
visit adjusted	**1.32**	[1.26, 1.37]	9.22E-42	**1.28**	[1.23, 1.34]	1.11E-32
visit/demographics adjusted	**1.27**	[1.22, 1.32]	1.12E-27	**1.28**	[1.23, 1.33]	1.51E-31

D UCDDP: Osteoporosis Exposure, AD Diagnosis Outcome

	No Strata			Strata: recruitment age		
Model	**Hazard Ratio**	**95% CI**	**p-value**	**Hazard Ratio**	**95% CI**	**p-value**
Unadjusted	**1.81**	[1.70, 1.92]	5.20E-82	**1.71**	[1.61, 1.82]	7.10E-67
demographics adjusted	**1.61**	[1.52, 1.72]	1.52E-52	**1.70**	[1.60, 1.81]	6.98E-07
visit adjusted	**1.68**	[1.56, 1.80]	4.34E-47	**1.59**	[1.48, 1.72]	4.57E-34
visit/demographics adjusted	**1.57**	[1.45, 1.70]	7.96E-29	**1.57**	[1.46, 1.69]	1.05E-31

Supplementary Figure 3.8 UCDDP hyperlipidemia and osteoporosis survival curve numbers and cox proportional hazard model results

A. Hyperlipidemia sex-stratified combined Kaplan-Meyer survival curves with counts. 95% confidence interval are shown. Log rank test comparison results are below.
B. Osteoporosis sex-stratified combined Kaplan-Meyer survival curves with counts. 95% confidence interval are shown. Log rank test comparison results are below.

(Figure caption continued on the next page.)

(Figure caption continued from the previous page.)

C. Hyperlipidemia exposure cox proportional hazard models for AD as the outcome, shown are the hazard ratios and 95% confidence intervals obtained from the exposure coefficient for unadjusted, demographic adjusted (gender, age, race, ethnicity), visit adjusted (first visit age, log(number of visits)), and demographic/visit adjusted. Right group shows computed hazard ratios with stratification by recruitment or starting age (age strata: <55, 55-60, 60-65, 65-70, 70-75, 75-80, >80).
D. Osteoporosis exposure cox proportional hazard models for AD as the outcome, shown are the hazard ratios and 95% confidence intervals obtained from the exposure coefficient for unadjusted, demographic adjusted, visit adjusted, and demographic/visit adjusted. Right group shows computed hazard ratios with stratification by recruitment or starting age (age strata: <60, 60-65, 65-70, 70-75, 75-80, >80).

3.10 Supplementary Tables

Supplementary Table 3.1 Control exclusion codes

List of mappings from ICD-10 codes G[123]* to OMOP codes for determining exclusion of Controls. The mapping was generated and manually reviewed to white-list certain codes and approve exclusion of dementia-related codes.
Download at this link:
https://www.medrxiv.org/content/medrxiv/early/2023/03/19/2023.03.14.23287224/DC2/embed/media-2.xlsx?download=true

Supplementary Table 3.2 Dementia codes

List of mappings from Dementia/FTD related condition concepts to SNOMED OMOP mappings and N06D ATC code to RxNorm OMOP mappings for identifying index time 0 for AD patients.
Download at this link:
https://www.medrxiv.org/content/medrxiv/early/2023/03/19/2023.03.14.23287224/DC3/embed/media-3.xlsx?download=true

Supplementary Table 3.3 Matching results for time point models on matched cohorts

Demographics of matched cohorts (propensity-score matched by demographics and visit-related factors, see Methods) on the training set for hypothesis generation models.
Download at this link:
https://www.medrxiv.org/content/medrxiv/early/2023/03/19/2023.03.14.23287224/DC4/embed/media-4.xlsx?download=true

Supplementary Table 3.4 Male and female demographics and matching result

Demographics of male and female cohorts (combined train and test set). The same patients for train/test set split in the general model are utilized for the sex-stratified models. Matched cohorts on the sex-strata training sets are also shown for hypothesis generation models.
Download at this link:
https://www.medrxiv.org/content/medrxiv/early/2023/03/19/2023.03.14.23287224/DC5/embed/media-5.xlsx?download=true

Supplementary Table 3.5 Matched cohort trained model comparison between logistic regression and random forest

Mean and standard deviations AUROC was computed for bootstrapped samples of the held-out evaluation set for both the random forest and logistic regression models for comparability.

Model Time	Bootstrapped mean AUROC	Bootstrapped std AUROC	Model Type	Features
-1 day	0.771667	0.139762	random forest	clinical
-1 yr	0.738797	0.16183	random forest	clinical
-1 day	0.70313	0.20803	logistic regression	clinical
-3 yr	0.695912	0.183331	random forest	clinical
-1 yr	0.674981	0.189248	logistic regression	clinical
-3 yr	0.637145	0.187386	logistic regression	clinical
-5 yr	0.598022	0.207104	random forest	clinical
-5 yr	0.589743	0.197771	logistic regression	clinical
-7 yr	0.583837	0.206874	random forest	clinical
-7 yr	0.549061	0.192763	logistic regression	clinical

Model Time	Bootstrapped mean AUROC	Bootstrapped std AUROC	Model Type	Features
-1 day	0.738133	0.170242	random forest	clinical + demo/visits
-1 day	0.728514	0.182411	logistic regression	clinical + demo/visits
-1 yr	0.710432	0.183543	random forest	clinical + demo/visits
-3 yr	0.700636	0.175239	random forest	clinical + demo/visits
-1 yr	0.663312	0.187115	logistic regression	clinical + demo/visits
-3 yr	0.657146	0.198652	logistic regression	clinical + demo/visits
-5 yr	0.619819	0.201044	logistic regression	clinical + demo/visits
-7 yr	0.60619	0.181737	random forest	clinical + demo/visits
-5 yr	0.599069	0.199292	random forest	clinical + demo/visits
-7 yr	0.59507	0.199938	logistic regression	clinical + demo/visits

Supplementary Table 3.6 Balanced accuracy performance of models

Balanced accuracy (defined as average recall for both the positive and negative class) performance on the held-out evaluation set for both the full model and the matched cohort trained model.

Model Time	Full or matched cohort model	Held-out evaluation set balanced accuracy	Features
-7 yr	full	0.66333567	clinical
-5 yr	full	0.68713887	clinical
-3 yr	full	0.70504803	clinical
-1 yr	full	0.72020014	clinical
-1 day	full	0.73342356	clinical
-7 yr	matched	0.56751959	clinical
-5 yr	matched	0.59464602	clinical
-3 yr	matched	0.64351437	clinical
-1 yr	matched	0.65277962	clinical
-1 day	matched	0.67079665	clinical

Model Time	Full or matched cohort model	Held-out evaluation set balanced accuracy	Features
-7 yr	full	0.79420735	clinical + demo/visits
-5 yr	full	0.79553055	clinical + demo/visits
-3 yr	full	0.80227966	clinical + demo/visits
-1 yr	full	0.79940679	clinical + demo/visits
-1 day	full	0.81155760	clinical + demo/visits
-7 yr	matched	0.56836172	clinical + demo/visits
-5 yr	matched	0.58730455	clinical + demo/visits
-3 yr	matched	0.64186021	clinical + demo/visits
-1 yr	matched	0.65705140	clinical + demo/visits
-1 day	matched	0.67644997	clinical + demo/visits

Supplementary Table 3.7 UCDDP AD patient concepts and demographics

Top table shows the specific concepts utilized to identify Alzheimer's Disease as the outcome in the UCDDP database, with breakdown by number of patients per concept. Due to deidentification, only a patient's birth year is known for age estimation.

Term	**# patients**
Alzheimer's disease	20562
Primary degenerative dementia of the Alzheimer type, senile onset	9327
Primary degenerative dementia of the Alzheimer type, presenile onset	2530

		Overall
n		24389
estimated_age, mean (SD)		45.6 (23.5)
gender, n (%)	**FEMALE**	12915 (53.0)
	MALE	11391 (46.7)
	UNKNOWN	83 (0.3)
race, n (%)	**Native**	78 (0.3)
	Asian	2069 (8.5)
	Black	1079 (4.4)
	Multirace	494 (2.0)
	NHPI	108 (0.4)
	Other Race	3413 (14.0)
	Unknown	6535 (26.8)
	White	10613 (43.5)
ethnicity, n (%)	**Hispanic or Latino**	3815 (15.6)
	Not Hispanic or Latino	13869 (56.9)
	Unknown	6705 (27.5)
# visits, mean (SD)	missing = 3092	21.1 (51.8)

Supplementary Table 3.8 Hyperlipidemia UCDDP concepts and demographics

Top table shows the specific concepts utilized to identify HLD as the exposure in the UCDDP database, with breakdown by number of patients per concept. Due to deidentification, only a patient's birth year is known for age estimation. Recruitment age is utilized as the starting age for survival analysis, with HLD group as the age of HLD diagnosis, and unexposed group as the age of first EHR visit. H/L: Hispanic/latino.

Term	# patients
Hyperlipidemia	702142
Mixed hyperlipidemia	169316

		Overall	No HLD	HLD	SMD
n		728578	364289	364289	
gender, n (%)	**FEMALE**	371050 (50.9)	186259 (51.1)	184791 (50.7)	0.037
	MALE	357255 (49.0)	177768 (48.8)	179487 (49.3)	
	UNKNOWN	273 (0.0)	262 (0.1)	11 (0.0)	
race, n (%)	**Native**	3278 (0.4)	1762 (0.5)	1516 (0.4)	0.113
	Asian	69432 (9.5)	32466 (8.9)	36966 (10.1)	
	Black	35072 (4.8)	16512 (4.5)	18560 (5.1)	
	Multirace	17486 (2.4)	7635 (2.1)	9851 (2.7)	
	NHPI	2972 (0.4)	1270 (0.3)	1702 (0.5)	
	Other Race	81646 (11.2)	44093 (12.1)	37553 (10.3)	
	Unknown	81062 (11.1)	44889 (12.3)	36173 (9.9)	
	White	437630 (60.1)	215662 (59.2)	221968 (60.9)	
ethnicity, n (%)	**H/L**	102163 (14.0)	53581 (14.7)	48582 (13.3)	0.126
	Not H/L	560067 (76.9)	271574 (74.5)	288493 (79.2)	
	Unknown	66348 (9.1)	39134 (10.7)	27214 (7.5)	
estimated_age, mean (SD)		69.7 (10.8)	69.6 (11.0)	69.8 (10.7)	0.012
recruitment_age, mean (SD)		63.9 (10.5)	63.4 (10.5)	64.3 (10.5)	0.087

Supplementary Table 3.9 Osteoporosis UCDDP concepts and demographics

Top table shows the specific concepts utilized to identify osteoporosis as the exposure in the UCDDP database with inclusion of children concepts, and breakdown by number of patients per concept. Due to deidentification, only a patient's birth year is known for age estimation. Recruitment age is utilized as the starting age for survival analysis, with osteoporosis group as the age of osteoporosis diagnosis, and unexposed group as the age of first EHR visit

Term	# patients
Osteoporosis	145608
Senile osteoporosis	30611
Osteoporotic fracture	7772
Osteoporotic fracture of vertebra	3987
Localized osteoporosis - Lequesne	3126
Osteoporotic fracture of femur	2971
Idiopathic osteoporosis	1231
Disuse osteoporosis	309
Osteoporotic fracture of humerus	186
Osteoporotic fracture of hand	39

