Spring 2021

Clinical Treatment Human Disease Networks and Comparative Effectiveness Research: Analyses of the Medicare Administrative Data

Hao Mei
Yale University Graduate School of Arts and Sciences, sherrymei0330@gmail.com

Abstract

Clinical Treatment Human Disease Networks and Comparative Effectiveness Research: Analyses of the Medicare Administrative Data

Hao Mei

2021

As the nation's largest healthcare payer, the Medicare program generates an unimaginable vast volume of medical data. With an increasing emphasis on evidence-based care, how to effectively handle and make inferences from the heterogeneous and noisy healthcare data remains an important question. High-quality analysis could improve the quality, planning, and administrations of health services, evaluate comparative therapies, and forward research on epidemiology and disease etiology. This is especially true for older adults since this population's health condition is generally complicated with multimorbidity, and the healthcare system for older adults is riddled with administrative and regulatory complexities. Taking advantage of the scaled and comprehensive Medicare data, this dissertation focuses on outcome research, human disease networks, and comparative effectiveness research for older adults.

Healthcare outcome measures such as mortality, readmission, length of stay (LOS), and medical costs have been extensively studied. However, existing analysis generally focuses on one single disease (or at most a few pre-selected and closely related diseases) or all diseases combined. It is increasingly evident that human diseases are interconnected with each other. Motivated by the emerging human disease network (HDN) analysis, we conduct network analysis of disease interconnections on healthcare outcomes measures.

First, we propose a clinical treatment HDN that analyzes inpatient LOS data. In the network graph, one node represents one disease, and two nodes are linked with an edge if their disease-specific LOS are correlated (conditional on LOS of all other diseases). To accommodate zero-inflated LOS data, we propose a network construction approach based on the multivariate Hurdle model. We analyze the Medicare inpatient data for the period of January 2008 to December 2018. Based on the constructed network, key network properties such as connectivity, module/hub, and temporal variation are analyzed. The results are found to be biomedically sensible, especially from a treat-

ment perspective. A closer examination also reveals novel findings that are less/not investigated in the individual-disease studies. This work has been published in *Statistics in Medicine*.

Second, considering that many healthcare outcomes are closely related to each other, we propose a high-dimensional clinical treatment HDN that can incorporate multiple outcomes. We construct a clinical treatment HDN on LOS and readmission and note that the proposed method can be easily generalized to other outcomes of different data types. To deal with uniquely challenging data distributions (high-dimensionality and zero-inflation), a new network construction approach is developed based on the integrative analysis of generalized linear models. Data analysis is conducted using the Medicare inpatient data from January 2010 to December 2018. Network structure and properties are found to be similar to that of the LOS HDN (in Chapter 2) but provide additional insights into disease interconnections considering both LOS and readmission. The proposed clinical treatment of HDNs can promote a better understanding of human diseases and their interconnections, guide a more efficient disease management and healthcare resources allocation, and foster complex network analysis. The manuscript of this work has been drafted and is ready for submission.

Comparative effectiveness research aims to directly compare the outcomes of two or more healthcare strategies to address a particular medical condition. Such analysis can provide information about the risks, benefits, and costs of different treatment options, thus guide better clinical decisions. While conducting a randomized controlled trial is the gold-standard approach, there are several limitations. Efforts have been made to utilize healthcare record data in comparative effectiveness research. To estimate and compare causal effects of treatments/interventions, we use the Medicare data to emulate target clinical trials and develop a deep learning-based analysis approach.

Under emulation, target clinical trials are explicitly "assembled" using the Medicare data. As such, statistical methods for clinical trials can be directly applied to estimate causal effects. With emulation analysis, we evaluate the effectiveness and safety outcomes of rivaroxaban versus dabigatran for Medicare patients with atrial fibrillation. The results show that dabigatran is superior in terms of time to any primary event (including ischemic stroke, other thromboembolic events, major bleeding, and death), major bleeding, and mortality. This work has been submitted to *Clinical Epidemiology*. Considering that many regression-based statistical methods (e.g., Cox proportional hazards model for survival data) have too strict data assumptions, we further develop an innovative deep learning-based analysis strategy. With the "emulation + deep learning" approach, we study

the survival outcomes of endovascular repair versus open aortic repair for Medicare patients with abdominal aortic aneurysms. It is found that endovascular repair has survival advantages in both short- and long-term mortality. This work has been published in *Entropy*.

Significantly different and advancing from the existing literature, this dissertation extends the scope of outcome research, human disease networks, and comparative effectiveness research. The findings in this dissertation are shown to have scientific merits, and the methodological developments may have other applications and serve as prototypes for future analysis.

Clinical Treatment Human Disease Networks and Comparative Effectiveness Research: Analyses of the Medicare Administrative Data

A Dissertation
Presented to the Faculty of the Graduate School
of
Yale University
in Candidacy for the Degree of
Doctor of Philosophy

by
Hao Mei

Dissertation Director: Shuangge Ma

June, 2021

Contents

List of Figures

List of Tables

Acknowledgements

It has been eight years since I joined the Yale community: two years as a master's student, two years as a Yale staff, then four years as a Ph.D. student. It is quite a long and rewarding journey, and my experience has been nothing colorful and memorable. While I leveraged the unique opportunities presented during my research and my life at Yale, much appreciation goes to every great individual who agreed to participate in this process. Your help and accompany cannot go unrecognized.

First and foremost, I would like to express my deepest appreciation to my academic advisor, Dr. Shuangge Ma. Since my very first day at Yale, Dr. Ma has substantially contributed to my amazing experience at research. The support I obtained through his motivation, patience, and immense knowledge has been pivotal in my learning and academic inquiry. In my experience developing this dissertation, he has been most significant in enabling me to excel in my research projects. He has been a true mentor and role model not only academically but also in every aspect of my life. His enthusiasm, foresight, and sagacity have boosted my morals to pursue my future career and to live as a better man.

I would like to thank my dissertation committee members, Dr. Arjun Venkatesh and Dr. Joshua Warren, for their endless support in developing this dissertation. Dr. Venkatesh is an expert in emergency medicine and outcome research. His unique insight and expertise have guided me in choosing research directions that can solve real-world healthcare problems. Dr. Warren, as an expert in high-dimensional models, has provided great support in the methodological development of my research projects. Overall, the committee's contribution in generating my thesis topics, guiding synthesis of research materials, and determining research methods, have aided my progress in plausible ways. Thank you for your positive critiques and moral support, without which this dissertation would be impossible.

At the same time, my sincere gratitude goes to my colleagues at Yale Center for Outcomes Research and Evaluation (CORE). During my six years at CORE, I have gained unique opportunities to collaborate with talented and interdisciplinary researchers, which have enriched my practical experience at research and deepened my understanding of healthcare analytics. I would like to thank my mentor at CORE, Dr. Zhenqiu Lin, especially. As the analyst team leader, he always explores rewarding opportunities for us, emphasizes our personal learning and growth, and encourages us to think critically and creatively. His tremendous help and support have been a great inspiration to me in several ways. I would also like to thank Dr. Harlan Krumholz, Dr. Arjun Venkatesh, and Dr. Lisa Suter, whose unique insights in healthcare research have taught me to think beyond statistics, consider the big picture, and focus on what can make a real impact. Their deep enthusiasm in the field also inspired me to pursue a future career in public health research.

Further, I would like to thank all members of Dr. Shuangge Ma's lab, which have been a source of love and optimism. Special thanks go to Yaqing Xu, Ruofan Jia, and Jiping Wang. As the co-authors of my research projects, you are my most trusted lab-mates and collaborators. Your excellent professional skills, detail-oriented contributions, and passion for work have upheld the projects in a significant way. I really enjoy the moments we brainstorm, develop a working algorithm, stay up late drafting manuscripts, and of course, when the papers finally got published. Dozens of people in the academic sphere, including all faculties and staffs in the Yale Department of Biostatistics, have been a great source of help and support. My sincere appreciation to Mrs. Melanie Elliot and Dr. Christian Tschudi, who have helped me meet every academic milestone, linked me with helpful resources such as workshops and seminars, and addressed my every concern promptly. Thank you for always been trustworthy and supportive. I would also like to thank Dr. Maria Ciarleglio, Dr. Xiaomei Ma, and Dr. Lei Liu, who have agreed to serve as my dissertation readers. I really appreciate your time and efforts. I am looking forward to your comments and suggestions, which I believe would significantly improve my dissertation and guide my future research.

Importantly, my warmest appreciations go to my friends who have made my everyday life at Yale so joyful and memorable. Xinyue, Wenlan, and Yaqing, I feel so blessed to have friends like you. You are all master chefs. Thank you for many delicious meals that have made me feel so happy and fulfilled. I remember all your signature dishes, such as Xinyue's cakes, Wenlan's Sichuan cuisine, and Yaqing's bun and dumplings, all of which are a perfect match for my outstanding

dishwashing skill. I also remember all the great moments we cherished together, including the days of rock-climbing, grocery shopping, exploring new restaurants, and playing Animal Crossing and Timi. Thank you for sharing cheerful moments and accompanying me through difficult times. Your friendship is irreplaceable. Shuling and Ji-young, you are my best officemates. All the office chats were fun, relaxing, and intellectually stimulating. Shuling, thank you for being my best lunch buddy and sharing all those great statistical articles and books. Ji-young, thank you for teaching me the Korea-style makeups and all the long-distance and delightful greetings after you moved to Germany.

Last, I want to direct my appreciation to my most fundamental source of energy and power, my family. Heartful thanks go to my boyfriend, Xuefeng, who can always make me laugh when I feel down. I enjoy cooking with you, playing basketball with you, and traveling with you to see the world. Thank you for all the helpful suggestions and assistance in creating impressive scientific figures with lucid interpretations, not forgetting your tremendous support emotionally in developing this dissertation. Most importantly, I am deeply grateful to my parents, who have provided massive teaching and unconditional support in every step of my life. You always encourage me with a true belief in me when the times get rough. It is always a great comfort and relief to know that you are willing to provide everything you can, even though I know that you are not superman and you have struggled times too. It was because of your wholehearted encouragement and support that I can finish my degree.

Chapter 1

Introduction

There is a growing consciousness among healthcare institutions for adopting the electronic health record (EHR) system to improve care quality through evidence-based practices. As of 2015, more than 84% of the U.S. non-federal acute care hospitals have adopted a basic EHR system [1], which has generated an unimaginably vast volume of patient data. As healthcare data becomes more voluminous and convoluted, there exist both opportunities and challenges. Advancements in the statistical field to effectively handle and make inferences from the heterogeneous and noisy health-care data are required for the transition from information to knowledge. As administrative and billing data is one essential component of the EHR system, this dissertation focuses on the Medicare database, aiming to answer critical biomedical questions that have been previously unanswered.

Medicare is a federal health insurance program for adults aged 65 years and above, certain younger people with disabilities, and people with end-stage renal disease (permanent kidney failure requiring dialysis or a transplant). As the single largest payer of healthcare in the U.S., it covers 98% of adults aged 65 and above [2]. It also accounts for over 99% of death for the elderly population and about 40% of all inpatient discharges [2, 3]. The Centers for Medicare & Medicaid Services (CMS) offers a wide range of datasets that follow Medicare beneficiaries across multiple care settings. Specifically, it collects over two billion data points per year through reimbursement to hospital care (Medicare PartA), physician and outpatient services (Medicare Part B), drug prescription (Medicare PartD), and other health care claims. It also collects billions of other data points through enrollment information, beneficiary eligibility checks, quality metrics, and calls to 1-800-MEDICARE [4], which can fulfill many research purposes. In addition to its universal coverage and rich information

contained, the Medicare data is generally of high quality since it is the basis of determination of reimbursing eligibility. Demographic information is mostly reliable and valid because it comes from the Social Security Administration [3]. In the literature, there have been enormous efforts exploring the Medicare database to answer various questions. This dissertation focuses mainly on outcome research, disease interconnections, and comparative effectiveness research.

1.1 Literature review

1.1.1 Outcome research

Healthcare outcome measures such as mortality, readmission, length of stay (LOS), and medical costs are the quality and cost targets healthcare organizations are trying to improve. The importance of delivering high-value care for patients is increasingly emphasized, with value defined as the health outcomes achieved per dollar spent [5]. For most if not all diseases, hundreds of these outcomes have been extensively studied and reported to stakeholders at all levels. Such analysis can inform a more efficient allocation of resources, identify variations of care, and reveal areas in which interventions could improve care. The Institute for Healthcare Improvement describes outcome measurement as "a critical part of testing and implementing changes. Measuring tells a team whether the changes they are making lead to improvement [6]." In other words, outcome research fosters improvement and adoption of better practices, thus further improves outcomes. For example, Piedmont Healthcare's evidence-based care standardization for pneumonia patients resulted in a 56.5% relative reduction in mortality rate and a 9.3% relative reduction in LOS [7]. By improving the analytic platform and advancing its applications, the University of Texas Medical Branch achieved a 14.5% relative decrease in their 30-day all-cause readmission rate, resulting in $1.9 million in cost reduction [8]. Also, since outcomes informatively reflect intrinsic disease properties (prevalence, severity, trend, etc.), outcome analysis can advance disease etiology research [9]. Accordingly, understanding outcomes is essential in providing patient-centered and value-based healthcare services. This statement is especially true for older adults since the health condition for this population is generally complicated with multimorbidity, and the healthcare system for this population is riddled with administrative and regulatory complexities [10].

In the literature, extensive analyses on healthcare outcomes have been done using healthcare

record data, especially the Medicare data. They generally fall into two families. The first family focuses on a single disease. For example, with a Saskatchewan Ministry of Health hospital administrative database, Feng and Li [11] analyzed the LOS of ischaemic heart disease. The zero-inflation nature of data was observed, which is typical in outcome data. For most conditions, only a proportion of the general population may be susceptible and have treatment [11]. Chronic diseases not needing any hospital treatment is another contributing factor [12]. Feng and Li [11] conducted a two-part regression analysis to accommodate zero-inflation and identified aboriginal status, age, and gender as risk factors. Gupta et al. [13] evaluated the effect on a federal program's outcomes to reduce condition-specific readmission rates for the Medicare population. With interrupted time-series and survival analyses, Gupta et al. [13] identified a reduction in readmission rate and an increase in mortality associated with implementing the program for Medicare patients hospitalized with heart failure. Other studies include Rinne et al. [14], who investigated the association between LOS and readmission for chronic obstructive pulmonary disease (COPD); and Petersen et al. [15], who compared multiple outcomes of myocardial infarction in Veterans Health Administration patients versus Medicare patients.

The second family focuses on a general population and studies all diseases combined. For example, Huling et al. [16] proposed a framework for estimating individualized treatment rules (ITRs) that optimize all disease combined medical costs in the Medicare population. Observing the zero-inflation nature of the medical costs data, Huling et al. [16] employed a two-part semicontinuous modeling, wherein the ITR is estimated by separately targeting the zero part of the outcome and the strictly positive part. Also, they employed a cooperative LASSO penalty to simultaneously select variables and encourage the signs of coefficients for each variable to agree between the two components of the ITR. In healthcare outcome studies, penalized optimization is commonly used since modeling often involves a large number of covariates, and many of them are naturally grouped with variables in the same group being systematically related or statistically correlated. In analyzing all diseases combined number of doctor visits data, Chatterjee et al. [17] also noted the zero-inflation nature of data and group-wise correlation in covariates. Under such settings, they propose a unified algorithm (Gooogle: Group Regularization for Zero-inflated Count Regression Models) using a least-squares approximation of the mixture likelihood and a variety of group-wise penalties on the coefficients. Other examples of all diseases combined outcome studies include Jung et al. [18], who

compared general readmission rates in two populations: Medicare Advantage and Medicare Fee-For-Service; and Dall et al. [19], who examined the increasing trend of healthcare resource demand and identified the overall shortage of hospital beds in the U.S.

There are also a few "scattered" studies analyzing treatment measures of a few pre-selected and closely related diseases, such as cardiovascular diseases [20] and certain cancers [21]. Correlations between outcomes of different diseases have been noted in such studies. For example, in a study investigating LOS variations within the Diagnosis Related Groups, Berki et al. [22] argued that patients with similar diagnostic and other case-management-relevant characteristics had correlated LOS measures and emphasized the necessity of jointly analyzing LOS behaviors for diseases homogeneous in clinically relevant characteristics. In a study examining the association between COPD readmission and other quality measures, Rinne et al. [23] discovered modest correlations between COPD readmission rate and readmission rates for other medical conditions, including heart failure, acute myocardial infarction, pneumonia, and stroke; low correlations between COPD readmission rate and mortality rates for all measured conditions; and significant correlations between COPD readmission rate and all patient experience measures. The authors suggested that although pay-for-performance programs generally focused on individual disease outcomes, there may be common organizational factors that influenced multiple disease-specific outcomes. Similar arguments have been made in follow-up studies. However, existing studies on outcome interconnections are limited to a small number of diseases.

1.1.2 Human disease network

There is increasing evidence that human diseases are not isolated from each other. Instead, diseases are related through multiple dimensions. For example, correlation between diseases can be caused by 1) shared or causal disease etiology (e.g., between breast cancer, ovarian cancer, and other hormone-related diseases; between diabetes and its complications), 2) shared environmental, dietary, socioeconomic, and other risk factors (e.g., between multiple respiratory diseases that share air pollution as a common risk factor), and 3) shared prevention, diagnosis, and treatment strategies (e.g., between diseases that need radiological treatments). The emerging human disease network (HDN) analysis in the past two decades offers a platform to explore disease interconnections using a single graph-theoretic framework. In an HDN, one node represents one disease, and

two nodes are connected with an edge if they are determined to be associated or correlated by an underlying model. Disease interconnection studies, especially the HDN ones, have significantly advanced biomedical research beyond the individual-disease ones.

Goh et al. [24] were the first to develop the concept of HDN and establish pan-disease interconnections. They proposed an HDN that linked disorders and disease genes by the known disorder–gene associations, indicating many diseases' common genetic origin. The authors found that, while a few essential human genes play a central role in the human interactome, the vast majority of disease genes are nonessential, show no tendency to encode hub proteins, and are localized in the functional periphery of the network. A closer examination showed that the network structure supported distinct disease-specific functional modules. This study and a few other follow-up HDNs are gene-centric, under which two diseases are interconnected if they share common genetic risk factors. For example, Lee et al. [25] built a network that linked two diseases if the mutated enzymes associated with them catalyzed adjacent metabolic reactions. They found that connected disease pairs displayed higher correlated reaction flux rate, corresponding enzyme-encoding gene coexpression, and higher comorbidity than those with no metabolic link between them. Li and Agarwal [26] constructed an HDN by linking diseases based on shared pathways where disease-associated genes are enriched. Examining the network properties, the authors provided examples of novel disease relationships that cannot be readily captured through simple literature search or gene overlap analysis. Such HDNs representing various molecular relationships have provided new insights into disease etiology, classification, and shared biological mechanisms.

