From Metabolites to Myocardial Infarction: Translational Bioinformatics of Atherosclerotic Cardiovascular Disease

by

Aaron S. Eisman

Sc. B., Brown University, 2008

A dissertation submitted in partial fulfillment of the
requirements for the Degree of Doctor of Philosophy
at the Center for Computational Molecular Biology at Brown University

Providence, Rhode Island

May 2023

This dissertation by Aaron S. Eisman is accepted in its present form by
the Center for Computational Molecular Biology as satisfying the dissertation requirement
for the degree of Doctor of Philosophy.

Date _____ _____
Indra Neil Sarkar, Ph.D., MLIS, Advisor

Recommended to the Graduate Council

Date _____ _____
Elizabeth Chen, Ph.D., Reader

Date _____ _____
Sorin Istrail, Ph.D., Reader

Date _____ _____
Wen-Chih Hank Wu, M.D., Reader

Approved by the Graduate Council

Date _____ _____
Thomas A. Lewis, Ph.D.
Dean of the Graduate School

Curriculum Vitae

Aaron S. Eisman

The Warren Alpert Medical School | Center for Biomedical Informatics | Brown University

EDUCATION

MD-PhD Candidate, Computational Biology, Brown University
ScB, Applied Mathematics, Brown University, 2008

RESEARCH INTERESTS

My current research interests include the development of biomedical informatics methods that leverage population-level data to improve the accuracy of cardiovascular disease risk estimation. My work aims to improve adherence to clinical practice guidelines, enhance the precision of preventative medical therapies, better explain observed health outcome disparities in racial and ethnic minority populations, and develop mechanisms for translating omics research into clinical medical practice.

RESEARCH EXPERIENCE

Doctoral Dissertation, Brown University, 2015-

From Metabolites to Myocardial Infarction: Translational Bioinformatics of Atherosclerotic Cardiovascular Disease

Clinical Research Coordinator, Massachusetts General Hospital, 2013-2015

Coordinated NIH and industry-sponsored heart failure and exercise physiology clinical trials. Used clinically indicated cardiopulmonary exercise testing with right heart catheterization to elucidate potential pre-clinical manifestations of heart failure in patients with dyspnea on exertion.

Research Assistant, Koleske Lab, Yale University, 2005-2007

Demonstrated first in vivo phenotypic dysregulation of neurogenesis in Arg -/- mice by observing that they do not undergo normal morphological maturation.

Teaching

Teaching Assistant, Biomedical Informatics and Data Science Skills - Brown University, Summer 2019

Designed and gave a lecture with a live programming demonstration on Plotting and Data Visualisation. Assisted students with programming during class and office hours. Met with individual students to design and implement summer research projects.

Teaching Assistant, Methods in Informatics and Data Science for Health - Brown University, Spring 2019

Assisted students with course material during live programming exercises in class, during office hours, and via electronic communication. Met with teams to design and implement their Course-based Undergraduate Research Experience projects.

Instructor, Introduction to Exercise Physiology - Brown University STEM II Summer Program, July 2016

Designed and taught a two-week course to rising 9th and 10th-grade students introducing them to exercise physiology. Daily interactive 3-hour lectures with four 3-hour afternoon sessions in the Brown exercise lab.

Volunteer

Brown Student Free Clinic @ Rhode Island Free Clinic, Board Member, 2015-2017

Patient Aide, Yale New Haven Hospital CCU, 2012-2013

Providence Science Outreach Co-Coordinator, Brown University, 2004-2006

PUBLICATIONS

*indicates equal contribution in work and shared co-first authorship

Usman A Tahir, Daniel H Katz, Julian Avila-Pachecho, Alexander G Bick, Akhil Pampana, Jeremy M Robbins, Zhi Yu, Zsu-Zsu Chen, Mark D Benson, Daniel E Cruz, Debby Ngo, Shuliang Deng, Xu Shi, Shuning Zheng, **Aaron S Eisman**, Laurie Farrell, Michael E Hall, Adolfo Correa, Russell P Tracy, Peter Durda, Kent D Taylor, Yongmei Liu, W Craig Johnson, Xiuqing Guo, Jie Yao, Yii-Der Ida Chen, Ani W Manichaikul, Frederick L Ruberg, William S Blaner, Deepti Jain, NHLBI Trans-Omics for Precision Medicine 1 Consortium, Claude Bouchard, Mark A Sarzynski, Stephen S Rich, Jerome I Rotter, Thomas J Wang, James G Wilson, Clary B Clish, Pradeep Natarajan, Robert E Gerszten. Whole Genome Association Study of the Plasma Metabolome Identifies Metabolites Linked to Cardiometabolic Disease in Black Individuals. Nat Commun [Internet]. 2022 Aug 22;13(1):4923. Available from: http://dx.doi.org/10.1038/s41467-022-32275-3

Aaron S Eisman, Katherine A. Brown, Elizabeth S. Chen, Indra Neil Sarkar. Clinical Note Section Detection Using a Hidden Markov Model of Unified Medical Language System Semantic Types. AMIA Annu Symp Proc 2021.

Aaron S Eisman, Nishant R Shah, Carsten Eickhoff, George Zerveas, Elizabeth S Chen, Wen-Chih Wu, Indra Neil Sarkar. Extracting Angina Symptoms from Clinical Notes Using Pre-Trained Transformer Architectures. AMIA Annu Symp Proc 2020.

Nishant R Shah*, **Aaron S Eisman***, David Winchester, Alan R Morrison, Reema Qureshi, Indra Neil Sarkar, Wen-Chih Wu. E-Consult Protocoling to Improve the Quality of Cardiac Stress Tests. JACC Cardiovasc Imaging [Internet]. 2020 Sep 26; Available from: http://dx.doi.org/10.1016/j.jcmg.2020.08.009

Jennifer E Ho, Emily K Zern, Emily S Lau, Luke Wooster, Cole S Bailey, Thomas Cunningham, **Aaron S Eisman**, Kathryn M Hardin, Robyn Farrel, John A Sbarbaro, Mark W Schoenike, Nicholas E Houstis, Aaron L Baggish, Ravi V Shah, Matthew Nayor, Rajeev Malhotra, Gregory D Lewis. Exercise Pulmonary Hypertension Predicts Clinical Outcomes in Patients With Dyspnea on Effort. J Am Coll Cardiol [Internet]. 2020 Jan 7;75(1):17–26. Available from: http://dx.doi.org/10.1016/j.jacc.2019.10.048

Jennifer E Ho, Emily K Zern, Luke Wooster, Cole S Bailey, Thomas Cunningham, **Aaron S Eisman**, Kathryn M Hardin, Giovanna A Zampierollo, Petr Jarolim, Paul Pappagianopoulos, Rajeev Malhotra, Matthew Nayor, Gregory D Lewis. Differential Clinical Profiles, Exercise Responses, and Outcomes Associated With Existing HFpEF Definitions. Circulation. 2019 Jul 30;140(5):353–65.

Aaron S Eisman*, Ravi V Shah*, Bishnu P Dhakal, Paul P Pappagianopoulos, Luke Wooster, Cole S Bailey, Thomas F Cunningham, Kathryn M Hardin, Aaron L Baggish, Jennifer E Ho, Rajeev Malhotra, Gregory D Lewis. Pulmonary Capillary Wedge Pressure Patterns During Exercise Predict Exercise Capacity and Incident Heart Failure. *Circulation: Heart Failure* May 2018.

Doreen DeFaria Yeh, Ada C Stefanescu Schmidt, **Aaron S Eisman**, John D Serfas, Mariam Naqvi, Mohamed A Youniss, Aaron D Ryfa, Asaad Khan, Lucy Safi, Sara R Tabtabai, Ami B Bhatt, Gregory D Lewis. Impaired Right

Ventricular Reserve Predicts Adverse Cardiac Outcomes in Adults with Congenital Right Heart Disease. *BMJ: Heart*. 2018 Jul 20.

Aaron S Eisman, Rory B. Weiner, Elizabeth S. Chen, Paul C. Stey, Rishi K. Wadhera, Aaron P. Kithcart, Indra Neil Sarkar. An Automated System for Categorizing Transthoracic Echocardiography Indications According to the Echocardiography Appropriate Use Criteria. AMIA Annu Symp Proc. 2017.

Nick E Houstis, **Aaron S Eisman**, Paul P Pappagianopoulos, Luke Wooster, Cole S Bailey, Peter D Wagner, Gregory D Lewis. Exercise Intolerance in HFpEF: Diagnosing and Ranking its Causes Using Personalized O2 Pathway Analysis. Circulation. 2017 Oct 9.

Ravi V Shah, Venkatesh L Murthy, Laura A Colangelo, Jared Reis, Bharath Ambale Venkatesh, Ravi Sharma, Siddique A Abbasi, David C Goff, J Jeffrey Carr, Jamal S Rana, James G Terry, Claude Bouchard, Mark A Sarzynski, **Aaron S Eisman**, Tomas Neilan, Saumya Das, Michael Jerosch-Herold, Cora E Lewis, Mercedes Carnethon, Gregory D Lewis, Joao AC Lima. "Association of Fitness in Young Adulthood With Survival and Cardiovascular Risk: The Coronary Artery Risk Development in Young Adults (CARDIA) Study." *JAMA Internal Medicine*: (2016) 1-9.

Rajeev Malhotra, Bishnu P Dhakal, **Aaron S Eisman**, Paul P Pappagianopoulos, Ashley Dress, Rory B Weiner, Aaron L Baggish, Gregory D Lewis. Pulmonary Vascular Distensibility Predicts Pulmonary Hypertension Severity, Exercise Capacity, and Survival in Heart Failure. Circ Heart Fail [Internet]. 2016;9(6).

Ravi V Shah, Shingo Kato, Sebastien Roujol, Venkatesh Murthy, Steven Bellm, Abyaad Kashem, Tamer Basha, Jihye Jang, **Aaron S Eisman**, Warren J Manning, Reza Nezafat. "Native Myocardial T 1 as a Biomarker of Cardiac Structure in Non-Ischemic Cardiomyopathy." *The American Journal of Cardiology* (2016).

Luiza H Degani-Costa, Barbara Levarge, Subba R Digumarthy, **Aaron S Eisman**, R Scott Harris, Gregory D Lewis. "Pulmonary vascular response patterns during exercise in interstitial lung disease." *European Respiratory Journal* (2015): ERJ-01910-2014.

Bishnu P Dhakal, Rajeev Malhorta, Ryan M Murphy, Paul P Pappagiaanopoulos, Aaron L Baggish, Rory B Weiner, Nick E Houstis, **Aaron S Eisman**, Stacyann S Hough, Gregory D Lewis. "Mechanisms of exercise intolerance in heart failure with preserved ejection fraction: the role of abnormal peripheral oxygen extraction." *Circulation: Heart Failure* (2014): CIRCHEARTFAILURE. 114.001825.

Meagan M Wasfy, James Deluca, Brant Berkstresser, Kathryn E Ackerman, **Aaron S Eisman**, Gregory D Lewis, Adolph M Hutter, Rory B Weiner, Aaron L Baggish. "ECG findings in competitive rowers: normative data and the prevalence of abnormalities using contemporary screening recommendations." *British journal of sports medicine* (2014): bjsports-2014-093919.

Mindan K Sfakianos, **Aaron S Eisman**, Shannon D Gourley, William D Bradley, AJ Scheetz, Jeffrey Settleman, Jane R Taylor, Charlie R Greer, Anne Williamson, Anthony J Koleske. "Inhibition of Rho via Arg and p190RhoGAP in the postnatal mouse hippocampus regulates dendritic spine maturation, synapse and dendrite stability, and behavior." *The Journal of Neuroscience* 27.41 (2007): 10982-10992.

Aaron S Eisman and Monty Robson. "Lightcurve of asteroid (21652) 1999 OQ2." *Minor Planet Bulletin* 31 (2004): 84.

ABSTRACTS AND PRESENTATIONS

NHLBI Trans-Omics for Precision Medicine Annual Meeting 2023, Rockville, MD
Protein-Metabolite Association Studies Identify Novel Proteomic Determinants of Metabolite Levels in Human Plasma (Multi-Omics Session, Jan 2023)

American Medical Informatics Association Symposium 2021, San Diego, CA
Clinical Note Section Detection Using a Hidden Markov Model of Unified Medical Language System Semantic Types (Oral Paper Presentation, Nov 2021)

American Medical Informatics Association Symposium 2020, Virtual
Extracting Angina Symptoms from Clinical Notes Using Pre-Trained Transformer Architectures (Oral Paper Presentation, Nov 2020)

American Medical Informatics Association Symposium 2017, Washington DC
An Automated System for Categorizing Transthoracic Echocardiography Indications According to the Appropriate Use Criteria (Oral Paper Presentation, Nov 2017)

American Heart Association Scientific Sessions 2015, Orlando, FL
Exercise Pulmonary Capillary Wedge Pressure Patterns Predict Heart Failure Outcomes (Oral Abstract Presentation, Nov 2015)

American Heart Association Scientific Sessions 2014, Chicago, IL
Left Ventricular Mass Predicts Left Sided filling Pressures and Exercise Capacity in Patients with Preserved Left Ventricular Function and Normal Resting Hemodynamics (Poster, Nov 2014)

Pulmonary Arterial Pressure Recovery Patterns Reflect Right Ventricular Function and Pulmonary Vascular Reserve (Poster, Nov 2014)

Heart Failure Society of America 2014, Las Vegas, NV
Pulmonary Arterial Pressure Recovery Patterns Reflect Right Ventricular Function and Pulmonary Vascular Reserve (Poster, Sep 2014)

RESEARCH GRANTS

Brown Scholarly Concentration Summer Research Fellow (2016) - $5,000
National Research Service Award, F30LM013320 (2020-2023) - $149,100

REVIEW ACTIVITIES

North East Computational Health Summit Abstract Reviewer (2019)
AMIA Annual Symposium Reviewer (2019-2023)
Journal of Biomedical Informatics, Ad Hoc (2020)

for Kira

Acknowledgements

In 2015 I arrived in Providence to enroll at Brown for the second time. I expected to spend four years as a medical student and then head off to residency. Neil Sarkar and Elizabeth Chen had grander plans and it has been a true privilege to be a part of them.

As the first MD/PhD student based at the Brown Center for Biomedical Informatics, I was fortunate to have Neil and Liz's unwavering support in forging both a professional and a personal path toward becoming a physician-scientist. They instilled the foundations and methods of the discipline and its many applications. Neil shepherded me through each step of the doctoral process, all while trusting my instincts, even when they were lofty. Liz reined me in when possibilities seemed limitless and time far too limited. Crucially, she helped me to obtain the data I needed to complete this project– perhaps the loftiest goal of all. Sorin Istrail's instruction grounded the entire range of bioinformatics methods that I implemented in this dissertation. Hank Wu ensured that clinical applications centered my theoretical endeavors. Nishant Shah and Ross Hilliard each reminded me that all of this work is for the patients.

Just when I thought I might have enough on my plate, I met with Mark Benson (and almost every Monday at 1 pm since). Mark has been the model of the physician-scientist I hope to be, and he enriched both my graduate training and this project far beyond the confines of a single chapter. Working through the finer points of multi-omics research problems week after week, Mark became both a mentor and a friend.

I am fortunate to have many homes at Brown. At BCBI, Carsten Eickhoff introduced me to large

language models for information retrieval that have been instrumental to this project. Hamish Fraser shared his passion for informatics of clinical symptoms. Karen Crowley whittled away at code mapping lists with me. Ashlin Harris figured out how to load the data. Paul Stey brought me into the machine learning fold. George Zerveus worked through the details of transformer architectures. Katherine Brown traversed all of graduate school (and new parenthood) nearly in lockstep and was who I went to whenever I needed help telling a complicated data story visually. Dilum Althuge helped untangle some of my knottiest problems. Hopefully one day he'll let me return the favor. Nicholas Jones brought enthusiasm and insight to lab meetings. Tarah Johnson was reliable and supportive as I navigated bureaucratic terrain.

At the Center for Computational Molecular Biology, Jeremy Bigness, Pinar Demetci, and Qing Wu were my coursework companions. Alan DenAdel knew the answer. Sahar Shahamatdara showed me the way. And, Nathanael Gardener was always smiling – often with technology, plants, or CCMB swag in hand.

At the Alpert Medical School, Allan Tunkel provided essential support at a critical juncture. Jonathan Kurtis and Tom Bartnikas championed a new incarnation of the MD/PhD program. Luba Dumenco and Paul George's passion for medical education was contagious and kept my sights on the horizon. Emily Green and Diane Green handed me some oars when I needed them.

There are some mentors who are present in every project. Tony Koleske believed I was a scientist, even when I thought that ship had sailed. Gregory Lewis gave me a second chance at a life in science and medicine. Robert Gerzsten taught me to "remember the biology." Laurie Farrel bridged connections across time and place.

There would be no dissertation were it not for some of the phenomenal educators that taught me the basics of science and visual literacy at Staples High School. Barbara Rose ignited my ambitious curiosity. Michael Lazaroff showed me that experiencing wonder and taking care of my whole self were crucial. David Scrofani remains the model educator I strive to be. Sara Scrofani trained my attention on the smallest details of visual communication. AJ Scheetz, first a teacher, then a mentor, and now a close friend, taught me how to embark on scientific inquiry.

Liz Peri's care for our new family was unparalleled and her enthusiasm helped me push this dissertation over the finish line. Lakshyan Shanzer gave me the tools to finish what I started.

Within this project, I see many traits that I have learned from my parents. My Mom's sense of imagination and passion for a life well lived is only surpassed by her ability to bring them to fruition. My Dad's eternal optimism helped me know that my path was possible. Lisa's studiousness and work ethic are hard to beat. Tom and Linda always told me how much they believed in me. Tom's "do-it-yourself" attitude and Linda's insistence on mindfulness encouraged me forward.

There are two people who were instrumental in my decision to set out on this journey. JP's patience and dependability helped make it a safe passage. My grandmother Blanche taught me that dreams and education were worth almost everything. Their images have looked over my shoulder as I wrote this dissertation, soaking in each word and figure on my computer screens.

Eve, you accelerated my research progress while stealing both my attention and my heart. And Kira, you are my *why* and my *how*. I dedicate this work to you. You make all of my dreams come true.

I am extremely grateful to the reviewers at the National Library of Medicine who supported my fellowship application. This project was funded in large part by NIH F30LM013320.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

A 39-year-old man with hyperlipidemia, hypertension, a 46 pack-year history of tobacco use, and positive family history of premature coronary artery disease presented with his first ST-segment elevation myocardial infarction in 1970. Despite smoking cessation and a 20-pound intentional weight loss, worsening exertional angina led to coronary artery bypass surgery in 1971. Twenty years later, the patient passed away in his sleep, secondary to graft failure at the age of 60.

This patient's story exemplifies the Coronary Profile proposed by Abraham Kagan in 1963. The concept of coronary artery disease as an epidemiological entity for which there exist risk factors including dyslipidemia, elevated blood pressure, smoking history, and family history (i.e., genetics) was coming into focus just as this patient experienced his first myocardial infarction. His lifestyle changes, including engaging in a regular exercise regimen and smoking cessation, had not yet been proven to be preventative (i.e., in the causal pathway) of future events, but would later become mainstay recommendations for both primary and secondary prevention of atherosclerotic cardiovascular disease (ASCVD). Near the end of his life, cardiovascular medicine ushered in an era of primary prevention. The first statin for lowering cholesterol [Alberts et al., 1980] was approved in 1987 [McKenney, 1988], and subsequent evidence that statins stabilize coronary plaques and prevent incident cardiovascular disease paved the way for the widespread adoption of statins for primary prevention in high-risk individuals. The Framingham risk score [Wilson et al., 1998] and

later the pooled cohort equations [Goff et al., 2014] were employed to guide shared decision-making discussions between doctors and patients around daily statin therapy and recommended lifestyle changes.

Seventy-five years after the first epidemiological study of coronary heart disease began in Framingham, Massachusetts, cardiovascular diseases remain the leading cause of morbidity and mortality in the United States [World Health Organization, 2018]. More than forty million Americans are prescribed statins annually [Salami et al., 2017] as a first-line pharmacotherapy to lower circulating low-density lipoprotein cholesterol, stabilize atherosclerotic plaques, and prevent myocardial infarction and stroke. Yet some individuals classified as "high-risk" never experience a major adverse cardiovascular event and others who are "low-risk" do. Clearly, significant gaps remain in the understanding of the biology of atherosclerosis and interactions between risk factors and outcomes.

Clinical guidelines for primary and secondary prevention that are informed by risk estimates rely on ten-year-old prediction models built on decades-old data that do not reflect the populations to which they are being applied. At the same time, the past decade has seen exponential growth in data generation from doctor-patient interactions, as well as from the laboratory. Electronic health records have developed from niche homegrown products [Pryor et al., 1983] that housed a tiny fraction of clinical data to comprehensive data management solutions that store nearly every documentable aspect of patient-healthcare interactions. These electronic health records are a natural substrate for investigating the human clinical "phenome." In addition, the advancement of high throughput methods for generating omics data that capture increasingly higher resolution snapshots of intricate biologically-based molecular network interactions between the genome, proteome, and metabolome, combined with advancing analytical methods and computing resources, enable the teasing apart of meaningful biological inter-omic relationships.

Despite these promising developments, data fragmentation – largely attributable to the fragmentation of healthcare delivery in the United States – constrains our potential to find causal connections between biomarkers and clinical outcomes. The natural state of big data is not sufficiently organized to facilitate data aggregation within a particular layer of abstraction (e.g., metabolome), let alone across the trans-omics spectrum (e.g., genome to phenome). As such, the organization of data according to standards is required for meaningful interdisciplinary knowledge discovery.

1.1 Contributions

This dissertation addresses these challenges by developing a translational bioinformatics framework for advancing the clinical practice of primary prevention of cardiovascular disease. This framework integrates (1) health informatics solutions to characterize population ASCVD risk and prevention, (2) bioinformatics methods to probe the underlying biology of atherosclerosis, and (3) genomic-linked electronic health records to propose potential biomarkers for incident ASCVD.

Healthcare data fragmentation and complexity are barriers to leveraging big data for understanding current clinical care practices and for characterizing local population risk profiles. This requires the development of new information retrieval tools as well as the integration of data using common data model standards. This dissertation presents novel applications for information retrieval from natural language data, including the answering of clinically relevant questions using large language models and transformer architectures. Domain expertise is applied for the modeling of hidden states to identify structure within natural language clinical notes.

In addition, this work presents the first-ever standardization of health information exchange clinical data into a common data model. The organization afforded by the common data model enables the understanding of population cardiovascular risk and primary prevention practices at the intersection of a range of care settings and providers. This demonstrated fragmentation in cardiovascular risk factor data collection across healthcare sites for individual patients and a significant opportunity to improve population statin guideline adherence.

Advancements in ASCVD prediction and patient risk stratification will be informed by causal biomarkers of atherosclerosis. Towards this goal, whole genomic, aptamer-based proteomic, and mass spectrometry-based metabolomic data derived from banked samples were studied to identify novel proteomic determinants of metabolite levels in human plasma. Mendelian randomization identified previously uncharacterized causal relationships that involve known ASCVD biomarkers. The top findings from this study were validated by performing serum metabolomics on murine knockout models. The gene-protein-metabolite associations from the bioinformatics study were integrated with whole genome sequencing linked electronic health records from the All of Us Research Program. This analysis rediscovered the relationship between apolipoprotein E and the development

of ASCVD and proposed a role for both the surface protein CD36 and plasmalogen glycerophospholipids in the causal pathway with incident myocardial infarction. Together, these studies lay the groundwork for future investigation of the biological basis of atherosclerosis and improved primary prevention of cardiovascular disease.

1.2 Specific Aims and Research Hypotheses

Specific Aims

This dissertation integrates and analyzes observational biomedical datasets to motivate and facilitate the identification of candidate causal biomarkers for incident cardiovascular disease.

In pursuit of this goal, the Specific Aims are to:

1. Characterize population atherosclerotic cardiovascular disease risk and primary prevention practices using a health information exchange.

2. Identify novel genomic and proteomic determinants of circulating levels of clinically relevant plasma metabolites.

3. Leverage genomics as a translational bridge to propose candidate causal biomarkers of incident cardiovascular disease.

Research Hypotheses

The aforementioned Specific Aims are addressed by testing the following three research hypotheses:

- Aim 1 Hypothesis: *Health information exchange aggregates cardiovascular disease risk factors that inform primary prevention care decisions for individual patients that would otherwise be missed by a single healthcare institution.*

- Aim 2 Hypothesis: *Instrumental variable analysis of multi-omics data derived from banked human plasma samples will identify novel causal inter-biomarker associations.*

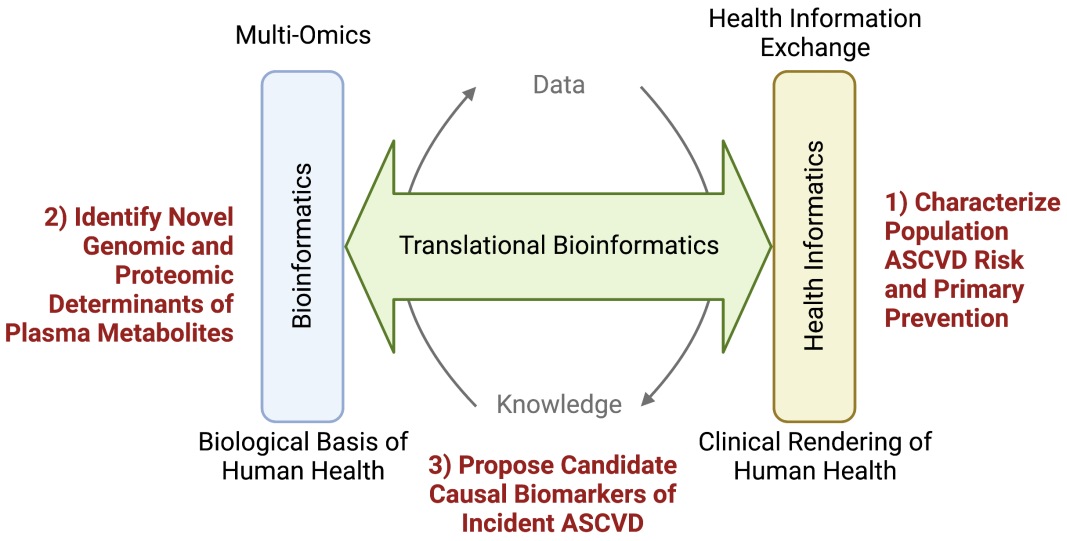

Figure 1.1: Specific Aims of the dissertation.

- Aim 3 Hypothesis: *Genomics is a translational bridge from bioinformatics knowledge to clinical phenotypes for the identification of candidate incident disease causal biomarkers.*

1.3 Translational Bioinformatics

The proverbial interpretation of translational medicine is the practical application of basic research findings in the direction "from bench to bedside." This paradigm supposes that there are significant advances in basic research that require additional research efforts to allow them to be applied in clinical settings (e.g., in a pill bottle). The field of translational bioinformatics [Butte, 2008, Shah and Tenenbaum, 2012, Tenenbaum, 2016], which is designed to bridge bioinformatics findings with applied health informatics domains, also presupposes that research advances from the molecular to the patient level.

These translational paradigms are rarely presented in the opposite direction – basic research can take inspiration from working closely with translational medical and clinical researchers. For this reason, it is important to note that the origins of this dissertation rest firmly in clinical applications and began as a quest to answer health informatics questions surrounding cardiovascular disease

risk and primary prevention. It was from this inspiration that further investigation ventured into the bioinformatics of molecular interactions to identify causal protein-to-metabolite pathway partners. Finally, this work traversed the translational research barrier a second time by translating those bioinformatics findings into a clinically relevant context using genomic-linked electronic health records. It should also be noted that this dissertation is central to the training of a dual-degree candidate for the degrees of Doctor of Philosophy and Doctor of Medicine in the hope that this combined training will be synergistically generative for both future research and clinical care.

1.4 Outline of Research and Dissertation

This dissertation is organized into six chapters. The first places the work that follows into both clinical and historical contexts, identifies existing knowledge gaps, and sets dissertation Specific Aims. Chapter 2 is a literature review that covers three broad topics. The first is a history of cardiovascular disease as an epidemiological entity that examines the development of the concept of risk factors for chronic disease as well as the shift towards population risk-based primary prevention. The second is an overview of biomarkers as risk factors, in particular those that are measured in blood and have been demonstrated to be associated with incident cardiovascular diseases. Next is a theoretical discussion of what can be learned from observational data and some foundational mathematical theories that are fundamental for distinguishing between correlation and causation.

There are three research chapters. Chapter 3 considers information retrieval from and the organization of electronic health records. The second half of this chapter focuses on health information exchange to characterize population cardiovascular risk and primary prevention practices. Chapter 4 is a departure from clinical data and instead focuses on deep multi-omic phenotyping of plasma samples collected as part of epidemiological studies. Here, the integration of genomics, proteomics, and metabolomics profiling identifies interaction partners between known and unknown clinically relevant molecular (e.g., protein and metabolite) biochemical actors. Chapter 5 integrates the work of Chapters 3 and 4 by translating experimentally validated bioinformatics findings into a clinical context using genomics-linked electronic health records to identify novel candidate biomarkers of incident cardiovascular disease. Chapter 6 provides a summary of these findings, discusses the limitations of work presented to date, and outlines planned future directions.

Chapter 2

Translational Medicine of Atherosclerotic Cardiovascular Disease

Cardiovascular diseases are the leading cause of morbidity and mortality in developed countries, including the United States [World Health Organization, 2018]. The primary mechanism for cardiovascular disease development is atherosclerosis. Atherosclerosis is the gradual build-up of plaque on the walls of arteries. This results in a narrowing of the lumen, increased turbulence of blood flow, and risk of plaque rupture and downstream total arterial occlusion (Figure 2.1). When this occurs in the heart it is known as myocardial infarction (MI), and it is a stroke when it occurs in the brain. Complete elimination of blood flow for an extended period leads to cell death due to a lack of oxygen supply. Depending on the size of the affected tissue, this can result in irreversible heart failure, functional impairment, or death.

The accumulation of plaque occurs over decades, and it is common for older asymptomatic adults to have arteries in their heart that are more than seventy percent occluded. These obstructions can be detected with a range of tests, including stress testing where a patient performs exercise

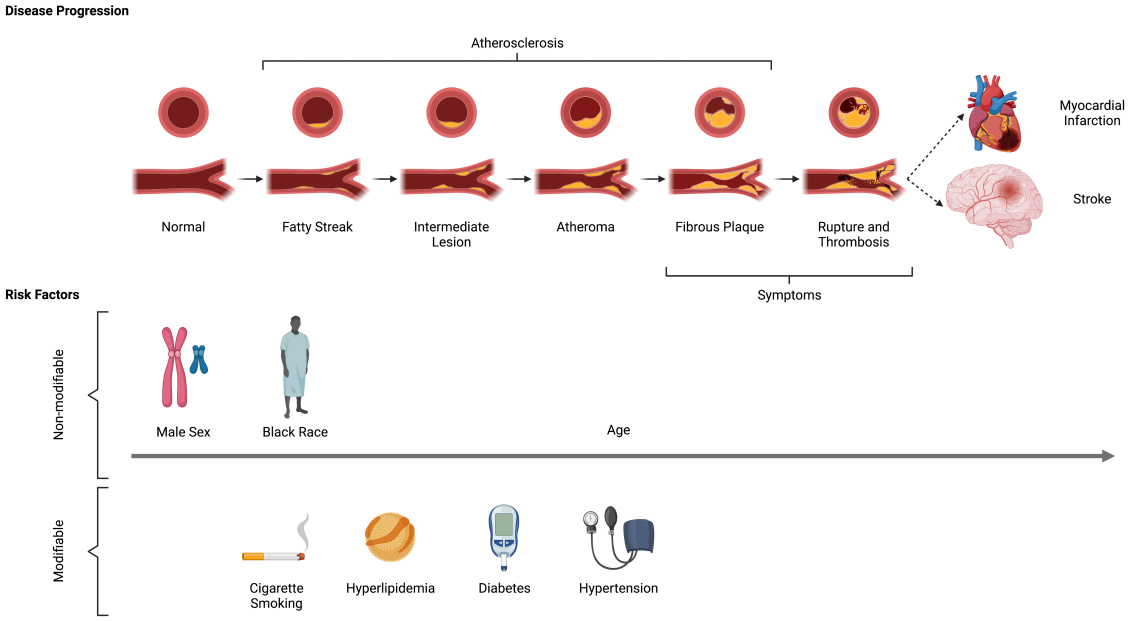

Figure 2.1: Atherosclerosis is the gradual build-up of plaque on the walls of arteries that initiates as a fatty streak and progresses through an intermediate lesion and atheroma without producing clinical symptoms. Fibrous plaques can cause symptoms (e.g., angina) and rupture or thrombosis can result in major adverse events such as myocardial infarction or stroke. Both non-modifiable and modifiable risk factors have been identified through epidemiological studies including male sex, Black race, cigarette smoking, hyperlipidemia, diabetes, and hypertension.

and an electrocardiogram is concurrently recorded. Associated changes in blood flow to cardiac muscle during exercise can also be measured with a radioactive nuclear tracer. The gold standard for diagnosing atherosclerotic occlusion is catheterization, which involves the percutaneous insertion of a small wire, release of radioopaque contrast dye, and direct visualization of stenosis.

Treatment options for patients with known or suspected atherosclerosis include medical therapies and interventions. HMG-CoA reductase inhibitors (i.e., statins) are the most effective and well-known for reducing cholesterol and stabilizing plaques, reducing their likelihood of rupture. Ezetimibe, bile-acid resins, and PCSK-9 inhibitors each play a second-line role in treating elevated cholesterol. Interventions include stent placement and coronary artery bypass surgery. Clinically, the goal and challenge are to identify patients who have the underlying disease before a catastrophic event, such as MI or stroke, so that they can be treated appropriately.

Coronary artery disease, MI, and stroke are collectively responsible for more than half of all cardiovascular deaths in the United States but are considered largely preventable by the Centers for Disease Control and Prevention (CDC) [Ritchey et al., 2018, Wright et al., 2018]. Black, Asian, and non-White Hispanic Americans are disproportionately impacted by atherosclerotic cardiovascular disease (ASCVD), present with symptoms considered "atypical," and are less likely to receive evidence-based preventative treatment [Hicks et al., 2004, Kurian and Cardarelli, 2007, McSweeney et al., 2006, Mensah et al., 2005, Ong et al., 2007, Redmond et al., 2011]. Collectively, marginalized populations have a higher prevalence of unrecognized or traditionally unmeasured risk factors, and many belong to populations too small to have been studied individually [Kurian and Cardarelli, 2007, Mensah et al., 2005]. These disparities motivate the need to identify risk factors that do not rely on imprecise socially constructed race categories [Fields and Fields, 2014] to better estimate risk across all populations.

2.1 Chronic Disease Management by Risk Estimation

The Framingham Heart Study

At the inception of the Framingham Heart Study (FHS), it was not obvious that atherosclerotic cardiovascular disease was appropriate for study as an epidemiological entity. The domain of epidemiology has its foundations in the study of infectious diseases and its modern origins are credited to John Snow and his identification of contaminated water sources as the primary factor in the spread of cholera [Snow, 1849, Snow, 1856]. For one hundred years, subsequent epidemiology of cardiovascular disease remained in the domains of infectious disease and nutritional deficiency. This included identifying the association between *Streptococcus pyogenes* with rheumatic heart disease [Coburn and Pauli, 1932], Syphilis with aortic aneurysm [Brockbank, 1925], and Rubella with congenital patent ductus arteriosis [Rutstein et al., 1952]. In addition, thiamine and niacin deficiencies and their associations with dilated cardiomyopathy (i.e., wet beriberi syndrome and pellagra, respectively) [Rachmilewitz and Braun, 1945]. Despite these well-established relationships between environmental factors and cardiovascular diseases, almost nothing was known about the factors associated with the primary and rising causes of morbidity and mortality due to hypertensive and atherosclerotic cardiovascular diseases. It was however believed that atherosclerotic cardiovascular disease was a complex multi-factorial disease that developed over a long period of time [Dawber

et al., 1951, Dawber et al., 1963]. FHS was designed to address these challenges and determine the important factors that lead to the development of atherosclerotic cardiovascular disease.

FHS is a community-wide study of the general adult population in the town of Framingham, Massachusetts. It was originally designed as a prospective study of the host and environmental factors that may contribute to the development of coronary heart disease. FHS was motivated by the increasing proportion of coronary heart disease-attributable deaths in the United States and indications by prior research that there were factors, including sex differences, that appeared to alter the probability of developing coronary heart disease. Starting in 1948, 5,209 men and women between the ages of 30 and 62 were recruited to be followed every two to six years for a total of 20 years. The initial examination included a comprehensive collection of medical history, medication use, past medical history, family and social history, a review of symptoms, and estimated retrospective average weights at five-year intervals. Physical exam was also conducted and included comprehensive body measurements, skin examination, cardiovascular physical exam, abdominal exam, chest x-ray, electrocardiogram, and blood draw [Dawber et al., 1951, Dawber and Kannel, 1958, Dawber et al., 1963]. These baseline data allowed for prevalence phenotyping of the population along with the definition of a subpopulation of "normals" in which the natural progression of atherosclerotic cardiovascular disease could be observed.

FHS defined coronary heart disease as the presence of one of three clinical syndromes [Dawber and Kannel, 1963]. The first is the presence of angina pectoris - substernal chest pain that clinical judgment determines to likely be cardiac in origin. The second is the clinical, laboratory, or electrocardiographic evidence of MI. Third, is the experience of sudden unexpected death either due to known cardiovascular disease or for an unknown reason. Prior data indicated that the vast majority of sudden death in the United States is due to cardiovascular disease and therefore assuming unexplained sudden death is a manifestation of coronary heart disease was a reasonable simplifying assumption.

With only four years of follow-up available, some of the most important risk factors for the development of coronary heart disease already started to come into focus [Dawber et al., 1957] (Figure 2.2). The incidence of CHD among men and women was roughly similar in the fourth and fifth decade of life but a significant increase in incident CHD was seen in men compared to women in the

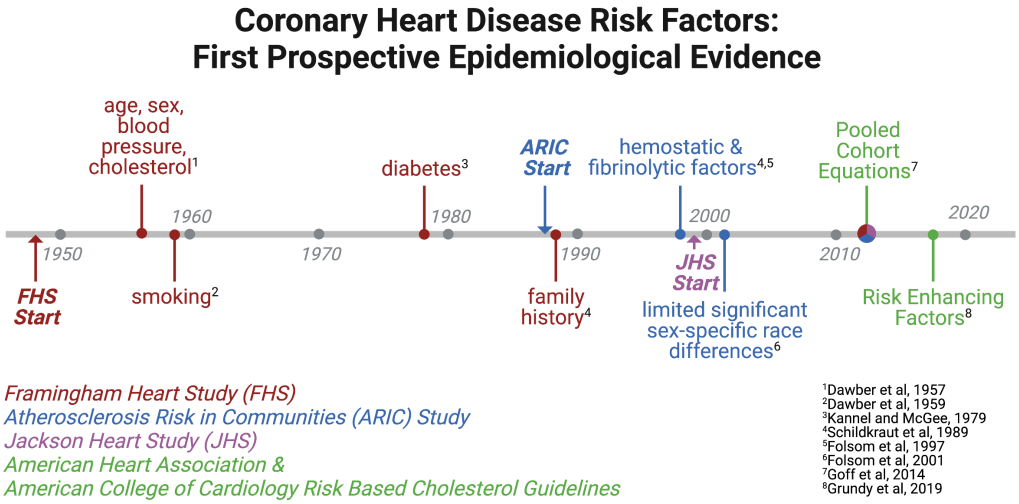

Figure 2.2: Timeline of coronary heart disease risk factor evidence

sixth decade of life. Even more stark was the difference in incident MI (15-to-1) where women were more likely to experience angina pectoris as their incident presentation of disease at rates similar to men. These differences in the prevalence of CHD and MI in particular between men and women would later drive decades of analyses of FHS data because an increase in the number of events at each analysis point meant that the study had more power to reveal significant epidemiologic factors in men compared to women and therefore evidence for the factors mounted earlier in men than women. Subsequent follow-up bolstered the evidence for sex differences in incident CHD [Dawber et al., 1962]. Sex was not the only strong CHD-associated factor identified.

Both elevated blood pressure and serum cholesterol levels were found to be significant factors associated with the incidence of CHD [Dawber et al., 1957, Dawber et al., 1962]. The finding that left ventricular hypertrophy secondary to high blood pressure was also found to be associated with incident CHD was generative of mechanistic questions regarding the causal relationship between all three. Obesity also appeared to be related to incident CHD, however, when considering the interaction between high blood pressure, cholesterol, and obesity, it became clear that while blood pressure and cholesterol are independent risk factors, the relationship between obesity and incident CHD is confounded by cholesterol. Subsequent joint dependence analysis of cholesterol and blood

pressure revealed that the association between blood pressure and cholesterol with incident CHD is independent and therefore risk is conferred that is multiplicative [Cornfield, 1962]. In particular, the relative risk of developing CHD in high versus low blood pressure and cholesterol was about five-fold. Having both high cholesterol and high blood pressure conferred a twenty-five-fold increase in risk compared to low levels of both risk factors. Furthermore, no critical values could be identified for increased incident CHD risk.

The association between smoking and CHD was evident but weak compared to blood pressure and cholesterol [Dawber et al., 1959]. This was in part due to the complex nature of recording smoking history including quantity (e.g., packs per day) and type of exposure (e.g., pipe, cigar, and cigarette). In addition to both of these factors affecting CHD outcomes, evidence also suggested that there might be a differential effect on subclasses of CHD [Dawber et al., 1962]. In particular, that cigarette smoking was the most harmful and may increase the risk of MI but not angina pectoris. Additional follow-up data was needed before definitive conclusions could be made. Eighteen years of follow-up enabled the observation of how changes in behavior, in particular the cessation of cigarette smoking, impacted incident CHD [Gordon et al., 1974, Gordon et al., 1975]. Subjects who were observed to have stopped cigarette smoking during the study and those who reported quitting cigarette smoking before the study began experienced subsequent incident CHD risk profiles substantially reduced compared to those who never quit smoking.

The epidemiological relationship between diabetes and coronary heart disease proved to be more complicated than blood pressure, cholesterol, or smoking. People with prior evidence of diabetes mellitus were observed to have a two to a three-fold higher risk of cardiovascular diseases than those without, but the elevated risk for CHD adjusted for other risk factors was less pronounced at 1.7 to two-fold [Kannel and McGee, 1979, Kannel, 1985]. The effects of diabetes on CHD were observed to be greater in women than men. In addition, subjects with evidence of diabetes had a higher prevalence of known risk factors. The relatively low prevalence of diabetes in the first generation of FHS combined with these other confounding factors contributed to the difficulty in interpreting the role of diabetes in incident CHD.

In addition to coronary heart disease, FHS data provided a rich insight into the comparative contribution of risk factors for incident vascular disease of the brain. Twelve-year follow-up comparing the

most important risk factors for MI with non-hemorrhagic thrombotic stroke demonstrated substantial overlap with notable differences [Kannel et al., 1965]. Evidence of thrombotic stroke was defined as the occurrence of new-onset focal neurologic deficit in the setting of a potential embolism source, such as atrial fibrillation or recent MI. Age appeared to have a differential effect on incident embolic stroke by sex. Women were observed to have a similar incidence of stroke and MI by age whereas stroke incidence in men lagged that of MI by 20 years. Blood pressure and electrocardiographic abnormalities both contributed to the risk of incident stroke. Cigarette smoking demonstrated a dose-response relationship in men while results were equivocal in women. Elevated serum cholesterol, a primary risk factor for coronary heart disease appeared to pose no increased risk of incident stroke.

FHS did not demonstrate an association between other measured factors and incident CHD. Measures of social and economic factors were inherently flawed [Dawber et al., 1959]. The place of residence as measured by voting precinct failed to demonstrate any meaningful differences in CHD incidence. Levels of education were strongly confounded by age, with younger study subjects being more likely to attain higher levels of education. Demographic homogeneity among Framingham subjects may not have provided sufficient ancestral variance to provide meaningful conclusions about this factor's potential role in the development of CHD.

The success of FHS led to a second cohort, the children of the original cohort and their spouses, to be subsequently enrolled and followed longitudinally starting in 1971 [Kannel et al., 1979]. The Framingham Offspring Study aimed to determine if there were changes in CHD risk factors between generations and determine the familial and genetic effects on those risk factors. Most recently, starting in the year 2002, the grandchildren of the original cohort have been studied as well [Splansky et al., 2007]. The Third Generation Cohort set out to supplement pedigrees and expand observations of the prior two cohorts. Over 70 years and thousands of publications, FHS has defined the modern notion of health risk factors [Kagan et al., 1963] and specifically translated the relationship between sex, cholesterol, blood pressure, smoking, and diabetes with coronary heart disease from hypothesis to lay health knowledge.

Familial aggregation of ischemic heart disease was long reported before rigorous study was possible. A well-conducted case-control study comparing reported family histories of consecutive patients

admitted for MI compared to other admitted patients demonstrated excess mortality from ischaemic heart disease in the MI patients' parents compared to controls [Rose, 1964]. FHS was able to build on these findings by collecting vital status and cause of death information on subjects at regular intervals throughout follow-up. This demonstrated that family history is an independent risk factor for incident CHD and does not interact with other known CHD risk factors [Schildkraut et al., 1989, Myers et al., 1990]. These studies go through significant effort to demonstrate the quality of family history data collected in FHS. While these efforts add to the voracity of the study findings they raise significant questions about the feasibility of collecting this kind of information reliably in a clinical setting.

The 2013 Guideline on the Assessment of Cardiovascular Risk

Lessons and insights from the FHS led to the conducting of several other large epidemiological studies of atherosclerotic cardiovascular diseases in the United States. These studies aimed to fill gaps in understanding resulting from the homogenous population of Framingham and leverage new technologies to measure additional biomarkers and subclinical disease.

The Atherosclerosis Risk in Communities (ARIC) study recruited subjects from four separate geographic regions including Forsyth Country, North Carolina, the city of Jackson, Mississippi, suburbs of Minneapolis, Minnesota, and Washington County, Maryland starting in 1987 [The ARIC investigators, 1989]. All sites attempted to recruit populations that represented the diversity of the recruitment area other than Jackson, where recruitment was limited to Black participants. The stated goals of ARIC included the investigation of the etiology and natural history of atherosclerosis and atherosclerotic diseases, and how incidence, associated risk factors, and medical care vary by race, sex, place, and time. In addition to the investigation of the coronary heart disease endpoints investigated in FHS, ARIC used arterial ultrasound as an indicator of preclinical disease by measuring wall thickness and distensibility, two physiologic properties that precede stenosis. Biomarkers beyond total cholesterol were also measured from baseline, including specific lipoproteins (e.g., Apolipoprotein B, low-density lipoprotein cholesterol, high-density lipoprotein cholesterol, apolipoprotein A-I, and apolipoprotein E), proteins involved in the clotting cascade (e.g., fibrinogen, factors VII and VIII, von Willibrand factor, antithrombin, and protein C), platelet function, and thromboxane B2.

ARIC demonstrated that fibrinolytic factors [Folsom et al., 2001], hemostatic factors [Folsom et al., 1997], and carotid arterial wall thickness [Chambless et al., 1997] are all associated with coronary heart disease outcomes independent of traditional risk factors but that the added information provided by these measurements may not be practically significant. In addition, an examination of sex and race differences in multi-variable adjusted risk factors for coronary heart disease failed to demonstrate statistically significant within-sex race differences in a variety of risk factors after adjustment. One notable exception was a significantly increased hazard of diabetes in White compared to Black women [Jones et al., 2002]. Other adjusted hazard ratios for risk factors including hypertension, smoking, and left ventricular hypertrophy were each observed to have notable but statistically insignificant differences by race.

Out of ARIC grew the desire to further study risk factors, especially hypertension, in Black individuals as well as develop institutional infrastructure and a pipeline for Black public health and epidemiology researchers. The Jackson Heart Study (JHS) was initiated as a continuation of the Jackson, Mississippi cohort from ARIC and sought to investigate these aims driven by epidemiological data that demonstrated unequal burden of cardiovascular risk factors and mortality in Black versus White individuals [Sempos et al., 1999]. In addition, longitudinal data for the United States and Mississippi in particular demonstrated that differences in age-adjusted rates of cardiovascular mortality by race increased from 1980 to 1995.

In parallel to ARIC and JHS, two additional population studies aimed to assess risk factors and outcomes in both young [Friedman et al., 1988] and older [Fried et al., 1991] age groups. The Coronary Artery Risk Developing in Young Adults (CARDIA) study looked at a biracial cohort of 18-30 year-olds to better understand how coronary artery disease risk factors evolve during young adulthood and inform the understanding of lifetime cardiovascular risk. The Cardiovascular Health Study (CHS) was motivated by the significant burden of cardiovascular disease in older individuals (65+), the aging United States population, to address an information gap in how traditional risk factors relate to incident cardiovascular disease in this cohort, and the desire to determine if any risk factors are unique to older individuals. pooled cohort equations In 2008, the National Heart, Lung, and Blood Institute (NHLBI) initiated the creation of a set of four guidelines including the assessment of cardiovascular risk, management of blood cholesterol, lifestyle modification to reduce

cardiovascular risk, and management of overweight and obesity in adults [Goff et al., 2014]. The American Heart Association and the American College of Cardiology in a joint effort convened expert panels to review the highest quality evidence available to develop these guidelines. A systematic review of the literature including randomized controlled trials, meta-analyses, and observational studies was considered. The remainder of this section will focus on the development of a set of pooled cohort equations for the assessment of cardiovascular risk and the use of these equations for the management of blood cholesterol.

The Risk Assessment Work Group developed an approach to quantitative risk assessment for ten-year incident ASCVD events. The definition of ASCVD events used for this purpose was a nonfatal MI or coronary heart disease death or fatal or nonfatal stroke [Goff et al., 2014]. They focused on data that primary care physicians were likely to have readily available about their patients and decided to formalize the use of absolute risk assessment that could be used to determine the appropriateness of preventative pharmacotherapy. The risk factors that became part of the pooled cohort equations (PCE) include sex, age, race, total cholesterol, high-density lipoprotein cholesterol, systolic blood pressure, hypertension treatment, diabetes, and tobacco use. In doing so, the Work Group acknowledged that applying any kind of average risk to an individual is flawed (e.g., no patient will ever experience a partial MI). However, as an approach to incident disease risk reduction on a population level, the paradigm was deemed effective. There was an attempt to match the derivation population to individual patients by developing four separate risk equations by sex (male or female) and race (White or Black). Insufficient data were available to develop additional race-specific risk prediction models.

Participant-level data from FHS, ARIC, CARDIA, CHS, and JHS were combined for the formation of the PCE [Goff et al., 2014]. Excluded from this analysis were participants who identified as a race other than White or Black. In addition, those identifying as Hispanic were also excluded. The insufficient numbers of participants in these subgroups and the inability to make additional race and ethnicity-specific prediction equations were the justification given for excluding them from the final analysis. However, the Work Group proposed that the same equations could be used in populations not represented in their derivation with a *moderate* level of supporting evidence.

Guidance for the use of these equations to inform clinical decision-making includes that as continuous

risk factor values get more extreme (e.g., systolic blood pressure near 200), the precise risk estimation should be interpreted with caution. However, risk factor values near or beyond the ranges studied should be subject to logical inference. For example, if a clinician is evaluating the ASCVD risk for an individual patient who has total cholesterol higher than the range accepted by the PCE, a combination of common sense and clinical judgment would allow the use of the maximum allowable value for risk estimation using the PCEs while knowing the individual patient's risk is likely higher than the equations report.

The motivating assumption for using risk factor-driven prediction of incident ASCVD is that the modifiability of some of the risk factors used for prediction (e.g., blood pressure and cholesterol) may inform clinical decision-making and reduce incident ASCVD. The corollary to this assumption is the suggestion by the Work Group that clinical decisions around risk factor modification match the intensity of conferred risk [Goff et al., 2014]. For example, a patient with relatively higher cholesterol should be considered for more aggressive cholesterol-lowering intervention while a patient with modestly elevated cholesterol may only require motivational interviewing around lifestyle changes.

The PCE predicts that one-third of the United States population has at least a 7.5% chance of experiencing a hard ASCVD event within the next ten years [Goff et al., 2014]. The prediction of ten-year ASCVD risk is the cornerstone of the 2013 guidelines on the management of blood cholesterol [Stone et al., 2014] as well as the more recently updated version of the guidelines [Grundy et al., 2019]. The algorithm for primary prevention in the most recent guidelines is reprinted as Figure 2.3. Briefly, patients with very high ($\geq 190 mm/dL$) low-density lipoprotein cholesterol (LDL-C) should be considered for a high-intensity statin. Any patient with diabetes mellitus 40-75 years old should be considered for a moderate-intensity statin and evaluated for incident ASCVD risk $\geq 7.5\%$ for potential high-intensity statin therapy. Patients 40 to 75 years old with LDL-C between 70 and $190 mg/dL$ without diabetes mellitus and a ten-year risk of incident ASCVD $\geq 7.5\%$ should be considered for moderate-intensity statin therapy. High-risk individuals ($\geq 20\%$) should be started on high-intensity statins. Patients of all ages and risk levels should be engaged in discussions about lifestyle modifications that can prevent or reduce the risk of developing ASCVD. Additional risk enhancers including (e.g., family history of early ASCVD and chronic kidney disease) should also inform clinical judgment shared decision-making with patients around statin appropriateness and

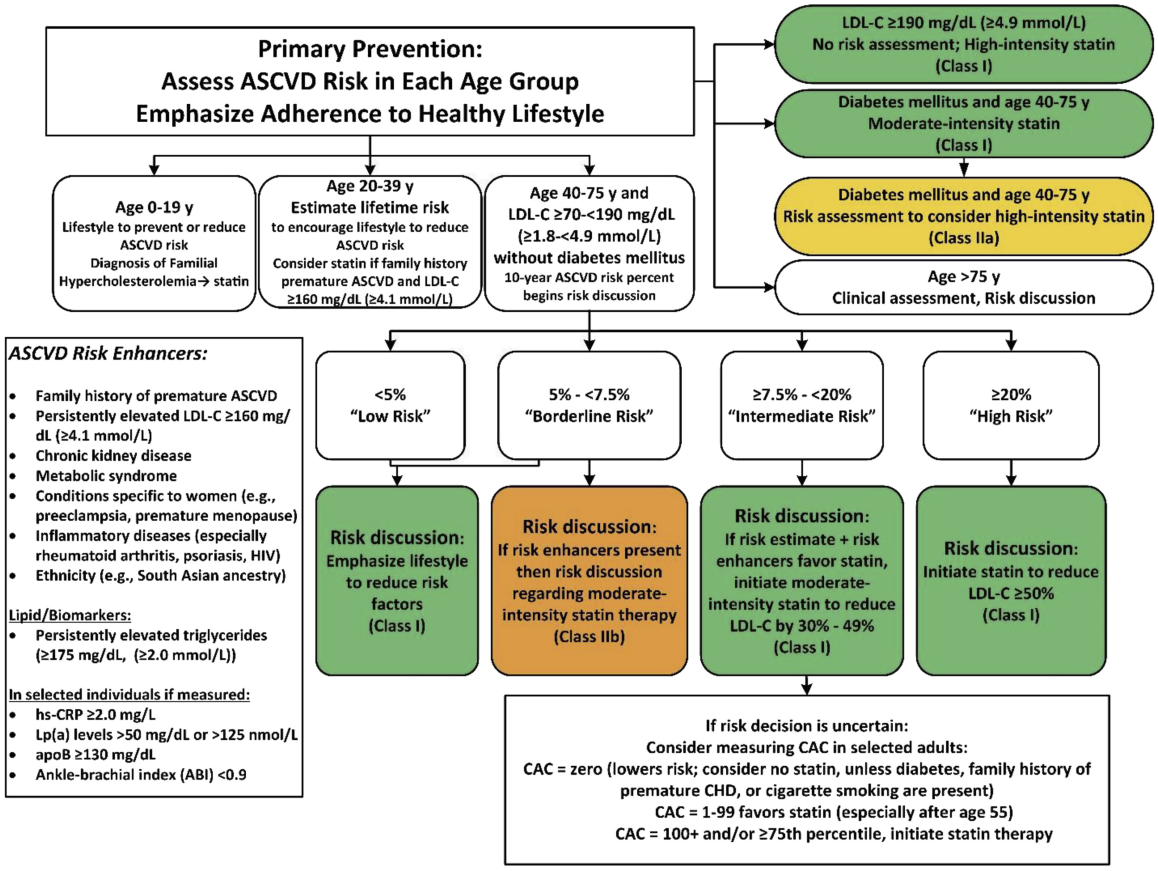

Figure 2.3: Algorithm for primary prevention of ASCVD [Grundy et al., 2019].

intensity.

Improving on the Pooled Cohort Equations

The significant burden of cardiovascular diseases has led to approach prevention on a population health level. That is, the patients who are at high risk compared to the general population receive testing and treatment.

The importance of accurate atherosclerotic cardiovascular disease (ASCVD) risk estimation and subsequent medical therapy recommendation has been underscored by the ISCHEMIA trial results, which demonstrate that optimal lipid-lowering therapy is equivalent to invasive procedures for ASCVD prevention [Reynolds et al., 2021]. The PCE define lipid-lowering therapy guidelines for the prevention of ASCVD in patients without known cardiovascular disease [Chou et al., 2016, Grundy

et al., 2019, Last et al., 2017, Stone et al., 2014, Taylor et al., 2013]. Overall, they have been validated to provide "moderate to good" discriminatory power (c-index = 0.71) [Muntner et al., 2014]. However, these equations have notable limitations. Validation studies suggest that the PCE may require cohort-specific calibration [Colantonio et al., 2017, DeFilippis et al., 2017, DeFilippis et al., 2015, Emdin et al., 2017, Loprinzi and Joyner, 2016, Zhang et al., 2017]. The use of single time point data for risk estimation results in therapy recommendations impacting sensitivity to single blood pressure measurements in up to 24% of patients at low and medium risk boundaries [Gupta et al., 2016]. In addition, simulation analysis of the PCE identifies paradoxical therapy impacting risk estimation for White women smokers and older Black women that are artifacts of logistic modeling of interaction terms [Schiros et al., 2015]. While the PCE improve upon risk estimation in Black patients over prior risk models [Qureshi et al., 2016, Topel et al., 2018], they accomplish this by computing separate equations for Blacks and side-step disparities that are likely a result of social factors of ASCVD risk [Morris et al., 2013]. This method assigns false precision to racial categories [Fields and Fields, 2014] and ignores observed ASCVD disparities in other racial and ethnic groups [Graham, 2015]. Finally, the burden of calculating ASCVD risk remains a barrier to optimizing lipid-lowering therapy [Bakhai et al., 2018], and improved pharmacotherapy-based primary prevention could reduce associated deaths by 12.6% annually [Yang et al., 2017]. Therefore, there is a need to streamline the calculation and increase the accuracy of ASCVD risk estimation at a population level. Doing so will further enable the primary prevention of incident ASCVD events.

2.2 Biomarkers as Intermediates of Disease Relevant Biochemical Processes

Identification of the next generation of predictors and modifiable risk factors of atherosclerotic cardiovascular disease will require an improved understanding of plasma biochemistry. Several well-characterized biochemical pathways are related to atherosclerosis including cholesterol metabolism, coagulation, inflammation, and glucose homeostasis. The summary of these biochemical processes over the next few pages is informed by The Pathologic Basis of Disease [FRCPath et al., 2020]. On the trans-omics spectrum from the genome to phenome, the proteome and metabolome are the biochemical actors most in proximity to clinical phenotype (Figure 2.4). Compared to genetics, the proteome and metabolome are significantly lower dimensionality and represent a real-time indication

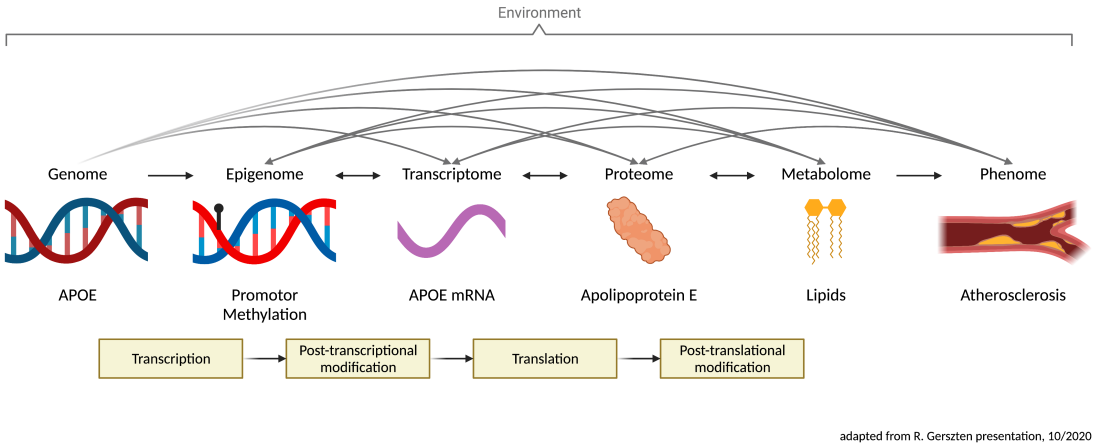

Figure 2.4: Trans-omics interaction with human phenotypes. Broad classes (-omes) of structurally and functionally related biologically important molecules exist on a spectrum from genome to phenome. These include the epigenome, transcriptome, proteome, and metabolome. Each class participates in a complex network of intra- and inter-ome interactions through biochemical processes that include transcription, post-transcriptional modification, translation, and post-translational modification. The constellation of these interactions manifests in physiological and pathophysiological phenotypes.

of genetics interacting with the environment. For this reason, they are theoretically well-positioned as predictive risk factors and to elucidate mechanisms of subclinical disease processes.