		Overall	**No osteo**	**osteo**	**SMD**
n		137880	68940	68940	
gender, n (%)	**FEMALE**	119637 (86.8)	60386 (87.6)	59251 (85.9)	0.049
	MALE	18241 (13.2)	8554 (12.4)	9687 (14.1)	
	UNKNOWN	2 (0.0)		2 (0.0)	
race, n (%)	**Native**	496 (0.4)	272 (0.4)	224 (0.3)	0.134
	Asian	15784 (11.4)	7364 (10.7)	8420 (12.2)	
	Black	4611 (3.3)	2546 (3.7)	2065 (3.0)	
	Multirace	3564 (2.6)	1737 (2.5)	1827 (2.7)	
	NHPI	419 (0.3)	198 (0.3)	221 (0.3)	
	Other Race	13032 (9.5)	7427 (10.8)	5605 (8.1)	
	Unknown	13670 (9.9)	7552 (11.0)	6118 (8.9)	
	White	86304 (62.6)	41844 (60.7)	44460 (64.5)	
ethnicity, n (%)	**H/L**	15530 (11.3)	8509 (12.3)	7021 (10.2)	0.133
	Not H/L	112474 (81.6)	54548 (79.1)	57926 (84.0)	
	Unknown	9876 (7.2)	5883 (8.5)	3993 (5.8)	
estimated_age, mean (SD)		74.8 (9.2)	75.2 (9.1)	74.5 (9.3)	-0.074
recruitment_age, mean (SD)		68.7 (8.9)	68.2 (8.7)	69.2 (9.1)	0.12

3.11 Supplementary Data

Supplementary Data 3.1 Model Inputs and Model Importances

Excel sheet with a list of the number of model inputs and tabs with input OMOP concept for each model. 'Model Inputs' tabs lists the number of input features per model, and a description of demographic or visit-related features. 'Model Importance' tab lists all trained random forest model importance, full and matched, and sex-stratified models, and the mapped phecodes. 'Top Feature Prevalences' tab shows the prevalence of some of the top conditions utilized in prediction. The rest of the excel sheets lists all model inputs and associated OMOP concept_ids.
The data can be downloaded at www.synapse.org/AD_EHR_Prediction or Synapse repository ID syn52816091.

3.12 References

1. 2022 Alzheimer's disease facts and figures. *Alzheimers Dement.* **18**, 700–789 (2022).
2. Rasmussen, J. & Langerman, H. Alzheimer's Disease – Why We Need Early Diagnosis. *Degener. Neurol. Neuromuscul. Dis.* **Volume 9**, 123–130 (2019).
3. Kivipelto, M. Midlife vascular risk factors and Alzheimer's disease in later life: longitudinal, population based study. *BMJ* **322**, 1447–1451 (2001).
4. Niculescu, A. B. *et al.* Blood biomarkers for memory: toward early detection of risk for Alzheimer disease, pharmacogenomics, and repurposed drugs. *Mol. Psychiatry* **25**, 1651–1672 (2020).
5. Alena V. Savonenko, Philip C. Wong, & Tong Li. Alzheimer diseases. (2023) doi:10.1016/b978-0-323-85654-6.00022-8.
6. Neugroschl, J. & Wang, S. Alzheimer's Disease: Diagnosis and Treatment Across the Spectrum of Disease Severity. *Mt. Sinai J. Med. N. Y.* **78**, 596–612 (2011).
7. Tang, A. S. *et al.* Deep phenotyping of Alzheimer's disease leveraging electronic medical records identifies sex-specific clinical associations. *Nat. Commun.* **13**, 675 (2022).
8. Taubes, A. *et al.* Experimental and real-world evidence supporting the computational repurposing of bumetanide for APOE4-related Alzheimer's disease. *Nat. Aging* **1**, 932–947 (2021).
9. Ben Miled, Z. *et al.* Predicting dementia with routine care EMR data. *Artif. Intell. Med.* **102**, 101771 (2020).
10. Tang, A., Woldemariam, S., Roger, J. & Sirota, M. Translational Bioinformatics to Enable Precision Medicine for All: Elevating Equity across Molecular, Clinical, and Digital Realms. *Yearb. Med. Inform.* **31**, 106–115 (2022).

11. Xu, J. *et al.* Data-driven discovery of probable Alzheimer's disease and related dementia subphenotypes using electronic health records. *Learn. Health Syst.* **4**, e10246 (2020).
12. Park, J. H. *et al.* Machine learning prediction of incidence of Alzheimer's disease using large-scale administrative health data. *Npj Digit. Med.* **3**, 46 (2020).
13. Qiu, S. *et al.* Multimodal deep learning for Alzheimer's disease dementia assessment. *Nat. Commun.* **13**, 3404 (2022).
14. Diogo, V. S., Ferreira, H. A., Prata, D., & for the Alzheimer's Disease Neuroimaging Initiative. Early diagnosis of Alzheimer's disease using machine learning: a multi-diagnostic, generalizable approach. *Alzheimers Res. Ther.* **14**, 107 (2022).
15. Ding, Y. *et al.* A Deep Learning Model to Predict a Diagnosis of Alzheimer Disease by Using ^{18}F-FDG PET of the Brain. *Radiology* **290**, 456–464 (2019).
16. Popuri, K., Ma, D., Wang, L. & Beg, M. F. Using machine learning to quantify structural MRI neurodegeneration patterns of Alzheimer's disease into dementia score: Independent validation on 8,834 images from ADNI, AIBL, OASIS, and MIRIAD databases. *Hum. Brain Mapp.* **41**, 4127–4147 (2020).
17. Chang, C.-H., Lin, C.-H. & Lane, H.-Y. Machine Learning and Novel Biomarkers for the Diagnosis of Alzheimer's Disease. *Int. J. Mol. Sci.* **22**, 2761 (2021).
18. Stamate, D. *et al.* A metabolite-based machine learning approach to diagnose Alzheimer-type dementia in blood: Results from the European Medical Information Framework for Alzheimer disease biomarker discovery cohort. *Alzheimers Dement. Transl. Res. Clin. Interv.* **5**, 933–938 (2019).

19. Dubal, D. B. Chapter 16 - Sex difference in Alzheimer's disease: An updated, balanced and emerging perspective on differing vulnerabilities. in *Handbook of Clinical Neurology* (eds. Lanzenberger, R., Kranz, G. S. & Savic, I.) vol. 175 261–273 (Elsevier, 2020).

20. Hampel, H. *et al.* Precision medicine and drug development in Alzheimer's disease: the importance of sexual dimorphism and patient stratification. *Front. Neuroendocrinol.* **50**, 31–51 (2018).

21. Nelson, C. A., Bove, R., Butte, A. J. & Baranzini, S. E. Embedding electronic health records onto a knowledge network recognizes prodromal features of multiple sclerosis and predicts diagnosis. *J. Am. Med. Inform. Assoc.* **29**, 424–434 (2022).

22. Belonwu, S. A. *et al.* Sex-Stratified Single-Cell RNA-Seq Analysis Identifies Sex-Specific and Cell Type-Specific Transcriptional Responses in Alzheimer's Disease Across Two Brain Regions. *Mol. Neurobiol.* (2021) doi:10.1007/s12035-021-02591-8.

23. Carlos A. Saura, Angel Deprada, Maria Dolores Capilla-López, & Arnaldo Parra-Damas. Revealing cell vulnerability in Alzheimer's disease by single-cell transcriptomics. *Semin. Cell Dev. Biol.* (2022) doi:10.1016/j.semcdb.2022.05.007.

24. Leonenko, G. *et al.* Polygenic risk and hazard scores for Alzheimer's disease prediction. *Ann. Clin. Transl. Neurol.* **6**, 456–465 (2019).

25. Alzheimer's Disease Neuroimaging Initiative *et al.* Multimodal Phenotyping of Alzheimer's Disease with Longitudinal Magnetic Resonance Imaging and Cognitive Function Data. *Sci. Rep.* **10**, 5527 (2020).

26. Himmelstein, D. S. *et al.* Systematic integration of biomedical knowledge prioritizes drugs for repurposing. *eLife* **6**, e26726 (2017).

27. Nelson, C. A., Butte, A. J. & Baranzini, S. E. Integrating biomedical research and electronic health records to create knowledge-based biologically meaningful machine-readable embeddings. *Nat. Commun.* **10**, 3045 (2019).

28. Morris, J. H. *et al.* The scalable precision medicine open knowledge engine (SPOKE): a massive knowledge graph of biomedical information. *Bioinformatics* **39**, btad080 (2023).

29. Bastarache, L. Using Phecodes for Research with the Electronic Health Record: From PheWAS to PheRS. *Annu. Rev. Biomed. Data Sci.* **4**, 1–19 (2021).

30. Schwartzentruber, J. *et al.* Genome-wide meta-analysis, fine-mapping and integrative prioritization implicate new Alzheimer's disease risk genes. *Nat. Genet.* **53**, 392–402 (2021).

31. Global Lipids Genetics Consortium. Discovery and refinement of loci associated with lipid levels. *Nat. Genet.* **45**, 1274–1283 (2013).

32. Morris, J. A. *et al.* An atlas of genetic influences on osteoporosis in humans and mice. *Nat. Genet.* **51**, 258–266 (2019).

33. Jansen, W. J. *et al.* Association of Cerebral Amyloid-beta Aggregation With Cognitive Functioning in Persons Without Dementia. *JAMA Psychiatry* **75**, 84 (2018).

34. Yagis, E. *et al.* Effect of data leakage in brain MRI classification using 2D convolutional neural networks. *Sci. Rep.* **11**, 22544 (2021).

35. You, J. *et al.* Development of a novel dementia risk prediction model in the general population: A large, longitudinal, population-based machine-learning study. *eClinicalMedicine* **53**, 101665 (2022).

36. Littlejohns, T. J. *et al.* Vitamin D and the risk of dementia and Alzheimer disease. *Neurology* **83**, 920–928 (2014).

37. Elbejjani, M. *et al.* Depression, depressive symptoms, and rate of hippocampal atrophy in a longitudinal cohort of older men and women. *Psychol. Med.* **45**, 1931–1944 (2015).

38. Goveas, J. S., Espeland, M. A., Woods, N. F., Wassertheil-Smoller, S. & Kotchen, J. M. Depressive Symptoms and Incidence of Mild Cognitive Impairment and Probable Dementia in Elderly Women: The Women's Health Initiative Memory Study: DEPRESSION AND INCIDENT MCI AND DEMENTIA. *J. Am. Geriatr. Soc.* **59**, 57–66 (2011).

39. Swerdlow, R. H. Is aging part of Alzheimer's disease, or is Alzheimer's disease part of aging? *Neurobiol. Aging* **28**, 1465–1480 (2007).

40. Kosyreva, A. M., Sentyabreva, A. V., Tsvetkov, I. S. & Makarova, O. V. Alzheimer's Disease and Inflammaging. *Brain Sci.* **12**, 1237 (2022).

41. Wallace, L. M. K. *et al.* Investigation of frailty as a moderator of the relationship between neuropathology and dementia in Alzheimer's disease: a cross-sectional analysis of data from the Rush Memory and Aging Project. *Lancet Neurol.* **18**, 177–184 (2019).

42. Kojima, G., Taniguchi, Y., Iliffe, S. & Walters, K. Frailty as a Predictor of Alzheimer Disease, Vascular Dementia, and All Dementia Among Community-Dwelling Older People: A Systematic Review and Meta-Analysis. *J. Am. Med. Dir. Assoc.* **17**, 881–888 (2016).

43. Wallace, L., Theou, O., Rockwood, K. & Andrew, M. K. Relationship between frailty and Alzheimer's disease biomarkers: A scoping review. *Alzheimers Dement. Diagn. Assess. Dis. Monit.* **10**, 394–401 (2018).