Noting that genetic information reflects only part of disease interconnections, efforts have been made to utilize the healthcare record data in constructing HDNs. Working on the clinical history of 30 million patients, Hidalgo et al. [27] developed the first phenotypic disease network that connects two diseases if their phenotypes are correlated (e.g., comorbidity). This kind of phenotypic HDN has important clinical implications, as the network structure is closely relevant to understanding illness progression. Hidalgo et al. showed that 1) diseases progressed preferentially along with the links of the network; 2) this progression is different for patients with different genders and racial backgrounds, and 3) patients affected by highly connected diseases had a higher mortality risk. Follow-up studies along this direction include Jiang et al. [28], who explored the Taiwan National Health Insurance Research Database and calculated disease comorbidity probability using

the ϕ-correlation; and Roque et al. [29], who used text mining and data extraction techniques on unstructured electronic health record data. Phenotypic HDNs utilize largely accessible healthcare administrative data, have lucid interpretability, and complement existing genetic HDNs by accommodating non-molecular connections. There are also studies that go multi-layer, under which one layer describes the genetic interconnections among diseases, and another layer describes the phenotypic interconnections. Building a multi-layer HDN, Halu et al. [30] showed that diseases with common genetic constituents tended to share symptoms and uncovered how phenotype information helped boost genotype information.

There emerges a relatively new family of HDN that examines disease interconnections in healthcare outcomes. Ma et al. [31] were the first to construct a clinical treatment HDN by analyzing medical costs data in the Taiwan National Health Insurance Research Database. They used a two-part model for the marginal distribution of disease-specific medical costs and a copula-based approach to identify unconditional pairwise interconnection between medical costs of two diseases. With medical costs informatively reflect the financial burden of care and disease severity, such a clinical treatment HDN has unique implications for disease management and resource allocation. Healthcare outcome measures are "correlated with" but not equivalent to genetic risk factors and phenotypes. The interconnections in disease-associated outcomes cannot be derived directly from the existing genetic and phenotypic HDNs. In other words, a clinical treatment HDN captures disease interconnections from not only shared causal disease etiology, shared environmental, dietary, and socioeconomic risk factors, but also shared prevention, diagnosis, and treatment strategies.

1.1.3 Comparative effectiveness research

Comparative effectiveness research aims to evaluate and compare the outcomes of two or more healthcare strategies to address a particular medical condition. The goal of comparative effectiveness research is to provide more evidence to guide clinical decisions. Ideally, this should be achieved through the gold standard double-blinded randomized clinical trials. However, there exist several practical limitations. A randomized clinical trial is generally expensive and sometimes infeasible. Limited sample size and inadequate follow-ups are common scenarios that diminish the statistical power of a randomized controlled trial. Moreover, a randomized controlled trial tests treatment efficacy under rigorous experimental conditions rather than real-world effectiveness. Results from

a randomized clinical trial may have limited generalizability to other than designed populations (e.g., older adults who are underrepresented in clinical trials).

With the concerns mentioned above, efforts have been made to focus on observational data in developing comparative effectiveness evidence. As one of the richest sources of treatment information in the country, the Medicare data plays an essential role in observational comparative effectiveness research. For example, Graham et al. [32] examined the effectiveness and safety outcomes of dabigatran versus warfarin for Medicare patients with nonvalvular atrial fibrillation. They found that dabigatran was associated with a reduced risk of ischemic stroke, intracranial hemorrhage, death, and increased risk of major gastrointestinal bleeding. The authors used propensity scores to match the two treatment arms and Cox proportional hazards regression to compare time to events. Even with randomization in clinical trials, propensity scores are generally used to weigh or match cohorts to account for potential confounding in comparative effectiveness research. In a study of survival outcomes after endovascular repair versus open repair of abdominal aortic aneurysm in Medicare beneficiaries, Schermerhorn et al. [33] also analyzed propensity score-matched cohorts with Cox proportional hazards regression. They found that endovascular repair was associated with a substantial early survival advantage that gradually decreased over time.

Observational comparative effectiveness studies using healthcare records data are more cost-effective, have a large sample size and statistical power, and implicate actual practices by assessing real-world effectiveness. However, it is well established that observational studies generate results on associations instead of the desired causality. To tackle this problem, statistical techniques have been developed to conduce causal inferences using observational data. For example, van der Laan and Rubin [34] proposed the targeted maximum likelihood estimation (TMLE), which is an automated and iterative procedure to analyze censored observational data in a way that allows effect estimation in the presence of confounding factors. Chipman et al. [35] proposed a Bayesian additive regression trees (BART) model where each tree is constrained by a regularization prior, and fitting and inference are accomplished via an iterative Bayesian backfitting MCMC algorithm that generates samples from a posterior. Effectively, BART is a nonparametric Bayesian regression approach that enables full posterior inference on average causal effects. As the existing literature on this topic is vast, we note that different methods have different advantages and disadvantages [36]. There exists no dominating approach.

In this dissertation, we focus on the emulation approach proposed by Hernan et al. [37]. Under emulation, target clinical trials are explicitly "assembled" using observational data, and statistical techniques for randomized clinical trials can be directly applied. As such, compared to causal inference techniques such as the TMLE and BART, emulation has much more lucid interpretations. Using EHR data, especially the Medicare data, this approach has been employed on various illness conditions and treatment strategies. Petito et al. [38] evaluated the suitability of emulation analysis for assessing the effectiveness of adding a drug (fluorouracil or erlotinib) to an existing treatment regimen on the overall survival of elderly patients with cancer. They utilized the Surveillance, Epidemiology, and End Results (SEER)–Medicare linked database to emulate two existing clinical trials. The emulation analysis findings were not meaningfully different from those in elderly subgroup analyses reported from randomized trials. However, naive observational estimates were not compatible with those from previous trials. In the situation that a real randomized trial is not practical, García-Albéniz et al. [39] utilized the Medicare data to emulate a hypothesized trial that evaluates the effectiveness and safety of screening colonoscopy prevent colorectal cancer. They concluded a modest benefit of screening colonoscopy in preventing colorectal cancer for beneficiaries aged 70 to 74 years and a smaller benefit for older beneficiaries.

1.2 Background studies

In addition to a throughout literature review, I have collaborated with interdisciplinary researchers at Yale Center for Outcome Research and Evaluation on a variety of outcomes studies, mostly involving analyses of the Medicare/Medicaid administrative data. These studies are limited to disease-specific or all diseases combined outcome analysis, which have ignored the complex disease interconnection. However, they demonstrate the significance of outcome research and are essential motivations of this dissertation.

In a series of three papers investigating emergency department (ED) utilization in the Medicare population, Venkatesh et al. first constructed an operational definition for ED visitation using comprehensive Medicare data sets [40], then examined ED utilization in vulnerable older adult subpopulations [41, 42]. Despite the high profile of ED visits in analyses using administrative claims, little work had evaluated the degree to which existing definitions based on claims data accurately

capture conventionally defined hospital-based ED services. Based on expert consensus and clinician review, we applied several modifications to existing definitions to construct a new operational definition for ED visits using both provider- and facility-based Medicare administrative claims. A comparison between our definition and three existing definitions showed significant differences in ED visitation estimates, as our definition identified ED visits not captured by previous definitions. This work has provided several points of guidance to researchers seeking to use administrative claims data for ED research. Based on the newly developed definition, we further investigated ED utilization in the vulnerable sub-populations, including Medicare beneficiaries with multiple chronic conditions (MCCs), dual eligibility, hospice enrollment, and skilled nursing facility (SNF) residence. We found that vulnerable sub-populations disproportionately visit the ED compared to physician offices for unscheduled care. For example, 1) dual-eligible beneficiaries demonstrated higher ED visit rates and lower office visit rates for unscheduled care; 2) the sub-population with MCCs uses both the ED and the office settings for unscheduled care more so than any other groups, and 3) a higher proportion of unscheduled care were made to EDs by beneficiaries after a SNF stay in comparison to those actively residing in a SNF and those without SNF utilization. These findings have important implications for improving care coordination, measuring quality, or reform payment to influence ED visitation.

Another work involving ED visits evaluated the effectiveness of an ED-initiated Patient Navigation program (ED-PN) in improving health care access for Medicaid beneficiaries [43]. Medicaid enrollees frequently utilized the ED due to barriers to access health care services in other settings. The ED-PN program was designed to have patient navigators who met with the ED patients and linked them to financial assistance programs and other resources. They helped patients identify a primary care provider, assisted them in scheduling primary care and specialist appointments, and accompanied them for up to three visits. They also educated patients on why they shouldn't use the ED for primary care and taught them how to navigate the healthcare system. To evaluate the ED-PN program's effectiveness, we conducted a prospective, randomized controlled trial comparing ED-PN with usual care among 100 Medicaid-enrolled frequent ED users (defined as 4–18 ED visits in the prior year). The primary outcome was ED utilization during the 12 months pre and post-enrollment. Secondary outcomes included hospitalizations, outpatient utilization, hospital costs, and Medicaid costs. Using the difference-in-difference approach, we found that the ED-PN pro-

gram demonstrated significant reductions in ED visits and hospitalizations in the 12 months after enrollment. We also compared characteristics between ED-PN patients with and without reduced ED utilization and found that older patients and patients with lower health literacy reduced utilization more than younger patients and patients with higher health literacy. This study provides high-quality evidence to support ED-PN's effectiveness, which has important clinical implications for improving care for the high-need and high-cost Medicaid patients.

Collaborative analyses have also been done on other outcome measures. Considering that the accrual of final Medicare claims can take up to a year, Li et al. [44] developed real-time reporting (within two months from admission) models for national trends of multiple outcomes. Specifically, with incomplete and non-final claims data, we developed time series models for real-time estimation of national admission, readmission, and observation-stay rates in Medicare patients with acute myocardial infarction, heart failure, or pneumonia. It was shown that these models provided validated estimates and predictions. This work allows policymakers to track policy decisions' impact in real-time and enable hospitals to better monitor their performance compared to a national benchmark. In response to the tremendous impact of coronavirus disease 2019 (COVID-19), Janke et al. [45] examined the relationship between hospital resources and mortality among hospital referral regions from March 1 to July 26, 2020. Using American Hospital Association data and COVID-19 data from the New York Times, we found that geographic areas with fewer intensive care unit beds, nurses, and general medicine/surgical beds were significantly associated with more deaths. The association was found stronger in the early pandemic period. Our findings underscore the potential impact of innovative hospital capacity protocols and care models to create resource flexibility and limit system overload early in a pandemic.

1.3 Summary

Through a comprehensive literature review and a variety of collaborative studies, the importance of healthcare outcome analysis has been well established. However, it is noted that existing analyses are generally limited to a single disease (or at most a few pre-selected and closely related diseases) or all diseases combined. The complex interconnections in human diseases have been ignored. Motivated by the emerging of HDN analysis, the first two aims (Chapter 2 and Chapter 3) of this

dissertation are to investigate interconnections in healthcare outcome measures at a pan-disease level. Specifically, in Chapter 2, we construct a clinical treatment HDN on Medicare LOS data. To accommodate zero-inflated data, we develop a network construction approach based on the multivariate Hurdle approach. In Chapter 3, considering many outcomes are closely related to each other, we expand the clinical treatment HDN to incorporate multiple outcomes: LOS and readmission. We propose an innovative modeling approach based on integrative analysis of generalized linear models to accommodate high-dimensionality and zero-inflation. In both chapters, novel modeling and estimation approaches are developed to accommodate uniquely challenging data distributions. In addition, while most existing HDNs are based unconditional pairwise relationships, our models quantify conditional disease interconnections, which are more informative and statistically more challenging. Based on the constructed networks, fundamental network properties such as connectivity, module/hub, and temporal trend are examined to answer important biomedical questions. The proposed clinical treatment HDNs can promote a better understanding of human diseases and their interconnections, guide a more efficient disease management and healthcare resources allocation, and foster complex network analysis.

It is also noted that comparative effectiveness research analyzing the observational Medicare data can only lead to conclusions on association instead of causation. Causal inference is often desired to make definite conclusions on the relative performance of treatment strategies. Ideally, this should be achieved using randomized controlled trials, as we have done in evaluating the effects of the ED-PN program on ED utilization and other outcomes [43]. However, as discussed earlier, a randomized controlled trial is not always feasible and has limitations. To take advantage of the scaled Medicare data, Chapter 4 aims to conduct causal inference with the emulation approach. Under emulation, we use the observational Medicare data to explicitly "assemble" a hypothetical clinical trial with rigorously defined inclusion/exclusion criteria, treatment regimens, and analysis procedures. Statistical methods for clinical trials can be directly applied, and causal conclusions can thus be drawn. We first conduct a case study to evaluate the effectiveness and safety outcomes of rivaroxaban versus dabigatran for Medicare patients with atrial fibrillation. We analyze the Medicare data using propensity score and inverse probability of treatment (IPT) weighting Cox proportional hazards regression, which are commonly used statistical techniques in clinical trials. Considering that the Cox proportional hazards model has too strict data assumptions, we further

develop a deep learning-based analysis strategy with relaxed assumptions for survival data. This work is the first to introduce deep learning to the emulation paradigm. Building on the existing deep learning components, we propose an innovative analysis pipeline that mimics the "propensity score + IPT weighting Cox regression" approach and conduct a bootstrap-type procedure for variation assessment. With the proposed "emulation + deep learning" approach, we conduct another case study to compare survival outcomes of endovascular repair versus open aortic repair for Medicare patients with abdominal aortic aneurysms.

Chapter 2

Clinical Treatment Human Disease Network: Analysis of the Medicare Inpatient Length of Stay Data

2.1 Introduction

For most if not all diseases, clinical treatment measures, such as inpatient LOS, number of outpatient treatments, and inpatient/outpatient treatment cost, have been extensively studied. Such analysis has important implications for stakeholders at all levels. For example, the analysis of LOS can inform hospitals how to effectively and efficiently allocate beds and human resources and plan forward [46]. The analysis of treatment cost can inform government agencies and individuals/households how to allocate financial resources as well as guide insurance agencies on more accurately determining premium [19]. Clinical treatment measures can also informatively reflect prevalence, trend, severity, and other disease properties, and as such, the analysis can also advance research on disease etiology [9].

In this article, we focus on LOS because of its high clinical relevance and note that other clinical treatment measures can be analyzed in a similar manner. LOS is one of the most watched clinical treatment measures. In the U.S., hospital stays cost the health system at least $377.5 billion per year [19]. There have been extensive clinical and managerial efforts that target LOS as an indicator

of high-value care [46]. Researchers have also established significant positive correlations between LOS and disease severity. For example, according to Hicks et al. [47], using the Medis Groups which allocates patients into five admission severity groups, there is an increase in mean LOS from 2.7 days to 14.3 days from the lowest to the highest severity group.

Existing studies on LOS and other clinical treatment measures mostly belong to two families. The first family analyzes the treatment of a single disease. For example, with a Saskatchewan Ministry of Health hospital discharge administrative database, Feng and Li [11] analyzed the LOS of ischaemic heart disease. The zero-inflation nature of data was observed, which is common as for most diseases, only a proportion of the general population may be susceptible and have treatment. Chronic diseases not needing any hospital treatment is another contributing factor [11, 12]. Feng and Li [11] conducted a two-part regression analysis to accommodate zero-inflation and identified aboriginal status, age, and gender as risk factors. The second family of analysis considers diseases all together. For example, studies have examined the increasing trend of healthcare resource demand and identified the overall shortage of hospital beds in the U.S. and other countries [19]. There are also a few "scattered" studies analyzing treatment measures of a few pre-selected and closely related diseases, such as cardiovascular diseases [20] and certain cancers [21].

Significantly different and advancing from the existing literature, in this article, we will study the interconnections in LOS among diseases. This has been partly motivated by the surge of disease interconnection studies in the past two decades. For example, Goh et al. [24] was the first to develop the concept of human disease network (HDN) and establish pan-disease interconnections. This and a few other followup HDNs [26, 48] are gene-centric, under which two diseases are interconnected if they share common genetic risk factors. Building on the clinical history of 30 million patients, Hidalgo et al. [27] developed the first phenotypic disease network. Followup studies include Jiang et al. [28], Roque et al. [29], and others. Under such a network, two diseases are interconnected if their phenotypes/outcomes are correlated (e.g., comorbidity). There are also studies that go multi-layer [30], under which one layer describes the genetic interconnections among diseases, and another layer describes the phenotypic interconnections. Disease interconnection studies especially the HDN ones have significantly advanced biomedical research beyond the individual-disease ones. LOS and other clinical treatment measures are "correlated with" but not equivalent to genetic risk factors and phenotypes. As such, the interconnections in LOS (or other clinical treatment measures)

among diseases cannot be derived from the molecular and phenotypic HDNs. The interconnection in LOS among diseases has been noted. For example, in a study investigating LOS variations within the Diagnosis Related Groups, Berki et al. [22] argued that patients with similar diagnostic and other case-management-relevant characteristics had correlated LOS measures and emphasized the necessity of jointly analyzing LOS behaviors for diseases homogeneous in clinically relevant characteristics. Similar arguments have been made in followup studies. However, the existing studies on LOS interconnections are limited to a small number of diseases. This study, with a global perspective, is expected to have higher significance, in the same way as the HDN ones.

Interconnections in clinical treatment measures can be caused by shared or causal disease etiology (for example, between breast cancer, ovarian cancer, and other hormone-related diseases; between diabetes and its complications), shared environmental, dietary, socioeconomic, and other risk factors (for example, between multiple respiratory diseases that share air pollution as a common risk factor), shared prevention, diagnosis, and treatment strategies (for example, between diseases that need radiological treatments), and other factors. In Figure 2.1, we showcase the marginal and joint distributions of LOS of two common diseases: essential hypertension and diabetes mellitus without complication. The nonparametric fit (blue line in the scatter plot) clearly shows a positive correlation. In addition, from the marginal distributions, the zero-inflation nature of data is clear.

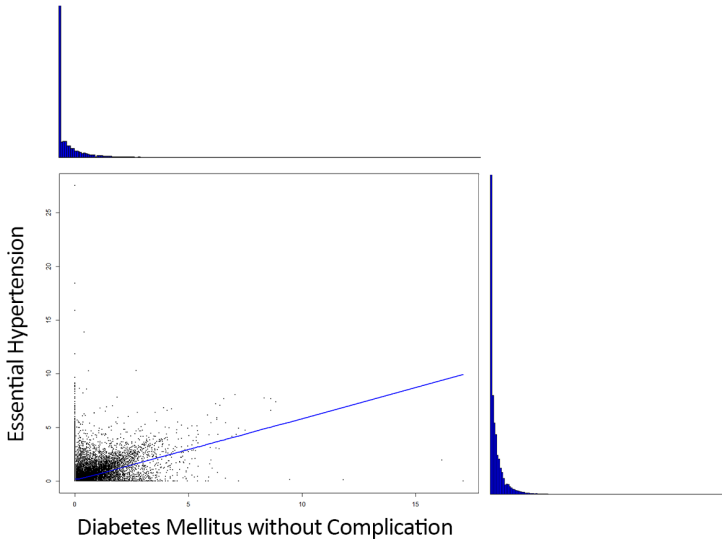


Figure 2.1: Joint and marginal distributions of LOS of essential hypertension and diabetes mellitus without complication

Government agencies, insurance companies, hospitals, and individuals all manage a large number of diseases with limited healthcare, financial, and human resources. To more efficiently allocate and plan hospital beds and other resources, which has public health, economic, and ethical importance, it is critical to go beyond individual diseases, take a global view, and account for the interconnections among diseases in clinical treatment measure research. For example, as suggested in other HDN studies, tightly interconnected diseases should be considered together, diseases interconnected with a large number of others should receive higher priority, and interventions targeting factors that affect a group of interconnected diseases can be more efficient and cost-effective. Last but not least, interconnections in treatment can also informatively reflect intrinsic disease properties.