Cholesterol Metabolism

Cholesterol is absorbed through the diet and endogenously produced as needed. Ingested cholesterol and triglycerides form chylomicrons in the small intestine and are absorbed through the gut mucosa. Chylomicrons are transported via the lymphatic system, avoiding first-pass metabolism by the liver, before being released into the bloodstream for circulation. They are hydrolyzed by endothelial lipoprotein lipases in the capillaries of smooth muscle and fat tissue. Chylomicron remnants rich in cholesterol are transported to the liver where they either enter into the metabolic pool or are excreted as bile acid or free cholesterol.

The liver secretes very-low-density lipoproteins (VLDL) which are comprised of a large number of triglycerides and significant cholesterol, as well as surface proteins apolipoprotein C (Apo-C), apolipoprotein E (Apo-E), and apolipoprotein B-100 (Apo-B). Similar to chylomicrons, VLDLs

undergo lipolysis in the capillaries of muscle and fat tissue which in turn uptake triglycerides. The VLDL-remnant, also known as intermediate density lipoprotein (IDL), differ from VLDL in that they contain relatively fewer triglycerides and lack Apo-C surface protein. About half of IDL returns to the liver where it is taken up via the LDL-receptor (LDL-r) which recognizes Apo-B or Apo-E to initiate endocytosis of IDL. The remaining IDL is converted to LDL, which is primarily comprised of cholesterol with Apo-B surface protein. Most LDL is returned to the liver and taken up by the LDL-r. However, the process is achieved with lower efficiency due to the absent Apo-E surface protein. A minority of LDL is taken up by LDL-r on cells other than the liver throughout the body.

Low-density lipoprotein receptors are found within clathrin-coated pits on the surface of cells. These receptors bind Apo-B on the surface of LDL, the binding of which initiates the invagination of the pit as a coated vesicle and subsequent endosome. The LDL-containing endosome fuses with lysosome for processing where LDL-r are hydrolyzed into amino acids, and cholesteryl esters are converted to free cholesterol, all of which are excreted into cellular cytoplasm. Free cholesterol can then be used by the cell, for example, as a building block for cell membranes and steroid hormone production, or can be stored as cholesteryl esters. Oversupply of cytoplasmic cholesterol acts as a regulator, inhibiting the endogenous production of cholesterol by β-hydroxy β-methylglutaryl coenzyme A (HMG-CoA), as well as inhibiting the synthesis of new LDL-r. While low LDL-r expression reduces cholesterol uptake by the cell, it results in a net increase of circulating LDL. In addition to de-novo LDL-r synthesis, LDL-r recycling is inhibited by proprotein convertase subtilisin/kexin type 9 (PCSK9). These are two mechanisms that pharmacotherapy targets for lowering LDL cholesterol, a modifiable risk factor associated with the development of incident ASCVD. Statins, also known as HMG-CoA reductase inhibitors, reduce the endogenous production of cholesterol as a first-line treatment for elevated LDL cholesterol. PCSK9 inhibitors inhibit the inhibition of LDL-r recycling, leading to increased cellular uptake of LDL and reduced circulating LDL cholesterol. This is a second-line therapy for patients who do not achieve adequate blood cholesterol control from statins and lifestyle management alone.

Coagulation Cascade

The lumen of arteries is lined by a single layer of endothelial cells. A layer of smooth muscle controls vascular resistance, dynamically adjusting to maintain blood pressure and shunt blood flow.

Pathologic narrowing of the lumen occurs by two primary mechanisms relevant to atherosclerotic cardiovascular disease events. The first is secondary to prolonged exposure to hypertension. Hyaline arteriosclerosis is the process by which there is a hyaline thickening with luminal narrowing. Injured endothelial cells release plasma proteins and smooth muscle synthesizes additional matrix in response to abnormal hemodynamics. These pathologic changes may be a mechanism by which hypertension is a risk factor for atherosclerosis

Atherosclerosis is a complex, poorly understood process, by which exposure to modifiable and unmodifiable risk factors over time contributes to the development of intimal atherosclerotic plaques. These plaques are raised lesions predominantly comprising coagulated cholesterol and cholesterol esters with a fibrous cap. Atherosclerotic plaques can result in clinical ASCVD as a result of direct obstruction or rupture leading to thrombosis.

Ulcerated atherosclerotic plaques expose subendothelial von Willibrand Factor (vWF) and tissue factor (TF), which can result in turbulent blood flow. Platelets bind to vWF via glycoprotein Ib with the initiation of primary hemostasis. These activated platelets then release adenosine diphosphate and thromboxane A2 which recruit and activate additional platelets. Platelets aggregate and bind together with surface glycoprotein IIb/IIIa crosslinked by fibrin forming a hemostatic plug.

Secondary hemostasis is the process of fibrin polymerization to stabilize a clot in a cascade of reactions. Proenzyme clotting factors are activated into enzymes by other activated clotting factors. Exposed tissue factor catalyzes the conversion of coagulation factor VII (factor VII) into activated factor VII (factor VIIa). Factor VIIa converts factor IX to IXa, and IXa activates factor X. Factor Xa is responsible for converting prothrombin (factor II) to thrombin (factor IIa), which converts fibrinogen (factor I) into fibrin (factor Ia). Thrombin is also responsible for positive feedback indirectly promoting IX to IXa via XI to XIa intermediate, as well as stimulating factor VIIIa and Va influence on IXa and Xa respectively. These reactions are facilitated by an activated platelet phospholipid surface, the product of primary hemostasis.

Inflammation

Chronic inflammation plays a direct role in atherosclerosis. In particular, macrophages, foam cells, and lymphocytes are implicated in the development and expansion of atherosclerotic plaques. Macrophages in particular are responsible for the release of cytokines including interleukins and chemokines, as well as growth factors, which can be pro- or anti-inflammatory depending on the pathway by which the macrophage is activated. A particularly well-characterized node of inflammatory mediators is arachidonic acid and its metabolites.

Arachidonic acid is a polyunsaturated fatty acid derived from linoleic acid. Its primary metabolic pathways are via cyclooxygenase and 5-lipoxygenase. Cycloxygenase is responsible for converting arachidonic acid into prostaglandins, prostacyclins, and thromboxane, each of which plays a different role in mediating inflammation. As noted previously, thromboxane A2 is responsible for platelet activation and promotes aggregation, while prostacyclin I2 inhibits platelet aggregation. The prostaglandins mediate vascular permeability and leukocyte chemotaxis. All cyclooxygenase pathway arachidonic acid metabolites mediate vascular resistance. The 5-lipoxygenase pathway produces cytokines including leukotrienes responsible for mediating chemotaxis and vascular permeability.

Acetylation of the cyclooxygenases by aspirin results in irreversible inactivation. In this way, aspirin is used as a preventative treatment for atherosclerotic cardiovascular disease events secondary to thrombus formation. Aspirin is primarily employed in this capacity for secondary prevention of ASCVD events in patients who have survived occlusive cardiovascular disease (e.g., MI, or ischemic stroke).

Additional markers of inflammation synthesized by the liver have also been shown to be associated with atherosclerosis. Cytokines including interleukin-6, secreted by smooth muscle cells, stimulate the liver to produce c-reactive protein, a serum marker for inflammation and elevated risk of MI.

Glucose Homeostasis

Dysfunctional insulin secretion or action results in hyperglycemia, the pathognomonic characteristic of diabetes mellitus. Sustained exposure to elevated glucose levels in the blood leads to endothelial dysfunction and worsening insulin resistance. In particular, macrovascular damage is implicated in

accelerating the process of atherosclerosis and incident atherosclerotic cardiovascular disease events with the same morphology as in individuals without diabetes.

The mechanism by which this occurs is the production of elevated levels of advanced glycosylated end products (AGE). These are formed by the binding of protein amine groups to glucose metabolites, which circulate in elevated concentrations secondary to hyperglycemia. Advanced glycosylated end products then bind receptors on inflammatory, endothelial, and vascular smooth muscle cells to stimulate the release of cytokines and growth factors (e.g., TGF-β). Vascular endothelial growth factor and elevated basement membrane deposition combined with reactive oxygen species in the endothelium contribute to a state of chronic inflammation. Chronic inflammation supports pro-coagulation activity of the endothelium, increased extracellular matrix production, and vascular smooth muscle hypertrophy. Collectively, this macrovascular disease is associated with ASCVD.

Serum Biomarkers as Predictors of Incident ASCVD

Biomarkers are discretely measurable, biologically relevant indicators of underlying physiology. The quantitative level of many proteins and metabolites in human plasma is measured routinely as part of clinical care. These biomarkers are employed to screen patients for underlying pathology, diagnose suspected disease, and inform clinical prognosis. The well-described processes of cholesterol metabolism, coagulation, inflammation, and glucose homeostasis are all mediated by proteins and metabolites, as described above, that can be measured within the human plasma. Many of these are known predictors of atherosclerotic cardiovascular disease. While these biochemical processes are comprised of key reactions between enzymes, substrates, signaling molecules, and the catabolism and anabolism of small molecules, the in vivo behavior is undoubtedly more complicated and dynamic. This is particularly true concerning the manifestation of clinical pathology and the source of individual variation that manifests as pathology in normal physiology in one person and pathology in another.

Plasma biomarkers currently used in clinical practice are derived from well-characterized biochemical pathways and therefore are often highly correlated to other known risk factors. Investigation of classic biomarkers for atherosclerotic cardiovascular disease has likewise demonstrated a high correlation with known risk factors, which limits their contribution to predictive power for screening purposes

[Gerszten et al., 2008]. Simulation studies of theoretical biomarkers with a range of correlations to traditional risk factors have demonstrated that discriminatory power is not significantly improved when biomarkers are highly correlated with traditional risk factors [Pepe and Thompson, 2000]. These observations suggest that there should be a high bar for including additional biomarkers in the prediction of incident ASCVD.

In particular, the addition of new risk factors, including biomarkers, into existing risk calculators for predicting incident ASCVD must clear two hurdles [Wilson et al., 2004]. First, they must demonstrate benefit in the context of the prior probability of disease. This means that they should improve upon the positive and negative predictive discriminatory value of the models within the population to which they will be applied. In addition, the measurement of such biomarkers must be robustly measurable and readily available. While this second criterion is necessary for the inclusion of a new biomarker into clinical practice, it should not limit the investigation of new potential risk factors.

2.3 Towards Data-Driven Causal Models

The methods for interrogation of how two variables are related are dependent on the kinds of data available and how they were collected. Ultimately, this also bounds the types of conclusions that can be drawn from these measures of association. Observation is a foundational tool for scientific inquiry and observational data can reveal statistical relationships between variables known as correlation. Two variables are correlated when the expected value of one is informed by the observation of another. Relationships of this nature can be enumerated as the coefficients of both linear and nonlinear equations. Most famous, for continuous variables with a linear association, is the Pearson correlation coefficient, ρ (Eq: 2.1), which describes the linear association between two variables, Y_1 and Y_2, with standard deviations σ_1 and σ_2.

$$\rho = \frac{\mathrm{Cov}(Y_1, Y_2)}{\sigma_1, \sigma_2} \tag{2.1}$$

Data for investigating the association between dichotomous variables, E and O, are best expressed

in a 2 x 2 table:

	O	$\bar{O}$
E	a	b
$\bar{E}$	c	d

Table 2.1: Generic 2 x 2 matrix.

In this context, the odds ratio (Eq: 2.2), is the corollary of the Pearson coefficient. In particular, the odds ratio is the ratio of O and $\bar{O}$ ("not O") for E and $\bar{E}$ respectively and represents the effect observing E has on the expected value of O:

$$OR_{EO} = \frac{a/b}{c/d} \tag{2.2}$$

Variables that are strongly correlated with another measure of interest are useful for both classification and prediction tasks. Classification is the grouping of objects into categories based on observable attributes, such as the diagnosis of a patient based on available clinical information. Prediction, as done with the PCE, refers to the assignment of a probability of a future event occurring based on currently available data. Reliable prediction requires the satisfaction of several assumptions including that the data from which the correlation was observed is drawn from the same distribution in which the prediction will be applied. In addition, both the data in which the correlation is observed and applied must be sufficiently large. In practice, it can be challenging or even impossible to satisfy these assumptions. Finally, the observed correlation between two variables may indicate the existence of a causal relationship and therefore useful for hypothesis generation or may add support to prior knowledge for such an association.

Associations, Confounding, and Selection Bias in Observational Data

Understanding the behavior of measures of association is facilitated by numerical examples, particularly when exploring counterintuitive results. The following 2 x 2 table will motivate the examples discussed in the remainder of this section.[1]

The marginal association measure for E and O can be computed from Table 2.2 according to

[1]The values used in the contingency tables are the same as proposed in [Simpson, 1951] while the discussed hypothetical examples associated with the tables have been altered to be more relevant to this dissertation.

	O	$\bar{O}$
E	20	20
$\bar{E}$	6	6

Table 2.2: Example 2 x 2 matrix.

Equation 2.2 as,

$$OR_{EO} = \frac{20/20}{6/6} = 1 \tag{2.3}$$

That OR_{EO} is exactly equal to one is a peculiarity of the numbers chosen by [Simpson, 1951] and should be interpreted as no association between E and O. The observation of only two variables is rarely sufficient for classification and prediction tasks. However, the introduction of additional covariates introduces the possibility of conflicting statistical findings that require expert knowledge of the system being modeled for proper interpretation. Consider the addition of a third variable, C. The following table is derived from Table 2.2 conditioned on a third variable C.

C

	O	$\bar{O}$
E	5	8
$\bar{E}$	3	4

Table 2.3: Example 2 x 2 matrix conditional on C.

That the conditional odds ratio, $OR_{XY|C}$, differs from the marginal odds ratio, OR_{EO}, is unsurprising. The explanation for such a finding could easily be explained as a case of confounding whereby the third variable, C, is also associated with the original two variables, E and O, or effect modification whereby E behaves differently in the two subpopulations stratified by C.

$$OR_{XY|C} = \frac{5/8}{3/4} = \frac{5}{6} \tag{2.4}$$

Using both Table 2.2, and Table 2.3 it is possible to compute a third 2 x 2 conditional table for the case of $\bar{C}$ where $C + \bar{C} = 1$ by definition.

	$\bar{C}$	
	O	$\bar{O}$
E	15	12
$\bar{E}$	3	2

Table 2.4: Example 2 x 2 matrix conditional on $\bar{C}$.

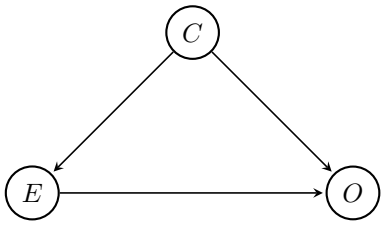

Figure 2.5: Directed acyclic graph with C as a confounder. The existence of a third variable C, that is associated with both E and O, interferes with (i.e., confounds) the ability to directly observe the effect E has on O.

From Table 2.4 is then possible to calculate an associated conditional probability on $\bar{C}$,

$$OR_{XY|\bar{C}} = \frac{15/12}{3/2} = \frac{5}{6} \tag{2.5}$$

Taken together with the prior calculated odds ratios, this is where the example posed by Simpson becomes particularly interesting. How can conditioning on either C or $\bar{C}$ produce equivalent odds ratios (5/6) that are different from the marginal odds ratio ($OR_{EO} = 1$) of Table 2.2? More importantly, which odds ratio is the correct one to report?

The only way to disentangle these seemingly conflicting results is to apply expert knowledge about the underlying system to decide which is the correct interpretation of the data. Adapted from [Hernán et al., 2011] are two possible model systems, diagramed as directed graphs in Figures 2.5 and 2.6.

These directed acyclic graphs represent two of the possible arrangements by which E, O, and C may be associated. In Figure 2.5, C is a confounding variable resulting in a spurious association measure between E and O. A hypothetical example of this arrangement would be the association of systolic blood pressure (E) with incident atherosclerotic cardiovascular disease (O). If only these

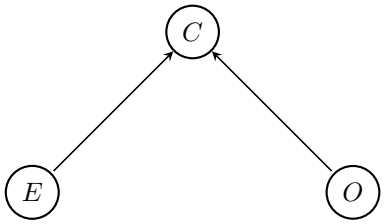

Figure 2.6: Directed acyclic graph with C as a collider. The existence of a third variable C, that both E and O are associated with, introduces a "back-door path" [Pearl, 2009]. The existence of a back-door path means that conditioning on C when evaluating the relationship between E and O will produce a spurious estimate of their association.

two variables are considered, it may appear that the total effect of systolic blood pressure on incident atherosclerotic cardiovascular disease is different from its true effect due to confounding by a third variable, age (C) [Pearl, 2022]. For the purpose of this discussion, it is reasonable to consider binary transformations of these variables, for example, normal or elevated blood pressure (e.g., systolic blood pressure less than 130 mmHg) and old age (e.g., $\geq$ 65 years old). In this case, it would be more appropriate to report the conditional rather than the marginal odds ratio as age is associated with both systolic blood pressure and atherosclerotic cardiovascular disease. The difference in the odds ratios ($1 - \frac{5}{6} = \frac{1}{6}$) is spurious. Instead, the true total effect of normal systolic blood pressure on incident atherosclerotic cardiovascular disease would be 5/6 and the marginal association of 1 would be incorrect.[2]

Alternatively, Figure 2.6 represents a model whereby C is a collider variable. In this case, both E and O are associated with C but not each other. The existence of this collider means that conditioning on C will reveal a spurious association between E and O. Broadly speaking, this is the structural model for selection bias [Hernán et al., 2004]. A classic example of this is the identification of a spurious association between two diseases among hospitalized patients [Berkson, 1946], where E and O and two unrelated diseases such as diabetes and cholecystitis. Differential rates of community prevalence and hospitalization for the two diseases lead to an association between them among hospitalized patients that do not generalize to the community population. Applying the odds ratio conditional on hospitalization to the general population could result in the errant practice of prophylactic gallbladder removal in patients with diabetes to prevent incident cholecystitis when in fact the marginal odds ratio is the appropriate one to consider given expert knowledge of the

[2]The use of Simpson's numbers for this example required the finding that normal blood pressure is protective (e.g., $OR < 1$) rather than the more traditional orientation of high blood pressure being a risk factor (e.g., $OR > 1$).

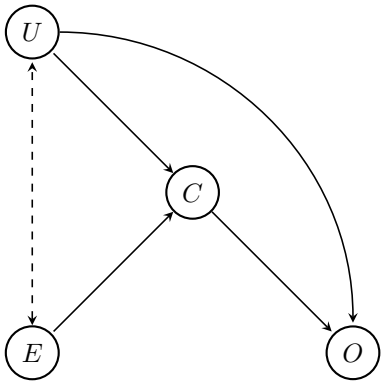

Figure 2.7: Unmeasured confounder, U, creates a back-door path from E to U to E, via C as a collider.

world.

The problem of colliders is not always as simple as graphed in 2.6. While they can occur due to arbitrarily complex systems of causal effects, they do not have to be much more complicated to reveal apparent paradoxes that require causal inference to tease apart. Figure 2.7 demonstrates a system whereby there is an unmeasured confounder U of both C and O. In this case, C is still a collider but one of the input arrows is a factor not previously considered by the study. This opens up a backdoor path from E to O via C and U ($E \rightarrow C \leftarrow U \rightarrow O$). This is precisely the arrangement that accounts for the paradoxical finding that maternal smoking is protective against mortality for low birth weight infants [Schisterman and Platt, 2011]. Nutrition is one of many plausible unmeasured risk factors that would both exert a causal effect on birth weight (C) and mortality (O) [Wilcox, 2001]. The existence of this confounder even when not measured or conditioned upon, given the causal structure that makes low birth weight a collider, will result in a spurious association between smoking E, observed in the data as protective, and infant mortality O.

These examples demonstrate some of the potential pitfalls of drawing conclusions from observational data and demonstrate the need for particular care when interpreting observed statistical relationships. In particular, it is necessary to understand the sources of data and the systems in which those data are generated [Hernán et al., 2002]. In addition to applying care in interpretation, it is important to consider how others might falsely extend observations, and the tendency to over-interpret correlation as causation.

Causality

Causality is a peculiar concept. It is both intuitively familiar to a toddler – throwing food on the floor *causes* the dog to come – and yet abstract enough that a parsimonious definition is elusive. What does it mean for one event to cause another? For Pearson, cause and effect starts and ends with the contingency table. However, this equates correlation with causation. No matter how consistent the observation of consecutive events is, regularity is not enough to assert causality unless it is also possible to intervene to change the outcome.

In the context of this work, the ability of event A to cause event B is defined by the existence of a counterfactual. That is, the natural history of event B would be different, but for the occurrence of event A. In this way, event A causes event B. This is a very practical paradigm for causation in the context of health care where clinicians and patients are routinely confronted with decisions of whether or not to proceed with an intervention with the premise that these interventions are expected to alter the natural course of human health. Causality in this sense also injects a certain level of agency for both clinicians and patients, who without it, would simply be observers of complex systems over which they have no control.

Randomized Controlled Trials

Randomized controlled trials are the gold standard method for testing causality experimentally. By design, these experiments elegantly eliminate confounding with the use of randomization. Fisher details the use of randomized experiments for this purpose in The Design of Experiments [Fısher, 1935]. He famously describes a Lady who asserts the ability to discriminate between tea that has been prepared by pouring milk into a cup of tea versus tea that has been poured into a cup of milk. The causal question at hand is, does the method of preparing tea affect its flavor in such a way that is discernable by the Lady? Fisher proposes two key components that must be considered to determine such a relationship experimentally. First, the Lady, well initiated to the task at hand (i.e., identifying how the tea has been prepared), is presented with cups of tea blinded to how they have been prepared in random order. Second, the experiment must be repeated sufficient times such that the result of the experiment is unlikely to be due to chance alone. Fisher suggests a 1 in 20 probability as a reasonable convention for the threshold, in the case of this experiment it corresponds to 8 cups of tea, four prepared in each of the two ways.

The principles of blind randomization with sufficient sample size as proposed by Fisher are cornerstones of modern randomized controlled trials. Randomized controlled trials have demonstrated causal relationships between statin therapy and the reduction of cardiovascular events [Scandinavian Simvastatin Survival Study Group, 1994, Nakamura et al., 2006]. The reduction in events conferred by statin therapy results in decreased mortality from those events [Shepherd et al., 1995]. This has been true in a variety of populations [Baigent et al., 2010, Collins et al., 2016], including those with [Intervention with Pravastatin in and others, 1998, Cannon et al., 2004, LaRosa et al., 2005, Pedersen et al., 2005] and without known ASCVD [Downs et al., 1998, Yusuf et al., 2016].

Mendelian Randomization

Despite the elegant design and success of randomized controlled trials in identifying causal relationships between pharmacotherapy and disease outcomes, they pose significant challenges and subsequent limitations. Results of randomized trials of relatively small populations will always suffer from questions of generalizability to the broader population. Selection bias potentially plagues comparisons between subpopulations in ways that are difficult to quantify. Many interesting questions do not lend themselves to randomized trials due to impracticality, immorality, or both. For these questions, the only available human *in vivo* data are observational and with careful creative integration of those observations, it is sometimes possible to make causal inferences.

One such method is instrumental variable analysis. Instrumental variables are those that cause a change in an explanatory variable and have no independent effect on a dependent variable [Angrist et al., 1996]. In other words, any change observed in the dependent variable observed in relation to the instrument is due to the instrument's effect on the explanatory variable. A source of instrumental variables and "naturally occuring randomized controlled trials" in observational health research is single nucleotide polymorphisms [Lawlor et al., 2008] (Figure 2.8). Another name for instrumental variable analysis using single nucleotide polymorphisms is Mendelian randomization.

Mendelian randomization leverages the random assignment of genes at conception to ensure that observed relationships between genetics and explanatory variables are free from confounding. The methods available for estimating the causal effect of exposure on outcome are dependent on the number of instrumental variables available and how they are related. All center around the ratio of

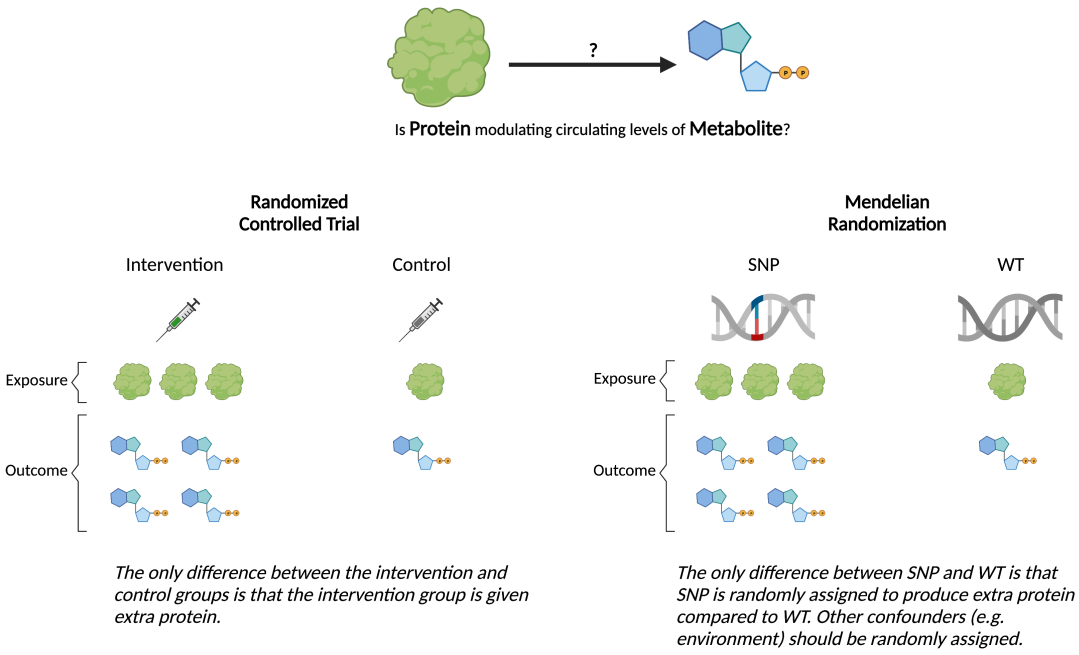

Figure 2.8: Mendelian randomization: naturally occurring randomized controlled trials. In a randomized controlled trial (left) designed to determine if a protein has a causal effect on the concentration of a metabolite, individuals would be recruited for the study and randomly assigned to receive additional protein (intervention syringe) or placebo (control syringe). If the outcome was significantly different in the intervention group, the conclusion would be made that the intervention caused the observed change. In comparison, with Mendelian randomization (right), instead of a randomly assigned intervention syringe, genetics assigned at birth (SNP) differentially exposed a random group of individuals to increased levels of the protein compared to wild-type (WT). If those same individuals have systematically different levels of metabolite, then it is reasonable to conclude that the increased protein is responsible for the observed difference. In this way, Mendelian randomization is analogous to a naturally occurring randomized controlled trial.

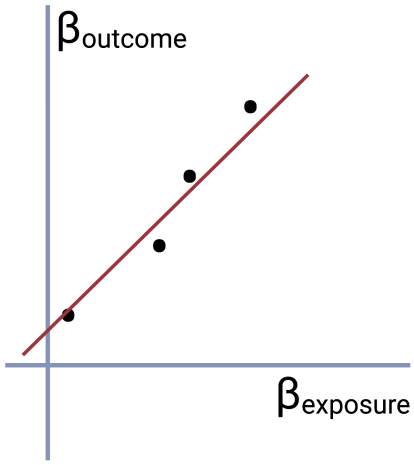

Figure 2.9: Mendelian randomization Beta-Beta plot. The causal estimate for the effect of the exposure on the outcome is the slope of the line formed by the Beta-Beta plot.

the instrumental variable effect size on the outcome compared to exposure.

With a single instrumental variable, the estimate is the ratio between the beta of the outcome and the beta of the exposure both in relation to the polymorphism (Eq: 2.6). This is equivalent to the slope of the line between the origin and the single point on a Beta-Beta plot.

$$\beta_{IV} = \frac{\beta_{Outcome}}{\beta_{Exposure}} \tag{2.6}$$

Additional uncorrelated instruments allow for a few additional methods to be considered. The median method removes outliers by looking at the central tendency of all available point estimates. Taking the median point is effective at reducing bias in estimation while sacrificing reduced variance. Reduction in the variance of the estimate is possible with regression analysis that includes all available valid instruments (Figure 2.9). The inverse-weighted method is a regression that weights point estimates with lower variance higher and uses the origin as a fixed intercept. Fixing the origin as the intercept is interpretable as including the data point that no change in exposure results in no change in the outcome.

A potential pitfall of including the origin is that it can mask bias in the rest of the data. For this

reason, some methods including MR-Egger allow the intercept to float at the expense of additional variance in the causal estimate [Bowden et al., 2015, Burgess and Thompson, 2017]. In the case where many instruments are available that are correlated with each other, two options are possible. Single nucleotide polymorphisms can be pruned sequentially to identify sentinel polymorphisms until the only remaining instruments are in linkage equilibrium [Lawlor et al., 2008]. Alternatively, all instruments can be used with a modified estimate that accounts for these correlations using a linkage disequilibrium matrix [Swerdlow et al., 2016, Burgess et al., 2016, Burgess et al., 2017b]. All of these methods are implemented in the MendelianRandomization R package [Yavorska and Burgess, 2017]. The combination of expert knowledge of the system being studied and Mendelian randomization methods are a powerful combination for the identification of causal associations within genomic-linked observational data.

2.4 Summary

This chapter provided a theoretical background for the following three research chapters. The first section reviewed the history of the epidemiology of cardiovascular diseases. This included a 75-year literature review of the first evidence of each traditional risk factor that informs current risk-based guidelines for managing cholesterol. Included in these risk factors are now commonly measured biomarkers such as total and HDL cholesterol. The next section of this chapter reviewed the biochemistry of atherosclerosis and the motivation for identifying additional serum biomarkers of incident ASCVD. Finally, this chapter ended with a theoretical discussion of causality. It reviewed fundamental considerations when making causal inferences from the kinds of observational data that are the focus of this dissertation.

Chapter 3

Health Information Exchange Powered Analysis of Population Cardiovascular Health

Specific Aim 1 of this dissertation is to characterize population ASCVD risk and primary prevention practices using health information exchange (Figure 3.1). This aim hypothesizes that health information exchange (HIE) aggregates cardiovascular disease risk factors that inform primary prevention care decisions for individual patients that would otherwise be missed by a single healthcare institution. This chapter presents the work that has been completed toward this goal. The first section (3.1) contains three papers that have previously been presented at the American Medical Informatics Association (AMIA) Annual Symposium (i.e., 2017, 2020, and 2021). They each develop natural language processing methods for information retrieval and information extraction from electronic health records (EHR). The first (3.1.1) classifies transthoracic echocardiography (TTE) imaging orders according to appropriate use guidelines. The second (3.1.2) presents a generalizable method for determining section boundaries in clinical notes to contextualize information derived from natural language by taking advantage of structure within "unstructured data" using a hidden Markov model. The final information extraction paper (3.1.3) builds on the first two motivated by medical

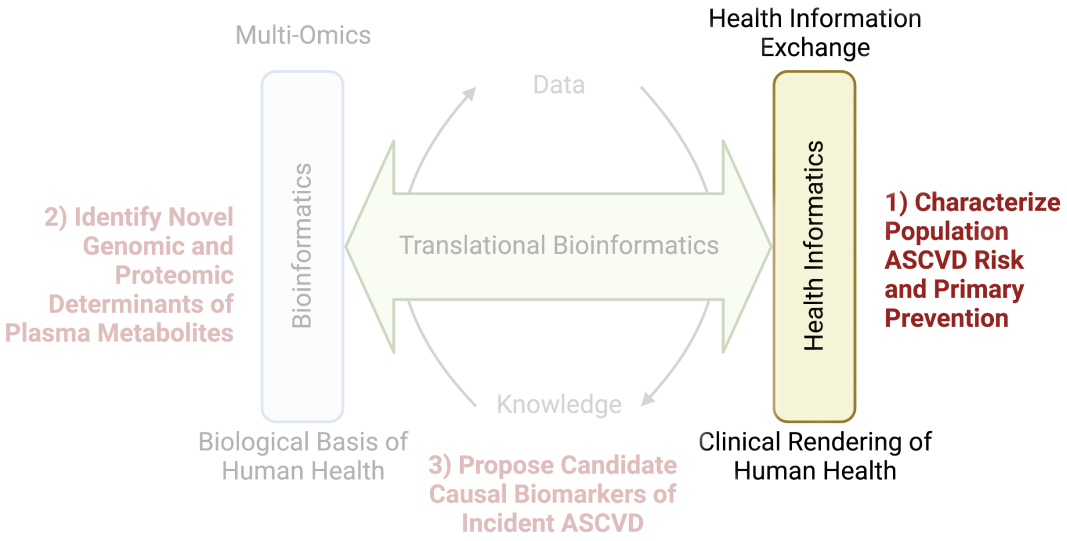

Figure 3.1: Overview of Specific Aim 1

knowledge (i.e., that "typical" chest pain portends cardiac ischemia) and the ability to identify note sections. It uses a large language model classifier to extract chest pain symptoms from the *History of Present Illness* section of clinical notes. While these papers did not directly contribute to the characterization of population ASCVD in this dissertation, they represent advances in methods and their applications that could be applied to large sets of clinical notes in the future that could inform the assessment of cardiovascular risk. At the time of writing this dissertation, clinical notes were not available as part of the Rhode Island (RI) HIE. The second section (3.2) directly addresses Specific Aim 1 by investigating structured data within the RI HIE. It begins with the standardization of the HIE to a common data model (CDM) and then continues with the extraction of known ASCVD risk factors for individual patients across time and care locations. These data are aggregated to estimate longitudinal 10-year ASCVD risk and adherence to primary prevention practices with statins according to current cholesterol-lowering guidelines. In total, this chapter presents both a current assessment of ASCVD risk and primary prevention practices in RI and methods that may allow for further elucidation of population ASCVD risk in the future using clinical notes.

3.1 Electronic Health Record Information Retrieval to Evaluate Clinical Practices

The electronic health record is the digital ledger of patient interaction with the health system. Within the electronic health record (EHR), are information about a patient's conditions, exposures to drugs, and procedures. Laboratory, imaging, anatomical, and vital signs measurements are also stored within the EHR. This information is organized in structured format as continuous, categorical, and binary data points using standard codes as well as in unstructured interpretive free text in clinical notes and reports. The constellation of these data for a single patient represents a clinical rendering of that individual's health. In aggregate and through a clinical healthcare lens, these data describe the functioning of the healthcare system and are a window into the health and healthcare of a population.

3.1.1 Categorizing Transthoracic Echocardiography Indications

A version of this section has been previously presented at the 2017 American Medical Informatics Association Symposium and published in the associated conference proceedings [Eisman et al., 2017].

Medicare alone spends more than one billion dollars annually on transthoracic echocardiography (TTE) and the per capita growth rate in the performance of TTE increased steadily by 8% per year from the mid 1990s to early 2000s, a rate that exceeded any plausable increase in commensurate clinical value [MedPAC, 2014, Levitt et al., 2016]. Transthoracic echocardiography is a powerful, non-invasive, low-cost imaging tool for evaluating cardiac structure and function. Trends towards evidence-based, targeted diagnostic testing highlight the potential for optimization of TTE utilization. In response to this trend, a joint task force published a set of consensus echocardiography appropriate use criteria (EAUC) to define indications for TTE in 2007 and revised criteria were published in 2011 [American College of Cardiology Foundation Appropriate Use Criteria Task, Force et al., 2011]. In the years since the EAUC were revised their use has been a component of echocardiography laboratory accreditation, necessary for insurance reimbursement, and also a way to promote quality improvement. However, the guidelines for how EAUC should be used are non-specific, only requiring that labs be aware of the EAUC, have access to them, and make concerted effort to improve the rates of appropriate testing while maintaining the caveat that the EAUC are "expert opinion"

that are non-exhaustive in scope [Picard et al., 2011]. Changes to prior authorization requirements implemented with the Affordable Care Act were in part responsible for a seventeen percent decline in spending between 2009 and 2013, however, there is little evidence that this decline resulted in a higher rate of appropriate testing and there still exists an eight-and-a-half fold variation in testing rates by geographic region in the United States [Singh and Ward, 2016, Fonseca et al., 2015]. All of this suggests the need to characterize physician TTE ordering practices to improve rates of appropriate clinical testing.

There have been concerted efforts to characterize TTE ordering behavior using the EAUC since they were published [Bhatia et al., 2012]. These studies have revealed insights into the rates of inappropriate testing and variability across centers, as well as the relative effectiveness of different interventional methods for impacting physician behavior and reducing rates of inappropriate testing. However, these studies have been limited by the need for manual classification of TTE appropriateness. This is a labor intensive process which is time consuming and limits the sample size available for analysis. The largest appropriate use TTE study consisted of approximately 3,000 TTEs classified in a manual fashion [Dudzinski et al., 2016].

To date, transthoracic echocardiography appropriate use studies have relied on manual annotation of patient medical records to determine EAUC classifications; however, natural language processing (NLP) has been successfully applied to extract information from the medical record in other related ways. For example, one study analyzed the medical record for necessary information required to calculate the Framingham Risk Score [Byrd et al., 2014]. Others have developed methods to extract clinical findings from echocardiography reports [Nath et al., 2016a, Toepfer et al., 2015]. Tools like these have been shown to be able to be applied across health systems demonstrating their broad utility beyond the electronic health record system within which they were developed [Carroll et al., 2012]. Applying such technology to the study of the appropriate use of TTE could have far-reaching implications as it would allow for the classification of TTEs on a much larger scale and provide the opportunity to more readily study clinical practice patterns and the impact of educational and other interventions. This study aimed to develop an automated rule-based method for processing the "indication" listed in a TTE report and classifying it into one of the broad, major EAUC categories. The technique was evaluated relative to a manually annotated reference standard derived from a

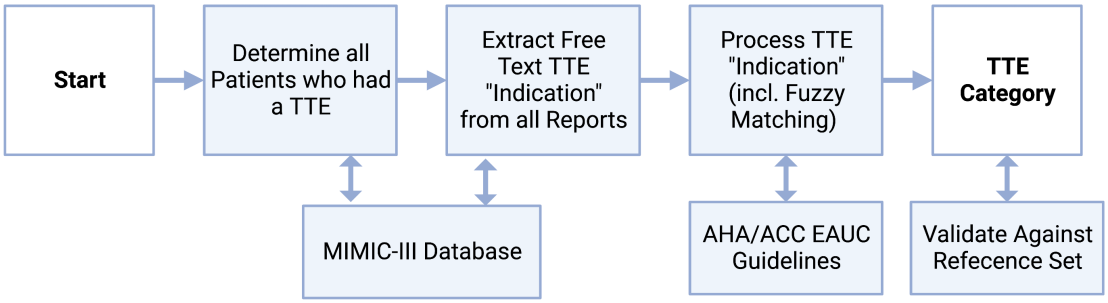

Figure 3.2: Overview of transthoracic echocardiography indication study. All patients who received a transthoracic echocardiogram (TTE) within the MIMIC-III database were included in the study. The indication field was extracted from each TTE report. The indications for TTE were processed as described including the use of fuzzy matching and classifying according to the American Heart Association (AHA) / American College of Cardiology (ACC) appropriate use criteria (EAUC) [American College of Cardiology Foundation Appropriate Use Criteria Task, Force et al., 2011]. Classifier performance was evaluated against a manually adjudicated reference set.

publicly available dataset.

Method for Developing a System to Classify TTE Orders

Transthoracic echocardiography indications were extracted from a publically available dataset. The indication text was processed and then matched against a thesaurus of terms to determine the EAUC indication category for the TTE. This method was validated on a sample of expertly adjudicated indication categorizations and then applied to the entire dataset. Most of the processing was implemented using the Julia v.0.5 [Bezanson et al., 2012] programming language, with statistical calculations done using R v3.3.1 [Team and Others, 2018].

MIMIC-III Database

The Medical Information Mart for Intensive Care III (MIMIC-III) is a publicly available de-identified database of critical care unit patient health information from a tertiary care hospital [Johnson et al., 2016]. MIMIC-III data range from provider notes and charts to imaging reports and survival information on approximately 60,000 critical care unit admissions at Beth Israel Deaconess Medical Center from 2001 to 2012. This study focused on patients who had at least one transthoracic echocardiography report available in the “noteevents” table and associated demographic information

from the "patients" table. TTEs were identified using the "category" column from "noteevents" where values containing "Echo" AND either the phrase "TTE (Complete)" OR "TTE(Complete)" were selected for analysis. Free text indications for performing each TTE were extracted from the reports as the text between the standardized header "Indication:" and special character marking the end of the line "\n."

Manual Classification of TTEs

A reference standard was developed to evaluate the performance of the automated classification program. Using the EAUC and specific annotation guidelines (shown in Table 3.1), two cardiology fellows with prior experience researching TTE ordering practices performed independent classification of a random sample of TTE indications. Three overlapping training rounds were used to ensure consistency between reviewers. Discrepancies were discussed and adjudicated by a cardiology attending physician, and the annotation guidelines were updated after each training round to reflect consensus classification and a Cohen's Kappa statistic was calculated. Once the training rounds were performed, each reviewer independently classified 450 TTEs to complete the reference set (n = 1200).

Automated Classification of TTEs

A rule-based computer program was developed in the Julia programming language to automatically classify TTEs into the categories outlined in the EAUC for echocardiography. Phrases and acronyms likely to indicate TTE ordering reason were generated using a combination of sampling MIMIC-III TTE indications (different from the reference set), consulting clinical experts, and SNOMED-CT concepts. These phrases were then associated with EAUC indication categories and collectively make up a "TTE thesaurus" of 355 phrase-category associations, which were mapped to 179 clinical concepts. TTE indications were pre-processed and compared to each phrase in the thesaurus using a string matching algorithm, which were consequently mapped to one of the clinical concepts contained in the thesaurus.

Pre-processing prior to string matching consisted of converting text to lowercase and removing non-alphabetic characters (except for hyphens, which are commonly used in clinical text). Finally, "stop

Annotation Classification	Scenario
(1) General	EAUC: Suspected Cardiac Etiology [or] Arrhythmias [or] Lightheadedness/Presyncope/Syncope [or] Evaluation of Ventricular Function [or] Perioperative Evaluation [or] Pulmonary Hypertension Additional Guidance: Ventricular Function as the only indication
(2) Acute	EAUC: Hypotension or Hemodynamic Instability [or] Myocardial Ischemia/Myocardial Infarction [or] Evaluation of Ventricular function after ACS [or] Respiratory Failure [or] Pulmonary Embolism [or] Cardiac Trauma Additional Guidance: Mention of Chest Pain [or] MI [or] Pulmonary Embolism
(3) Valvular	EAUC: Murmur or Click [or] Native Valvular Stenosis [or] Native Valvular Regurgitation [or] Prosthetic Valves [or] Infective Endocarditis Additional Guidance: Any mention of valve disease (incl. murmur) [or] Endocarditis
(4) Cardiac Structure	EAUC: Suspected Cardiac Mass [or] Source of Embolism [or] Pericardial Effusion [or] Guidance for Percutaneous Coronary Intervention Additional Guidance: Stroke [or] Pericardial Effusion [or] TIA/Stroke/Embolism
(5) Aortic	EAUC: Evaluation of the ascending aorta
(6) Hypertension / Heart Failure / Cardiomyopathy	EAUC: Hypertension [or] Heart Failure [or] Device Evaluation [or] Ventricular Assist Device and Cardiac Transplantation [or] Cardiomyopathies Additional Guidance: Chemotherapy or other Cardiotoxic Agent [or] Myocarditis
(7) Adult Congenital Heart Disease	EAUC: Initial Evaluation or ongoing Surveillance of Adult Congenital Heart Disease
Cannot be Classified	Does not clearly fit into one of the categories outlined by the EAUC

Table 3.1: Transthoracic echocardiography annotation guidelines. Using tables 1-6 (pages 7-10) of the transthoracic echocardiography (TTE) appropriate use criteria guidelines [American College of Cardiology Foundation Appropriate Use Criteria Task, Force et al., 2011] and the TTE Indication from each TTE report, annotators classified each TTE into a single annotation classification.

words" (common indefinite articles and prepositions such as "the", "at", or "on") were also removed.

Processed TTE indications were then compared against the TTE thesaurus to determine likely matches (demonstrated in Table 3.2). For each processed TTE indication string, the matching algorithm employed a moving window of m words, where m is the number of words in the thesaurus phrase string (note that for the purposes of this study, an abbreviation is considered a "word"). The window was moved across the TTE indication phrase to compare the potential thesaurus phrases against substrings of the TTE indication. Comparisons of processed TTE strings to thesaurus substrings were done using Damerau-Levenshtein distance to evaluate the similarity of strings. The similarity threshold for declaring a match was dependent on the length of the strings being compared, with shorter strings requiring near-perfect similarly and longer strings allowing for some disagreement. A constraint was also imposed such that a match could not be declared unless the first character of each word in one string matched that of its positional counterpart in the other.

Sample TTE Indications	Processed String → Thesaurus Match	Classification
"Mitral stenosis. Aortic regurgitation."	"mitral stenosis" → Valvular "aortic regurgitation" → Valvular	Valvular
"?endocardtitis"	"endocarditis" → Valvular	Valvular
"NSTEMI Myocardial infarction. Shortness of breath."	"nstemi" → Acute "myocardial infarction" → Acute "shortness breath" → General	Acute
" Left ventricular function. Pericardial effusion."	"left ventricular function" → General "pericardial effusion" → Cardiac Structure	Cardiac Structure
"Evaluate for Left ventricular function / wall motion abnormalities / diastolic functiuon. Coronary artery disease. Hypertension. Diastolic Congestive heart failure."	"left ventricular function" → General "wall motion" → General "diastolic function" → General "coronary artery disease" → General "hypertension" → HTN/HF/CM "congestive heart failure" → HTN/HF/CM	HTN/HF/CM
"Left ventricular function. New onset tachycrdia / ekg changes."	"left ventricular function" → General "tachycardia" → General "ekg changes" → General	General

Table 3.2: Sample of transthoracic echocardiography classification. Typographical errors are highlighted. HTN/HF/CM - hypertension / heart failure / cardiomyopathy.

Transthoracic echocardiography indication matches were then reduced to distinct clinical concepts using the TTE thesaurus and the frequency of concept-category matches for each TTE indication was

tabulated. TTE indications were then assigned an EAUC-associated category based on the highest number of matches with the following exceptions. The General category was only chosen when no more precise category matches were found and due to the specificity of valvular indications, any mention of valve disease or endocarditis was annotated as "Valvular." The automated rule-based method for classifying TTE indications according to the EAUC was evaluated compared to the expert derived reference set. The level of agreement was quantified using Cohen's Kappa. Precision, recall, and F_1 were all calculated for each classification category. The program was then used to categorize all 30,320 TTE reports in MIMIC-III database.

Comparison Against a Machine Learning Classifier

The rule-based classification approach was compared to a random forest classifier. The target variable was the human-generated, adjudicated TTE classifications. The features were engineered using a bag-of-words approach on individual words from the indication text after the removal of "stop words" (same "stop words" as above). Term frequency-inverse document frequency (TF*IDF) was used for normalization. This yielded 603 sparse features. The 1200 adjudicated TTE classifications were split into training and test sets (80% and 20%, respectively). Five-fold cross-validation was used to select the following meta-parameters: the number of features to sample at each node, the proportion of training set to be sampled for each tree. All trials used 1000 trees.

Transthoracic Echocardiography Classification Findings

Data for 30,320 TTEs from 18,487 different patients were analyzed with a mean of 1.6 (95% CI: [1-2]) TTEs per patient. Less than half (38%) of the TTEs were performed inpatient. Patients had a mean age of 68 years (95% CI: [56-79]), 56% were female, with a mean height of 67 inches (95% CI: [64-70]), weighing 170 lbs (95% CI: [144-201]). The mean patient BMI was 26.8 (95% CI: [23.4 - 31.3]).

Classification validation results are reported in Table 3.3. Two physicians classified a sample of TTE indications according to the EAUC across three training rounds in order to ensure consistency between them. Once consistency was achieved they each independently classified an additional 450 TTE indications. Collectively, 1200 expertly classified TTE indications comprise a reference set. The automated method was then tested against a small sample of the expertly adjudicated

classifications before being applied to the balance of the reference set. Physician v. automated classification validation round discrepancies are shown in Table 3.4. Precision, recall, and F_1 were calculated for each class and are reported in Table 3.5.

	Comparison	Weighted Kappa
Training Round 1 (n = 100)	Physician v. Physician	0.76 (0.64 - 0.89)
Training Round 2 (n = 100)	Physician v. Physician	0.78 (0.65 - 0.90)
Training Round 3 (n = 100)	Physician v. Physician	0.84 (0.72 - 0.96)
Program Testing (n = 50)	Physician v. Computer	0.90 (0.79 - 1.00)
Validation Round (n = 1150)	Physician v. Computer	0.89 (0.86 - 0.91)

Table 3.3: Validation statistics of transthoracic echocardiography classification.

		Physician Classification							
		General	Acute	Valvular	Cardiac Structure	Aortic Disease	HTN, HF, or CM	ACHD	Cannot be Classified
Automated Classification	General	317	2	0	5	1	0	0	0
	Acute	30	126	0	8	0	12	0	0
	Valvular	4	7	325	7	1	14	0	0
	Cardiac Structure	8	5	0	111	1	3	0	0
	Aortic Disease	1	0	0	0	1	0	0	0
	HTN, HF, or CM	69	4	0	1	0	0	1	0
	ACHD	1	0	0	4	0	0	1	0
	Cannot be Classified	3	0	0	0	0	0	0	4

Table 3.4: Validation round discrepancies for transthoracic echocardiography classification. HTN - Hypertension, HF, Heart Failure, CM - Cardiomyopathy, ACHD - Adult Congenital Heart Disease

A random forest classifier was generated to benchmark the rule-based method. The final model used 1000 trees, randomly sampled 250 features for splitting at each node, had a maximum of five observations in leaf nodes and drew bootstrap samples from 70% of the training set for each tree. This model achieved a classification weighted Cohen's Kappa of 0.89 (0.83 - 0.95).

After the automated classification program was validated against the reference set, it was applied

	Precision	Recall	F_1
General	0.98	0.73	0.84
Acute	0.72	0.88	0.79
Valvular	0.91	1.00	0.95
Cardiac Structure	0.87	0.82	0.84
Aortic Disease	0.5	0.25	0.33
HTN / HF / CM	0.63	0.81	0.71
ACHD	0.17	1.00	0.29
Cannot be Categorized	0.57	1.00	0.73

Table 3.5: Performance statistics for transthoracic echocardiography classification. HTN - Hypertension, HF, Heart Failure, CM - Cardiomyopathy, ACHD - Adult Congenital Heart Disease.

to the entire MIMIC-III database. The frequency of each TTE category is reported in Table 3.6, showing that more than half of the TTEs were ordered for General (27%) or Valvular (30%) reasons. HTN, HF, or CM (17%), Acute (15%), and Cardiac Structure (10%) accounted for most of the remaining TTEs with Aortic Disease and ACHD at less than 1% combined. Less than 1% of the TTE indications could not be classified automatically.

To assess the effect of accepting fuzzy matches in the algorithm, the program was rerun to only accept exact matches. Out of the entire dataset of 30,320 TTE indications, there were 4,511 with missed thesaurus phrase matches when fuzzy matching was not implemented. This resulted in 364 changes to classification categories, indicating redundancy in the information available in the TTE indication despite typographical errors.

TTE Category	N (%)
(1) General	8334 (27%)
(2) Acute	4420 (15%)
(3) Valvular	8967 (30%)
(4) Cardiac Structure	3100 (10%)
(5) Aortic Disease	114 (<1%)
(7) ACHD	132 (<1%)
Cannot be Categorized	211 (<1%)

Table 3.6: Automated transthoracic echocardiography classifications.

Evaluation of Automated Transthoracic Echocardiography Classification

An automated rule-based method was developed in this study for classifying TTEs into one of seven categories outlined by the EAUC for echocardiography. The most common indication for TTE was Valvular assessment closely followed by General. Hypertension / Heart Failure / Cardiomyopathy

(HTN/HF/CM), Acute, and Cardiac Structure assessment each contributed more than ten percent within this patient population. The system performed at a comparable level as trained physicians, allowing the classification of more than 30,000 TTE indications from a public database in less than ten minutes. This task would have taken the same physicians more than 45 hours.

Validation testing elucidated that areas of disagreement between the computer program and reference set established by trained physicians were mostly limited to four out of 56 possible discrepancy types, accounting for more than 65% of discrepancies. Two of these included the computer selecting the more specific categories of Acute or HTN/HF/CM, whereas a trained physician selected the General category. The remaining common discrepancy areas included the computer selecting Acute or Valvular where the trained physician classified them as HTN/HF/CM. The tight clustering of discrepancies demonstrates known overlaps in the EAUC categories. For example, determining whether the assessment of ventricular function belongs in the General or HTN/HF/CM depends on whether or not there is an associated clinical suspicion of heart failure. This information is not always available from the TTE report indication and therefore would require additional associated clinical data to make this classification with certainty.

It was found that a machine learning classifier (built using Random Forests) performed comparably to the rule-based method. However, machine learning approaches may be difficult to fully interpret and may therefore be viewed as being less transparent (i.e., the rules for a particular classification are not readily discernible). In addition, rule-based approaches may be more computationally efficient than machine-learning approaches, which often require large training sets. For example, the Random Forest approach used in this study required 80% of the adjudicated data set for training. The success of the rule-based approach demonstrated in this study, suggests the potential scalability to the development of similar methods for use in other clinical contexts.

Applications for Automated Transthoracic Echocardiography Classification

Classification of transthoracic echocardiography indications according to the EAUC has previously involved relatively small sample sizes largely due to the time required for manual classification. Research to date has utilized the EAUC in order to track rates of appropriate testing [Bhatia et al., 2012] or as an outcome measure for quality improvement [Dudzinski et al., 2016, Bhatia et al., 2014].

These studies have reported that a significant number (15-25%) of TTEs are ordered inappropriately, with the majority of rarely appropriate TTEs being ordered in the outpatient setting [Bhatia et al., 2014, Ward, 2012]. Randomized control trials of various combinations of educational intervention, clinical decision support, and audit/feedback mechanisms have demonstrated the ability to improve rates of appropriate use with limited sustained effect [Bhatia et al., 2013]. Applying an automated tool such as the one developed in the current study would allow for the evaluation of the appropriate use of TTE to occur on a much larger scale. This would likely have several important clinical implications, including a more accurate characterization of current practice patterns in various settings (e.g., academic, community, and inpatient/outpatient) and determining the impact of educational and other interventions.

Limitations and Generalizability of Findings

The generalizability of the breakdown of TTE ordering indications reported in this study is limited by the patient population. All TTE data within the MIMIC-III database were either obtained during or subsequent to a hospitalization in which the patient received care in an ICU. Other limitations to this study are the function of the data being used and how the EAUC is written. TTE indications are brief notes that lack clinical context provided by access to a complete medical record. Accurately distinguishing between some overlapping indications would require this additional information when TTE indications are non-specific.

Future research will use this classification algorithm to guide automated EAUC classification using additional clinical data including physician notes, discharge summaries, and lab testing results to determine appropriateness. The development of a system to automatically determine TTE appropriateness according to the EAUC will enable sustainable audit/feedback intervention that can be applied to any electronic medical record system at the point of care. Finally, an automated tool may facilitate appropriate use tracking for individual laboratories and improve the ability to comply with standards set by accrediting bodies.

More than 30,000 TTEs were classified according to the EAUC for echocardiography using a computer program that was validated against trained physicians. The promising results suggest that there is potential for an automated approach to track the appropriate use of TTE, as well as guide

the development of systems for prospectively identifying when TTE use is recommended. Automated assessment of EAUC adherence may enable effective and sustainable intervention to improve appropriate use rates and efficiency in the application of diagnostic echocardiography in the care of patients.

3.1.2 Identifying Clinical Note Section Transitions

A version of this section has been previously presented at the 2021 American Medical Informatics Association Symposium and published in the associated conference proceedings [Eisman et al., 2021].

Natural language processing of clinical notes promises to elucidate nuances of clinical care including what brings a patient to interact with the healthcare system. As suggested in the previous section, this includes the reasons for ordering clinical imaging studies such as transthoracic echocardiography. The interpretation of physician intentions of clinical notes is potentially facilitated by the uncovering of *structure* within the *unstructured* free text. The motivation for this work is informed by knowledge of how clinicians are trained to document encounters such that other humans can efficiently interact with recorded information. It is conceivable that computer programs can be designed to take advantage of this structure as well.

The electronic health record (EHR) is a rich source of biomedical data. With the passing of the Health Information Technology for Economic and Clinical Health Act of 2009 to incentivize the adoption of EHRs in medical practice, penetration of systems with core functionality, including clinician notes, has gone from about 27% to almost 100% of health systems over the past twelve years [Everson et al., 2020]. There is significant potential to leverage these data for the advancement of biomedical science and clinical medicine.

EHR data can be broadly separated into two categories. Structured data are stored as standardized values or codes. These include vital signs, laboratory results, medications, demographic information, and problem lists, as well as diagnosis, procedure, and billing codes. The remainder of clinical data is stored in an unstructured form, such as in clinician narrative admission, discharge, and progress notes. Clinical notes contain a wealth of information that corroborates, contextualizes, and supplements what is available in a structured form. However, challenges surrounding the extraction

of information from clinical notes (e.g., using natural language processing [NLP] techniques) have limited their use in large studies.

A notable portion of the EHR consists of unstructured free-text data. The clinical note is deliberately organized to enable efficient information retrieval by human readers. Each note category includes a series of expected section types. For example, the possible sections of a clinical note can be broken down into subjective information, objective information, and assessment and plan [Jacobs, 2009]. Subjective information is gathered using the interview and includes *Medical History*, *Review of Systems*, *Medications*, and *Family and Social History*. Objective information is obtained by the clinician's in-person interaction with the patient or the ordering of diagnostic testing. In the clinical note, it comprises the *Physical Exam* and *Labs and Imaging* sections. The final section of *Assessment and Plan* represents clinicians' summative conclusions about everything that precedes it and is where the next steps for the patient's care are documented. Answers to individual clinical questions (e.g., the presence of chest pain), if present, can be reasonably expected to be confined to a small number of sections (e.g., Medical History or Review of Systems) [Eisman et al., 2020]. While it is possible to further subdivide these section definitions [Burke et al., 2015], in practice, those subdivisions are often combined inside a single paragraph or even within a single sentence (e.g., *Assessment and Plan*). Therefore, section identification at the level of granularity outlined is an effective first step to significantly increase the signal-to-noise ratio for downstream NLP tasks without significantly increasing section boundary ambiguity.

The method for note section identification for research purposes is almost exclusively performed using rule-based or machine learning approaches to identify institution-specific section headers along with formatting tokens like colons or bolded text that indicate the section header (e.g., "<b>Medications:<b>") [Dai et al., 2015]. Prior work that modeled the sequence of note sections headings using a hidden Markov model (HMM) found that the majority of notes (67%) in some clinical settings do not possess any section naming headers, and efforts to identify note sections were limited to the minority of notes that contained standardized section indicating tokens [Li et al., 2010]. Additional studies have reported high accuracy of granular section detection using section headers in notes to identify unlabeled sections [Denny et al., 2009]. While this has been shown to be effective, it is challenged when either non-standard headers are used, the headers do

not reflect the content that follows, or when seeking to develop generalizable approaches that can work across multiple clinical care settings with minimal training and configuration.

Generalizable approaches for identifying sections of the text using NLP tools include the identification of repeated words in adjacent text blocks [Hearst, 1994, Hearst, 1997] and the identification of lexical chains using a thesaurus to link together related concepts [Morris and Hirst, 1991]. The brevity of clinical notes and lack of redundancy make direct application of these techniques that rely on repeated words unlikely to work. The most successful projects to automatically segment clinical text have used a shotgun approach of attempting to identify section breaks based on whitespace and variations in section headers along with markers for header definition such as colons and bolded type with good but not perfect accuracy [Apostolova et al., 2009]. The success of this is highly dependent on the institution and EHR system-specific structure of clinical notes.