44. Barnes, L. L. *et al.* Sex Differences in the Clinical Manifestations of Alzheimer Disease Pathology. *Arch. Gen. Psychiatry* **62**, 685 (2005).

45. Davis, E. J. *et al.* Sex-Specific Association of the X Chromosome With Cognitive Change and Tau Pathology in Aging and Alzheimer Disease. *JAMA Neurol.* (2021) doi:10.1001/jamaneurol.2021.2806.

46. Campion, D. *et al.* Early-onset autosomal dominant Alzheimer disease: prevalence, genetic heterogeneity, and mutation spectrum. *Am. J. Hum. Genet.* **65**, 664–670 (1999).

47. Liew, T. M. Subjective cognitive decline, APOE e4 allele, and the risk of neurocognitive disorders: Age- and sex-stratified cohort study. *Aust. N. Z. J. Psychiatry* (2022) doi:10.1177/00048674221079217.

48. He, Z. *et al.* Genome-wide analysis of common and rare variants via multiple knockoffs at biobank scale, with an application to Alzheimer disease genetics. *Am. J. Hum. Genet.* (2021) doi:10.1016/j.ajhg.2021.10.009.

49. Nandar, W. & Connor, J. R. HFE Gene Variants Affect Iron in the Brain1–3. *J. Nutr.* **141**, S729–S739 (2011).

50. Wang, Z. *et al.* Deep post-GWAS analysis identifies potential risk genes and risk variants for Alzheimer's disease, providing new insights into its disease mechanisms. *Sci. Rep.* **11**, 20511 (2021).

51. Iivonen, S. *et al.* Heparan sulfate proteoglycan 2 polymorphism in Alzheimer's disease and correlation with neuropathology. *Neurosci. Lett.* **352**, 146–150 (2003).

52. Talwar, P. *et al.* Genomic convergence and network analysis approach to identify candidate genes in Alzheimer's disease. *BMC Genomics* **15**, 199 (2014).

53. Talwar, P. *et al.* Validating a Genomic Convergence and Network Analysis Approach Using Association Analysis of Identified Candidate Genes in Alzheimer's Disease. *Front. Genet.* **12**, 722221 (2021).

54. Zhu, M. *et al.* Mutations in the γ-Actin Gene (ACTG1) Are Associated with Dominant Progressive Deafness (DFNA20/26). *Am. J. Hum. Genet.* **73**, 1082–1091 (2003).

55. Vasilopoulos, Y., Gkretsi, V., Armaka, M., Aidinis, V. & Kollias, G. Actin cytoskeleton dynamics linked to synovial fibroblast activation as a novel pathogenic principle in TNF-driven arthritis. *Ann. Rheum. Dis.* **66**, iii23–iii28 (2007).

56. Lee, W.-C., Guntur, A. R., Long, F. & Rosen, C. J. Energy Metabolism of the Osteoblast: Implications for Osteoporosis. *Endocr. Rev.* **38**, 255–266 (2017).

57. Wang, F., Han, L. & Hu, D. Fasting insulin, insulin resistance and risk of hypertension in the general population: A meta-analysis. *Clin. Chim. Acta Int. J. Clin. Chem.* **464**, 57–63 (2017).

58. James, D. E., Stöckli, J. & Birnbaum, M. J. The aetiology and molecular landscape of insulin resistance. *Nat. Rev. Mol. Cell Biol.* **22**, 751–771 (2021).

59. Schrijvers, E. M. C. *et al.* Insulin metabolism and the risk of Alzheimer disease: The Rotterdam Study. *Neurology* **75**, 1982–1987 (2010).

60. Ferreira, L. S. S., Fernandes, C. S., Vieira, M. N. N. & De Felice, F. G. Insulin Resistance in Alzheimer's Disease. *Front. Neurosci.* **12**, 830 (2018).

61. Rahman, S. O. *et al.* Association between insulin and Nrf2 signalling pathway in Alzheimer's disease: A molecular landscape. *Life Sci.* **328**, 121899 (2023).

62. Ataie-Ashtiani, S. & Forbes, B. A Review of the Biosynthesis and Structural Implications of Insulin Gene Mutations Linked to Human Disease. *Cells* **12**, 1008 (2023).

63. Gillespie, M. *et al.* The reactome pathway knowledgebase 2022. *Nucleic Acids Res.* **50**, D687–D692 (2022).

64. Bowman, G. L., Kaye, J. A. & Quinn, J. F. Dyslipidemia and Blood-Brain Barrier Integrity in Alzheimer's Disease. *Curr. Gerontol. Geriatr. Res.* **2012**, 1–5 (2012).

65. Reitz, C. Dyslipidemia and the risk of Alzheimer's disease. *Curr. Atheroscler. Rep.* **15**, 307 (2013).

66. Goldstein, F. C. *et al.* Effects of hypertension and hypercholesterolemia on cognitive functioning in patients with alzheimer disease. *Alzheimer Dis. Assoc. Disord.* **22**, 336–342 (2008).

67. Sáiz-Vazquez, O., Puente-Martínez, A., Ubillos-Landa, S., Pacheco-Bonrostro, J. & Santabárbara, J. Cholesterol and Alzheimer's Disease Risk: A Meta-Meta-Analysis. *Brain Sci.* **10**, 386 (2020).

68. Bertram, L. & Tanzi, R. E. Genome-wide association studies in Alzheimer's disease. *Hum. Mol. Genet.* **18**, R137–R145 (2009).

69. Corder, E. H. *et al.* Gene Dose of Apolipoprotein E Type 4 Allele and the Risk of Alzheimer's Disease in Late Onset Families. *Science* **261**, 921–923 (1993).

70. Garcia, A. R. *et al.* APOE4 is associated with elevated blood lipids and lower levels of innate immune biomarkers in a tropical Amerindian subsistence population. *eLife* **10**, e68231 (2021).

71. Mahley, R. W. & Rall, S. C. Apolipoprotein E: Far More Than a Lipid Transport Protein. *Annu. Rev. Genomics Hum. Genet.* **1**, 507–537 (2000).

72. Kimura, R. *et al.* Albumin gene encoding free fatty acid and β-amyloid transporter is genetically associated with Alzheimer disease: Albumin gene and Alzheimer's disease. *Psychiatry Clin. Neurosci.* **60**, S34–S39 (2006).

73. Lv, X.-L. *et al.* Association between Osteoporosis, Bone Mineral Density Levels and Alzheimer's Disease: A Systematic Review and Meta-analysis. *Int. J. Gerontol.* **12**, 76–83 (2018).

74. Amouzougan, A. *et al.* High prevalence of dementia in women with osteoporosis. *Joint Bone Spine* **84**, 611–614 (2017).

75. Liu, Y., Jin, G., Wang, X., Dong, Y. & Ding, F. Identification of New Genes and Loci Associated With Bone Mineral Density Based on Mendelian Randomization. *Front. Genet.* **12**, 728563 (2021).

76. Fan, C. C. *et al.* Sex-dependent autosomal effects on clinical progression of Alzheimer's disease. *Brain* **143**, 2272–2280 (2020).

77. Deming, Y. *et al.* The MS4A gene cluster is a key modulator of soluble TREM2 and Alzheimer's disease risk. *Sci. Transl. Med.* **11**, eaau2291 (2019).

78. Chen, Y.-H. & Lo, R. Y. Alzheimer's disease and osteoporosis. *Ci Ji Yi Xue Za Zhi Tzu-Chi Med. J.* **29**, 138–142 (2017).

79. Li, S., Liu, B., Zhang, L. & Rong, L. Amyloid beta peptide is elevated in osteoporotic bone tissues and enhances osteoclast function. *Bone* **61**, 164–175 (2014).

80. Gale, S. A. *et al.* Preclinical Alzheimer Disease and the Electronic Health Record: Balancing Confidentiality and Care. *Neurology* **99**, 987–994 (2022).

81. Serrano-Pozo, A. *et al.* Mild to moderate Alzheimer dementia with insufficient neuropathological changes. *Ann. Neurol.* **75**, 597–601 (2014).

82. Nelson, P. T. *et al.* Alzheimer's disease is not 'brain aging': neuropathological, genetic, and epidemiological human studies. *Acta Neuropathol. (Berl.)* **121**, 571–587 (2011).

83. Jack, C. R. *et al.* NIA-AA Research Framework: Toward a biological definition of Alzheimer's disease. *Alzheimers Dement. J. Alzheimers Assoc.* **14**, 535–562 (2018).

84. Data Equity Taskforce sponsored by the Health Equity Council at UCSF Health. *UCSF Health's equity-related variables user's guide.* (2021).

85. Austin, P. C. An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies. *Multivar. Behav. Res.* **46**, 399–424 (2011).

86. Karlin, L. *et al.* Use of the Propensity Score Matching Method to Reduce Recruitment Bias in Observational Studies: Application to the Estimation of Survival Benefit of Non-Myeloablative Allogeneic Transplantation In Patients with Multiple Myeloma Relapsing after a First Autologous Transplantation. *Blood* **112**, 1133–1133 (2008).

87. Tipton, E. *et al.* Sample Selection in Randomized Experiments: A New Method Using Propensity Score Stratified Sampling. *J. Res. Educ. Eff.* **7**, 114–135 (2014).

88. Bingenheimer, J. B., Brennan, R. T. & Earls, F. J. Firearm violence exposure and serious violent behavior. *Science* **308**, 1323–1326 (2005).

89. Xia, Y. *et al.* Association between dietary patterns and metabolic syndrome in Chinese adults: a propensity score-matched case-control study. *Sci. Rep.* **6**, 34748 (2016).

90. Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* (2012) doi:10.48550/ARXIV.1201.0490.

91. scikit-learn developers. Scikit-Learn Documentation: Random Forest Parameters. https://scikit-learn.org/stable/modules/ensemble.html#random-forest-parameters.

92. Breiman, L. Random Forests. *Mach. Learn.* **45**, 5–32 (2001).

93. Azodi, C. B., Tang, J. & Shiu, S.-H. Opening the Black Box: Interpretable Machine Learning for Geneticists. *Trends Genet.* **36**, 442–455 (2020).

94. Wu, P. *et al.* Mapping ICD-10 and ICD-10-CM Codes to Phecodes: Workflow Development and Initial Evaluation. *JMIR Med. Inform.* **7**, e14325 (2019).

95. Assenov, Y., Ramírez, F., Schelhorn, S.-E., Lengauer, T. & Albrecht, M. Computing topological parameters of biological networks. *Bioinformatics* **24**, 282–284 (2008).

96. Ghoussaini, M. *et al.* Open Targets Genetics: systematic identification of trait-associated genes using large-scale genetics and functional genomics. *Nucleic Acids Res.* **49**, D1311–D1320 (2021).

97. Mountjoy, E. *et al.* An open approach to systematically prioritize causal variants and genes at all published human GWAS trait-associated loci. *Nat. Genet.* **53**, 1527–1533 (2021).

98. Sun, B. B. *et al.* Genomic atlas of the human plasma proteome. *Nature* **558**, 73–79 (2018).

99. Suhre, K. *et al.* Connecting genetic risk to disease end points through the human blood plasma proteome. *Nat. Commun.* **8**, 14357 (2017).

100. Neale Lab. UK Biobank GWAS Round 2. http://www.nealelab.is/uk-biobank/.

101. Giambartolomei, C. *et al.* Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. *PLoS Genet.* **10**, e1004383 (2014).