This study can complement the existing literature and fill knowledge gaps in multiple ways. Specifically, it differs from the existing LOS analyses by uniquely focusing on the interconnections among diseases. It advances from studies that analyze a small number of pre-selected diseases by conducting analysis at the pan-disease level. It differs from the existing molecular and phenotypic HDNs by being "closer" to clinical treatments, thus having a higher practical value. In the literature, the most relevant is perhaps the HDN analysis by Ma et al. [31], which analyzed treatment costs and demonstrated the significance of analyzing disease interconnections in treatment. The present study differs from Ma et al. [31] in multiple ways. In particular, Ma et al. [31] analyzed treatment costs, while we examine LOS. The Taiwan population analyzed in Ma et al. [31] dramatically differs from the Medicare population. In terms of methodology, Ma et al. [31] examines unconditional interconnections, which can be much easier and less informative than conditional interconnections studied here. In the literature, there is a lack of conditional network analysis approach tailored to LOS data and, as such, new methodological development is needed. Overall, this study is warranted beyond the existing literature.

2.2 Data

Medicare is a federal health insurance program for adults aged 65 years and above, certain younger people with disabilities, and people with end-stage renal disease (permanent kidney failure requiring dialysis or a transplant). It is estimated that 98% of adults aged 65 and older in the U.S. are enrolled

in Medicare [2]. The Medicare program accounts for over 99% of death for adults aged 65 and older and about 40 percent of all inpatient discharges [2, 3]. The Medicare claims are bills for services provided to the Medicare enrollees. The Center for Medicare & Medicaid Services (CMS) offers a wide range of claims data that is derived from reimbursement or payment of bills.

In analysis, we first retrieve records on 133 million hospital inpatient admissions for the period of January 2008 to December 2018. These admissions cover Medicare fee-for-services utilized by 35 million Medicare beneficiaries. We further randomly select data on 100,000 subjects aged 65 years and above.Published studies [49, 50] suggest that the sample size needed for an accurate network estimation depends on the complexity of network structure (e.g., number of nodes and sparsity level). For the proposed analysis, in simulation (details in Section 2.5), we find that a sample size of 10,000 is sufficient to generate accurate estimation for a variety of settings with 100 nodes. In data analysis, with 108 nodes, we conservatively choose a sample size of 100,000, which is expected to be sufficiently large while still being computationally affordable.

For each inpatient claim, there can be up to 25 (1 primary and 24 secondary) diagnosis codes, which are defined under the International Classification of Diseases, Ninth Revision, Clinical Modification (ICD-9-CM) for discharges prior to October 1, 2015 and under the International Classification of Diseases, Tenth Revision, Clinical Modification (ICD-10-CM) otherwise. For each inpatient treatment episode, LOS is calculated as the length between dates of admission and discharge. To accommodate multiple possible conditions per visit, we allocate 60% of the LOS to the primary diagnosis and divide the rest evenly among all secondary diagnoses. For example, if an inpatient treatment episode involved a LOS of ten days and five diagnosis codes, six days will be attributed to the primary diagnosis, and one day will be attributed to each of the four secondary diagnoses. We note that in the literature [51], there is still a lack of consensus on how to allocate among multiple disease conditions. The adopted allocation ensures that for each visit, the primary diagnosis dominates and that all comorbidity conditions contribute equally. Then for each subject, LOS for each disease is summed over all inpatient treatment episodes within the study period. The resulted analysis dataset has 100,000 observations (one for each subject), and each observation contains the summed LOS for each disease. A zero-entry means that the subject had not been admitted to hospitals because of the corresponding disease within the study period.

As can be seen from the marginal distributions in Figure 2.1, in addition to the zero-inflation

nature, LOS data are also skewed. Following the literature [11,52], we conduct marginal logarithm transformations for the non-zero data entries. Figure 2.2 shows the distributions of the non-zero data before and after transformation for essential hypertension, diabetes mellitus without complication, and glaucoma (as an example of rare diseases). From Figure 2.2 and other alike (omitted here), we conclude that, although there may still be slight skewness after transformation, normality can be reasonably assumed.

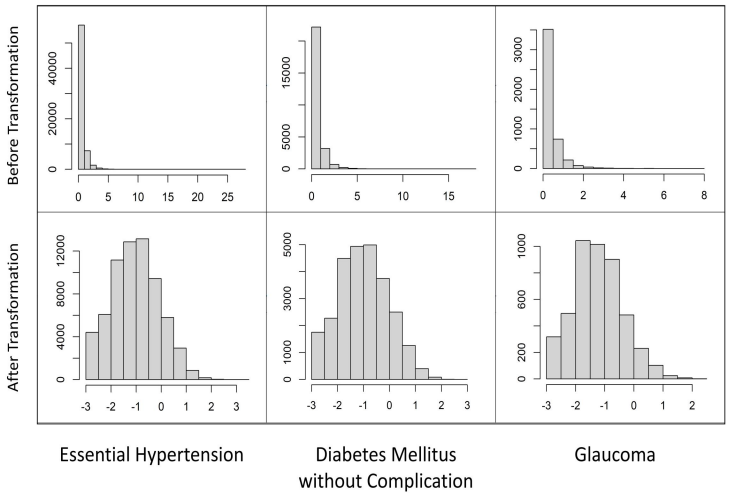


Figure 2.2: Distributions of non-zero LOS data before and after transformation

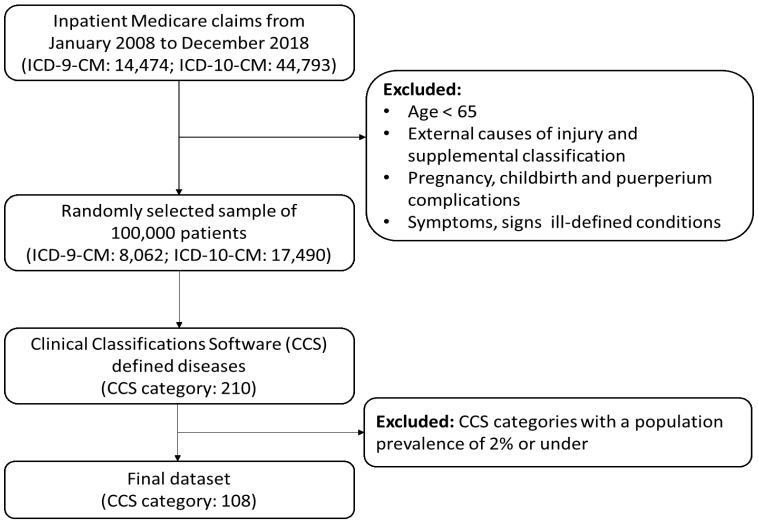
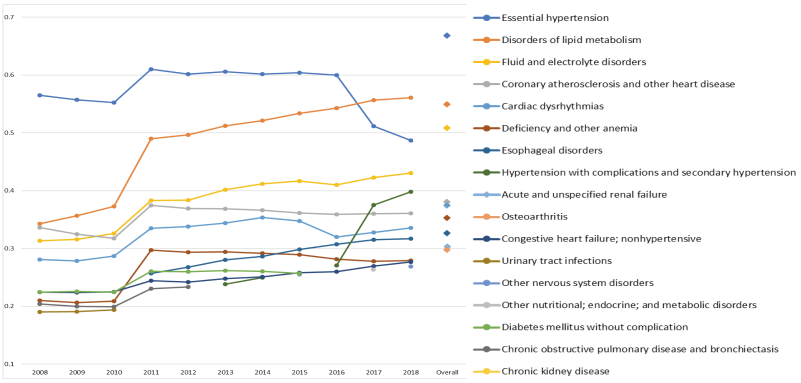


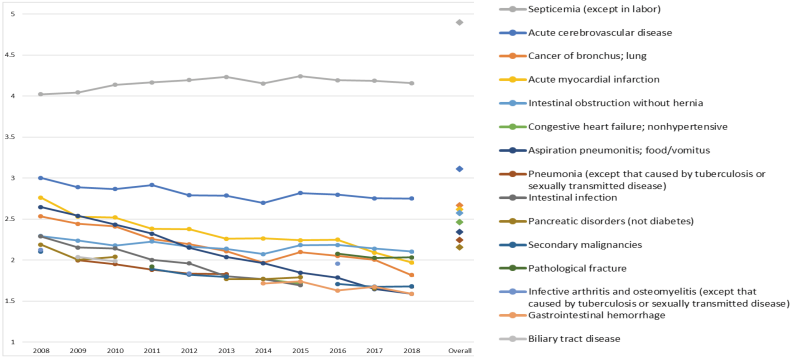
Figure 2.3: Flowchart of data processing

Following the literature [28], we exclude the following diseases (codes) from analysis: 1) external causes of injury and supplemental classification (the E and V codes in ICD-9-CM and V00-Z99 in

ICD-10-CM); 2) pregnancy, childbirth and puerperium complications (630 – 679 in ICD-9-CM and O00-O9A in ICD-10-CM); and 3) symptoms, signs & ill-defined conditions (760-999 in ICD-9-CM and P00-P96, R00-R99, and S00-T88 in ICD-10-CM). This leads to 8,062 ICD-9-CM codes and 17,490 ICD-10-CM codes. To better classify and define human diseases, these diagnosis codes are further grouped using the Clinical Classifications Software (CCS) developed by the Agency for Healthcare Research and Quality. CCS is a tool for clustering disease diagnoses into a manageable number of clinically meaningful categories and has been widely used in a variety of studies related to diagnoses [53]. To generate more reliable estimates, we focus on common diseases defined as having a population prevalence of 2% or greater over the 11-year study period, leading to 108 diseases for analysis. More detailed information on these diseases is available in Appendix A.1. Data processing is also presented in Figure 2.3.



(a) Prevalence



(b) Average LOS

Figure 2.4: Top 10 diseases with the highest prevalence and average LOS

To have an overview of the analyzed data and diseases, Figure 2.4a shows the top 10 diseases with the highest prevalence (population percentage of non-zero LOS), and Figure 2.4b shows the

top 10 diseases with the highest average LOS among patients with non-zero LOS. The two plots include both overall (summarized over 11 years) and yearly values. Prevalence and average LOS depict from two different perspectives the medical burden from hospitalization. It is observed from Figure 2.4 that chronic conditions tend to have higher prevalence, and acute conditions tend to have higher average LOS. Congestive heart failure; nonhypertensive is the only condition that ranks in the top 10 for both prevalence and average LOS. Temporal variations are also clearly observed from Figure 2.4. While some diseases consistently rank high, others may only appear in the top 10 lists for a few years. For example, hypertension with complications and secondary hypertension ranks out of the top 10 in prevalence in early years, but rises in rank rapidly from 2016.

There have been extensive studies on LOS using the Medicare data. In the analysis of individual diseases, for example, Bueno et al. [54] depicted the trends of LOS and its correlation with other short-term outcomes of Medicare patients hospitalized for heart failure. In the analysis of diseases overall, for example, Jencks et al. [55] investigated the correlation between LOS and hospital-associated mortality for all Medicare patients. As described in the above section, the proposed analysis fundamentally differs from these two types of analysis.

2.3 Methods

Our goal is to construct a disease network under which each node corresponds to the LOS of one disease. Two nodes are connected with an edge if the LOS values of the corresponding diseases are "connected". In general, there are two families of network constructions: unconditional and conditional. In unconditional analysis, when investigating whether two diseases are interconnected, the other diseases are "ignored". Most of the existing HDNs, including the gene-centric HDN by Goh et al. [24], phenotypic HDN by Hidalgo et al. [27], and treatment costs network by Ma et al. [31], belong to this category. In conditional analysis, the goal is to quantify whether two diseases are interconnected *conditional on* the other diseases. As established in the literature [56], conditional networks can be more informative and, at the same time, more challenging.

Graphical models have been widely used in the literature for modeling conditional dependence among a set of variables. A graphical model is associated with a graph $G = (V; E)$, where the node set V represents the variables of interest, and the edge set E encodes the corresponding conditional

dependence relationships for the node set V. Let $nb(v) = \{w \in V : \{w, v\} \in E\}$ be the neighbors of node $v \in V$, then X_v – measurement of this node – is conditionally independent of its non-neighbors $X_{V \setminus (nb(v) \cup v)}$ given its neighbors $X_{nb(v)}$. In a sense, a graphical model is a set of multivariate joint distributions that exhibit certain conditional dependence. In the statistical literature, the most popular and developed graphical model is perhaps the Gaussian Graphical Model (GGM), which assumes that the nodes have a multivariate normal distribution, and two nodes are identified to be conditionally independent if the corresponding element in the precision matrix (i.e., the inverse of the covariance matrix) is zero. Another commonly used approach, the Ising model, applies to the case where all nodes are binary. With the consideration that both assumptions are too restrictive, multiple efforts have been taken, assuming alternative data distributions. For example Yang et al. [57] assumed that the node-wise conditional distributions arise from some commonly used univariate exponential families. Voorman et al. [58] proposed a semi-parametric method and estimated the graph structure with joint additive models, which allow the conditional means to take on an arbitrary additive form. Fellinghauer et al. [59] used a Graphical Random Forests method to estimate the pairwise conditional independence relationships among mixed-type, i.e. continuous and discrete, variables. Several normal copula and nonparanormal models [60] have also been developed to accommodate non-Gaussian data. Despite extensive effort, however, a closer examination of the existing methods shows that *they are all not directly applicable to zero-inflated data*, as observed in Figure 2.1.

2.3.1 Modeling

To accommodate the zero-inflation nature of LOS data, and also motivated by individual-disease LOS analysis [11], we propose employing a multivariate Hurdle model [61]. In general, univariate Hurdle models arise from modification of a density through exclusion of points in the support and assignment of positive masses to these points. In our case, to accommodate zero-inflation, the origin is excluded from the density. Let $v_y = I\{y \neq 0\}$. Then the Hurdle model derived from a Normal distribution with mean μ and precision η^2 has density:

$$f(y) = exp\{v_y[1/2log(\eta^2/(2\pi)) + logp/(1-p) - \mu^2\eta^2/2] + y\mu\eta^2 - y^2\eta^2/2 + log(1-p)\},$$

where $p = P(v_y = 1) \in (0, 1)$ is the probability of observing a non-zero value.

Built on the univariate Hurdle model, under the multivariate Hurdle model, each conditional distribution of one node (disease) given the others is a mixture of a point mass at zero and a normal distribution. Here we note that the original LOS measurements are count data. However, with the splitting (between the primary and secondary diagnoses) and quite "spread" measurements, it is reasonable to model LOS as continuous distributions.

Let $\mathbf{v} = (v_1, ..., v_d)^T$ be a d-dimensional binary vector, where d is the number of diseases, and v_j indicates whether a subject has inpatient treatment for the jth disease. Assume that $\mathbf{V}$ follows an Ising model with joint probability:

$$p_{\mathbf{v}} \equiv P(\mathbf{V} = \mathbf{v}) \propto exp(\mathbf{v}^T \mathbf{G} \mathbf{v}), \quad \mathbf{v} \in \{0, 1\}^d,$$

where $\mathbf{G}$ is a symmetric interaction matrix in $\mathbb{R}^{d \times d}$, and zero entries indicate conditional independence in the occurrence of disease treatment.

Let $\mathbf{y} \in \mathbb{R}^d$ be the vector of LOS of the d diseases. Logarithm transformation for non-zero data entries is conducted. With a slight abuse of notation, we still use the same notation for the transformed LOS values, for which normality can be assumed. Consider the distribution:

$$\mathbf{Y}|\mathbf{V} \sim N(\mu(\mathbf{v}), \Sigma(\mathbf{v}))$$

with mean vector $\mu(\mathbf{v}) \in \mathbb{R}^2$ and covariance matrix $\Sigma(\mathbf{v}) \in PD(\mathbf{v})$. Then the conditional distribution of $\mathbf{Y}$ given $\mathbf{V} = \mathbf{v}$ has log-density:

$$logf(\mathbf{y}|\mathbf{V} = \mathbf{v}) = \mathbf{v}^T \mathbf{H} \mathbf{y} - \frac{1}{2} \mathbf{y}^T \mathbf{K} \mathbf{y} - C'(\mathbf{H}, \mathbf{K}), \quad \mathbf{y} \in \mathbb{R}^{\mathbf{v}},$$

where $\mathbf{H}$ and $\mathbf{K}$ are two $d \times d$ interaction matrices that do not vary with $\mathbf{v}$ and measure the dependence between the presence of treatment and LOS and between LOS values respectively, $C'(\mathbf{H}, \mathbf{K})$ is a normalization constant, and $\mathbb{R}^{\mathbf{v}} = \prod_{j=1}^{d} \mathbb{R}^{v_j}$ with $\mathbb{R}^{\mathbf{0}} = \{0\}$. $\mathbf{K}$ is symmetric and positive definite, but there is no constraint on $\mathbf{H}$. The joint log-density of $\mathbf{Y}$ and $\mathbf{V}$ is:

$$logf(\mathbf{y}, \mathbf{v}) = \mathbf{v}^T \mathbf{G} \mathbf{v} + \mathbf{v}^T \mathbf{H} \mathbf{y} - \frac{1}{2} \mathbf{y}^T \mathbf{K} \mathbf{y} - C(\mathbf{G}, \mathbf{H}, \mathbf{K}), \quad y \in \mathbb{R}^d.$$

To quantify pairwise conditional interconnections, the conditional density is derived from the joint density as follows. For a fixed coordinate $b \in \{1, ..., d\}$, define its complement $A = \{1, ..., d\} \setminus \{b\}$. Noting that $v_i y_i = y_i$ and $v_i^2 = v_i$, we have the conditional log-density of Y_b given Y_A as:

$$log f_{[b|A]}(y_b, v_b | \mathbf{y}_A, \mathbf{v}_A) = v_b g_{[b|A]} + y_b h_{[b|A]} - \frac{1}{2} y_b^2 k_{[b|A]} - C_{[b|A]},$$

where the normalization constant $C_{[b|A]}$ does not depend on (y_b, v_b) and

$$g_{[b|A]} = g_{bb} + 2\mathbf{g}_{bA}\mathbf{v}_A + \mathbf{h}_{bA}\mathbf{y}_A,\ h_{[b|A]} = h_{bb} + \mathbf{h}_{Ab}^T\mathbf{v}_A + \mathbf{k}_{bA}\mathbf{y}_A,\ k_{[b|A]} = k_{bb}.$$

Based on the conditional distribution, elements in the three interaction matrices $\mathbf{G}$, $\mathbf{H}$, and $\mathbf{K}$, which quantify pairwise interconnections between diseases conditional on all other diseases, can be obtained. Following the literature [61], we conclude conditional independence between diseases i and j if and only if all four ij interaction elements are zero. That is,

$$g_{ij} = h_{ij} = h_{ji} = k_{ij} = 0.$$

2.3.2 Estimation

Considering the fact that not all diseases are interconnected, and to generate sparse and interpretable networks, we propose conducting sparse estimation. In particular, we adopt the penalization technique [56], which has been widely used in structure learning of the GGM and other network constructions. It has also been used in the analysis of clinical treatment measures of individual diseases [62]. More specifically, for a fixed coordinate $b \in \{1, ..., d\}$, consider the conditional distribution of Y_b given the other variables in Y_A for $A = \{1, ..., d\} \setminus \{b\}$. For $a \in A$, define the parameter vector $\boldsymbol{\theta}_a = (g_{ba}, h_{ba}, h_{ab}, k_{ba})^T$. Then $Y_b \perp\!\!\!\perp Y_a | \mathbf{Y}_{A\setminus\{a\}}$ if and only if $\boldsymbol{\theta}_a = \mathbf{0}$. Consider the penalized objective function:

$$log f_{[b|A]}(y_b, v_b | \mathbf{y}_A, \mathbf{v}_A) - P_\lambda(\boldsymbol{\theta}),$$

where $P_\lambda(\boldsymbol{\theta}) = \lambda ||\boldsymbol{\theta}_a||_1$ is the LASSO penalty and $\lambda > 0$ is the data-dependent tuning parameter. Note that the above estimation is defined for each coordinate. The estimations of different coordi-

nates are connected via sharing the same tuning parameter, ensuring a comparable ground for all nodes.