Inspired by these concepts, this study set out to use the Unified Medical Language System (UMLS) Metathesaurus to identify semantic concepts within sections of clinical text to characterize the distribution of substantive content of note sections [Caviedes and Cimino, 2004, Shivade et al., 2015]. UMLS concepts were chosen as a domain-appropriate knowledge structure to serve as a proxy for informative data contained within note sections. The concept-based semantic types further reduced data dimensionality such that they could be modeled as a Markov process. It was thus hypothesized that a heterogeneous HMM using semantic type distribution and section transition probabilities would provide a generalizable note section detection algorithm.

Developing a Generalizable System for Note Section Identification

Training and testing sets of admissions notes were randomly sampled from a publicly-accessible EHR dataset. The notes were then manually labeled for component sections based on content. These annotations were used to calculate the probability of transitioning from one section to another. The note text was also processed by the MetaMap NLP system to identify UMLS semantic types [Aronson, 2001]. The probabilities of observing the semantic types were calculated for each note section type. A subset of semantic types was selected for inclusion based on their ability to uniquely identify a note section type. Collectively, these defined a heterogeneous HMM that was trained and cross-validated on the first set of notes and tested on the second. The performance of the HMM

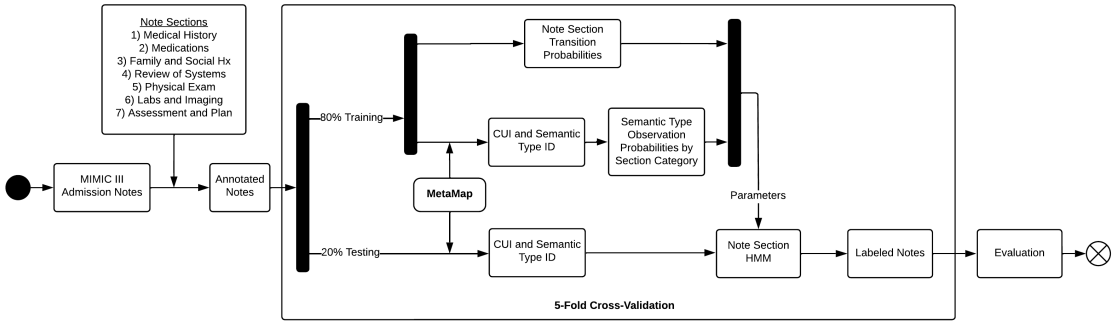

Figure 3.3: Overview of note section identification study. Admission notes extracted from the MIMIC-III database were annotated for one of seven note section types. Unified Medical Language System (UMLS) semantic types were identified using the National Library of Medicine's MetaMap natural language processing tool. Note section transition and semantic type observation probabilities from a training set defined a note section hidden Markov model and was tested with five-fold cross-validation on held-out sets of notes from the same corpus. Note section boundaries were evaluated for the percent of interesting content (i.e., UMLS semantic types) captured.

was evaluated and compared to a template to identify note sections. The overview of these steps is outlined in Figure 3.3.

MIMIC-III Database

Clinical notes from the Medical Information Mart for Intensive Care III (MIMIC-III) database were used for this analysis [Johnson et al., 2016]. MIMIC-III is a publicly available de-identified dataset of intensive care patients treated at Beth Israel Deaconess Medical Center from 2001 to 2012 and includes approximately 60,000 critical care admissions. A set of 100 notes that used the single most common admission note template was randomly selected as a training set. The single most common template was selected for training in order to allow out-of-sample evaluation of different note templates to demonstrate how sensitive the HMM method is to template changes. A testing set of 100 notes was selected that comprised a variety of templates. Many of these contained no relevant changes, and those that had changes often had differences in one or two headers (e.g., "HPI:" vs "History of Present Illness:").

Annotation of Clinical Note Sections

A reference standard was developed to train and test the automated identification of note sections.

Using written guidelines, two human annotators (KAB[1] and INS[2]) with clinical and informatics expertise independently identified note sections in the same two hundred clinical notes with standardized annotations at the beginning of new sections. Disagreement between the two annotators was adjudicated by a third annotator (ASE[3]). Up to seven different section types were identified in each note. (1) Medical History was defined as including the chief complaint, history of present illness, past medical history, and review of systems if continuous with the rest of medical history. (2) Medications also included allergies if the two appeared contiguously in the note. (3) Review of systems was labeled separately if not adjacent to a medical history section. (4) Family and social history were combined into a single section type due to the fact that they are sometimes combined into a single line of text making it impossible to distinguish them at a line-level granularity. (5) Physical exam was defined to include vital signs and other "flowsheet" information in addition to the standard physical exam. (6) Labs and Imaging including any objective testing information results included in the note. (7) Assessment and Plan was defined as a summative analysis of the patient's status at the time of note writing followed by a description of the next steps for medical care.

Identification of Unified Medical Language System Semantic Types with MetaMap

MetaMap is an NLP tool maintained by the National Library of Medicine for the processing of text for the identification of biology and medicine UMLS concepts. The full set of notes was processed using the MetaMap Docker image in batch. MetaMap was configured to use the NLMSubSyn vocabulary. UMLS concept unique identifiers (CUI) and their associated semantic types were tracked, along with their associated line number from the original note text file. Absolute and relative frequencies of semantic types identified within each section type were calculated for the notes in the training set. A subset of semantic types (n=24) was selected for the HMM that (1) occurred at least 1.5X more frequently in a single section type compared to the note as a whole, and (2) occurred in at least 90% of the sections of the type for which it is overrepresented.

Modeling Note Section Transitions as a Series of Hidden States

A hidden Markov model (HMM) is a simple dynamic Bayesian network of a Markov process with

[1]Katherine A. Brown
[2]Indra Neil Sarkar
[3]Aaron S. Eisman

underlying unobserved states that influence the observed output. There are three main parameters of an HMM. The first is an initial probability array, π, which contains the probability of starting in each of the possible unobserved states. The second is a state transition probability matrix, a, which contains the probability of changing from each possible unobserved state to another unobserved state (including remaining in the same state) after each observation. Finally, there is an observation probability matrix, b, that contains the probabilities of each observation for each state [Rabiner, 1989]. In the context of the task of note section identification, the note sections are the unobserved hidden states and the UMLS semantic types associated with MetaMap-identified UMLS CUIs are the observations. The model was constrained to only allow for the identification of a state transition after the end of a new line in the note. This was accomplished by having an identity matrix, a_2, that represents the probability of state changes between observations on a single line of the note text.

The parameters for the HMM were trained using the manually annotated clinical notes from the training set. The initial state probability array, π, was defined as the proportion of notes that started with each section type. State transition probabilities for a were empirically determined on a per-line basis. For each note, the transition probabilities were determined for all sections within that note with each new line being considered an opportunity for state transition. The arithmetic mean of each state transition probability was then calculated across all notes and normalized to sum to a total probability of one. Observation probabilities were also calculated within each note for each section. They were defined as the proportion of each semantic type divided by the total semantic types in each section. The observation probabilities matrix was then calculated as the mean of the probabilities calculated for each note.

Defining a Template Boundary Model for Evaluation

A template boundary model (TBM) was developed using the same most common admission note template applied in the training set. This model looked for the words or phrases that indicate the start of each of the corresponding note sections using regular expression matching of standardized section headers in Table 3.7. In the event that consecutive template sections were assigned to the same "mapped section," those were considered a single merged section. The TBM was also applied to the testing set. The results of this model were used to contextualize the evaluation of the HMM method and allow a comparison of each method's sensitivity to small changes in the template.

Template String	Mapped Section
Chief Complaint:	Medical History
HPI:	Medical History
Allergies:	Medications
Infusions:	Medications
Other ICU medications:	Medications
Other medications:	Medications
Past medical history:	Medical History
Family history:	Family and Social History
Social history:	Family and Social History
Review of systems:	Review of Systems
Flowsheet Data:	Physical Exam
Vital Signs:	Physical Exam
Labs / Radiology:	Labs and Imaging
Assessment and Plan:	Assessment and Plan

Table 3.7: Template boundary model.

Measuring the Quality of Note Section Identification

Five-fold cross-validation of the HMM method was performed by randomly separating the data into five equal bins. The model was trained on four of the bins (i.e., HMM parameters were empirically determined) and subsequently tested on the remaining bin. Using this process, each original note was labeled by the HMM exactly one time. This process was performed once on the training set. The content of a note section in the reference standard was defined as the line number occurrence of an annotated section up to but not including the line of the next label or the end of the document. The average of the parameters from the five-fold cross-validation was carried forward as a single model applied to the testing set.

The performance of the identified note section boundaries was measured by the sensitivity and specificity of the content those boundaries define. The model was evaluated as the number of UMLS concepts (e.g., a proxy for the interesting data) captured by the identified section boundaries. Concepts for a given section were considered true positives if they were within the manually annotated and model boundaries for the section. False positives were defined as concepts inside the model boundaries but outside the annotated ones. Concepts outside of the model boundaries but within the annotated boundaries were considered false negatives. UMLS concepts outside of both the model and annotated boundaries were considered true negatives. Full confusion matrices were calculated using these definitions for the HMM and the TBM for both the training and testing datasets and

then used to calculate sensitivity, specificity, precision, recall, and F_1-score.

Note Section Identification Findings

Two hundred clinical notes were manually annotated into 1,336 distinct sections based on the annotation guidelines and reference standard as described above. Overall, the notes were organized and formatted in a chronological fashion from the time prior to the note being written to planned future clinical care. The notes were formatted in the original dataset such that lines were limited in length to about twelve words. As a result, paragraphs of text and even individual sentences were often interrupted by a newline character. In addition, many parts of the notes did not contain complete sentences. For example, lists were commonly employed to record medications, symptoms, and medical history not directly pertaining to the chief complaint. Each note typically began with a history of present illness and medical history. This was followed by varying combinations of *Medications*, *Review of Systems*, *Family and Social History*, *Physical Exam*, and *Labs and Imaging* results. Notes almost always ended with an assessment of the patient's status and a plan going forward, including medications, testing, and follow-up care. Most of these sections contained discrete labels within the text.

There was variation in how sections were labeled, and at times they were entirely mislabeled. For example, the most commonly mislabeled sections included *Medical History*, and *Family and Social History*. These section headers were often listed out, and then the content for all three sections was merged together in the lines that followed the header labels. Another less frequent example included content from distinct sections that were embedded within one another (e.g., home medication list followed by medical history/surgical history, and ending with a list of more or duplicate home medications). Other observations throughout manual annotation included extra text that did not pertain to a particular section or additional notes as an addendum added at the end that included brief summaries of *Medical History*, *Physical Exam*, and *Assessment and Plan* written by another clinician (e.g., specialist, attending, or medical student). The annotators labeled sections based on content and not the section headers. If a section header corresponded to the following content, it was marked as the start of the section. As expected, it was evident that the unique characteristics, language, content, length, and formatting of the clinical note varied based on individual provider and specialty.

Note Section Transition Probabilities

A transition state probability matrix was calculated using training data for each fold in the cross-validation. This was repeated five times. Mean transition probabilities after a newline character are diagrammed as a Markov process in Figure 3.4a. For all sections, it was found to be most common to stay within that section. Beyond that, *Medical History* can transition into *Medications* or *Family and Social History*. *Family and Social History* can be followed by *Medications* or *Physical Exam*. The remaining sections of *Review of Systems*, *Physical Exam*, *Labs and Imaging*, and *Assessment and Plan* always follow in sequential order. An example portion of the heterogeneous HMM, including both observations and hidden states, is shown in Figure 3.4b. Observations are UMLS CUIs converted to UMLS semantic types (e.g., Metronidazole → Antibiotic). State transitions are only allowed between observations separated by a newline character with probabilities shown in Figure 3.4a. Between observations without a newline character, the hidden state is maintained. The most likely path is determined using expectation maximization. A hypothetical example of this is bolded in Figure 3.4b as Metformin → Pharmacologic Substance (Medications) to Tobacco → Plant (Family and Social History). The most common progression of note sections is visualized as an alluvial diagram in Figure 3.4c. One-half of the notes followed the most common section progression of Medical History, Medications, additional Medical History, Family and Social History, Physical Exam, Labs and Imaging, and then Assessment and Plan. About one-quarter of the notes also had a Review of Systems section, most commonly between Family and Social History and Physical Exam.

Semantic Type Feature Distribution

The 24 UMLS semantic type features were calculated for all 1,366 note-sections in the reference standard. Semantic type distributions by note section type are depicted in Figure 3.5. The analysis of semantic type distribution represented by different sections was in line with expectations. *Medications*, *Family and Social History*, *Physical Exam*, *Labs and Imaging*, and *Review of Systems* all contain distinct information represented by specialized language that was expected to generate unique probability distributions.

It was found that more than two-thirds of the identified concepts within the *Medications* sections were “Pharmacologic Substance”, “Organic Chemical”, or “Antibiotic”. *Family and Social History*

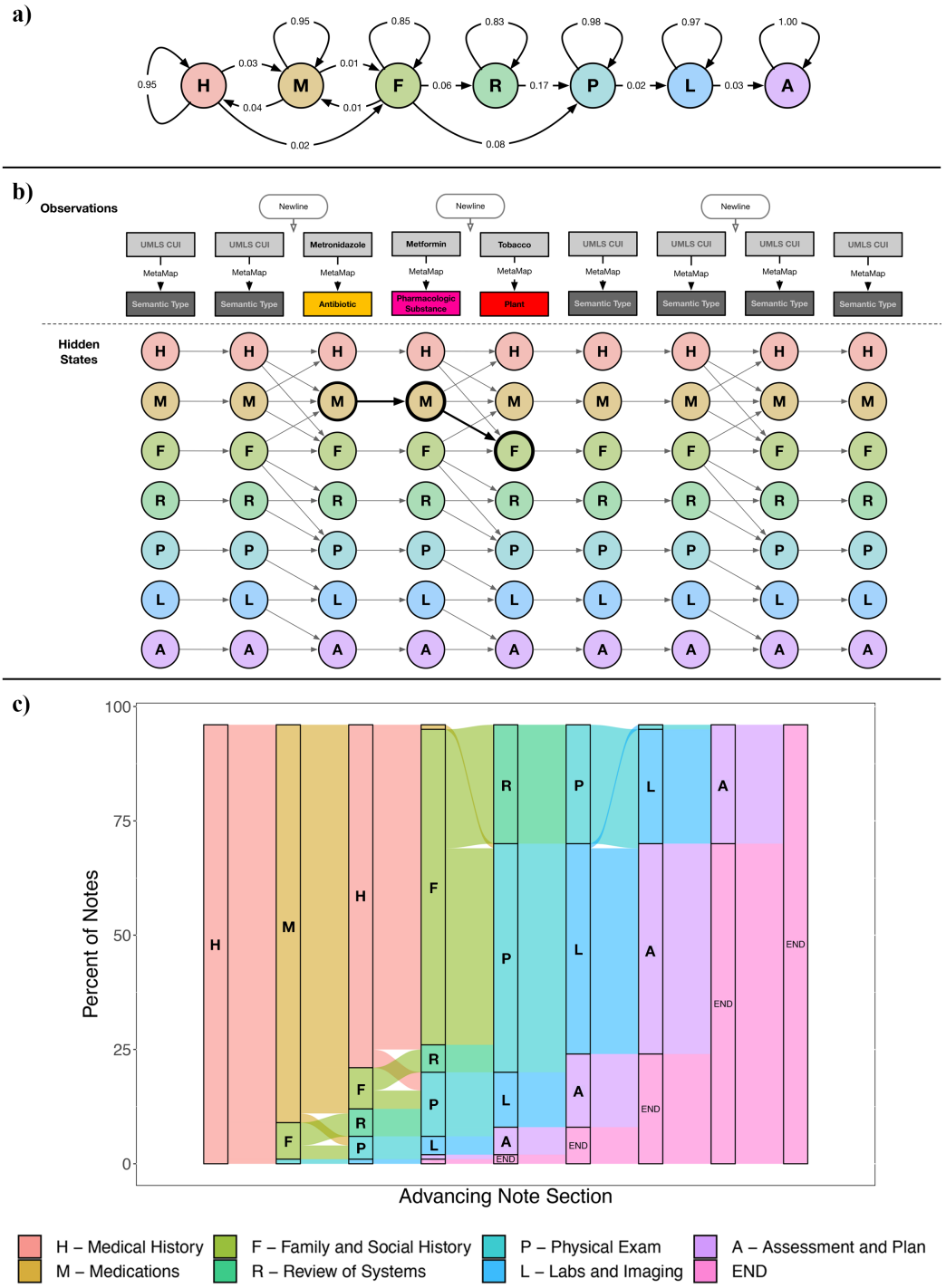

Figure 3.4: Note section transitions. a) Note section transitions diagramed as a Markov process with nodes as each section type and edges specifying section transition probabilities, b) Illustrative example of note section transitions as a heterogeneous hidden Markov model, c) Observed note section transition distributions as an alluvial diagram.

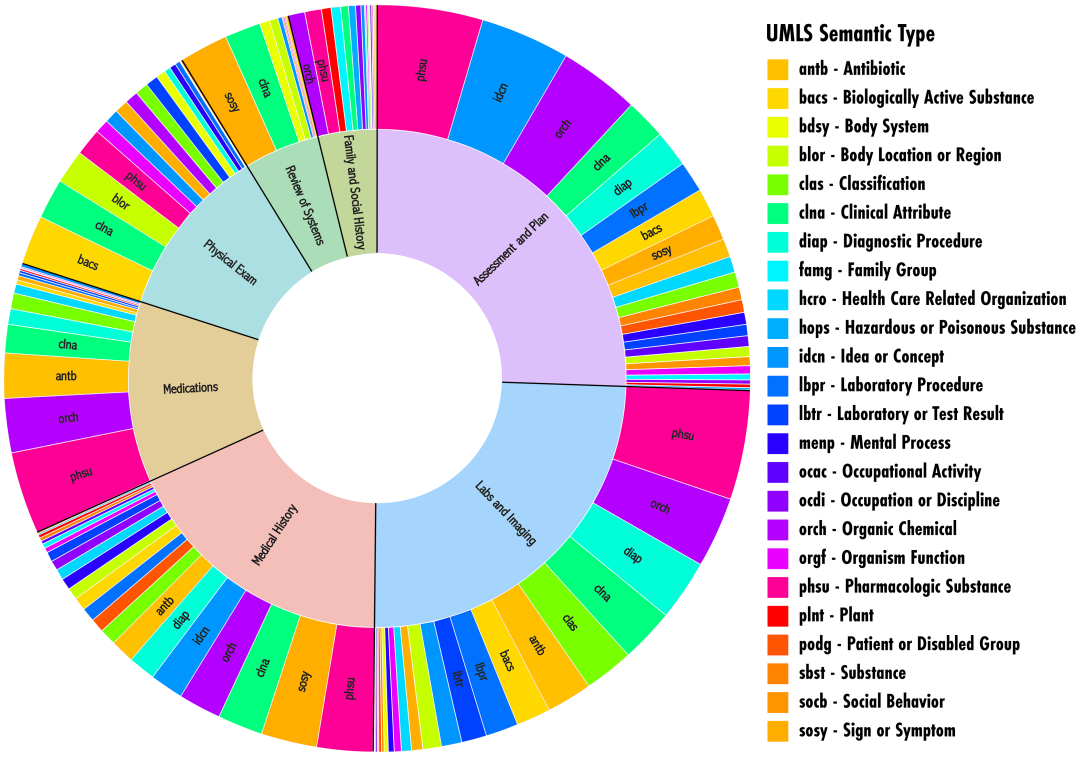

Figure 3.5: Distribution of UMLS semantic types by note section.

was dominated by "Organic Chemical", Pharmacologic Substance", and "Plant", all referring to substance use habits (e.g., Tobacco). In addition, "Family Group" is significantly overrepresented, likely reflecting the common format for family history (e.g., "Father had a myocardial infarction at 64 years old") and description of living arrangements, marital, and family status that is commonly included in *Social History*. *Review of Systems* is expectedly overweight in "Body System" and "Sign or Symptom". Similarly, an outsized proportion of "Body Location or Region" within the *Physical Exam* sections is indicative of the kinds of notations that are made about the physical exam, which represents in-person observations about a patient's body. Finally, the ambiguity presented by *Medical History* and *Assessment and Plan* was an expected challenge, but the latter includes an expectedly higher proportion of "Idea or Concept" words (e.g., "probably" and "consistent with") that reflect the author's thought process when assessing the patient's condition. In addition, differentiating these two sections is accounted for by the transition probabilities matrix of the HMM.

Evaluation of Hidden Markov Model Identification of Note Section Transitions

Both the HMM and TBM were used to identify note sections in two-hundred admission notes. The training was performed on one-hundred notes that used the single most common template within the MIMIC-III database. The performance of the HMM on these notes compared to the TBM is reported as five-fold cross-validation of the HMM in Table 3.8. Overall, both models performed well. The HMM captured greater than 85% of the semantic types in all sections, including better than 92% in Review of Systems, Physical Exam, and Assessment and Plan with the exception of Labs and Imaging, which performed relatively poorly (70%). The specificity of the HMM was at least 93% across all note section types. The HMM significantly outperformed the TBM in identifying Medical History concepts (88% vs. 75%) and significantly underperformed the TBM in identifying Labs and Imaging concepts (70% vs. 99%). The HMM modestly underperformed the TBM ($\Delta \leq 10\%$) in the remaining note section types. Specificity was excellent for both and only significantly differed for two section types where the HMM outperformed the TBM for Family and Social History (99% vs. 93%) and underperformed the TBM for Assessment and Plan (93% vs. 100%).

The performance of the HMM compared to the TBM on the testing set is reported in Table 3.9. The HMM achieved a sensitivity of at least 77% across all note section categories. It performed at 84% for Medical History and Review of Systems and at least 93% for *Physical Exam* and *Assessment and*

	Specificity			Precision			Recall/Sensitivity			F_1		
	HMM	TBM	Δ	HMM	TBM	Δ	HMM	TBM	Δ	HMM	TBM	Δ
H	0.98	0.97	0.01	0.94	0.90	0.04	0.88	0.75	0.13	0.91	0.82	0.09
M	0.99	1.00	-0.01	0.61	0.95	-0.34	0.85	0.92	-0.07	0.83	0.94	-0.11
F	0.99	0.93	0.06	0.81	0.28	0.53	0.85	0.95	-0.10	0.83	0.43	0.40
R	0.99	0.99	0.00	0.67	0.75	-0.08	0.92	1.00	-0.08	0.77	0.85	-0.08
P	1.00	1.00	0.00	0.98	1.00	-0.02	0.92	0.99	-0.07	0.95	0.99	-0.04
L	0.99	1.00	-0.01	0.86	0.99	-0.13	0.70	0.99	-0.29	0.77	0.99	-0.22
A	0.93	0.99	-0.06	0.87	0.98	-0.11	0.95	1.00	-0.05	0.91	0.99	-0.08

Table 3.8: Five-fold cross-validation (100 single template notes). H - Medical History, M - Medications, F - Family and Social History, R - Review of Systems, P - Physical Exam, L - Labs and Imaging, A - Assessment and Plan.

Plan. The sensitivity of the HMM outperformed the TBM in every note section category except *Family and Social History*; however, this was achieved by the TBM with only 20% precision. The HMM achieved very high specificity ($\geq 94\%$) for all note section types, comparable to the TBM.

Between the training and testing sets, the HMM had an average sensitivity performance reduction of less than 3% per section type compared to a 14% degradation in the TBM. For the HMM, no single section degraded more than 8% (Family and Social History and Review of Systems), and the worst-performing section for the HMM in the testing set, Labs and Imaging, improved by 7%. Overall, the HMM performance was consistent between training and testing sets when compared to the TBM, where every section had reduced sensitivity of at least 10%, with the exception of Medical History, which was nearly unchanged.

Admission notes from MIMIC-III represent real-world examples of clinical notes for ICU patients.

	Specificity			Precision			Recall/Sensitivity			F_1		
	HMM	TBM	Δ	HMM	TBM	Δ	HMM	TBM	Δ	HMM	TBM	Δ
H	0.96	0.94	0.02	0.86	0.78	0.08	0.84	0.74	0.10	0.85	0.76	0.09
M	0.96	0.99	-0.03	0.61	0.86	-0.25	0.79	0.66	0.13	0.69	0.75	-0.06
F	0.99	0.91	0.08	0.69	0.20	0.49	0.77	0.86	-0.09	0.72	0.32	0.40
R	0.99	0.99	0.00	0.74	0.78	-0.04	0.84	0.83	0.01	0.78	0.81	-0.03
P	0.99	1.00	-0.01	0.95	1.00	-0.05	0.93	0.87	0.06	0.94	0.93	0.01
L	0.98	1.00	-0.02	0.83	0.95	-0.12	0.77	0.72	0.05	0.80	0.82	-0.02
A	0.94	0.99	-0.05	0.91	0.98	-0.07	0.94	0.86	0.08	0.92	0.92	0.00

Table 3.9: Out of sample testing (100 with Minor Template Modifications). H - Medical History, M - Medications, F - Family and Social History, R - Review of Systems, P - Physical Exam, L - Labs and Imaging, A - Assessment and Plan.

The manual annotation process clearly demonstrated that while these notes possessed the expected structure that domain experience suggested, they were far from ideal. The clinical note text was commonly untidy and poorly arranged. For example, many of the *Labs and Imaging* sections were intended to be in a tabular format but were generally misaligned. In addition, some sections were routinely mislabeled or labeled without any relevant content to follow. In addition, there were frequent line breaks that interrupted sentences and paragraphs. These are likely artifacts of data storage processing and the result of irregular use of note templates. All of these findings reestablished the importance of note section detection informed by content rather than formatting tokens.

It is important to acknowledge that the results presented here represent a single note type tested from a single institution. Furthermore, all patients within MIMIC III were treated in an ICU, which likely both increases the complexity of the recorded information in the associated clinical notes but also necessitates structure. Significantly shorter notes that contain no underlying structure are unlikely to benefit from section detection using the methods described in this study.

The HMM method performed favorably to the TBM in several important ways. The goal of this comparison was to evaluate the model's robustness to support portability to changes in templates over time or to other electronic health record systems. The most dramatic example was with *Family and Social History*. In this dataset, while the headings were consistently used, they frequently did not appropriately precede relevant content. Therefore, while the TBM appropriately found the *Family and Social History* section label, the section itself was essentially mislabeled. *Family and Social History* are particularly important as they represent the most basic form of genetic information and can impact guideline-based clinical care decisions that rely on risk stratification [Hughes Halbert et al., 2016]. The selected evaluation criteria account for these errors by measuring the amount of likely relevant content (i.e., the number of UMLS concepts) captured. The result of the findings presented is that the identification of *Family and Social History* information had very poor precision when using a method that looks for section header labels. Additionally, the HMM performed consistently from training to testing sets compared to TBM, indicating that it may be a more portable method for note section detection across different templates and note types.

When considering the NLP tasks that are downstream of note section identification, the HMM results are particularly promising. Two of the three note section types for which the HMM performed the

worst are for data that are relatively low yield in clinical notes. Both *Medications* and *Labs and Imaging* data within notes are often copied from structured sources or dedicated reports. These data are likely more useful in the structured form from which they were derived than extracted from clinical notes.

Despite these findings, the methods presented have significant room for improvement, and analyzing common errors can inform future model development. Improved detection of candidate section transitions would improve the accuracy of the model by decreasing the number of state change decision points. This could be accomplished with additional text preprocessing to, for example, reformat the notes so that narrative text that is broken up into multiple lines intra-paragraph and intra-sentence is considered as a single block inside which transitioning of states cannot occur. In addition, staging the identification of sections by starting with an HMM that only considers the three categories of "Subjective", "Objective", and "Assessment and Plan," and then further processing each of those into component sections may improve performance.

The promising results of this study suggest that clinical note sections can be identified using an HMM based on UMLS semantic types associated with MetaMap-identified UMLS CUIs. The relatively standard organization of clinical notes and the expected information contained within each section are both the motivation for identifying clinical text sections and the reason why these methods worked. It is likely that these methods would be portable across EHR systems, as preliminary testing on an additional corpus of publicly available notes [MTSamples, 2020] demonstrates better performance than reported here. The ability to identify note sections across EHR systems and medical specialties using minimal training data in this way would enable the development of tools to aggregate clinical data about an individual patient across health systems. For example, the aggregation of sparsely recorded information like *Family and Social History* could be collected and synthesized into a single best summary for a patient. This information may be more likely to be recorded by a primary care physician but could then be made available to physicians during all of a patient's clinical encounters. In addition, the note section contains contextual information about the meaning of the underlying information. Most importantly, distinguishing between *Medical History* and *Assessment and Plan* is needed in order to assemble a chronology of a patient's clinical information. For example, a medication that appears as part of a standard Medications section is

generally confirmed as active therapy, while a medication that appears in the *Assessment and Plan* could go unfilled.

The hidden Markov model and results presented are the first steps in the development of a generalizable method for the identification of sections based on content in clinical notes. The findings demonstrate that semantic types associated with MetaMap-identified UMLS CUIs differ among clinical note sections, and the probability of observing consecutive content within notes can be used to determine note sections. Additional work is required to improve the performance of these models and test this method on a variety of note types across different healthcare settings.

3.1.3 Extracting Angina Symptoms from Clinical Notes

A version of this section has been previously presented at the 2020 American Medical Informatics Association Symposium and published in the associated conference proceedings [Eisman et al., 2020].

The semi-structured nature of clinical notes means that the answers to many interesting questions are often constrained to one or two sections. Therefore, reliable extraction of an individual clinical note section increases the signal-to-noise ratio of the natural language processing information retrieval task – that is, it reduces the size of the proverbial haystack from which the needle of information is to be found. One such question that is potentially relevant for the task of understanding the risk of incident myocardial infarction is the symptom of chest pain that clinical judgment suspects are cardiac in origin.

Angina pectoris is a constellation of symptoms that portends inadequate oxygenation of cardiac muscle due to either a decrease in coronary blood supply, an increase in myocardial oxygen demand, or both [Lilly and Braunwald, 2012]. These symptoms are classically described as substernal chest pain that worsens with exertion and improves with rest or administration of nitroglycerin [Diamond, 1983]. Concomitant shortness of breath is also frequently reported. The presence of anginal symptoms has been demonstrated to increase the likelihood of underlying atherosclerotic cardiovascular disease substantially and is a key input variable along with traditional cardiovascular risk factors in estimating patients' pretest probability of coronary artery disease (CAD). This pretest probability, in turn, informs the appropriateness of providers' referral of patients for cardiac testing to

identify obstructive CAD [Wolk et al., 2014]. The CAD Consortium Score is a commonly used tool to estimate the pretest probability of CAD [Genders et al., 2011]. In this score, the presence of typical versus atypical anginal symptoms connotes a more than three-fold increase in the pre-test probability of CAD.

While traditional cardiovascular risk factors (e.g., age, sex, race, hypertension, hyperlipidemia, diabetes, and smoking status) are generally available within the electronic health record (EHR) as structured data, angina symptoms are typically recorded as unstructured natural language free-text descriptions within physician notes [Goff et al., 2014]. A prior study found that chest pain history is recorded as structured International Classification of Disease (ICD) codes in the EHR only half of the time [Pakhomov et al., 2007]. This prevents symptoms from being incorporated into automated clinical decision support systems, limits the evaluation of guideline adherence, and makes it difficult to do large-scale research about the prognostic ability of symptoms outside of clinical trials [Genders et al., 2011] and natural language processing (NLP) challenges [Stubbs et al., 2015].

Prior work to extract both anginal and other symptoms from clinician-written natural language has primarily utilized complex annotation procedures and specialized software requiring both clinical and linguistic expertise [Byrd et al., 2014, Koleck et al., 2019, Pakhomov et al., 2008, Pakhomov et al., 2007, Vijayakrishnan et al., 2014]. In addition, portability assessments outside of the environment in which these methods have been developed show performance reduction [Sohn et al., 2018]. Challenges include the generalized detection of negation, [Wu et al., 2014] interpretation of time expression [Velupillai et al., 2015], and retraining to account for institution-specific differences [Carroll et al., 2012]. However, clinical note content documenting similar events exhibit important similarities that should enable the implementation of NLP tasks across institutions [Sohn et al., 2018]. The combination of significant upfront work to develop NLP tools and the likelihood that they will require testing and retraining to be applied to new environments has made them impractical to implement [Chapman et al., 2011]. Therefore, it is necessary to explore methods that minimize the need for manual feature labeling by both clinical and linguistic experts.

Pre-trained transformer encoder architectures are large language models that address several of these shortcomings and have been shown to produce state-of-the-art results on NLP tasks [Devlin et al., 2018]. Trained to perform language modeling (e.g., next word prediction) on large corpora

of text, they extract contextual representations of words by considering all words in the input sequence at once and selectively focusing on specific parts of their context, based on several aspects such as syntactic structure and semantic similarity. Transformer encoders, which typically comprise hundreds of millions of parameters or more, can thus extract rich, general linguistic information from text and then be fine-tuned to specific tasks, including sequence classification, requiring several orders of magnitude fewer data than would otherwise be needed to train a model from scratch. This study aimed to apply a pre-trained transformer model for the task of angina symptom detection in clinical notes that were written by primary care physicians referring patients for cardiac stress testing [Vaswani et al., 2017]. The specific publicly available pre-trained neural network was built on the Bidirectional Encoder Representation from Transformers (BERT) language model. It was first pre-trained for language modeling on a large-scale biomedical corpus and then on publicly available medical notes [Alsentzer et al., 2019, Johnson et al., 2016]. This study tested the hypothesis that such a language model can be fine-tuned on a small number of annotated physician notes for the purpose of extracting clinically relevant symptom information associated with angina pectoris.

Building a Transformer-Based Classifier for Extracting Angina Symptoms

Patient Population

Consecutive patients (n=459) without known CAD referred for cardiac testing by primary care physicians were included in the study cohort. This population was expected to be enriched for cardiovascular symptoms, including angina and anginal equivalents (e.g., dyspnea on exertion) as they are a common impetus for cardiac stress test referral. Baseline cardiac risk factors were collected including demographics (e.g., age and sex, and race), comorbidities (e.g., diabetes), blood pressure (e.g., systolic and diastolic), medications (e.g., statins, aspirin, and anti-hypertension), and serum cholesterol (e.g., total and high-density lipoprotein). Patients with a known history of prior percutaneous coronary intervention or coronary artery bypass graft surgery were excluded. This study was approved by the Providence VA Medical Center Institutional Review Board.

Angina Definitions

Typical angina is pain or discomfort that is: (1) substernal (i.e., the center of the chest), (2) provoked by exertion or emotional stress, and (3) relieved by rest or nitroglycerin. The presence of two out of

three of these symptoms constitutes atypical angina, and pain is considered nonanginal when it only meets a single criterion [Diamond et al., 1980]. The category of nonanginal pain for the purpose of risk estimation has been broadened clinically to include other non-specific symptoms of potential cardiac origin including vague chest discomfort, shortness of breath, and dyspnea on exertion [Wolk et al., 2014].

Manual Note Annotation

Medical notes are routinely organized into sections that indicate expected content. Clinical symptoms can most frequently be found in the history of present illness (HPI), review of systems (ROS), or assessment/plan (A/P). This study focused on the HPI section, which represents the interval medical history as expressed by the patient and distilled by the note writer. For each patient in the study cohort, the primary care note most proximal to the cardiac stress test order date was identified using the VA's Computerized Patient Record System, and the relevant sections were extracted. HPI sections were selected as one single continuous portion of text from within the note, the beginning of which was usually indicated by a section header (e.g., "HPI" or "Subjective") and the end of which was indicated by the start of another note section (e.g., "Medications" or "Objective").

Study data were collected and managed using instruments implemented in the Research Electronic Data Capture (REDCap) tool [Harris et al., 2019, Harris et al., 2009]. HPIs were stored in free text fields and reviewed from within REDCap. Each HPI was annotated using symptom fields regarding the positive, negative, or absent mention of (1) chest pain or discomfort, (2) substernal chest pain, (3) chest pain provoked by exertion or emotional stress, (4) chest pain relieved by rest or nitroglycerin, (5) shortness of breath, and (6) dyspnea on exertion (Table 3.10). The definition of 'substernal' was broadened from the original Diamond and Forrester description [Diamond et al., 1980] to include imprecise anatomical descriptions that are colloquially meant as substernal or for all practical purposes are more likely than not a description of substernal pain or discomfort. This includes left-sided, left anterior, and descriptions of heaviness, tightness, heartburn, and precordial pain. Definitions of exertion included almost any association of the pain with a description of the activity. Two examples that were deemed not enough activity to warrant exertion were "with standing" and "while washing the dishes." The emotional stress criteria were considered met with the direct connection made between anxiety or an emotionally stressful event to the onset of chest

pain at the same time. A patient who generally reported being anxious or under a lot of stress (e.g., family, job, or homelessness) but did not make the explicit link to chest pain symptoms was not annotated as having emotional stress-induced chest pain.

Symptom	Positive Examples	Negative Examples	Ambiguous Cases
Chest Pain or Discomfort	Chest: pain, pressure, tightness, ache, discomfort, heaviness, sitting on, burning, something "not quite right" Described as: sharp, dull, severe, mild, stabbing Acronyms: cp, sscp	any negation of a positive	The patient has had two episodes of chest pain in the past month. Patient denies chest pain today. → positive
Substernal Chest Pain	CP location: substernal, sternal, central, center, middle, left, anterior, precordial, epigastric, across CP quality that implies location: pressure, tightness, heaviness, sitting on, radiates to the left arm/shoulder Acronyms: sscp, ssc, ss, l, l ant	right anterior, right, another body part only (shoulder, neck, arm, jaw, leg)	
Chest Pain Provoked by Exertion or Emotional Stress	Provocation: exertion, stress, walking, running, stairs, shoveling, mowing lawn, not specifically deemed "negative"	(w/out, w. out, without) "positive" OR (at/while) rest, laying down, night, sitting, standing up, washing dishes	"with rest and exertion" → positive
Chest Pain Relieved by Rest or Nitroglycerin	Palliation: rest, stopping, sitting, laying down, taking a few breaths	"positive" with time modifier that is excessively long (e.g., > 1hr)	"relieved by rest or continuing to walk" → positive
Shortness of Breath	Synonyms: shortness of breath, sob, dyspnea, any mention of breathing difficulty Acronym: sob	any negation of a positive	
Dyspnea on Exertion	Defined with respect to dyspnea as exertional chest pain is defined with respect to chest pain. Acronym: doe		

Table 3.10: Angina symptoms annotations guidelines.

It was not uncommon for there to be internal inconsistencies within a single note for these categories. For example, the HPI may state that a patient experienced exertional chest pain symptoms recently but then go on to state that the patient had no chest pain at the time of the visit. Similarly, patients sometimes report that pain occurs both at rest and with exertion. Cases with both a positive and negative mention of a label were considered positive mentions unless the positive case was stated to have been resolved for a clearly explained reason. It took an average of two minutes to annotate each HPI for these symptoms. All annotations were performed by [ASE[4]] and then reviewed by [NRS[5]]. Disagreements were adjudicated between the two annotators with the help of [INS[6]]. The final reference standard was the result of a consensus between the three coders.

Transfer Learning with Clinical BERT

BioBERT is a language representation model trained on both general (Wikipedia and BooksCorpus) and domain-specific (PubMed Abstracts and PMC Full-text articles) text [Lee et al., 2020]. The model has been fine-tuned on a language modeling task using discharge summaries from the Beth Israel Deaconess Medical Center Medical Information Mart for Intensive Care III database (MIMIC III). Bio+Discharge Summary BERT is publicly available within the Python HuggingFace transformers library as part of the BertForSequencClassification transformer class [Alsentzer et al., 2019, Wolf et al., 2019]. Bio+Discharge Summary BERT was chosen as a domain-appropriate language model for the downstream task of detecting anginal symptoms, which was cast as a classification task.

HPI text from each primary care note was embedded using Bio+Discharge Summary BERT. Each embedding was prepended with a [CLS] and padded with [PAD] tokens until reaching 512 total tokens, the maximum allowed by BERT models. Attention masks were constructed to enable batch processing of sequences by identifying which tokens should be attended to and which should be ignored as padding. Separate models were constructed to identify each of the six symptoms. Models were configured to identify three classes including positive, negative, and absent mentions of the symptom. The training was performed over 2, 4, 8, 16, 32, 64, and 128 epochs with a batch size of 8. The batch size was selected as the maximum number of 512 token embeddings that can be processed on one 12 GB GPU at a time. Data was split into training, validation, and testing sets (80/10/10)

[4]Aaron S. Eisman
[5]Nishant R. Shah
[6]Indra Neil Sarkar

and then cross-validated ten-fold such that all data was represented in a collective validation and a testing set exactly once each. Matthews Correlation Coefficient (MCC) was calculated on pooled validation sets and plotted versus the number of training epochs for each symptom. The number of training epochs corresponding to the maximum of each of these curves was selected as the number of epochs to be used for reporting aggregate performance on the test set.

Measuring the Performance of Angina Symptoms Retrieval

For each model, we reported the aggregate performance on the test set, after previously optimizing the number of training epochs based on aggregate performance on the validation set. In order to assess the performance of detecting the positive presence of symptoms within the notes, the confusion matrices for each symptom were reduced to two-by-two such that the absent and negative classes were collapsed into a single class. Precision, recall, F_1-score, specificity, and Matthews Correlation Coefficient (MCC) were calculated. MCC was chosen in addition to the other information retrieval evaluation metrics that are not informed by true negative cases [Matthews, 1975]. This was accomplished by balancing the ratios of the four quadrants of the binary confusion matrix (i.e., true positives, false positives, true negatives, and false negatives).

Computational Resources

This research was conducted using computational resources and services at the Center for Computation and Visualization at Brown University. 12GB NVIDIA GeForce Titan V and Tesla P100 GPUs were used for computation. Each epoch had a runtime of approximately 17-26 seconds (Titan vs P100 respectively) for a total of between 2.5 minutes to 18 minutes to fine-tune one model on the Titan V (8 and 64 epochs respectively).

Angina Symptoms in Cardiac Stress Test Referral Notes

Patient Characteristics

After annotation of HPIs extracted from 459 consecutive patients without known CAD referred for cardiac testing by primary care physicians, 243 patients were noted to be experiencing chest pain and 205 were noted to be experiencing shortness of breath (SOB). Baseline clinical characteristics

are reported by chest pain class (Table 3.11).

The median age of all patients was 66 years [57-71]. Patients were predominantly male (92%) and self-identified as White race (88%). Thirty percent had a history of diabetes. About half of the patients were on a statin (58%), aspirin (44%), or anti-hypertensive (58%) therapies. Ten-year pooled cohort equation risk of developing atherosclerotic cardiovascular disease based on traditional risk factors was 18.1% (9.9-27.8) [Goff et al., 2014].

Characteristic	All	Typical Chest Pain	Atypical Chest Pain	Non-Specific Symptoms	Asymptomatic
N	459	24	73	255	107
Age (years)	66 [57-71]	66.5 [57.5-71]	62 [54-71]	66 [59-70]	65 [58-71]
Sex (male)	423 (92%)	24 (100%)	65 (89%)	231 (91%)	103 (96%)
Race (white)	406 (88%)	22 (92%)	64 (88%)	230 (90%)	90 (84%)
Diabetes	137 (30%)	9 (38%)	16 (22%)	78 (31%)	34 (32%)
Smoker (current)	94 (20%)	3 (13%)	15 (21%)	51 (20%)	25 (23%)
SBP (mmHg)	132 [124-141]	130 [121.8-141]	129 [121-140]	132 [124-141]	132 [124-142]
DBP (mmHg)	78 [73-82]	76 [74-82.3]	76 [70-83]	78 [73-82]	78 [74-82]
Cholesterol					
Total (mg/dL)	179 [153-213]	191 [159-216.5]	179 [153-227]	181 [153-212]	177 [147.5-206.5]
HDL-C (mg/DL)	44 [37-54]	44 [39.5-47.5]	46 [37-53]	45 [37-54]	43 [38-50.5]
Chest Pain	243 (53%)	24 (100%)	73 (100%)	146 (57%)	–
Substernal	179 (39%)	24 (100%)	65 (89%)	90 (35%)	–
Exertional	108 (24%)	24 (100%)	65 (89%)	7 (19%)	–
Improves with Rest	42 (9%)	24 (100%)	16 (22%)	2 (1%)	–
Shortness of Breath	205 (45%)	16 (67%)	35 (48%)	154 (60%)	–
Exertional	155 (34%)	16 (67%)	29 (40%)	110 (43%)	–
Pharmacotherapy					
Statin Therapy	265 (58%)	12 (50%)	43 (59%)	149 (58%)	61 (57%)
High intensity	121 (26%)	7 (29%)	21 (29%)	67 (26%)	26 (24%)
Moderate intensity	134 (29%)	5 (21%)	20 (27%)	77 (30%)	32 (30%)
Low intensity	10 (2%)	–	2 (3%)	5 (2%)	3 (3%)
Aspirin	201 (44%)	11 (46%)	28 (38%)	112 (44%)	50 (47%)
Anti-HTN	264 (58%)	13 (54%)	33 (45%)	146 (57%)	72 (67%)
Cardiovascular Risk					
10-Year ASCVD Risk	18.1% [9.9-27.8]	20.9% [9.1-27.9]	14.3 [7.5-22.8]	18.4 [10.7-28.0]	20.5% [9.9-28.6]
CAD Consortium Pretest Probability					
High	–	–	–	–	–
Moderate	389 (85%)	22 (92%)	58 (79%)	214 (84%)	95 (89%)
Low	70 (15%)	2 (8%)	15 (21%)	41 (16%)	12 (11%)

Table 3.11: Stress test referral patient characteristics. SBP - systolic blood pressure, DBP - diastolic blood pressure, HDL-C - high-density lipoprotein cholesterol, HTN - hypertension, ASCVD - atherosclerotic cardiovascular disease, CAD - coronary artery disease.

Note Annotations

Notes were annotated for chest pain and shortness of breath symptoms. The majority of HPIs contained information about chest pain (77%) and shortness of breath (64%). Of the notes that contained positive mentions of chest pain, three-quarters mentioned pain location while a minority described pain provocation (34%) or palliation (12%). Positive or negative mentions of shortness of breath were in 64% of HPIs, more than half of which specifically mentioned the presence or absence of exertional symptoms. Overall, there were three times as many positive as negative symptom mentions. Despite the classification of distinct presence and absence of symptoms, clinical documentation of anginal symptoms often described complex narratives that required interpretation in order to be correctly classified. One example of this was a patient whose chest heaviness and shortness of breath had been linked to the non-cardiac source of mold exposure and since moving apartments, the patient reports the chest heaviness as being fully resolved and shortness of breath improved. In this case, the note would be correctly classified as negative for substernal chest pain and positive for shortness of breath. However, a prior note (not part of this study) that hypothesized mold as a potential cause of the patient's symptoms would be considered positive for both. The final model classified the patient as being positive for both symptoms (correct for shortness of breath and incorrect for substernal chest pain).

Anginal Symptom Extraction

Embedded HPI sections were truncated to BERT's 512 token limit. The majority of tokenized notes fit within the limit resulting in only 3.3% (n = 16) being truncated. A Chi-Square test demonstrated that the frequency of positive symptoms within the truncated samples was in line with the overall dataset (p = 0.95).

BioBERT+Discharge Summary models were each fine-tuned to detect one of six different anginal symptoms using the labeled embedded HPI sections. The epoch number parameter was determined using an aggregated validation set. The graphs in Figure 3.6 were inspected visually to determine the best early stopping parameter for each symptom. Parameter selection was straightforward given the limited number of epochs tested and the absence of local maximums observed. The chosen epoch number parameter was then used to test combined held-out test data from each of the ten

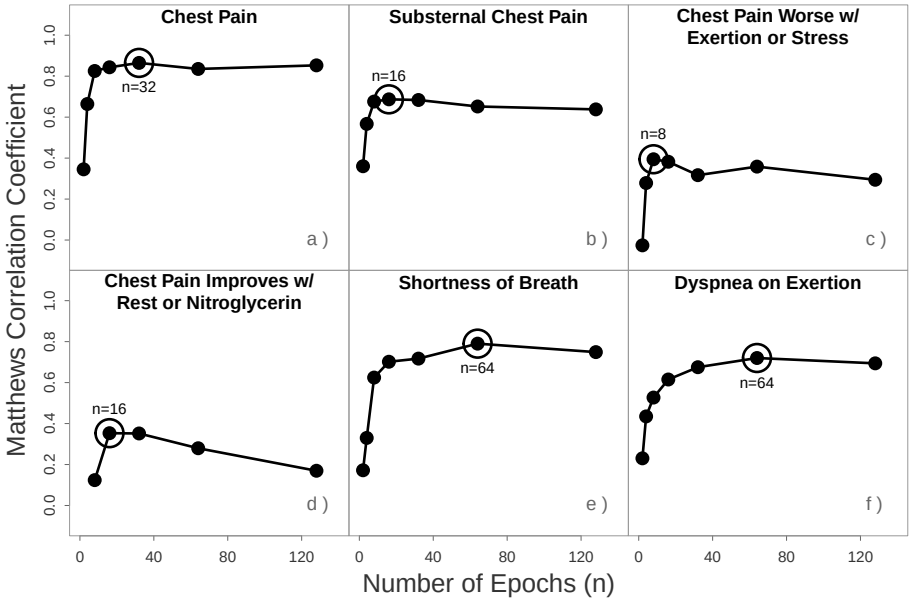

Figure 3.6: Epoch number selection by angina symptom. Individual models for each symptom: (a) chest pain or discomfort, (b) substernal chest pain, (c) chest pain provoked by exertion or emotional stress, (d) chest pain relieved by rest or nitroglycerin, (e) shortness of breath, and (f) dyspnea on exertion, were trained using a range of epochs. Model performance (Matthews Correlation Coefficient) was evaluated as a function of epochs. The best-performing number of epochs for each symptom was selected as the final model parameter for evaluation on held-out testing data.

		Adjudicated Manual Annotation								
		Absent	+	–	Absent	+	–	Absent	+	–
Predicted Annotation		Chest Pain			SS CP			EX CP		
	Absent	90	4	1	223	26	4	274	47	23
	+	8	226	10	39	152	5	30	59	22
	–	14	9	97	0	0	0	0	2	2
		RE CP			SOB			DOE		
	Absent	389	29	12	136	7	2	245	14	6
	+	8	12	2	18	186	11	38	137	6
	–	5	1	1	12	12	75	7	3	3

Table 3.12: Angina symptom retrieval confusion matrices. SS CP – substernal chest pain, EX CP – exercise/stress-induced chest pain, RE CP – chest pain improves with rest or nitroglycerin, SOB – shortness of breath, DOE – dyspnea on exertion.

cross-folds.

The final raw model performance is presented in a three-by-three confusion matrix for each symptom (Table 3.12). These include predicted annotation of absent, positive, and negative symptom mentions versus consensus manual annotation. For chest pain, the sequence classification model's most common errors included negative identification of chest pain when it was absent (n = 14) and positive identification of chest pain when it was negative (n = 10). Negative assertions of chest pain characterization variables were under-classified by the models. False positives of dyspnea on exertion were the most common breathing-related misclassification (n = 38).

The confusion matrices were reduced to two-by-two tables where "Absent" and "–" assertions were collapsed into a single category to evaluate the model's performance to detect the presence or absence of symptom documentation. Precision, recall, F_1, specificity, and Matthews Correlation Coefficient were calculated for each symptom and are reported in Table 3.13. Chest pain was extracted better than any other symptom (F_1 = 0.94, MCC = 0.87). Chest pain subtype extraction performance ranged from good (substernal) to poor (improved with rest). Shortness of breath and dyspnea on exertion was extracted with good performance (F_1 = 0.83, MCC = 0.70 and F_1 = 0.82, MCC = 0.72 respectively).

Evaluation of Angina Symptom Retrieval using a Pre-Trained Transformer

This study examined the potential to use a pre-trained transformer architecture to extract anginal symptom information from clinical notes [Koleck et al., 2019]. A publicly available BioBERT+Discharge

	Chest Pain	SS CP	EX CP	RE CP	SOB	DOE
Precision	0.926	0.776	0.532	0.545	0.865	0.757
Recall/Sensitivity	0.946	0.854	0.546	0.286	0.907	0.890
F_1	0.936	0.812	0.539	0.375	0.886	0.818
Specificity	0.918	0.843	0.852	0.976	0.886	0.856
MCC	0.865	0.690	0.394	0.353	0.790	0.72
Training Epochs	32	16	8	16	64	64

Table 3.13: Evaluation statistics of angina symptom retrieval. SS CP - substernal chest pain, EX CP - exercise/stress-induced chest pain, RE CP - chest pain improves with rest/nitroglycerin, SOB - shortness of breath, DOE - dyspnea on exertion, MCC - Mathews Correlation Coefficient.

language model was fine-tuned using HPI sections from 459 consecutive patients referred for cardiac testing. The models detected mention of chest pain with greater than 90 percent sensitivity and specificity. Chest pain location, shortness of breath, and dyspnea on exertion were similarly extracted with high sensitivity and specificity. These represent excellent starting points for the characterization of anginal symptoms from clinical notes of patients referred for cardiac testing.

Prior literature on the extraction of symptom information from clinical notes is sparse. A recent systematic review identified 27 articles related to the use of NLP for clinical symptom detection and noted that study reproducibility is overall poor due to: (1) lack of detail about the patient population, (2) failure to report detailed performance, and (3) poorly described NLP methodology [Koleck et al., 2019]. One study extracted chest pain and dyspnea symptoms with reported sensitivity on a small subset of their data, similar to the performance reported here [Pakhomov et al., 2008]. However, no additional metrics, including false-positive rates, were provided for comparison, and methods lack details that would enable the system to be reproduced or implemented.

The performance of NLP systems to extract the provocation and palliation of chest pain has not been previously reported. The presented results indicate that the granular characterization of chest pain symptoms can be ruled out with high specificity. Reliable detection of these features requires additional consideration. The strength of the note for documentation is that it allows the clinician freedom to express complexity, ambiguity, and uncertainty of how symptoms are experienced and remembered. This leads to significant variability in how patient symptoms are ultimately documented. For example, it is useful to document that a patient has both experienced chest pain in the recent past and is pain-free at the time of the clinical encounter. However, these seemingly

contradictory statements may be a challenge for a general language model attempting to determine if a patient has experienced chest pain at all. In addition, through the annotation process, it became clear that there is a wide range of ways that patients report, and clinicians document provocation and palliation of chest pain.

Fractional withholding of training data for both chest pain and substernal models indicates that performance is strongly related to the number of positive examples. Training models for these two symptoms with a comparable number of positive examples as the results reported for provocation and palliation of symptoms indicate similar performance. Small sample size for the provocation and palliation descriptions combined with narrative complexity is likely responsible for low sensitivity. Based on these findings, performance appears to plateau between 200 and 250 positive examples which may represent the necessary number to encompass the semantic variability of how chest pain is documented within the studied population.

The six classification objectives (i.e., angina symptoms) are not independent. Therefore, instead of training separate models for predicting each independently, it is possible to train a model that jointly predicts all objectives. Not only would deploying and maintaining a single model instead of six be more practical but this approach, known as multi-task learning, also offers theoretical benefits, such as the regularization of model weights [Baxter, 1997, Collobert and Weston, 2008]. The multi-task model shares the same architecture as the task-specific models, but with six distinct output layers instead of only one. The loss to be minimized during training is a weighted average of the six component losses. The multi-task model showed inferior predictive performance compared to the task-specific models. The potential reasons for these findings include: (a) the difficulty in stratifying the small training and test sets across 10 cross-validation folds in a way that preserves class balances and (b) the well-documented sensitivity of training to the weights of the individual tasks [Kendall et al., 2018].

Identifying clinically actionable anginal symptoms was the focus of this study. The developed models were designed to answer questions about patients that aid in the diagnosis [Diamond et al., 1980] and determine the pretest probability [Genders et al., 2011] of coronary artery disease. Automating the extraction of anginal symptom information as presented in this study is potentially deployable as

part of clinical decision support systems. It would enable health system-wide identification of high-risk patients and coordinate diagnostic cardiac testing for symptomatic ones [Wolk et al., 2014]. In addition, automated symptom extraction is required for research on the relationship between anginal symptoms and downstream clinical testing (e.g., catheterization with and without percutaneous coronary intervention, coronary artery bypass graft surgery) and outcomes (e.g., myocardial infarction, cardiac death, and all-cause mortality).

Several limitations deserve mention that would need to be addressed for this work to be applied in a production environment. The process of identifying appropriate clinical notes for symptom extraction was performed manually and relies on institution-dependent EHR knowledge, including note types and physician relationship to the patient (e.g., primary care). In addition, relevant note section identification was also performed manually. This has been previously automated using institution-specific template headers [Denny et al., 2009], as well as by modeling sections as hidden states of a Markov process [Eisman et al., 2021]. Finally, the transferability of these models to other clinical contexts and healthcare systems needs to be evaluated.

This study presents a promising method for detecting anginal and anginal-equivalent symptoms from clinical texts using pre-trained transformer architectures. The findings demonstrate that a generalized language model can be fine-tuned on a limited set of annotated physician notes to enable the extraction of clinically actionable anginal symptoms. The extracted symptoms align with data inputs for validated risk models and clinical guidelines and may allow the development of automated decision support and quality assessment. Additional work is required to assess the additional fine-tuning needed to apply these models in other clinical settings.

3.2 Characterizing Population Incident Atherosclerotic Cardiovascular Disease and Primary Prevention with Centralized Health Information Exchange

The Rhode Island Quality Institute manages the statewide health information exchange CurrentCare. CurrentCare aggregates clinical data from Rhode Island healthcare institutions for individuals

that receive healthcare in Rhode Island and is the source data for the population studied in the following section.

Statewide Adherence to 10-Year Atherosclerotic Cardiovascular Disease Risk-Based Primary Prevention Guidelines with Statins in Rhode Island

Health information exchange (HIE) evokes care coordination through timely point-of-care access to electronic health records (EHR) across time and place about a single patient [Menachemi et al., 2018]. Centrally managed HIE also has the potential to enable population health monitoring [Feldman et al., 2022] and identify gaps between medical science and population health care [Cochrane et al., 2007, Metser et al., 2021, Spranger et al., 2004]. In this study, we demonstrate this potential by analyzing population adherence to risk-based cholesterol-lowering guidelines with statins. Our work also provides insights into atherosclerotic cardiovascular disease (ASCVD) risk and statin therapy practices across diverse care settings.

ASCVD is the leading cause of morbidity and mortality in the United States [Murphy et al., 2021, World Health Organization, 2018]. Since 2013, guidelines for the primary prevention of ASCVD with statins have been informed by an individual's estimated 10-year risk of incident disease [Goff et al., 2014, Grundy et al., 2019, Stone and Lloyd-Jones, 2014]. The risk factors that form these models are routinely collected as part of clinical care [Eisman et al., 2020, Pakhomov et al., 2007]. Still, little is known about the patterns of risk factor collection across healthcare institutions for individual patients and how these risk factors translate to primary prevention practices for a geographically defined population. Existing studies that leverage EHR data [Metser et al., 2021] for this purpose are limited to single healthcare organizations [Brown et al., 2023], and attempts to characterize populations rely on national surveys [Patel et al., 2019].

The overall goal of this study was to characterize population 10-year ASCVD risk and primary prevention practices across Rhode Island (RI) using HIE data. We standardized data from the RI statewide designated HIE, CurrentCare, to the Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM), which ensured data quality and consistency across data-sharing partners [Makadia and Ryan, 2014]. We report the aggregation of standardized ASCVD risk factors by HIE for RI. We then assessed guideline-directed statin adherence using patient-level

medication fill data provided by the medication data aggregator SureScripts. Finally, motivated by previously reported disparities [Hicks et al., 2004, Kurian and Cardarelli, 2007, Mensah et al., 2005, Ong et al., 2007, Redmond et al., 2011], ASCVD risk factors, 10-year ASCVD risk, and statin adherence were evaluated for potentially vulnerable individuals at the intersection of care at Federally Qualified Health Centers (FQHC) and racial and ethnic minority subpopulations.

We hypothesized that individuals in RI have pharmacotherapy-determining ASCVD risk factors regularly measured at multiple healthcare institutions and that ASCVD risk factors, 10-year risk, and primary prevention practices will differ by primary care institution type.

Methods for Characterizing Population ASCVD Risk and Primary Prevention Practices

Study Population

This study analyzed CurrentCare HIE data from 2013 to 2023 on 62,999 adults. Adults aged 40-75 years with the minimum required data within CurrentCare to calculate 10-year ASCVD risk and without known clinical ASCVD were included in the cohort for this study. Individuals who experienced incident ASCVD were censored at the time of diagnosis. A primary care provider (PCP) site was assigned to each cohort member. Individuals receiving care at PCP practices or FQHCs were assigned those as their PCP sites. Otherwise, the PCP site (e.g., a health system with known ambulatory primary care services) was defined as the most common outpatient care site.