102. Võsa, U. *et al.* Large-scale cis- and trans-eQTL analyses identify thousands of genetic loci and polygenic scores that regulate blood gene expression. *Nat. Genet.* **53**, 1300–1310 (2021).

Chapter 4: Learnings and Considerations in Designing, Implementing and Interpreting Electronic Medical Record-based Informatics Studies

4.1 Abstract

With the increasing availability of rich longitudinal real world clinical data across millions of patients recorded in electronic medical records (EMR), there is a growing interest in leveraging these records for scientific questions and applications to improve understanding and treatment of human health and disease. While EMR datasets provide great opportunity for deriving insights into disease and treatments, there is also a need to carefully consider limitations due to biases from clinical diagnostic or treatment behaviors and missing information from data collection. These limitations can pose great challenges in EMR-based informatics studies that can lead to confusing or incorrect conclusions. Here, we discuss the potential applications of EMR datasets as well as considerations in the design, implementation, and interpretation of EMR-based informatics studies and draw from examples in the literature across hypothesis generation and hypothesis-driven studies. This paper aims to provide a perspective and starting point for researchers and engineers tackling EMR-based clinical informatics studies.

4.2 Introduction

Since the beginning of the practice of medicine, record-keeping has been an important aspect of clinical care[1]. These records include relevant health history of a patient, such as diagnoses, procedures, treatments, and monetary exchanges, across multiple care delivery sites. The FDA now defines these clinical records as part of real world data (RWD)[2], which spans across sources like health records from hospitals and clinics, databases of medications sold at pharmacies, registry and public health databases, insurance claims, and data acquired from sensors and mobile devices that may inform health status for an individual or a population. With the advancements of computer technologies, databases and records have transitioned from manual physical documentation to electronic and automated forms, including the use of electronic medical records (EMR) in care site settings. In 2009 President Obama prioritized and financially incentivized the transition to digital records and implementation of software for EMRs in hospitals[3,4]. This adoption has been useful to the medicine workflow by decreasing medication errors and improving billing, with various impacts on healthcare delivery quality and cost effectiveness[5–7].

EMRs are now mainstream in the healthcare setting, with over 75% of office-based practices and over 90% of hospitals with EMRs adopted and utilized[8]. With longitudinal health data on millions of patients collected, and billions of invested dollars over a decade of this big data collection and data storage effort, there is much recent rising interest in leveraging real world datasets, especially EMRs, for research applications within biological and healthcare spaces[9,10]. These datasets are now available at many institutions and healthcare systems within the US and worldwide[11,12], and some are publicly accessible for research, such as MIMIC[13], AllofUs[14], and UK Biobank[15].

There are many exciting opportunities for EMR datasets to give insights into disease phenotyping, characterize treatment pathways and outcomes for drug approvals and drug repurposing, disentangle disease heterogeneity, and even advance the understanding of disease biology. Some of the benefits of research with EMRs include flexibility in defining inclusions and exclusions for cohort selection, unlike many existing observational studies that apply extensive screening criteria and thereby not capturing the complexity among real-world contexts. The flexibility in cohort selection and individual representation among a large sample size further increases the potential for utilizing these datasets to investigate diverse cohorts or otherwise under-studied populations and allowing an opportunity to answer questions on populations with otherwise sparse or unavailable data. Therefore, by utilizing EMRs and other real-world datasets, resulting analyses may better capture real world populations or real world measures of disease and outcomes.

Nevertheless, despite the exponential increase in the interest in RWD and EMR datasets for advancing human health, there are also many drawbacks, biases, and considerations for utilizing and interpreting EMR data in research. Since the original purpose of EMR was not for research, but instead for billing and record-keeping, many drawbacks arise from biases in the data collection, data representation, and data preprocessing pathway. For example, some diseases may exist in a patient, but are not measured or recorded. Timing associated with a chronic disease like diabetes may represent the acknowledgement of a disease instead of biological onset. These considerations are essential to account for in designing methods and approaches for EMR utilization and interpretation of results, particularly if the insights will be utilized to inform treatment and care that can impact patient lives. Mitigating these challenges entails understanding the data collection pipeline and employing data pre-processing methods, such as models to account

for data missingness, disease encoding and representation across sites, and factoring in the influence of social or environmental exposures. Collaborating with practitioners can provide insight to account for these factors that may impact data collection in clinical practice, such as the order of medications given in a treatment pathway or insurance reimbursement incentives. Considerations for timing between biological disease and timing of EMR records may also be acknowledged during the process of model selection and statistical approaches.

In this perspective chapter, we will first describe the pathway of EMR data collection and opportunities and examples for research use. We provide examples of leveraging clinical data for phenotyping, hypothesis generation, and specific hypothesis-driven studies in the context of disease diagnostics and therapeutics. We also provide considerations for researchers tackling EMR-based research and discuss current advances impacting the field and conclude with ongoing advancements in the EMR research field.

4.3 Data Collection to Data Insights

Before clinical datasets can be utilized for answering questions related to disease biology, many decisions impact the workflow from data collection, normalization, preprocessing, simplification, and de-identification. Understanding this data flow is essential for making model decisions and evaluating insights based on biases that may be introduced in the data flow pathway.

Data Collection and Representation. To understand the possibilities and limitations of how EMR data can be utilized to answer questions in health, it is important to recognize the process of EMR data collection. In turn, this will guide the process of hypothesis development, method selection, and interpretation of results. Heterogeneity in data collection may arise as a result of varying clinical practices between providers, insurance coverages, patient population, location, existence of scribes, and even the EMR software and database storage approach for a healthcare

setting[16]. For example, when a clinical visit is scheduled, there may be diagnoses given, medications prescribed, diagnostic tests ordered, and/or procedures delivered. These may be recorded with codes (e.g., ICD (International Classification of Diseases) for diagnoses; RxNorm and NDC (National Drug Code) for medications; LOINC (Logical Observation Identifiers Names and Codes) for laboratory tests; CPT (Current Procedural Terminology) for procedures; and SNOMED CT (Systematized Nomenclature of Medicine - Clinical Terms) which includes codes for diagnoses, symptoms, medications, procedures to indicate a service given and ensure billing or payment for the personnel, medication, and equipment[16]. The choice of a coding system may differ between clinical sites and may not encompass the same underlying biological phenomenon due to differences in clinical workflows or treatment pathways. Furthermore, there is no fully standardized way for providers to adhere to a terminology when assigning diagnosis (e.g., two providers may assign different ICD codes for the same condition).

Care sites in the United States may also choose to utilize different EMR softwares like Epic, Cerner, Athena, OpenEMR, etc., which may result in differences in coding approaches and database structure (e.g., EMR software may come with their own diagnostic codes instead of using ICD codes). Other considerations between care sites include that a few community clinics may still primarily utilize paper records, while other sites may only have recently adopted EMR systems due to cost and administrative concerns[17], which may lead to different data starting times or differences in captured intervals of care. Without standardization, each record system may have heterogeneous ways of storing and representing health data. In other countries, the coding and database structure may be consistent in a centralized healthcare setting, but variations still exist between national borders[18,19]. These differences therefore contribute to heterogenous starting points in data storage and representation.

In order for EMR data to be utilized for research, they need to exist in a form that can be leveraged for analysis. After data collection, data extraction can take the form of manual labor including populating tables from EMR chart review to database queries and machine learning-based concept extraction. In all cases, extraction takes the form of converting raw data representations encoded by the EMR software or user interface into standardized representations of data, for example through a Clinical Data Warehouse (CDW) or the Observational Medical Outcomes Partnership (OMOP) Common Data Model[20].

Data Preparation and Standardization. Data preprocessing into a database structure requires personnel such as database managers and infrastructure engineers to be involved in the decision-making process of mapping raw data to usable formats and representations and identifying the appropriate hardware and software required for future data manipulation. Furthermore, in order to ensure the data can be utilized for research purposes, de-identification methods are performed on the preprocessing step to ensure proper privacy is maintained[13,21]. Each one of these decisions to transform the originally recorded data into a format that maintains relevant information influences model selection and interpretation of results downstream for the researcher.

In 2007, the FDA in collaboration with industry and academia introduced OMOP[20,22], which has since been expanded to the ODHSI (Observational Health Data Sciences and Informatics) suite of tools and applications to enable clinical evidence to be used for research purposes worldwide[23]. This structure and paradigm have the additional benefit of utilizing a standard vocabulary and structure that can enable data representation and code sharing between sites. While the goals of standardization has enabled greater data harmonization, transportability, and federated analyses, there are still site-based biases and limitations that may result from

individual site decisions for treatment and mappings between the source EMR data structure to standardized structures and SNOMED CT ontology, resulting in loss of specificity from a diagnostic code to a broader category in SNOMED and many-to-many mappings that may lead to further loss or addition of clinical codes.

Another pre-processing step important in increasing the availability of EMR data for researchers is the process of de-identification[13,24]. Methods and approaches to achieve de-identification of EMR data must be done in accordance with the Health Insurance Portability and Accountability Act (HIPAA) of 1996, including the Safe Harbor Method and the Expert Determination Method[25]. The de-identification process may contribute to new biases in the data due to removal of patients or loss of accuracy from shifting of dates and exclusion of some information such as ages over 89 years, which may especially impact the records of vulnerable populations and the elderly[26,27]. Text information may require further automated means of de-identification that can introduce biases from perceived lower relevance of concepts that may come from removal of information as opposed to lack of association[28]. Every step in data collection, data representation, and data preprocessing is therefore important to consider for careful use of EMR data for research purposes (**Figure 4.1**).

4.4 EMR For Hypothesis Generation

In leveraging EMR data, one of the greatest opportunities includes improved understanding of the magnitude and implications of disease burdens, as well patterns of healthcare service utilization within specific demographics or locations. This can allow for an approach to characterize the impact of a disease without any influences from prior beliefs or inferences from potentially non-representative studies. Moreover, a thorough phenotyping can provide a relatively unbiased view of the EMR dataset for investigating comorbid conditions and formulate hypotheses

pertaining to underlying trends. It is, however, crucial to recognize that EMR data captures only discrete intervals of an individual's health trajectory. Therefore, understanding the limitations and representation of the EMR dataset becomes imperative in determining the breadth of measurable data and the feasibility of specific research goals.

Phenotyping of Disease, Drugs, and Clinical Behaviors. EMR data provide a great way to characterize and phenotype many aspects of health, including gaining a broad understanding of the prevalence of diseases, drugs, and higher order combinations of comorbid disorders or concomitant medications within a disease[29–33]. Many researchers view real world data and EMR data as a data source for exploratory research and scientific inquiry pipeline. Understanding and characterizing patterns in the data can also give insight into understanding disease populations, practices for treatment prescription and current clinical behaviors[34–36], as well as identify potential differences in subgroups and disparities in current care[37–39]. Data exploration approaches can also better quantify any issues in data quality, such as the extent of data missingness, as well as understand potential provider or patient behaviors that may contribute to variations in data quality. Since clinical datasets may contain potential misclassifications or lack of specificity, exploratory analysis of the data can aid in the development of phenotyping algorithms and cohort selection approaches that may best represent the biological characteristics of a desired population (e.g., increasing confidence in Alzheimer's Disease diagnosis with more diagnostic occurrences or by identifying a neurologist's diagnosis as trustworthy)[40]. Augmentation of EMR datasets with molecular datasets, such as genomics data or imaging datasets, can aid in biological support of cohort selection approaches and phenotyping[41].