Computation The penalized objective function is the sum of a smooth convex function $g(\boldsymbol{\theta})$ and a non-differentiable convex function $h(\boldsymbol{\theta})$. Generically, the optimization problem $min_{\boldsymbol{\theta}} L(\boldsymbol{\theta}) = g(\boldsymbol{\theta}) + h(\boldsymbol{\theta})$ can be tackled using the proximal gradient descent technique [61]. Specifically, the proximity operator for $\boldsymbol{\theta}$ related to t and h is defined as:

$$prox_{h,t}(\boldsymbol{\theta}) = \arg\min_{\mathbf{x}} \frac{1}{2t}\|\boldsymbol{\theta} - \mathbf{x}\|_2^2 + h(\boldsymbol{\theta}),$$

where t is the step size and h is the LASSO penalty. Then the update rule can be described as

$$\boldsymbol{\theta}^{k+1} = prox_{h,t_k}(\boldsymbol{\theta}^k - t_k \nabla g(\boldsymbol{\theta}^k)).$$

For a fixed coordinate $b \in \{1, ..., d\}$, let $\boldsymbol{\theta}_0 = (g_{bb}, h_{bb}, k_{bb})^T$, $\boldsymbol{\theta}_a = (g_{ba}, h_{ba}, h_{ab}, k_{ba})^T$, and $\boldsymbol{\theta} = (\boldsymbol{\theta}_0^T, \boldsymbol{\theta}_a^T)^T$. The gradients of the log-likelihood function $g(\boldsymbol{\theta}) = logf_{[b|A]}(y_b, v_b|\mathbf{y}_A, \mathbf{v}_A)$ with respect to $\boldsymbol{\theta}$ can be calculated using the chain rule. As such, the proximal gradient descent-based neighborhood selection procedure can be conducted for each node. The estimates for $\mathbf{G}$, $\mathbf{H}$, and $\mathbf{K}$ can thus be obtained. More details are available from the authors as well as realized in our code, which is available at *www.github.com/shuanggema*. The proposed approach involves the tuning parameter λ. It is chosen using cross-validation, which has been adopted in penalized GGM and other network analyses.

With a large sample size, a large number of diseases, and complex model structures (with multiple interconnection matrices), computation is unfortunately expensive. With a fixed tuning parameter, analysis takes about two hours. Luckily, estimation under different tunings can be run in a parallel manner to reduce computer time. For larger data, further parallelization may be needed to make the analysis affordable.

2.3.3 Analysis of network properties

In a sense, the parameter estimates and model obtained above can fully describe data properties. However, they may not have lucid interpretations. In network analysis, key properties are usually

summarized through adjacency, connectivity, module/hub, and other measures [56, 63], which can provide more meaningful understandings than model parameters only. Similar analysis has been conducted in the phenotypic and treatment costs HDNs [28, 31].

Adjacency matrix The most fundamental property of a network is reflected in the $d \times d$ adjacency matrix, whose elements describe whether a pair of diseases are interconnected conditional on the others (i.e., whether a linking edge exists). In our model, we conclude conditional independence between diseases i and j if and only if $\mathbf{G}_{ij} = \mathbf{H}_{ij} = \mathbf{H}_{ji} = \mathbf{K}_{ij} = 0$. More specifically, consider the estimated interaction matrices $\hat{\mathbf{G}}$, $\hat{\mathbf{H}}_1$ (the estimate of $\mathbf{H}$ fitted by column), $\hat{\mathbf{H}}_2$ (the estimate of $\mathbf{H}$ fitted by row), and $\hat{\mathbf{K}}$. The adjacency matrix $\mathbf{A} = [a_{ij}]$ is defined as

$$a_{ij} = \begin{cases} 0, & \text{if } \hat{G}_{ij} < \tau \ \& \ \hat{H}_{1ij} < \tau \ \& \ \hat{H}_{2ij} < \tau \ \& \ \hat{K}_{ij} < \tau \\ 1, & \text{otherwise} \end{cases}$$

where we further impose a threshold τ to remove spurious small interconnections and generate sparser and more interpretable networks. τ can be determined data-dependently (for example, using cross validation) or based on the specific contexts.

Connectivity For node (disease) $i \in \{1, ..., d\}$, its connectivity is defined as

$$K_i = \sum_{j \neq i} a_{ij},$$

which measures how tightly it is connected to the other nodes. For many networks including the HDNs [28,31], it has been suggested that nodes with higher connectivity values are more important in a network sense since they have a higher impact overall.

Module Modules have also been referred to as network communities or clusters in the literature. A module consists of tightly interconnected nodes, and different modules are relatively weakly interconnected. It has been suggested that nodes within the same module tend to "behave similarly" and should be considered as a group. In module construction, first consider the topological overlap matrix (TOM), whose (i, j)th element is defined as

$$TOM_{ij} = \frac{l_{ij} + a_{ij}}{minK_iK_j + 1 - a_{ij}},$$

where $l_{ij} = \sum_u a_{iu}a_{uj}$ measures how many neighbors that nodes i and j share. Accordingly, the TOM-based dissimilarity matrix is defined as dissTOM = 1-TOM [63]. Based on dissTOM, which measures the dissimilarity of any pairs of diseases, modules can then be constructed by hierarchical clustering with a dynamic tree cutting approach [63]. After module construction, for a node, its intramodular connectivity, which is connectivity limited to the module it belongs to, can be computed to reflect "more local" connection properties.

2.4 Data analysis

2.4.1 Network properties

We first consider the "whole picture" with all 11 years combined. With 108 nodes (diseases), the resulted network has 1,049 edges, with a disease on average connected to 19.4 others. The network is graphically presented in Figure 2.5 using *Gephi*. In the plot, a line represents an edge, the size of a node is proportional to its connectivity, and the color of a node represents its module membership. Similar constructions and visualizations have been done for individual years. Additional details are provided in Appendix A.2. Figure 2.5 shows extensive interconnections among nodes. There are only two diseases that are not linked to any other diseases. This is expected since multimorbidity is the most common clinical picture of the elderly population [64]. Simply eyeballing the network suggests that it differs fundamentally from the molecular HDN in Goh et al. [24], phenotypic HDN in Hidalgo et al. [27], and treatment costs HDN in Ma et al. [31]. This is attributable to differences in disease definition, approach in network construction, population characteristics, and outcome measure.

Connectivity Figure 2.6 shows the top 10 diseases with the highest overall and year-specific connectivity values. Comparing it to Figure 2.4 shows that the diseases with the highest connectivity also tend to have the highest prevalence and possibly the highest LOS. Loosely speaking, these diseases are the most important. More specifically, all diseases in Figure 2.6 are well-known common diseases in the elderly population, which provides some support to the validity of our network analysis. For example, cardiovascular disorders, including congestive heart failure, cardiac dysrhythmias, and coronary atherosclerosis, are estimated to be the most common chronic conditions in the population aged 65 years and above [64]. Systemic diseases like fluid and electrolyte disorders,

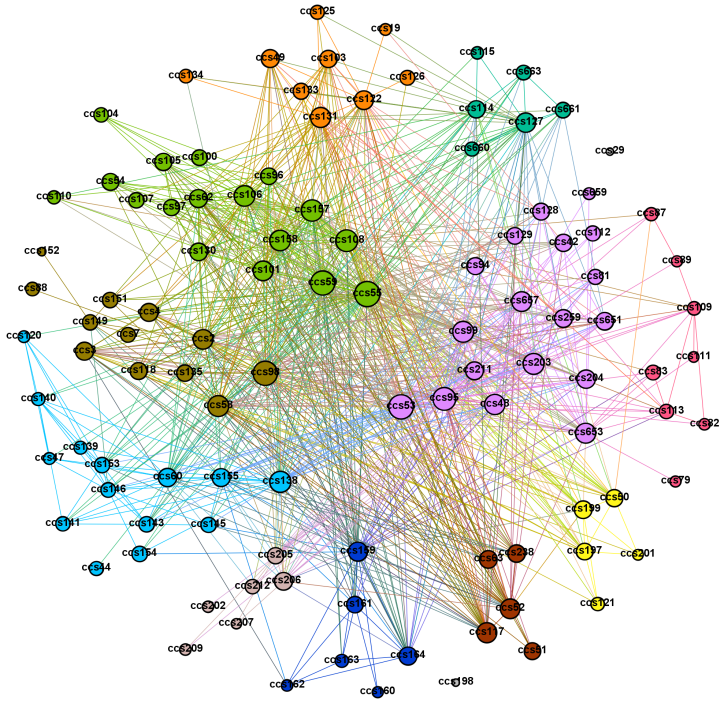


Figure 2.5: Disease LOS network for the period of January 2008 to December 2018

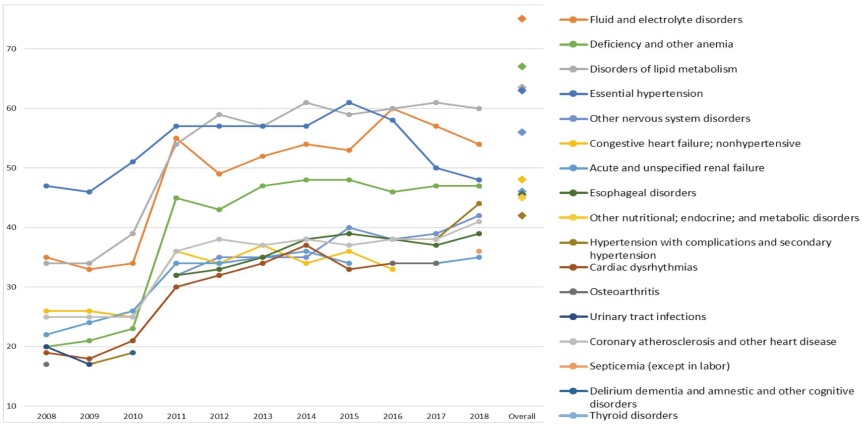


Figure 2.6: Top 10 diseases with the highest overall and yearly connectivity

nervous system disorders, and nutritional, endocrine, and metabolic disorders are also common in the elderly populations and are expected to have high connectivity because they affect the general functioning of the entire body [65]. From a treatment perspective, diseases such as fluid and electrolyte disorders, acute and unspecified renal failure, and hypertension with complications and secondary hypertension require extensive care through hospitalizations [66, 67] and are expected to have high LOS connectivity. In contrast, diseases such as essential hypertension, disorders of

lipid metabolism, osteoarthritis, and urinary tract infections often do not require hospitalization but have high connectivity in the constructed LOS network, indicating them to be important comorbidities that are connected with other diseases that require higher LOS. For example, serious osteoarthritis can lead to fractures [68], disorders of lipid metabolism increase the risk of coronary artery disease [69], and urinary tract infections are highly correlated to septicemia especially in elderly females [70]. While some individual relationships between a few selected diseases have been studied in the literature, our results can be informative by looking at all diseases "globally" and also examining disease pairs less/not investigated. For example, from etiological studies, secondary hypertension has a closer connection to many other heart and cerebral diseases, urinary diseases, and endocrine diseases, compared to primary hypertension [71]. However, we show that in the LOS network, primary hypertension has a higher connectivity. Hypertension is a complicated condition that individuals generally have different pathogenesis, progression, etiology, and pathophysiology, and studies on it remain inclusive [67]. Our analysis results may provide additional insights and also suggest potential future research directions related to hypertension. Such results cannot be obtained from the individual-disease based.

Module A total of 11 modules are identified, with sizes ranging from 18 to 5. In Figure 2.5, different modules are distinguished using colors. More details on year-specific results are provided in Appendix A.2. By construction, modules should consist of tightly interconnected diseases. An enrichment analysis is conducted to examine the representative diseases of different modules. It is found that the 11 modules are enriched with the following diseases: nervous system diseases and comorbidities (pink), cardiovascular diseases and complications (light green), gastrointestinal diseases (celeste), infectious diseases (yellowish brown), respiratory system diseases (orange), central nervous system diseases (vermeil), substance abuse disorders and complications (aqua green), osteoarticular diseases (champagne), urinary system diseases (blue), skin and arthritis infections (brown), and residual diseases (yellow), respectively.

Module construction can provide an alternative way of disease characterization and classification. Classic disease classifications are based on organ, symptom, and phenotype. It has been recognized that new classifications are needed to serve new purposes [72]. For example, gene-centric disease classifications have been developed [24] and integrated with classic classifications. In our LOS disease network, which is closer to clinical practice, diseases form modules not only

because they share common genetic and environmental risk factors but also because they have shared treatment strategies. Take the pink module as an example, which is enriched with nervous system diseases. Osteoarthritis, other connective tissue disease, and other non-traumatic joint disorders have a relatively "long distance" from nervous system diseases based on etiological studies. However, there is an increasing trend of listing these disorders together with neurological disorders as an indicator of disease progression and poor prognosis, especially in the elderly population [73]. To be more specific, it is common for elderly patients with multiple chronic diseases to be mobility-impaired. This could cause them to develop various nervous system diseases as well as osteoarthritis, other connective tissue disease, and other non-traumatic joint disorders.

Within each module, the hub disease is defined as the one with the highest intramodular connectivity. Generically in network analysis, hubs may play central roles in their respective modules and should be "targeted" first [56]. The hub diseases for the 11 modules are: other nervous system disorders (pink), coronary atherosclerosis and other heart disease (light green), diverticulosis and diverticulitis (celeste), septicemia (yellowish brown), pneumonia (orange), acute cerebrovascular disease (vermeil), chronic obstructive pulmonary disease and bronchiectasis (aqua green), spondylosis; intervertebral disc disorders; other back problems (champagne), urinary tract infections (blue), skin and subcutaneous tissue infections (brown), and nutritional deficiencies (yellow), respectively. All of these diseases have been suggested in the literature as having essential importance for older adults. We note that connectivity and intramodular connectivity are not the same concept. In particular, connectivity can be potentially affected by a large number of weak connections, whereas intramodular connectivity better reflects "local properties". In our analysis, for example, pneumonia has a low overall connectivity but the highest intramodular connectivity in the module enriched with respiratory system diseases.

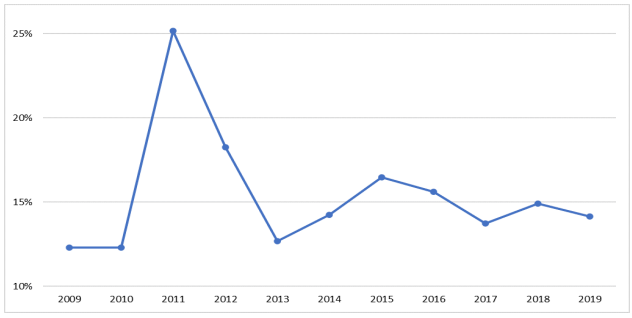
2.4.2 Temporal variation

It is easy to observe variations over time from the above summary statistics. Here we move on with conducting the above network analysis for each year separately and examining differences across time. We randomly select the 100,000 samples for each year separately. This way, we can ensure that representative samples are obtained for all years. We note that if we select and fix a cohort in year one, it may or may not be representative in the subsequent years. It is also noted that

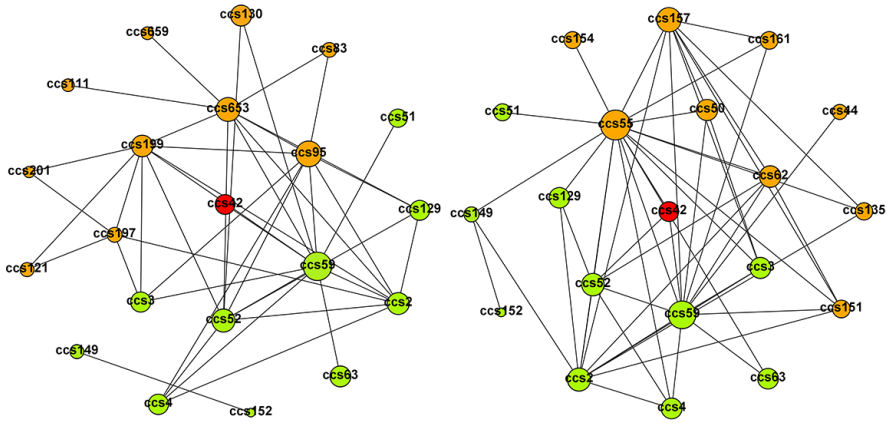
temporal variation is not presented with the molecular HDNs (as the molecular basis of most if not all diseases is stable) and ignored in most of the existing phenotypic HDNs. Temporal variation is also examined with the treatment costs HDN [31].