Data Source and Mapping HIE Data Warehouse to the OMOP CDM

CurrentCare, the RI statewide HIE is owned and operated by the Rhode Island Quality Institute (RIQI). CurrentCare aggregates more than 102M patient care transactions, 90% of RI prescription data, and 90% of RI laboratory data across 400 sites (e.g., primary care and specialty practices, community health centers, hospitals, long-term care facilities, and visiting nurse agencies) and 520 distinct sources for over 540,000 enrolled individuals (>50% of the RI population).

The Observational Health Data Sciences and Informatics (OHDSI) community maintains standard vocabularies and database schemas that define the OMOP CDM. The National Institutes of Health has chosen this patient-oriented longitudinal data structure as the standard for the All of Us Research

Program [All of Us Research Program Investigators et al., 2019] and national collaborations such as the National COVID Cohort Collaborative [Haendel et al., 2020]. We selected it as the CDM for the RI HIE and mapped the CurrentCare data warehouse to OMOP CDM version 5.3.1 [OHDSI, 2021b]. The OMOP observation start date was defined as the date the individual first enrolled in CurrentCare, and the end date was set to the data extraction date or disenrollment. Disenrollment from CurrentCare occurs upon request or death.

We standardized data by domain according to the OHDSI CDM and its vocabularies [OHDSI, 2021a, OHDSI, 2022a]. Existing Structured Query Language (SQL) code [OHDSI, 2022b] was used to standardize more than 99% of problems, diagnoses, drugs, and observations. We converted non-standard condition codes (78%) from ICD-9-CM and ICD-10-CM to SNOMED CT and non-standard drug codes (75%) from National Drug Codes to RxNorm. Observations and procedures were already coded using the required Logical Observation Identifiers Names and Codes (LOINC) and Current Procedural Terminology (CPT), respectively. We manually mapped 1000 of the most common measurement codes (e.g., SBP and HDL-cholesterol) not already represented with LOINC (10%) using the OHDSI Usagi [OHDSI, 2021a] tool to cover more than 95% of measurements. The OMOP CDM associates each data point with an encounter type (e.g., inpatient, outpatient, and emergency).

In accordance with the RIQI policies and procedures, data provided for this research study were de-identified based on expert determination in accordance with HIPAA guidelines.

Estimating 10-Year ASCVD Risk

We queried ASCVD risk factors (hereafter, risk factors) from the OMOP person (e.g., age, sex, and race), measurement (e.g., SBP, total cholesterol, and HDL-cholesterol), condition (e.g., diabetes and smoking status), and drug exposure (e.g., anti-hypertension medication use [Eisman, 2022, Whelton et al., 2018]) domains. We limited risk factor data to those derived from outpatient visits on days where no emergency or inpatient visits were recorded to avoid capturing acute events that may artificially elevate estimated risk (e.g., emergency visits for pain resulting in atypical SBP). We calculated the HIE-based 10-year ASCVD risk (hereafter, ASCVD risk) for each individual using the pooled cohort equations (PCE) [Goff et al., 2014]. We tracked ASCVD risk longitudinally at each date

for which an updated risk factor value was available throughout an individual's observation period or until an incident ASCVD event. We defined ASCVD events as a new condition of myocardial infarction (SNOMED CT 22298006), cerebrovascular accident (SNOMED CT 230690007), or any of their child codes (Appendix B).

Statin Guideline Adherence

Statins are clinically indicated for the primary prevention of incident ASCVD in individuals aged 40-75 who are at intermediate ($\geq$7.5% - <20%) or high ($\geq$20%) 10-year risk [Grundy et al., 2019]. We defined adherence to these guidelines as a dichotomous variable (e.g., on or off statin therapy) at each date for which an updated risk factor was available. We considered individuals to be 'on' statins for that date if they filled a prescription within 90 days of the updated risk factor. We maintained this status for each day until the next ASCVD risk update. We did not evaluate supply sources (e.g., pre-supply, hospital stays, or switching). Dosing information, days supply, and instructions were unavailable for this study. The reporting of this method to assess statin adherence was informed by the TEN-SPIDERS tool [Dalli et al., 2022].

Statistical Analysis

Baseline statistics are reported as mean (sd) for normally distributed (e.g., age) and median [IQR] for skewed (e.g., ASCVD risk) data. Average risk factors were computed on individual median values such that each individual contributed to the average exactly once. Comparisons of risk factors and statin therapy between subpopulations were adjusted for non-modifiable risk factors (i.e., age, sex, and Black race). Race was not adjusted for when comparing subpopulations of a single race. Adjusted β-coefficients were calculated using logistic regression. All analyses were performed with a significance threshold of $p < 0.05$ in R version 4.2.1 and RStudio version 2022.07.1.

Rhode Island Population ASCVD Risk and Primary Prevention Practices

Baseline Characteristics

A cohort of 62999 individuals met the inclusion and exclusion criteria for this study (Figure 3.7). Data for this cohort were derived from the RI HIE CurrentCare, which aggregates EHR data from RI health data partners into a consolidated unified care record for all HIE participants (Figure

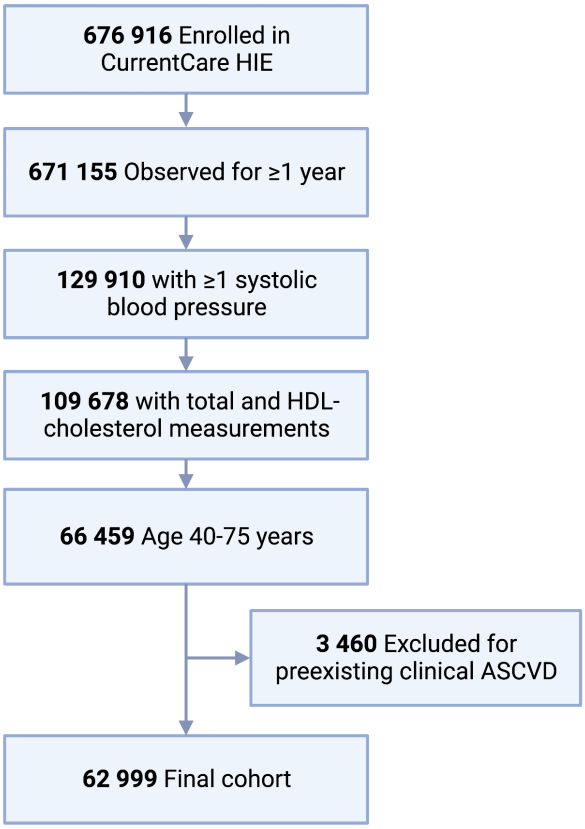

Figure 3.7: Flow diagram of inclusion and exclusion criteria

3.8A). This is the first reported statewide HIE mapped to the OMOP CDM to enable population observational health research. The cohort contained complete ASCVD risk factors required to estimate 10-year risk with a median [IQR] HIE risk estimate follow-up of 3.5 [1.6-5.7] years (Figure 3.8B). The mean age (SD) was 59.1 (10.7) years, and 40319 (64%) were women. By self-report, 4156 (6.6%) were Black, and 8971 were Hispanic (14.2%), which are proportionally consistent with RI Census Data. Black (55.9 years, $p<0.001$) and Hispanic (55.7 years, $p<0.001$) individuals were younger, and Hispanic individuals had a higher proportion (70%, $p<0.001$) of women compared to the population as a whole. Over half (65%) of the cohort had risk factors contributed by more than one data partner, and 30% had risk factors reported by three or more (Figure 3.8C).

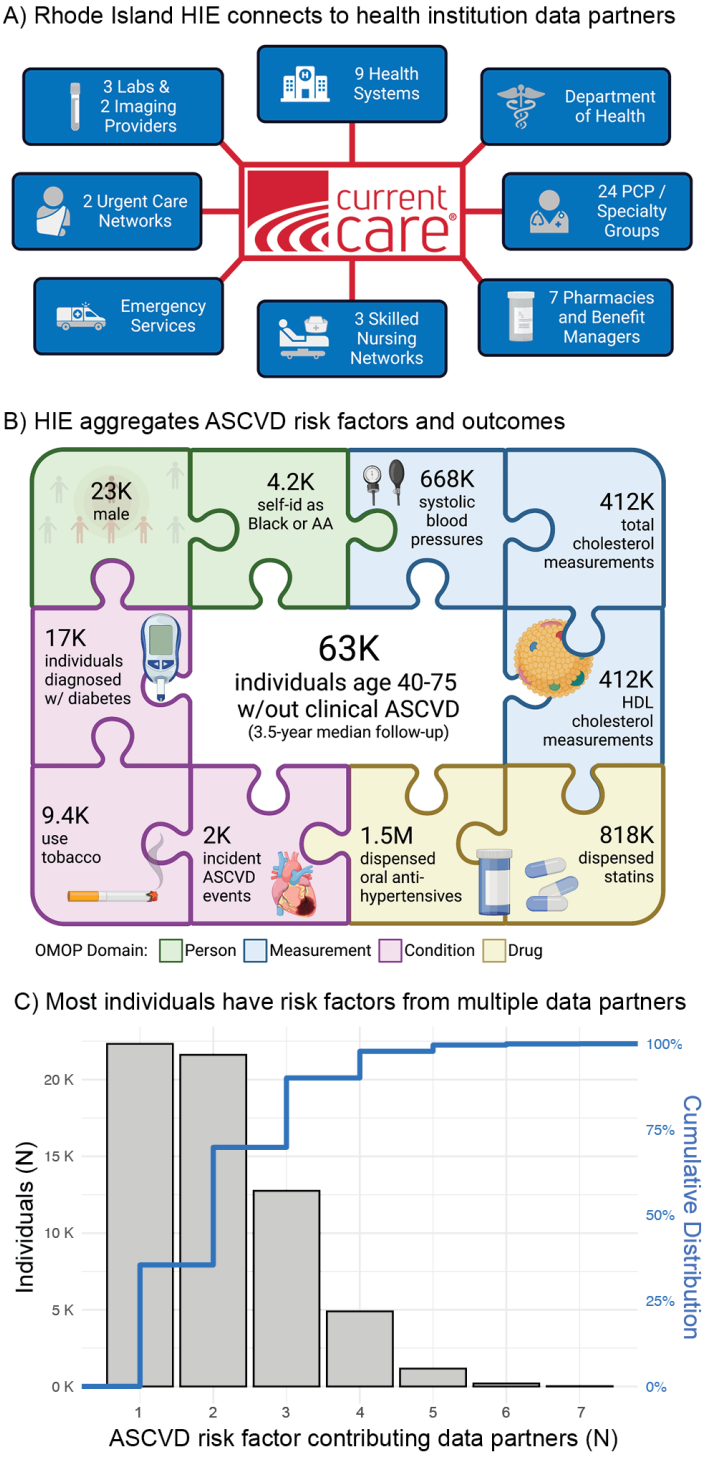

Figure 3.8: Health information exchange, ASCVD risk factors, and outcomes

	Overall
N	62999
Patient Days	87076619
Age	59.11 (10.71)
Female	64%
Black	7%
SBP	126.74 (13.44)
Total Cholesterol	185.79 (36.81)
HDL-Cholesterol	52.43 (15.33)
Diabetes	27%
Tobacco Use	15%
10-Year ASCVD Risk	7.0% [2.4-15.6]
Annual Risk Updates (PCP Only)	0.88 [0.28-1.83]
Annual Risk Updates (HIE)	1.81 [1.06-3.23]
Days at Intermediate Risk (+On Statin)	31% (31%)
Days at High Risk (+On Statin)	17% (47%)

Table 3.14: Baseline characteristics of the cohort

Assessment of Population ASCVD Risk and Primary Prevention Practices

We tracked HIE-based population 10-year ASCVD risk from 2013 to 2023 (Figure 3.9A). The median individual's risk was 7.0% [2.4%-15.6%], with mean component risk factors reported in Table 3.14. Population prevalence of diabetes and tobacco use were 27% and 15%, respectively. Incident non-fatal ASCVD events occurred at a rate of 1 every 118 years (n=2020), after which they were censored from follow-up analysis.

More than half of the cohort (56%, n=34695) were eligible for risk-based (ASCVD risk $\geq$7.5%) primary prevention of incident ASCVD with a statin at some point during the study, and 66% (n=23175) of those individuals had at least one recorded statin prescription filled. Of those at high risk ($\geq$20%, n=15138), 78% (n=11810) had at least one statin prescription filled. Longitudinal tracking of individual risk, statin therapy, and incident ASCVD events enabled the granular tracking of risk-based statin guideline adherence (Figure 3.9B).

This study followed the population ASCVD risk for 87M patient days. Of those, 34M (39%) were spent at low and 10M (11%) at borderline ASCVD risk. 43M (49%) days were spent eligible for statin therapy, with 27M (31%) and 15M (17%) at intermediate and high risk, respectively. Population risk-based statin guideline adherence increased monotonically by ASCVD risk category (Figure 3.9C), with 32% of days spent at intermediate risk and 47% at high risk on statin therapy. Excluding

individuals with diabetes (an additional indication for statin therapy) reduced this adherence for those at intermediate (27%) and high (38%) risk.

ASCVD Risk Factors, 10-Year Risk, and Statin Use by Race and Ethnicity

ASCVD risk was elevated in Black (β=3.6%, p<0.001) and Hispanic (β=0.9%, p<0.001) subpopulations after adjusting for non-modifiable risk factors (i.e., age, sex, and race for ethnicity comparison). This translated to a higher adjusted proportion of days spent at high risk for Black (β=0.11, p<0.001) and Hispanic (β=0.04, p<0.001) individuals. SBP was elevated in Black (β=4.1mmHg, p<0.001) and Hispanic (β=1.4mmHg, p<0.001) individuals after adjustment. Diabetes was 30% more prevalent in Black and Hispanic subpopulations. As a percentage, Black individuals were more likely (21%, p<0.001) and Hispanic individuals were less likely (12%, p<0.001) to smoke tobacco. Risk-based statin guideline adherence was higher in the high (β=1%, p=0.014) risk Black and the intermediate (β=4%, p<0.001) and high (β=5%, p<0.001) risk Hispanic subpopulations. Accordingly, adjusted total cholesterol was lower in Black (β=-4.1mmol/L, p<0.001) and Hispanic (β=-2.5mmol/L, p<0.001) individuals compared to the population as a whole.

Primary Prevention at the Intersections of FQHC Primary Care, Race, and Ethnicity

The aggregation of health data across institution types for a population affords the opportunity to explore disparities in healthcare system interactions, subpopulation risk characteristics, and treatment by healthcare institution type. Known disparities in cardiovascular risk factors, primary prevention practices, and outcomes by race, ethnicity, and socioeconomic status led us to evaluate the intersection of primary care at FQHCs (n=9251) with race and ethnicity compared to those who receive primary care at Health Systems (n=39555, Table 3.15). Compared to those at Health Systems (HS), FQHC patients were younger (56 vs. 60 years, p<0.001), more likely to be female (68% vs. 63%, p<0.001), and more likely to identify as Black (15% vs. 6%, p<0.001).

Adjusted for non-modifiable risk factors, all other ASCVD risk factors were unfavorable in FQHC patients. SBP was higher (β=2.7mmHg, p<0.001), total cholesterol was higher (β=5.9mmol/L, p<0.001), HDL-cholesterol was lower (β=-1.3mmol/L, p<0.001), diabetes was more prevalent (33% vs. 26%, p<0.001), and tobacco use more common (20% vs. 13%, p<0.001). This translated to

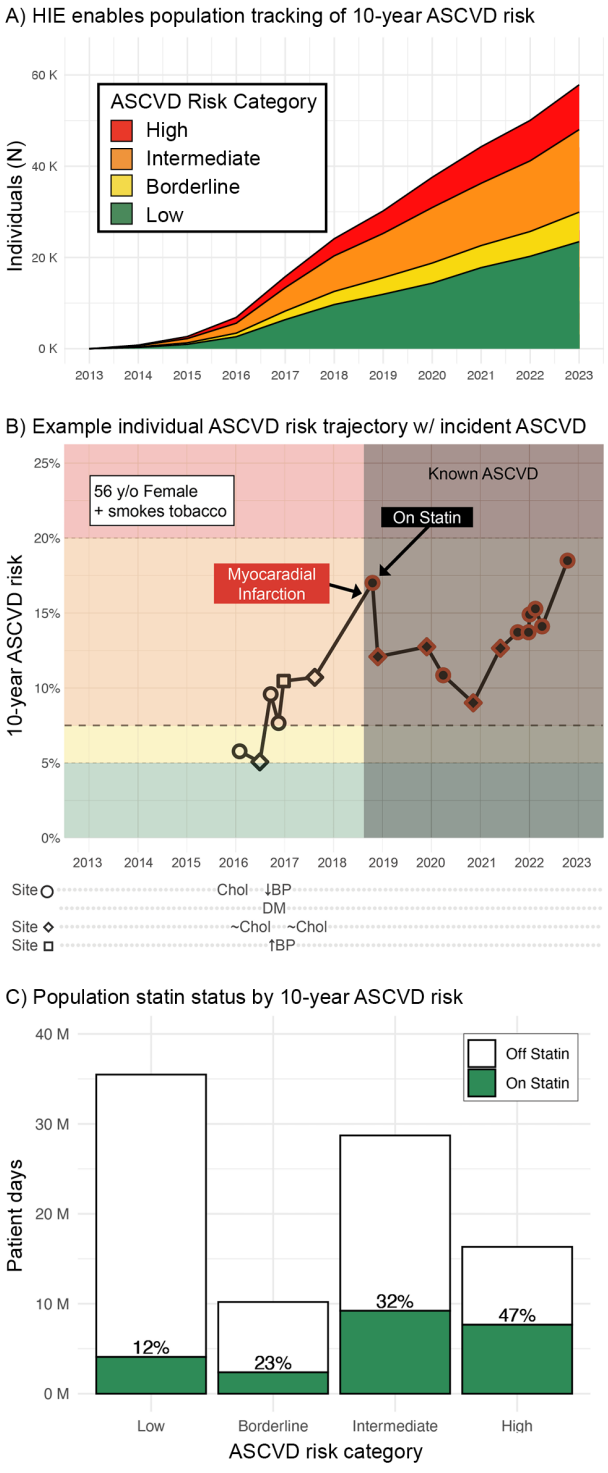

Figure 3.9: HIE enables granular tracking of patient ASCVD risk, primary prevention, and incident disease

worse adjusted ASCVD risk (β=0.78%, p<0.001). FQHC patients spent a higher adjusted number of days at intermediate (β=0.01, p<0.001) and high (β=0.04, p<0.001) ASCVD risk, and a higher proportion of those days on risk-based guideline-directed statin therapy (both β=0.08, both p<0.001). Analyzed at the intersection of race and ethnicity subpopulations, FQHC patients had unfavorable adjusted risk factors compared to hospital system patients of the same demographic subpopulations. Subsequently, FQHC patients spent more days at intermediate and high ASCVD risk and were more likely to be on statin therapy than hospital system patients. This difference remained for intermediate but not high-risk patients at FQHCs after controlling for diabetes diagnosis, an additional indication for statins.

	HS (Overall)	FQHC (Overall)	HS (non-Black & non-Hispanic)	FQHC (non-Black & non-Hispanic)	HS (+Black)	FQHC (+Black)	HS (+Hispanic)	FQHC (+Hispanic)	HS (+Black & +Hispanic)	FQHC (+Black & +Hispanic)
N	39555	9251	33997	3742	2227	1432	3586	4403	255	326
Patient Days	52387672	13108664	45131487	5311448	2784139	2035179	4824126	6228009	352080	465972
Median Risk Follow-Up (days)	1241 [533-2040]	1286 [606.5-2035.5]∗	1248 [531.25-2046]	1295 [640-2040]∗	1098 [446.5-1948]	1272 [596.75-2069]∗	1271.5 [610.5-2029]	1288 [586-2005]∗	1303 [484-2061]	1300 [679.75-1945]
Age	59.56 (10.61)	55.93 (10.39)∗	60.16 (10.52)	56.71 (10.30)∗	56.6 (10.57)	54.99 (10.29)∗	55.56 (10.39)	55.65 (10.43)	57.10 (10.52)	56.95 (10.05)
Female	63%	68%∗	65%	64%	65%	68%	69%	73%∗	67%	73%
Black	6%	15%∗	-	-	-	-	7%	7%	-	-
SBP	126.56 (13.52)	127.95 (13.58)∗{2.68}	126.42 (13.44)	127.74 (13.76)∗{2.24}	129.55 (14.11)	130.08 (14.4)∗{5.12}	126.34 (13.82)	127.53 (13.09)∗{2.59}	130.17 (15.10)	129.31 (13.72)
Total Cholesterol	185.55 (37.04)	185.16 (36.29)∗{5.91}	185.96 (36.89)	185.27 (37.21)∗{5.30}	181.67 (40.62)	184.59 (36.84)∗{3.64}	183.94 (36.19)	185.39 (35.31)∗{5.46}	183.03 (39.57)	187.16 (36.39)∗{9.46}
HDL-Cholesterol	52.64 (15.51)	50.53 (13.82)∗	53 (15.82)	51.08 (14.98)	52.31 (13.89)	52.53 (14.68)	49.3 (12.75)	49.46 (12.38)∗{-2.50}	50.23 (13.46)	51.02 (14.3)∗{-1.77}
Diabetes	26%	33%∗	24%	30%∗	36%	34%∗	36%	36%∗	40%	40%
Tobacco Use	13%	20%∗	12%	29%∗	19%	22%∗	12%	12%∗↓	15%	17%
10-Year ASCVD Risk	6.77% [2.31-15.02]	5.34% [1.79-12.9]∗{0.78}	6.94% [2.41-15.2]	5.9% [2.06-13.43]∗{0.63}	8.28% [3.29-16.72]	7.42% [2.73-15.44]∗{3.4}	4.37% [1.48-11.71]	4.40% [1.5-11.52]∗{0.67}	9.00% [2.72-18.17]	8.59% [3.29-15.97]
Risk-Based Guideline Adherance										
Annual Risk Updates (PCP Only)	1.13 [0.54-2.15]	0.79 [0.16-2]	1.13 [0.53-2.12]	0.58 [0-1.64]∗{-0.38}	1.17 [0.54-2.4]	0.82 [0.25-2.12]	1.2 [0.57-2.35]	1 [0.26-2.29]	1.23 [0.57-2.57]	0.99 [0.31-2.39]
Annual Risk Updates (HIE)	1.75 [1.02-3.08]	2.21 [1.31-3.79]	1.73 [1.01-3.03]	2.04 [1.19-3.52]	1.87 [1.06-3.48]	2.31 [1.33-3.84]	1.87 [1.09-3.33]	2.39 [1.41-4.04]	2 [1.09-3.58]	2.67 [1.45-4.09]
Days at Intermediate Risk (+On Statin)	32% (30%)	29%∗{1%} (37%∗{8%})	32% (30%)	31%∗{3%} (36%∗{5%})	35% (28%)	35%∗{9%} (36%∗{5%})	24% (30%)	25%∗{-1%} (40%∗{10%})	29% (25%)	36%∗{2%} (40% {7%})
Days at High Risk (+On Statin)	17% (45%)	15%∗{4%} (52%∗{8%})	17% (45%)	15%∗{1%} (48%∗{2%})	20% (45%)	19%∗{13%} (53%∗{7%})	14% (45%)	15%∗{3%} (56%∗{12%})	25% (55%)	18% (57%∗{4%})

Table 3.15: ASCVD risk and primary prevention in FQHCs and Health Systems. HS = Health System, FQHC = Federally Qualified Health Center, ∗p<0.05, and {adj. Beta}

Implications of Findings

In the United States, healthcare delivery and data fragmentation make population health monitoring with clinical data particularly challenging. This cohort study of individuals enrolled in CurrentCare demonstrates the population health monitoring capabilities of a centralized HIE mapped to the OMOP CDM. We present the first geographically defined evaluation of cardiovascular risk and primary prevention practices. We found that the risk factors that inform the estimation of 10-year ASCVD risk and risk-based guideline-directed cholesterol-lowering primary prevention with statins are collected and stored longitudinally at the patient level across time and healthcare institutions in RI. Aggregating and standardizing these data enabled granular longitudinal tracking of ASCVD risk and statin therapy using risk factors collected at multiple healthcare institutions for a single patient. The number of individuals included in the cohort increased over time with increasing HIE enrollment and the onboarding of additional data partners. We expect this trend to accelerate with the change from an 'opt-in' to an 'opt-out' consent model for CurrentCare, resulting in nearly all RI individuals in the system.

Compared to the RI population, our cohort included more female (64% vs. 51%) and fewer Black (7.0% vs. 8.8%) individuals. Diabetes prevalence in the cohort was significantly higher than the general population (27% vs. 10.3%), likely due to sampling bias (i.e., patients with diabetes are more likely to interact with the healthcare system). Orthogonal data for other risk factors were unavailable for comparison. Cohort tobacco use was similar to the RI population (15.0% vs. 14.6%) and is an indication of the quality of this variable within CurrentCare, especially given its importance in estimating ASCVD risk.

Our cholesterol guideline adherence findings concur with the only modern United States-based population study contemporary to current guidelines. In that study of the National Health and Nutrition Examination Survey (NHANES), found nearly identical risk-based statin guideline adherence compared to our cohort for intermediate (29% vs. 27%) and high (38% vs. 38%) risk populations without diabetes [Jacobs et al., 2023]. All other recent studies pertaining to modern guideline adherence are limited to a single healthcare system. These report a wide range of statin guideline adherence, including substantially lower (e.g., 25%) [Dorsch et al., 2019] to substantially higher (e.g., 72%) [Rashid et al., 2022] rates of statin use among risk-eligible patients. These comparisons

demonstrate the substantial limitations of single-system studies for imputing regional healthcare system practice patterns. In contrast, the concordance of our statin-adherence findings with those derived in NHANES indicates that centrally managed HIE is likely a representative data source for population health research and monitoring. Regardless, low adherence to statin guidelines highlights the importance of developing effective alternative therapies for statin-intolerant individuals [Nissen et al., 2023].

Prior studies of ASCVD risk and statin therapy that explored disparities by race and ethnicity have found unfavorable risk factors [Howard et al., 2011], elevated rates of ASCVD events [Akinyelure et al., 2021, Jacobs et al., 2023, Zakai et al., 2022], and reduced adherence to risk-based primary prevention in Black and Hispanic individuals [Dorsch et al., 2019, Metser et al., 2021]. Here, we report risk factors and 10-year ASCVD risk in Black and Hispanic individuals that are in accordance with these findings. Notably, we found that statin guideline adherence in Black and Hispanic individuals was the same or higher than in the cohort as a whole. This differs from prior reports. There are a few interpretations of this finding. One is that disparities in statin guideline adherence from survey data (e.g., NHANES) could result from disparities in healthcare access instead of healthcare delivery. In other words, once patients access a clinician, statin care disparities may diminish or be eliminated entirely. Secondly, guideline-directed primary prevention with statins for Black and Hispanic individuals in RI with healthcare access may be better than for the United States population.

This is the first reported analysis that compares cardiovascular risk and primary prevention practices according to care location type within a single geographic region. Comparing populations receiving primary care at FQHCs to those at Health Systems, we found, as expected [Nath et al., 2016b], that FQHCs serve higher proportions of Black and Hispanic individuals, and their patient population has elevated ASCVD risk. This finding held across the intersection of FQHC care with race and ethnicity, which indicates that care location and local population factors may provide better insight into individual ASCVD risk than socially constructed racial categories [Fields and Fields, 2014]. For example, Black patients of FQHCs had unfavorable risk factor profiles compared to Black patients of Health Systems after controlling for non-modifiable risk factors. FQHC patients were more likely to be on statins in accordance with risk-based guidelines. Despite this, the adjusted total cholesterol of Black patients remained unfavorable compared to Health Systems patients. These findings

demonstrate that FQHCs provide high-quality guideline-directed care to vulnerable populations that remain predicted to have elevated rates of ASCVD events in the future.

The aggregation of ASCVD risk factors and standardization to a CDM in the context of a centrally managed HIE form the basis for population clinical decision support. For example, an HIE-based ASCVD risk dashboard could be deployed to aggregate the same risk factors, 10-year risk estimates, and statin guideline recommendations reported to clinicians through the HIE web portal or made available directly within a clinician's EHR using technologies such as SMART-on-FHIR to access the HIE data. Support for such a population health intervention includes reducing barriers to risk estimation for clinicians [Bakhai et al., 2018, Rashid et al., 2022] and addressing the patient risk perception gap [Fung et al., 2018], which have been shown to improve statin use. There are promising potential applications for several chronic diseases managed with pharmacotherapy, including hypertension and diabetes, using centralized HIE.

Finally, the standardized collection of ASCVD risk factors and outcomes using aggregate population health data is ripe for developing risk models. Population-specific models would avoid many of the known pitfalls of the PCE, including population-dependent calibration [Colantonio et al., 2017, DeFilippis et al., 2017, Emdin et al., 2017, Zhang et al., 2017], and could begin to address existing race-based differences in treatment [Vasan and van den Heuvel, 2022].

Study Limitations

This study has some noteworthy limitations. Blood pressure measurements are a relatively new contribution to CurrentCare by many data partners, which explains why the median risk follow-up is significantly shorter than the study length. While this limits our ability to analyze statin adherence beyond a few years, the demonstrated population health monitoring utility of HIE and population statin guideline adherence remain strong. With the continued maturation of the HIE through data partner onboarding, the ability to perform comprehensive population assessment will only improve. Ideally, our assessment of statin adherence would have been performed using data that included days of supply dispensed. This would have enabled a more granular (as opposed to binary) assessment of statin adherence. While this is an area of ongoing interest, the concordance of our findings at a population level with those in the previously published NHANES study is reassuring.

An additional lack of precision in our results can be traced to deidentification requirements. In particular, self-identified races other than "White" and "Black or African American" (e.g., Asian) were redacted, and birth years were binned into five-year ranges leading to reduced precision of risk estimation. Future studies with narrower data scopes may allow these data to be uncensored while complying with deidentification concerns. Finally, CurrentCare does not have a method for identifying individuals who have moved out of the RI healthcare catchment area. In our study, these individuals remain under observation without additional data being collected, reducing the accuracy of pertinent negative database queries. However, we believe the impact of this to be small.

Study Conclusions

This cohort study followed CurrentCare enrollees from 2013 to 2023 for ASCVD risk factors and statin use patterns across RI healthcare institutions. Disparities in risk factors, 10-year risk, and statin therapy were explored at the intersections of primary care location and racial and ethnic minority populations. Overall, statin guideline adherence was poor for both intermediate and high-risk groups. Individuals receiving care at FQHCs were at higher risk of incident ASCVD, more likely self-identify as Black and Hispanic, and more likely to be on guideline-directed statin therapy. Despite increased adherence to statin therapy guidelines, on average, FQHC patients' total cholesterol remained elevated compared to those receiving primary care at Health Systems.

3.3 Summary

Specific Aim 1 of this dissertation hypothesized that health information exchange (HIE) aggregates cardiovascular disease risk factors that inform primary prevention care decisions for individual patients that would otherwise be missed by a single healthcare institution. This hypothesis was affirmed by the work presented in the preceding chapter. The Rhode Island HIE was transformed into the observational medical outcomes partnership common data model. Atherosclerotic cardiovascular disease (ASCVD) risk factors were analyzed, 10-year ASCVD risk was estimated, and primary prevention practices with statins were assessed on 62,999 individuals from 2013 to 2023. The majority of these individuals had risk factors measured at more than one healthcare institution. Analysis of risk-based primary prevention with statins demonstrated that population cholesterol guideline adherence was poor at intermediate and high 10-year risk of incident ASCVD. These analyses demonstrated

the utility of centrally managed HIE for population health monitoring and informing population clinical decision support systems.

Chapter 4

Unsupervised Multi-Omics Analysis to Identify Disease Driving Biochemistry

Specific Aim 2 of this dissertation is to identify novel genomic and proteomic determinants of circulating levels of clinically relevant plasma metabolites (Figure 4.1). This aim hypothesizes that instrumental variable analysis of multi-omics data derived from banked human plasma samples will identify novel causal inter-biomarker associations. Toward this goal, this chapter presents an unsupervised analysis of protein-metabolite associations measured in human plasma across three population studies, including, the Jackson Heart Study (JHS), the Multi-Ethnic Study of Atherosclerosis (MESA), and the Health, Risk Factors, Exercise Training and Genetics (HERITAGE) Family study. In aggregate, multi-omics measurements (i.e., genomics, proteomics, and metabolomics) on 3,626 individuals were analyzed. Single nucleotide polymorphisms (SNPs) that influence circulating levels of proteins were used as instrumental variables for Mendelian randomization to test for causal associations between proteins and metabolites. Top associations were validated *in vivo* by performing metabolomics on knockout mice. The aim demonstrates the use of observational data from epidemiology studies to triage experimental validation of clinically relevant biochemical relationships.

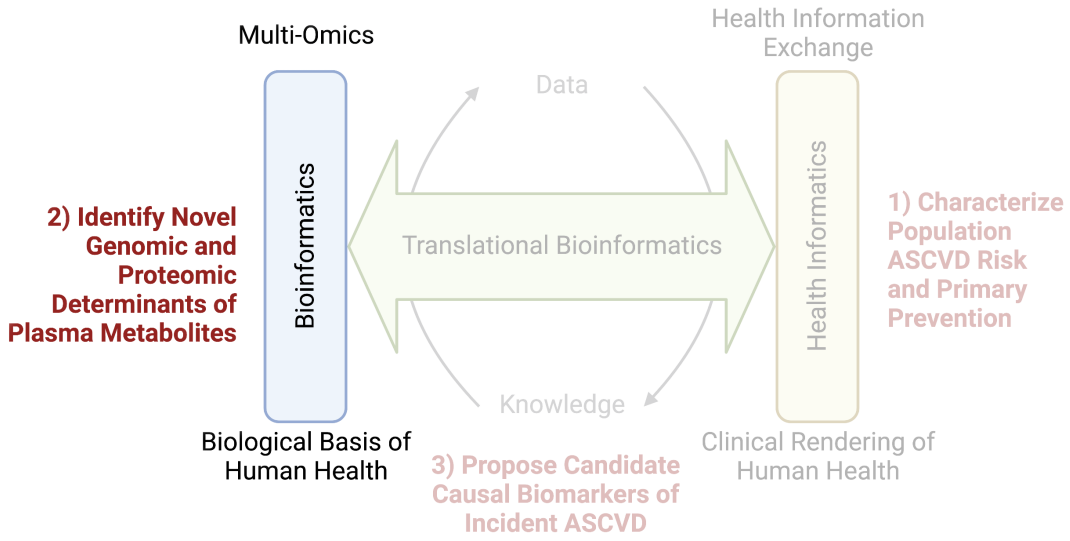

Figure 4.1: Overview of Specific Aim 2

4.1 Proteomic and Metabolomics of Human Plasma Samples

This chapter is a derivative of work that has been submitted to and is under revision by Cell Metabolism (12/2022). The work has been submitted as equal contribution and co-first authorship with Mark Benson and under senior mentorship of Robert Gerszten.

The integration of metabolomic and proteomic profiling data from large-scale population studies offers the opportunity to connect circulating proteins and metabolites as pathway partners in human physiology. Mass spectrometry-based metabolomics approaches measure low-molecular weight lipids, organic acids, nucleic acids, and other key chemical mediators of central metabolic and signaling pathways. Affinity-based proteomics techniques measure many of the secreted enzymes, transporters, cytokines, peptide hormones, and other proteins that catalyze and regulate these pathways. For example, small molecule profiling techniques can measure thyroxine, while proteomics can measure the transporter thyroxine-binding globulin that regulates circulating levels of this metabolite in the thyroid hormone signaling pathway. Similarly, the amino acids aspartate and glutamate, as well as the enzyme aspartate transaminase that catalyzes the interconversion of these two compounds can be assayed by these complementary techniques in the same biological sample. Combining metabolomics

and proteomics data may provide insights into new transporter-ligand, enzyme-substrate, and other protein-metabolite pairs that can be used for pathway discovery. However, the systematic integration of metabolomics and proteomics data from large-scale population studies to identify these biological relationships is still in its nascent stages.

Genome-wide association studies (GWAS) of plasma metabolites [Gieger et al., 2008, Illig et al., 2010, Kettunen et al., 2012, Long et al., 2017, Lotta et al., 2021, Rhee et al., 2013, Rhee et al., 2016, Shin et al., 2014, Suhre et al., 2011, Tahir et al., 2022, Yin et al., 2022, Yousri et al., 2018] and protein levels [Benson et al., 2017, Di Narzo et al., 2017, Emilsson et al., 2018, Ferkingstad et al., 2021, Folkersen et al., 2017, Gudjonsson et al., 2022, Katz et al., 2022, Lourdusamy et al., 2012, Pietzner et al., 2021, Pietzner et al., 2020, Png et al., 2021, Suhre et al., 2017, Sun et al., 2018, Yao et al., 2018, Zhong et al., 2021] in large-scale population studies have been increasingly leveraged to identify causal determinants of circulating factors in human plasma. For example, plasma proline levels are strongly associated with genetic variants in the PRODH locus that encodes proline dehydrogenase, as well as variants in other enzymes that are important in the catabolism of this metabolite [Kettunen et al., 2012, Rhee et al., 2013]. Similarly, GWAS of thrombin protein levels and other blood clotting factors have confirmed pathway relationships within the coagulation cascade [Katz et al., 2022, Olson et al., 2015]. While novel gene-metabolite or gene-protein associations have been identified and then validated in experimental model systems [Benson et al., 2017, Carayol et al., 2017, Kraus et al., 2015, Rhee et al., 2013, Solomon et al., 2016], systematic studies that leverage human genetics to illuminate novel protein-metabolite associations are lacking.

To determine causal associations between circulating proteins and metabolites, mass spectrometry-based metabolomics and aptamer-based proteomics profiling of plasma samples were analyzed from 3,626 individuals in three cohorts: the Jackson Heart Study (JHS),[1] the Multi-Ethnic Study of

[1]The Jackson Heart Study (JHS) is supported and conducted in collaboration with Jackson State University (HHSN268201800013I), Tougaloo College (HHSN268201800014I), the Mississippi State Department of Health (HHSN268201800015I/HHSN26800001) and the University of Mississippi Medical Center (HHSN268201800010I, HHSN268201800011I and HHSN268201800012I) contracts from the National Heart, Lung, and Blood Institute and the National Institute for Minority Health and Health Disparities (NIMHD). Molecular data for the Trans-Omics in Precision Medicine (TOPMed) program was supported by the National Heart, Lung and Blood Institute. Genome sequencing for "NHLBI TOPMed: The Jackson Heart Study" (phs000964.v1.p1) was performed at the Northwest Genomics Center (HHSN268201100037C). Core support including centralized genomic read mapping and genotype calling, along with variant quality metrics and filtering were provided by the TOPMed Informatics Research Center (3R01HL-117626-02S1; contract HHSN268201800002I). Core support including phenotype harmonization, data management, sample-identity QC, and general program coordination were provided by the TOPMed Data Coordinating Center (R01HL-120393; U01HL-120393; contract HHSN268201800001I). We gratefully acknowledge the studies and participants who provided biological samples and data for TOPMed. The authors wish to thank the staff and

Atherosclerosis (MESA),[2] and the Health, Risk Factors, Exercise Training and Genetics (HERITAGE)[3] Family study. In total, the relationships between circulating levels of 1302 proteins and 365 metabolites were measured in the same banked plasma samples. Toward this goal, every pairwise protein-metabolite combination was evaluated for correlation and then correlations were examined in aggregate for enrichment by specific metabolite class. Genetic data from each study was leveraged to perform Mendelian randomization (MR) analyses to identify putative causal relationships of circulating proteins with plasma metabolite levels. Top protein-to-metabolite MR associations were experimentally validated in proof-of-concept plasma metabolomics studies in three murine knockout models. Further, all protein-metabolite association results have been provided as a publicly available dataset for pathway discovery in human metabolism and disease (linked from: https://github.com/aeisman/protein-metabolite).

4.2 Triaging Proteomic and Metabolomic Association Studies with Enrichment Analyses

We studied the relationships between circulating levels of 1302 proteins and 365 metabolites measured in fasting plasma samples from participants of the Jackson Heart Study (JHS, n=1985), the Multi-Ethnic Study of Atherosclerosis (MESA, n=983), and the Health, Risk Factors, Exercise Training and Genetics (HERITAGE) Family study (n=658). An overview of the study design is

participants of the JHS.

[2]Whole genome sequencing (WGS) for the Trans-Omics in Precision Medicine program was supported by the National Heart, Lung and Blood Institute. WGS for "NHLBI TOPMed: Multi-Ethnic Study of Atherosclerosis (MESA)" (phs001416.v1.p1) was performed at the Broad Institute of MIT and Harvard (3U54HG003067-13S1). Centralized read mapping and genotype calling, along with variant quality metrics and filtering were provided by the TOPMed Informatics Research Center (3R01HL-117626-02S1). Phenotype harmonization, data management, sample-identity QC, and general study coordination, were provided by the TOPMed Data Coordinating Center (3R01HL-120393-02S1). The MESA projects are conducted and supported by the National Heart, Lung, and Blood Institute in collaboration with MESA investigators. Support for the Multi-Ethnic Study of Atherosclerosis (MESA) projects are conducted and supported by the National Heart, Lung, and Blood Institute in collaboration with MESA investigators. Support for MESA is provided by contracts 75N92020D00001, HHSN268201500003I, N01-HC-95159, 75N92020D00005, N01-HC-95160, 75N92020D00002, N01-HC-95161, 75N92020D00003, N01-HC-95162, 75N92020D00006, N01-HC-95163, 75N92020D00004, N01-HC-95164, 75N92020D00007, N01-HC-95165, N01-HC-95166, N01-HC-95167, N01-HC-95168, N01-HC-95169, UL1-TR-000040, UL1-TR-001079, and UL1-TR-001420. Also supported in part by the National Center for Advancing Translational Sciences, CTSI grant UL1TR001881, and the National Institute of Diabetes and Digestive and Kidney Disease Diabetes Research Center (DRC) grant DK063491 to the Southern California Diabetes Endocrinology Research Center.

[3]We thank Drs. Arthur S. Leon, D.C. Rao, James S. Skinner, Tuomo Rankinen, Jacques Gagnon, and the late Jack H. Wilmore for contributions to the planning, data collection, and conduct of the HERITAGE project. This research was partially funded by National Heart, Lung, and Blood Institute Grants HL-45670, HL-47317, HL-47321, HL-47323, and HL-47327, all in support of the HERITAGE Family Study. C. B. is partially funded by the John W. Barton Sr. Chair in Genetics and Nutrition, and NIH COBRE grant (NIH P30GM118430-01). Dr. Sarzynski is supported by R01HL146462.

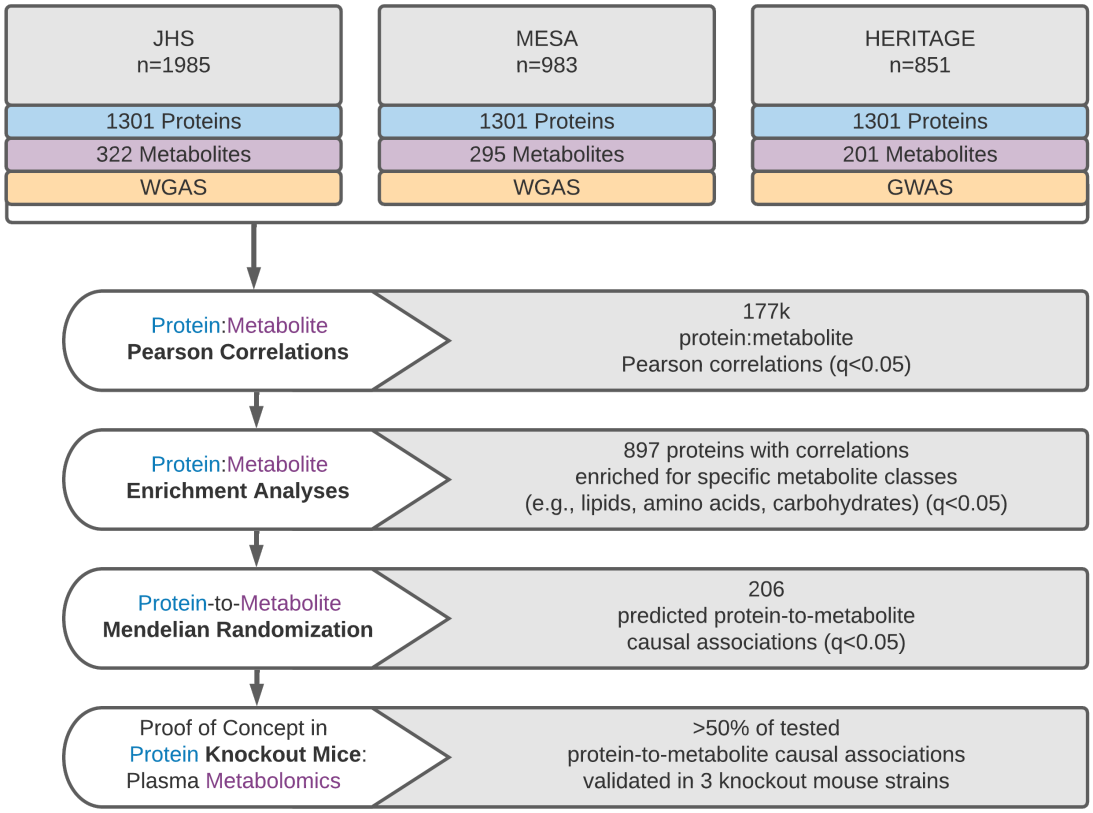

Figure 4.2: The integration of human plasma proteomic, metabolomic, and genomic profiling data for pathway discovery. Flow diagram detailing the experimental pipeline and main results from the integration of plasma proteomic, metabolomic, and genomic profiling datasets in the Jackson Heart Study (JHS), Multi-Ethnic Study of Atherosclerosis (MESA), and Health, Risk Factors, Exercise Training and Genetics Study (HERITAGE Family study). WGAS - whole genome association study, GWAS - genome wide association study.

provided in Figure 4.2, which included 1) correlation analyses between every pairwise combination of each protein and metabolite, 2) enrichment analyses to detect proteins that are highly associated with specific classes of metabolites, 3) Mendelian randomization (MR) analyses to identify putative causal relationships of circulating proteins and metabolite plasma levels, and 4) experimental validation of a subset of the top protein-to-metabolite MR associations in proof of concept plasma metabolomics studies in three murine knockout models.

Protein-metabolite correlations in human plasma

Pearson correlation coefficients were calculated for every pairwise protein-metabolite combination within the JHS, MESA, and HERITAGE Family study using age- and sex-adjusted, log-normalized, and standardized protein and metabolite levels. We identified a set of protein and metabolite pairs in each cohort that were significantly correlated with an FDR-adjusted q-value ≤ 0.05 (Figure 4.3a). Meta-analyses of these correlations across the three studies identified 171,800 significant protein-metabolite correlations (q-value ≤ 0.05). Ninety-seven percent of these correlations remained significant when further adjusted for body mass index (BMI), and 87% remained significant when additionally adjusted for estimated glomerular filtration rate (eGFR) to account for potential effects of kidney function on the circulating proteins in multivariable adjusted (MVA) analyses (Figure 4.3b). The magnitude and directionality of correlation coefficients were consistent across the individual studies (Figure 4.3c).

As anticipated, several of the most significant correlations reflected well-characterized protein-metabolite biological relationships. These included associations between plasma binding proteins such as thyroxine binding globulin and thyroxine (correlation coefficient = 0.51, q-value $\leq$ 1.0 x 10E-300), plasma transporters such as apolipoprotein E (APOE) and lipids including diacylglycerol C36:3 (correlation coefficient = 0.51, q-value $\leq$ 1.0 x 10E-300), and plasma enzymes such as aspartate transaminase (AST) and its canonical substrate aspartate (correlation coefficient = -0.07, q-value = 8.4 x 10E-5) and product glutamate (correlation coefficient = 0.23, q-value = 5.3 x 10E-48). Visualization of the tens of thousands of additional, previously unexplored protein-metabolite correlations using a heat map demonstrated distinct patterns of correlations between individual proteins and members of specific metabolite classes (Figure 4.4d). For example, the protein hormone insulin demonstrated positive correlations with metabolites within the carbohydrate, glycerolipid, acyl carnitine, branched chain amino acid, and aromatic amino acid classes and inverse correlations with metabolites within the lysophosphatidylethanolamine (LPE), lysophosphatidylcholine (LPC), and polar uncharged amino acid classes (Figure 4.4e). The hormones adiponectin and ghrelin similarly demonstrated strong metabolite correlations within lipid and amino acid class lines. Finally, the hormone fibroblast growth factor 19 (FGF19), a regulator of bile acid synthesis [Holt et al., 2003], demonstrated marked positive correlations within the bile acid metabolite class. While the clustering of these correlations within specific metabolite classes was consistent with known functions

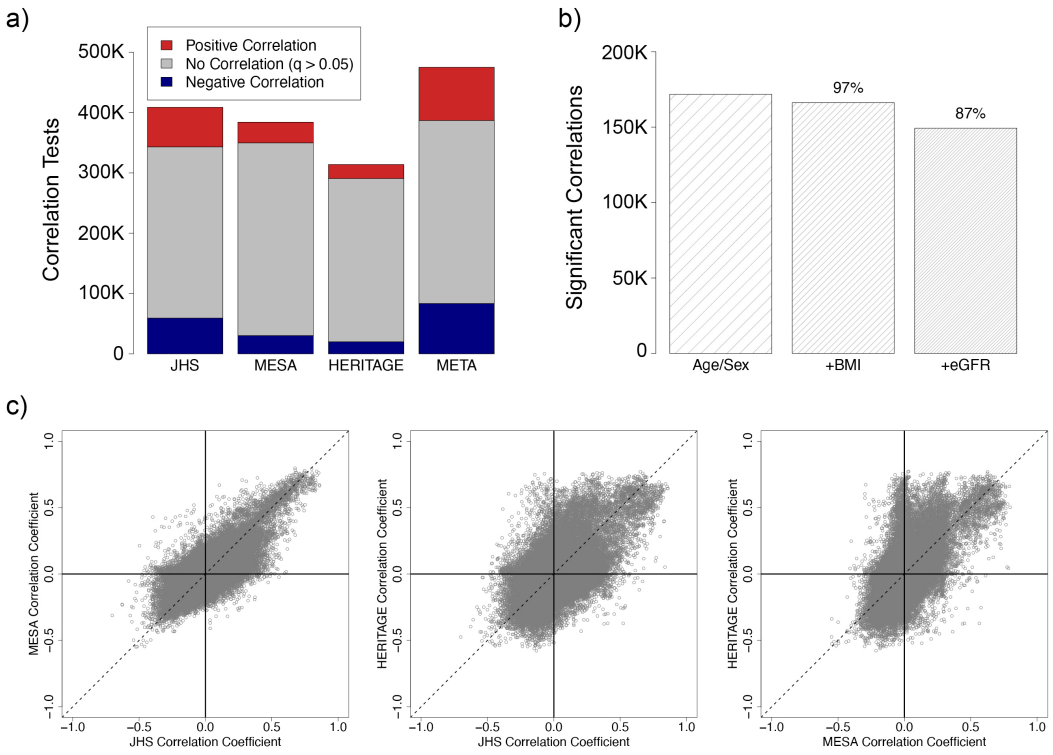

Figure 4.3: Protein-metabolite correlations in human plasma (panel 1). Pearson correlation coefficients were calculated for every pairwise protein-metabolite combination within the JHS, MESA, and HERITAGE Family study using age- and sex-adjusted, log-normalized, and standardized protein and metabolite levels. (A) A subset of proteins and metabolites were positively (red) or negatively (blue) correlated with an FDR-adjusted q-value $\leq$0.05 in each study, and in a meta-analysis of the three studies. (B) Ninety-seven percent of the meta-analyzed age- and sex-adjusted protein-metabolite correlations remained significant with an FDR-adjusted q-value $\leq$0.05 when further adjusted for BMI, and 87% remained significant when additionally adjusted for eGFR. (C) The magnitude and directionality of the meta-analyzed age- and sex-adjusted protein-metabolite correlations were consistent across studies.

of each of these central metabolic hormones, these analyses also provided novel details regarding interactions with specific subclasses of metabolite species for each protein, particularly in the context of human physiology. For example, insulin demonstrated strong positive correlations with several saturated fatty acids (e.g., butyric acid, q-value 5.6 x 10E-3) and quinolone carboxylic acids (e.g., xanthurenic acid, q-value 7.3 x 10E-10), as well as inverse correlations with unsaturated fatty acids (e.g., linoleic acid, q-value 3.7 x 10E-3), amino fatty acids (e.g., 2-aminoisobutyric acid, q-value 4.9 x 10E-8), N-acyl amines (e.g., N-oleoyl-glycine, q-value 8.1 x 10E-8) and dicarboxylic acids (e.g., malonic acid, q-value 2.9 x 10E-3).

Protein correlations are enriched for specific classes of metabolites in human plasma

To characterize protein-metabolite correlations more systematically, we investigated whether ranked metabolite correlations for each protein were significantly enriched for specific sets of metabolite classes, analogous to Gene Set Enrichment Analyses (GSEA) [Mootha et al., 2003, Subramanian et al., 2005]. As an initial proof of concept, we confirmed that the most significantly correlated metabolites with plasma levels of APOE protein were members of the lipid metabolite class, consistent with the well-established role of APOE in lipid transport (Figure 4.5a). To quantitate this overrepresentation of lipid metabolites among the strongest correlations with APOE, a plot of the running sum statistic for these lipids in the ranked correlation data was used to generate an enrichment score (ES)(Figure 4.5b). The enrichment score could be assessed for statistical significance from the null hypothesis of no enrichment of correlations for lipid metabolites for APOE using a permutation test as detailed in Methods. This analysis was then repeated for each of the 1302 measured plasma proteins in the dataset.

As shown in Figure 4.5c, we identified 241 proteins that were significantly enriched for correlations with plasma lipids (q-value ≤ 0.05). These included proteins with well-established roles in lipid metabolism such as proprotein convertase subtilisin/kexin type 9, apolipoprotein B, low-density lipoprotein receptor-related protein 1B, and angiopoietin-like 4. Many novel protein-lipid findings included associations with several members of the cathepsin proteases (CTSA, CTSB, and CTSF), serpin peptidase inhibitors (AGT, SERPING1, and SERPIND1), and secreted glycoproteins (NID1,

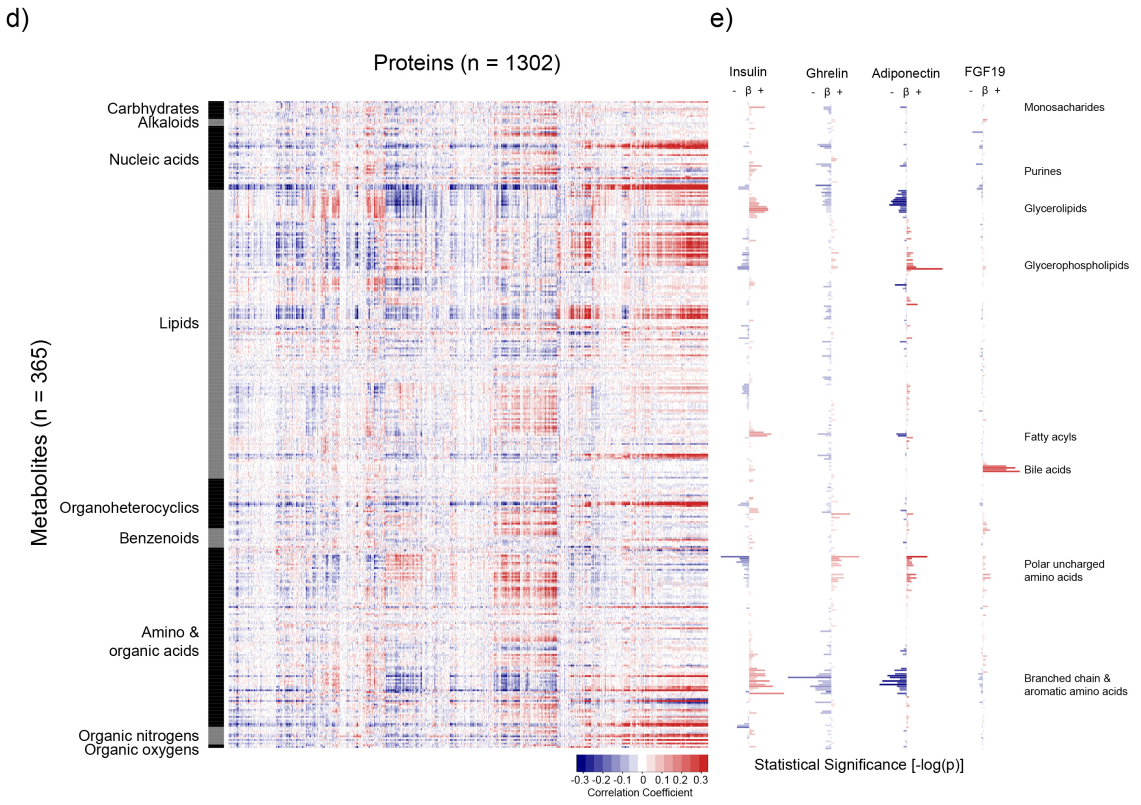

Figure 4.4: Protein-metabolite correlations in human plasma (panel 2). Pearson correlation coefficients were calculated for every pairwise protein-metabolite combination within the JHS, MESA, and HERITAGE Family study using age- and sex-adjusted, log-normalized, and standardized protein and metabolite levels. (D) Visualization of the protein-metabolite correlations using a heat map demonstrated distinct patterns of associations between individual proteins and members of specific classes of metabolites. Individual metabolites were ordered according to RefMet class along the y-axis, and proteins were allowed to order using a hierarchical cluster analysis along the x-axis. The magnitude and directionality of the correlation coefficient for each protein-metabolite association is depicted by color, as indicated in the legend. (E) The statistical significance (-log(p-value)) of the correlation between each metabolite and four representative protein hormones is shown.

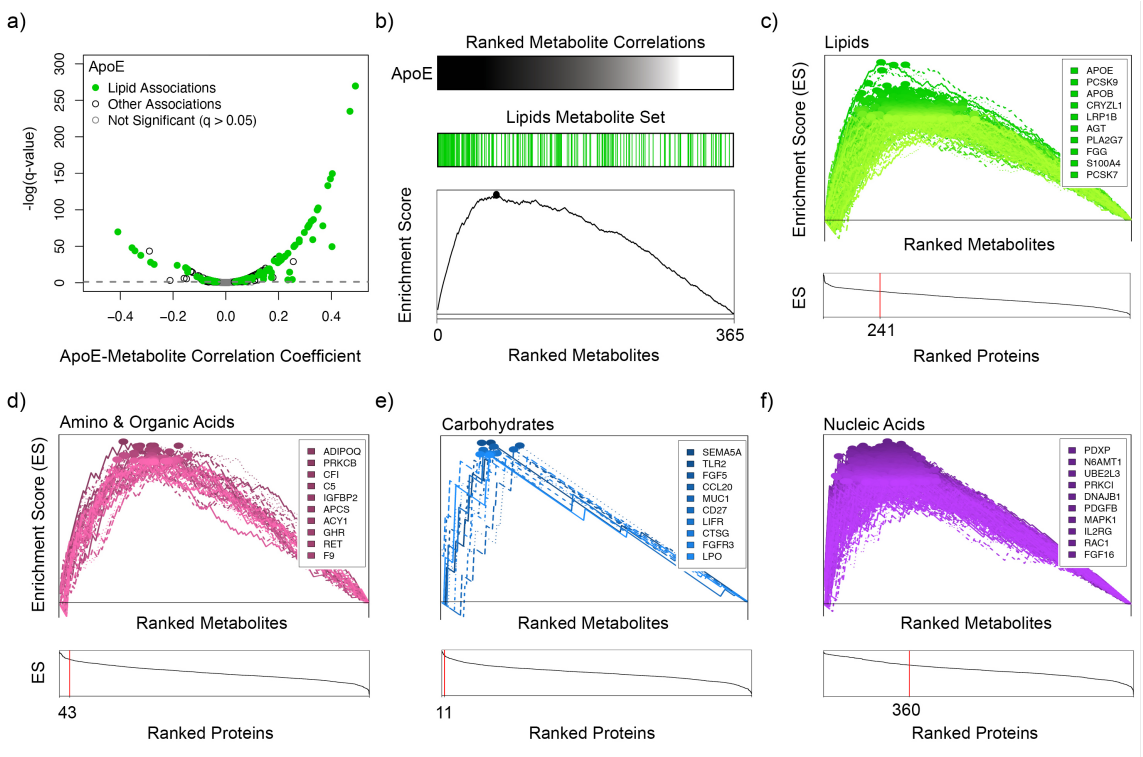

Figure 4.5: Protein correlations are significantly enriched for specific metabolite classes. (A) A volcano plot demonstrates that the most significantly correlated metabolites with plasma levels of APOE protein were members of the lipid metabolite class. (B) To evaluate this enrichment for lipids more quantitatively, an enrichment score (ES) was computed and visualized as the maximum point (marked with •) of a running sum statistic that increases proportionally with each lipid and decreases with non-lipids along the ranked list of metabolite correlations with plasma levels of APOE. (C) 241 proteins were significantly enriched for correlations with plasma lipids (FDR-adjusted q-value ≤ 0.05). In the top panel, the ES tracings are shown for each individual protein, and the ten proteins with the highest enrichment scores are listed. In the bottom panel, proteins are ordered in descending order of calculated enrichment scores (horizontal axis), and the number of proteins significantly enriched for correlations with lipids (FDR-adjusted q-value ≤ 0.05) is indicated with the vertical red line. Similar enrichment analyses are shown for Amino and Organic Acids (D), Carbohydrates (E), and Nucleic Acids (F).