Once a researcher gains a deeper understanding of the characteristics and patterns in the dataset, they can leverage EMR to perform deep phenotyping of a disease for improved

characterization of associated comorbid conditions, medications, and lab results that can give unbiased insight into disease associations through characterization of multiple patient disease profiles. For example, large-scale association analyses in Alzheimer's Disease (AD) identified sex-specific and racialized population specific comorbidities that may highlight heterogeneous differences in the clinical profile, risk or manifestation of AD, enabling hypotheses for downstream analyses[31,42]. To investigate associations, choosing a cohort, comparison cohort, and statistical approach are needed. In the above AD association studies, this took the form of a cross-sectional case-control study between AD and matched controls, where matching (i.e. propensity-score matching) was utilized to account for associations that may not be useful for disease interpretation. Many approaches exist for matching[43–45], with each approach having its benefits and limitations. Ultimately, the covariates utilized in matching or adjustment should be considered when contextualizing the high-level understanding of associations in the data set for hypothesis generation (for example, whether an association is independent of age or not). Low-dimensional embeddings can also aid in visualizing clusters of pathways or differences in patient populations due to the observed clinical profile[29,30,46].

Hypothesis Generation. Hypothesis generation approaches can take many forms depending on available data and give insight into behaviors, biological underpinnings of disease, risk factors, or drug outcomes. From a cross-sectional view of disease, broad associations can aid in exploratory research through deep phenotyping of disease[31]. Medication and procedure associations can be explored to understand patient populations and generate hypotheses that may impact these populations[47,48]. With availability of genetic data, genetic association with phenotypes can also be explored to elucidate biological hypotheses (e.g., PheWAS[41,49]). Beyond genetics, other datasets that can be mapped to patients in the EMRs can also help augment hypothesis generation and

elucidate potential biological associations with clinical phenotypes (e.g, the Rheumatology Informatics System for Effectiveness database[50]). Excitingly, augmentation of clinical phenotypes with known disease-molecular associations from heterogenous knowledge databases are evolving as one potential opportunity for biological hypothesis prioritization[51,52].

Hypothesis generation approaches can also include temporal filters in the analyses to derive insights into the temporal nature of an association, based on time to an index event or index date. For example, one study identified associations prior to pregnancy in order to identify hypotheses related to mechanisms of recurrent pregnancy loss, such as menstrual abnormalities and infertility diagnoses[53]. Beyond hypothesis generation, if the dataset and features are utilized for machine learning model training, then understanding associations can provide explainability in model performance based on available associations. With an available index time, careful interpretation is needed to understand what the associated time represents in order to interpret the data and define what hypotheses are possible (e.g., whether a time represents the onset of a disease, or the clinical acknowledgement of a disease). If the time represents a clinical acknowledgement of a health event, then any associations should be considered as a means to interpret existing clinical behaviors with potential utility in identifying red herrings and clinical biases in either diagnostic, decision-making, or inequities in care[42,54].

Exploratory approaches to EMR data can provide unbiased insight into disease phenotypes, trajectories, subphenotypes, and associations with any other available dataset (measurements, genetics, imaging, biomarkers), improving understanding and generating hypotheses that can advance scientific inquiry into health (**Figure 4.2**).

4.5 EMR For Hypothesis-Driven Studies

Electronic medical records also provide an opportunity to extend traditional epidemiological clinical research and biostatistics into large patient settings, although modifications may need to be made to account for the retrospective and nature of available data. As mentioned in the previous sections, it is imperative to understand the data through exploratory approaches and ensure a hypothesized question can be answered. In particular, biases from data missingness or clinical behaviors should be considered before publicizing conclusions, which may impact regulatory decision making, clinical care, or even patient health behaviors.

Identifying a question or hypothesis. The first step for a hypothesis-driven study includes identifying a hypothesis, including expanding on findings from prior literature or identifying a cohort of interest that can answer the hypothesis as robustly as possible. Time is also an important consideration in the choice of a cohort, whether the time is relative to an index date or based on a year or societal event. The choice of a cohort may also need to include adjustments or acknowledgements of potential biases or quality control steps in retrospective real world datasets in order to understand limitations in cohort specificity. Ultimately, both the existence of a cohort and relevant data for a hypothesis will impact the answerability of a question and choice of analysis approach. While there is extensive literature in the epidemiological fields on classifying studies that are applied to retrospective data, some considerations are currently in development for the advancement of methods applicable to EMR data that address issues such as data missingness, biases, and data noise[55–58]. A portion of studies currently published include case-control, cross-sectional, and cohort study designs, as well as other designs not mentioned here.

Case-control design. Case-control designs encompass not only the selection of a target cohort but also a control or comparison cohort. In numerous studies, the selection of controls often

involves adjusting for confounders, either in the choice of the control or within the analytical model itself[59]. Presently, there is increased emphasis on control selection, including the utilization of matching methods, propensity score models, or regression to account for multiple covariates or identify controls from external databases[55,60–62]. Since certain diseases may manifest concurrently and single datasets may have sample size limitations, some studies are expanding to include multiple cohorts or controls of various complexities[56,62–64].

The choice of a control or comparator cohort influences the interpretation of analyses, especially in scenarios where specific predefined covariates have been accounted for. While unadjusted analyses remain useful in discerning overarching patterns, adjusted or matched analyses can better pinpoint trends that persist even after controlling for other potential confounders. Since case-control and case-comparator designs are inherently retrospective, causal conclusions cannot be made. This caveat holds true analogously for machine learning models designed with a similar approach in the selection of the training cohort.

Within cohort selection, temporal assignments to data are important in determining whether the analysis is completely retrospective or cross-sectional, and whether a specific date of an acute event is relevant for a cohort. Given a specific hypothesis of interest, general prevalence and associations with outcomes can be identified (e.g., exposure to drug and death), but a consideration of the temporal relationships is needed to identify the degree of support for causal conclusions (e.g., the time assigned to chronic diseases like hypertension is not representative of biological onset). As an example, a cross-sectional study validating bumetanide as a drug repurposing candidate for Alzheimer's Disease explored the association between bumetanide use and Alzheimer's Disease prevalence[65]. The real-world evidence in this study supports a lower prevalence of Alzheimer's Disease in those with bumetanide exposure, but causal conclusions

about the drug's effect could not be made without the support of experimental evidence in mice, which was also included in the study.

Retrospective Cohort Study Design. Another approach to leveraging EMR for hypothesis-driven studies includes the use of a retrospective cohort study design or survival style design. Questions may investigate the relationship between drug exposure or risk factor exposure and outcomes or prognosis of a disease and labeled with terms such as 'emulated trials'[66]. These designs are best approached with consideration of what a cohort study may look like, and selection of the cohort must be done with information up to a certain point in the EMR. This study design was utilized to investigate statin and antidepressant usage in COVID-19 mortality[67,68] as well as antidepressant use in pregnancy outcomes[69]. With this design, a temporal association can be determined to understand disease progression, drug exposure outcomes, or risk factor associations. It is nevertheless important to consider that cohort study design is still retrospective in nature and while causal inference techniques can help in the interpretation of an association, there are still large issues in terms of large attrition rates and missing data. For example, studies have identified antidepressant use associated with preterm birth as an adverse outcome, with possible confounding by depression severity and medication adherence[69].

Machine learning and biases. These fundamental designs and considerations also apply to sophisticated machine learning models. The selection of a cohort, temporal covariates of inclusion, impact of biases, and hypothesis of interest all play into the interpretation of a model's performance and the causal nature of an identified association[51]. Furthermore, as mentioned previously, the behavior of clinicians and patients must not be ignored in interpreting associations, as the presence of a diagnosis may represent the acknowledgement of a disease and not biological onset. Also, the presence of abnormal lab results may indicate a symptom or indication bias that influences the

presence of a measurement. As such, sometimes these underlying behaviors that lead to a diagnosis or measurement may be associated with an outcome as opposed to the diagnosis or measurement itself[57].

In summary, hypothesis-driven approaches can provide valuable insight into specific questions, such as risk factors, disease prognosis, and drug exposure outcomes. Currently, causal inference methods and prospective studies are currently being implemented as a means to aim for more precise causal conclusions[70]. Nevertheless, due to the convoluted nature of the data, conclusions from EMRs studies are suggestive, and confidence in conclusions should be further supported by similar conclusions across study sites, study designs, and omics modalities[31,53] (**Figure 4.2**).

4.6 Considerations for EMR Studies

Data appropriateness for the question. When considering the utilization of EMR as a data source for a study, the first step is determining whether the EMR data is the right data source to answer the question, hypothesis, or exploratory question. This includes ensuring that the specificity of the cohort desired (e.g., ICD-10 code G35 may not be sufficient to identify multiple sclerosis subtypes) and measurements of interest (e.g., genotype information may be unavailable or sparse in EMR) are available. Data visualization tools such as ATLAS[71] and PatientExploreR[33] provide a means to understand data availability at either an individual or group level. Depending on the specificity, quality, or type of data desired, other real world data sources may be more relevant to the question of interest (e.g., claims database for medication use, research databases for improved specificity, social media for patient-oriented viewpoints)[9]. Mappings between real world data sources may exist to help augment EMR information with improved specificity or map phenotypic information with molecular information but careful consideration of selection biases and timing of

data is essential to consider for possible results and analyses that can be obtained. In the future, we anticipate the development of methods that map between databases for improved multimodal data availability as well as methods that account for missing data. Nevertheless, it is important to consider how much noisy preprocessing or incomplete mappings will impact the results of a potential analysis.

Sociopolitical and behavioral impacts on data. Often, real world datasets have a lot of influences due to societal exposures (e.g., racism, sexism, homophobia, transphobia, policies that negatively impact vulnerable populations) and biases that can impact patient behaviors or clinical decision making. In terms of societal exposures, patient language, background, and identities can influence the label assignment in the demographic tables of real-world datasets. These identities can differ between regions and countries and may represent the impact of an exposure as opposed to an inherent biological phenomenon[72]. If a question is aimed at studying the impact of sociopolitical groupings, it is recommended to utilize a "one vs rest" reference group or to identify controls of the same grouping[42], as opposed to a comparative analysis between two groups where one group is chosen as the reference group, similar to comparison of cell types in bioinformatics studies[73].

Biases in data, timing, and clinical decisions. In terms of biases that may impact patient or clinical decision making, often one should consider that the timing and presence of a diagnosis or record in the EMR often indicates a clinical acknowledgement of the entry, with both monetary and legal incentives in play. For example, due to potential impacts on a patient's mental capacity or driving rights when diagnosed with certain neurological disorders (e.g., epilepsy[74]), the timing of a diagnosis may be delayed. Prior temporal associations and predictive models may therefore pick up nonspecific diagnoses or other clues that indicate the diagnostic pathway of a clinician as

opposed to true biological risk. For example, some prior association analyses and models for Alzheimer's Disease pick up the prescription of an anti-dementia drug prior to an official Alzheimer's Disease diagnosis[75]. The existence of a lab or measurement may also indicate biases in ordering tests from a clinician (indication bias), while missing data may not indicate the lack of an abnormal measurement. Biases may also exist in clinical decision making, such as clinician biases towards certain identities. For example, race correction in interpreting glomerular filtration rate (GFR) or deciding on a C-section may lead to increased rates of procedures within a demographic group, which may be picked up by association studies or predictive models[38,76,77]. Those results should therefore be interpreted with potential clinical decision biases in mind before considering biological differences between groups. Predictive models may predict and identify causes of clinician misclassification or diagnostic errors, which can then be corrected in disease prediction models[54].