Connectivity Figure 2.6 shows the top 10 diseases with the highest yearly connectivity values. While the structures of high-connectivity diseases for adjacent years are similar, the overall differences are obvious. Some diseases, such as essential hypertension, fluid and electrolyte disorders, and disorders lipid metabolism, are stable with high connectivity in almost all years. In contrast, some other diseases, such as urinary tract infections, osteoarthritis, and septicemia, only appear in the top 10 lists for a few years. Moreover, a jump in individual disease connectivity between 2010 and 2011 is observed, indicating more intensive interconnections between diseases in the later year. A closer examination of data and literature suggests that this is at least partly caused by the change in CMS' electronic transaction standards that hospitals use to submit Medicare claims. Starting in January 2011, institutional providers were able to enter up to 25 diagnosis codes for a single claim, while previously only 9 were allowed [74]. In the retrieved dataset, the average number of per claim diagnosis codes increases from 8.03 in 2010 to 13.93 in 2011. This can lead to a higher number of comorbidities per patient and hence a higher disease connectivity. We note that, for the overall network constructed using all 11 years of data combined, the numbers of comorbidities are accrued for the entire 11 years of study period, and the policy change in 2011 might not make a big impact. This is confirmed by repeating the overall network analysis using only the first 9 diagnosis codes per claim. Assuming physicians enter important diagnoses first, the first 9 existing codes would be the 9 codes entered if the coding policy had not changed. Therefore, using only the first 9 diagnosis codes mimic the situation that the policy change in 2011 never happened. Comparing this with the overall network constructed in Section 2.4.1, we observe only ignorable changes in the network structure (details omitted and available from the authors).

Module To examine yearly variations in disease modules, we compare pairwise disease module relationships. More specifically, if two diseases are in the same module for one year but not for the later year, we say their disease module relationship has changed. Figure 2.7a shows the yearly percentages of pairs of diseases that have changed their module relationships. Overall, we observe a moderate degree of changes, except for year 2011 where 26% of the pairwise module relationships have changed. The big change in 2011 is likely also due to the change in CMS'



(a) Yearly changes in disease module relationships



ccs42	Secondary malignancies	ccs135	Intestinal infection	ccs2	Septicemia (except in labor)
ccs44	Neoplasms of unspecified nature or uncertain behavior	ccs151	Other liver diseases	ccs3	Bacterial infection; unspecified site
ccs50	Diabetes mellitus with complications	ccs154	Noninfectious gastroenteritis	ccs4	Mycoses
ccs55	Fluid and electrolyte disorders	ccs157	Acute and unspecified renal failure	ccs51	Other endocrine disorders
ccs62	Coagulation and hemorrhagic disorders	ccs161	Other diseases of kidney and ureters	ccs52	Nutritional deficiencies
ccs83	Epilepsy; convulsions	ccs197	Skin and subcutaneous tissue infections	ccs59	Deficiency and other anemia
ccs95	Other nervous system disorders	ccs199	Chronic ulcer of skin	ccs63	Diseases of white blood cells
ccs111	Other and ill-defined cerebrovascular disease	ccs201	Infective arthritis and osteomyelitis	ccs129	Aspiration pneumonitis; food/vomitus
ccs121	Other diseases of veins and lymphatics	ccs653	Delirium dementia and amnestic and other cognitive disorders	ccs149	Biliary tract disease
ccs130	Pleurisy; pneumothorax; pulmonary collapse	ccs659	Schizophrenia and other psychotic disorders	ccs152	Pancreatic disorders (not diabetes)

(b) Modules that contain secondary malignancies in 2014 (left) and 2018 (right)

Figure 2.7: Temporal variations of module structure

electronic transaction standards mentioned earlier. Variations in module relationships can be caused by multiple reasons, such as changes in risk factors, progress in disease diagnosis, and advances in treatment strategies. As an example of how disease module memberships may change, Figure 2.7b shows the modules that secondary malignancies (ccs42) belongs to in 2014 and 2018. Thanks to the innovative cancer immunotherapy technique – immune checkpoint inhibitors (ICPIs), the landscape of advanced cancer treatment has significantly changed in the last decade. ICPIs has shown clinically significant antitumor response in multiple cancer types [75]. Since 2015, ICPIs has been widely used upon the Food and Drug Administration's approvals and led to significant improvement in cancer survival [75]. Comparing the module that secondary malignancies belongs

to in 2014 with that in 2018, we see a big change in the module structure. In Figure 2.7b, common disease shared by the two modules are colored in green and different diseases are colored in orange. In 2018, secondary malignancies is no longer clustered with multiple types of skin infections (ccs197 and ccs199) and neuropsychiatric disorders (ccs95, ccs 653, and ccs659), but is clustered with several intestinal (ccs135 and ccs154), liver (ccs151), and renal disorders (ccs157 and ccs161), all of which are potential side effects from ICPIs [76].

Remarks Under simpler settings, there are more sophisticated approaches for studying temporal variations, for example, the time-varying coefficient model. We suspect that it is possible to couple such techniques with the proposed HDN analysis. However, significant new developments may be needed. The above examination is simple and can be easier to interpret.

2.4.3 Sensitivity analysis

In the above analysis, for claims with multiple diagnosis codes, we attribute 60% of the LOS to the primary diagnosis, and the rest 40% evenly to all secondary diagnoses. We would like to note again that this assignment, although slightly subjective, is sensible, and that there is still no consensus on splitting LOS (and other treatment measures) among diagnoses. To examine the effects of this allocation on the findings, we re-conduct the above analysis with all 11 years data combined using allocations with 70% or 50% to the primary diagnosis and the rest evenly to all secondary diagnoses. Only ignorable changes in the network structure are observed (details omitted and available from the authors).

2.4.4 Alternative analysis

For the same CCS diseases, in Appendix A.3, we construct the molecular HDN (using information extracted from *https://phewascatalog.org/phewas*) and phenotypic HDN (using the same data as analyzed above). It is observed that these two HDNs and their key properties differ significantly from those of the proposed, which further justifies that the proposed disease interconnection analysis is warranted beyond the existing HDN studies.

2.5 Simulation

Simulation is conducted to get more insights into performance of the proposed modeling and estimation approach. The sample size is fixed as 10,000, and there are 100 nodes (diseases). Note that the real data has a much larger sample size, and hence more reliable estimation is expected. We consider a variety of settings with different graph topologies, sparsity levels, and parametric models. Specifically, with the chain graph topology, the interaction matrices $\mathbf{G}$, $\mathbf{H}$, and $\mathbf{K}$ have banded structures. The numbers of nonzero diagonals depend on the specified sparsity levels. For example, a 1.5% sparsity results in a band width of three, and a 5% sparsity results in a band width of seven. With the random graph topology, the interaction matrices $\mathbf{G}$, $\mathbf{H}$, and $\mathbf{K}$ have certain numbers of random off-diagonal entries being nonzero, where the numbers are determined based on the prespecified sparsity levels. For parametric models, we consider the Hurdle model, Hurdle model under contamination with t_8 noise, and Gaussian/logistic selection model. The Hurdle model is said to be $\mathbf{G}-minimal$ when only $\mathbf{G}$ has nonzero off-diagonal elements and $\mathbf{H}$ and $\mathbf{K}$ are diagonal matrices. In this case, the Hurdle model reduces to the logistic/Ising model. The Hurdle model is said to be *complete* if for each edge present in the graph, all of the corresponding entries in each of the three interaction matrices are nonzero [61]. The Gaussian/logistic selection model is defined as:

$$\tilde{\mathbf{Y}} \sim \mathcal{N}(\mu, \mathbf{K}),\ P(\tilde{V}_j|\tilde{\mathbf{Y}} = \tilde{\mathbf{y}}) = logit(a + b\tilde{y}_j),\ \mathbf{Y} = \tilde{\mathbf{Y}}\tilde{\mathbf{V}},$$

where a and b can be modified to achieve different levels of zero in $\mathbf{Y}$. To make the simulated data close to the real data (which has 60% - 98% zeros in $\mathbf{Y}$), we choose the off-diagonal values equal to -2 for $\mathbf{G}$, -1 for $\mathbf{H}$, and 0.5 for $\mathbf{K}$ when applicable. a is randomly selected from Uniform[-2, 0] and b equals 1. Detailed simulation settings are described in Table 3.1.

For each setting, 100 replicates are simulated. In our analysis, the key is the identification of edges. In Table 3.2, we summarize the mean (standard deviation) percentage of false negative (FN) and false positive (FP) rates. Satisfactory performance is observed. With the much larger sample size of the real data, the simulation provides strong confidence in the validity of our findings.

1.	Hurdle model with a **G**-minimal chain graph and 1.5% sparsity.
2.	Hurdle model with a $\mathbf{G}-\mathbf{H}-\mathbf{K}$-complete chain graph and 1.5% sparsity.
3.	Hurdle model with a **G**-minimal chain graph and 5% sparsity.
4.	Hurdle model with a $\mathbf{G}-\mathbf{H}-\mathbf{K}$-complete chain graph and 5% sparsity.
5.	Hurdle model with a **G**-minimal random graph and 5% sparsity.
6.	Hurdle model with a $\mathbf{G}-\mathbf{H}-\mathbf{K}$-complete random graph and 5% sparsity.
7-12.	The same as 1-6, respectively, with nonzero observations contaminated with t_8 noise.
13.	Gaussian/logistic selection with a diagonal interaction matrix **K**.
14.	Gaussian/logistic selection with a chain graph **K** and 1.5% sparsity.
15.	Gaussian/logistic selection with a chain graph **K** and 5% sparsity.
16.	Gaussian/logistic selection with a random graph **K** and 5% sparsity.

Table 2.1: Simulation settings

Setting	FN(sd)	FP(sd)	Setting	FN(sd)	FP(sd)
1	0(0)	0(0)	9	0(0)	0(0)
2	0(0)	0(0)	10	0(0)	2.55%(0.003)
3	0(0)	0(0)	11	2.98%(0.021)	6.09%(0.009)
4	0(0)	1.08%(0.002)	12	1.92%(0.021)	2.30%(0.006)
5	3.19%(0.021)	2.21%(0.003)	13	0(0)	0(0)
6	2.47%(0.026)	0.74%(0.002)	14	0.77%(0.020)	1.56%(0.005)
7	0(0)	0.45%(0.008)	15	3.67%(0.015)	1.81%(0.003)
8	0(0)	0.56%(0.003)	16	7.54%(0.016)	8.44%(0.004)

Table 2.2: Simulation: mean (sd) percentage of FN and FP

2.6 Discussion

In this study, we have developed a HDN built on diseases' inpatient LOS. To the best of our knowledge, this is the first of its kind. This analysis can complement the existing individual-disease and overall LOS analyses as well as the molecular and phenotypic HDNs. This is especially true considering the significant differences in methodology and findings. As discussed in the first section, LOS analysis has important implications for managing and planning healthcare and other resources and can also be informative for reflecting intrinsic disease properties. This network analysis can provide insights into disease interconnections, assist more efficiently manage/plan resources, and advance our understanding of diseases. It also complements the existing HDN analysis, in particular the recent disease treatment costs HDN. It will be important to develop policymaking based on our findings. However, that is beyond the scope of this article and will be postponed to future research. Beyond delivering the first disease LOS network and its snapshot and dynamic properties, this study has also developed a network analysis approach which may have other applications.

This study inevitably has limitations. LOS only reflects part of the treatment picture. Many disease episodes are treated outpatient. Chronic diseases are often handled with drug refillings that do not involve any hospital treatment. It will be of interest to expand the analysis scope to accommodate other types of treatment. Nevertheless, we note that inpatient treatment, as the most serious type of treatment, has its own unique value and has been the focus of a large number of studies. We have considered the whole Medicare population. It will also be of interest to conduct stratified analysis to better accommodate heterogeneity. As in many other HDN analysis, our analysis can only infer associations (with undirected networks). Causal analysis would demand significant additional data. Our examination suggests that the findings are biomedically sensible to a large extent. However, we are unable to examine all findings considering the large number of diseases and interconnections. It will also be of interest to examine additional data and provide more interpretations on the findings.

Chapter 3

High-Dimensional Clinical Treatment Human Disease Network: Analysis of the Medicare Inpatient Length of Stay and Readmission Data

3.1 Introduction

Clinical treatment outcomes are defined to measure healthcare quality (e.g., mortality, readmission, and complication) and efficiency (e.g., length of stay (LOS), outpatient utilization, and medical costs). For most medical conditions, these outcomes have been extensively studied. Outcome analysis helps healthcare practitioners carry out more effective and efficient practices, hence further improve outcomes. For example, Piedmont Healthcare's evidence-based care standardization for pneumonia patients resulted in a 56.5% relative reduction in the pneumonia mortality rate and a 9.3% relative reduction in LOS [7]. By improving the analytic platform and advancing its applications, the University of Texas Medical Branch achieved a 14.5% relative decrease in their 30-day all-cause readmission rate, resulting in $1.9 million in cost reduction [8]. Therefore, understanding outcomes is essential in providing patient-centered and value-based healthcare services.

The analysis of clinical treatment outcomes can be generally classified into two families. The

first family focuses on a single disease. For example, Feng and Li [11] analyzed LOS for patients with ischaemic heart disease and identified important risk factors. They observed the zero-inflation nature of LOS data and conducted a two-part regression analysis to accommodate zero-inflation. Healthcare outcome data is generally zero-inflated since, for most conditions, only a small portion of the population would be susceptible and have clinical treatments. The second family focuses on a general population and studies all diseases combined. For example, using conditional logistic regression clustered by hospital and generalized estimating equations, Jung et al. [18] compared all diseases combined readmission rates in two populations: Medicare Advantage and Medicare Fee-For-Service. Even though both families are considered effective, they ignore the complex interconnections among human diseases. By examining LOS correlations within the Diagnosis Related Groups, Berki et al. [22] indicated a need to analyze and understand LOS behaviors for multiple diseases. Rinne et al. [23] also identified correlations between readmission rates of chronic obstructive pulmonary disease and other medical conditions (e.g., heart failure, acute myocardial infarction (AMI), pneumonia, and stroke). However, these studies are generally limited to a few pre-selected and closely related diseases.

Another major characteristic of analyzing clinical outcomes is that the practices generally focus on one single outcome. For instance, Huling et al. [16] proposed a framework to estimate individualized treatment rules targeting to optimize medical costs. They observed the zero-inflation nature of medical costs data and developed an approach based on the two-part semicontinuous modeling. Chatterjee et al. [17] proposed group regularization for zero-inflated count regression models using a least-squares approximation of the mixture likelihood and applied it to number of doctor office visits data. Noting that many clinical treatments have complicated effects that can only be effectively captured on multiple outcome scales, studies have focused on the condition-specific correlation between multiple outcomes. With interrupted time-series and survival analysis, Gupta et al. [13] identified a negative relationship between readmission rate and mortality for Medicare patients hospitalized with heart failure. With multivariate regression analysis, Rinne et al. [14] found that longer LOS was associated with a higher risk of readmission for veterans with chronic obstructive pulmonary disease. However, analysis on multiple outcomes is again limited to a single disease and ignore the complex interconnections among different diseases. To our best knowledge, there is a lack of work that jointly analyzes multiple outcomes of multiple diseases.

This article fills the research gaps by developing a clinical treatment human disease network (HDN) that quantifies pan-disease level interconnections on multiple outcomes. This is partially motivated by the emerging HDN analysis in the past two decades. Existing HDN analyses fall into three families. The first family is referred to as gene-centric such that diseases are interconnected if they share common genetic risk factors. Goh et al. [24] was the first to develop the concept of HDN by connecting two diseases if they are associated with the same gene. Follow-up studies include Li and Agarwal [26] and Barabasi et al. [48], in which links between human diseases represent various molecular relationships. The second family is based on disease phenotypes such that two diseases are connected if their phenotypes are correlated (e.g., comorbidity). Hidalgo et al. [27] was the first to build a phenotypic HDN upon patients' clinical histories. Follow-up studies include Roque et al. [29], Jiang et al. [28], and others. The third family, which is relatively new, examines disease interconnections based on healthcare outcomes. Ma et al. [31] was the first to construct a clinical treatment HDN by analyzing disease interconnections in medical costs. Mei et al. [77] expanded the analysis scope to disease interconnections in LOS. Unlike genetic and phenotypic HDNs, a clinical treatment HDN captures disease interconnections from not only shared causal disease etiology, shared environmental risk factors, but also shared prevention, diagnosis, and treatment strategies.

Significantly different and advancing from the existing literature, this article builds a high-dimensional clinical treatment HDN that focuses on disease interconnections in multiple outcomes. Specifically, we build a HDN that analyzes the Medicare LOS and readmission data. A third pseudo outcome, which is a binary indicator of treatment presence, is also included in the graphical model to accommodate zero-inflation. LOS and readmission are two important outcomes that have been extensively studied. Prolonged LOS and avoidable readmissions impose huge medical burdens on both hospitals and patients. In the U.S., hospital stays cost the health system at least $377.5 billion per year [19]. While a 30-day all-cause readmission follows 18% of all hospitalizations, 37% of them are estimated to be avoidable, which cost the healthcare system $25 to $45 billion per year [78, 79]. There have been substantive efforts that target LOS and readmission as indicators of high-value care [13,46]. Our choice of outcomes is mainly motivated by the high clinical importance of LOS and readmission, but we note that the proposed method can be easily generalized to other outcomes.

As the first network tool studying disease interconnections in multiple outcomes, our approach

differs from existing clinical treatment HDNs not only in outcome measures but also in statistical techniques. Ma et al. [31] used a two-part model for the marginal distribution of disease-specific medical costs and a copula-based approach to identify unconditional pairwise interconnection between medical costs of two diseases. Mei et al. [77] employed the multivariate Hurdle model to estimate conditional pairwise correlation between LOS of two diseases. While they both accommodate the zero-inflation nature of data, the modeling techniques are not readily applicable to high-dimensional outcome data. To deal with uniquely challenging data distributions (high-dimensionality and zero-inflation), this article develops an innovative network modeling strategy based on the integrative analysis of generalized linear models. Since the proposed approach is regression-based, it has much more lucid interpretations and can be easily generalized to incorporate more outcomes of different data types.

The proposed clinical treatment HDN can inform more efficient management of hospital beds and other resources, which has public health, economic, and ethical importance. For example, as suggested in other HDN studies, interventions targeting diseases interconnected with many others can be more efficient and cost-effective. Moreover, tightly interconnected diseases should be considered together, and factors that affect a group of interconnected diseases should receive higher priority [28, 31, 48]. This study complements the existing literature in multiple aspects. First, it advances current outcome analyses by focusing on multiple outcomes at a pan-disease level. Second, it fosters complex network research. As most developed HDNs estimate unconditional relationships on one single outcome, we develop a novel graphical model to estimate conditional interconnections considering multiple outcomes. It also differs from existing molecular and phenotypic HDNs by being "closer" to clinical treatments, thus having a higher practical value. Last, it provides new insights into using the large-scaled Medicare database for outcome research and promotes a better understanding of the Medicare population.

3.2 Data

This article analyzes the Medicare inpatient data from January 2010 to December 2018. Medicare is a federal health insurance program for adults aged 65 years and above, certain younger people with disabilities, and people with end-stage renal disease (permanent kidney failure requiring dialysis or

a transplant).The Medicare data is considered an excellent source for healthcare analytics because of its universal coverage and high quality. More than 98% of adults over 65 years old in the U.S. are enrolled in the Medicare program. It is estimated that the program accounts for over 99% of death for older adults and about 40% of all inpatient discharges [2, 3]. Centers for Medicare & Medicaid Services (CMS) offers a wide range of claims data that is derived from reimbursement or payment of bills. Since it is the basis of determination of reimburse eligibility, the Medicare data is generally of high quality. It also contains broad information such as beneficiary demographics, provider information, healthcare service utilization, and medical payments, which can fulfill many research purposes.