NID2, LAMA1, and GPC6). Further, the broad survey of proteins in this analysis identified additional protein pathway partners that demonstrated enriched correlations for lipid metabolites. A notable example of this included the identification of a pathway node centered on the scavenger receptor CD36, which functions as a high-affinity receptor for long chain fatty acids and other ligands in rodent models and has been implicated in fat metabolism traits in humans [Coburn et al., 2000, Harmon and Abumrad, 1993, Ibrahimi et al., 1999, Melis et al., 2017, Yanai et al., 2000]. The CD36 receptor was not only itself highly enriched for correlations with plasma lipid metabolites (q-value = 1.1 x 10E-3), but two well-established regulatory protein ligands of the receptor, thrombospondin 1 [Asch et al., 1987] and CD5 Molecule Like [Kurokawa et al., 2010], were also highly enriched for correlations with plasma lipids (THBS1, q-value = 8.3 x 10E-3; CD5L, q-value = 4.5 x 10E-3), highlighting the potential role for this receptor and associated ligands in human lipid metabolism.

We next expanded our analysis to identify proteins with correlations that were enriched for additional major classes of metabolites (Figure 4.5d-f). Interestingly, we identified 360 proteins with correlations that were significantly enriched for circulating nucleic acids. Among these were proteins with well-established roles in nucleotide metabolism, such as nucleoside diphosphate kinase B, nucleoside diphosphate kinase A, ectonucleoside triphosphate diphosphohydrolase 1, thymidine kinase 1, and adenylate kinase isoenzyme 1. These nucleoside kinases maintain the balance between nucleoside mono-, di- and triphosphates (e.g., AMP, ADP, and ATP) and several are known to circulate in human plasma with 1 nM concentrations [Desiere et al., 2006]. While they have been demonstrated to be secreted and to regulate extracellular ATP synthesis in model systems [Choo et al., 2008, Romani et al., 2018, Yokdang et al., 2011], their role in human plasma has not previously been fully elucidated.

We also identified 43 proteins with correlations that were significantly enriched for amino and organic acids. Several of these included proteins with well-established roles in the regulation of protein metabolism such as growth hormone receptor, insulin-like growth factor binding protein 2, and adiponectin. Interestingly, one of the top proteins enriched for correlations with amino and organic acids was the enzyme aminoacylase-1 (ACY1, q-value = 2.9 x 10E-3) which can hydrolyze N-acetyl-amino acids to free amino acids in isolated human and murine plasma [Ngo et al., 2021].

These enrichment data suggest that ACY1 may play a broader role in human plasma amino acid homeostasis, extending prior observations [Corey et al., 2022, Emilsson et al., 2018, Sass et al., 2006, Van Coster et al., 2005].

4.3 Protein-Metabolite Association Studies Identify Novel Proteomic Determinants of Metabolite Levels in Human Plasma

To identify potential causal relationships of circulating proteins on metabolite levels in human plasma, we next performed Mendelian randomization (MR) analyses. This approach leveraged whole genome- or genome-wide association studies (WGAS, GWAS) of each plasma protein and metabolite level in JHS, MESA, and HERITAGE Family study participants. Genetic variants within 1 mega-base (Mb) of the coding gene for each protein ("cis" variants) that were independent (linkage disequilibrium r2 ≤ 0.001) and strongly associated with circulating levels of the protein (Bonferroni-adjusted p-value ≤ 0.05) were used as instrumental variables to assess whether plasma levels of each protein (exposure) had a causal effect on correlated plasma metabolite levels (outcome) using the inverse-variance weighted (IVW) method [Burgess et al., 2017a, Burgess et al., 2013, Burgess et al., 2019, Burgess and Thompson, 2015, Davey Smith and Hemani, 2014, Smith and Ebrahim, 2003]. Several robust methods, including the limited information maximum likelihood (LIML) [Burgess et al., 2011], as well as the MR-Egger [Bowden et al., 2015], median [Han, 2008], and median-weighted [Bowden et al., 2016] robust methods when instruments contained more than two genetic variants, were used to assess the sensitivity of these analyses, as described in Methods.

We found that 547 of the 1302 proteins had cis variants that could be used as instrumental variables in MR analyses (Figure 4.6a). Proteins with available cis instruments spanned the genome (Figure 4.6b), and the majority of instruments were located in very close proximity to the transcriptional start site (TSS) of the protein coding gene (Figure 4.6c). We restricted our analyses to include only instruments in cis to the coding gene for each protein so that the effect of these instruments on the metabolite was likely to run through the protein exposure, rather than through an alternative, potentially pleiotropic biochemical pathway [Mokry et al., 2015, Sarwar et al., 2012, Wensley et al.,

2011].

In total, we identified 224 putative protein-to-metabolite causal associations between 52 proteins and 146 metabolites that were highly significant (q-value $\leq$ 0.05) (Figure 4.6d). Among the top MR findings were several examples of the well-established causal role that APOE protein plays in modulating plasma levels of lipids, fat-soluble vitamins, and cholesterol. These included MR associations between APOE protein and several glycerophospholipids (e.g., C38:5 PE plasmalogen; IVW beta -0.29, q-value 1.9x10E-7, LIML q-value 1.0x10E-6), phosphosphingolipids, and the lipid-soluble vitamin retinol.

These analyses also detected strong putative causal associations for several other proteins that were identified to have correlations with lipids, amino and organic acids, and nucleic acids in the enrichment analyses above. For example, we detected strong MR associations between the CD36 scavenger receptor and numerous lipid species not previously associated with this protein, including glycerophospholipids (e.g., C38:7 PE plasmalogen; IVW beta -0.28, q-value 1.2x10E-15, LIML q-value 9.6x10E-15), acyl carnitines, sphingomyelins, ceramides, and steroids (Figure 4d). Notably, we detected strong MR associations between CD36 and several polyunsaturated fatty acids, including eicosapentaenoic acid (EPA), docosahexaenoic acid (DHA), and arachidonic acid. These polyunsaturated fatty acids play a key role in eicosanoid signaling. CD36 has been implicated in the cellular uptake of polyunsaturated fatty acids using *in vitro* models [Franekova et al., 2015, Glatz and Luiken, 2015]. However, these findings may suggest a broader role for CD36 as a central regulator of lipid homeostasis in human plasma.

There were also strong MR associations between the enzyme ACY1 and several N-acetyl amino acids, including N-acetyl-glutamate (IVW beta -1.25, q-value 1.9x10E-33, LIML q-value 4.5x10E-12), N-acetyl-alanine, N-acetyl-glutamine, and N-acetyl-serine (Figure 4d). Circulating levels of N-acetyl amino acids are known to be tightly regulated and have recently been tied to several cardiometabolic phenotypes in human population studies, such as insulin resistance47, incident coronary artery disease [Cruz et al., 2022], and incident heart failure [Tahir et al., 2021]. These MR results are consistent with the known role of ACY1 in modulating the levels of these N-acetyl-amino acids in human plasma.

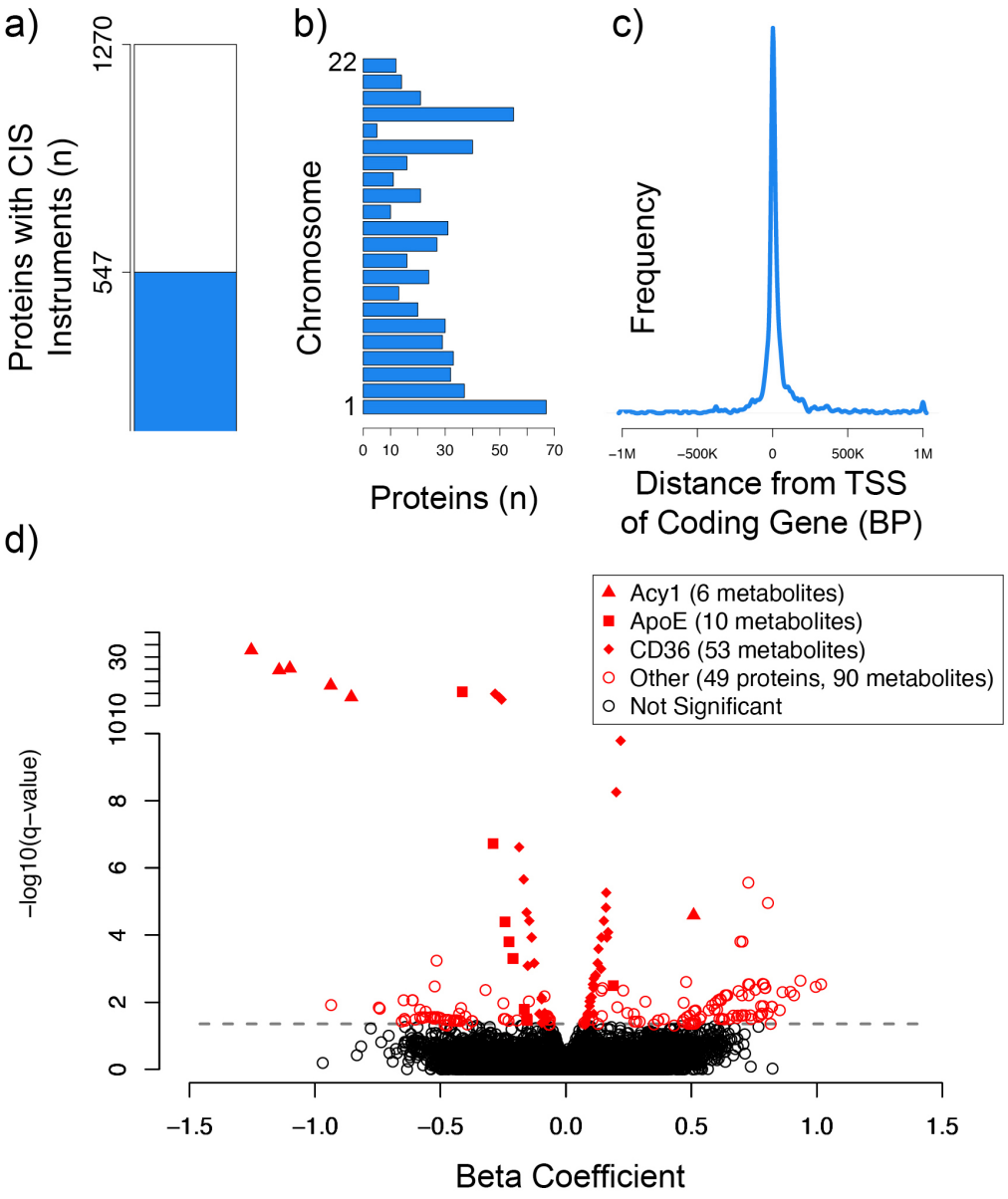

Figure 4.6: Mendelian randomization analyses identify putative causal protein-to-metabolite associations in humans. Proteins with at least one pQTL in cis (located within 1Mb) of the protein coding gene that could be used as an instrumental variable (IV) in Mendelian randomization (MR) analyses are depicted in blue (A). The proteins with available cis instruments were distributed evenly across the genome (B). pQTLs used in IVs were generally located near the transcriptional start site (TSS) of the protein coding gene (C). A volcano plot depicts the 224 significant MR associations between 52 proteins and 146 metabolites with an FDR-adjusted q-value ≤ 0.05 (D). The three proteins with the most significant MR metabolite associations are depicted by distinct shapes described in the figure legend.

Intriguingly, proprotein convertase subtilisin/kexin type 9 (PCSK9) demonstrated putative causal associations with multiple acyl carnitines, including CAR 18:1 (IVW beta -0.52, q-value 5.8 x 10E-4, LIML q-value 1.5 x 10E-3), CAR 18:2, CAR 16:0, and CAR 14:1 (Figure 4d). Acyl carnitines were among the most strongly associated metabolites with PCSK9 protein in pairwise Pearson correlation analyses (e.g., CAR 18:1 correlation coefficient -0.26, q-value 6.5 x 10E-64). Carnitines play a key role in the regulation of energy metabolism by facilitating the transport of long-chain fatty acids from adipose tissues to target cells. While PCSK9 has recently been shown to reduce the uptake of long-chain fatty acids by adipocytes in a cell culture system [Demers et al., 2015], these MR association data suggest a potential role for PCSK9 in regulating the carnitine transport system in human plasma.

Protein-to-metabolite causal associations predicted by MR analyses in human plasma experimentally validated in three murine knockout models.

As a proof of concept to test the causal protein-to-metabolite associations predicted by MR analyses above, we conducted plasma metabolomics on available C57BL/6 murine knockout (KO) strains for the three proteins that had the most significant MR metabolite associations (CD36, APOE, and ACY1) and compared these to wild-type (WT) controls.

The CD36 scavenger receptor was predicted to have a causal association with 68 metabolites in the human MR analyses above (q-value ≤ 0.10), 50 of which were also measured in the CD36 KO mouse. We identified significant differences in the plasma levels of 27 of these metabolites (54%) in metabolomic profiling studies of the CD36 KO animals (n=6) versus WT controls (n=8; $p \leq 0.05$) (Figure 4.7). These included experimental validation of predicted causal relationships with specific glycerophospholipids, sphingolipids, and fatty acyls, including causal associations of CD36 with circulating levels of the central signaling fatty acids docosahexaenoic acid (DHA; CD36 KO/WT fold-change = 0.73 $\pm$ 0.06, p-value = 0.01) and arachidonic acid (CD36 KO/WT fold-change = 0.74 $\pm$ 0.04, p-value = 0.01).

Similarly, APOE was predicted to have causal associations with 13 metabolites in the human MR analyses (q-value ≤ 0.10), and we detected significant differences with the expected directionality in the plasma levels of four of these metabolites in APOE KO animals (n=6) versus WT controls

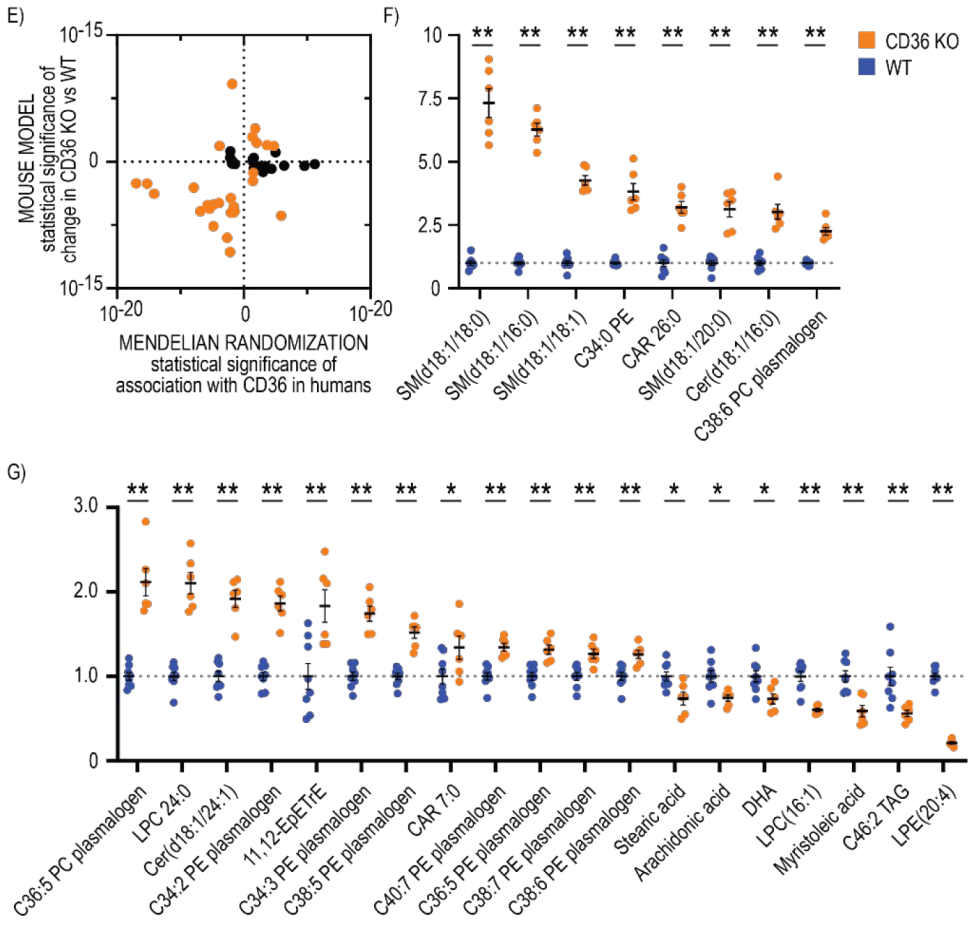

Figure 4.7: Protein-to-metabolite causal associations predicted by Mendelian randomization analyses in humans experimentally validate in murine knockout models (panel 1). Plasma metabolomics was conducted on C57BL/6 murine knockout (KO) strains for CD36 (n=6), APOE (n=6), and ACY1 (n=6) and compared to wild type (WT) controls (n=8). Scatterplots depict the number of predicted protein-to-metabolite MR associations in humans (with $q \leq 0.1$) that validated in each murine model (with p$\leq$0.05, highlighted in color), as well as the concordance in directionality of these associations (a, c, e). The position on the x axis represents the p-value of the predicted MR association between each protein and metabolite level in the human studies. The position on the y axis represents the p-value of the difference in metabolite level between KO and WT mice. Metabolites on the right half of the scatterplots are predicted to be positively associated with each protein by MR, whereas metabolites on the left half are predicted to be inversely associated with each protein. Similarly, metabolites on the top half of the scatterplots were higher in KO vs WT animals, whereas metabolites on the bottom half were lower in KO vs. WT animals. The northwest and southeast quadrants of each scatter plot are shaded to depict consistent directionality between the MR predictions and KO experiments. The levels of each metabolite that were significantly different in KO vs. WT animals ($p \leq 0.05$) are shown as fold changes compared to WT animals (b, d, f). * indicates P<0.05, and ** indicates $P < 0.01$ by students two tailed t-test.

(n=8; p≤0.05) (Figure 4.7). Notably, each of the four experimentally validated metabolites were phosphatidylethanolamine (PE) plasmalogens, including C38:6 (APOE KO/WT fold-change = 1.30 ± 0.05, p-value = 9.47x10E-4), C38:5, C36:3, and C40:7 PE plasmalogens. PEs have been speculated to interact with APOE following the hepatic secretion of nascent very low density lipoprotein (VLDL) particles in cell-based *in vitro* studies [Agren et al., 2005, Hamilton and Fielding, 1989]. These findings may suggest a causal role for APOE in the regulation of circulation PE plasmalogen levels in human plasma.

Finally, the circulating enzyme ACY1 was predicted to have causal associations with five metabolites in human MR studies, and we detected directionally-consistent significant differences in the plasma levels of four of these metabolites (80%) in ACY1 KO animals (n=6) versus WT controls (n=8; $p \leq 0.05$)(Figure 4.7), including N-acetyl-glutamate (ACY1 KO/WT fold-change = 7.19 ± 0.57, p-value = 1.11x10E-6), N-acetyl-glutamine, N-acetyl-serine, and N-acetyl-alanine.

In total, we experimentally validated 35 of the 68 (51%) predicted protein-to-metabolite MR associations. Further, we experimentally validated 62 additional protein-metabolite associations with significant pairwise Pearson correlations (q ≤ 0.05) that either did not have an available MR instrumental variable or were not captured by MR analyses.

4.4 Integrated Multi-Omic Analysis of Banked Human Plasma Samples Augments Existing *in vitro* Evidence of Protein-Metabolite Associations

This study leveraged metabolomic and proteomic profiling of plasma samples from three human cohorts to determine if the integration of these datasets may identify novel causal relationships between specific circulating proteins and metabolites. The profiling data from the JHS, MESA, and HERITAGE Family studies were "harmonized," in that we used the same mass spectrometry-based metabolomics and aptamer-based proteomics platforms in parallel across each of the 3,626 samples, providing an ideal opportunity to perform protein-metabolite association studies. Additionally, genomic data were available for each participant, allowing for the study of putative protein-to-metabolite causal associations with Mendelian randomization analyses and follow-up studies in a

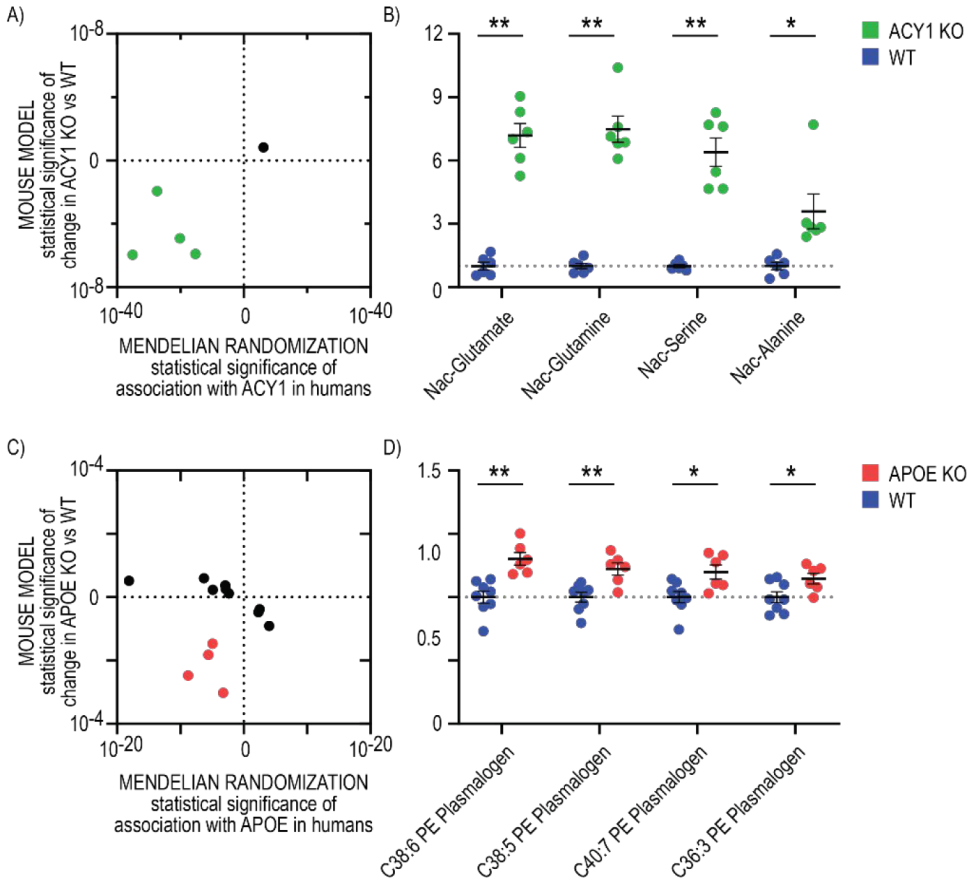

Figure 4.8: Protein-to-metabolite causal associations predicted by Mendelian randomization analyses in humans experimentally validate in murine knockout models (panel 2). Plasma metabolomics was conducted on C57BL/6 murine knockout (KO) strains for CD36 (n=6), APOE (n=6), and ACY1 (n=6) and compared to wild type (WT) controls (n=8). Scatterplots depict the number of predicted protein-to-metabolite MR associations in humans (with $q \leq 0.1$) that validated in each murine model (with p$\leq$0.05, highlighted in color), as well as the concordance in directionality of these associations (a, c, e). The position on the x axis represents the p-value of the predicted MR association between each protein and metabolite level in the human studies. The position on the y axis represents the p-value of the difference in metabolite level between KO and WT mice. Metabolites on the right half of the scatterplots are predicted to be positively associated with each protein by MR, whereas metabolites on the left half are predicted to be inversely associated with each protein. Similarly, metabolites on the top half of the scatterplots were higher in KO vs WT animals, whereas metabolites on the bottom half were lower in KO vs. WT animals. The northwest and southeast quadrants of each scatter plot are shaded to depict consistent directionality between the MR predictions and KO experiments. The levels of each metabolite that were significantly different in KO vs. WT animals ($p \leq 0.05$) are shown as fold changes compared to WT animals (b, d, f). * indicates P<0.05, and ** indicates $P < 0.01$ by students two tailed t-test.

select group of knockout mice. This analysis provides several important initial insights into the integration of metabolomic and proteomic profiling data for pathway discovery.

First, we show that known protein-metabolite associations that are key to established metabolic and signaling pathways (e.g., thyroxine binding globulin protein and thyroxine metabolite, aspartate transaminase (AST) protein and enzymatic substrate/products aspartate and glutamate metabolites) can be detected in banked samples from population studies. Further, as might be expected for proteins and metabolites that are related though a shared biological pathway, these associations persist despite adjustment for broad baseline characteristics of study participants (e.g., age, sex, BMI, eGFR), are reproducible across studies conducted at different geographical locations at different times, and in participants of diverse race and ethnicity.

Second, this study leverages protein-metabolite association data to link protein-metabolite relationships previously identified in model systems to human biology. We used protein-metabolite correlation data to perform enrichment analyses and identify several examples of protein-lipid, protein-amino acid, and protein-nucleic acid associations that have been studied in cell- and animal-based systems, but that have not to our knowledge been previously demonstrated in human plasma. For example, the secreted protease Cathepsin B has emerged as a potential novel lipid regulatory protein in several experimental model systems. Knockout of the Cathepsin B gene results in marked improvements in liver triglyceride and blood total cholesterol levels in a murine model of nonalcoholic fatty liver disease [Fang et al., 2020]. Mechanistically, Cathepsin B has been shown in cultured cell-based model systems to regulate very-low-density lipoprotein (VLDL) secretion and free fatty acid uptake by cleaving liver fatty acid-binding protein (LFAB) [Thibeaux et al., 2018]. The protein-metabolite enrichment analyses presented here extend these experimental findings and provide strong rationale for further study of Cathepsin B in human plasma lipid homeostasis.

Similarly, extensive experimental data have demonstrated a key role for adenylate kinase and ecto-nucleoside diphosphokinase nucleotide conversion enzymes in the regulation of extracellular ATP levels in cultured hepatocytes [Fabre et al., 2006], endothelial cells [Yegutkin et al., 2002], and lymphoid cells [Yegutkin et al., 2002], as well as in human vitreous fluid [Zeiner et al., 2019]. The protein-metabolite association data in the current study build on these mechanistic observations and suggest a role for these enzymes in the regulation of extracellular ATP and purine signaling that

is reflected in human plasma. Importantly, while we used an enrichment analysis strategy based on GSEA [Mootha et al., 2003, Subramanian et al., 2005] to interrogate the protein-metabolite association data in this study, additional methods could be employed to query these extensive data for pathway discovery. Thus, we have made the complete protein-metabolite association study dataset publicly available for further analyses.

Third, this study demonstrates how Mendelian randomization analyses can be leveraged to "triage" putative causal protein-to-metabolite associations from protein-metabolite correlation data for further experimental study. By integrating genetic data with our protein-metabolite association findings, we were able to use genetic variants located in or near the coding gene for a measured protein to examine the causal effect of that protein exposure on a metabolite outcome in human plasma. Approximately 40% of the measured proteins in our studies had strong, independent associations with genetic variants in cis to the protein cognate gene that could be used as instruments in MR analyses. Notably, the use of three cohorts representing diverse race/ethnicities and minor allele frequencies improved the ability to identify MR instruments. For example, the inclusion of African Americans in the JHS highlighted the variant rs2229152 (JHS MAF=1.7%) as an instrument for circulating ACY1 protein. This missense variant has been linked to the rare autosomal recessive inborn error of metabolism ACY1 deficiency that often manifests with neurologic symptoms in humans (MIM 609924) [Sommer et al., 2011], was strongly associated with circulating levels of ACY1 protein in JHS, and provided an MR instrument to support putative causal roles for ACY1 protein on plasma levels of N-acetyl-glutamate, N-acetyl-glutamine, N-acetyl-serine, and N-acetyl-alanine in JHS, each of which was experimentally validated in our murine ACY1 knockout studies. This variant is too rare in European populations (ALFA European MAF=0.0003) to have been captured for study in the MESA (MAF=0.006) or HERITAGE Family study (MAF=0.004), and thus would not have been available for analysis if not for inclusion of the JHS. This suggests that as an increasing number of multi-omics studies in diverse populations become available, our ability to study biologically significant protein-metabolite relationships will also improve.

It is also notable that over half (51%, 35 of 68) of the tested protein-to-metabolite MR associations experimentally validated in our three murine knockout models with at least nominal levels of significance ($p \leq 0.05$). This suggests that a significant fraction of the 224 total protein-to-metabolite MR

associations that we have identified point to biological relationships that can be further elucidated in model systems. Further, 62 additional protein-metabolite associations that were identified in the pair-wise Pearson correlation analyses but either did not have an available MR instrumental variable or were not captured by MR analyses validated in the murine models. This indicates that the protein-metabolite association data may highlight a substantial number of additional, biologically significant relationships, and that MR and Pearson correlation data can identify both overlapping and distinct causal associations.

Finally, we have identified several specific examples of protein-metabolite associations that suggest novel regulatory mechanisms affecting plasma metabolites. These included experimental validation of 27 (54%) of the predicted MR associations between the scavenger receptor CD36 protein and levels of plasma lipid metabolites. CD36 has been well-documented to regulate fatty acid uptake in a number of cell types in rodent models, including hematopoietic stem cells [Mistry et al., 2021], leukemic stem cells [Ye et al., 2016], cardiac myocytes [Coort et al., 2004], adipocytes [Daquinag et al., 2021], endothelial cells [Daquinag et al., 2021], and macrophages [Podrez et al., 2002]. CD36 is further known to regulate plasma levels of non-esterified fatty acids in murine models [Guy et al., 2007]. Our experimentally-validated protein-metabolite association data extend these findings and suggest that CD36 may play a central role in regulating plasma levels of a range of glycerophospholipid, sphingolipid, and fatty acyl metabolites. These include the polyunsaturated fatty acids arachidonic acid and docosahexaenoic acid (DHA), two key metabolites that participate in a wide array of eicosanoid-mediated biological pathways important in inflammation, nociception, the immune response, cell growth, atherosclerosis, and blood pressure regulation. DHA has furthermore been used clinically to reduce the risk of coronary heart disease, hypertension, and hypertriglyceridemia. While CD36 has been implicated in the regulation of polyunsaturated fatty acid cellular uptake using cultured cell-based *in vitro* models [Franekova et al., 2015, Glatz and Luiken, 2015], these findings provide a rationale for further studies of the role of CD36 in regulating key lipid metabolites in human plasma.

Our study had several limitations. Although the metabolomics and proteomics platforms applied in these studies provide broad coverage, they are targeted platforms that include sentinel proteins

and metabolites designed to survey a wide array of biological processes and do not provide a complete catalog of every circulating protein and metabolite species in human plasma. The proteomics platform is further agnostic to post-translational changes in proteins. We expect that our ability to refine these initial association data for pathway discovery will improve as platform coverage improves. Similarly, although our study used "harmonized" metabolomics and proteomics data across three human cohort studies, the sample size is relatively modest compared with many GWAS. We expect that insight from protein-metabolite association studies will improve as increasing numbers of multi-omics studies become available, especially in populations including diverse ethnicities. Finally, it should be noted that we were only able to perform MR analyses on the 42% of proteins with available cis instrumental variables. We limited our analyses to include only cis instruments to limit the risk for horizontal pleiotropy and to validate the specificity of the affinity-based aptamer, but note that many additional causal protein-to-metabolite relationships likely exist within our data.

In summary, we demonstrate that the integration of proteomic and metabolomic profiling data can be used to identify novel protein determinants of circulating metabolite levels in human plasma. We provide proof-of-concept that these insights can be tested successfully in model systems. Finally, we are making all protein-metabolite association data from three large human cohorts available as a public resource for the further study of human metabolism.

4.5 Methods for Protein-Metabolite Association Studies

Data Availability

Individual-level metabolomic, proteomic, and genomic data from JHS, MESA, and the HERITAGE Family study are available through application to the respective cohorts. All protein-metabolite association data, including pairwise Pearson correlation, enrichment, and Mendelian randomization results, are included in the article.

Study Approval

The JHS human study protocol was approved by the Jackson State University, Tougaloo College, and the University of Mississippi Medical Center Institutional Review Boards, and all participants

provided written informed consent. The MESA human protocol was approved by The Lundquist Institute (formerly Los Angeles BioMedical Research Institute) at Harbor-University of California, Los Angeles Medical Center, University of Washington, Wake Forest School of Medicine, Northwestern University, University of Minnesota, Columbia University, Johns Hopkins University, and University of California, Los Angeles Institutional Review Boards, and all participants provided written informed consent. The HERITAGE Family study human study protocol was approved by the Institutional Review Boards at the Beth Israel Deaconess Medical Center, University of Washington, and the four clinical centers of the HERITAGE Family study, and all participants provided written informed consent. All animal experiments were approved by the Institutional Animal Care and Use Committee at Beth Israel Deaconess Medical Center.

Human Cohort Study Participants

The JHS, MESA, and the HERITAGE Family studies have been previously described [Bouchard et al., 1995, Bild et al., 2002, Isezuo, 2005]. Briefly, JHS is a community-based longitudinal cohort study that started in 2000 and included 5306 self-identified Black individuals from the Jackson, Mississippi metropolitan area. Samples from 1985 individuals with available baseline metabolomics and proteomics data from Visit 1 were included in the present study. MESA is a population-based study that started in 2000 and included 6814 self-identified White, Black, Hispanic, and Asian individuals recruited from six clinical centers across the United States. Samples from 983 individuals with available baseline metabolomics and proteomics data from Visit 1 were included in the present study. The HERITAGE Family study is an exercise training study that started in 1994 and included 763 self-identified White and Black individuals in family units recruited from four clinical centers across the United States and Canada. Samples from 658 individuals with available baseline metabolomics and proteomics data from Visit 1 were included in the present study.

Proteomic Profiling

Aptamer-based proteomic profiling methods using the SOMAscan platform have been described previously [Ngo et al., 2016, Raffield et al., 2020, Robbins et al., 2021]. Briefly, in each study, proteomics was performed on baseline plasma samples that were collected during Visit 1 in EDTA tubes and subsequently stored at -70 degrees C. The SOMAscan 1.3k platform was used in JHS and

MESA studies, and the SOMAscan 5k platform was used in the HERITAGE Family study.

Metabolomics Profiling

Metabolomics profiling was performed using liquid chromatography mass spectrometry (LC-MS) on fasting baseline plasma samples that were collected during Visit 1 in JHS, MESA, and the HERITAGE Family study, as previously described [Robbins et al., 2019, Tahir et al., 2021]. Briefly, amino acids, amines, acylcarnitines, lipids, and other water-soluble, polar metabolites were measured using a Nexera X2 U-HPLC (Shimadzu) equipped with a 150 x 2 mm, 3 μm Atlantis hydrophilic interaction LC column (Waters) coupled to a Q Exactive hybrid quadrupole Orbitrap MS (ThermoFisher Scientific). Metabolites were extracted from 10 μl plasma by adding 90 μl of Acetonitrile:Methanol:Formic acid (74.9:24.9:0.2,v/v/v) solution spiked with valine-d8 (Sigma) and Phenylalanine-d8 (Cambridge Isotope Laboratories). The metabolites were eluted at 0.25 ml/min with 5% buffer A (10 mM Ammonium-Formate, 0.1% formic acid in water) for 0.5 minutes followed by a linear gradient to 40% buffer B (0.1% formic acid in acetonitrile) over 10 minutes. MS analyses were carried out using electrospray ionization in the positive mode and full scan spectra were acquired over 70-800 m/z. Raw data were processed using Trace Finder (v3.3, Thermo Fisher Scientific) and Progenesis QI (Waters). Sugars, purines, pyrimidines, organic acids and other intermediary metabolites were measured using a 1290 Infinity LC system (Agilent Technologies) equipped with a 100 x 2.1 mm XBridge amide column (Waters) coupled to a 6490 Triple Quad MS (Agilent Technologies) in negative ionization mode via multiple reaction monitoring (MRM) scanning. Data were quantified using MassHunter Quantitative Analysis software (V10.1, Agilent).

To ensure quality control, a mixture of 150 reference standards was analyzed before, during periodic intervals throughout, and after each MS run to ensure reproducibility of LC retention times, LC peak shapes, and MS sensitivity. Isotope labeled internal standards were monitored in each sample throughout the duration of each run. Pooled plasma samples were monitored after every 10 participant samples to standardize for MS drift over time using "nearest neighbor" normalization and between batches. Separate pooled plasma samples were monitored after every 20 participant samples to determine coefficient of variation (CV) for each metabolite. Metabolite identities were confirmed using authentic reference standards. All metabolite peaks were manually reviewed for peak quality in a blinded manner. None of the included metabolites had poor peak quality or CVs

$\geq$ 30% averaged across batches.

Animal Studies

Plasma was collected for LC-MS studies by cardiac puncture from the following fasting 5-week old, male mice maintained on a normal chow diet: B6.129P2-Apoetm1Unc/J (RRID:IMSR_JAX:002052, obtained from Jackson Labs), B6.129S1-Cd36tm1Mfe/J (RRID:IMSR_JAX:019006, obtained from Jackson Labs), C57BL/6N-Acy1em1(IMPC)J/Mmucd (RRID:MMRRC_046467-UCD, obtained from The Knockout Mouse Project (KOMP)), and C57BL/6J (RRID:IMSR_JAX:000664, obtained from Jackson Labs). Homozygous APOE and CD36 knockout animals were compared to wild-type C57BL/6J controls. Homozygous ACY1 knockout animals were compared to wild-type littermates. LC-MS was conducted using the same methods as described above.

Statistical Analyses

Correlation Analyses

Metabolite and protein levels were log-transformed, scaled to a mean of zero and standard deviation of 1, and residualized on age, sex, batch (for metabolites), plate (for proteins). Additional models included further adjustment for body mass index (BMI) and estimated glomerular filtration rate (eGFR; not available in HERITAGE Family study), where indicated. Pearson correlation coefficients were calculated for each pairwise protein-metabolite combination within each study and meta-analyzed across studies using the metacor function within the General Package for Meta-Analysis in R [Schwarzer et al., 2015]. Reported p-values were estimated using the Fisher transformation. FDR-adjusted q-values were computed for each protein using the Bioconductor qvalue package in R [Gentleman et al., 2004]. Correlations with q-values $\leq$ 5% were reported as significant. A correlation heat map was generated using the Heatmap3 R package [Zhao et al., 2014], in which the organization of metabolites was fixed by RefMet superclass, main class, and subclass [Fahy and Subramaniam, 2020], and proteins were allowed to self-organize using the default complete linkage method of the hierarchical clustering function.

Enrichment Analyses

Enrichment analyses were performed to identify proteins with pairwise metabolite correlations that were enriched for a specific metabolite class using a method analogous to Gene Set Enrichment Analysis (GSEA) [Mootha et al., 2003, Subramanian et al., 2005]. Sets of metabolites were generated according to RefMet superclasses, with fatty acyls, glycerolipids, glycerophospholipids, prenol lipids, sphingolipids, and sterol lipids combined into a single, combined lipid set. Members of the lipid set were evaluated in the lipid enrichment analysis and not included in other metabolite set enrichment analyses due to the large size of this set. Meta-analyzed metabolite correlation results for each protein were ranked by p-value, annotated by metabolite RefMet class set, and a running sum statistic was calculated to generate an enrichment score (ES). The running sum statistic increased when it encountered a member of the analyzed RefMet metabolite class set and decreased when it encountered a nonmember of the analyzed set. Each increase was weighted by the strength of the metabolite correlation with the protein (according to p-value) normalized by the sum of the correlations over all the metabolites (p=1 in equation 1) [Subramanian et al., 2005]. The significance of each ES was assessed by comparison to the null distribution of calculated ES for each metabolite class generated by running 100,000 simulations of the analysis. FDR-adjusted q-values were calculated for each protein to metabolite class ES, with q-values $\leq 5\%$ reported as significant.

Genotyping

Whole-genome sequencing (WGS) in JHS and MESA has been described [Raffield et al., 2017]. Participant samples underwent >30× WGS through the Trans-Omics for Precision Medicine project at the Northwest Genome Center at University of Washington and joint genotype calling with participants in Freeze 6. Genotype calling was performed by the Informatics Resource Center at the University of Michigan. Genotyping in HERITAGE was performed on the Illumina Infinium Global Screening Array, and genotypes were called using Illumina's GenCall based on the TOP/BOT strand method. Genotype imputation to the TOPMed Freeze5 reference panel was performed using the University of Michigan Imputation Server Minimac4. Phasing was performed with Eagle v2.4. Sites with call rate <90%, mismatched alleles, or invalid alleles were excluded.

Genome-Wide Association Studies

Metabolite and protein levels in the JHS, MESA, and HERITAGE Family Study were log-transformed,

scaled to a mean of zero and standard deviation of 1, and residualized on age, sex, batch (for metabolites), plate (for proteins), and principal components of ancestry 1-10 as determined by the Genetic Estimation and Inference in Structured samples (GENESIS) [Raffield et al., 2017]. These values were inverse normalized and tested for association with genetic variants using linear mixed effects models adjusted for age, sex, the genetic relationship matrix, and principal components of ancestry 1-10 using the fastGWA model implemented in the GCTA software package.

Mendelian Randomization Analyses

Genetic instrumental variables (IV) for MR were selected from variants that were located in cis to the coding gene for each protein ($\leq$ 1 million bases upstream or downstream of the transcriptional start site for the protein cognate gene, or to the transcription end site for genes $>$ 1 million bases) that had a study-specific observed minor allele frequency (MAF) $\geq$ 0.01 and were associated with circulating levels of the measured protein with a p $\leq$ 0.05 that was Bonferroni-adjusted for the total number of variants within this cis window. Candidate instruments that met these criteria were then pruned using a study-specific linkage disequilibrium (LD) threshold of 0.001 with PLINK 1.996-9810–12.

One-sample MR using individual level data was performed since genetic variants, plasma protein levels, and plasma metabolite levels were available in the same individuals, and because samples between studies represented different ethnic groups with different patterns of linkage disequilibrium, minor allele frequencies, and population characteristics. Pairwise significant associations from meta-analyzed Pearson correlation and enrichment analyses were considered candidate protein-to-metabolite causal associations for MR. Causal effect estimates were obtained using the inverse-variance weighted (IVW) method using the MendelianRandomization R package [Yavorska and Burgess, 2017] and then meta-analyzed across all three studies using a fixed effect model [Schwarzer et al., 2015]. The limited information maximum likelihood (LIML) robust method was performed to assess the sensitivity for findings, and further supplemented with MR-Egger, median, and median-weighted methods when instruments contained more than two genetic variants. MR was performed on each protein that had $\geq$ 1 available IV and associated metabolites with q-value $\leq$ 0.05 (either by pairwise correlation analysis or enrichment analysis). Significance levels were adjusted for multiple hypothesis testing by computing FDR-adjusted q-values for each protein using the Bioconductor q-value package in R

[Gentleman et al., 2004]. Correlations, with q-values $\leq 5\%$ reported as significant.

4.6 Summary

Specific Aim 2 of this dissertation hypothesized that instrumental variable analysis of multi-omics data derived from banked human plasma samples would identify novel causal inter-biomarker associations. This hypothesis was affirmed by the work presented in the preceding chapter. Multi-omics data (i.e., genomics, proteomics, and metabolomics) were integrated from 3,626 individuals across three epidemiology studies. Causal protein-to-metabolite associations were identified using Mendelian randomization instrumental variable analysis, including associations between the scavenger receptor CD36 and circulating levels of dozens of lipid species such as plasmalogens. Top associations were validated by performing metabolomics on serum samples from three strains of murine knockout models. More than 50% of tested associations were validated in this *in vivo* experiment. This work demonstrated the potential for bioinformatics methods to triage biochemical relationships for mechanistic study.

Chapter 5

Genomics as a Bio-to-Health Informatics Translational Bridge

Specific Aim 3 of this dissertation is to leverage genomics as a translational bridge to propose candidate causal biomarkers of incident cardiovascular disease (Figure 5.1). This aim hypothesizes that genomics is a translational bridge from bioinformatics knowledge to clinical phenotypes for the identification of candidate incident disease causal biomarkers. The prior two chapters detailed the application of informatics methods on (1) clinical data aggregated from health information exchange and on (2) multi-omics data derived from samples collected as part of longitudinal epidemiological studies. In Chapter 3, information retrieval, health information exchange, and a common data model were leveraged to organize health data across healthcare organizations. The standardized data was then used to understand clinical medical practice behaviors and hypothesize sources of healthcare disparities. Chapter 4 employed a multi-omics approach to genomic, proteomic, and metabolomic data derived from population-based serum samples to first hypothesize and then validate, using *in vivo* mouse knockout models, causal relationships between proteins and metabolites in human plasma. This chapter explores a method for translating the bioinformatics findings of Chapter 4 into a clinically relevant context through the identification of biomarkers in the causal pathway toward clinical phenotypes associated with atherosclerotic cardiovascular disease.

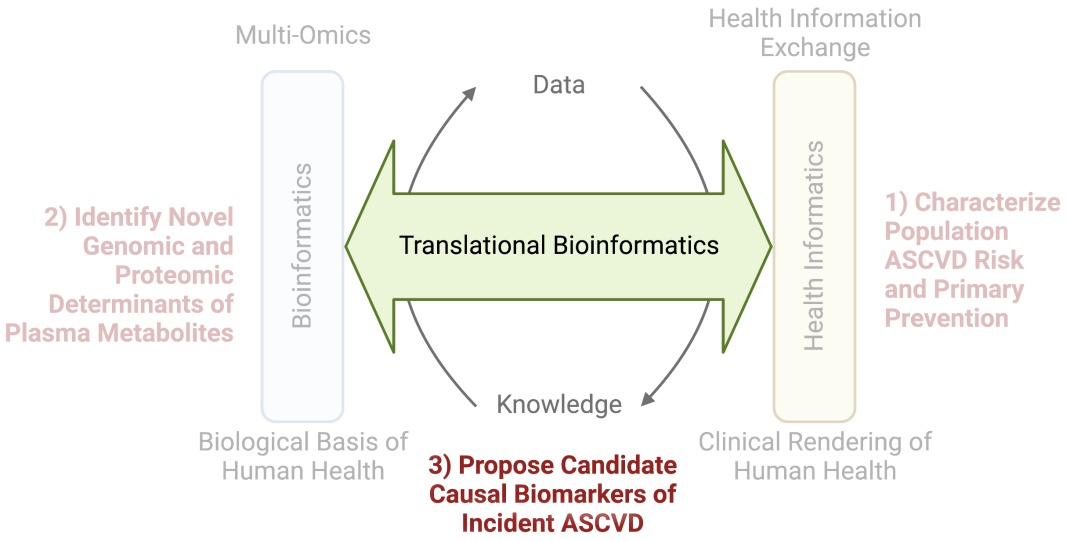

Figure 5.1: Overview of Specific Aim 3

5.1 Genomics as a Translational Bridge

Two-sample MR

Data available for health research are fragmented across multiple dimensions. The trade-off between these dimensions is analogous to the classic bias-variance tradeoff well known to statisticians – that it is often not possible to improve accuracy without introducing bias. This manifests in biological and health datasets as a tension between data breadth and depth. Broad population-level data such as those from the Rhode Island Health Information Exchange comprises hundreds of millions of data points across hundreds of thousands of individuals. These data represent individual interactions with the healthcare system and collectively render a clinical image of human health. The measurements available on such a large number of people are limited to well-studied physiology (e.g., blood pressure, complete blood count) that led to the collection of the data for clinical practice in the first place. On the other hand, individuals followed as part of detailed epidemiological studies like the Jackson Heart Study have had in-depth phenotyping performed where data such as those collected as part of routine clinical care have been carefully adjudicated. Furthermore, a subset of the participants of these studies has been phenotyped further using state-of-the-art purpose-built technologies for performing

multi-omics measurements on collected biospecimens. These include whole genome sequencing, as well as aptamer-based proteomics and liquid-chromatography mass spectroscopy metabolomics both using multiple orthogonal platforms to enhance both sensitivity and specificity. While the size of multi-omics datasets is expected to grow dramatically over the next few years, thus far it has been impractical for this level of phenotyping to be performed on more than a few thousand individuals from a single cohort.

Of all of the multi-omics data currently available, the most practical to obtain on a large number of individuals is genomics where the cost of performing whole genome sequencing has come down from $25M per individual in 2006, to less than $2,000 today, thanks to advances in next-generation sequencing. This means that it is increasingly reasonable to expect that genomics data will be available on both large population datasets and small deeply phenotyped ones. Whole genome sequencing has the potential to act as a translational bridge between these two types of studies.

Genomics has the advantage of being assigned before birth and static throughout an individual's lifetime.[1] The temporal nature of genomics enables the assertion of genomics as a causal factor in models of physiology. In other words, genetics is a valid source of instrumental variables for causal analysis in naturally occurring experiments.

In Chapter 4, single nucleotide polymorphisms located in and around the coding region for proteins were used as instrumental variables for Mendelian randomization to identify putative causal associations between proteins and metabolites in human plasma samples. In this study, both the exposure (e.g., protein) and outcome (e.g., metabolite) were measured in the same individuals. One-sample Mendelian randomization is the method that described causal effect estimation using measurement of both exposure and outcome in a single population.

The correlation analysis that was used in Chapter 4 to triage protein-metabolite pairs for downstream Mendelian randomization required both the protein and metabolite to be measured from the same sample. Unlike correlation, Mendelian randomization is performed using the summary statics of regression of the measurement of interest as a function of individual alleles. It is therefore possible

[1]The statement of genomics being static throughout an individual's lifetime is a simplifying assumption that does not take into account epigenetics or micro-RNAs. However, at the level of data acquired by whole genome sequencing and with notable exceptions (e.g. cancers, gene therapy, organ transplant) the genome is set at conception and remains constant for an individual's lifetime.

to use summary statistics from two different studies to estimate the causal effect of an exposure on an outcome where the exposure and outcome are measured in different populations. This is called two-sample Mendelian randomization (Figure 5.2). This method has been used to demonstrate, among other things, that observed associations between HDL-cholesterol and ischemic cardiovascular disease [Haase et al., 2012], HDL-cholesterol and type-2 diabetes [Haase et al., 2015], and LDL-cholesterol and cancer [Benn et al., 2011] are not causal.

All of Us Research Program

The All of Us Research Program is a longitudinal health study in the United States that is part of the National Institute of Health's Precision Medicine Initiative [Sankar and Parker, 2017, Galea and Abdalla, 2018, All of Us Research Program Investigators et al., 2019, Ramirez et al., 2022]. The program aims to enroll one million people living in the United States to contribute data by a combination of answering survey questions, providing biospecimens (e.g., blood and urine), as well as electronic health record and mobile health data. Participants are recruited or able to self-enroll. The intake process involves one to two hours of health surveys that can be taken on a smartphone, personal computer, or at the intake appointment. The intake appointment involves taking biometric measurements including weight, blood pressure, and height, as well as the collection of biospecimens. It is the goal of the program to enroll participants that represent the diversity of the United States population with particular attention to populations historically underrepresented in medical research. To contribute electronic health record data, participants must grant access to their home health institutions using an authentication protocol. Once a link has been made, All of Us downloads historical health data for the participant. Electronic health data associated with the recruiting institution will remain updated for as long as consent remains active. Health data from outside that institution or health data for self-enrolled individuals must be manually linked and periodically granted authentication requests to stay up to date.

Unique to the All of Us Research Program is the implementation of a purpose-built informatics workbench platform. Researcher access is currently limited to those who have affiliations with approved institutions using a data passport model. Data access is broken down into three tiers. Summary counts data for the entire dataset are publically available. This includes the number of participants with particular diagnosis codes or measurements performed. It also includes genetic

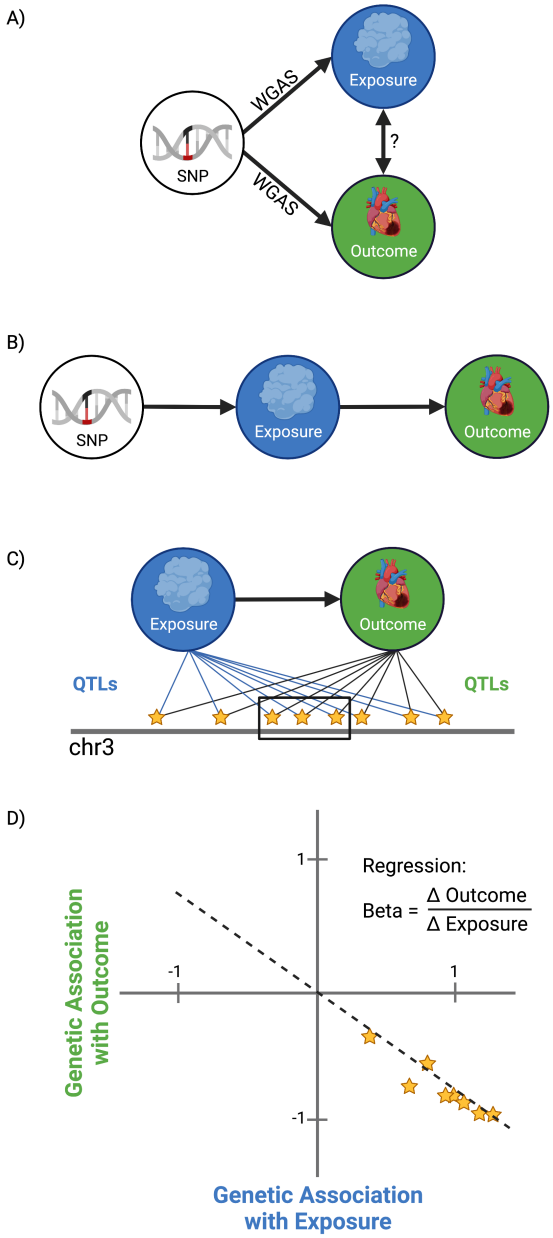

Figure 5.2: Mendelian randomization leverages the genetic associations of both an exposure and outcome variable in order to estimate the causal effect that the exposure has on the outcome. (A) Summary statistics from whole genome association or genome-wide association studies are the input estimates for Mendelian randomization. (B) Satisfaction of the Mendelian randomization assumptions described in Chapter 2 allows for the construction of a causal model that places the exposure between the genetic association and the outcome variable. (C) Shared quantitative trait loci between exposure and outcome result in a significant causal association between exposure and outcome. (D) The causal association between exposure and outcome is estimated by regressing the genetic association of the exposure as a function of the genetic association with the outcome.

ancestry estimates as well as frequencies of individual single nucleotide polymorphisms and how they are distributed by ancestry. Researchers are granted access to one of two additional tiers of data beyond what is publicly available depending on the level of data the institution is approved for and the training the researcher has completed. This includes the "registered tier" and "controlled tier." Registered tier data includes individual-level electronic health records, survey responses, physical measurements and wearable device data. The controlled tier adds genomic and additional demographic data.

According to the public data browser,[2] the dataset includes 372,380 participants as of this writing. Electronic health records for these individuals contain more than 24K discrete codes for conditions, nearly 30K for drug exposures, more than 15K for labs and measurements, and more than 29K for procedures. Whole genome sequencing with an average read depth of 30X has been performed on more than 98K participants. Hundreds of survey fields have been filled out including more than 57K individuals' responses to 80 questions regarding social factors of health and 142K participants' responses to questions about their personal medical history.

All of these data are stored in a central database that is accessible to researchers via the All of Us Workbench. The Workbench is a Google Cloud-based programming platform. Users program in either Python or R Jupyter notebooks. Depending on researcher requirements, analysis can be performed on small inexpensive virtual machines costing pennies per hour to use up to the most powerful 96-core supercomputers with nearly limitless RAM and numbers of workers for Dataproc tasks. The implementation of a centrally managed dataset and workspace breaks new ground for the democratization of researcher access to data and tools. This framework promises to promote open science in a field that has historically struggled with reproducibility in part due to sensitivity surrounding data access controls. Easy remote installation of software packages is a strength of the Workbench platform, a task that can prove challenging in secure environments. In addition, the data passport model makes collaboration both within and between institutions seamless.

[2] accessed on 1/23/2023, as of 6/6/2022

5.2 Identifying Biomarker Candidates in the Causal Pathway with Clinical Phenotypes

Recalling from Chapter 2, the proteome and metabolome were introduced as existing on the trans-omics spectrum from genome to phenome (Figure 2.4). Proteins, and in particular metabolites, are excellent candidates for biomarkers of disease as they may be proximal causal biochemical actors of pathology. The sections that follow will leverage the protein-to-metabolite causal associations that were validated using murine knockout models in Chapter 4 to identify protein-to-phenotype and metabolite-to-phenotype relationships. The particular protein-to-metabolite relationships identified are an excellent case study for this analysis as they include both well-characterized and novel physiology. Apolipoprotein E is known to be involved in the transport of lipids and apolipoprotein variants have been implicated for their role in the development of cardiovascular [Mahley and Rall, 2000, Phillips, 2014] and Alzheimer's diseases [Strittmatter et al., 1993, Corder et al., 1993].

In the prior chapter, several members of the glycerophospholipid subclass, plasmalogens, were identified by Mendelian randomization and experimentally confirmed to be modulated by both plasma levels of apolipoprotein E and CD36. Plasmalogens are lipid species characterized by a *cis* vinyl ether bond that links an alkyl chain to the sn-1 position of a glycerol backbone (Figure 5.3). They are a primary structural component of cell membranes and constitute eight to twenty percent of the total phospholipid mass in humans [Lohner, 1996, Nagan and Zoeller, 2001]. Plasmalogens are synthesized from alkyl phospholipids and alkyl phosphatidylethanolamines (PE) via desaturase enzyme [Brosche and Platt, 1998] and prior studies have postulated that they are atheroprotective via their anti-oxidant properties [Rasmiena et al., 2015, Paul et al., 2019]. They are believed to be a reservoir of polyunsaturated fatty acids including arachidonic and docosahexaenoic acids [Nagan and Zoeller, 2001]. For all of these reasons apolipoprotein E, CD36, and the plasmalogens they modulate are interesting targets for interrogation of their relationships with atherosclerotic cardiovascular risk factors and outcome phenotypes.

A total of eight distinct single nucleotide polymorphisms were identified across three population studies that were in or near the coding region of the proteins aminoacylase-1, CD36, and apolipoprotein E that were significantly associated with circulating levels of the corresponding protein (Table 5.1).

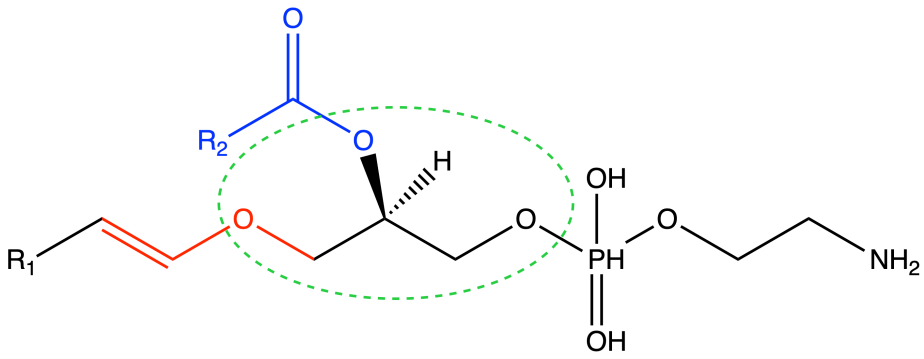

Figure 5.3: Phosphatidylethanolamine plasmalogen structure. R_1 - vinyl ether-linked fatty alcohol. R_2 is aldehyde linked polyunsaturated fatty acid. (e.g., arachadonic acid)

The SNPs were linkage-disequilibrium pruned within each study as discussed in the prior chapter. A single significant exonic SNP was found to be associated with aminoacylase-1 using data from the Jackson Heart Study. Four SNPs were associated with the circulating level of CD36 in human plasma. One intergenic SNP was found in both JHS and MESA, while two additional unique SNPs (one exonic and the other intergenic) were found in JHS and one in HERITAGE. Finally, three distinct SNPs (one from each study) were found to be significantly related to the circulating level of apolipoprotein E.