Interpretation is complex. Covariates and features utilized in patient representation are also important in the study design, and the choice of features that are balanced between groups also impact contextualization and interpretation of results[57]. Since EMR data may be inaccurate, incomplete, insufficiently granular, or transformed in ways that introduce new biases, the specificity of a phenotype may not be sufficient for accurate biological representation of a cohort. It is also important to consider the difference between characteristic information (e.g., demographic information), chronic conditions (e.g., hypertension), and acute conditions (e.g., fractured bone). Timing associated with chronic conditions may not be accurate due to the chronic nature of a disease. Furthermore, even "persistent" information may change (e.g., patient moves location). All these considerations in the accuracy and temporal nature of a feature or covariate can impact the interpretation of a result (**Figure 4.3**).

In the future, sophisticated methods with causal diagrams may be utilized to account for the influence of behaviors. ML models may still learn temporal patterns that are behavior-based rather than biology-based due to the retrospective nature of the data and the impact of indication biases. Heterogeneity due to data missingness and biases can not be completely accounted for in methods, so one should consider EMR informatics as one approach among many for understanding health and disease evidence. Augmentation of EMR data with specialty databases (e.g., use of a memory and aging center database for neurodegeneration[51]) or molecular datasets (e.g., genomics data[41]) provide the ability to identify links between biological and phenotypic signals. Furthermore, identifying an association or signal among alternative EMRs, diverse cohorts, or across omics modalities can help fully generalize what is identified in EMRs and improve biological plausibility. Follow-up studies will still be needed to help further support a hypothesized causal biological mechanism. Therefore, EMR analysis is only a step out of many in the process of clinical decision making, treatment identification, and biological understanding of a disease.

4.7 Conclusions

Electronic medical records provide an extensive, rich, longitudinal dataset with great opportunity for answering scientific questions, developing AI models, and advancing therapeutics in human health and disease[9,58,78,79]. Nevertheless, clinical behavioral biases, data missingness, data preprocessing, and societal impact of a conclusion should be considered when designing, implementing, and interpreting EMR-based studies. Currently, advancements in the EMR informatics fields include causal inference methods[45,80], state transition models[81], and transformer models[82,83] to account for temporal relationships in the records. Furthermore, methodological developments allow for combined analysis with both structured and unstructured datasets, including clinical imaging, clinical notes, and even inclusion of molecular omics datasets such as

genetics, gene expression, and proteomic measurements. Of importance is the identification of corroborated findings across datasets and omics modalities, even if on different patients, as similar signals across heterogeneous data collection and measurement methodologies can help validate a potential biological conclusion identified from a single dataset.

With improved data collection and methodological advancements, there is great potential for exciting applications in the future for deriving insight into biology based on EMR trends and molecular-phenotype associations, which allows for improved predictive modeling,[51], subtyping[46], drug repurposing, and therapeutic response investigations[65,78]. Prospective methods to evaluate algorithmic performance and biases are also developed as part of the implementation process to allow for iterative evaluation and improvement of algorithms or models, ensuring equitable performance across diverse cohorts[84]. With these considerations and advancements in both data collection and methodologies, EMR-based informatics research will provide support to the understanding and treatment of complex diseases in the future.

4.8 Figures

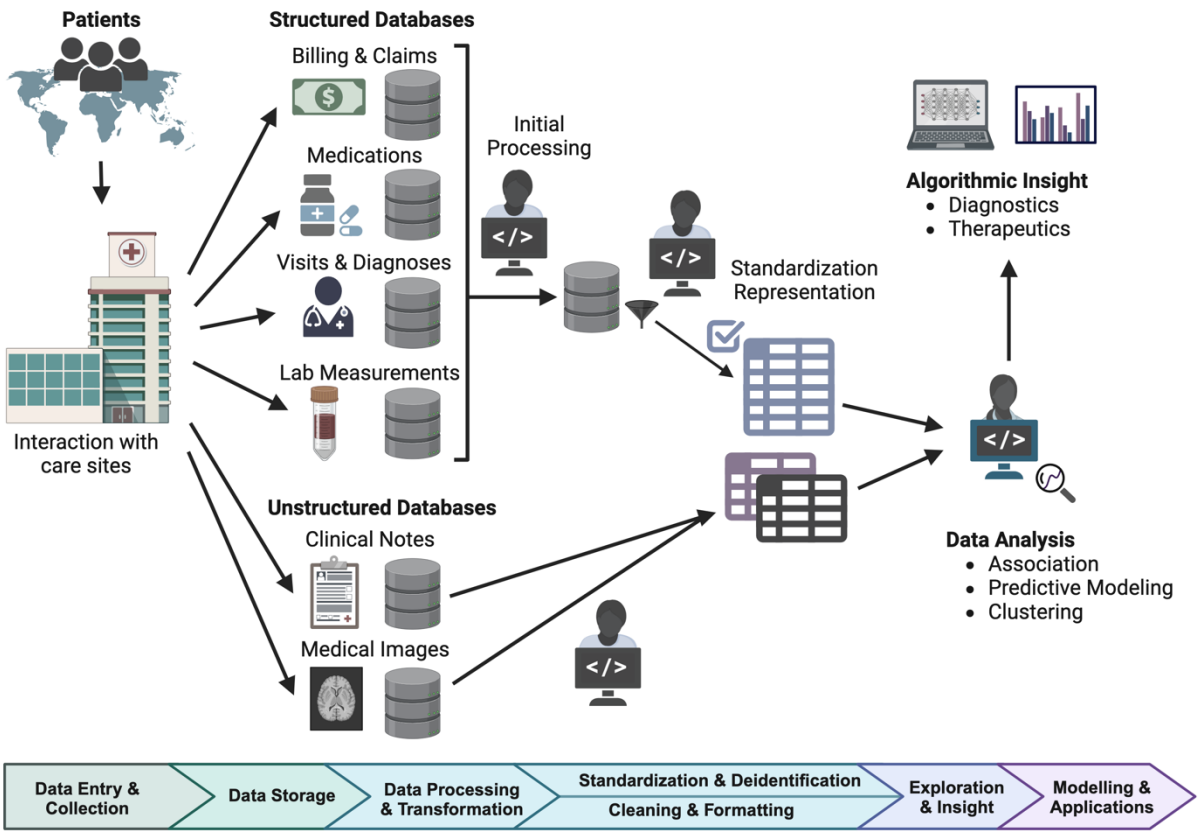


Figure 4.1 Electronic medical record data collection, storage, and processing for research applications

Patients first interact with a health-related care site, such as a hospital, primary care clinic, pharmacy, or other relevant sites. These intervals of care will have electronic footprints due to scheduling and billing, as well as provider inputted data about visits, diagnoses, and labs. Both structured and unstructured data may be obtained. Depending on the structure of the underlying databases, initial preprocessing must be performed by an information technology team to combine, simplify, standardize, and de-identify the data in order to make it available for researcher use. When the researcher accesses the data, the study goals and models will impact data processing and analysis decisions. Ultimately this data flow will impact the insights and algorithms obtained for various scientific inquiry or application purposes.

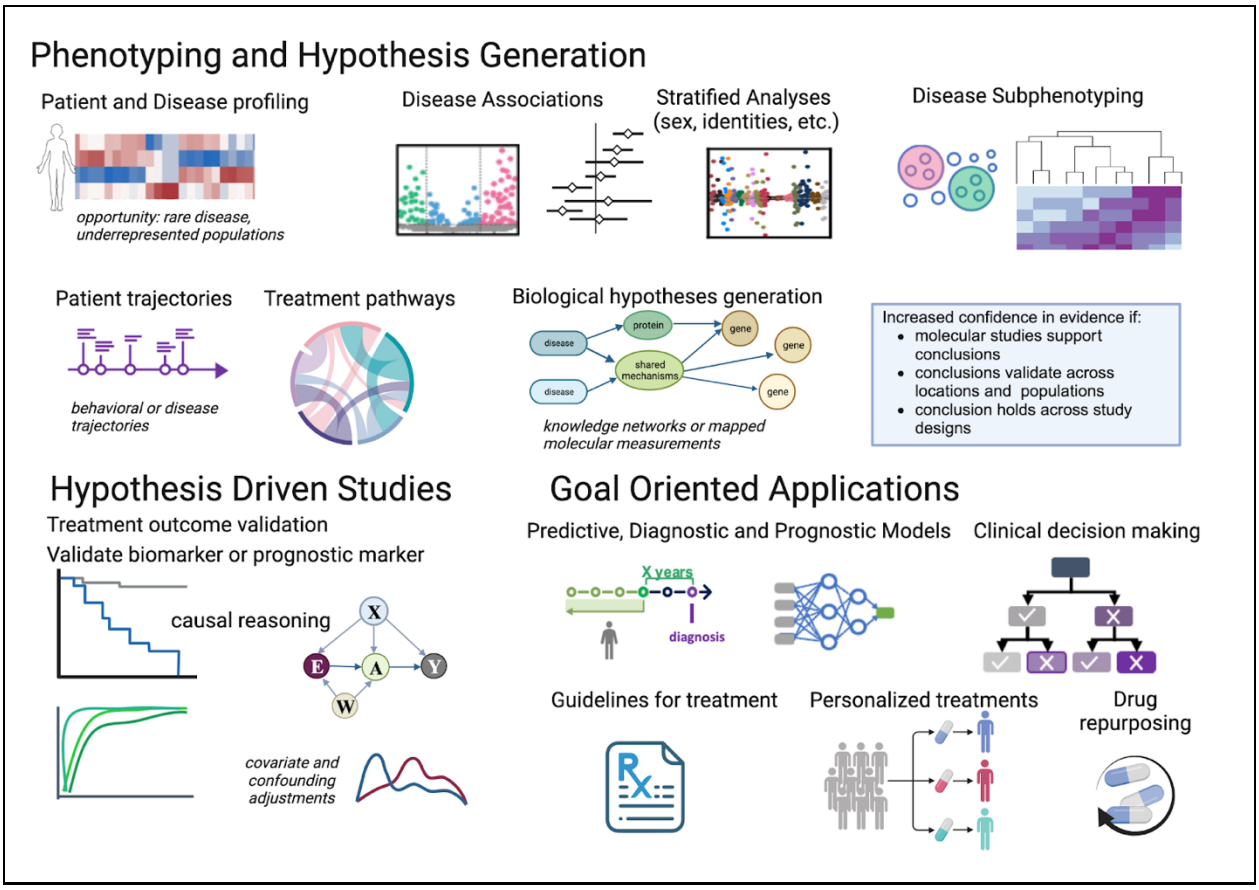


Figure 4.2 Potential EMR informatics study approaches include phenotyping and hypothesis generation, hypothesis-driven studies, and goal-oriented applications

Exploratory phenotyping and hypothesis generation provide unbiased profiling of a disease, including subphenotypes and relevant pathways and associations. Augmentation with biological datasets and knowledge networks can aid in biological hypothesis generation. Hypothesis-driven studies can further investigate a suggested relationship through careful selection of study design and adjustment methods. In both approaches, further evidence may be required to support or strengthen interpretations of the findings. Ultimately, applications of EMR informatics include predictive and diagnostic models, clinical decision making, drug repurposing and emulated trials, and ultimately support improvement of human health.