3.2.1 Data preparation

In the analysis, we first retrieve records on 108 million inpatient treatment episodes for the period of January 2010 to December 2018, which cover Medicare inpatient services utilized by 35 million beneficiaries. Following the literature [80], we exclude patients who were less than 65 years old at admission, who had less than 30 days of post-discharge Medicare enrollment, and who died in hospital. We further select a random sample of 100,000 eligible subjects to reduce computation burden. In network analysis, the complexity of network structure determines the sample size needed for accurate estimation. We conduct simulation (details in Section 5) and conclude that the proposed method achieves satisfactory performance with a sample size of 10,000 for 100 nodes, under a variety of network structures. For real data analysis of 106 nodes (diseases), we conservatively choose a sample size of 100,000, which is expected to be sufficiently large.

For each inpatient treatment episode, LOS is calculated as the length between dates of admission and discharge. Readmission is defined following CMS' public reported measure on the 30-day risk-standardized hospital-wide readmission [80]. Specifically, the measure captures all-cause unplanned readmissions that arise from acute clinical events requiring urgent rehospitalization within 30 days of discharge. Planned readmissions, which are not a signal of care quality and are generally unavoidable, are not considered in the measure. The detailed algorithm on defining 30-day all-cause unplanned readmissions is available in CMS' measure methodology report [80]. For one index admission, if a patient had multiple unplanned admissions within 30 days of discharge, only the first is considered readmission. However, the readmission could be an index admission for subsequent

readmissions.

Diseases are defined by diagnosis codes under the International Classification of Diseases, Ninth Revision, Clinical Modification (ICD-9-CM) for discharges before October 1, 2015, and under the International Classification of Diseases, Tenth Revision, Clinical Modification (ICD-10-CM) otherwise. Following the literature [28, 31], we exclude the following diseases (codes) from the analysis: 1) external causes of injury and supplemental classification (the E and V codes in ICD-9-CM and V00-Z99 in ICD-10-CM); 2) pregnancy, childbirth and puerperium complications (630 – 679 in ICD-9-CM and O00-O9A in ICD-10-CM); and 3) symptoms, signs & ill-defined conditions (760-999 in ICD-9-CM and P00-P96, R00-R99, and S00-T88 in ICD-10-CM). This leads to 7,462 ICD-9-CM codes and 16,682 ICD-10-CM codes. These diagnosis codes are further grouped into clinically meaningful categories using the Clinical Classifications Software (CCS). CCS is developed by the Agency for Healthcare Research and Quality to better classify ICD diagnosis codes into conventionally defined human diseases. It has been widely adopted in various studies involving analysis of ICD codes [53]. To generate reliable estimates, we exclude rare diseases with a population prevalence of less than 2% over the nine years of study period. This leads to 106 CCS diseases for downstream analysis. Data processing is presented in Figure 3.1. More detailed information on these diseases is available in Appendix B.1.

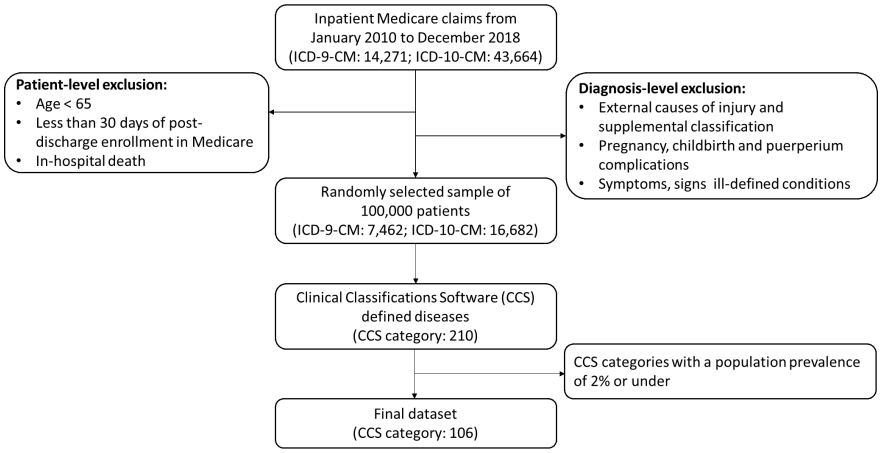


Figure 3.1: Flowchart of data processing

For each inpatient claim, there can be up to 25 (1 primary and 24 secondary) diagnosis codes. As mortality rates have declined and the population has aged, multimorbidity has become prevalent among older adults. As a result, it is common for one inpatient treatment episode to deal with

multiple conditions. To accommodate multiple possible conditions per claim, we allocate 60% of LOS to the primary diagnosis and divide the rest evenly among all secondary diagnoses. For example, if an inpatient treatment episode involved a LOS of ten days and five diagnosis codes, six days will be attributed to the primary diagnosis, and one day will be attributed to each of the four secondary diagnoses. Allocation of number of readmissions is conducted in the same manner. In the literature, there is no consensus on how to allocate among multiple disease conditions [51]. The adopted allocation ensures that for each visit, the primary diagnosis dominates and all comorbidity conditions contribute equally.

After LOS and number of readmissions are defined for each inpatient treatment episode and allocated to each diagnosis, for each subject and each CCS disease, LOS and number of readmissions are summed over all inpatient treatment episodes within the study period. To accommodate the zero-inflation nature of data, we introduce a pseudo outcome, which is a binary indicator of treatment presence. As such, the resulted analysis datasets contain three $n \times p$ matrices, where n is the number of study subjects and p is the number of CCS diseases. The ij-th entry in each of the three matrices represents treatment presence indicator (binary), total LOS (continuous), and total number of readmissions (count), respectively, for subject i and disease j.

3.2.2 Data summary

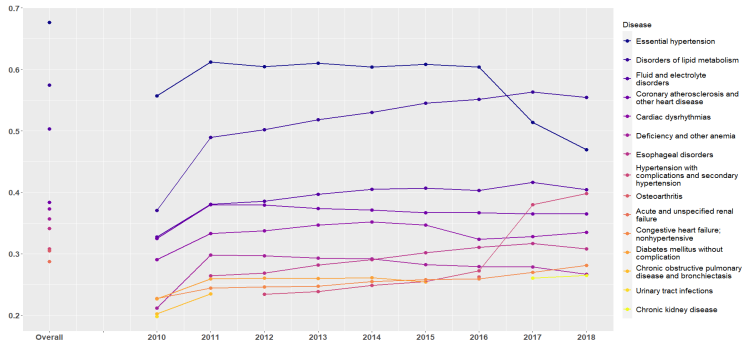
To have an overview of the analyzed data and diseases, Figure 3.2a shows the top 10 diseases with the highest prevalence (population percentage of subjects who had clinical treatments), Figure 3.2b shows the top 10 diseases with the highest average LOS among subjects with clinical treatments, and Figure 3.2c shows the top 10 diseases with the highest average number of readmissions among subjects with clinical treatments. The three plots include both overall (summarized over nine years) and yearly values. Prevalence, LOS, and number of readmissions depict from different perspectives the medical burden from hospitalization. From Figure 3.2, all diseases on the top 10 lists are well-known common diseases for the elderly population. It is also observed that chronic conditions tend to have higher prevalence, and acute conditions tend to have higher LOS and number of readmissions. While the top 10 lists for LOS and readmission are more similar to each other, congestive heart failure; nonhypertensive is the only condition that ranks in the top 10 for all three outcomes. Temporal variations are also clearly observed. For both LOS and number of

readmissions, a decreasing trend is clearly shown in Figure 3.2b and 3.2c, respectively.

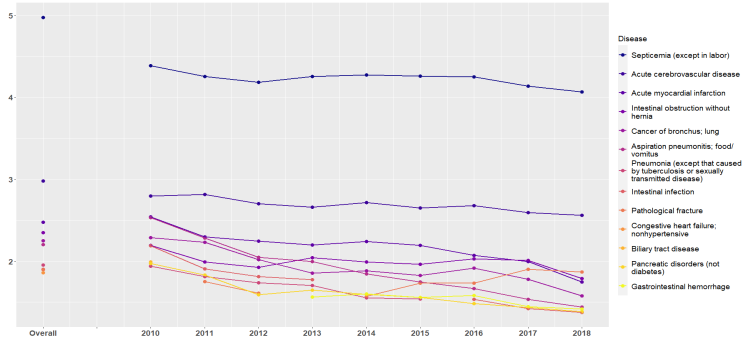
As discussed in Section 1, although limited to a few pre-selected and closely related diseases, interconnections between LOS and readmission among multiple diseases have been noted in the literature [22, 23]. In Figure 3.3, we showcase the marginal and joint distributions of LOS and number of readmissions of two common diseases: essential hypertension and disorders of lipid metabolism. In all four scatter plots, the nonparametric fit (blue line) clearly shows a positive correlation. In addition, from the marginal distributions, the zero-inflation nature and skewness of data are obvious. Following the literature [11], we conduct marginal logarithm transformations for non-zero LOS data entries. Here we note that the original LOS measurements are count data. However, with the splitting between the primary and secondary diagnoses and summarizing over the nine years of study period, it is reasonable to model LOS as continuous distributions. Figure 3.4 shows the distributions of non-zero LOS data before and after transformation for essential hypertension, disorders of lipid metabolism, and other inflammatory condition of skin. Other inflammatory condition of skin is included as an example of rare diseases. From Figure 3.4 and other alike (omitted here), we conclude that, although there may still be slight skewness after transformation, normality can be reasonably assumed. We model the number of readmission data as zero-inflated count data, and no data transformation is applied.

3.3 Methods

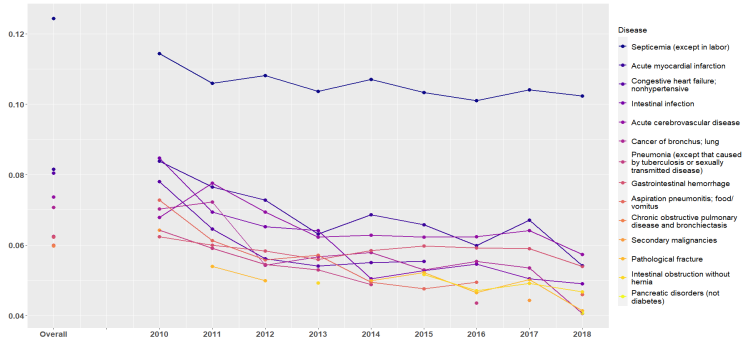
This article aims to construct an undirected (i.e., no causal relationship between diseases) clinical treatment HDN that each node represents one disease and two nodes are linked with an edge if they are associated with correlated LOS or number of readmissions. Methods for learning the structure of undirected networks can be broadly categorized into methods based on unconditional and conditional associations among variables. Most existing HDN analysis evaluates unconditional relationships such that when determining whether two diseases are interconnected, the other diseases are "ignored." For example, the phenotypic HDN by Jiang et al. [28] employed the ϕ-correlation to evaluate unconditional pairwise relationships between phenotypes of two diseases. The clinical treatment HDN by Ma et al. [31] used a two-part model for the marginal distribution of zero-inflated medical payments data and a copula-based approach to identify unconditional pairwise



(a) Prevalence



(b) Average LOS



(c) Average number of readmissions

Figure 3.2: Top 10 diseases with the highest prevalence, average LOS, and average number of readmissions

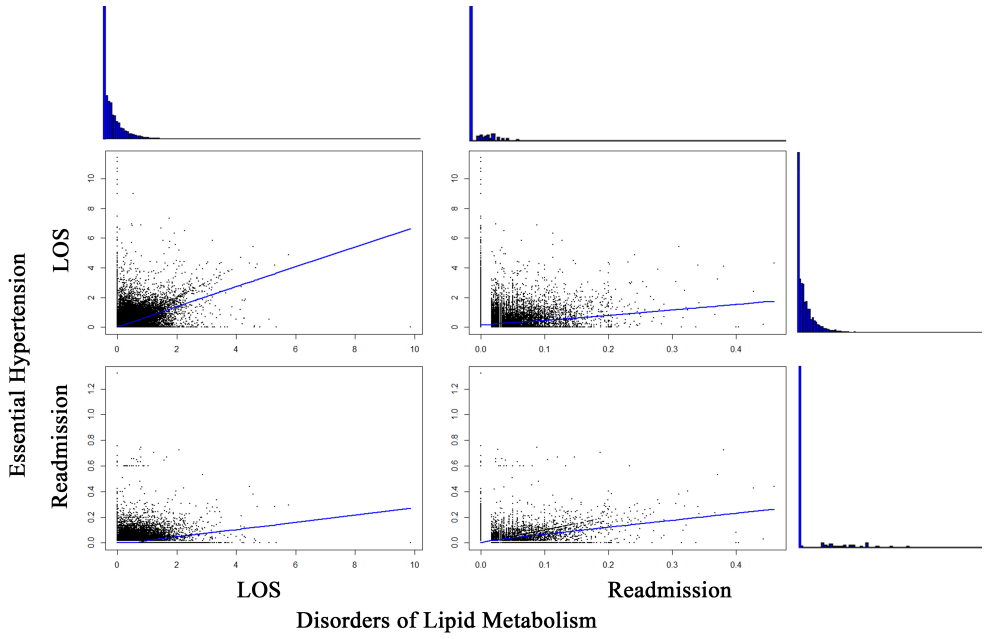


Figure 3.3: Joint and marginal distributions

relationships between the medical payments of two diseases. Considering that unconditional associations cannot distinguish between direct and indirect relationships, here we aim to conduct conditional network analysis, by which the interconnection relationship of two diseases is evaluated conditioning on all other diseases. As established in the literature [56], conditional networks can be more informative but more challenging.

In the literature, graphical models are widely used for deriving conditional dependence in network analysis. A graphical model is associated with a graph $G = (V; E)$, where the node set $V = \{1, ..., d\}$ represents the variables of interest, and the edge set $E \subseteq V \times V$ encodes the corresponding conditional dependence relationships for the node set V. Let $nb(v) = \{w \in V : \{w, v\} \in E\}$ be the neighbors of node $v \in V$. Then if node $u \notin nb(v)$, X_v and X_u (measurements of node v and u), are conditionally independent given other variables in the graph. That is:

$$(u, v) \notin E \quad \text{iff} \quad u \perp v | V/u, v.$$

This is the pairwise Markov property [81], and graph satisfying this property is called conditional independence graph (CIG).

A CIG is a set of multivariate joint distributions that depicts conditional independence relation-

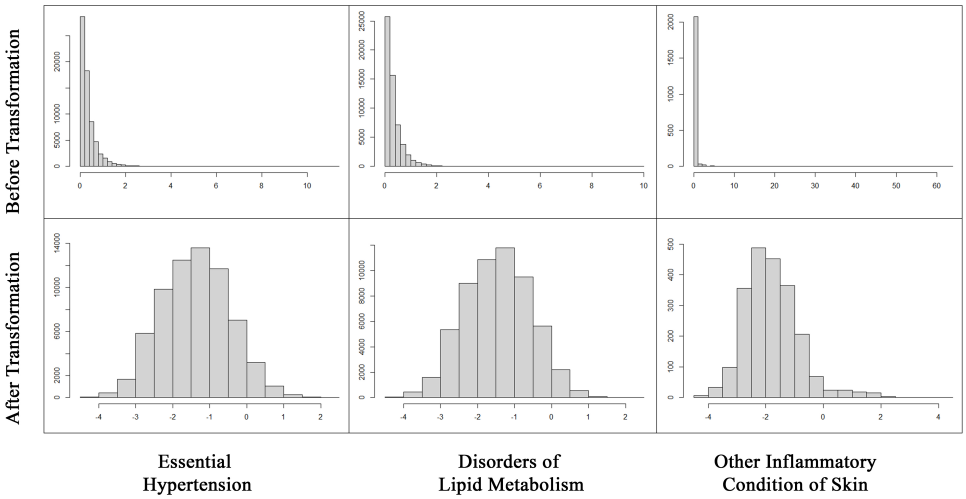


Figure 3.4: Distributions of non-zero LOS data before and after transformation

ships. It is advantageous in terms of its good interpretability and visualization capability. There are various approaches to estimate the structure of a CIG. When the dimension of the feature space is high, an efficient and popular method is developed on LASSO [81]. Through estimating the structure of the conditional distribution of each node given others in the graph, the overall graph structure could be obtained based on the neighborhood of each node. As such, the estimation could be viewed as a series of variable selection tasks. This is the so-called neighbourhood selection method. Meinshausen and Buhlmann [81] proposed this idea for the Gaussian Graphical Model, and Ravikumar et al. [82] extended the application to binary data, wherein the neighborhood of each node can be estimated independently via linear models or logistic models. Allen and Liu [83] then used log-linear regressions to estimate the CIG of count data. Efforts has been made to accommodate more data types. For example, Yang et al. [57] employed generalized linear models (GLMs), and Voorman et al. [58] used penalized generalized additive models for avoiding strict data assumptions. However, a careful examination shows that when we have multiple information sources (e.g, multiple outcomes, in our case, LOS and number of readmissions), none of the existing methods is directly applicable. Herein, we propose a shared CIG that can incorporate all information available.

3.3.1 Modeling

To jointly analyze multiple outcomes of different data types and to accommodate zero-inflation, our method is developed based on the integrative analysis of GLMs. Integrative analysis jointly analyzes a set of models with similar model structures and closely related response variables. By assuming model structures to be similar but not completely equivalent, integrative analysis can detect heterogeneity and utilize information across various data sets, thus lead to more reliable estimations [84]. We propose an integrative analysis of GLMs, which can accommodate various data types thus have strong versatility. Let d be the number of diseases, p be the number of outcomes, and $\mathbf{y}^{(k)} = (y_1^{(k)}, ..., y_d^{(k)})^T$ $(k = 1, ..., p)$ be the observed values for the $k-th$ outcome. Consider the conditional distribution of the $k-th$ outcome of disease j, given all other diseases, belongs to the exponential family:

$$p(y_j^{(k)}|\mathbf{y}_{-j}) = C_k(y_j^{(k)})exp[\theta_j^{(k)} y_j^{(k)} - b_k(y_j^{(k)})], \quad j = 1, ..., d \quad \text{and} \quad k = 1, ..., p,$$

where $\theta_j^{(k)}$ is the natural parameter, and $b_k(y_j^{(k)})$ is a know function. From the properties of exponential family, it can be shown that $\mu_{y_j^{(k)}} = E(y_j^{(k)}|\mathbf{y}_{-j}) = \dot{b_k}(y_j^{(k)})$, with $\dot{b_k}$ being the first derivative of b_k.