Protein	Study	SNP	A1	A2	N	AF1	BETA	P	Genecode Cat.	Genecode Info (dist)
Acy-1	JHS	chr3:51989004:C:T	T	C	1852	0.018	-1.11	3.22E-19	exonic	ABHD14A-ACY1; ACY1
CD36	JHS	chr7:79658618:G:A	A	G	1852	0.016	-0.72	3.50E-08	ncRNA_intronic	AC073048.1
CD36	HERITAGE	chr7:80671133:T:G	G	T	708	0.026	-1.56	2.20E-19	exonic	CD36
CD36	JHS	chr7:80690622:G:C	C	G	1852	0.093	-1.29	5.40E-109	intergenic	CD36(11348); SEMA3C(51916)
CD36	MESA	chr7:80690622:G:C	C	G	980	0.023	-1.49	6.37E-27	intergenic	CD36(11348); SEMA3C(51916)
CD36	JHS	chr7:80753528:G:A	A	G	1852	0.014	-1.16	9.48E-16	intronic	SEMA3C
Apo-E	JHS	chr19:44908822:C:T	T	C	1852	0.111	-0.78	2.23E-49	exonic	APOE
Apo-E	MESA	chr19:44909976:G:T	T	G	980	0.082	-0.95	8.52E-37	ncRNA_exonic	AC011481.3
Apo-E	HERITAGE	chr19:44927023:C:G	C	G	708	0.477	-0.33	4.27E-08	intergenic	APOC1(7674); APOC4(15214)

Table 5.1: Annotated protein instrumental variables. Instrumental variables from the three proteins that have validated protein-to-metabolite causal associations from Chapter 4. The three proteins have eight instrumental variables across three studies with JHS and MESA sharing one SNP. The SNPs are all within 1M base pairs of the transcription start site of the protein's cognate gene and include exonic, intronic, or intergenic SNPs.

Unlike for proteins, metabolites do not have an analgous "cognate gene" region to focus the identification of SNPs for instrumental variable analysis. The closest approximation of this for metabolites are known enzyme-metabolite pairs, where the cognate gene for the enzyme would plausably be an excellent candidate for an instrumental variable. However, a complete annotation of gene-enzyme-metabolite pairs does not exist. Therefore, the search for SNP-metabolite associations was performed genome-wide. The universe of metabolites was limited to those validated in the prior chapter to be modulated by aminoacylase-1, CD36, or apolipoprotein E.

A genome-wide search for SNP-metabolite associations elicited a total of 33 unique SNPs across the JHS, MESA, and HERITAGE cohorts (Table 5.2). The majority of them were identified using the JHS cohort, likely due to it having more than twice the number of particants of the other two studies combined. While this issue was present in the identification of protein instrumental variables, it was exacerbated with metabolites due to the need for both genome-wide significance and the increased functional distance between snp and metabolite compared to snp and protein on the trans-omics spectrum. Of the 33 SNPs, two are exonic. Thirteen of the SNPs are intronic and fifteen are intergenic. Two of them are non-coding and one is from an untranslated region. The mechanisms by which these SNPs modulate levels of these metabolites in human plasma are likely multifactorial. Some of these SNPs are near proteins that were demonstrated in the previous chapter to be causally associated with the metabolite indicating that protein-metabolite interactions could play a role. In addition, SNPs near genes for enzymes not measured on the Somalogic proteomics platform but have known roles in regulating metabolite levels were identified. For example, SNPs near the genes for the fatty acid desaturase family of enzymes [FADS1-Gene-NCBI, 2023, FADS2-Gene-NCBI, 2023], including Fads1 and Fads2 were associated with levels of arachadonic acid and plasmalogens [Paul et al., 2019]. Plasmalogens are known to be a reservoir for polyunsaturated fatty acids including arachidonic acid and desaturases act on plasmalogens. Additional SNP-metabolite associations include SNPs near genes for proteins that regulate transcription [TADA2A-Gene-NCBI, 2023], promote growth and angiogenesis [HGF-Gene-NCBI, 2023], and regulate post-translational modification [TADA2A-Gene-NCBI, 2023]. Further investigation is required to understand the mechanisms by which these SNPs may be modulating circulating metabolite levels.

Metabolite	Study	SNP	A1	A2	N	AF1	BETA	P	Genecode Cat.	Genecode Info (dist)
Arachidonic acid	HERITAGE	chr5:138262667:T:G	G	T	659	0.02	-1.18	3.67E-08	intronic	GFRA3
Arachidonic acid	JHS	chr11:61806068:C:T	T	C	1563	0.02	-0.69	3.19E-08	UTR5	FADS1
Arachidonic acid	HERITAGE	chr11:61820833:A:G	G	A	659	0.28	-0.37	7.53E-09	intronic	FADS1; FADS2
Arachidonic acid	MESA	chr11:61828850:CA:C	C	CA	993	0.36	-0.39	1.54E-20	intronic	FADS1; FADS2
C34:0 PE	JHS	chr7:81469016:G:C	C	G	2466	0.076	-0.31	1.55E-08	intergenic	AC005008.2(269901); HGF(229994)
C36:5 PC plasmalogen	MESA	chr11:61825533:A:G	G	A	995	0.392	-0.24	3.53E-08	intronic	FADS1; FADS2
C36:5 PC plasmalogen	JHS	chr14:67498735:G:A	A	G	2466	0.17	-0.23	2.62E-09	intronic	TMEM229B
C36:5 PE plasmalogen	JHS	chr7:80671133:T:G	G	T	2466	0.092	0.29	9.05E-09	exonic	CD36
C38:6 PC plasmalogen	JHS	chr10:113241444:C:A	A	C	2466	0.037	-0.42	4.51E-08	intergenic	TCF7L2(73766); RNU7-165P(111981)
C38:6 PC plasmalogen	JHS	chr11:61820833:A:G	G	A	2466	0.114	-0.28	8.58E-10	intronic	FADS1; FADS2
C38:6 PC plasmalogen	JHS	chr14:67498735:G:A	A	G	2466	0.17	-0.28	8.93E-13	intronic	TMEM229B
C38:6 PC plasmalogen	JHS	chr17:37475328:C:T	T	C	2465	0.023	0.54	3.45E-08	intronic	TADA2A
C38:6 PE plasmalogen	JHS	chr7:80671133:T:G	G	T	2466	0.092	0.39	1.60E-14	exonic	CD36
C38:7 PE plasmalogen	JHS	chr3:108141165:G:A	A	G	2466	0.369	-0.17	6.55E-09	intergenic	LINC01215(2555); IFT57(19647)
C38:7 PE plasmalogen	JHS	chr7:80671133:T:G	G	T	2466	0.092	0.4	1.20E-15	exonic	CD36
C40:7 PE plasmalogen	JHS	chr7:80690622:G:C	C	G	2466	0.094	0.35	2.30E-12	intergenic	CD36(11348); SEMA3C(51916)
C46:2 TAG	JHS	chr19:9135907:A:G	G	A	2464	0.031	-0.46	4.24E-08	intergenic	OR7G3(8957); ZNF317(4490)
CAR 26:0	MESA	chr10:103284400:C:T	T	C	995	0.444	0.25	3.75E-08	intronic	INA
CAR 26:0	JHS	chr11:68633911:A:G	G	A	2390	0.231	0.28	4.64E-15	intergenic	AP003096.1(17200); GAL(50633)
Cer(d18:1/16:0)	JHS	chr2:26830034:G:T	T	G	2465	0.052	0.36	4.65E-08	intergenic	CENPA(35445); DPYSL5(17713)
DHA	HERITAGE	chr1:14819275:G:A	A	G	659	0.011	1.63	2.32E-08	intronic	KAZN
N-Acetylalanine	JHS	chr3:51959148:C:A	A	C	2465	0.017	1.23	1.73E-28	intronic	PCBP4
N-Acetylglutamic acid	JHS	chr2:73577581:A:G	G	A	2466	0.548	0.17	6.06E-09	intronic	ALMS1
N-Acetylglutamic acid	JHS	chr3:51959148:C:A	A	C	2466	0.017	1.4	3.04E-36	intronic	PCBP4
N-Acetylglutamic acid	JHS	chr5:14093354:C:T	T	C	2466	0.072	0.32	2.84E-08	intergenic	DNAH5(148811); TRIO(49988)
N-Acetylglutamine	JHS	chr2:73631838:T:C	C	T	2101	0.528	0.84	3.87E-161	intergenic	ALMS1(6672); NAT8(8994)
N-Acetylglutamine	JHS	chr2:73675421:A:G	G	A	2101	0.011	0.87	1.43E-08	intergenic	NAT8(33028); NAT8B(25155)
N-Acetylglutamine	JHS	chr2:73704764:G:A	A	G	2101	0.405	-0.23	2.16E-13	intergenic	NAT8B(3424); TPRKB(25066)
N-Acetylglutamine	JHS	chr2:77336858:C:G	G	C	2101	0.09	-0.3	2.89E-08	intronic	LRRTM4
N-Acetylglutamine	JHS	chr3:51932445:C:T	T	C	2101	0.018	0.97	1.08E-16	ncRNA_exonic	AC115284.4
N-Acetylserine	JHS	chr3:51959148:C:A	A	C	2466	0.017	1.06	1.17E-21	intronic	PCBP4
N-Acetylserine	JHS	chr9:80667356:T:C	C	T	2466	0.062	-0.34	2.05E-08	intergenic	LINC01507(632801); AL138749.1(202402)
SM(d18:1/16:0)	JHS	chr1:55063542:C:A	A	C	2466	0.011	-0.98	1.28E-12	exonic	PCSK9
SM(d18:1/16:0)	JHS	chr17:75617001:T:C	C	T	2466	0.056	0.35	4.07E-08	intronic	MYO15B
SM(d18:1/18:0)	JHS	chr1:55063542:C:A	A	C	2466	0.011	-0.83	2.18E-09	exonic	PCSK9
SM(d18:1/18:1)	JHS	chr8:20912046:G:A	A	G	2466	0.18	-0.22	1.15E-08	intergenic	AC018541.1(211894); AC015468.3(29382)
SM(d18:1/18:1)	HERITAGE	chr12:56309938:G:A	A	G	658	0.036	-0.83	3.97E-08	ncRNA_intronic	AC073896.2
SM(d18:1/18:1)	JHS	chr13:104672706:G:A	A	G	2466	0.826	-0.21	2.96E-08	intergenic	AL136524.1(538449); AL138954.1(738110)
SM(d18:1/20:0)	JHS	chr1:55063542:C:A	A	C	2466	0.011	-0.76	4.31E-08	exonic	PCSK9
SM(d18:1/20:0)	JHS	chr14:63766999:A:G	G	A	2466	0.399	0.18	1.06E-09	intergenic	RNU7-116P(30494); RF00019(81262)
Stearic acid	JHS	chr10:100315722:G:A	A	G	1563	0.668	0.33	3.61E-18	intronic	PKD2L1

Table 5.2: Annotated metabolite instrumental variables. Genome-wide quantitative trait loci for metabolites validated to be causally modulated by apolipoprotein E, aminoacylase-1, or CD36 were identified as instrumental variables for metabolite-to-phenotype Mendelian randomization. SNP-metabolite association statistics are reported by study and SNP locations are annotated for the nearest gene(s) and SNP category.

The set of SNP-protein and SNP-metabolite associations discussed represent the set of instruments to be queried using All of Us for SNP-phenotype associations. The SNP-phenotype associations from All of Us will then be integrated with the SNP-protein and SNP-metabolite associations from the protein-metabolite study to attempt to identify protein-to-phenotype and metabolite-to-phenotype causal relationships. These interrogations represent a natural experiment. That is, individuals randomly assigned these SNPs at birth have lived their lives with genetically determined differences in associated protein and metabolite concentrations circulating in their plasma that differ from the average. If those differences are associated with atherosclerotic cardiovascular disease risk factors or outcome phenotypes, they may represent a causal mechanism for atherosclerosis.

Electronic Health Record SNP-Phenotype Associations

Biomedical phenotypes are the observable physical or physiological characteristics of an individual that are relevant to health and disease. These include measurable features such as height, weight, blood pressure, and laboratory values. Expert assessment of aggregate clinical traits, signs, and symptoms, that are summarized as diagnoses, are an important class of biomedical phenotypes that dictate future downstream prognosis and treatment. The phenome is the collection of all possible observable phenotypes. In practice, it is limited to phenotypes that can be expected to be observed with some minimum frequency and captured, for example, by the health system as part of electronic medical records. This study of SNP-phenotype associations will focus on both risk factors and outcomes related to atherosclerotic cardiovascular disease (Table 5.3) that inform the risk-based cholesterol guidelines discussed in Chapter 3. In particular, the risk factors for which genetic associations will be investigated are hypertension, hyperlipidemia, and diabetes. The outcomes of interest include myocardial infarction and stroke.

Phenotype	Phecode(s)	ICD-9	ICD-10
Myocardial infarction	411.2	410; 410.01; 410.02; 410.1; 410.11; 410.12; 410.2; 410.21; 410.22; 410.3; 410.31; 410.32; 410.4; 410.41; 410.42; 410.5; 410.51; 410.52; 410.6; 410.61; 410.62; 410.7; 410.71; 410.72; 410.8; 410.81; 410.82; 410.9; 410.91; 410.92; 411; 412; 429.7; 429.71; 429.79	I21; I21.0; I21.1; I21.2; I21.3; I21.4; I21.9; I22; I22.0; I22.1; I22.8; I22.9; I23; I23.0; I23.1; I23.2; I23.3; I23.6; I23.8; I24.1; I25.2; I51.0; I51.3
Ischemic Stroke	433.1; 433.11; 433.2; 433.21; 433.3; 433.31	433; 433.01; 433.1; 433.11; 433.2; 433.21; 433.3; 433.31; 433.8; 433.81; 433.9; 433.91; 434; 434.01; 434.1; 434.11; 434.9; 434.91; 435; 435.1; 435.2; 435.3; 435.8; 435.9; 437.1; 437.6; V12.54	G45.0; G45.1; G45.2; G45.8; G45.9; G46.0; G46.1; G46.2; I63; I63.0; I63.1; I63.2; I63.2; I63.3; I63.4; I63.5; I63.5; I63.6; I63.8; I63.9; I64; I65; I65.0; I65.1; I65.2; I65.3; I65.8; I65.9; I66; I66.0; I66.1; I66.2; I66.3; I66.4; I66.8; I66.9; I67; I67.6; I67.8
Hyperlipidemia	272.1	272.2; 272.4	E78.2; E78.4; E78.5
Hypertension	401.1; 401.2; 401.21; 401.22; 401.3	401; 401.1; 401.9; 402; 402.01; 402.1; 402.11; 402.9; 402.91; 403; 403.01; 403.1; 403.11; 403.9; 403.91; 404; 404.01; 404.02; 404.03; 404.1; 404.11; 404.12; 404.13; 404.9; 404.91; 404.92; 404.93; 405; 405.01; 405.09; 405.1; 405.11; 405.19; 405.9; 405.91; 405.99; 437.2	I11; I11.0; I11.9; I12; I12.0; I12.9; I13; I13.0; I13.1; I13.2; I13.9; I15; I15.0; I15.1; I15.2; I15.8; I15.9; I67.4
Diabetes	250.1; 250.11; 250.12; 250.13; 250.14; 250.15; 250.2; 250.21; 250.22; 250.23; 250.24; 250.25	250; 250.01; 250.02; 250.03; 250.1; 250.11; 250.12; 250.13; 250.2; 250.21; 250.22; 250.23; 250.3; 250.31; 250.32; 250.33; 250.4; 250.41; 250.42; 250.43; 250.5; 250.51; 250.52; 250.53; 250.6; 250.61; 250.62; 250.63; 250.7; 250.71; 250.72; 250.73; 250.8; 250.81; 250.82; 250.83; 250.9; 250.91; 250.92; 250.93	E10; E10.0; E10.1; E10.2; E10.3; E10.3; E10.4; E10.4; E10.6; E10.7; E10.8; E10.9; E11; E11.0; E11.1; E11.2; E11.3; E11.4; E11.6; E11.7; E11.8; E11.9; E12.3; E13; E13.1; E13.4; E13.5; E13.6; E13.7; E13.8; E13.9; E14.9; G59.0

Table 5.3: Atherosclerotic cardiovascular disease risk factor and outcome phenotype codes. Phenotype codes and their associated ICD-9 and ICD-10 diagnosis codes are reported for the outcomes of myocardial infarction and ischemic stroke as well as the risk factors hyperlipidemia, hypertension, and diabetes. For ischemic stroke, hypertension, and diabetes, multiple phenotype codes were aggregated to define a single phenotype.

A total of 39 distinct SNPs were queried for their association with the five electronic health record phenotypes using All of Us. An investigation of the minor allele frequencies of the SNPs by African and European ancestry (the two most common ancestries in All of Us) reveals that they differ significantly by ancestry (Table 5.4). In particular, there are ten out of eleven SNPs that are low-frequency (MAF $1 - 5\%$) in the African ancestry population that are rare ($< 1\%$) in the European population. There are an additional three SNPs that are common ($> 5\%$) in the African ancestry population that are rare in the European population. This is unsurprising as the derivation population for these protein and metabolite quantitative trait loci were either entirely (JHS) or approximately 50% self-identified as Black or African American. The increased genetic diversity present in a population enriched for African ancestry significantly increases the potential quantitative trait loci available for instrumental variable analysis. In addition, even if SNPs are present at reasonable frequencies in populations of European ancestry, the mechanisms by which these loci result in altered circulating concentrations of proteins and metabolites are poorly understood and it is possible that compensatory mechanisms exist in subpopulations that would cause them to be invalid instruments for Mendelian randomization in those populations. For this reason, the protein-phenotype and metabolite-phenotype Mendelian randomization presented has been limited to the population of individuals in All of Us who self-identify as Black or African American for this study. This does not imply that the findings presented are likely to be only relevant in those that are studied but rather that the natural experiment that has occurred due to the random assignment of different levels of circulating proteins and metabolites is most readily detectable in populations with the identified allele frequencies. There is no reason to believe that protein-phenotype and metabolite-phenotype relationships are not conserved across ancestral populations even if they are modulated by different SNPs.

SNP Quality Control in All of Us

A standard quality control process used for WGAS was applied to the 39 SNPs under consideration in this study. Multi-allelic loci were separated into multiple SNPs with the reference allele being related to each of the minor alleles. All considered SNPs were present in the vast majority of individuals being studied ($> 99\%$). A minor allele threshold was set to 0.5% and Hardy-Weinberg equilibrium ($p < 10E - 10$) was required. A relatedness matrix supplied by All of Us was used to exclude all but one of related individuals. Principle component adjustment was used to account for

SNP	MAF (African Ancestry)	MAF (European Ancestry)
chr1:14819275:G:A	0.001	0.006
chr1:55063542:C:A	0.008	0.000
chr2:26830034:G:T	0.047	0.000
chr2:73577581:A:G	0.569	0.229
chr2:73631838:T:C	0.546	0.220
chr2:73675421:A:G	0.012	0.000
chr2:73704764:G:A	0.417	0.637
chr2:77336858:C:G	0.091	0.498
chr3:108141165:G:A	0.385	0.506
chr3:51932445:C:T	0.019	0.000
chr3:51959148:C:A	0.019	0.000
chr3:51989004:C:T	0.019	0.000
chr5:138262667:T:G	0.004	0.027
chr5:14093354:C:T	0.074	0.071
chr7:79658618:G:A	0.018	0.000
chr7:80671133:T:G	0.090	0.000
chr7:80690622:G:C	0.092	0.000
chr7:80753528:G:A	0.017	0.000
chr7:81469016:G:C	0.072	0.009
chr8:20912046:G:A	0.177	0.217
chr9:80667356:T:C	0.061	0.090
chr10:100315722:G:A	0.677	0.192
chr10:103284400:C:T	0.447	0.447
chr10:113241444:C:A	0.039	0.000
chr11:61806068:C:T	0.021	0.075
chr11:61820833:A:G	0.115	0.336
chr11:61825533:A:G	0.263	0.339
chr11:61828850:CA:C	0.148	0.333
chr11:68633911:A:G	0.217	0.109
chr12:56309938:G:A	0.009	0.059
chr13:104672706:G:A	0.170	0.055
chr14:63766999:A:G	0.401	0.169
chr14:67498735:G:A	0.161	0.486
chr17:37475328:C:T	0.025	0.000
chr17:75617001:T:C	0.056	0.050
chr19:44908822:C:T	0.105	0.080
chr19:44909976:G:T	0.115	0.081
chr19:44927023:C:G	0.449	0.454
chr19:9135907:A:G	0.033	0.000

Table 5.4: Minor allele frequencies in the All of Us Research Program by ancestry. Minor allele frequencies are reported for 39 candidate instrumental variable SNPs by the two most common ancestries in the All of Us Research Program. Notably, thirteen SNPs that are either common ($> 5\%$) or low-frequency ($1 - 5\%$) in the African ancestry population are rare ($< 1\%$) in the European ancestry population.

population stratification.

The current[3] release of All of Us includes a total of 70,000 individuals that have at least one electronic health record diagnosis code and whole genome sequencing data. Of these, 13,000 self-identified as Black or African American on their intake survey. Relatedness analysis resulted in the removal of 400 individuals, leaving 12,600 unrelated individuals in the sample for analysis. Quality control of the 39 SNPs, removed two SNPs due to low minor allele frequency. The remaining 37 SNPs met the threshold for adhering to Hardy-Weinberg equilibrium. SNP-phenotype associations are presented in Table 5.5. Setting a false-discovery rate threshold to 5%, yield no significant SNP-phenotype relationships. However, there are seven nominally significant ($p < 0.05$) relationships. Four of them are with myocardial infarction, and one each with ischemic stroke, hyperlipidemia, and hypertension. Despite these underwhelming findings, the addition of these summary data to Table 5.1 and Table 5.2 are sufficient for protein-to-phenotype and metabolite-to-phenotype causal analysis respectively. The integration of directionally consistent SNP-phenotype relationships across multiple SNPs associated with a single protein or metabolite across multiple studies may reveal more convincing relationships with a combination of two-sample Mendelian randomization and meta-analysis.

Protein-to-Phenotype Causal Associations

Two-sample Mendelian randomization was performed using protein quantitative trait loci identified in three distinct population studies (JHS, MESA, and HERITAGE). The loci from each study were considered separately for association with phenotypes in All of Us. Where multiple SNPs were available for a protein from a single study, the inverse-variance weighted (IVW) method was used for the estimation of a causal association. All of Us phenotypes as outcome variables for causal analysis are defined in Table 5.3.

Meta-analyzed causal protein-to-phenotype estimates are presented in Tables 5.6, 5.7, and 5.8. Briefly, elevated circulating levels of apolipoprotein E were found to be protective against myocardial infarction ($\beta = -0.13$, $p = 0.017$), hyperlipidemia ($\beta = -0.07$,$p = 0.029$), and ischemic stroke ($\beta = -0.17$,$p = 0.044$). This is directionally and mechanistically consistent with known functions of Apo-E including that a well-studied mouse model for atherosclerosis is APOE -/- [Piedrahita et al., 1992, Plump et al., 1992, Hansson et al., 2002, Meir and Leitersdorf, 2004]. Elevated circulating

[3] 6/6/2022 as of 3/9/2023

SNP	Myocard. Infarct.		Ischemic Stroke		Hyperlipidemia		Hypertension		Diabetes	
	Beta	P	Beta	P	Beta	P	Beta	P	Beta	P
chr1:55063542:C:A	-0.28	0.47	-0.21	0.548	-0.19	0.235	-0.16	0.282	-0.15	0.434
chr2:26830034:G:T	-0.03	0.846	-0.15	0.293	-0.06	0.387	-0.2	0.0014*	-0.06	0.44
chr2:73577581:A:G	-0.01	0.872	-0.04	0.471	-0.02	0.413	-0.04	0.11	-0.03	0.341
chr2:73631838:T:C	0.02	0.786	-0.02	0.706	-0.01	0.667	-0.03	0.243	-0.01	0.677
chr2:73675421:A:G	-0.07	0.802	0.14	0.566	-0.02	0.899	0.1	0.409	0.09	0.538
chr2:73704764:G:A	0.04	0.544	0	0.95	0.03	0.265	0.03	0.339	0.01	0.721
chr2:77336858:C:G	-0.03	0.794	-0.06	0.537	-0.04	0.384	-0.04	0.345	-0.01	0.837
chr3:108141165:G:A	0.06	0.318	-0.02	0.673	-0.01	0.625	0.01	0.739	-0.02	0.597
chr3:51932445:C:T	-0.13	0.612	0.21	0.286	0.01	0.961	-0.1	0.32	-0.14	0.299
chr3:51959148:C:A	-0.12	0.62	0.25	0.195	0.01	0.93	-0.09	0.394	-0.21	0.129
chr3:51989004:C:T	-0.14	0.571	0.23	0.23	-0.01	0.948	-0.1	0.336	-0.17	0.194
chr5:138262667:T:G	0.14	0.735	0.12	0.766	-0.3	0.155	-0.03	0.888	-0.58	0.0502
chr5:14093354:C:T	0.16	0.139	0.09	0.396	-0.03	0.588	0.01	0.863	0.03	0.62
chr7:79658618:G:A	-0.26	0.329	-0.38	0.132	0.15	0.139	0.1	0.318	-0.07	0.579
chr7:80671133:T:G	-0.26	0.0313*	0.11	0.23	0.07	0.168	-0.05	0.28	-0.02	0.793
chr7:80690622:G:C	-0.32	0.0078*	0.12	0.209	0.05	0.29	-0.05	0.238	-0.03	0.658
chr7:80753528:G:A	0.1	0.664	0.14	0.509	-0.08	0.471	-0.02	0.829	-0.25	0.0662
chr7:81469016:G:C	0.18	0.107	-0.01	0.925	-0.04	0.414	-0.03	0.527	-0.12	0.0687
chr8:20912046:G:A	-0.04	0.613	0.03	0.677	0.04	0.239	0.03	0.42	-0.07	0.123
chr9:80667356:T:C	-0.01	0.967	0.04	0.728	-0.03	0.55	0.08	0.122	0.06	0.381
chr10:100315722:G:A	0.13	0.0496*	0.01	0.933	0.01	0.657	0.03	0.312	0.06	0.108
chr10:103284400:C:T	0.02	0.742	-0.02	0.741	-0.03	0.212	-0.04	0.169	-0.01	0.722
chr11:61806068:C:T	-0.15	0.515	-0.2	0.359	-0.02	0.864	-0.14	0.157	0	0.975
chr11:61820833:A:G	-0.13	0.191	-0.1	0.279	-0.02	0.622	-0.01	0.738	-0.01	0.789
chr11:61825533:A:G	-0.07	0.344	-0.03	0.654	0.02	0.477	-0.02	0.453	0.06	0.138
chr11:61828850:CA:C	-0.1	0.258	-0.01	0.888	0.06	0.106	-0.01	0.826	0.01	0.836
chr11:68633911:A:G	-0.01	0.924	0.06	0.402	0.01	0.657	-0.05	0.0912	-0.01	0.776
chr12:56309938:G:A	-0.21	0.57	0.28	0.303	-0.22	0.156	0.01	0.926	0.12	0.484
chr13:104672706:G:A	0.02	0.832	-0.03	0.662	-0.05	0.198	-0.05	0.19	-0.03	0.476
chr14:63766999:A:G	0.04	0.48	-0.03	0.548	-0.01	0.8	0.01	0.771	-0.01	0.672
chr14:67498735:G:A	-0.08	0.363	0.16	0.0265*	-0.03	0.495	-0.05	0.148	0.03	0.565
chr17:37475328:C:T	-0.01	0.94	0.12	0.492	-0.03	0.759	-0.13	0.121	-0.01	0.909
chr17:75617001:T:C	-0.07	0.605	-0.05	0.674	-0.12	0.0477*	-0.06	0.273	-0.12	0.116
chr19:44908822:C:T	0.12	0.235	0.13	0.156	0.08	0.0609	0.05	0.29	0.03	0.572
chr19:44909976:G:T	0.11	0.218	0.13	0.132	0.07	0.085	0.03	0.512	0.06	0.218
chr19:44927023:C:G	0.14	0.0233*	0.02	0.782	0	0.981	-0.01	0.778	0.02	0.577
chr19:9135907:A:G	-0.28	0.143	-0.06	0.702	-0.07	0.344	-0.03	0.65	-0.06	0.544

Table 5.5: All of Us SNP to atherosclerotic cardiovascular disease outcome and risk factor phenotypes. Thirty-seven SNPs passed quality control steps and are included in this table. No SNP-phenotype relationships met a 5% false-discovery rate threshold. P-values indicated with a "*" are nominally significant ($p < 0.05$).

Protein	Phenotype	Beta	P
Apo-E	Myocardial Infarction	-0.17	0.017
Apo-E	Hyperlipidemia	-0.07	0.029
Apo-E	Ischemic Stroke	-0.13	0.044
Apo-E	Diabetes	0.02	0.292
Apo-E	Hypertension	-0.03	0.371

Table 5.6: Mendelian randomization of Apolipoprotein E to ASCVD risk factor and outcome phenotypes

Protein	Phenotype	Beta	P
CD36	Myocardial Infarction	0.19	3.46e-5
CD36	Ischemic Stroke	-0.08	0.0397
CD36	Hyperlipidemia	-0.04	0.0438
CD36	Hypertension	0.03	0.0666
CD36	Diabetes	0.02	0.292

Table 5.7: Mendelian randomization of CD36 to ASCVD risk factor and outcome phenotypes

levels of CD36 in human plasma were found to cause myocardial infarction ($\beta = 0.19$, $p = 3.46e-5$) but were protective against ischemic stroke ($\beta = -0.08$, $p = 0.0397$) and hyperlipidemia ($\beta = -0.04$, $p = 0.0438$). Finally, aminoacylase-1 was not significantly associated with any of the investigated phenotypes but interestingly its top association was with diabetes which is at least provocative given its known interactions with diabetes-associated acetylated branched-chain amino acids [Ngo et al., 2021]. While the significance levels of these protein-to-phenotype associations would not stand up to phenome-wide association interrogation, they are at least hypothesis-generating as a targeted study.

Metabolite-to-Phenotype Causal Associations

With myocardial infarction as the top protein-to-phenotype causal association for both apolipoprotein E and CD36, metabolite-to-myocardial infarction relationships were queried to identify potential

Protein	Phenotype	Beta	P
Acy-1	Diabetes	0.16	0.194
Acy-1	Ischemic Stroke	-0.21	0.23
Acy-1	Hypertension	0.09	0.336
Acy-1	Myocardial Infarction	0.13	0.571
Acy-1	Hyperlipidemia	0.01	0.948

Table 5.8: Mendelian randomization of Aminoacylase-1 to ASCVD risk factor and outcome phenotypes

mechanisms for these relationships. All metabolites with genome-wide significant instruments from Table 5.2 were tested for a causal association with myocardial infarction by study and meta-analyzed in Table 5.9 using the same method described for protein-to-phenotype associations. Four out of six queried plasmalogen species including all phosphatidylethanolamine were nominally significant for their causal association with myocardial infarction ($p < 0.05$). All four of these plasmalogens were experimentally confirmed to be modulated by CD36 by performing metabolomics on CD36 -/- mice in Chapter 4. Two of the four were confirmed to be causally associated with apolipoprotein E.

Metabolite	Phenotype	BETA	P	IV Chromosome(s)	Assoc. Protein(s)
C40:7 PE plasmalogen	Myocardial Infarction	-0.92	0.0078	7	Apo-E; CD36
C38:7 PE plasmalogen	Myocardial Infarction	-0.52	0.0217	3; 7	CD36
C38:6 PE plasmalogen	Myocardial Infarction	-0.67	0.0313	7	Apo-E; CD36
C36:5 PE plasmalogen	Myocardial Infarction	-0.89	0.0313	7	CD36
Stearic acid	Myocardial Infarction	0.40	0.0496	10	CD36
C34:0 PE	Myocardial Infarction	-0.59	0.107	7	CD36
Arachidonic acid	Myocardial Infarction	0.22	0.127	5; 11	CD36
C46:2 TAG	Myocardial Infarction	0.61	0.143	19	CD36
C36:5 PC plasmalogen	Myocardial Infarction	0.31	0.191	11; 14	Apo-E; CD36
C38:6 PC plasmalogen	Myocardial Infarction	0.25	0.209	10; 11; 14; 17	Apo-E; CD36
SM(d18:1/20:0)	Myocardial Infarction	0.28	0.322	1; 14	Apo-E; CD36
SM(d18:1/18:0)	Myocardial Infarction	0.34	0.470	1	CD36
N-Acetylalanine	Myocardial Infarction	-0.09	0.486	3	Acy-1
SM(d18:1/18:1)	Myocardial Infarction	0.11	0.633	8; 12; 13	CD36
N-Acetylserine	Myocardial Infarction	-0.08	0.689	3; 9	Acy-1
Cer(d18:1/16:0)	Myocardial Infarction	-0.08	0.846	2	CD36
CAR 26:0	Myocardial Infarction	0.03	0.858	10; 11	CD36
SM(d18:1/16:0)	Myocardial Infarction	0.04	0.879	1; 17	CD36
N-Acetylglutamic acid	Myocardial Infarction	0.02	0.885	2; 3; 5	Acy-1
N-Acetylglutamine	Myocardial Infarction	0.00	0.971	2; 3	Acy-1

Table 5.9: Mendelian randomization of experimentally confirmed metabolites from protein-to-metabolite study and myocardial infarction.

In order to contextualize these findings, metabolite-to-phenotype Mendelian randomization was performed for all plasmalogens against the five phenotypes (Table 5.10). The four significant plasmalogen-to-myocardial infarction relationships remain the top hits across all phenotypes. While not significant, the two phosphatidylcholine plasmalogen-to-myocardial infarction associations were stronger than the majority of other metabolite-to-phenotype relationships. Taking the plasmalogen-to-phenotype associations in total, it is clear that they are enriched for myocardial infarction. Like the protein-to-phenotype associations, while these relationships would not be considered significant phenome-wide, the findings are an indication of potential metabolite-to-phenotype associations that may warrant further interrogation. Together with the protein-to-phenotype findings, these two sample Mendelian randomizations represent a successful proof of concept integration of data from deeply phenotypes epidemiology studies with population-representative data using genomics as a

translational bridge.

Metabolite	Phenotype	BETA	P	IV Chromosome(s)	Assoc. Protein(s)
C40:7 PE plasmalogen	Myocardial Infarction	-0.92	0.0078	7	Apo-E; CD36
C38:7 PE plasmalogen	Myocardial Infarction	-0.52	0.0217	3; 7	CD36
C38:6 PE plasmalogen	Myocardial Infarction	-0.67	0.0313	7	Apo-E; CD36
C36:5 PE plasmalogen	Myocardial Infarction	-0.89	0.0313	7	CD36
C36:5 PC plasmalogen	Diabetes	-0.19	0.131	11; 14	Apo-E; CD36
C36:5 PC plasmalogen	Hypertension	0.15	0.134	11; 14	Apo-E; CD36
C38:7 PE plasmalogen	Hyperlipidemia	0.13	0.161	3; 7	CD36
C38:6 PE plasmalogen	Hyperlipidemia	0.17	0.168	7	Apo-E; CD36
C36:5 PE plasmalogen	Hyperlipidemia	0.23	0.168	7	CD36
C36:5 PC plasmalogen	Myocardial Infarction	0.31	0.191	11; 14	Apo-E; CD36
C40:7 PE plasmalogen	Ischemic Stroke	0.34	0.209	7	Apo-E; CD36
C38:6 PC plasmalogen	Myocardial Infarction	0.25	0.209	10; 11; 14; 17	Apo-E; CD36
C38:7 PE plasmalogen	Ischemic Stroke	0.23	0.222	3; 7	CD36
C38:6 PE plasmalogen	Ischemic Stroke	0.30	0.230	7	Apo-E; CD36
C36:5 PE plasmalogen	Ischemic Stroke	0.40	0.230	7	CD36
C40:7 PE plasmalogen	Hypertension	-0.16	0.238	7	Apo-E; CD36
C36:5 PC plasmalogen	Ischemic Stroke	-0.23	0.262	11; 14	Apo-E; CD36
C38:6 PE plasmalogen	Hypertension	-0.13	0.280	7	Apo-E; CD36
C36:5 PE plasmalogen	Hypertension	-0.17	0.280	7	CD36
C38:7 PE plasmalogen	Hypertension	-0.1	0.287	3; 7	CD36
C40:7 PE plasmalogen	Hyperlipidemia	0.14	0.290	7	Apo-E; CD36
C38:6 PC plasmalogen	Hyperlipidemia	0.05	0.584	10; 11; 14; 17	Apo-E; CD36
C38:6 PC plasmalogen	Ischemic Stroke	-0.09	0.612	10; 11; 14; 17	Apo-E; CD36
C40:7 PE plasmalogen	Diabetes	-0.07	0.658	7	Apo-E; CD36
C38:6 PC plasmalogen	Hypertension	0.02	0.772	10; 11; 14; 17	Apo-E; CD36
C38:6 PC plasmalogen	Diabetes	-0.03	0.778	10; 11; 14; 17	Apo-E; CD36
C38:6 PE plasmalogen	Diabetes	-0.04	0.793	7	Apo-E; CD36
C36:5 PE plasmalogen	Diabetes	-0.05	0.793	7	CD36
C36:5 PC plasmalogen	Hyperlipidemia	-0.01	0.908	11; 14	Apo-E; CD36
C38:7 PE plasmalogen	Diabetes	0.01	0.916	3; 7	CD36

Table 5.10: Mendelian randomization of plasmalogens experimentally confirmed to be modulated by CD36 or Apo-E and ASCVD phenotypes.

Constructing a Trans-Omics Causal Graph

In aggregate, the protein-to-metabolite study from Chapter 4 and the protein and metabolite-to-phenotype relationships elucidated in this chapter define directional inter-omic edges that can begin to reveal a multi-omic to phenotype causal graph. The edges include SNP to protein and SNP to metabolite quantitative trait loci. Mendelian randomization of these SNPs combined with metabolomics performed on murine mouse models allows the drawing of causal edges between proteins and metabolites. In addition, the Mendelian randomization of SNPs associated with proteins and metabolites with electronic health record-based phenotypes from the All of Us Research Program are hypothesis-generating for protein-to-phenotype and metabolite-to-phenotype relationships that may identify important biochemical mechanisms for the development of atherosclerotic cardiovascular disease. A schematic of the edges between SNPs, the proteins apolipoprotein E and CD36, plasmalogens as a family of metabolites, and myocardial infarction are diagrammed in Figure 5.4.

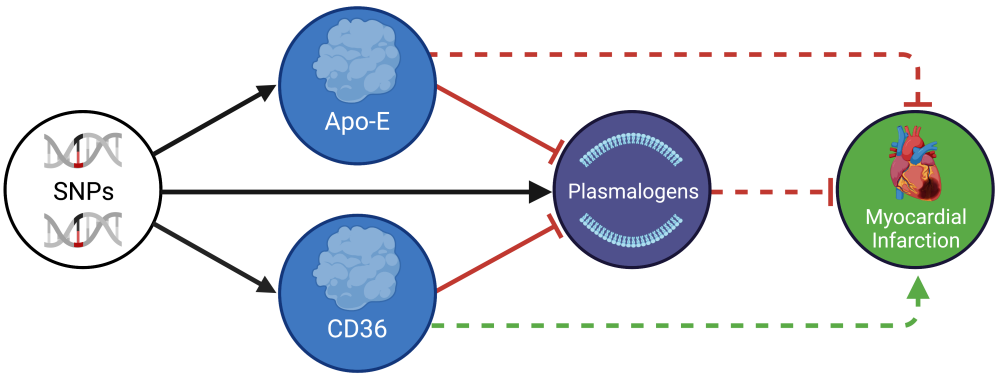

Figure 5.4: Multi-omics network. Genetics modulate circulating levels of plasma proteins (solid black arrows). Causal protein (Apo-E and CD36) to metabolite (plasmalogens) relationships validated in murine knockout models (solid red inhibitory arrows). Positive (green dashed arrow) and negative (red dashed arrows) protein and metabolite to myocardial infarction causal associations from All of Us.

In particular, both apolipoprotein E and CD36 were demonstrated to inhibit circulating levels of both phosphatidylethanolamine and phosphatidylcholine plasmalogens. Mendelian randomization identified that elevated levels of both circulating apolipoprotein E and plasmalogens are protective against myocardial infarction while elevated levels of CD36 increase the incidence of myocardial infarction. Together, these findings indicate significant potential for the integration of multi-omics datasets for the identification of causal biochemical-to-phenotype causal relationship pathways.

5.3 The Case for an HIE-Linked Genomic Data Repository

The methods and results presented have demonstrated that data derived from deeply phenotyped individuals can be combined with large datasets collected as part of clinical medical practice using genomics as a translational bridge. In doing so, this dissertation has presented a framework for the integration of big data sets from across the translational bioinformatics and multi-omics spectrums to identify candidate causal biomarkers. The focus of leveraging observational data to make causal inferences about disease incidence and mechanisms expands the types of target questions that can be considered from correlation to causation. In particular, the associations revealed in this chapter indicate that plasmalogens may play a causal role in the development of atherosclerotic cardiovascular disease as an intermediary between the carrier protein apolipoprotein E, the cell surface protein

CD36, and myocardial infarction.

The use of gene-phenotype relationships from the All of Us Research Program is a significant demonstration that phenotypes derived from electronic health records in the United States are of sufficient quality for the integration of multi-omics data, such as those presented in Chapter 4, for the development of trans-omics causal graphs. While they are some of the largest datasets of their kind available to date, the numbers represented in the presented analyses are small and the power to identify additional new trans-omics relationships will greatly increase as larger cohorts are phenotyped.

The All of Us Research Program is in a nascent stage. However, the findings in this chapter demonstrate the potential application of population-level genomics linked to electronic health records to integrate bioinformatics knowledge into a clinical context. Deployed on a wide scale, the natural repository and interface of this would be centrally managed health information exchanges. Chapter 3 demonstrated real-world fragmentation in clinical care in the context of cardiovascular risk factor collection. To avoid replicating this problem in the genomic era, the static nature of whole genome sequencing data for an individual's lifetime makes it a particularly strong candidate data type that should be centrally managed for both clinical and research applications.

5.4 Summary

Specific Aim 3 of this dissertation hypothesized that genomics could act as a translational bridge from bioinformatics knowledge to clinical phenotypes for the identification of candidate incident disease causal biomarkers. This hypothesis was affirmed by the work presented in the preceding chapter. Electronic health record phenotypes developed in Aim 1, and SNP-to-protein and SNP-to-metabolite associations identified in Aim 2, were integrated with whole genome-associated electronic health records from the All of Us Research Program. In doing so, a multi-omic directed graph was proposed that spans the trans-omics spectrum from genome to proteome to metabolome to phenome. This included the rediscovery of the inverse relationship between Apolipoprotein E and incident myocardial infarction. In addition, negative associations between circulating levels of plasmalogens and positive associations with circulating levels of CD36 were found with myocardial infarction. This work bridged the translational gap between bio- and health informatics using genomics as a

translational bridge.

Chapter 6

Conclusion

6.1 Summary

Cardiovascular diseases are the leading cause of morbidity and mortality in the United States, and atherosclerosis is the primary mechanism for cardiovascular disease development. Atherosclerosis is the gradual build-up of plaque on the intimal walls of arteries that results in the narrowing of the lumen, increased turbulence of blood flow, and risk of plaque rupture and downstream total arterial occlusion. When this happens in the heart, it is a myocardial infarction. In the brain, it is a stroke. Complete elimination of blood flow for an extended period leads to cell death due to lack of oxygen supply. Depending on the size of the affected tissues, this can result in irreversible heart failure, functional impairment, or death. The accumulation of plaque occurs over a lifetime, and it is common for older asymptomatic adults to have arteries in their heart that are more than seventy percent occluded. Symptoms from ischemia occur late in this process.

Longitudinal epidemiology studies going back 75 years have shown that exposure to risk factors over decades increases an individual's chance of a major adverse cardiovascular event. These include but are not limited to age, sex, elevated blood pressure, cholesterol, smoking, diabetes, and family history. Some risk factors, including blood pressure and cholesterol, are modifiable using pharmacotherapy and have informed guidelines for treating risk factors with drugs as primary prevention. Guidelines

for lowering cholesterol with statins in individuals without known atherosclerotic cardiovascular disease rely on the predicted risk of incident myocardial infarction or stroke. In particular, individuals aged 40-75 years with an intermediate or high risk of an ASCVD event should take daily statin therapy for primary prevention.

The biological basis for this intervention is rooted in a complex web of trans-omics interactions. Omics refers to the collective study of biological molecules of a particular class, and trans-omics interactions are inter- and intra- class networks of associations. For example, the APOE gene, through transcription, post-translational modification, and translation, instructs cellular machinery to produce the protein Apolipoprotein E. This is a carrier protein responsible for the transport of lipid species throughout the body in the bloodstream. Mice genetically modified not to produce Apolipoprotein E lack this transport function and develop very high levels of lipids in their blood, leading to atherosclerosis. Mutations in the APOE gene in humans also lead to dyslipidemia and increased rates of atherosclerosis. In this way, small perturbations on one end of the trans-omics spectrum can have significant phenotypic effects on the other. Motivated by the goal of improving the primary prevention of atherosclerosis through the detailed longitudinal investigation of population-level clinical data, this dissertation iterated on a translational bioinformatics framework (Figure 6.1) that bridges: (1) health informatics where aggregated electronic health records provide a clinical rendering of human health and healthcare practices, (2) bioinformatics by leveraging multi-omics data to elucidate the biological basis of human health, and (3) a translational mechanism whereby the knowledge produced from one end of the translational bioinformatics spectrum becomes the data for investigation in the other.

The presented framework took significant cues and motivation from the theoretical foundation laid by Sarkar et al. in 2011, where they proposed genetics would be a translational mechanism between bio- and health informatics at a time when the possibility of a $1000 genome was coming into focus. This dissertation took advantage of more than a decade of advancements in next-generation sequencing and multi-omics platform development to apply Mendelian randomization for the causal inference between proteins, metabolites, and cardiovascular electronic health record phenotypes. To this end, this dissertation demonstrates that translational bioinformatics is not simply a shared set of methods or long-term goals but a field that can integrate discovery from across the informatics

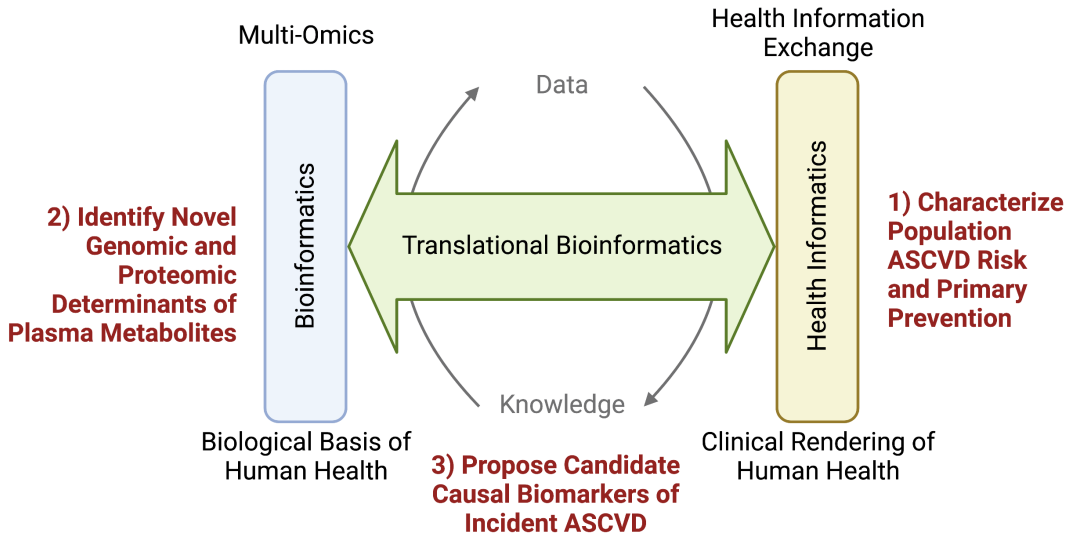

Figure 6.1: Specific Aims of the dissertation

spectrum today using concrete examples across three Specific Aims and associated hypotheses.

This dissertation set out to integrate and analyze observational biomedical datasets to motivate and facilitate the identification of candidate causal biomarkers for incident cardiovascular disease. The work presented made significant contributions toward this goal (Figure 6.2).

In Aim 1, working with the Rhode Island Quality Institute, aggregate electronic health record data assembled through information exchange from healthcare institutions and providers across Rhode Island were transformed into a common data model and then leveraged to characterize population ASCVD risk and primary prevention practices. This work demonstrated the public health monitoring potential of centrally managed health information exchange and gaps between medical knowledge and clinical care. It also explored disparities in baseline risk factors, preventative care, and long-term risk of incident cardiovascular disease in vulnerable populations. The combination of poor population statin guideline adherence and a significant proportion of high-risk individuals off of statins who never go on to develop incident ASCVD motivate a deeper understanding of the underlying biology of ASCVD and the search for more precise causal biomarkers to identify truly high-risk populations for primary prevention.

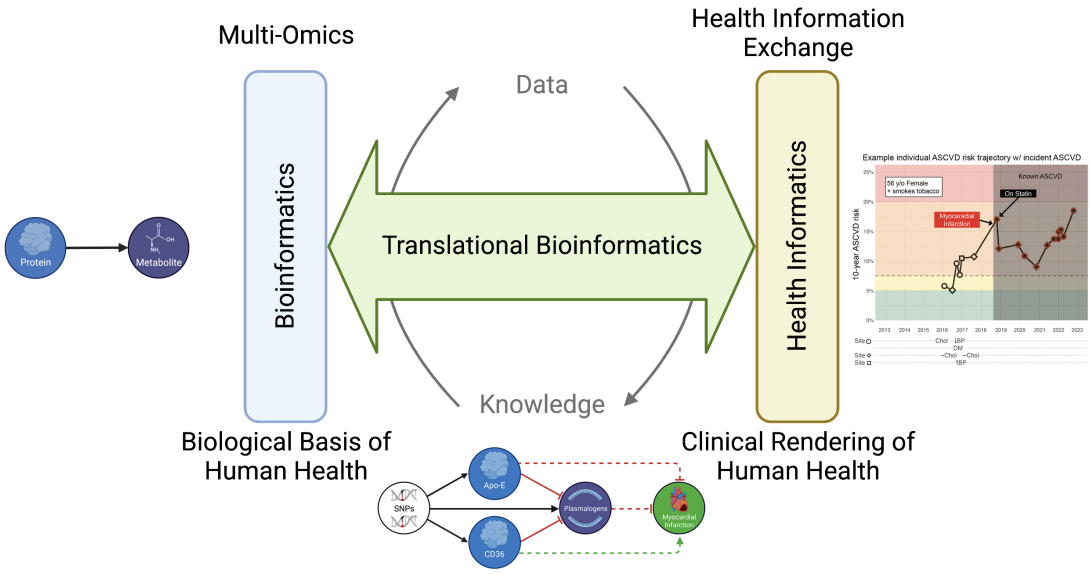

Figure 6.2: Summary of the dissertation

In Aim 2, in collaboration with the National Heart Lung and Blood Institute's Trans-Omics for Precision Medicine Initiative, multi-omics data measured on banked plasma samples from three epidemiology studies were leveraged to identify novel genomic and proteomic determinants of plasma metabolite levels that included known ASCVD biomarkers. Computational methods for inferring causal associations using single nucleotide polymorphisms as instrumental variables allowed for the causal inference of relationships between circulating molecules measured in human plasma samples. The top findings were validated by performing metabolomics on murine knockout models engineered not to produce causal proteins of interest. Aim 2 demonstrated this bioinformatics method's ability to triage pairwise molecular relationships for downstream mechanistic study in the lab.

In Aim 3, whole genome sequencing linked electronic health records from the All of Us Research Program were used to connect findings from Aims 1 and 2. Novel candidate causal biomarkers, including the scavenger protein C36 and several plasmalogen lipid species from Aim 2, were found to be related to ASCVD risk factors and outcomes by Mendelian randomization (e.g., myocardial infarction) from Aim 1.

This dissertation integrated data from large observational datasets. The datasets used in each aim

are simultaneously some of the largest of their kind and yet barely large enough to power these analyses. However, they are each poised to grow in the coming years, promising to reveal significantly more biologically and clinically important relationships. The Rhode Island health information exchange (CurrentCare) is moving from opt-in, where individuals need to enroll for their data to be collected, to opt-out, which will significantly increase the number of people that can be followed for longitudinal ASCVD risk and primary prevention practices. CurrentCare is the first HIE in the country to standardize to a common data model, which enables the kind of work shown here. As additional HIEs are standardized to a CDM, it will be possible to resolve more complete population health pictures for various clinical applications and share purpose-built informatics tools across the United States.

The multi-omics findings presented in Aim 2 were accomplished using data from a few thousand individuals. Proteomics and metabolomics measurements continue to be performed on additional samples. Analogous datasets with more than ten times the number of individuals will be available for analysis within a few years, and it is expected that similar analyses applied to them will reveal significantly more causal associations that may further elucidate the biological basis of atherosclerosis. Finally, All of Us is in its infancy. Over the next decade, additional electronic health records will be collected longitudinally, and more than ten times the number of individual participants will be available for research. In addition, there is active ongoing work for All of Us to collaborate with state and regional health information exchanges to recruit individuals receiving care in community settings. As demonstrated in Aim 1, health information exchange aggregated records will provide improved electronic health record coverage for individual subjects compared to single health systems alone.

6.2 Limitations

This dissertation has some noteworthy limitations that correspond to each of its main contributions. These are summarized below and organized by contribution area.

Natural Language Processing of Clinical Notes

The two main contributions of this dissertation in the field of information retrieval and information extraction from clinical notes are the identification of note sections using a hidden Markov model of UMLS semantic types and the extraction of clinical symptoms of angina from the history of present illness sections of clinical notes using a large language model classifier. The evaluation of the note section identification model was limited by using notes from the same institution in which the model was developed. To avoid overfitting, the notes for out-of-sample evaluation originated from different templates, but a more robust evaluation would have included notes from other institutions. There are indications in the results presented that this may have demonstrated a more impressive performance over the template model comparison. However, the presented investigation was unable to evaluate this. In addition, to improve the portability of this tool, the model would have benefited from being trained on notes from multiple institutions and across a range of specialties.

Extraction of angina symptoms from the history of present illness sections of clinical notes was performed successfully. The work demonstrated that this task is well suited to take advantage of transformer architectures and large language models to enable the training on relatively small amounts of labeled data. The results, however, identified an unexpected weakness in the original goal of these efforts. In particular, the documentation of angina symptoms in the notes studied was poor in that there was not sufficient detail regularly documented (e.g., improving with rest) to enable classification distinction between typical and atypical angina. This distinction is required to interpret the clinical significance of chest pain. The finding that this documentation was poor in the clinical scenario where documentation is required for an ordered test indicates that it is unlikely that angina symptoms would be well documented in a general corpus of clinical notes. Therefore the future utility of this application is likely limited.

Health Information Exchange and the Primary Prevention of Cardiovascular Disease

This dissertation simultaneously contributes to the literature about health information exchange and population primary prevention of cardiovascular disease practices. It presents the first known transformation of data originating from a health information exchange to a common data model. This

enabled detailed evaluation of electronic health record data collected for individual patients across time and place and facilitated longitudinal tracking of cardiovascular risk factors, 10-year ASCVD risk, and primary prevention practices with statins. Through this analysis, it became clear that there are several important areas for improvement in the health information exchange data partner network. In particular, many data partners do not reliably contribute blood pressure measurements to the exchange. This limited the ability to track cardiovascular risk at a true population level. In addition, the omission of blood pressure was primarily from small primary care practices and small group specialty practices. This limited the conclusions that could be made with regard to comparing guideline adherence in those care settings. Instead, the presentation of this work focuses on the comparison between care at Federally Qualified Health Centers (FQHCs) and Health Systems.

A second limitation of this work stemmed from limitations in data available from the prescription data aggregator SureScripts. While SureScripts provides reliable information about a prescription being filled, it does not currently provide a readily computable way to calculate the "days of supply" for that fill (i.e., the information is embedded in short text phrases that would require natural language processing). This means that it was not possible to calculate the proportion of days covered to assess the quality of individual statin coverage as a continuous variable. Instead, for this study, statin coverage was approximated by identifying statin prescription fills near updated ASCVD risk factors, and statin therapy status was evaluated as a binary variable.

Bioinformatics of Multi-Omics Data

The investigation of protein-to-metabolite associations expanded the scope of this dissertation to include the investigation of the biological basis of atherosclerosis and was intermediary work towards identifying causal biomarkers of incident ASCVD. Despite the identification and *in vivo* validation of novel protein-to-metabolite associations, the number of identified causal interactions is small and is concentrated in a few causal proteins. The reason for this is a relatively small sample size for whole-genome association studies (i.e., the number of individuals with proteomic and metabolomic measurements) and limited numbers of independent instrumental variables. It is expected that the volume of multi-omic data suitable for these kinds of analyses will grow significantly in the coming years, and the work presented in this dissertation will form the theoretical foundation for more comprehensive analyses.

Identification of Causal Biomarkers

The rediscovery of known (i.e., Apolipoprotein E), additional evidence for purported (i.e., plasmalogens), and proposal of a completely novel (i.e., CD36), biomarkers for incident cardiovascular disease are simultaneously one of the most novel and least developed portions of this dissertation. This work is essential in closing the translational barrier between bio- and health informatics using genetics, as discussed using a translational bioinformatics framework. These findings leveraged SNP-to-phenotype associations within the All of Us Research Program that would not be considered otherwise whole-genome significant. However, they stand out as potentially exciting findings because of the targeted nature of this inquiry enabled by the protein-to-metabolite bioinformatics investigations. Additional study with larger cohorts and potential validation in animal models is required to contextualize the impact of these findings better.

6.3 Future Directions

Several exciting potential future investigations stem from the projects presented in this dissertation. The aggregation of population-level cardiovascular risk factors, primary prevention practices, and ASCVD events lays the foundation for population-derived ASCVD risk prediction. This would provide a distinct advantage over the current pooled cohort equation models, including population-based calibration. Applying models directly to the population from which they are derived enables the focus on internal (rather than external) model validity. This was unimaginable ten years ago when the current risk equations were being developed. Two additional data components are actively being curated for inclusion into the HIE dataset to enable this project. The first is integrating the cause of death from the Department of Health. This is required to model all incident cardiovascular diseases rather than just non-fatal diseases as was used for the monitoring of population primary prevention practices. The second is the inclusion of social factors of health through census tract-based social vulnerability index components. The findings in Aim 1 that individuals receiving primary care at FQHCs have unfavorable cardiovascular risk factors when compared to those receiving care at health systems across the intersection of race and ethnicity indicate that it may be possible to identify predictors of incident ASCVD that are causal and ultimately reduce the reliance on Black race as a determinant of guideline-directed therapies.

The second potential project on the horizon that extends the health information exchange work is creating a population ASCVD risk and primary prevention dashboard integrated into the HIE web portal. This population clinical decision support tool would leverage aggregated electronic health records standardized across time and place with a common data model to estimate 10-year ASCVD risk and suggest primary prevention for individuals at intermediate and high risk who are not already on a statin. This dashboard could include decision support for additional guideline-directed medical therapies, including diabetes management and anti-hypertensive treatment.

The protein-metabolite project demonstrated the potential of using instrumental variable analysis to triage causal relationships between circulating molecules measured in human plasma. There are three natural follow-ups to this project, some already underway. This includes the identification of protein-to-protein, metabolite-to-protein, and metabolite-to-metabolite causal relationships. Projects that put the metabolite on the causal exposure side of instrumental variable analysis pose an additional challenge of identifying single nucleotide polymorphisms in *cis* to the metabolite. Current attempts to do this involve querying known enzyme-to-metabolite relationships. Genetic instruments in or near enzymes known to modulate a metabolite may provide sufficient theoretical backing as instrumental variables for metabolite-to-protein and metabolite-to-metabolite causal relationships.

Finally, with regard to the translational bioinformatics aim, the proposed causal gene-to-protein-to-metabolite-to-phenotype trans-omics network warrants further inquiry. This includes replicating the analysis in orthogonal datasets available from additional population studies to add to the voracity of the findings. With additional bioinformatics support, animal studies would be warranted, for example, in mice genetically engineered to overexpress CD36 to determine if they experience high rates of atherosclerotic phenotypes.

6.4 Conclusion

The presentation of this dissertation comes at an exciting time. The projects presented here have bared witness and taken advantage of an apparent inflection point across biomedical data and methods. Recent innovations in data collection, organization, and sharing have made many of these analyses possible. At the same time, advancements in methods and frameworks for understanding

health disparities, biological data, and causal relationships from the molecular to the population level have inspired this work and informed its conclusions. Towards the goals of Aim 1, a statewide health information exchange standardized to a common data model was leveraged to characterize population ASCVD risk and primary prevention practices in Rhode Island. For Aim 2, novel causal associations between circulating levels of proteins and metabolites in human plasma were identified using human plasma from banked blood samples. The bioinformatics findings were validated experimentally using serum metabolomics on knockout mouse models. Finally, these data types were integrated using genetics in Aim 3. These investigations rediscovered known and identified potential new protein and metabolite causal biomarkers for incident myocardial infarction.