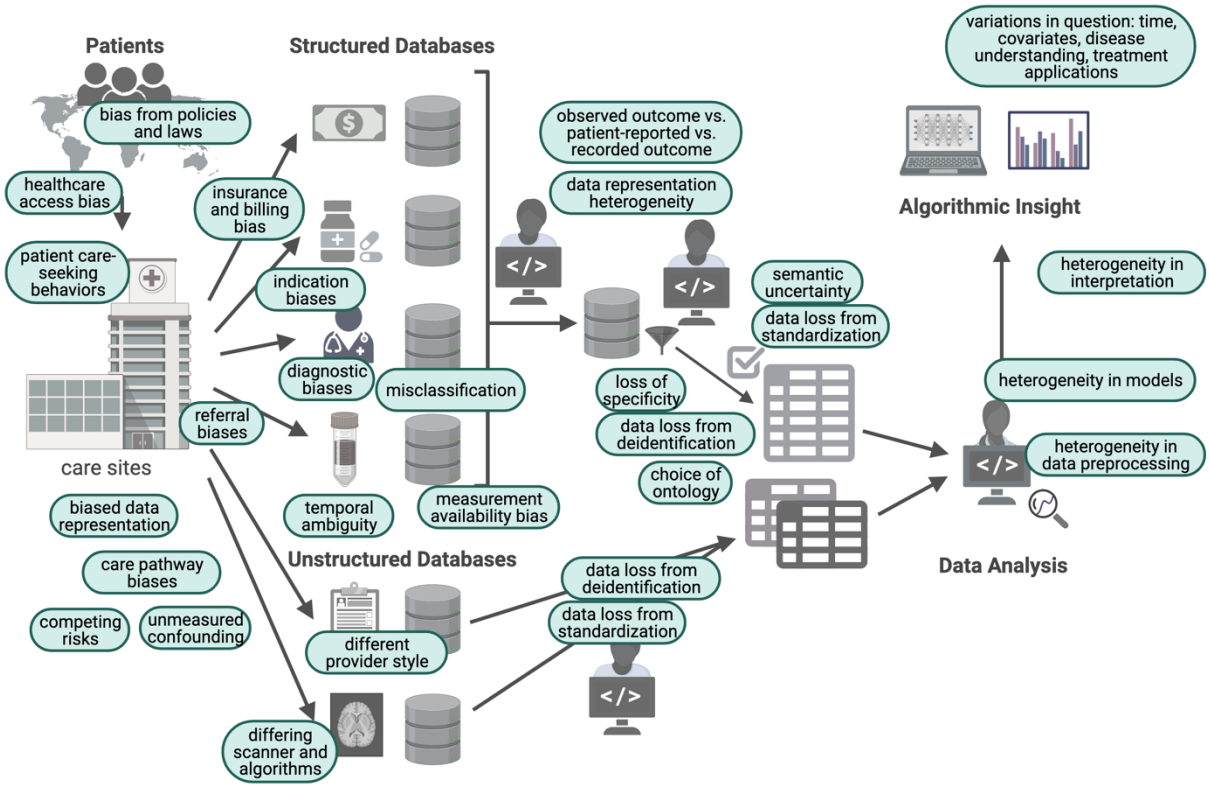


Figure 4.3 Sources of heterogeneity and bias in EMR-based informatics studies

EMR data, while providing extensive opportunity with a large unbiased sample of patients, also suffer from extensive biases and inaccuracies across the data flow pathway. These biases start from limitations in sampling of patients that seek healthcare, to data entry and representation heterogeneity, to choices of preprocessing and de-identification that can even introduce new biases to the data. Understanding these potential biases and sources of error are essential in the choice of data analysis methods and interpretation of results, or evaluation of models.

4.9 References

1. Gillum, R. F. From Papyrus to the Electronic Tablet: A Brief History of the Clinical Medical Record with Lessons for the Digital Age. *Am. J. Med.* **126**, 853–857 (2013).
2. U.S. Food and Drug Administration. *Real-World Evidence*. www.fda.gov/science-research/science-and-research-special-topics/real-world-evidence.
3. Gunter, T. D. & Terry, N. P. The emergence of national electronic health record architectures in the United States and Australia: models, costs, and questions. *J. Med. Internet Res.* **7**, e3 (2005).
4. Webster, P. C. Electronic health records a 'strong priority' for US government. *CMAJ Can. Med. Assoc. J. J. Assoc. Medicale Can.* **182**, E315-316 (2010).
5. Reis, Z. S. N. *et al.* Is There Evidence of Cost Benefits of Electronic Medical Records, Standards, or Interoperability in Hospital Information Systems? Overview of Systematic Reviews. *JMIR Med. Inform.* **5**, e26 (2017).
6. Agrawal, A. Medication errors: prevention using information technology systems. *Br. J. Clin. Pharmacol.* **67**, 681–686 (2009).
7. Greenhalgh, T., Potts, H. W. W., Wong, G., Bark, P. & Swinglehurst, D. Tensions and paradoxes in electronic patient record research: a systematic literature review using the meta-narrative method. *Milbank Q.* **87**, 729–788 (2009).
8. Office of the National Coordinator for Health Information Technology. *National Trends in Hospital and Physician Adoption of Electronic Health Records*. https://www.healthit.gov/data/quickstats/national-trends-hospital-and-physician-adoption-electronic-health-records.

9. Liu, F. & Panagiotakos, D. Real-world data: a brief review of the methods, applications, challenges and opportunities. *BMC Med. Res. Methodol.* **22**, 287 (2022).

10. Cowie, M. R. *et al.* Electronic health records to facilitate clinical research. *Clin. Res. Cardiol. Off. J. Ger. Card. Soc.* **106**, 1–9 (2017).

11. Kierkegaard, P. Electronic health record: Wiring Europe's healthcare. *Comput. Law Secur. Rev.* **27**, 503–515 (2011).

12. Wen, H.-C., Chang, W.-P., Hsu, M.-H., Ho, C.-H. & Chu, C.-M. An Assessment of the Interoperability of Electronic Health Record Exchanges Among Hospitals and Clinics in Taiwan. *JMIR Med. Inform.* **7**, e12630 (2019).

13. Johnson, A. E. W. *et al.* MIMIC-IV, a freely accessible electronic health record dataset. *Sci. Data* **10**, 1 (2023).

14. The All of Us Research Program Investigators. The "All of Us" Research Program. *N. Engl. J. Med.* **381**, 668–676 (2019).

15. Sudlow, C. *et al.* UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age. *PLOS Med.* **12**, e1001779 (2015).

16. Sinha, P., Sunder, G., Bendale, P., Mantri, M. & Dande, A. *Electronic Health Record: Standards, Coding Systems, Frameworks, and Infrastructures*. (Wiley, 2012). doi:10.1002/9781118479612.

17. El-Yafouri, R., Klieb, L. & Sabatier, V. Psychological, social and technical factors influencing electronic medical records systems adoption by United States physicians: a systematic model. *Health Res. Policy Syst.* **20**, 48 (2022).

18. Wang, Z. Data integration of electronic medical record under administrative decentralization of medical insurance and healthcare in China: a case study. *Isr. J. Health Policy Res.* **8**, 24 (2019).

19. Government of India. *EHR Standards for India*. https://www.nrces.in/standards/ehr-standards-for-india.

20. Overhage, J. M., Ryan, P. B., Reich, C. G., Hartzema, A. G. & Stang, P. E. Validation of a common data model for active safety surveillance research. *J. Am. Med. Inform. Assoc. JAMIA* **19**, 54–60 (2012).

21. Murugadoss, K. *et al.* Building a best-in-class automated de-identification tool for electronic health records through ensemble learning. *Patterns* **2**, 100255 (2021).

22. Stang, P. E. *et al.* Advancing the Science for Active Surveillance: Rationale and Design for the Observational Medical Outcomes Partnership. *Ann. Intern. Med.* **153**, 600 (2010).

23. Hripcsak, G. *et al.* Observational Health Data Sciences and Informatics (OHDSI): Opportunities for Observational Researchers. *Stud. Health Technol. Inform.* **216**, 574–578 (2015).

24. California Department of Health Care Services. *List of HIPAA Identifiers*. https://www.dhcs.ca.gov/dataandstats/data/Pages/ListofHIPAAIdentifiers.aspx.

25. Office for Civil Rights. Guidance Regarding Methods for De-identification of Protected Health Information in Accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule.

26. Yogarajan, V., Pfahringer, B. & Mayo, M. A review of Automatic end-to-end De-Identification: Is High Accuracy the Only Metric? *Appl. Artif. Intell.* **34**, 251–269 (2020).

27. Mandl, K. D. & Perakslis, E. D. HIPAA and the Leak of "Deidentified" EHR Data. *N. Engl. J. Med.* **384**, 2171–2173 (2021).

28. Norgeot, B. *et al.* Protected Health Information filter (Philter): accurately and securely de-identifying free-text clinical notes. *Npj Digit. Med.* **3**, 57 (2020).

29. Li, L. *et al.* Identification of type 2 diabetes subgroups through topological analysis of patient similarity. *Sci. Transl. Med.* **7**, 311ra174 (2015).

30. De Freitas, J. K. *et al.* Phe2vec: Automated disease phenotyping based on unsupervised embeddings from electronic health records. *Patterns* **2**, 100337 (2021).

31. Tang, A. S. *et al.* Deep phenotyping of Alzheimer's disease leveraging electronic medical records identifies sex-specific clinical associations. *Nat. Commun.* **13**, 675 (2022).

32. Su, C. *et al.* Clinical subphenotypes in COVID-19: derivation, validation, prediction, temporal patterns, and interaction with social determinants of health. *Npj Digit. Med.* **4**, 110 (2021).

33. Glicksberg, B. S. *et al.* PatientExploreR: an extensible application for dynamic visualization of patient clinical history from electronic health records in the OMOP common data model. *Bioinformatics* **35**, 4515–4518 (2019).

34. Huang, Z., Dong, W., Bath, P., Ji, L. & Duan, H. On mining latent treatment patterns from electronic medical records. *Data Min. Knowl. Discov.* **29**, 914–949 (2015).

35. Zaballa, O., Pérez, A., Gómez Inhiesto, E., Acaiturri Ayesta, T. & Lozano, J. A. Identifying common treatments from Electronic Health Records with missing information. An application to breast cancer. *PLOS ONE* **15**, e0244004 (2020).

36. Lou, S. S., Liu, H., Harford, D., Lu, C. & Kannampallil, T. Characterizing the macrostructure of electronic health record work using raw audit logs: an unsupervised action embeddings approach. *J. Am. Med. Inform. Assoc.* **30**, 539–544 (2023).

37. Glicksberg, B. S. *et al.* Comparative analyses of population-scale phenomic data in electronic medical records reveal race-specific disease networks. *Bioinformatics* **32**, i101–i110 (2016).

38. Obermeyer, Z., Powers, B., Vogeli, C. & Mullainathan, S. Dissecting racial bias in an algorithm used to manage the health of populations. *Science* **366**, 447–453 (2019).

39. Smith, M. A. *et al.* Insights into measuring health disparities using electronic health records from a statewide network of health systems: A case study. *J. Clin. Transl. Sci.* **7**, e54 (2023).

40. Swerdel, J. N., Hripcsak, G. & Ryan, P. B. PheValuator: Development and evaluation of a phenotype algorithm evaluator. *J. Biomed. Inform.* **97**, 103258 (2019).

41. Denny, J. C. *et al.* PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene-disease associations. *Bioinformatics* **26**, 1205–1210 (2010).

42. Woldemariam, S. R., Tang, A. S., Oskotsky, T. T., Yaffe, K. & Sirota, M. Similarities and differences in Alzheimer's dementia comorbidities in racialized populations identified from electronic medical records. *Commun. Med.* **3**, 50 (2023).

43. Austin, P. C. An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies. *Multivar. Behav. Res.* **46**, 399–424 (2011).