To further model the zero-inflation, we employ the idea of two-part models [11] and introduce a pseudo outcome $\mathbf{y}^{(0)} = (y_1^{(0)}, ..., y_d^{(0)})^T$ with each element being a binary indicator of treatment presence of the corresponding disease. Then for disease j, the $k-th$ outcome value is conditional on the pseudo outcome in a sense that:

$$\begin{cases} p(y_j^{(k)}|y_j^{(0)} = 0, \mathbf{y}_{-j}) = 0 \\ p(y_j^{(k)}|y_j^{(0)} = 1, \mathbf{y}_{-j}) = C_k(y_j^{(k)})exp[\theta_j^{(k)} y_j^{(k)} - b_k(y_j^{(k)})] \end{cases}, \quad j = 1, ..., d \quad \text{and} \quad k = 0, ..., p.$$

Since elements in $\mathbf{y}^{(k)}$ are very closely related, to capture the connectivity between them, we assume their GLMs share similar model structures. Let $l_j = \sum_{k=0}^{d} \mathbf{y}_{-j}^{(k)^T} \boldsymbol{\beta}_j^{(k)}$, we define the link function as:

$$g_k(\mu_{y_j^{(k)}}) = \alpha_j^{(k)} + \tau_j^{(k)} l_j, \quad j = 1, ..., d, \quad k = 0, ..., p, \quad \text{and} \quad \tau_j^{(0)} = 1,$$

where $\alpha_j^{(k)}$, $\boldsymbol{\beta}_j^{(k)}$, and $\tau_j^{(k)}$ are the parameters to be estimated. Note that the GLMs for different outcomes of the same disease are also integrated with similar model structures. That is, except for the intercepts, we let the linear predictors be proportional and controlled by the parameter $\tau_j^{(k)}$.

Assume that g_k is the natural link (i.e., $g_k = \dot{b_k}^{-1}$) and $\theta_j^{(k)} = u(g_k(\mu_{y_j^{(k)}}))$. Since $\mu_{y_j^{(k)}} = \dot{b_k}(\theta_j^{(k)})$, it can be shown that $u(t) = \dot{b_k}^{-1}(g^{-1}(t)) = t$. Also assume that $Y_j^{(0)}|\mathbf{Y}_{-j} \sim Bernoulli(p_j)$. Then the log-likelihood of disease j can be derived as:

$$L_j(\boldsymbol{\alpha}_j, \boldsymbol{\beta}_j, \boldsymbol{\tau}_j, \mathbf{Y}) = -\sum_{i=1}^{n} log[1 + exp(\alpha_j^{(0)} + l_{ij})] + \sum_{i=1}^{n} y_{ij}^{(0)}[\alpha_j^{(0)} + l_{ij} + \sum_{k=1}^{p} y_{ij}^{(k)}(\alpha_j^{(k)} + \tau_j^{(k)} l_{ij}) - b_k(\alpha_j^{(k)} + \tau_j^{(k)} l_{ij})],$$

where n is the number of subjects.

This article aims is to jointly analyze LOS and number of readmissions. Therefore, we have $d = 2$. Let $\mathbf{y}^{(0)}$, $\mathbf{y}^{(1)}$, and $\mathbf{y}^{(2)}$ be values of treatment presence indicator, LOS, and number of readmissions, respectively. With the assumptions that:

$$\begin{aligned}
&Y_j^{(0)}|\mathbf{Y}_{-j}^{(0)}, \mathbf{Y}_{-j}^{(1)}, \mathbf{Y}_{-j}^{(2)} \sim Bernoulli(p_j),\\
&Y_j^{(1)}|Y_j^{(0)} = 1, \mathbf{Y}_{-j}^{(0)}, \mathbf{Y}_{-j}^{(1)}, \mathbf{Y}_{-j}^{(2)} \sim N(\mu_j, \sigma_j),\\
&Y_j^{(2)}|Y_j^{(0)} = 1, \mathbf{Y}_{-j}^{(0)}, \mathbf{Y}_{-j}^{(1)}, \mathbf{Y}_{-j}^{(2)} \sim Poisson(\lambda_j),
\end{aligned}$$

we have the link functions for each outcome as:

$$\begin{aligned}
&log(\frac{p_j}{1-p_j}) = \alpha_j^{(0)} + l_j,\\
&\mu_j = \alpha_j^{(1)} + \tau_j^{(1)} l_j,\\
&log(\lambda_j) = \alpha_j^{(2)} + \tau_j^{(2)} l_j,
\end{aligned}$$

respectively. Then the log-likelihood can be derived as:

$$L_j(\boldsymbol{\alpha}_j, \boldsymbol{\beta}_j, \boldsymbol{\tau}_j, \mathbf{Y}) = -\sum_{i=1}^{n} log[1 + exp(\alpha_j^{(0)} + l_{ij})] + \sum_{i=1}^{n} y_{ij}^{(0)}[\alpha_j^{(0)} + l_{ij} + y_{ij}^{(1)}(\alpha_j^{(1)} + \tau_j^{(1)} l_{ij}) - \frac{\sigma_j^2}{2}(\alpha_j^{(1)} + \tau_j^{(1)} l_{ij})^2 + y_{ij}^{(2)}(\alpha_j^{(2)} + \tau_j^{(2)} l_{ij}) - exp(\alpha_j^{(2)} + \tau_j^{(2)} l_{ij})].$$

3.3.2 Estimation

With the developed node-wise log-likelihood, we can estimate the structure of the conditional dependence for each node given all others in the graph, and thereby obtain the overall graph structure. Specifically, for a fix node $j \in \{1, ..., d\}$, consider the conditional distribution of $\mathbf{Y}_j$ given all other nodes in $\mathbf{Y}_\mathbf{M}$ for $\mathbf{M} = \{1, ..., d\} \setminus \{j\}$. For $m \in \mathbf{M}$, node m and node j are defined as conditionally independent if and only if all the related interactive parameters are zeros. That is,

$$\beta_{mj}^{(0)} = \beta_{mj}^{(1)} = ... = \beta_{mj}^{(p)} = 0$$

Our estimation method could be viewed as a generalization of the Quasi-likelihood estimation method. Even if the models are misspecified, consistency could still be achieved if the conditional expectation and variance are set correctly for $\mathbf{y}^{(k)}(k = 1, ..., p)$.

When analyzing disease interconnection relationships at a pan-disease level, it is reasonable to assume that not all disease are correlated. Therefore, we propose a sparse estimation such that the estimated network is sparse and interpretable. Here we employ a group LASSO penalty for sparse estimation, as it has been well established and widely used in HDN analysis [56, 77]. For $m \in \mathbf{M}$, define the parameter vector $\boldsymbol{\theta}_m = (\beta_{mj}^{(0)}, \beta_{mj}^{(1)}, ..., \beta_{mj}^{(p)})^T$. As established, $\mathbf{Y}_j \perp\!\!\!\perp \mathbf{Y}_m | \mathbf{Y}_{\mathbf{M}\setminus\{m\}}$ if and only if $\boldsymbol{\theta}_m = \mathbf{0}$, thus we want the elements of $\boldsymbol{\theta}_m$ to be all zeros or non-zeros at the same time. To guarantee the uniformity of graph structure, we apply a group LASSO penalty and propose the penalized objective function:

$$\min_{\boldsymbol{\alpha}_j, \boldsymbol{\beta}_j, \boldsymbol{\tau}_j} -L_j(\boldsymbol{\alpha}_j, \boldsymbol{\beta}_j, \boldsymbol{\tau}_j, \mathbf{Y}) + \lambda \sum_{m \neq j} ||\boldsymbol{\theta}_m||_2,$$

where λ is the tuning parameter controlling penalty strength and is selected data dependently using cross validation.

Define the parameter vector $\boldsymbol{\psi} = (\boldsymbol{\alpha}_j, \boldsymbol{\beta}_j, \boldsymbol{\tau}_j)^T$. It can be observed that the penalized objective function is a sum of a differentiable convex function $g(\boldsymbol{\psi}) = -L_j(\boldsymbol{\alpha}_j, \boldsymbol{\beta}_j, \boldsymbol{\tau}_j, \mathbf{Y})$ and a non-differentiable convex function $h(\boldsymbol{\psi}) = \lambda \sum_{m \neq j} ||\boldsymbol{\theta}_m||_2$. Given the structure of the penalized objective function, it can be optimized using the proximal gradient descent technique, which has been largely used in other network studies [61, 77]. Specifically, the proximity operator for $\boldsymbol{\psi}$ is defined as:

$$prox_{h,t}(\boldsymbol{\psi}) = \arg\min_{\mathbf{x}} \frac{1}{2t}\|\boldsymbol{\psi} - \mathbf{x}\|_2^2 + h(\boldsymbol{\psi}),$$

where t is the step size. Then the update rule can be described as

$$\boldsymbol{\psi}^{k+1} = prox_{h,t_k}(\boldsymbol{\psi}^k - t_k \nabla g(\boldsymbol{\psi}^k)).$$

The optimum could be achieved by finding the solution $\boldsymbol{\psi}^\star$ of:

$$\boldsymbol{\psi} = prox_{h,t_k}(\boldsymbol{\psi} - t_k \nabla g(\boldsymbol{\psi})).$$

Thus the algorithm terminates when $\boldsymbol{\psi}^{k+1}$ is very close to $\boldsymbol{\psi}^k$, i.e., $|\boldsymbol{\psi}^{k+1} - \boldsymbol{\psi}^k| < \epsilon$, with ϵ being a prespecified threshold.

Since $h(\boldsymbol{\psi})$ only involves $\boldsymbol{\beta}_j$, the gradient descent algorithm could be used for updating $\boldsymbol{\alpha}_j$ and $\boldsymbol{\tau}_j$. Set $\phi(\boldsymbol{\theta}_m^k) = \boldsymbol{\theta}_m^k - t_k \partial g(\boldsymbol{\theta})$. For $\boldsymbol{\theta}_m (m \neq j)$, the detailed update rule is:

$$\boldsymbol{\theta}_m^{k+1} = \begin{cases} 0, & \text{if } ||\phi(\boldsymbol{\theta}_m^k)|| \leq t_k\lambda \\ \frac{(||\phi(\boldsymbol{\theta}_m^k)|| - t_k\lambda)\phi(\boldsymbol{\theta}_m^k)}{||\phi(\boldsymbol{\theta}_m^k)||}, & \text{if } ||\phi(\boldsymbol{\theta}_m^k)|| > t_k\lambda \end{cases}$$

For node $j = 1, ..., d$, the gradient of the log-likelihood $g(\boldsymbol{\psi}) = -L_j(\boldsymbol{\alpha}_j, \boldsymbol{\beta}_j, \boldsymbol{\tau}_j, \mathbf{Y})$ with respect to $\boldsymbol{\psi}$ can be easily calculated with the chain rule. As such, the neighborhood selection procedure based on proximal gradient descent can be performed for each node, and the overall graph structure can thus be estimated. Pseudocode are presented in Algorithm 1.

Algorithm 1 Neighborhood selection algorithm

Input:
Data sets $\mathbf{Y}$;
Tuning parameter λ;
Threshold ϵ;
Output:
$\boldsymbol{\theta}_m(m \neq j, j = 1, ..., p)$;

for $j = 1, ..., p$ **do**
  $k = 0$;
  Initialize $/\psi^0$;
  loop
    Calculate $\partial g(\boldsymbol{\psi}^k)$;
    $\boldsymbol{\alpha}_j^{k+1} = \boldsymbol{\alpha}_j^k - t_k \partial g(\boldsymbol{\alpha}_j^k)$;
    $\boldsymbol{\tau}_j^{k+1} = \boldsymbol{\tau}_j^k - t_k \partial g(\boldsymbol{\tau}_j^k)$;
    $\boldsymbol{\theta}_m^{k+1} = \begin{cases} 0, ||\phi(\boldsymbol{\theta}_m^k)|| \leq t_k\lambda \\ \frac{(||\phi(\boldsymbol{\theta}_m^k)|| - t_k\lambda)\phi(\boldsymbol{\theta}_m^k)}{||\phi(\boldsymbol{\theta}_m^k)||}, ||\phi(\boldsymbol{\theta}_m^k)|| > t_k\lambda \end{cases}$;
    EXIT WHEN $|\boldsymbol{\psi}^{k-1} - \boldsymbol{\psi}^k| < \epsilon$
    $k = k + 1$
  end loop
  return $\boldsymbol{\theta}_m(m \neq j)$;
end for

3.3.3 Analysis of network properties

With the graphical model developed in Section 3.3.1 and parameter estimations obtained in Section 3.3.2, we can construct and visualize the desired clinical treatment HDN. In addition, key properties such as adjacency, connectivity, and module/hub are summarized [56, 63]. Analysis of theses properties can provide more meaningful interpretations of the network than model parameters only. Similar analysis has been conducted in the prior phenotypic and clinical treatment HDNs [28,31,77].

Adjacency matrix Elements in the $d \times d$ adjacency matrix describe pairwise conditional independence, which indicates whether an edge exists between two nodes. In our model, since conditional independence structure is estimated separately for each node, we conclude conditional independence between two diseases i and j if and only if all ij interactive parameters are zeros. That is, $\beta_{ij}^{(0)} = ... = \beta_{ij}^{(p)} = \beta_{ji}^{(0)} = ... = \beta_{ij}^{(p)} = 0$. In our case, with $p = 2$, the adjacency matrix $\mathbf{A} = [a_{ij}]$ is

defined as

$$a_{ij} = \begin{cases} 0, & \text{if } \hat{\beta}_{ij}^{(0)} < \tau \ \& \ \hat{\beta}_{ij}^{(1)} < \tau \ \& \ \hat{\beta}_{ij}^{(2)} < \tau \ \& \ \hat{\beta}_{ji}^{(0)} < \tau \ \& \ \hat{\beta}_{ji}^{(1)} < \tau \ \& \ \hat{\beta}_{ji}^{(2)} < \tau \\ 1, & \text{otherwise} \end{cases}$$

where a threshold τ is imposed to further remove spurious small interconnections and generate more interpretable results. τ can be determined data-dependently (for example, using cross validation) or based on the specific contexts.

Connectivity To what extent one node is connected with other is defined as connectivity. For node (disease) $j \in \{1, ..., d\}$, its connectivity is defined as

$$K_j = \sum_{i \neq j} a_{ij}.$$

Since higher connectivity means higher overall impact in a network sense, the importance of nodes with high connectivity is emphasized in many network analysis including the HDNs [28, 31].

Module Network modules, also been referred to as network communities or clusters in some studies, consist of tightly interconnected nodes. There are many approaches in module construction. Here we adopt the topological overlap matrix (TOM) based method [63] because of its lucid interpretability. The ij-th element of TOM is defined as

$$TOM_{ij} = \frac{l_{ij} + a_{ij}}{minK_iK_j + 1 - a_{ij}},$$

where $l_{ij} = \sum_u a_{iu}a_{uj}$ measures how many neighbors that nodes i and j share. Accordingly, dissimilarity matrix, which is defined as dissTOM = 1-TOM, measures the dissimilarity of any pairs of diseases. Based on dissTOM, modules can then be constructed by hierarchical clustering with a dynamic tree cutting approach [63]. By definition, diseases within the same module tend to behave similarly in terms of the outcomes measured, and diseases from different module are relatively weakly connected. Therefore, module analysis provides fundamental information in disease characterization and classification. After module construction, a node's intramodular connectivity, is measured as its connectivity limited to the module it belongs to. Diseases with the highest intramodular connectivity are referred as the hub diseases. These diseases have a higher impact

within the module it belongs to.

3.4 Data analysis

Analyzing the Medicare inpatient data from January 2010 to December 2008, the constructed clinical treatment HDN on LOS and readmission is visualized using *Gephi* and presented in Figure 3.5. In the graph, each node represents one CCS disease. Two nodes are linked with a line if they are determined to be interconnected under the framework developed in Section 3.3. The node size is proportional to its connectivity, and the node color represents its module membership. There are 106 nodes linked with 1,008 edges and clustered into 9 modules. The network is highly-interconnected, with every node connected to at least one other. Similar network visualization is also conducted for each year within the study period and is provided in Appendix B.2.

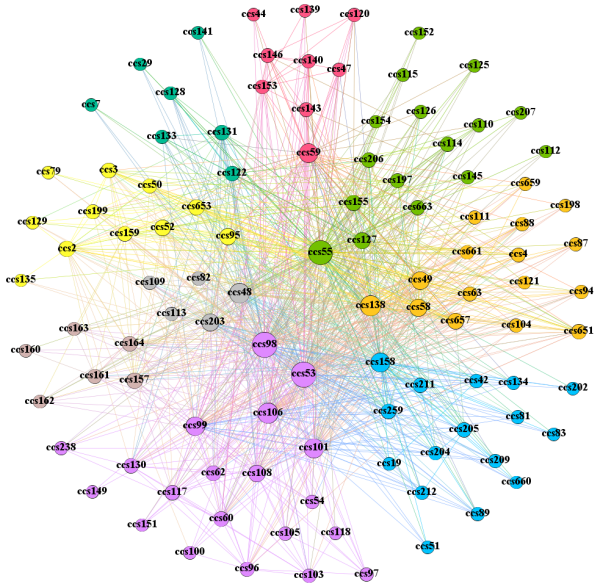


Figure 3.5: LOS and readmission HDN for the period of January 2010 to December 2018

3.4.1 Connectivity

Figure 3.6 shows the top ten diseases with the highest connectivity. Both overall and yearly values are provided. Comparing Figure 3.6 to Figure 3.2, diseases with high prevalence tend to have high connectivity. All diseases in Figure 3.6, except for thyroid disorders, also appear in Figure 3.2a.

On the contrary, the relationship between disease rank in connectivity and average LOS/number of readmissions is not apparent. All diseases with high average LOS and number of readmissions do not appear to have a high connectivity. This finding may provide critical insights into future policymaking since existing interventions generally focus on individual conditions with outstanding LOS and readmission rates. For example, in March 2010, the Affordable Care Act established the Hospital Readmissions Reduction Program (HRRP) to incentivize hospitals to reduce readmissions among Medicare beneficiaries. Starting in October 2012, the program began to penalize general acute care hospitals with high risk-adjusted 30-day readmission rates for three target conditions: AMI, heart failure, and pneumonia. The target conditions are well known to have high readmission rates, which is also confirmed in Figure 3.2c. However, our network analysis shows that they do not have a high connectivity in terms of LOS and readmission, even in years before the implementation of HRRP. Since conditions with higher connectivity values have a higher overall impact within the network system, it may be necessary for interventions like the HRRP to target diseases shown in Figure 3.6.