Appendix A

Abbreviations

Abbreviations	Meaning
AGE	Advanced Glycosylated End Products
AMIA	American Medical Informatics Association
Apo-B	Apolipoprotein B
Apo-C	Apolipoprotein C
Apo-E	Apolipoprotein E
ARIC	Atherosclerosis Risk in Communities study
ASCVD	Atherosclerotic Cardiovascular Disease
CARDIA	Coronary Artery Risk Developing in Young Adults
CDC	Centers for Disease Control and Prevention
CHS	Cardiovascular Health Study
FHS	Framingham Heart Study
FQHC	Federally Qualified Health Center
HDL-C	High-Density Lipoprotein Cholesterol
HIE	Health Information Exchange
HMG-CoA	B-hydroxy B-Methylglutaryl Coenzyme A
IDL	Intermediate-Density Lipoprotein
JHS	Jackson Heart Study
LDL-C	Low-Density Lipoprotein Cholesterol
LDL-r	Low-Density Lipoprotein Receptor
MI	Myocardial Infarction
PCE	Pooled Cohort Equations
PCSK9	Proprotein Convertase Subtilisin/Kexin Type 9
TF	Tissue Factor
TTE	Transthoracic Echocardiography
VLDL	Very-Low Density Lipoprotein
vWF	von-Willibrand Factor
WT	Wild Type
RI	Rhode Island
CDM	Common Data Model
OMOP	Observational Medical Outcomes Partnership
PCP	Primary Care Provider
RIQI	Rhode Island Quality Institute
OHDSI	Observational Health Data Sciences and Informatics

Table A.1: Abbreviations

Appendix B

ASCVD Codes

OHDSI CONCEPT ID	CONCEPT CODE	VOCABULARY ID	CONCEPT NAME
312327	57054005	SNOMED	Acute myocardial infarction
314666	1755008	SNOMED	Old myocardial infarction
319039	233838001	SNOMED	Acute posterior myocardial infarction
379778	56267009	SNOMED	Multi-infarct dementia
381316	230690007	SNOMED	Cerebrovascular accident
434376	54329005	SNOMED	Acute myocardial infarction of anterior wall
436706	58612006	SNOMED	Acute myocardial infarction of lateral wall
438170	73795002	SNOMED	Acute myocardial infarction of inferior wall
438438	70211005	SNOMED	Acute myocardial infarction of anterolateral wall
438447	65547006	SNOMED	Acute myocardial infarction of inferolateral wall
439693	194802003	SNOMED	True posterior myocardial infarction
441579	76593002	SNOMED	Acute myocardial infarction of inferoposterior wall
443790	25772007	SNOMED	Multi-infarct dementia with delusions
443864	14070001	SNOMED	Multi-infarct dementia with depression
444091	10349009	SNOMED	Multi-infarct dementia with delirium
444406	70422006	SNOMED	Acute subendocardial infarction
761785	16000471000119100	SNOMED	Cerebrovascular accident due to occlusion of left anterior cerebral artery
761792	16002231000119100	SNOMED	Cerebrovascular accident due to thrombus of right posterior cerebral artery
761793	16002271000119100	SNOMED	Left posterior cerebral artery thrombosis with stroke
761795	16002391000119100	SNOMED	Left vertebral artery embolism with stroke
761796	16002431000119100	SNOMED	Right carotid artery embolism with stroke
761797	16002471000119100	SNOMED	Left carotid artery embolism with stroke
761798	16002511000119100	SNOMED	Basilar artery embolism with stroke
762339	329391000119100	SNOMED	Right anterior cerebral artery embolism with stroke
762340	329401000119103	SNOMED	Left anterior cerebral artery embolism with stroke
762345	329631000119108	SNOMED	Left vertebral artery thrombosis with stroke
763095	436041000124107	SNOMED	Posterior inferior cerebellar artery occlusion with infarction
764721	5571000124103	SNOMED	Cerebrovascular accident with intracranial hemorrhage
765281	16000551000119100	SNOMED	Cerebrovascular accident due to occlusion of right anterior cerebral artery
765568	431411000124103	SNOMED	Chronic cerebrovascular accident
3654465	836293000	SNOMED	Acute myocardial infarction of right ventricle
3661503	840312002	SNOMED	Acute ST segment elevation myocardial infarction due to mid left anterior descending coronary artery occlusion
3661504	840316004	SNOMED	Acute ST segment elevation myocardial infarction due to distal left anterior descending coronary artery occlusion
4006295	111297002	SNOMED	Nonparalytic stroke
4023571	116288000	SNOMED	Paralytic stroke
4045734	230691006	SNOMED	CVA - cerebrovascular accident due to cerebral artery occlusion
4045748	230715005	SNOMED	Posterior circulation stroke of uncertain pathology
4046363	230713003	SNOMED	Stroke of uncertain pathology
4086178	24654003	SNOMED	Weber-Gubler syndrome
4099974	25133001	SNOMED	Completed stroke
4108217	194856005	SNOMED	Subsequent myocardial infarction
4108218	194858006	SNOMED	Subsequent myocardial infarction of inferior wall
4108677	194857001	SNOMED	Subsequent myocardial infarction of anterior wall
4110196	195217004	SNOMED	Right sided cerebral hemisphere cerebrovascular accident
4111710	195212005	SNOMED	Brainstem stroke syndrome
4111711	195213000	SNOMED	Cerebellar stroke syndrome
4112022	195216008	SNOMED	Left sided cerebral hemisphere cerebrovascular accident
4119948	233836002	SNOMED	Acute Q wave infarction - widespread
4119949	233839009	SNOMED	Old anterior myocardial infarction
4121467	233840006	SNOMED	Old inferior myocardial infarction
4121468	233842003	SNOMED	Old posterior myocardial infarction
4124686	233843008	SNOMED	Silent myocardial infarction
4145721	307140009	SNOMED	Acute non-Q wave infarction
4153352	371041009	SNOMED	Embolic stroke
4159140	371040005	SNOMED	Thrombotic stroke
4170094	418044006	SNOMED	Myocardial infarction in recovery phase
4173632	42531007	SNOMED	Microinfarct of heart
4189462	373606000	SNOMED	Occlusive stroke
4200113	314207007	SNOMED	Non-Q wave myocardial infarction
4211509	413758000	SNOMED	Cardioembolic stroke
4215259	394710008	SNOMED	First myocardial infarction
4270024	401314000	SNOMED	Acute non-ST segment elevation myocardial infarction
4296653	401303003	SNOMED	Acute ST segment elevation myocardial infarction
4301259	78569004	SNOMED	Posterior inferior cerebellar artery syndrome
4310996	422504002	SNOMED	Ischemic stroke
4329847	22298006	SNOMED	Myocardial infarction
35611571	12238151000119100	SNOMED	Acute ST segment elevation myocardial infarction of inferoposterior wall
36684840	457551000124104	SNOMED	Acute stroke
36716999	1637178100011910 0	SNOMED	Cerebellar stroke
37109288	329671000119106	SNOMED	Traumatic subarachnoid hemorrhage with loss of consciousness
37110238	724425005	SNOMED	Cerebral ischemic stroke due to intracranial large artery atherosclerosis
37110239	724426006	SNOMED	Cerebral ischemic stroke due to extracranial large artery atherosclerosis
37110241	724429004	SNOMED	Stroke co-occurrent with migraine
37110679	724994008	SNOMED	Cerebral ischemic stroke due to stenosis of extracranial large artery
37309626	16837681000119100	SNOMED	Myocardial infarction due to demand ischemia
37312015	788882003	SNOMED	Cerebral ischemic stroke due to global hypoperfusion with watershed infarct
37395562	10602100011910 5	SNOMED	Multi-infarct dementia due to atherosclerosis
37395574	16023911000119100	SNOMED	Cerebrovascular accident due to right carotid artery occlusion
37395575	16026951000119100	SNOMED	Cerebrovascular accident due to right carotid artery stenosis
37395576	16026991000119100	SNOMED	Cerebrovascular accident due to left carotid artery stenosis
42535110	16000391000119100	SNOMED	Cerebrovascular accident due to occlusion of right posterior cerebral artery
42535111	16000431000119100	SNOMED	Cerebrovascular accident due to occlusion of right middle cerebral artery
42535112	16000511000119100	SNOMED	Cerebrovascular accident due to occlusion of left middle cerebral artery
42535113	16002031000119100	SNOMED	Cerebrovascular accident due to thrombus of right middle cerebral artery
42535114	16002111000119100	SNOMED	Cerebrovascular accident due to thrombus of left middle cerebral artery
42535148	16024151000119100	SNOMED	Cerebrovascular accident due to occlusion of left cerebellar artery
42535149	16024271000119100	SNOMED	Cerebrovascular accident due to occlusion of right cerebellar artery
42535460	292681000119101	SNOMED	Cerebrovascular accident due to right vertebral artery occlusion
42535511	329641000119104	SNOMED	Cerebrovascular accident due to thrombus of basilar artery
42535512	329651000119102	SNOMED	Cerebrovascular accident due to thrombus of right carotid artery
42539166	330791000119108	SNOMED	Cerebrovascular accident due to thrombus of left carotid artery
42539195	16000351000119100	SNOMED	Cerebrovascular accident due to occlusion of left posterior cerebral artery
42539262	16024111000119100	SNOMED	Cerebrovascular accident due to occlusion of left carotid artery
43020460	285981000119103	SNOMED	Acute ST segment elevation myocardial infarction involving left anterior descending coronary artery
43530670	140921000119102	SNOMED	Ischemic stroke without coma
43531605	9901000119100	SNOMED	Occlusion of cerebral artery with stroke
43531607	99451000119105	SNOMED	Cerebral infarction due to stenosis of carotid artery
44782769	175310001191 05	SNOMED	Acute myocardial infarction due to left coronary artery occlusion
45766075	703164000	SNOMED	Acute anterior ST segment elevation myocardial infarction
45766114	703211006	SNOMED	Subsequent ST segment elevation myocardial infarction
45766116	703213009	SNOMED	Acute ST segment elevation myocardial infarction of inferior wall
45766151	703253007	SNOMED	Acute ST segment elevation myocardial infarction of inferior wall involving right ventricle
45766241	703360004	SNOMED	Subsequent non-ST segment elevation myocardial infarction
46270160	15712961000119100	SNOMED	Acute ST segment elevation myocardial infarction of anteroseptal wall
46270161	15713041000119100	SNOMED	Acute ST segment elevation myocardial infarction of posterior wall
46270162	15713081000119100	SNOMED	Acute ST segment elevation myocardial infarction due to left coronary artery occlusion
46270163	15713121000119100	SNOMED	Acute ST segment elevation myocardial infarction due to right coronary artery occlusion
46270164	15713161000119100	SNOMED	Acute ST segment elevation myocardial infarction of septum
46274044	15712921000119100	SNOMED	Acute ST segment elevation myocardial infarction of lateral wall

Table B.1: ASCVD codes

References

[Agren et al., 2005] Agren, J. J., Kurvinen, J. P., and Kuksis, A. (2005). Isolation of very low density lipoprotein phospholipids enriched in ethanolamine phospholipids from rats injected with triton WR 1339. *Biochimica et biophysica acta*, 1734(1):34–43.

[Akinyelure et al., 2021] Akinyelure, O. P., Jaeger, B. C., Moore, T. L., Hubbard, D., Oparil, S., Howard, V. J., Howard, G., Buie, J. N., Magwood, G. S., Adams, R. J., Bonilha, L., Lackland, D. T., and Muntner, P. (2021). Racial differences in blood pressure control following stroke: The REGARDS study. *Stroke; a journal of cerebral circulation*, 52(12):3944–3952.

[Alberts et al., 1980] Alberts, A. W., Chen, J., Kuron, G., Hunt, V., Huff, J., Hoffman, C., Rothrock, J., Lopez, M., Joshua, H., Harris, E., Patchett, A., Monaghan, R., Currie, S., Stapley, E., Albers-Schonberg, G., Hensens, O., Hirshfield, J., Hoogsteen, K., Liesch, J., and Springer, J. (1980). Mevinolin: a highly potent competitive inhibitor of hydroxymethylglutaryl-coenzyme a reductase and a cholesterol-lowering agent. *Proceedings of the National Academy of Sciences of the United States of America*, 77(7):3957–3961.

[All of Us Research Program Investigators et al., 2019] All of Us Research Program Investigators, Denny, J. C., Rutter, J. L., Goldstein, D. B., Philippakis, A., Smoller, J. W., Jenkins, G., and Dishman, E. (2019). The "all of us" research program. *The New England journal of medicine*, 381(7):668–676.

[Alsentzer et al., 2019] Alsentzer, E., Murphy, J. R., Boag, W., Weng, W.-H., Jin, D., Naumann, T., and McDermott, M. B. A. (2019). Publicly available clinical BERT embeddings.

[American College of Cardiology Foundation Appropriate Use Criteria Task, Force et al., 2011] American College of Cardiology Foundation Appropriate Use Criteria Task, Force, American Society of, Echocardiography, American Heart, Association, American Society of Nuclear, Cardiology, Heart Failure Society of, America, Heart Rhythm, Society, Society for Cardiovascular, Angiography, Interventions, Society of Critical Care, Medicine, Society of Cardiovascular Computed, Tomography, Society for Cardiovascular Magnetic, Resonance, Douglas, P. S., Garcia, M. J., Haines, D. E., Lai, W. W., Manning, W. J., Patel, A. R., Picard, M. H., Polk, D. M., Ragosta, M., Ward, R. P., and Weiner, R. B. (2011). ACCF/ASE/AHA/ASNC/HFSA/HRS/SCAI/SCCM/SCCT/SCMR 2011 appropriate use criteria for echocardiography. a report of the american college of cardiology foundation appropriate use criteria task force, american society of echocardiography, american heart association, american society of nuclear cardiology, heart failure society of america, heart rhythm society, society for cardiovascular angiography and interventions, society of critical care medicine, society of

cardiovascular computed tomography, and society for cardiovascular magnetic resonance endorsed by the american college of chest physicians. *Journal of the American College of Cardiology*, 57(9):1126–1166.

[Angrist et al., 1996] Angrist, J. D., Imbens, G. W., and Rubin, D. B. (1996). Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*, 91(434):444–455.

[Apostolova et al., 2009] Apostolova, E., Channin, D. S., Demner-Fushman, D., Furst, J., Lytinen, S., and Raicu, D. (2009). Automatic segmentation of clinical texts. *Conference proceedings: ... Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Conference*, 2009:5905–5908.

[Aronson, 2001] Aronson, A. R. (2001). Effective mapping of biomedical text to the UMLS metathesaurus: the MetaMap program. *Proceedings / AMIA ... Annual Symposium. AMIA Symposium*, pages 17–21.

[Asch et al., 1987] Asch, A. S., Barnwell, J., Silverstein, R. L., and Nachman, R. L. (1987). Isolation of the thrombospondin membrane receptor. *The Journal of clinical investigation*, 79(4):1054–1061.

[Baigent et al., 2010] Baigent, C., Blackwell, L., Emberson, J., Holland, L. E., Reith, C., Bhala, N., Peto, R., Barnes, E. H., Keech, A., Simes, J., and Collins, R. (2010). Efficacy and safety of more intensive lowering of LDL cholesterol: a meta-analysis of data from 170 000 participants in 26 randomised trials. *The Lancet*, 376(9753):1670–1681.

[Bakhai et al., 2018] Bakhai, S., Bhardwaj, A., Sandhu, P., and Reynolds, J. L. (2018). Optimisation of lipids for prevention of cardiovascular disease in a primary care. *BMJ open quality*, 7(3):e000071.

[Baxter, 1997] Baxter, J. (1997). A Bayesian/Information theoretic model of learning to learn via multiple task sampling. *Machine learning*, 28(1):7–39.

[Benn et al., 2011] Benn, M., Tybjærg-Hansen, A., Stender, S., Frikke-Schmidt, R., and Nordestgaard, B. G. (2011). Low-density lipoprotein cholesterol and the risk of cancer: a mendelian randomization study. *Journal of the National Cancer Institute*, 103(6):508–519.

[Benson et al., 2017] Benson, M. D., Yang, Q., Ngo, D., Zhu, Y., Shen, D., Farrell, L. A., Sinha, S., Keyes, M. J., Vasan, R. S., Larson, M. G., Smith, J. G., Wang, T. J., and Gerszten, R. E. (2017). The genetic architecture of the cardiovascular risk proteome. *Circulation*.

[Berkson, 1946] Berkson, J. (1946). Limitations of the application of fourfold table analysis to hospital data. *Biometrics*, 2(3):47–53.

[Bezanson et al., 2012] Bezanson, J., Karpinski, S., Shah, V. B., and Edelman, A. (2012). Julia: A fast dynamic language for technical computing.

[Bhatia et al., 2012] Bhatia, R. S., Carne, D. M., Picard, M. H., and Weiner, R. B. (2012). Comparison of the 2007 and 2011 appropriate use criteria for transthoracic echocardiography in various clinical settings. *Journal of the American Society of Echocardiography: official publication of the American Society of Echocardiography*, 25(11):1162–1169.

[Bhatia et al., 2014] Bhatia, R. S., Dudzinski, D. M., Malhotra, R., Milford, C. E., Yoerger Sanborn, D. M., Picard, M. H., and Weiner, R. B. (2014). Educational intervention to reduce outpatient inappropriate echocardiograms: a randomized control trial. *JACC. Cardiovascular imaging*, 7(9):857–866.

[Bhatia et al., 2013] Bhatia, R. S., Milford, C. E., Picard, M. H., and Weiner, R. B. (2013). An educational intervention reduces the rate of inappropriate echocardiograms on an inpatient medical service. *JACC. Cardiovascular imaging*, 6(5):545–555.

[Bild et al., 2002] Bild, D. E., Bluemke, D. A., Burke, G. L., Detrano, R., Diez Roux, A. V., Folsom, A. R., Greenland, P., Jacob, D. R., Kronmal, R., Liu, K., Nelson, J. C., O'Leary, D., Saad, M. F., Shea, S., Szklo, M., and Tracy, R. P. (2002). Multi-Ethnic study of atherosclerosis: objectives and design. *American journal of epidemiology*, 156(9):871–881.

[Bouchard et al., 1995] Bouchard, C., Leon, A. S., Rao, D. C., Skinner, J. S., Wilmore, J. H., and Gagnon, J. (1995). The HERITAGE family study. aims, design, and measurement protocol. *Medicine and science in sports and exercise*, 27(5):721–729.

[Bowden et al., 2015] Bowden, J., Davey Smith, G., and Burgess, S. (2015). Mendelian randomization with invalid instruments: effect estimation and bias detection through egger regression. *International journal of epidemiology*, 44(2):512–525.

[Bowden et al., 2016] Bowden, J., Davey Smith, G., Haycock, P. C., and Burgess, S. (2016). Consistent estimation in mendelian randomization with some invalid instruments using a weighted median estimator. *Genetic epidemiology*, 40(4):304–314.

[Brockbank, 1925] Brockbank, E. M. (1925). Syphilis and aneurysm: Age and sex incidence and cause of death in aortic aneurysm. *British medical journal*, 2(3379):606–607.

[Brosche and Platt, 1998] Brosche, T. and Platt, D. (1998). The biological significance of plasmalogens in defense against oxidative damage. *Experimental gerontology*, 33(5):363–369.

[Brown et al., 2023] Brown, C. J., Chang, L.-S., Hosomura, N., Malmasi, S., Morrison, F., Shubina, M., Lan, Z., and Turchin, A. (2023). Assessment of sex disparities in nonacceptance of statin therapy and Low-Density lipoprotein cholesterol levels among patients at high cardiovascular risk. *JAMA network open*, 6(2):e231047.

[Burgess et al., 2017a] Burgess, S., Bowden, J., Fall, T., Ingelsson, E., and Thompson, S. G. (2017a). Sensitivity analyses for robust causal inference from mendelian randomization analyses with multiple genetic variants. *Epidemiology*, 28(1):30–42.

[Burgess et al., 2013] Burgess, S., Butterworth, A., and Thompson, S. G. (2013). Mendelian randomization analysis with multiple genetic variants using summarized data. *Genetic epidemiology*, 37(7):658–665.

[Burgess et al., 2019] Burgess, S., Davey Smith, G., Davies, N. M., Dudbridge, F., Gill, D., Glymour, M. M., Hartwig, F. P., Holmes, M. V., Minelli, C., Relton, C. L., and Theodoratou, E. (2019). Guidelines for performing mendelian randomization investigations. *Wellcome Open Res*, 4:186.

[Burgess et al., 2016] Burgess, S., Dudbridge, F., and Thompson, S. G. (2016). Combining information on multiple instrumental variables in mendelian randomization: comparison of allele score and summarized data methods. *Statistics in medicine*, 35(11):1880–1906.

[Burgess and Thompson, 2015] Burgess, S. and Thompson, S. G. (2015). *Mendelian Randomization: Methods for Using Genetic Variants in Causal Estimation*. Chapman & Hall/CRC Interdisciplinary Statistics. CRC Press.

[Burgess and Thompson, 2017] Burgess, S. and Thompson, S. G. (2017). Interpreting findings from mendelian randomization using the MR-Egger method. *European journal of epidemiology*, 32(5):377–389.

[Burgess et al., 2011] Burgess, S., Thompson, S. G., and CRP CHD Genetics Collaboration (2011). Avoiding bias from weak instruments in mendelian randomization studies. *International journal of epidemiology*, 40(3):755–764.

[Burgess et al., 2017b] Burgess, S., Zuber, V., Valdes-Marquez, E., Sun, B. B., and Hopewell, J. C. (2017b). Mendelian randomization with fine-mapped genetic data: Choosing from large numbers of correlated instrumental variables. *Genetic epidemiology*, 41(8):714–725.

[Burke et al., 2015] Burke, H. B., Sessums, L. L., Hoang, A., Becher, D. A., Fontelo, P., Liu, F., Stephens, M., Pangaro, L. N., O'Malley, P. G., Baxi, N. S., Bunt, C. W., Capaldi, 2nd, V. F., Chen, J. M., Cooper, B. A., Djuric, D. A., Hodge, J. A., Kane, S., Magee, C., Makary, Z. R., Mallory, R. M., Miller, T., Saperstein, A., Servey, J., and Gimbel, R. W. (2015). Electronic health records improve clinical note quality. *Journal of the American Medical Informatics Association: JAMIA*, 22(1):199–205.

[Butte, 2008] Butte, A. J. (2008). Translational bioinformatics: coming of age. *Journal of the American Medical Informatics Association: JAMIA*, 15(6):709–714.

[Byrd et al., 2014] Byrd, R. J., Steinhubl, S. R., Sun, J., Ebadollahi, S., and Stewart, W. F. (2014). Automatic identification of heart failure diagnostic criteria, using text analysis of clinical notes from electronic health records. *International journal of medical informatics*, 83(12):983–992.

[Cannon et al., 2004] Cannon, C. P., Braunwald, E., McCabe, C. H., Rader, D. J., Rouleau, J. L., Belder, R., Joyal, S. V., Hill, K. A., Pfeffer, M. A., Skene, A. M., and Pravastatin or Atorvastatin Evaluation and Infection Therapy-Thrombolysis in Myocardial Infarction 22 Investigators (2004). Intensive versus moderate lipid lowering with statins after acute coronary syndromes. *The New England journal of medicine*, 350(15):1495–1504.

[Carayol et al., 2017] Carayol, J., Chabert, C., Di Cara, A., Armenise, C., Lefebvre, G., Langin, D., Viguerie, N., Metairon, S., Saris, W. H. M., Astrup, A., Descombes, P., Valsesia, A., and Hager, J. (2017). Protein quantitative trait locus study in obesity during weight-loss identifies a leptin regulator. *Nature communications*, 8(1):2084.

[Carroll et al., 2012] Carroll, R. J., Thompson, W. K., Eyler, A. E., Mandelin, A. M., Cai, T., Zink, R. M., Pacheco, J. A., Boomershine, C. S., Lasko, T. A., Xu, H., Karlson, E. W., Perez, R. G., Gainer, V. S., Murphy, S. N., Ruderman, E. M., Pope, R. M., Plenge, R. M., Kho, A. N., Liao, K. P., and Denny, J. C. (2012). Portability of an algorithm to identify rheumatoid arthritis in

electronic health records. *Journal of the American Medical Informatics Association: JAMIA*, 19(e1):e162–9.

[Caviedes and Cimino, 2004] Caviedes, J. E. and Cimino, J. J. (2004). Towards the development of a conceptual distance metric for the UMLS. *Journal of biomedical informatics*, 37(2):77–85.

[Chambless et al., 1997] Chambless, L. E., Heiss, G., Folsom, A. R., Rosamond, W., Szklo, M., Sharrett, A. R., and Clegg, L. X. (1997). Association of coronary heart disease incidence with carotid arterial wall thickness and major risk factors: the atherosclerosis risk in communities (ARIC) study, 1987-1993. *American journal of epidemiology*, 146(6):483–494.

[Chapman et al., 2011] Chapman, W. W., Nadkarni, P. M., Hirschman, L., D'Avolio, L. W., Savova, G. K., and Uzuner, O. (2011). Overcoming barriers to NLP for clinical text: the role of shared tasks and the need for additional creative solutions. *Journal of the American Medical Informatics Association: JAMIA*, 18(5):540–543.

[Choo et al., 2008] Choo, H. J., Kim, B. W., Kwon, O. B., Lee, C. S., Choi, J. S., and Ko, Y. G. (2008). Secretion of adenylate kinase 1 is required for extracellular ATP synthesis in C2C12 myotubes. *Experimental & molecular medicine*, 40(2):220–228.

[Chou et al., 2016] Chou, R., Dana, T., Blazina, I., Daeges, M., Bougatsos, C., Grusing, S., and Jeanne, T. L. (2016). *Statin Use for the Prevention of Cardiovascular Disease in Adults: A Systematic Review for the U.S. Preventive Services Task Force.* Agency for Healthcare Research and Quality (US), Rockville (MD).

[Coburn and Pauli, 1932] Coburn, A. F. and Pauli, R. H. (1932). Studies on the relationship of streptococcus hemolyticus to the rheumatic process : I. observations on the ecology of hemolytic streptococcus in relation to the epidemiology of rheumatic fever. *The Journal of experimental medicine*, 56(5):609–632.

[Coburn et al., 2000] Coburn, C. T., Knapp, F. F., Febbraio, M., Beets, A. L., Silverstein, R. L., and Abumrad, N. A. (2000). Defective uptake and utilization of long chain fatty acids in muscle and adipose tissues of CD36 knockout mice. *The Journal of biological chemistry*, 275(42):32523–32529.

[Cochrane et al., 2007] Cochrane, L. J., Olson, C. A., Murray, S., Dupuis, M., Tooman, T., and Hayes, S. (2007). Gaps between knowing and doing: understanding and assessing the barriers to optimal health care. *The Journal of continuing education in the health professions*, 27(2):94–102.

[Colantonio et al., 2017] Colantonio, L. D., Richman, J. S., Carson, A. P., Lloyd-Jones, D. M., Howard, G., Deng, L., Howard, V. J., Safford, M. M., Muntner, P., and Goff, Jr, D. C. (2017). Performance of the atherosclerotic cardiovascular disease pooled cohort risk equations by social deprivation status. *Journal of the American Heart Association*, 6(3).

[Collins et al., 2016] Collins, R., Reith, C., Emberson, J., Armitage, J., Baigent, C., Blackwell, L., Blumenthal, R., Danesh, J., Smith, G. D., DeMets, D., Evans, S., Law, M., MacMahon, S., Martin, S., Neal, B., Poulter, N., Preiss, D., Ridker, P., Roberts, I., Rodgers, A., Sandercock, P., Schulz, K., Sever, P., Simes, J., Smeeth, L., Wald, N., Yusuf, S., and Peto, R. (2016). Interpretation of the evidence for the efficacy and safety of statin therapy. *The Lancet*, 388(10059):2532–2561.

[Collobert and Weston, 2008] Collobert, R. and Weston, J. (2008). A unified architecture for natural language processing: Deep neural networks with multitask learning. *Proceedings of the 25th international conference*.

[Coort et al., 2004] Coort, S. L., Hasselbaink, D. M., Koonen, D. P., Willems, J., Coumans, W. A., Chabowski, A., van der Vusse, G. J., Bonen, A., Glatz, J. F., and Luiken, J. J. (2004). Enhanced sarcolemmal FAT/CD36 content and triacylglycerol storage in cardiac myocytes from obese zucker rats. *Diabetes*, 53(7):1655–1663.

[Corder et al., 1993] Corder, E. H., Saunders, A. M., Strittmatter, W. J., Schmechel, D. E., Gaskell, P. C., Small, G. W., Roses, A. D., Haines, J. L., and Pericak-Vance, M. A. (1993). Gene dose of apolipoprotein E type 4 allele and the risk of alzheimer's disease in late onset families. *Science*, 261(5123):921–923.

[Corey et al., 2022] Corey, K. E., Pitts, R., Lai, M., Loureiro, J., Masia, R., Osganian, S. A., Gustafson, J. L., Hutter, M. M., Gee, D. W., Meireles, O. R., Witkowski, E. R., Richards, S. M., Jacob, J., Finkel, N., Ngo, D., Wang, T. J., Gerszten, R. E., Ukomadu, C., and Jennings, L. L. (2022). ADAMTSL2 protein and a soluble biomarker signature identify at-risk non-alcoholic steatohepatitis and fibrosis in adults with NAFLD. *Journal of hepatology*, 76(1):25–33.

[Cornfield, 1962] Cornfield, J. (1962). Joint dependence of risk of coronary heart disease on serum cholesterol and systolic blood pressure: a discriminant function analysis. *Federation proceedings*, 21(4)Pt 2:58–61.

[Cruz et al., 2022] Cruz, D. E., Tahir, U. A., Hu, J., Ngo, D., Chen, Z. Z., Robbins, J. M., Katz, D., Balasubramanian, R., Peterson, B., Deng, S., Benson, M. D., Shi, X., Dailey, L., Gao, Y., Correa, A., Wang, T. J., Clish, C. B., Rexrode, K. M., Wilson, J. G., and Gerszten, R. E. (2022). Metabolomic analysis of coronary heart disease in an african american cohort from the jackson heart study. *JAMA Cardiol*, 7(2):184–194.

[Dai et al., 2015] Dai, H.-J., Syed-Abdul, S., Chen, C.-W., and Wu, C.-C. (2015). Recognition and evaluation of clinical section headings in clinical documents using Token-Based formulation with conditional random fields. *BioMed research international*, 2015:873012.

[Dalli et al., 2022] Dalli, L. L., Kilkenny, M. F., Arnet, I., Sanfilippo, F. M., Cummings, D. M., Kapral, M. K., Kim, J., Cameron, J., Yap, K. Y., Greenland, M., and Cadilhac, D. A. (2022). Towards better reporting of the proportion of days covered method in cardiovascular medication adherence: A scoping review and new tool TEN-SPIDERS. *British journal of clinical pharmacology*, 88(10):4427–4442.

[Daquinag et al., 2021] Daquinag, A. C., Gao, Z., Fussell, C., Immaraj, L., Pasqualini, R., Arap, W., Akimzhanov, A. M., Febbraio, M., and Kolonin, M. G. (2021). Fatty acid mobilization from adipose tissue is mediated by CD36 posttranslational modifications and intracellular trafficking. *JCI Insight*, 6(17).

[Davey Smith and Hemani, 2014] Davey Smith, G. and Hemani, G. (2014). Mendelian randomization: genetic anchors for causal inference in epidemiological studies. *Human molecular genetics*, 23(R1):R89–98.

[Dawber and Kannel, 1958] Dawber, T. R. and Kannel, W. B. (1958). An epidemiologic study of heart disease: the framingham study. *Nutrition reviews*, 16(1):1–4.

[Dawber and Kannel, 1963] Dawber, T. R. and Kannel, W. B. (1963). Coronary heart disease as an epidemiology entity. *American journal of public health and the nation's health*, 53:433–437.

[Dawber et al., 1963] Dawber, T. R., Kannel, W. B., and Lyell, L. P. (1963). An approach to longitudinal studies in a community: the framingham study. *Annals of the New York Academy of Sciences*, 107:539–556.

[Dawber et al., 1962] Dawber, T. R., Kannel, W. B., Revotskie, N., and Kagan, A. (1962). The epidemiology of coronary heart disease–the framingham enquiry. *Proceedings of the Royal Society of Medicine*, 55:265–271.

[Dawber et al., 1959] Dawber, T. R., Kannel, W. B., Revotskie, N., Stokes, 3rd, J., Kagan, A., and Gordon, T. (1959). Some factors associated with the development of coronary heart disease: six years' follow-up experience in the framingham study. *American journal of public health and the nation's health*, 49:1349–1356.

[Dawber et al., 1951] Dawber, T. R., Meadors, G. F., and Moore, Jr, F. E. (1951). Epidemiological approaches to heart disease: the framingham study. *American journal of public health and the nation's health*, 41(3):279–281.

[Dawber et al., 1957] Dawber, T. R., Moore, F. E., and Mann, G. V. (1957). Coronary heart disease in the framingham study. *American journal of public health and the nation's health*, 47(4 Pt 2):4–24.

[DeFilippis et al., 2015] DeFilippis, A. P., Young, R., and Blaha, M. J. (2015). Calibration and discrimination among multiple cardiovascular risk scores in a modern multiethnic cohort. *Annals of internal medicine*, 163(1):68–69.

[DeFilippis et al., 2017] DeFilippis, A. P., Young, R., McEvoy, J. W., Michos, E. D., Sandfort, V., Kronmal, R. A., McClelland, R. L., and Blaha, M. J. (2017). Risk score overestimation: the impact of individual cardiovascular risk factors and preventive therapies on the performance of the american heart Association-American college of Cardiology-Atherosclerotic cardiovascular disease risk score in a modern multi-ethnic cohort. *European heart journal*, 38(8):598–608.

[Demers et al., 2015] Demers, A., Samami, S., Lauzier, B., Des Rosiers, C., Ngo Sock, E. T., Ong, H., and Mayer, G. (2015). PCSK9 induces CD36 degradation and affects Long-Chain fatty acid uptake and triglyceride metabolism in adipocytes and in mouse liver. *Arteriosclerosis, thrombosis, and vascular biology*, 35(12):2517–2525.

[Denny et al., 2009] Denny, J. C., Spickard, 3rd, A., Johnson, K. B., Peterson, N. B., Peterson, J. F., and Miller, R. A. (2009). Evaluation of a method to identify and categorize section headers in clinical documents. *Journal of the American Medical Informatics Association: JAMIA*, 16(6):806–815.

[Desiere et al., 2006] Desiere, F., Deutsch, E. W., King, N. L., Nesvizhskii, A. I., Mallick, P., Eng, J., Chen, S., Eddes, J., Loevenich, S. N., and Aebersold, R. (2006). The PeptideAtlas project.

Nucleic acids research, 34(Database issue):D655–8.

[Devlin et al., 2018] Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding.

[Di Narzo et al., 2017] Di Narzo, A. F., Telesco, S. E., Brodmerkel, C., Argmann, C., Peters, L. A., Li, K., Kidd, B., Dudley, J., Cho, J., Schadt, E. E., Kasarskis, A., Dobrin, R., and Hao, K. (2017). High-Throughput characterization of blood serum proteomics of IBD patients with respect to aging and genetic factors. *PLoS genetics*, 13(1):e1006565.

[Diamond, 1983] Diamond, G. A. (1983). A clinically relevant classification of chest discomfort. *Journal of the American College of Cardiology*, 1(2 Pt 1):574–575.

[Diamond et al., 1980] Diamond, G. A., Forrester, J. S., Hirsch, M., Staniloff, H. M., Vas, R., Berman, D. S., and Swan, H. J. (1980). Application of conditional probability analysis to the clinical diagnosis of coronary artery disease. *The Journal of clinical investigation*, 65(5):1210–1221.

[Dorsch et al., 2019] Dorsch, M. P., Lester, C. A., Ding, Y., Joseph, M., and Brook, R. D. (2019). Effects of race on statin prescribing for primary prevention with high atherosclerotic cardiovascular disease risk in a large healthcare system. *Journal of the American Heart Association*, 8(22):e014709.

[Downs et al., 1998] Downs, J. R., Clearfield, M., Weis, S., Whitney, E., Shapiro, D. R., Beere, P. A., Langendorfer, A., Stein, E. A., Kruyer, W., Gotto, Jr, A. M., and for the AFCAPS/TexCAPS Research Group, for the AFCAPS/TexCAPS Research Group (1998). Primary prevention of acute coronary events with lovastatin in men and women with average cholesterol levels: Results of AFCAPS/TexCAPS. *JAMA: the journal of the American Medical Association*, 279(20):1615–1622.

[Dudzinski et al., 2016] Dudzinski, D. M., Bhatia, R. S., Mi, M. Y., Isselbacher, E. M., Picard, M. H., and Weiner, R. B. (2016). Effect of educational intervention on the rate of rarely appropriate outpatient echocardiograms ordered by attending academic cardiologists: A randomized clinical trial. *JAMA Cardiol*.

[Eisman, 2022] Eisman, A. (2022). DrugConceptSets: Guidelines driven standard concept sets for electronic health records research.

[Eisman et al., 2021] Eisman, A. S., Brown, K. A., Chen, E. S., and Sarkar, I. N. (2021). Clinical note section detection using a hidden markov model of unified medical language system semantic types. *AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium*, 2021:418–427.

[Eisman et al., 2020] Eisman, A. S., Shah, N. R., Eickhoff, C., Zerveas, G., Chen, E. S., Wu, W.-C., and Sarkar, I. N. (2020). Extracting angina symptoms from clinical notes using Pre-Trained transformer architectures. *AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium*, 2020:412–421.

[Eisman et al., 2017] Eisman, A. S., Weiner, R. B., Chen, E. S., Stey, P. C., Wadhera, R. K.,

Kithcart, A. P., and Sarkar, I. N. (2017). An automated system for categorizing transthoracic echocardiography indications according to the echocardiography appropriate use criteria. *AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium*, 2017:670–678.

[Emdin et al., 2017] Emdin, C. A., Khera, A. V., Natarajan, P., Klarin, D., Baber, U., Mehran, R., Rader, D. J., Fuster, V., and Kathiresan, S. (2017). Evaluation of the pooled cohort equations for prediction of cardiovascular risk in a contemporary prospective cohort. *The American journal of cardiology*, 119(6):881–885.

[Emilsson et al., 2018] Emilsson, V., Ilkov, M., Lamb, J. R., Finkel, N., Gudmundsson, E. F., Pitts, R., Hoover, H., Gudmundsdottir, V., Horman, S. R., Aspelund, T., Shu, L., Trifonov, V., Sigurdsson, S., Manolescu, A., Zhu, J., Olafsson, Ö., Jakobsdottir, J., Lesley, S. A., To, J., Zhang, J., Harris, T. B., Launer, L. J., Zhang, B., Eiriksdottir, G., Yang, X., Orth, A. P., Jennings, L. L., and Gudnason, V. (2018). Co-regulatory networks of human serum proteins link genetics to disease. *Science*, 361(6404):769–773.

[Everson et al., 2020] Everson, J., Rubin, J. C., and Friedman, C. P. (2020). Reconsidering hospital EHR adoption at the dawn of HITECH: implications of the reported 9% adoption of a "basic" EHR. *Journal of the American Medical Informatics Association: JAMIA*, 27(8):1198–1205.

[Fabre et al., 2006] Fabre, A. C., Vantourout, P., Champagne, E., Tercé, F., Rolland, C., Perret, B., Collet, X., Barbaras, R., and Martinez, L. O. (2006). Cell surface adenylate kinase activity regulates the F(1)-ATPase/P2Y (13)-mediated HDL endocytosis pathway on human hepatocytes. *Cellular and molecular life sciences: CMLS*, 63(23):2829–2837.

[FADS1-Gene-NCBI, 2023] FADS1-Gene-NCBI (2023). FADS1 fatty acid desaturase 1 [homo sapiens (human)] - gene - NCBI. `https://www.ncbi.nlm.nih.gov/gene?Db=gene&Cmd=ShowDetailView&TermToSearch=3992`. Accessed: 2023-2-14.

[FADS2-Gene-NCBI, 2023] FADS2-Gene-NCBI (2023). FADS2 fatty acid desaturase 2 [homo sapiens (human)] - gene - NCBI. `https://www.ncbi.nlm.nih.gov/gene/9415`. Accessed: 2023-2-14.

[Fahy and Subramaniam, 2020] Fahy, E. and Subramaniam, S. (2020). RefMet: a reference nomenclature for metabolomics. *Nature methods*, 17(12):1173–1174.

[Fang et al., 2020] Fang, W., Deng, Z., Benadjaoud, F., Yang, C., and Shi, G. P. (2020). Cathepsin B deficiency ameliorates liver lipid deposition, inflammatory cell infiltration, and fibrosis after diet-induced nonalcoholic steatohepatitis. *Translational research: the journal of laboratory and clinical medicine*, 222:28–40.

[Feldman et al., 2022] Feldman, K. A., Hanks, A., Williams, T. W., Blouse, B., Babcock, C., Goode, A., Dowling, S., Thakuria, A., Teichmann, J., Aljalahma, H., and Pearlowitz, M. (2022). A state health department and health information exchange partnership: an effective collaboration for a Data-Driven response for COVID-19 contact tracing in maryland. *Sexually transmitted diseases*.

[Ferkingstad et al., 2021] Ferkingstad, E., Sulem, P., Atlason, B. A., Sveinbjornsson, G., Magnusson, M. I., Styrmisdottir, E. L., Gunnarsdottir, K., Helgason, A., Oddsson, A., Halldorsson, B. V., Jensson, B. O., Zink, F., Halldorsson, G. H., Masson, G., Arnadottir, G. A., Katrinardottir, H.,

Juliusson, K., Magnusson, M. K., Magnusson, O. T., Fridriksdottir, R., Saevarsdottir, S., Gudjonsson, S. A., Stacey, S. N., Rognvaldsson, S., Eiriksdottir, T., Olafsdottir, T. A., Steinthorsdottir, V., Tragante, V., Ulfarsson, M. O., Stefansson, H., Jonsdottir, I., Holm, H., Rafnar, T., Melsted, P., Saemundsdottir, J., Norddahl, G. L., Lund, S. H., Gudbjartsson, D. F., Thorsteinsdottir, U., and Stefansson, K. (2021). Large-scale integration of the plasma proteome with genetics and disease. *Nature genetics*, 53(12):1712–1721.

[Fields and Fields, 2014] Fields, K. E. and Fields, B. J. (2014). *Racecraft: The Soul of Inequality in American Life*. Verso Books.

[Fısher, 1935] Fısher, R. A. (1935). The design of experiments. *Oliver and Boyd, Edinburgh*.

[Folkersen et al., 2017] Folkersen, L., Fauman, E., Sabater-Lleal, M., Strawbridge, R. J., Frånberg, M., Sennblad, B., Baldassarre, D., Veglia, F., Humphries, S. E., Rauramaa, R., de Faire, U., Smit, A. J., Giral, P., Kurl, S., Mannarino, E., Enroth, S., Johansson, Å., Enroth, S. B., Gustafsson, S., Lind, L., Lindgren, C., Morris, A. P., Giedraitis, V., Silveira, A., Franco-Cereceda, A., Tremoli, E., Gyllensten, U., Ingelsson, E., Brunak, S., Eriksson, P., Ziemek, D., Hamsten, A., Mälarstig, A., and Group, I. S. (2017). Mapping of 79 loci for 83 plasma protein biomarkers in cardiovascular disease. *PLoS genetics*, 13(4):e1006706.

[Folsom et al., 2001] Folsom, A. R., Aleksic, N., Park, E., Salomaa, V., Juneja, H., and Wu, K. K. (2001). Prospective study of fibrinolytic factors and incident coronary heart disease: the atherosclerosis risk in communities (ARIC) study. *Arteriosclerosis, thrombosis, and vascular biology*, 21(4):611–617.

[Folsom et al., 1997] Folsom, A. R., Wu, K. K., Rosamond, W. D., Sharrett, A. R., and Chambless, L. E. (1997). Prospective study of hemostatic factors and incidence of coronary heart disease: the atherosclerosis risk in communities (ARIC) study. *Circulation*, 96(4):1102–1108.

[Fonseca et al., 2015] Fonseca, R., Negishi, K., and Marwick, T. H. (2015). What is the evidence status of appropriate use criteria (AUC)? insight from a matching exercise with the guidelines for echocardiography. *Internal medicine journal*, 45(8):864–869.

[Franekova et al., 2015] Franekova, V., Angin, Y., Hoebers, N. T., Coumans, W. A., Simons, P. J., Glatz, J. F., Luiken, J. J., and Larsen, T. S. (2015). Marine omega-3 fatty acids prevent myocardial insulin resistance and metabolic remodeling as induced experimentally by high insulin exposure. *American journal of physiology. Cell physiology*, 308(4):C297–307.

[FRCPath et al., 2020] FRCPath, V. K. M., Abbas MBBS, A. K., and Aster, J. C. (2020). *Robbins & Cotran Pathologic Basis of Disease (Robbins Pathology)*. Elsevier, 10 edition.

[Fried et al., 1991] Fried, L. P., Borhani, N. O., Enright, P., Furberg, C. D., Gardin, J. M., Kronmal, R. A., Kuller, L. H., Manolio, T. A., Mittelmark, M. B., and Newman, A. (1991). The cardiovascular health study: design and rationale. *Annals of epidemiology*, 1(3):263–276.

[Friedman et al., 1988] Friedman, G. D., Cutter, G. R., Donahue, R. P., Hughes, G. H., Hulley, S. B., Jacobs, Jr, D. R., Liu, K., and Savage, P. J. (1988). CARDIA: study design, recruitment, and some characteristics of the examined subjects. *Journal of clinical epidemiology*, 41(11):1105–1116.

[Fung et al., 2018] Fung, V., Graetz, I., Reed, M., and Jaffe, M. G. (2018). Patient-reported adherence to statin therapy, barriers to adherence, and perceptions of cardiovascular risk. *PloS one*, 13(2):e0191817.

[Galea and Abdalla, 2018] Galea, S. and Abdalla, S. M. (2018). Precision medicine approaches and the health of populations: Study design concerns and considerations. *Perspectives in biology and medicine*, 61(4):527–536.

[Genders et al., 2011] Genders, T. S. S., Steyerberg, E. W., Alkadhi, H., Leschka, S., Desbiolles, L., Nieman, K., Galema, T. W., Meijboom, W. B., Mollet, N. R., de Feyter, P. J., Cademartiri, F., Maffei, E., Dewey, M., Zimmermann, E., Laule, M., Pugliese, F., Barbagallo, R., Sinitsyn, V., Bogaert, J., Goetschalckx, K., Schoepf, U. J., Rowe, G. W., Schuijf, J. D., Bax, J. J., de Graaf, F. R., Knuuti, J., Kajander, S., van Mieghem, C. A. G., Meijs, M. F. L., Cramer, M. J., Gopalan, D., Feuchtner, G., Friedrich, G., Krestin, G. P., Hunink, M. G. M., and CAD Consortium (2011). A clinical prediction rule for the diagnosis of coronary artery disease: validation, updating, and extension. *European heart journal*, 32(11):1316–1330.

[Gentleman et al., 2004] Gentleman, R. C., Carey, V. J., Bates, D. M., Bolstad, B., Dettling, M., Dudoit, S., Ellis, B., Gautier, L., Ge, Y., Gentry, J., Hornik, K., Hothorn, T., Huber, W., Iacus, S., Irizarry, R., Leisch, F., Li, C., Maechler, M., Rossini, A. J., Sawitzki, G., Smith, C., Smyth, G., Tierney, L., Yang, J. Y., and Zhang, J. (2004). Bioconductor: open software development for computational biology and bioinformatics. *Genome biology*, 5(10):R80.

[Gerszten et al., 2008] Gerszten, R. E., Accurso, F., Bernard, G. R., Caprioli, R. M., Klee, E. W., Klee, G. G., Kullo, I., Laguna, T. A., Roth, F. P., Sabatine, M., Srinivas, P., Wang, T. J., and Ware, L. B. (2008). Challenges in translating plasma proteomics from bench to bedside: update from the NHLBI clinical proteomics programs. *American journal of physiology. Lung cellular and molecular physiology*, 295(1):L16–22.

[Gieger et al., 2008] Gieger, C., Geistlinger, L., Altmaier, E., Hrabé de Angelis, M., Kronenberg, F., Meitinger, T., Mewes, H. W., Wichmann, H. E., Weinberger, K. M., Adamski, J., Illig, T., and Suhre, K. (2008). Genetics meets metabolomics: a genome-wide association study of metabolite profiles in human serum. *PLoS genetics*, 4(11):e1000282.

[Glatz and Luiken, 2015] Glatz, J. F. and Luiken, J. J. (2015). Fatty acids in cell signaling: historical perspective and future outlook. *Prostaglandins, leukotrienes, and essential fatty acids*, 92:57–62.

[Goff et al., 2014] Goff, Jr, D. C., Lloyd-Jones, D. M., Bennett, G., Coady, S., D'Agostino, Sr, R. B., Gibbons, R., Greenland, P., Lackland, D. T., Levy, D., O'Donnell, C. J., Robinson, J. G., Schwartz, J. S., Shero, S. T., Smith, Jr, S. C., Sorlie, P., Stone, N. J., Wilson, P. W. F., and American College of Cardiology/American Heart Association Task Force on Practice Guidelines (2014). 2013 ACC/AHA guideline on the assessment of cardiovascular risk: a report of the american college of Cardiology/American heart association task force on practice guidelines. *Journal of the American College of Cardiology*, 63(25 Pt B):2935–2959.

[Gordon et al., 1975] Gordon, T., Kannel, W. B., Dawber, T. R., and McGee, D. (1975). Changes associated with quitting cigarette smoking: the framingham study. *American heart journal*, 90(3):322–328.

[Gordon et al., 1974] Gordon, T., Kannel, W. B., McGee, D., and Dawber, T. R. (1974). Death and coronary attacks in men after giving up cigarette smoking. a report from the framingham study. *The Lancet*, 2(7893):1345–1348.

[Graham, 2015] Graham, G. (2015). Disparities in cardiovascular disease risk in the united states. *Current cardiology reviews*, 11(3):238–245.

[Grundy et al., 2019] Grundy, S. M., Stone, N. J., Bailey, A. L., Beam, C., Birtcher, K. K., Blumenthal, R. S., Braun, L. T., de Ferranti, S., Faiella-Tommasino, J., Forman, D. E., Goldberg, R., Heidenreich, P. A., Hlatky, M. A., Jones, D. W., Lloyd-Jones, D., Lopez-Pajares, N., Ndumele, C. E., Orringer, C. E., Peralta, C. A., Saseen, J. J., Smith, Jr, S. C., Sperling, L., Virani, S. S., and Yeboah, J. (2019). 2018 AHA/ACC/AACVPR/AAPA/ABC/ACPM/ADA/AGS/APhA/ASPC/NLA/PCNA guideline on the management of blood cholesterol: A report of the american college of cardiology/american heart association task force on clinical practice guidelines. *Circulation*, 139(25):e1082–e1143.

[Gudjonsson et al., 2022] Gudjonsson, A., Gudmundsdottir, V., Axelsson, G. T., Gudmundsson, E. F., Jonsson, B. G., Launer, L. J., Lamb, J. R., Jennings, L. L., Aspelund, T., Emilsson, V., and Gudnason, V. (2022). A genome-wide association study of serum proteins reveals shared loci with common diseases. *Nature communications*, 13(1):480.

[Gupta et al., 2016] Gupta, H., Schiros, C. G., Sharifov, O. F., Jain, A., and Denney, Jr, T. S. (2016). Impact of clinical input variable uncertainties on ten-year atherosclerotic cardiovascular disease risk using new pooled cohort equations. *BMC cardiovascular disorders*, 16(1):165.

[Guy et al., 2007] Guy, E., Kuchibhotla, S., Silverstein, R., and Febbraio, M. (2007). Continued inhibition of atherosclerotic lesion development in long term western diet fed CD36o /apoeo mice. *Atherosclerosis*, 192(1):123–130.

[Haase et al., 2015] Haase, C. L., Tybjærg-Hansen, A., Nordestgaard, B. G., and Frikke-Schmidt, R. (2015). HDL cholesterol and risk of type 2 diabetes: A mendelian randomization study. *Diabetes*, 64(9):3328–3333.

[Haase et al., 2012] Haase, C. L., Tybjærg-Hansen, A., Qayyum, A. A., Schou, J., Nordestgaard, B. G., and Frikke-Schmidt, R. (2012). LCAT, HDL cholesterol and ischemic cardiovascular disease: a mendelian randomization study of HDL cholesterol in 54,500 individuals. *The Journal of clinical endocrinology and metabolism*, 97(2):E248–56.

[Haendel et al., 2020] Haendel, M. A., Chute, C. G., Bennett, T. D., Eichmann, D. A., Guinney, J., Kibbe, W. A., Payne, P. R. O., Pfaff, E. R., Robinson, P. N., Saltz, J. H., and Others (2020). The national COVID cohort collaborative (N3C): rationale, design, infrastructure, and deployment. *Journal of the American Medical Informatics Association: JAMIA*.

[Hamilton and Fielding, 1989] Hamilton, R. L. and Fielding, P. E. (1989). Nascent very low density lipoproteins from rat hepatocytic golgi fractions are enriched in phosphatidylethanolamine. *Biochemical and biophysical research communications*, 160(1):162–173.

[Han, 2008] Han, C. (2008). Detecting invalid instruments using L1-GMM.

[Hansson et al., 2002] Hansson, G. K., Libby, P., Schönbeck, U., and Yan, Z.-Q. (2002). Innate and adaptive immunity in the pathogenesis of atherosclerosis. *Circulation research*, 91(4):281–291.

[Harmon and Abumrad, 1993] Harmon, C. M. and Abumrad, N. A. (1993). Binding of sulfosuccinimidyl fatty acids to adipocyte membrane proteins: isolation and amino-terminal sequence of an 88-kd protein implicated in transport of long-chain fatty acids. *The Journal of membrane biology*, 133(1):43–49.

[Harris et al., 2019] Harris, P. A., Taylor, R., Minor, B. L., Elliott, V., Fernandez, M., O'Neal, L., McLeod, L., Delacqua, G., Delacqua, F., Kirby, J., Duda, S. N., and REDCap Consortium (2019). The REDCap consortium: Building an international community of software platform partners. *Journal of biomedical informatics*, 95:103208.

[Harris et al., 2009] Harris, P. A., Taylor, R., Thielke, R., Payne, J., Gonzalez, N., Conde, J. G., and Others (2009). A metadata-driven methodology and workflow process for providing translational research informatics support. *Journal of biomedical informatics*, 42(2):377–381.

[Hearst, 1994] Hearst, M. A. (1994). Multi-paragraph segmentation of expository text. In *Proceedings of the 32Nd Annual Meeting on Association for Computational Linguistics*, ACL '94, pages 9–16, Stroudsburg, PA, USA. Association for Computational Linguistics.

[Hearst, 1997] Hearst, M. A. (1997). TextTiling: Segmenting text into multi-paragraph subtopic passages. *Computational Linguistics*, 23(1):33–64.

[Hernán et al., 2011] Hernán, M. A., Clayton, D., and Keiding, N. (2011). The simpson's paradox unraveled. *International journal of epidemiology*, 40(3):780–785.

[Hernán et al., 2004] Hernán, M. A., Hernández-Díaz, S., and Robins, J. M. (2004). A structural approach to selection bias. *Epidemiology*, 15(5):615–625.

[Hernán et al., 2002] Hernán, M. A., Hernández-Díaz, S., Werler, M. M., and Mitchell, A. A. (2002). Causal knowledge as a prerequisite for confounding evaluation: an application to birth defects epidemiology. *American journal of epidemiology*, 155(2):176–184.

[HGF-Gene-NCBI, 2023] HGF-Gene-NCBI (2023). HGF hepatocyte growth factor [homo sapiens (human)] - gene - NCBI. `https://www.ncbi.nlm.nih.gov/gene/3082`. Accessed: 2023-2-14.

[Hicks et al., 2004] Hicks, L. S., Fairchild, D. G., Horng, M. S., Orav, E. J., Bates, D. W., and Ayanian, J. Z. (2004). Determinants of JNC VI guideline adherence, intensity of drug therapy, and blood pressure control by race and ethnicity. *Hypertension*, 44(4):429–434.

[Holt et al., 2003] Holt, J. A., Luo, G., Billin, A. N., Bisi, J., McNeill, Y. Y., Kozarsky, K. F., Donahee, M., Wang, D. Y., Mansfield, T. A., Kliewer, S. A., Goodwin, B., and Jones, S. A. (2003). Definition of a novel growth factor-dependent signal cascade for the suppression of bile acid biosynthesis. *Genes & development*, 17(13):1581–1591.

[Howard et al., 2011] Howard, G., Cushman, M., Kissela, B. M., Kleindorfer, D. O., McClure, L. A.,

Safford, M. M., Rhodes, J. D., Soliman, E. Z., Moy, C. S., Judd, S. E., Howard, V. J., and REasons for Geographic And Racial Differences in Stroke (REGARDS) Investigators (2011). Traditional risk factors as the underlying cause of racial disparities in stroke: lessons from the half-full (empty?) glass. *Stroke; a journal of cerebral circulation*, 42(12):3369–3375.

[Hughes Halbert et al., 2016] Hughes Halbert, C., Welch, B., Lynch, C., Magwood, G., Rice, L., Jefferson, M., and Riley, J. (2016). Social determinants of family health history collection. *Journal of community genetics*, 7(1):57–64.

[Ibrahimi et al., 1999] Ibrahimi, A., Bonen, A., Blinn, W. D., Hajri, T., Li, X., Zhong, K., Cameron, R., and Abumrad, N. A. (1999). Muscle-specific overexpression of FAT/CD36 enhances fatty acid oxidation by contracting muscle, reduces plasma triglycerides and fatty acids, and increases plasma glucose and insulin. *The Journal of biological chemistry*, 274(38):26761–26766.

[Illig et al., 2010] Illig, T., Gieger, C., Zhai, G., Römisch-Margl, W., Wang-Sattler, R., Prehn, C., Altmaier, E., Kastenmüller, G., Kato, B. S., Mewes, H. W., Meitinger, T., de Angelis, M. H., Kronenberg, F., Soranzo, N., Wichmann, H. E., Spector, T. D., Adamski, J., and Suhre, K. (2010). A genome-wide perspective of genetic variation in human metabolism. *Nature genetics*, 42(2):137–141.

[Intervention with Pravastatin in and others, 1998] Intervention with Pravastatin in, L.-T. and others (1998). Prevention of cardiovascular events and death with pravastatin in patients with coronary heart disease and a broad range of initial cholesterol levels. *The New England journal of medicine*, 339(19):1349–1357.

[Isezuo, 2005] Isezuo, S. A. (2005). Is high density lipoprotein cholesterol useful in diagnosis of metabolic syndrome in native africans with type 2 diabetes? *Ethnicity & disease*, 15(1):6–10.

[Jacobs et al., 2023] Jacobs, J. A., Addo, D. K., Zheutlin, A. R., Derington, C. G., Essien, U. R., Navar, A. M., Hernandez, I., Lloyd-Jones, D. M., King, J. B., Rao, S., Herrick, J. S., Bress, A. P., and Pandey, A. (2023). Prevalence of statin use for primary prevention of atherosclerotic cardiovascular disease by race, ethnicity, and 10-year disease risk in the US: National health and nutrition examination surveys, 2013 to march 2020. *JAMA cardiology*.

[Jacobs, 2009] Jacobs, L. (2009). Interview with lawrence weed, MD- the father of the Problem-Oriented medical record looks ahead. *The Permanente journal*, 13(3):84–89.

[Johnson et al., 2016] Johnson, A. E. W., Pollard, T. J., Shen, L., Lehman, L.-W. H., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Celi, L. A., and Mark, R. G. (2016). MIMIC-III, a freely accessible critical care database. *Scientific data*, 3:160035.

[Jones et al., 2002] Jones, D. W., Chambless, L. E., Folsom, A. R., Heiss, G., Hutchinson, R. G., Sharrett, A. R., Szklo, M., and Taylor, Jr, H. A. (2002). Risk factors for coronary heart disease in african americans: the atherosclerosis risk in communities study, 1987-1997. *Archives of internal medicine*, 162(22):2565–2571.

[Kagan et al., 1963] Kagan, A., Kannel, W. B., Dawber, T. R., and Revotskie, N. (1963). The coronary profile. *Annals of the New York Academy of Sciences*, 97:883–894.

[Kannel, 1985] Kannel, W. B. (1985). Lipids, diabetes, and coronary heart disease: insights from the framingham study. *American heart journal*, 110(5):1100–1107.

[Kannel et al., 1965] Kannel, W. B., Dawber, T. R., Cohen, M. E., and Mcnamara, P. M. (1965). Vascular disease of the Brain–Epidemiologic aspects: The framingham study. *American journal of public health and the nation's health*, 55:1355–1366.