44. Karlin, L. *et al.* Use of the Propensity Score Matching Method to Reduce Recruitment Bias in Observational Studies: Application to the Estimation of Survival Benefit of Non-Myeloablative Allogeneic Transplantation In Patients with Multiple Myeloma Relapsing after a First Autologous Transplantation. *Blood* **112**, 1133–1133 (2008).

45. Ho, D., Imai, K., King, G. & Stuart, E. A. MatchIt: Nonparametric Preprocessing for Parametric Causal Inference. *J. Stat. Softw. Vol 1 Issue 8 2011* (2011).

46. Landi, I. *et al.* Deep representation learning of electronic health records to unlock patient stratification at scale. *Npj Digit. Med.* **3**, 1–11 (2020).

47. Bai, W. *et al.* A population-based phenome-wide association study of cardiac and aortic structure and function. *Nat. Med.* 1–9 (2020) doi:10.1038/s41591-020-1009-y.

48. Engels, E. A. *et al.* Comprehensive Evaluation of Medical Conditions Associated with Risk of Non-Hodgkin Lymphoma using Medicare Claims ("MedWAS"). *Cancer Epidemiol. Biomarkers Prev.* **25**, 1105–1113 (2016).

49. Bastarache, L., Denny, J. C. & Roden, D. M. Phenome-Wide Association Studies. *JAMA* **327**, 75 (2022).

50. Yazdany, J. *et al.* Rheumatology Informatics System for Effectiveness: A National Informatics-Enabled Registry for Quality Improvement. *Arthritis Care Res.* **68**, 1866–1873 (2016).

51. Tang, A. *et al.* Leveraging Electronic Medical Records and Knowledge Networks to Predict Disease Onset and Gain Biological Insight Into Alzheimer's Disease. 2023.03.14.23287224 Preprint at https://doi.org/10.1101/2023.03.14.23287224 (2023).

52. Nelson, C. A., Bove, R., Butte, A. J. & Baranzini, S. E. Embedding electronic health records onto a knowledge network recognizes prodromal features of multiple sclerosis and predicts diagnosis. *J. Am. Med. Inform. Assoc.* **29**, 424–434 (2022).

53. Roger, J. *et al.* Leveraging electronic health records to identify risk factors for recurrent pregnancy loss across two medical centers: a case-control study. *Res. Sq.* rs.3.rs-2631220 (2023) doi:10.21203/rs.3.rs-2631220/v2.

54. Mullainathan, S. & Obermeyer, Z. Diagnosing Physician Error: A Machine Learning Approach to Low-Value Health Care. *Q. J. Econ.* **137**, 679–727 (2022).

55. Makin, T. R. & Orban De Xivry, J.-J. Ten common statistical mistakes to watch out for when writing or reviewing a manuscript. *eLife* **8**, e48175 (2019).

56. Carrigan, G. *et al.* External Comparator Groups Derived from Real-world Data Used in Support of Regulatory Decision Making: Use Cases and Challenges. *Curr. Epidemiol. Rep.* **9**, 326–337 (2022).

57. Hersh, W. R. *et al.* Caveats for the Use of Operational Electronic Health Record Data in Comparative Effectiveness Research. *Med. Care* **51**, S30–S37 (2013).

58. Rudrapatna, V. A. & Butte, A. J. Opportunities and challenges in using real-world data for health care. *J. Clin. Invest.* **130**, 565–574 (2020).

59. Varga, A. N. *et al.* Dealing with confounding in observational studies: A scoping review of methods evaluated in simulation studies with single-point exposure. *Stat. Med.* **42**, 487–516 (2023).

60. Carrigan, G. *et al.* Using Electronic Health Records to Derive Control Arms for Early Phase Single-Arm Lung Cancer Trials: Proof-of-Concept in Randomized Controlled Trials. *Clin. Pharmacol. Ther.* **107**, 369–377 (2020).

61. Infante-Rivard, C. & Cusson, A. Reflection on modern methods: selection bias—a review of recent developments. *Int. J. Epidemiol.* **47**, 1714–1722 (2018).

62. Degtiar, I. & Rose, S. A Review of Generalizability and Transportability. *Annu. Rev. Stat. Its Appl.* **10**, 501–524 (2023).

63. Badhwar, A. *et al.* A multiomics approach to heterogeneity in Alzheimer's disease: focused review and roadmap. *Brain* **143**, 1315–1331 (2020).

64. Stuart, E. A. & Rubin, D. B. Matching With Multiple Control Groups With Adjustment for Group Differences. *J. Educ. Behav. Stat.* **33**, 279–306 (2008).

65. Taubes, A. *et al.* Experimental and real-world evidence supporting the computational repurposing of bumetanide for APOE4-related Alzheimer's disease. *Nat. Aging* **1**, 932–947 (2021).

66. Hernán, M. A. & Robins, J. M. Using Big Data to Emulate a Target Trial When a Randomized Trial Is Not Available: Table 1. *Am. J. Epidemiol.* **183**, 758–764 (2016).

67. Oskotsky, T. *et al.* Mortality Risk Among Patients With COVID-19 Prescribed Selective Serotonin Reuptake Inhibitor Antidepressants. *JAMA Netw. Open* **4**, e2133090 (2021).

68. Sperry, M. M. *et al.* Target-agnostic drug prediction integrated with medical record analysis uncovers differential associations of statins with increased survival in COVID-19 patients. *PLoS Comput. Biol.* **19**, e1011050 (2023).

69. Amit, G. *et al.* Antidepressant use during pregnancy and the risk of preterm birth – a cohort study. https://www.researchsquare.com (2023) doi:10.21203/rs.3.rs-3058509/v1.

70. Belthangady, C. *et al.* Causal deep learning reveals the comparative effectiveness of antihyperglycemic treatments in poorly controlled diabetes. *Nat. Commun.* **13**, 6921 (2022).

71. ATLAS. https://github.com/OHDSI/Atlas/wiki.

72. Lett, E., Asabor, E., Beltrán, S., Cannon, A. M. & Arah, O. A. Conceptualizing, Contextualizing, and Operationalizing Race in Quantitative Health Sciences Research. *Ann. Fam. Med.* **20**, 157–163 (2022).

73. Belonwu, S. A. *et al.* Sex-Stratified Single-Cell RNA-Seq Analysis Identifies Sex-Specific and Cell Type-Specific Transcriptional Responses in Alzheimer's Disease Across Two Brain Regions. *Mol. Neurobiol.* (2021) doi:10.1007/s12035-021-02591-8.

74. Krumholz, A. Driving and Epilepsy: A Review and Reappraisal. *JAMA* **265**, 622 (1991).

75. Xu, J. *et al.* Data-driven discovery of probable Alzheimer's disease and related dementia subphenotypes using electronic health records. *Learn. Health Syst.* **4**, e10246 (2020).

76. Vyas, D. A., Eisenstein, L. G. & Jones, D. S. Hidden in Plain Sight — Reconsidering the Use of Race Correction in Clinical Algorithms. *N. Engl. J. Med.* **383**, 874–882 (2020).

77. Gervasi, S. S. *et al.* The Potential For Bias In Machine Learning And Opportunities For Health Insurers To Address It: Article examines the potential for bias in machine learning and opportunities for health insurers to address it. *Health Aff. (Millwood)* **41**, 212–218 (2022).

78. Zhu, R. *et al.* Clinical Pharmacology Applications of Real-World Data and Real-World Evidence in Drug Development and Approval–An Industry Perspective. *Clin. Pharmacol. Ther.* **114**, 751–767 (2023).

79. Voss, E. A. *et al.* Accuracy of an automated knowledge base for identifying drug adverse reactions. *J. Biomed. Inform.* **66**, 72–81 (2017).

80. Stuart, E. A. Matching methods for causal inference: A review and a look forward. *Stat. Sci. Rev. J. Inst. Math. Stat.* **25**, 1–21 (2010).

81. Murali, L., Gopakumar, G., Viswanathan, D. M. & Nedungadi, P. Towards electronic health record-based medical knowledge graph construction, completion, and applications: A literature study. *J. Biomed. Inform.* **143**, 104403 (2023).

82. Li, Y. *et al.* BEHRT: Transformer for Electronic Health Records. *Sci. Rep.* **10**, 7155 (2020).

83. Guo, L. L. *et al.* EHR foundation models improve robustness in the presence of temporal distribution shift. *Sci. Rep.* **13**, 3767 (2023).

84. Gold, R. *et al.* Using Electronic Health Record-Based Clinical Decision Support to Provide Social Risk-Informed Care in Community Health Centers: Protocol for the Design and Assessment of a Clinical Decision Support Tool. *JMIR Res. Protoc.* **10**, e31733 (2021).

Conclusions

Increasing interests in real-world datasets, such as electronic medical records (EMRs), presents both opportunity and challenges in the study of complex disorders like Alzheimer's Disease (AD). The intricate interplay between molecular pathology and clinical manifestations poses significant hurdles in the development of effective disease modifying treatments. AD is characterized by considerable heterogeneity, hinting at the possibility of a spectrum of diverse mechanisms of disease onset, potentially influenced by factors such as biological sex. Our research leverages real-world data to demonstrate how deep phenotyping can facilitate a more accurate characterization of AD's real-world manifestations. This serves as a foundational step for comprehending variations attributable to characteristics like sex, and in identifying associations and hypotheses concerning disease risks, subgroup disparities, disease prognosis, and possible protective or harmful impacts of medications.

Furthermore, the rich data in EMRs pose opportunity for predictive modelling, which can pave the way for the creation of tools that could assist clinical decision-making through disease onset prediction and risk identification for possible early intervention. To address limitations inherent in EMR dataset quality, we enhanced EMR data with expert diagnoses from the UCSF Memory and Aging Center, ensuring a more accurate representation of potentially biological AD. Additionally, we carefully selected an index date to enhance the prediction accuracy of potential biological disease onset. Our work also underscores the value of integrating heterogenous knowledge networks, using human datasets as a starting point for deriving prioritized biological relationships. This approach led to notable findings, such as the identification of ACTB gene's role in the context of combined sensorineural hearing loss, arthropathy, and AD, offering avenues for targeted treatment strategies.

Beyond the prioritization of existing knowledge, our study showcases the benefits of enriching human dataset insights with molecular omics datasets. This approach has also facilitated novel discoveries, such as the high probability of causal association between MS4A gene family polymorphisms with low bone mineral density and AD, particularly among females. These findings open avenues for future research on the impact of MS4A variants in AD, especially in the context of known mechanisms in immune activation and homeostasis.

In summary, this dissertation demonstrates the effective utilization of EMR data, knowledge networks, and external omics databases in deepening our understanding of Alzheimer's disease and its heterogeneity. These insights are instrumental in guiding the future of personalized prevention and treatment strategies in AD, and the methodologies hold promise for furthering our understanding of other complex diseases within and beyond neurodegeneration.

Publishing Agreement

It is the policy of the University to encourage open access and broad distribution of all theses, dissertations, and manuscripts. The Graduate Division will facilitate the distribution of UCSF theses, dissertations, and manuscripts to the UCSF Library for open access and distribution. UCSF will make such theses, dissertations, and manuscripts accessible to the public and will take reasonable steps to preserve these works in perpetuity.

I hereby grant the non-exclusive, perpetual right to The Regents of the University of California to reproduce, publicly display, distribute, preserve, and publish copies of my thesis, dissertation, or manuscript in any form or media, now existing or later derived, including access online for teaching, research, and public service purposes.

DocuSigned by:

229125B24FFF409...

Author Signature

12/6/2023

Date