[Kannel et al., 1979] Kannel, W. B., Feinleib, M., McNamara, P. M., Garrison, R. J., and Castelli, W. P. (1979). An investigation of coronary heart disease in families. the framingham offspring study. *American journal of epidemiology*, 110(3):281–290.

[Kannel and McGee, 1979] Kannel, W. B. and McGee, D. L. (1979). Diabetes and cardiovascular disease. the framingham study. *JAMA: the journal of the American Medical Association*, 241(19):2035–2038.

[Katz et al., 2022] Katz, D. H., Tahir, U. A., Bick, A. G., Pampana, A., Ngo, D., Benson, M. D., Yu, Z., Robbins, J. M., Chen, Z. Z., Cruz, D. E., Deng, S., Farrell, L., Sinha, S., Schmaier, A. A., Shen, D., Gao, Y., Hall, M. E., Correa, A., Tracy, R. P., Durda, P., Taylor, K. D., Liu, Y., Johnson, W. C., Guo, X., Yao, J., Ida Chen, Y. D., Manichaikul, A. W., Jain, D., Bouchard, C., Sarzynski, M. A., Rich, S. S., Rotter, J. I., Wang, T. J., Wilson, J. G., Natarajan, P., Gerszten, R. E., National Heart, and for Precision Medicine) Consortium†, B. I. T. t.-O. (2022). Whole genome sequence analysis of the plasma proteome in black adults provides novel insights into cardiovascular disease. *Circulation*, 145(5):357–370.

[Kendall et al., 2018] Kendall, A., Gal, Y., and Cipolla, R. (2018). Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7482–7491. openaccess.thecvf.com.

[Kettunen et al., 2012] Kettunen, J., Tukiainen, T., Sarin, A. P., Ortega-Alonso, A., Tikkanen, E., Lyytikäinen, L. P., Kangas, A. J., Soininen, P., Würtz, P., Silander, K., Dick, D. M., Rose, R. J., Savolainen, M. J., Viikari, J., Kähönen, M., Lehtimäki, T., Pietiläinen, K. H., Inouye, M., McCarthy, M. I., Jula, A., Eriksson, J., Raitakari, O. T., Salomaa, V., Kaprio, J., Järvelin, M. R., Peltonen, L., Perola, M., Freimer, N. B., Ala-Korpela, M., Palotie, A., and Ripatti, S. (2012). Genome-wide association study identifies multiple loci influencing human serum metabolite levels. *Nature genetics*, 44(3):269–276.

[Koleck et al., 2019] Koleck, T. A., Dreisbach, C., Bourne, P. E., and Bakken, S. (2019). Natural language processing of symptoms documented in free-text narratives of electronic health records: a systematic review. *Journal of the American Medical Informatics Association: JAMIA*, 26(4):364–379.

[Kraus et al., 2015] Kraus, W. E., Muoio, D. M., Stevens, R., Craig, D., Bain, J. R., Grass, E., Haynes, C., Kwee, L., Qin, X., Slentz, D. H., Krupp, D., Muehlbauer, M., Hauser, E. R., Gregory, S. G., Newgard, C. B., and Shah, S. H. (2015). Metabolomic quantitative trait loci (mQTL) mapping implicates the ubiquitin proteasome system in cardiovascular disease pathogenesis. *PLoS genetics*, 11(11):e1005553.

[Kurian and Cardarelli, 2007] Kurian, A. K. and Cardarelli, K. M. (2007). Racial and ethnic differences in cardiovascular disease risk factors: a systematic review. *Ethnicity & disease*, 17(1):143–152.

[Kurokawa et al., 2010] Kurokawa, J., Arai, S., Nakashima, K., Nagano, H., Nishijima, A., Miyata, K., Ose, R., Mori, M., Kubota, N., Kadowaki, T., Oike, Y., Koga, H., Febbraio, M., Iwanaga, T., and Miyazaki, T. (2010). Macrophage-derived AIM is endocytosed into adipocytes and decreases lipid droplets via inhibition of fatty acid synthase activity. *Cell metabolism*, 11(6):479–492.

[LaRosa et al., 2005] LaRosa, J. C., Grundy, S. M., Waters, D. D., Shear, C., Barter, P., Fruchart, J.-C., Gotto, A. M., Greten, H., Kastelein, J. J. P., Shepherd, J., Wenger, N. K., and Treating to New Targets (TNT) Investigators (2005). Intensive lipid lowering with atorvastatin in patients with stable coronary disease. *The New England journal of medicine*, 352(14):1425–1435.

[Last et al., 2017] Last, A. R., Ference, J. D., and Menzel, E. R. (2017). Hyperlipidemia: Drugs for cardiovascular risk reduction in adults. *American family physician*, 95(2):78–87.

[Lawlor et al., 2008] Lawlor, D. A., Harbord, R. M., Sterne, J. A. C., Timpson, N., and Davey Smith, G. (2008). Mendelian randomization: using genes as instruments for making causal inferences in epidemiology. *Statistics in medicine*, 27(8):1133–1163.

[Lee et al., 2020] Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., and Kang, J. (2020). BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

[Levitt et al., 2016] Levitt, K., Shojania, K. G., and Bhatia, R. S. (2016). Point-of-care decision support for reducing inappropriate test use: easier said than done. *BMJ quality & safety*, 25(1):6–8.

[Li et al., 2010] Li, Y., Lipsky Gorman, S., and Elhadad, N. (2010). Section classification in clinical notes using supervised hidden markov model. In *Proceedings of the 1st ACM International Health Informatics Symposium*, IHI '10, pages 744–750, New York, NY, USA. ACM.

[Lilly and Braunwald, 2012] Lilly, L. S. and Braunwald, E. (2012). *Braunwald's Heart Disease: A Textbook of Cardiovascular Medicine*. Elsevier Health Sciences.

[Lohner, 1996] Lohner, K. (1996). Is the high propensity of ethanolamine plasmalogens to form non-lamellar lipid structures manifested in the properties of biomembranes? *Chemistry and physics of lipids*, 81(2):167–184.

[Long et al., 2017] Long, T., Hicks, M., Yu, H. C., Biggs, W. H., Kirkness, E. F., Menni, C., Zierer, J., Small, K. S., Mangino, M., Messier, H., Brewerton, S., Turpaz, Y., Perkins, B. A., Evans, A. M., Miller, L. A., Guo, L., Caskey, C. T., Schork, N. J., Garner, C., Spector, T. D., Venter, J. C., and Telenti, A. (2017). Whole-genome sequencing identifies common-to-rare variants associated with human blood metabolites. *Nature genetics*, 49(4):568–578.

[Loprinzi and Joyner, 2016] Loprinzi, P. D. and Joyner, C. (2016). Accelerometer-determined physical activity and mortality in a national prospective cohort study: Considerations by visual acuity. *Preventive medicine*, 87:18–21.

[Lotta et al., 2021] Lotta, L. A., Pietzner, M., Stewart, I. D., Wittemans, L. B. L., Li, C., Bonelli, R., Raffler, J., Biggs, E. K., Oliver-Williams, C., Auyeung, V. P. W., Luan, J., Wheeler, E., Paige, E., Surendran, P., Michelotti, G. A., Scott, R. A., Burgess, S., Zuber, V., Sanderson, E., Koulman, A., Imamura, F., Forouhi, N. G., Khaw, K. T., Griffin, J. L., Wood, A. M., Kastenmüller, G., Danesh, J., Butterworth, A. S., Gribble, F. M., Reimann, F., Bahlo, M., Fauman, E., Wareham, N. J., Langenberg, C., and Consortium, M. (2021). A cross-platform approach identifies genetic regulators of human metabolism and health. *Nature genetics*, 53(1):54–64.

[Lourdusamy et al., 2012] Lourdusamy, A., Newhouse, S., Lunnon, K., Proitsi, P., Powell, J., Hodges, A., Nelson, S. K., Stewart, A., Williams, S., Kloszewska, I., Mecocci, P., Soininen, H., Tsolaki, M., Vellas, B., Lovestone, S., Dobson, R., Consortium, A., and Initiative, A. D. N. (2012). Identification of cis-regulatory variation influencing protein abundance levels in human plasma. *Human molecular genetics*, 21(16):3719–3726.

[Mahley and Rall, 2000] Mahley, R. W. and Rall, Jr, S. C. (2000). Apolipoprotein e: far more than a lipid transport protein. *Annual review of genomics and human genetics*, 1:507–537.

[Makadia and Ryan, 2014] Makadia, R. and Ryan, P. B. (2014). Transforming the premier perspective hospital database into the observational medical outcomes partnership (OMOP) common data model. *EGEMS (Washington, DC)*, 2(1):1110.

[Matthews, 1975] Matthews, B. W. (1975). Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochimica et biophysica acta*, 405(2):442–451.

[McKenney, 1988] McKenney, J. M. (1988). Lovastatin: a new cholesterol-lowering agent. *Clinical pharmacy*, 7(1):21–36.

[McSweeney et al., 2006] McSweeney, J. C., O'Sullivan, P., Cody, M., Kovacs, M., Dunn, M., Webb-Corbett, R., Ramer, L., Crane, P., Green, A., and Lefler, L. (2006). Abstract 3306: Black, hispanic, and white women's symptoms of coronary heart disease. *Circulation*.

[MedPAC, 2014] MedPAC (2014). Healthcare spending and the medicare program. `https://www.medpac.gov`. Accessed: 2014-NA-NA.

[Meir and Leitersdorf, 2004] Meir, K. S. and Leitersdorf, E. (2004). Atherosclerosis in the apolipoprotein-e-deficient mouse: a decade of progress. *Arteriosclerosis, thrombosis, and vascular biology*, 24(6):1006–1014.

[Melis et al., 2017] Melis, M., Carta, G., Pintus, S., Pintus, P., Piras, C. A., Murru, E., Manca, C., Di Marzo, V., Banni, S., and Tomassini Barbarossa, I. (2017). Polymorphism. *Frontiers in physiology*, 8:1006.

[Menachemi et al., 2018] Menachemi, N., Rahurkar, S., Harle, C. A., and Vest, J. R. (2018). The benefits of health information exchange: an updated systematic review. *Journal of the American Medical Informatics Association: JAMIA*, 25(9):1259–1265.

[Mensah et al., 2005] Mensah, G. A., Mokdad, A. H., Ford, E. S., Greenlund, K. J., and Croft, J. B. (2005). State of disparities in cardiovascular health in the united states. *Circulation*, 111(10):1233–1241.

[Metser et al., 2021] Metser, G., Bradley, C., Moise, N., Liyanage-Don, N., Kronish, I., and Ye, S. (2021). Gaps and disparities in primary prevention statin prescription during outpatient care. *The American journal of cardiology*, 161:36–41.

[Mistry et al., 2021] Mistry, J. J., Hellmich, C., Moore, J. A., Jibril, A., Macaulay, I., Moreno-Gonzalez, M., Di Palma, F., Beraza, N., Bowles, K. M., and Rushworth, S. A. (2021). Free fatty-acid transport via CD36 drives β-oxidation-mediated hematopoietic stem cell response to infection. *Nature communications*, 12(1):7130.

[Mokry et al., 2015] Mokry, L. E., Ross, S., Ahmad, O. S., Forgetta, V., Smith, G. D., Goltzman, D., Leong, A., Greenwood, C. M., Thanassoulis, G., and Richards, J. B. (2015). Vitamin D and risk of multiple sclerosis: A mendelian randomization study. *PLoS medicine*, 12(8):e1001866.

[Mootha et al., 2003] Mootha, V. K., Lindgren, C. M., Eriksson, K. F., Subramanian, A., Sihag, S., Lehar, J., Puigserver, P., Carlsson, E., Ridderstråle, M., Laurila, E., Houstis, N., Daly, M. J., Patterson, N., Mesirov, J. P., Golub, T. R., Tamayo, P., Spiegelman, B., Lander, E. S., Hirschhorn, J. N., Altshuler, D., and Groop, L. C. (2003). PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nature genetics*, 34(3):267–273.

[Morris et al., 2013] Morris, A. A., Patel, R. S., Binongo, J. N. G., Poole, J., Al Mheid, I., Ahmed, Y., Stoyanova, N., Vaccarino, V., Din-Dzietham, R., Gibbons, G. H., and Quyyumi, A. (2013). Racial differences in arterial stiffness and microcirculatory function between black and white americans. *Journal of the American Heart Association*, 2(2):e002154.

[Morris and Hirst, 1991] Morris, J. and Hirst, G. (1991). Lexical cohesion computed by thesaural relations as an indicator of the structure of text. *Computational Linguistics*, 17(1):21–48.

[MTSamples, 2020] MTSamples (2020). Transcribed medical transcription sample reports and examples - MTSamples. `https://mtsamples.com/`. Accessed: 2020-1-1.

[Muntner et al., 2014] Muntner, P., Colantonio, L. D., Cushman, M., Goff, Jr, D. C., Howard, G., Howard, V. J., Kissela, B., Levitan, E. B., Lloyd-Jones, D. M., and Safford, M. M. (2014). Validation of the atherosclerotic cardiovascular disease pooled cohort risk equations. *JAMA: the journal of the American Medical Association*, 311(14):1406–1415.

[Murphy et al., 2021] Murphy, S. L., Kochanek, K. D., Xu, J., and Arias, E. (2021). Mortality in the united states, 2020. *NCHS data brief*, (427):1–8.

[Myers et al., 1990] Myers, R. H., Kiely, D. K., Cupples, L. A., and Kannel, W. B. (1990). Parental history is an independent risk factor for coronary artery disease: the framingham study. *American heart journal*, 120(4):963–969.

[Nagan and Zoeller, 2001] Nagan, N. and Zoeller, R. A. (2001). Plasmalogens: biosynthesis and functions. *Progress in lipid research*, 40(3):199–229.

[Nakamura et al., 2006] Nakamura, H., Arakawa, K., Itakura, H., Kitabatake, A., Goto, Y., Toyota, T., Nakaya, N., Nishimoto, S., Muranaka, M., Yamamoto, A., Mizuno, K., Ohashi, Y., and MEGA Study Group (2006). Primary prevention of cardiovascular disease with pravastatin in

japan (MEGA study): a prospective randomised controlled trial. *The Lancet*, 368(9542):1155–1163.

[Nath et al., 2016a] Nath, C., Albaghdadi, M. S., and Jonnalagadda, S. R. (2016a). A natural language processing tool for Large-Scale data extraction from echocardiography reports. *PloS one*, 11(4):e0153749.

[Nath et al., 2016b] Nath, J. B., Costigan, S., and Hsia, R. Y. (2016b). Changes in demographics of patients seen at federally qualified health centers, 2005-2014. *JAMA internal medicine*, 176(5):712–714.

[Ngo et al., 2021] Ngo, D., Benson, M. D., Long, J. Z., Chen, Z. Z., Wang, R., Nath, A. K., Keyes, M. J., Shen, D., Sinha, S., Kuhn, E., Morningstar, J. E., Shi, X., Peterson, B. D., Chan, C., Katz, D. H., Tahir, U. A., Farrell, L. A., Melander, O., Mosley, J. D., Carr, S. A., Vasan, R. S., Larson, M. G., Smith, J. G., Wang, T. J., Yang, Q., and Gerszten, R. E. (2021). Proteomic profiling reveals biomarkers and pathways in type 2 diabetes risk. *JCI Insight*, 6(5).

[Ngo et al., 2016] Ngo, D., Sinha, S., Shen, D., Kuhn, E. W., Keyes, M. J., Shi, X., Benson, M. D., O'Sullivan, J. F., Keshishian, H., Farrell, L. A., Fifer, M. A., Vasan, R. S., Sabatine, M. S., Larson, M. G., Carr, S. A., Wang, T. J., and Gerszten, R. E. (2016). Aptamer-Based proteomic profiling reveals novel candidate biomarkers and pathways in cardiovascular disease. *Circulation*, 134(4):270–285.

[Nissen et al., 2023] Nissen, S. E., Lincoff, A. M., Brennan, D., Ray, K. K., Mason, D., Kastelein, J. J. P., Thompson, P. D., Libby, P., Cho, L., Plutzky, J., Bays, H. E., Moriarty, P. M., Menon, V., Grobbee, D. E., Louie, M. J., Chen, C.-F., Li, N., Bloedon, L., Robinson, P., Horner, M., Sasiela, W. J., McCluskey, J., Davey, D., Fajardo-Campos, P., Petrovic, P., Fedacko, J., Zmuda, W., Lukyanov, Y., and Nicholls, S. J. (2023). Bempedoic acid and cardiovascular outcomes in Statin-Intolerant patients. *The New England journal of medicine*.

[OHDSI, 2021a] OHDSI (2021a). The book of OHDSI. `https://ohdsi.github.io/TheBookOfOhdsi/`. Accessed: 2022-7-26.

[OHDSI, 2021b] OHDSI (2021b). OMOP CDM 5.3.1.

[OHDSI, 2022a] OHDSI (2022a). ATHENA. `https://athena.ohdsi.org/search-terms/start`. Accessed: 2022-4-13.

[OHDSI, 2022b] OHDSI (2022b). OMOP CDM SQL scripts. `https://ohdsi.github.io/CommonDataModel/sqlScripts.html`. Accessed: 2022-4-13.

[Olson et al., 2015] Olson, N. C., Butenas, S., Lange, L. A., Lange, E. M., Cushman, M., Jenny, N. S., Walston, J., Souto, J. C., Soria, J. M., Chauhan, G., Debette, S., Longstreth, W. T., Seshadri, S., Reiner, A. P., and Tracy, R. P. (2015). Coagulation factor XII genetic variation, ex vivo thrombin generation, and stroke risk in the elderly: results from the cardiovascular health study. *Journal of thrombosis and haemostasis: JTH*, 13(10):1867–1877.

[Ong et al., 2007] Ong, K. L., Cheung, B. M. Y., Man, Y. B., Lau, C. P., and Lam, K. S. L. (2007). Prevalence, awareness, treatment, and control of hypertension among united states adults

1999-2004. *Hypertension*, 49(1):69–75.

[Pakhomov et al., 2008] Pakhomov, S., Jacobsen, S. J., Chute, C. G., and Roger, V. L. (2008). Agreement between patient-reported symptoms and their documentation in the medical record. *The American journal of managed care*, 14(8):530.

[Pakhomov et al., 2007] Pakhomov, S. S. V., Hemingway, H., Weston, S. A., Jacobsen, S. J., Rodeheffer, R., and Roger, V. L. (2007). Epidemiology of angina pectoris: role of natural language processing of the medical record. *American heart journal*, 153(4):666–673.

[Patel et al., 2019] Patel, N., Bhargava, A., Kalra, R., Parcha, V., Arora, G., and others (2019). Trends in lipid, lipoproteins, and statin use among US adults: impact of 2013 cholesterol guidelines. *Journal of the American*.

[Paul et al., 2019] Paul, S., Lancaster, G. I., and Meikle, P. J. (2019). Plasmalogens: A potential therapeutic target for neurodegenerative and cardiometabolic disease. *Progress in lipid research*, 74:186–195.

[Pearl, 2009] Pearl, J. (2009). *Causality*. Cambridge University Press.

[Pearl, 2022] Pearl, J. (2022). Direct and indirect effects. In *Probabilistic and Causal Inference: The Works of Judea Pearl*, volume 36, pages 373–392. Association for Computing Machinery, New York, NY, USA, 1 edition.

[Pedersen et al., 2005] Pedersen, T. R., Faergeman, O., Kastelein, J. J. P., Olsson, A. G., Tikkanen, M. J., Holme, I., Larsen, M. L., Bendiksen, F. S., Lindahl, C., Szarek, M., Tsai, J., and Incremental Decrease in End Points Through Aggressive Lipid Lowering (IDEAL) Study Group (2005). High-dose atorvastatin vs usual-dose simvastatin for secondary prevention after myocardial infarction: the IDEAL study: a randomized controlled trial. *JAMA: the journal of the American Medical Association*, 294(19):2437–2445.

[Pepe and Thompson, 2000] Pepe, M. S. and Thompson, M. L. (2000). Combining diagnostic test results to increase accuracy. *Biostatistics*, 1(2):123–140.

[Phillips, 2014] Phillips, M. C. (2014). Apolipoprotein E isoforms and lipoprotein metabolism. *IUBMB life*, 66(9):616–623.

[Picard et al., 2011] Picard, M. H., Adams, D., Bierig, S. M., Dent, J. M., Douglas, P. S., Gillam, L. D., Keller, A. M., Malenka, D. J., Masoudi, F. A., McCulloch, M., Pellikka, P. A., Peters, P. J., Stainback, R. F., Strachan, G. M., Zoghbi, W. A., and American Society of Echocardiography (2011). American society of echocardiography recommendations for quality echocardiography laboratory operations. *Journal of the American Society of Echocardiography: official publication of the American Society of Echocardiography*, 24(1):1–10.

[Piedrahita et al., 1992] Piedrahita, J. A., Zhang, S. H., Hagaman, J. R., Oliver, P. M., and Maeda, N. (1992). Generation of mice carrying a mutant apolipoprotein E gene inactivated by gene targeting in embryonic stem cells. *Proceedings of the National Academy of Sciences of the United States of America*, 89(10):4471–4475.

[Pietzner et al., 2021] Pietzner, M., Wheeler, E., Carrasco-Zanini, J., Cortes, A., Koprulu, M., Wörheide, M. A., Oerton, E., Cook, J., Stewart, I. D., Kerrison, N. D., Luan, J., Raffler, J., Arnold, M., Arlt, W., O'Rahilly, S., Kastenmüller, G., Gamazon, E. R., Hingorani, A. D., Scott, R. A., Wareham, N. J., and Langenberg, C. (2021). Mapping the proteo-genomic convergence of human diseases. *Science*, 374(6569):eabj1541.

[Pietzner et al., 2020] Pietzner, M., Wheeler, E., Carrasco-Zanini, J., Raffler, J., Kerrison, N. D., Oerton, E., Auyeung, V. P. W., Luan, J., Finan, C., Casas, J. P., Ostroff, R., Williams, S. A., Kastenmüller, G., Ralser, M., Gamazon, E. R., Wareham, N. J., Hingorani, A. D., and Langenberg, C. (2020). Genetic architecture of host proteins involved in SARS-CoV-2 infection. *Nature communications*, 11(1):6397.

[Plump et al., 1992] Plump, A. S., Smith, J. D., Hayek, T., Aalto-Setälä, K., Walsh, A., Verstuyft, J. G., Rubin, E. M., and Breslow, J. L. (1992). Severe hypercholesterolemia and atherosclerosis in apolipoprotein e-deficient mice created by homologous recombination in ES cells. *Cell*, 71(2):343–353.

[Png et al., 2021] Png, G., Barysenka, A., Repetto, L., Navarro, P., Shen, X., Pietzner, M., Wheeler, E., Wareham, N. J., Langenberg, C., Tsafantakis, E., Karaleftheri, M., Dedoussis, G., Mälarstig, A., Wilson, J. F., Gilly, A., and Zeggini, E. (2021). Mapping the serum proteome to neurological diseases using whole genome sequencing. *Nature communications*, 12(1):7042.

[Podrez et al., 2002] Podrez, E. A., Poliakov, E., Shen, Z., Zhang, R., Deng, Y., Sun, M., Finton, P. J., Shan, L., Febbraio, M., Hajjar, D. P., Silverstein, R. L., Hoff, H. F., Salomon, R. G., and Hazen, S. L. (2002). A novel family of atherogenic oxidized phospholipids promotes macrophage foam cell formation via the scavenger receptor CD36 and is enriched in atherosclerotic lesions. *The Journal of biological chemistry*, 277(41):38517–38523.

[Pryor et al., 1983] Pryor, T. A., Gardner, R. M., Clayton, P. D., and Warner, H. R. (1983). The HELP system. *Journal of medical systems*, 7(2):87–102.

[Qureshi et al., 2016] Qureshi, W. T., Michos, E. D., Flueckiger, P., Blaha, M., Sandfort, V., Herrington, D. M., Burke, G., and Yeboah, J. (2016). Impact of replacing the pooled cohort equation with other cardiovascular disease risk scores on atherosclerotic cardiovascular disease risk assessment (from the Multi-Ethnic study of atherosclerosis [MESA]). *The American journal of cardiology*, 118(5):691–696.

[Rabiner, 1989] Rabiner, L. R. (1989). A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286.

[Rachmilewitz and Braun, 1945] Rachmilewitz, M. and Braun, K. (1945). Electrocardiographic changes and the effect of niacin therapy in pellagra. *British heart journal*, 7(2):72–80.

[Raffield et al., 2020] Raffield, L. M., Dang, H., Pratte, K. A., Jacobson, S., Gillenwater, L. A., Ampleford, E., Barjaktarevic, I., Basta, P., Clish, C. B., Comellas, A. P., Cornell, E., Curtis, J. L., Doerschuk, C., Durda, P., Emson, C., Freeman, C. M., Guo, X., Hastie, A. T., Hawkins, G. A., Herrera, J., Johnson, W. C., Labaki, W. W., Liu, Y., Masters, B., Miller, M., Ortega, V. E., Papanicolaou, G., Peters, S., Taylor, K. D., Rich, S. S., Rotter, J. I., Auer, P., Reiner, A. P., Tracy, R. P., Ngo, D., Gerszten, R. E., O'Neal, W. K., Bowler, R. P., and Consortium, N.

T.-O. F. P. M. t. (2020). Comparison of proteomic assessment methods in multiple cohort studies. *Proteomics*, 20(12):e1900278.

[Raffield et al., 2017] Raffield, L. M., Zakai, N. A., Duan, Q., Laurie, C., Smith, J. D., Irvin, M. R., Doyle, M. F., Naik, R. P., Song, C., Manichaikul, A. W., Liu, Y., Durda, P., Rotter, J. I., Jenny, N. S., Rich, S. S., Wilson, J. G., Johnson, A. D., Correa, A., Li, Y., Nickerson, D. A., Rice, K., Lange, E. M., Cushman, M., Lange, L. A., Reiner, A. P., and NHLBI Trans-Omics for Precision Medicine (TOPMed) Consortium, H. M. . H. T. W. G. (2017). D-Dimer in african americans: Whole genome sequence analysis and relationship to cardiovascular disease risk in the jackson heart study. *Arteriosclerosis, thrombosis, and vascular biology*, 37(11):2220–2227.

[Ramirez et al., 2022] Ramirez, A. H., Sulieman, L., Schlueter, D. J., Halvorson, A., Qian, J., Ratsimbazafy, F., Loperena, R., Mayo, K., Basford, M., Deflaux, N., Muthuraman, K. N., Natarajan, K., Kho, A., Xu, H., Wilkins, C., Anton-Culver, H., Boerwinkle, E., Cicek, M., Clark, C. R., Cohn, E., Ohno-Machado, L., Schully, S. D., Ahmedani, B. K., Argos, M., Cronin, R. M., O'Donnell, C., Fouad, M., Goldstein, D. B., Greenland, P., Hebbring, S. J., Karlson, E. W., Khatri, P., Korf, B., Smoller, J. W., Sodeke, S., Wilbanks, J., Hentges, J., Mockrin, S., Lunt, C., Devaney, S. A., Gebo, K., Denny, J. C., Carroll, R. J., Glazer, D., Harris, P. A., Hripcsak, G., Philippakis, A., Roden, D. M., and All of Us Research Program (2022). The all of us research program: Data quality, utility, and diversity. *Patterns (New York, N.Y.)*, 3(8):100570.

[Rashid et al., 2022] Rashid, S., Suero-Abreu, G. A., Tysarowski, M., and others (2022). Increasing statin prescription rates to prevent cardiovascular disease among high-risk populations: a quality improvement intervention centred on a novel interactive .... *BMJ open*.

[Rasmiena et al., 2015] Rasmiena, A. A., Barlow, C. K., Stefanovic, N., Huynh, K., Tan, R., Sharma, A., Tull, D., de Haan, J. B., and Meikle, P. J. (2015). Plasmalogen modulation attenuates atherosclerosis in ApoE- and ApoE/GPx1-deficient mice. *Atherosclerosis*, 243(2):598–608.

[Redmond et al., 2011] Redmond, N., Baer, H. J., and Hicks, L. S. (2011). Health behaviors and racial disparity in blood pressure control in the national health and nutrition examination survey. *Hypertension*, 57(3):383–389.

[Reynolds et al., 2021] Reynolds, H. R., Shaw, L. J., Min, J. K., Page, C. B., Berman, D. S., Chaitman, B. R., Picard, M. H., Kwong, R. Y., O'Brien, S. M., Huang, Z., Mark, D. B., Nath, R. K., Dwivedi, S. K., Smanio, P. E. P., Stone, P. H., Held, C., Keltai, M., Bangalore, S., Newman, J. D., Spertus, J. A., Stone, G. W., Maron, D. J., and Hochman, J. S. (2021). Outcomes in the ISCHEMIA trial based on coronary artery disease and ischemia severity. *Circulation*, 144(13):1024–1038.

[Rhee et al., 2013] Rhee, E. P., Ho, J. E., Chen, M. H., Shen, D., Cheng, S., Larson, M. G., Ghorbani, A., Shi, X., Helenius, I. T., O'Donnell, C. J., Souza, A. L., Deik, A., Pierce, K. A., Bullock, K., Walford, G. A., Vasan, R. S., Florez, J. C., Clish, C., Yeh, J. R., Wang, T. J., and Gerszten, R. E. (2013). A genome-wide association study of the human metabolome in a community-based cohort. *Cell metabolism*, 18(1):130–143.

[Rhee et al., 2016] Rhee, E. P., Yang, Q., Yu, B., Liu, X., Cheng, S., Deik, A., Pierce, K. A., Bullock, K., Ho, J. E., Levy, D., Florez, J. C., Kathiresan, S., Larson, M. G., Vasan, R. S., Clish, C. B., Wang, T. J., Boerwinkle, E., O'Donnell, C. J., and Gerszten, R. E. (2016). An exome array study of the plasma metabolome. *Nature communications*, 7:12360.

[Ritchey et al., 2018] Ritchey, M. D., Wall, H. K., Owens, P. L., and Wright, J. S. (2018). Vital signs: State-Level variation in nonfatal and fatal cardiovascular events targeted for prevention by million hearts 2022. *MMWR. Morbidity and mortality weekly report*, 67(35):974–982.

[Robbins et al., 2019] Robbins, J. M., Herzig, M., Morningstar, J., Sarzynski, M. A., Cruz, D. E., Wang, T. J., Gao, Y., Wilson, J. G., Bouchard, C., Rankinen, T., and Gerszten, R. E. (2019). Association of dimethylguanidino valeric acid with partial resistance to metabolic health benefits of regular exercise. *JAMA Cardiol*, 4(7):636–643.

[Robbins et al., 2021] Robbins, J. M., Peterson, B., Schranner, D., Tahir, U. A., Rienmüller, T., Deng, S., Keyes, M. J., Katz, D. H., Beltran, P. M. J., Barber, J. L., Baumgartner, C., Carr, S. A., Ghosh, S., Shen, C., Jennings, L. L., Ross, R., Sarzynski, M. A., Bouchard, C., and Gerszten, R. E. (2021). Human plasma proteomic profiles indicative of cardiorespiratory fitness. *Nat Metab*, 3(6):786–797.

[Romani et al., 2018] Romani, P., Ignesti, M., Gargiulo, G., Hsu, T., and Cavaliere, V. (2018). Extracellular NME proteins: a player or a bystander? *Laboratory investigation; a journal of technical methods and pathology*, 98(2):248–257.

[Rose, 1964] Rose, G. (1964). Familial patterns in ischaemic heart disease. *British journal of preventive & social medicine*, 18:75–80.

[Rutstein et al., 1952] Rutstein, D. D., Nickerson, R. J., and Heald, F. P. (1952). Seasonal incidence of patent ductus arteriosus and maternal rubella. *A.M.A. American journal of diseases of children*, 84(2):199–213.

[Salami et al., 2017] Salami, J. A., Warraich, H., Valero-Elizondo, J., Spatz, E. S., Desai, N. R., Rana, J. S., Virani, S. S., Blankstein, R., Khera, A., Blaha, M. J., Blumenthal, R. S., Lloyd-Jones, D., and Nasir, K. (2017). National trends in statin use and expenditures in the US adult population from 2002 to 2013: Insights from the medical expenditure panel survey. *JAMA cardiology*, 2(1):56–65.

[Sankar and Parker, 2017] Sankar, P. L. and Parker, L. S. (2017). The precision medicine initiative's all of us research program: an agenda for research on its ethical, legal, and social issues. *Genetics in medicine: official journal of the American College of Medical Genetics*, 19(7):743–750.

[Sarwar et al., 2012] Sarwar, N., Butterworth, A. S., Freitag, D. F., Gregson, J., Willeit, P., Gorman, D. N., Gao, P., Saleheen, D., Rendon, A., Nelson, C. P., Braund, P. S., Hall, A. S., Chasman, D. I., Tybjærg-Hansen, A., Chambers, J. C., Benjamin, E. J., Franks, P. W., Clarke, R., Wilde, A. A., Trip, M. D., Steri, M., Witteman, J. C., Qi, L., van der Schoot, C. E., de Faire, U., Erdmann, J., Stringham, H. M., Koenig, W., Rader, D. J., Melzer, D., Reich, D., Psaty, B. M., Kleber, M. E., Panagiotakos, D. B., Willeit, J., Wennberg, P., Woodward, M., Adamovic, S., Rimm, E. B., Meade, T. W., Gillum, R. F., Shaffer, J. A., Hofman, A., Onat, A., Sundström, J., Wassertheil-Smoller, S., Mellström, D., Gallacher, J., Cushman, M., Tracy, R. P., Kauhanen, J., Karlsson, M., Salonen, J. T., Wilhelmsen, L., Amouyel, P., Cantin, B., Best, L. G., Ben-Shlomo, Y., Manson, J. E., Davey-Smith, G., de Bakker, P. I., O'Donnell, C. J., Wilson, J. F., Wilson, A. G., Assimes, T. L., Jansson, J. O., Ohlsson, C., Tivesten, Å., Ljunggren, Ö., Reilly, M. P., Hamsten, A., Ingelsson, E., Cambien, F., Hung, J., Thomas, G. N., Boehnke, M., Schunkert, H., Asselbergs, F. W., Kastelein, J. J., Gudnason, V., Salomaa, V., Harris, T. B., Kooner, J. S., Allin, K. H., Nordestgaard, B. G., Hopewell, J. C., Goodall, A. H., Ridker, P. M., Hólm, H., Watkins,

H., Ouwehand, W. H., Samani, N. J., Kaptoge, S., Di Angelantonio, E., Harari, O., Danesh, J., and Collaboration, I. G. C. E. R. F. (2012). Interleukin-6 receptor pathways in coronary heart disease: a collaborative meta-analysis of 82 studies. *The Lancet*, 379(9822):1205–1213.

[Sass et al., 2006] Sass, J. O., Mohr, V., Olbrich, H., Engelke, U., Horvath, J., Fliegauf, M., Loges, N. T., Schweitzer-Krantz, S., Moebus, R., Weiler, P., Kispert, A., Superti-Furga, A., Wevers, R. A., and Omran, H. (2006). Mutations in ACY1, the gene encoding aminoacylase 1, cause a novel inborn error of metabolism. *American journal of human genetics*, 78(3):401–409.

[Scandinavian Simvastatin Survival Study Group, 1994] Scandinavian Simvastatin Survival Study Group (1994). Randomised trial of cholesterol lowering in 4444 patients with coronary heart disease: the scandinavian simvastatin survival study (4s). *The Lancet*, 344(8934):1383–1389.

[Schildkraut et al., 1989] Schildkraut, J. M., Myers, R. H., Cupples, L. A., Kiely, D. K., and Kannel, W. B. (1989). Coronary risk associated with age and sex of parental heart disease in the framingham study. *The American journal of cardiology*, 64(10):555–559.

[Schiros et al., 2015] Schiros, C. G., Denney, Jr, T. S., and Gupta, H. (2015). Interaction analysis of the new pooled cohort equations for 10-year atherosclerotic cardiovascular disease risk estimation: a simulation analysis. *BMJ open*, 5(4):e006468.

[Schisterman and Platt, 2011] Schisterman, E. F. and Platt, R. W. (2011). Causal inference in perinatal epidemiology: Revisiting the birth weight paradox. *Reproductive and Perinatal Epidemiology*, page 276.

[Schwarzer et al., 2015] Schwarzer, G., Carpenter, J. R., Rücker, G., and SpringerLink (2015). *Meta-Analysis with R*. Use R! Springer International Publishing : Imprint: Springer, Cham, 1st 2015. edition.

[Sempos et al., 1999] Sempos, C. T., Bild, D. E., and Manolio, T. A. (1999). Overview of the jackson heart study: a study of cardiovascular diseases in african american men and women. *The American journal of the medical sciences*, 317(3):142–146.

[Shah and Tenenbaum, 2012] Shah, N. H. and Tenenbaum, J. D. (2012). The coming age of data-driven medicine: translational bioinformatics' next frontier. *Journal of the American Medical Informatics Association: JAMIA*, 19(e1):e2–4.

[Shepherd et al., 1995] Shepherd, J., Cobbe, S. M., Ford, I., Isles, C. G., Lorimer, A. R., Macfarlane, P. W., McKillop, J. H., and Packard, C. J. (1995). Prevention of coronary heart disease with pravastatin in men with hypercholesterolemia. *The New England journal of medicine*, 333(20):1301–1308.

[Shin et al., 2014] Shin, S. Y., Fauman, E. B., Petersen, A. K., Krumsiek, J., Santos, R., Huang, J., Arnold, M., Erte, I., Forgetta, V., Yang, T. P., Walter, K., Menni, C., Chen, L., Vasquez, L., Valdes, A. M., Hyde, C. L., Wang, V., Ziemek, D., Roberts, P., Xi, L., Grundberg, E., Waldenberger, M., Richards, J. B., Mohney, R. P., Milburn, M. V., John, S. L., Trimmer, J., Theis, F. J., Overington, J. P., Suhre, K., Brosnan, M. J., Gieger, C., Kastenmüller, G., Spector, T. D., Soranzo, N., and Consortium, M. T. H. E. R. m. (2014). An atlas of genetic influences on human blood metabolites. *Nature genetics*, 46(6):543–550.

[Shivade et al., 2015] Shivade, C., Malewadkar, P., Fosler-Lussier, E., and Lai, A. M. (2015). Comparison of UMLS terminologies to identify risk of heart disease using clinical notes. *Journal of biomedical informatics*, 58 Suppl:S103–10.

[Simpson, 1951] Simpson, E. H. (1951). The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society*, 13(2):238–241.

[Singh and Ward, 2016] Singh, A. and Ward, R. P. (2016). Appropriate use criteria for echocardiography: Evolving applications in the era of Value-Based healthcare. *Current cardiology reports*, 18(9):93.

[Smith and Ebrahim, 2003] Smith, G. D. and Ebrahim, S. (2003). 'mendelian randomization': can genetic epidemiology contribute to understanding environmental determinants of disease? *International journal of epidemiology*, 32(1):1–22.

[Snow, 1849] Snow, J. (1849). *On the Mode of Communication of Cholera*. John Churchill.

[Snow, 1856] Snow, J. (1856). Cholera and the water supply in the south districts of london in 1854. *Journal of public health, and sanitary review*, 2(7):239–257.

[Sohn et al., 2018] Sohn, S., Wang, Y., Wi, C.-I., Krusemark, E. A., Ryu, E., Ali, M. H., Juhn, Y. J., and Liu, H. (2018). Clinical documentation variations and NLP system portability: a case study in asthma birth cohorts across institutions. *Journal of the American Medical Informatics Association: JAMIA*, 25(3):353–359.

[Solomon et al., 2016] Solomon, T., Smith, E. N., Matsui, H., Braekkan, S. K., Wilsgaard, T., Njølstad, I., Mathiesen, E. B., Hansen, J. B., Frazer, K. A., and Consortium, I. (2016). Associations between common and rare exonic genetic variants and serum levels of 20 Cardiovascular-Related proteins: The tromsø study. *Circulation. Cardiovascular genetics*, 9(4):375–383.

[Sommer et al., 2011] Sommer, A., Christensen, E., Schwenger, S., Seul, R., Haas, D., Olbrich, H., Omran, H., and Sass, J. O. (2011). The molecular basis of aminoacylase 1 deficiency. *Biochimica et biophysica acta*, 1812(6):685–690.

[Splansky et al., 2007] Splansky, G. L., Corey, D., Yang, Q., Atwood, L. D., Cupples, L. A., Benjamin, E. J., D'Agostino, Sr, R. B., Fox, C. S., Larson, M. G., Murabito, J. M., O'Donnell, C. J., Vasan, R. S., Wolf, P. A., and Levy, D. (2007). The third generation cohort of the national heart, lung, and blood institute's framingham heart study: design, recruitment, and initial examination. *American journal of epidemiology*, 165(11):1328–1335.

[Spranger et al., 2004] Spranger, C. B., Ries, A. J., Berge, C. A., Radford, N. B., and Victor, R. G. (2004). Identifying gaps between guidelines and clinical practice in the evaluation and treatment of patients with hypertension. *The American journal of medicine*, 117(1):14–18.

[Stone and Lloyd-Jones, 2014] Stone, N. J. and Lloyd-Jones, D. M. (2014). Understanding the new atherosclerotic cardiovascular disease prevention guidelines: going beyond the headlines. *Clinical lipidology*, 9(2):121–123.

[Stone et al., 2014] Stone, N. J., Robinson, J. G., Lichtenstein, A. H., Merz, C. N. B., Blum, C. B.,

Eckel, R. H., Goldberg, A. C., Gordon, D., Levy, D., Lloyd-Jones, D. M., and Others (2014). 2013 ACC/AHA guideline on the treatment of blood cholesterol to reduce atherosclerotic cardiovascular risk in adults. *Circulation*, 129(25):S1–S45.

[Strittmatter et al., 1993] Strittmatter, W. J., Saunders, A. M., Schmechel, D., Pericak-Vance, M., Enghild, J., Salvesen, G. S., and Roses, A. D. (1993). Apolipoprotein e: high-avidity binding to beta-amyloid and increased frequency of type 4 allele in late-onset familial alzheimer disease. *Proceedings of the National Academy of Sciences of the United States of America*, 90(5):1977–1981.

[Stubbs et al., 2015] Stubbs, A., Kotfila, C., Xu, H., and Uzuner, Ö. (2015). Identifying risk factors for heart disease over time: Overview of 2014 i2b2/UTHealth shared task track 2. *Journal of biomedical informatics*, 58 Suppl:S67–77.

[Subramanian et al., 2005] Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S., and Mesirov, J. P. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America*, 102(43):15545–15550.

[Suhre et al., 2017] Suhre, K., Arnold, M., Bhagwat, A. M., Cotton, R. J., Engelke, R., Raffler, J., Sarwath, H., Thareja, G., Wahl, A., DeLisle, R. K., Gold, L., Pezer, M., Lauc, G., El-Din Selim, M. A., Mook-Kanamori, D. O., Al-Dous, E. K., Mohamoud, Y. A., Malek, J., Strauch, K., Grallert, H., Peters, A., Kastenmüller, G., Gieger, C., and Graumann, J. (2017). Connecting genetic risk to disease end points through the human blood plasma proteome. *Nature communications*, 8:14357.

[Suhre et al., 2011] Suhre, K., Shin, S. Y., Petersen, A. K., Mohney, R. P., Meredith, D., Wägele, B., Altmaier, E., Deloukas, P., Erdmann, J., Grundberg, E., Hammond, C. J., de Angelis, M. H., Kastenmüller, G., Köttgen, A., Kronenberg, F., Mangino, M., Meisinger, C., Meitinger, T., Mewes, H. W., Milburn, M. V., Prehn, C., Raffler, J., Ried, J. S., Römisch-Margl, W., Samani, N. J., Small, K. S., Wichmann, H. E., Zhai, G., Illig, T., Spector, T. D., Adamski, J., Soranzo, N., Gieger, C., and CARDIoGRAM (2011). Human metabolic individuality in biomedical and pharmaceutical research. *Nature*, 477(7362):54–60.

[Sun et al., 2018] Sun, B. B., Maranville, J. C., Peters, J. E., Stacey, D., Staley, J. R., Blackshaw, J., Burgess, S., Jiang, T., Paige, E., Surendran, P., Oliver-Williams, C., Kamat, M. A., Prins, B. P., Wilcox, S. K., Zimmerman, E. S., Chi, A., Bansal, N., Spain, S. L., Wood, A. M., Morrell, N. W., Bradley, J. R., Janjic, N., Roberts, D. J., Ouwehand, W. H., Todd, J. A., Soranzo, N., Suhre, K., Paul, D. S., Fox, C. S., Plenge, R. M., Danesh, J., Runz, H., and Butterworth, A. S. (2018). Genomic atlas of the human plasma proteome. *Nature*, 558(7708):73–79.

[Swerdlow et al., 2016] Swerdlow, D. I., Kuchenbaecker, K. B., Shah, S., Sofat, R., Holmes, M. V., White, J., Mindell, J. S., Kivimaki, M., Brunner, E. J., Whittaker, J. C., Casas, J. P., and Hingorani, A. D. (2016). Selecting instruments for mendelian randomization in the wake of genome-wide association studies. *International journal of epidemiology*, 45(5):1600–1616.

[TADA2A-Gene-NCBI, 2023] TADA2A-Gene-NCBI (2023). TADA2A transcriptional adaptor 2A [homo sapiens (human)] - gene - NCBI. `https://www.ncbi.nlm.nih.gov/gene/6871`. Accessed: 2023-2-14.

[Tahir et al., 2022] Tahir, U. A., Katz, D. H., Avila-Pachecho, J., Bick, A. G., Pampana, A., Robbins, J. M., Yu, Z., Chen, Z. Z., Benson, M. D., Cruz, D. E., Ngo, D., Deng, S., Shi, X., Zheng, S., Eisman, A. S., Farrell, L., Hall, M. E., Correa, A., Tracy, R. P., Durda, P., Taylor, K. D., Liu, Y., Johnson, W. C., Guo, X., Yao, J., Chen, Y. I., Manichaikul, A. W., Ruberg, F. L., Blaner, W. S., Jain, D., Bouchard, C., Sarzynski, M. A., Rich, S. S., Rotter, J. I., Wang, T. J., Wilson, J. G., Clish, C. B., Natarajan, P., Gerszten, R. E., and Consortium, N. T.-O. F. P. M. . (2022). Whole genome association study of the plasma metabolome identifies metabolites linked to cardiometabolic disease in black individuals. *Nature communications*, 13(1):4923.

[Tahir et al., 2021] Tahir, U. A., Katz, D. H., Zhao, T., Ngo, D., Cruz, D. E., Robbins, J. M., Chen, Z. Z., Peterson, B., Benson, M. D., Shi, X., Dailey, L., Andersson, C., Vasan, R. S., Gao, Y., Shen, C., Correa, A., Hall, M. E., Wang, T. J., Clish, C. B., Wilson, J. G., and Gerszten, R. E. (2021). Metabolomic profiles and heart failure risk in black adults: Insights from the jackson heart study. *Circulation. Heart failure*, 14(1):e007275.

[Taylor et al., 2013] Taylor, F., Huffman, M. D., Macedo, A. F., Moore, T. H. M., Burke, M., Davey Smith, G., Ward, K., and Ebrahim, S. (2013). Statins for the primary prevention of cardiovascular disease. *Cochrane database of systematic reviews*, (1):CD004816.

[Team and Others, 2018] Team, R. C. and Others (2018). R: A language and environment for statistical computing; 2018.

[Tenenbaum, 2016] Tenenbaum, J. D. (2016). Translational bioinformatics: Past, present, and future. *Genomics, proteomics & bioinformatics*, 14(1):31–41.

[The ARIC investigators, 1989] The ARIC investigators (1989). The atherosclerosis risk in communities (ARIC) study: design and objectives. *American journal of epidemiology*, 129(4):687–702.

[Thibeaux et al., 2018] Thibeaux, S., Siddiqi, S., Zhelyabovska, O., Moinuddin, F., Masternak, M. M., and Siddiqi, S. A. (2018). Cathepsin B regulates hepatic lipid metabolism by cleaving liver fatty acid-binding protein. *The Journal of biological chemistry*, 293(6):1910–1923.

[Toepfer et al., 2015] Toepfer, M., Corovic, H., Fette, G., Klügl, P., Störk, S., and Puppe, F. (2015). Fine-grained information extraction from german transthoracic echocardiography reports. *BMC medical informatics and decision making*, 15:91.

[Topel et al., 2018] Topel, M. L., Shen, J., Morris, A. A., Al Mheid, I., Sher, S., Dunbar, S. B., Vaccarino, V., Sperling, L. S., Gibbons, G. H., Martin, G. S., and Quyyumi, A. A. (2018). Comparisons of the framingham and pooled cohort equation risk scores for detecting subclinical vascular disease in blacks versus whites. *The American journal of cardiology*, 121(5):564–569.

[Van Coster et al., 2005] Van Coster, R. N., Gerlo, E. A., Giardina, T. G., Engelke, U. F., Smet, J. E., De Praeter, C. M., Meersschaut, V. A., De Meirleir, L. J., Seneca, S. H., Devreese, B., Leroy, J. G., Herga, S., Perrier, J. P., Wevers, R. A., and Lissens, W. (2005). Aminoacylase I deficiency: a novel inborn error of metabolism. *Biochemical and biophysical research communications*, 338(3):1322–1326.

[Vasan and van den Heuvel, 2022] Vasan, R. S. and van den Heuvel, E. (2022). Differences in estimates for 10-year risk of cardiovascular disease in black versus white individuals with identical

risk factor profiles using pooled cohort equations: an in silico cohort study. *The Lancet. Digital health*, 4(1):e55–e63.

[Vaswani et al., 2017] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł. U., and Polosukhin, I. (2017). Attention is all you need. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.

[Velupillai et al., 2015] Velupillai, S., Mowery, D. L., Abdelrahman, S., Christensen, L., and Chapman, W. W. (2015). Towards a generalizable time expression model for temporal reasoning in clinical notes. *AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium*, 2015:1252–1259.

[Vijayakrishnan et al., 2014] Vijayakrishnan, R., Steinhubl, S. R., Ng, K., Sun, J., Byrd, R. J., Daar, Z., Williams, B. A., deFilippi, C., Ebadollahi, S., and Stewart, W. F. (2014). Prevalence of heart failure signs and symptoms in a large primary care population identified through the use of text and data mining of the electronic health record. *Journal of cardiac failure*, 20(7):459–464.

[Ward, 2012] Ward, R. P. (2012). Appropriate use criteria for echocardiography: new applications for a new era of utilization. *Journal of the American Society of Echocardiography: official publication of the American Society of Echocardiography*, 25(6):599–602.

[Wensley et al., 2011] Wensley, F., Gao, P., Burgess, S., Kaptoge, S., Di Angelantonio, E., Shah, T., Engert, J. C., Clarke, R., Davey-Smith, G., Nordestgaard, B. G., Saleheen, D., Samani, N. J., Sandhu, M., Anand, S., Pepys, M. B., Smeeth, L., Whittaker, J., Casas, J. P., Thompson, S. G., Hingorani, A. D., Danesh, J., and (ccgc), C. R. P. C. H. D. G. C. (2011). Association between C reactive protein and coronary heart disease: mendelian randomisation analysis based on individual participant data. *BMJ*, 342:d548.

[Whelton et al., 2018] Whelton, P. K., Carey, R. M., Aronow, W. S., Casey, Jr, D. E., Collins, K. J., Dennison Himmelfarb, C., DePalma, S. M., Gidding, S., Jamerson, K. A., Jones, D. W., MacLaughlin, E. J., Muntner, P., Ovbiagele, B., Smith, Jr, S. C., Spencer, C. C., Stafford, R. S., Taler, S. J., Thomas, R. J., Williams, Sr, K. A., Williamson, J. D., and Wright, Jr, J. T. (2018). 2017 ACC/AHA/AAPA/ABC/ACPM/AGS/APhA/ASH/ASPC/NMA/PCNA guideline for the prevention, detection, evaluation, and management of high blood pressure in adults. *Journal of the American College of Cardiology*, 71(19):e127–e248.

[Wilcox, 2001] Wilcox, A. J. (2001). On the importance—and the unimportance—of birthweight. *International journal of epidemiology*, 30(6):1233–1241.

[Wilson et al., 1998] Wilson, P. W., D'Agostino, R. B., Levy, D., Belanger, A. M., Silbershatz, H., and Kannel, W. B. (1998). Prediction of coronary heart disease using risk factor categories. *Circulation*, 97(18):1837–1847.

[Wilson et al., 2004] Wilson, P. W. F., CDC, and AHA (2004). CDC/AHA workshop on markers of inflammation and cardiovascular disease: Application to clinical and public health practice: ability of inflammatory markers to predict disease in asymptomatic patients: a background paper. *Circulation*, 110(25):e568–71.

[Wolf et al., 2019] Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., and Others (2019). Huggingface's transformers: State-of-the-art natural language processing. *ArXiv, abs/1910. 03771*.

[Wolk et al., 2014] Wolk, M. J., Bailey, S. R., Doherty, J. U., Douglas, P. S., Hendel, R. C., Kramer, C. M., Min, J. K., Patel, M. R., Rosenbaum, L., Shaw, L. J., Stainback, R. F., Allen, J. M., and American College of Cardiology Foundation Appropriate Use Criteria Task Force (2014). ACCF/AHA/ASE/ASNC/HFSA/HRS/SCAI/SCCT/SCMR/STS 2013 multimodality appropriate use criteria for the detection and risk assessment of stable ischemic heart disease: a report of the american college of cardiology foundation appropriate use criteria task force, american heart association, american society of echocardiography, american society of nuclear cardiology, heart failure society of america, heart rhythm society, society for cardiovascular angiography and interventions, society of cardiovascular computed tomography, society for cardiovascular magnetic resonance, and society of thoracic surgeons. *Journal of the American College of Cardiology*, 63(4):380–406.

[World Health Organization, 2018] World Health Organization (2018). *World Health Statistics 2018: Monitoring Health for the SDGs Sustainable Development Goals*. World Health Organization.

[Wright et al., 2018] Wright, J. S., Wall, H. K., and Ritchey, M. D. (2018). Million hearts 2022: Small steps are needed for cardiovascular disease prevention. *JAMA: the journal of the American Medical Association*, 320(18):1857–1858.

[Wu et al., 2014] Wu, S., Miller, T., Masanz, J., Coarr, M., Halgrim, S., Carrell, D., and Clark, C. (2014). Negation's not solved: generalizability versus optimizability in clinical natural language processing. *PloS one*, 9(11):e112774.

[Yanai et al., 2000] Yanai, H., Chiba, H., Morimoto, M., Abe, K., Fujiwara, H., Fuda, H., Hui, S. P., Takahashi, Y., Akita, H., Jamieson, G. A., Kobayashi, K., and Matsuno, K. (2000). Human CD36 deficiency is associated with elevation in low-density lipoprotein-cholesterol. *American journal of medical genetics*, 93(4):299–304.

[Yang et al., 2017] Yang, Q., Zhong, Y., Gillespie, C., Merritt, R., Bowman, B., George, M. G., and Flanders, W. D. (2017). Assessing potential population impact of statin treatment for primary prevention of atherosclerotic cardiovascular diseases in the USA: population-based modelling study. *BMJ open*, 7(1):e011684.

[Yao et al., 2018] Yao, C., Chen, G., Song, C., Keefe, J., Mendelson, M., Huan, T., Sun, B. B., Laser, A., Maranville, J. C., Wu, H., Ho, J. E., Courchesne, P., Lyass, A., Larson, M. G., Gieger, C., Graumann, J., Johnson, A. D., Danesh, J., Runz, H., Hwang, S. J., Liu, C., Butterworth, A. S., Suhre, K., and Levy, D. (2018). Genome-wide mapping of plasma protein QTLs identifies putatively causal genes and pathways for cardiovascular disease. *Nature communications*, 9(1):3268.

[Yavorska and Burgess, 2017] Yavorska, O. O. and Burgess, S. (2017). MendelianRandomization: an R package for performing mendelian randomization analyses using summarized data. *International journal of epidemiology*, 46(6):1734–1739.

[Ye et al., 2016] Ye, H., Adane, B., Khan, N., Sullivan, T., Minhajuddin, M., Gasparetto, M., Stevens, B., Pei, S., Balys, M., Ashton, J. M., Klemm, D. J., Woolthuis, C. M., Stranahan, A. W., Park, C. Y., and Jordan, C. T. (2016). Leukemic stem cells evade chemotherapy by metabolic adaptation to an adipose tissue niche. *Cell stem cell*, 19(1):23–37.

[Yegutkin et al., 2002] Yegutkin, G. G., Henttinen, T., Samburski, S. S., Spychala, J., and Jalkanen, S. (2002). The evidence for two opposite, ATP-generating and ATP-consuming, extracellular pathways on endothelial and lymphoid cells. *Biochemical Journal*, 367(Pt 1):121–128.

[Yin et al., 2022] Yin, X., Chan, L. S., Bose, D., Jackson, A. U., VandeHaar, P., Locke, A. E., Fuchsberger, C., Stringham, H. M., Welch, R., Yu, K., Fernandes Silva, L., Service, S. K., Zhang, D., Hector, E. C., Young, E., Ganel, L., Das, I., Abel, H., Erdos, M. R., Bonnycastle, L. L., Kuusisto, J., Stitziel, N. O., Hall, I. M., Wagner, G. R., Kang, J., Morrison, J., Burant, C. F., Collins, F. S., Ripatti, S., Palotie, A., Freimer, N. B., Mohlke, K. L., Scott, L. J., Wen, X., Fauman, E. B., Laakso, M., Boehnke, M., and FinnGen (2022). Genome-wide association studies of metabolites in finnish men identify disease-relevant loci. *Nature communications*, 13(1):1644.

[Yokdang et al., 2011] Yokdang, N., Tellez, J. D., Tian, H., Norvell, J., Barsky, S. H., Valencik, M., and Buxton, I. L. (2011). A role for nucleotides in support of breast cancer angiogenesis: heterologous receptor signalling. *British journal of cancer*, 104(10):1628–1640.

[Yousri et al., 2018] Yousri, N. A., Fakhro, K. A., Robay, A., Rodriguez-Flores, J. L., Mohney, R. P., Zeriri, H., Odeh, T., Kader, S. A., Aldous, E. K., Thareja, G., Kumar, M., Al-Shakaki, A., Chidiac, O. M., Mohamoud, Y. A., Mezey, J. G., Malek, J. A., Crystal, R. G., and Suhre, K. (2018). Whole-exome sequencing identifies common and rare variant metabolic QTLs in a middle eastern population. *Nature communications*, 9(1):333.

[Yusuf et al., 2016] Yusuf, S., Bosch, J., Dagenais, G., Zhu, J., Xavier, D., Liu, L., Pais, P., López-Jaramillo, P., Leiter, L. A., Dans, A., Avezum, A., Piegas, L. S., Parkhomenko, A., Keltai, K., Keltai, M., Sliwa, K., Peters, R. J. G., Held, C., Chazova, I., Yusoff, K., Lewis, B. S., Jansky, P., Khunti, K., Toff, W. D., Reid, C. M., Varigos, J., Sanchez-Vallejo, G., McKelvie, R., Pogue, J., Jung, H., Gao, P., Diaz, R., Lonn, E., and HOPE-3 Investigators (2016). Cholesterol lowering in Intermediate-Risk persons without cardiovascular disease. *The New England journal of medicine*, 374(21):2021–2031.

[Zakai et al., 2022] Zakai, N. A., Minnier, J., Safford, M. M., Koh, I., Irvin, M. R., Fazio, S., Cushman, M., Howard, V. J., and Pamir, N. (2022). Race-Dependent association of High-Density lipoprotein cholesterol levels with incident coronary artery disease. *Journal of the American College of Cardiology*, 80(22):2104–2115.

[Zeiner et al., 2019] Zeiner, J., Loukovaara, S., Losenkova, K., Zuccarini, M., Korhonen, A. M., Lehti, K., Kauppinen, A., Kaarniranta, K., Müller, C. E., Jalkanen, S., and Yegutkin, G. G. (2019). Soluble and membrane-bound adenylate kinase and nucleotidases augment ATP-mediated inflammation in diabetic retinopathy eyes with vitreous hemorrhage. *Journal of molecular medicine*, 97(3):341–354.

[Zhang et al., 2017] Zhang, Z., Gillespie, C., Bowman, B., and Yang, Q. (2017). Prediction of atherosclerotic cardiovascular disease mortality in a nationally representative cohort using a set of risk factors from pooled cohort risk equations. *PloS one*, 12(4):e0175822.

[Zhao et al., 2014] Zhao, S., Guo, Y., Sheng, Q., and Shyr, Y. (2014). Advanced heat map and clustering analysis using heatmap3. *BioMed research international*, 2014:986048.

[Zhong et al., 2021] Zhong, W., Edfors, F., Gummesson, A., Bergström, G., Fagerberg, L., and Uhlén, M. (2021). Next generation plasma proteome profiling to monitor health and disease. *Nature communications*, 12(1):2493.