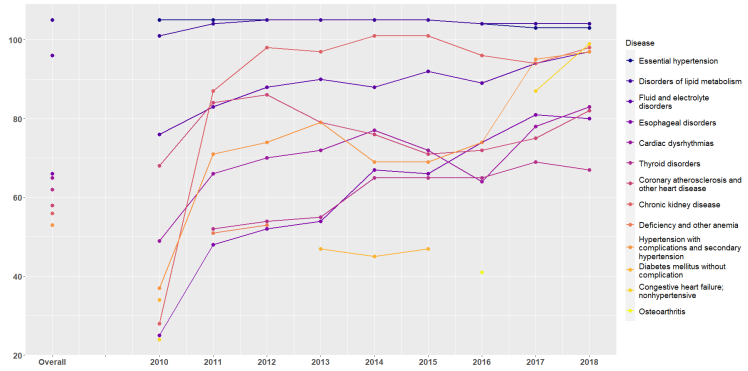


Figure 3.6: Top 10 diseases with the highest overall and yearly connectivity

3.4.2 Module

As shown in Figure 3.5, there are 9 modules whose size ranges from 5 to 20. Modules consist of a group of diseases that are closely interconnected with each other. Module analysis provides essential information for disease characterization and classification. An enrichment analysis shows that the 9 modules are enriched with the following diseases: cardiovascular and metabolic diseases (purple), acute conditions with bad prognosis (green), obstinate and recurrent disorders (blue), controllable

chronic disorders (orange), secondary diseases and complications (yellow); digestive system diseases (vermeil), genitourinary system diseases (champagne), cerebrovascular disease (brown), and residual disease (gray), respectively. Widely used classifications of diseases are generally based on 1) bodily region or system, 2) function or effect, 3) disease pathology, and 4) disease etiology. Significantly different from existing standards, our analysis provides an alternative way of disease characterization and classification from a treatment point of view. For example, the green, blue, and orange modules each contains diseases of various body systems. These diseases have a relatively long distance considering disease pathology and etiology. However, they share common characteristics in clinical treatments. Precisely, the green module consists of severe and acute conditions that require immediate treatments. These diseases generally have a poor prognosis, which would result in prolonged LOS and high readmission rates. The blue module is enriched with chronic conditions which are usually incurable and require repeated treatments. The orange module contains controllable chronic diseases, which do not require extensive inpatient treatments and are often handled with outpatient treatments and drug refillings.

3.4.3 Hub

Hub diseases are those with the highest intramodular connectivity. For each module, the hub diseases are: disorders of lipid metabolism and essential hypertension (purple), fluid and electrolyte disorders (green), chronic kidney disease (blue), esophageal disorders (orange), septicemia (yellow); deficiency and other anemia (vermeil), hyperplasia of prostate and other diseases of kidney and ureters (champagne), paralysis (brown), and pneumonia (gray), respectively. It is noted that overall connectivity and intramodular connectivity describe the importance of diseases from two different aspects. For example, hyperplasia of prostate, as the hub disease for the module (champagne) of genitourinary diseases, has a low overall connectivity but a high intramodular connectivity. This is expected since disorders of the genitourinary system are relatively independent in incidence, most of which are not significantly related to disorders of other systems [85]. A further literature search reveals that, while age and sex hormones are two important risk factors, the etiology of prostate hyperplasia has not been well established [86]. A closer examination of disease interconnected with prostate hyperplasia may provide additional insights and also suggest potential future research directions.

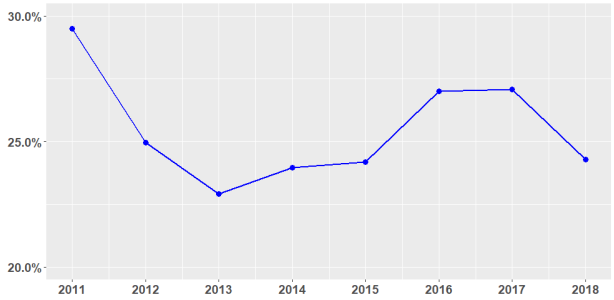
3.4.4 Temporal variation

Following the literature [27, 28, 77], we examine the temporal variations of network structure by repeating the above analysis for each year from 2010 to 2018. We select an independent random sample of 100,000 subjects for each year. The independent sampling ensures that the select sample represents cohort characteristics for the corresponding year.

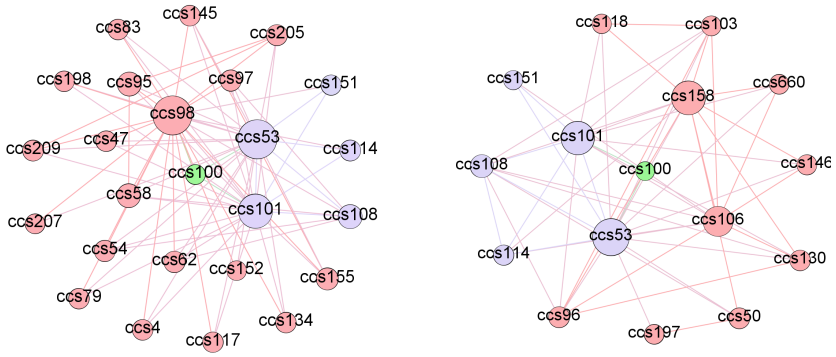
Connectivity Temporal variations in connectivity are apparent, as shown in Figure 3.6. Except for disorders of lipid metabolism and essential hypertension, which have continuously high connectivity in all years, all other diseases show a noticeable fluctuation in connectivity across time. For example, chronic kidney disease continuously ranks high (the third highest) in connectivity from 2011 to 2018. However, it has a lower (the eighth highest) overall connectivity due to the really low value in year 2010. This observation indicates that examining temporal variations, especially recent year trend, may provide valuable information that could not be obtained through an overall estimation. Generally speaking, there is an increasing trend in connectivity for most diseases from 2010 to 2018. Contributing factors may include the improvement in healthcare accessibility and technical advances in diagnosis and treatments. We also observe a jump in individual connectivity from 2010 to 2011. This is mainly due to the change in CMS' electronic transaction standards that providers use to submit Medicare claims. Starting in January 2011, institutional providers can enter up to 25 diagnosis codes for a single claim where previously only 10 were allowed [13]. A closer examination of the data reveals that the average number of diagnosis codes per Medicare inpatient claims increased from 8.03 in 2010 to 13.93 in 2011. This policy change causes an increased number of comorbidities and hence an increased connectivity. However, with the overall network building on all claims accumulated for 9 years, it is reasonable to believe that all comorbidities are captured. Therefore, the policy change may significantly impact yearly estimation but will not affect the overall analysis.

Module To get insights into temporal variations in module membership, we evaluate changes in pairwise module relationship. If two diseases belong to the same module in one year and belong to different modules in the following year, we say their module relationship has changed. The converse also holds true. Figure 3.7a shows the yearly percentage of disease pairs that have changed their module relationships. There is a moderate change (20% - 30%) in pairwise module relationships

in all years. The biggest change happens in year 2011, which is possibly also due to the change in CMS' electronic transaction standards mentioned above.



(a) Yearly changes in disease module relationships



(b) Modules that contain AMI in 2011 (left) and 2013 (right)

Figure 3.7: Temporal variations of module structure

To have a closer look at how module structures could change, Figure 3.7b shows the modules that AMI (ccs100) belongs to in 2011 and 2013. Diseases colored in purple are shared diseases in both years, and those colored in pink appear only in one of the two years. As one of the most common conditions in older adults, AMI is known to be associated with high medical expenses and poor prognosis. This is also confirmed in Figure 3.2b and 3.2c, where AMI has the second highest average LOS and average number of readmissions. For this reason, it is one of the three initial target conditions of HRRP, which aims to incentivize healthcare providers to reduce readmissions. The program went effective in October 2012 and resulted in a reduction in unadjusted readmission rates from 18.9% to 16.0% for AMI [87]. Therefore, we expect a noticeable change in the module structure containing AMI before and after the implementation of the HRRP. As shown in Figure 3.7b, the module containing AMI has 26 conditionals in 2011 and 16 conditions in 2013. For both years, AMI is highly interconnected with other cardiovascular diseases, including coronary

atherosclerosis and other heart disease (ccs101), congestive heart failure; nonhypertensive (ccs108), and peripheral and visceral atherosclerosis (ccs114). Fore year 2011, it is clustered with more diverse disorders, mainly secondary diseases and complications. Given the high prevalence and intensive treatments for AMI, this observation is very reasonable. However, for year 2013, with a significant decrease in readmission rates after implementing HRRP, AMI is no long clustered with various secondary diseases and complications. Instead, the module in 2013 include more heart-related conditions, including pulmonary heart disease (ccs103), cardiac dysrhythmias (ccs106), and heart valve disorders (ccs96).

3.4.5 Sensitivity analysis

In data preparation, we allocate 60% of outcome measures (i.e., LOS and number of readmissions) to the primary diagnosis code and the rest 40% evenly to all secondary diagnosis codes. This allocation ensures that the primary diagnosis dominates and all secondary diagnoses contribute evenly. To learn about the impact of this allocation approach on network structure, we conduct sensitivity analysis, in which we allocate 70% of outcome measures to the primary diagnosis and the rest 30% evenly to all secondary diagnoses. The above analysis in network structure and properties is repeated using combined data for the whole study period, and only ignorable differences are observed (details omitted and available from the authors).

3.4.6 Comparison to the LOS HDN

Comparing to the LOS HDN constructed in Chapter 2, the network incorporating both LOS and readmission has a similar structure. Specifically, both networks are highly interconnected with a comparable level of sparsity. High connectivity and hub diseases such as essential hypertension, disorders of lipid metabolism, and fluid and electrolyte disorders are consistently identified. Comparing Figure 2.6 with Figure 3.6, chronic kidney disease and diabetes mellitus without complication are the only two diseases that appear on the top 10 connectivity lists for the LOS and readmission HDN but not on that for the LOS HDN. Moreover, the jump in connectivity from 2010 to 2011 due to the change in CMS' electronic transaction standards are also captured by both networks. Despite the similarities, the two networks are different in several ways. Comparing Figure 2.5 with Figure 3.5, in the LOS and readmission HDN, there are several really big nodes with more number

of small sized nodes on the periphery of the network. This means that central diseases are interconnected with a larger number of other disease and are better recognized when adding readmission to the network construction. Moreover, modules constructed in Chapter 2 have more traditional interpretations in terms of bodily system, disease pathology, and disease etiology. Analyzing both LOS and readmission, modules constructed in this chapter consist of similar diseases from a treatment perspective. The yearly changes in pairwise module relationships are also observed to be higher in the network on LOS and readmission, compared to that on LOS only. These deviations are attributable to different outcomes evaluated and different modeling approaches utilized.

3.5 Simulation

Simulation is conducted to learn about performance of the proposed modeling and estimation approach. For simulated datasets, we consider a variety of settings with different graph theories (e.g., dependency network and star graph) and sparsity levels.

A dependency network [88] could be viewed as a mechanism for combining regression/classification models via Gibbs sampler to define a joint distribution. For example, in Bayesian network, joint distributions are defined using the product of local distributions. In our simulations, to generalize multiple interconnected data sets, integrated generalized linear models developed in section 3.3.1 are used here to define local distributions. For $\boldsymbol{\beta_j}$ $(j = 1, ..., p)$, we construct three interactive matrices $M(\boldsymbol{\beta^{(0)}})$, $M(\boldsymbol{\beta^{(1)}})$ and $M(\boldsymbol{\beta^{(2)}})$, indicating parameters in $\boldsymbol{\beta^{(0)}}, \boldsymbol{\beta^{(1)}}, \text{and} \boldsymbol{\beta^{(2)}}$, respectively. Noting that the diagonal elements have no actual impact, we set them to be 1 for convenience. We assume that the interactive matrices have a diagonal structure, a banded structures, or an exponential attenuation structure. For the banded structure, the numbers of nonzero diagonals depend on a prespecified sparsity levels. For example, a 1.5% sparsity level results in a band width of three (i.e., a tri-diagonal matrix), and a 5% sparsity level results in a band width of seven. Non-zero elements in matrices with a banded structure are set to be 0.1 or 0.5. For the exponential attenuation structure, we consider an attenuation factor of 0.5. We also impose blocks on the interactive matrices, such that nodes in different blocks are conditionally independent.

A star graph [82] consists of sets of mutually independent nodes linked to one seed nodes (i.e., a tree with one internal node and leaves). In the simulated data sets, we distributed the variables

equally based on the number of seeds. For example, if there are 5 seed nodes and 100 variables in total, we split the variables evenly into 5 groups with one seed per group. Again, the number of leaves in each group depends on a prespecified sparsity levels. For example, if the sparsity level is 0.5, then the seed node is expected to be linked with half of other nodes in its group. With exact sampling, data sets are sampled from models developed in section 3.3.1. $\boldsymbol{\beta}$ in each model are randomly generated from $N(0, 0.25)$.

Detailed simulation settings are described in Table 3.1. Coefficients other than $\boldsymbol{\beta}$ are randomly generated from prespecified distributions, which are presented in Table 3.2. To reduce computation burden, we fix the sample size as 10,000 and the node size as 100. We note that for real data analysis, a much larger sample size is used for a similar number of nodes, hence more reliable estimation is expected.

Dependency network	
1	$M(\boldsymbol{\beta}^{(0)})$ has a banded structure with sparsity level 1.5%; $M(\boldsymbol{\beta}^{(1)})$ and $M(\boldsymbol{\beta}^{(2)})$ are diagonal
2	$M(\boldsymbol{\beta}^{(0)})$, $M(\boldsymbol{\beta}^{(1)})$ and $M(\boldsymbol{\beta}^{(2)})$ all have a banded structure with sparsity level 1.5%
3	$M(\boldsymbol{\beta}^{(0)})$ has an exponential attenuation structure; $M(\boldsymbol{\beta}^{(1)})$ and $M(\boldsymbol{\beta}^{(2)})$ are diagonal
4	$M(\boldsymbol{\beta}^{(0)})$, $M(\boldsymbol{\beta}^{(1)})$ and $M(\boldsymbol{\beta}^{(2)})$ all have exponential attenuation structure
5	$M(\boldsymbol{\beta}^{(0)})$ has a banded structure with sparsity level 5%; $M(\boldsymbol{\beta}^{(1)})$ and $M(\boldsymbol{\beta}^{(2)})$ are diagonal
6	$M(\boldsymbol{\beta}^{(0)})$, $M(\boldsymbol{\beta}^{(1)})$ and $M(\boldsymbol{\beta}^{(2)})$ all have banded structure with sparsity level 5%
7-10	Five blocks of settings 3-6, respectively
Star graph	
11	One seed with a sparsity level of 0.7
12	Five seed with a sparsity level is 0.5
13	Five seed with a sparsity level is 0.7
14	Ten seed with a sparsity level is 0.7

Table 3.1: Simulation settings

$\boldsymbol{\alpha}$ for seeds in star graph	$N(0, 0.25)$
$\boldsymbol{\alpha}$ for dependency network and non-seeds in star graph	$N(0, 1)$
$\boldsymbol{\tau}$	$N(1, 0.01)$
$\boldsymbol{\sigma^2}$	1

Table 3.2: Coefficient generation

To gain more insights into the relative performance of our approach compared to two alternative approaches, we compute and compare the area under the receiver operating characteristic curve (AUC) using different methods. The first alternative approach conducts a regression analysis separately for each of the three outcomes: treatment presence, LOS, and number of readmissions.

Estimated coefficients indicate conditional pairwise relationship considering a single outcome. Conditional independence is then concluded between two diseases if the corresponding coefficients in all three models are estimated to be zero. The second alternative approach computes the unconditional pairwise Spearman correlation for measures of all three outcomes as a vector. For each setting, 100 replicates are simulated. In network analysis, the key is to accurately identify edges. In Table 3.3, we summarize the mean AUC and corresponding standard deviation, which shows that the proposed approach outperforms the two alternative approaches in all 14 settings. The simulation results provide strong confidence in the validity of our findings.

Setting	Proposed	Alternative1	Alternative2
1	94.61% (0.03)	87.81% (0.09)	85.90% (0.10)
2	96.50% (0.02)	85.86% (0.06)	83.00% (0.07)
3	94.75% (0.02)	89.24% (0.03)	87.13% (0.04)
4	94.99% (0.02)	85.44% (0.04)	81.12% (0.06)
5	95.40% (0.01)	95.03% (0.02)	91.48% (0.04)
6	87.11% (0.05)	79.94% (0.06)	67.70% (0.06)
7	94.53% (0.02)	86.77% (0.04)	85.40% (0.05)
8	93.14% (0.03)	81.57% (0.03)	76.83% (0.04)
9	94.88% (0.03)	92.35% (0.03)	88.20% (0.03)
10	87.96% (0.07)	81.86% (0.06)	73.10% (0.06)
11	84.90% (0.02)	76.48% (0.02)	81.42% (0.02)
12	94.17% (0.01)	72.14% (0.02)	78.54% (0.03)
13	83.43% (0.03)	77.80% (0.02)	80.14% (0.02)
14	86.90% (0.02)	73.59% (0.02)	79.23% (0.02)

Table 3.3: Simulation results: mean AUC (sd)

3.6 Discussion

In this article, we construct a clinical treatment HDN building on disease associated LOS and number of readmissions. The analysis can complement the existing literature in multiple ways. First, it complements the individual-disease and overall outcome analyses by focusing on pan-disease level interconnections. Second, the proposed clinical treatment HDN complement existing molecular and phenotypic HDNs by being "closer" to clinical treatments and thus having a higher practical value. Third, the network analysis on LOS and readmission complements existing clinical treatment HDNs (e.g., medical costs HDN by Ma et al. [31] and LOS HDN by Mei et al. [77]) by incorporating multiple outcomes. As discussed in Section 3.1, outcome analysis has important

implication for evaluating and improving healthcare quality and efficiancy, as well as infomatively reflect intrinsic disease properties. The proposed analysis on LOS and readmission can foster a better understanding of disease interconnections in term of healthcare outcomes, provide insights into a better disease management, and guide a more efficient allocation of hospital beds and other resources. In addition, the novel analytic approach developed for constructing a high-dimensional and conditional network using zero-inflated data may further complex network analysis.

This article has certain limitations. Due to limited data access, we focus on inpatient treatment only. Although inpatient treatment deals with the most serious diseases and consumes the most healthcare resources, it is only part of the healthcare system for the Medicare beneficiaries. In addition, LOS and readmission only describe part of the medical burden from inpatient treatment. With more data, it would be of interest to expand the analysis to incorporate other types of treatment (e.g., outpatient treatment and drug prescription) and other outcomes (e.g., number of office visits and medical costs). It is noted that the proposed modeling approach is regression-based and can be easily expanded to accommodate more outcomes of different data types. In the analysis, we examine the temporal variations in a cross-sectional manner. It is a common approach that has been largely used in studies on temporal variations in the healthcare domain [89,90], including existing phenotype [27, 28] and clinical treatment [31, 77] HDNs. With the consideration that we have observed remarkable dynamics in network structures across time, it will be of interest to develop a time-varying network framework. This will require extensive methodology development and will be postpone to future studies. Another limitation associated with the proposed model is that its computation is more expensive as compared to generally used unconditional models. The graphical algorithm presented in this article deals with a large sample size, a large number of unknown parameters, and a complex model structure. Although estimation for different nodes and under different tuning can be run in a parallel manner to reduce computation time, for large data, further parallelization and distributed computing should be carried out.

Chapter 4

Comparative Effectiveness Research via Clinical Trial Emulation and a Big Data Approach

4.1 Testing the effectiveness and safety of rivaroxaban and dabigatran for atrial fibrillation via an emulation analysis of the Medicare data

4.1.1 Introduction

Atrial fibrillation (AF) is an abnormal heart rhythm characterized by rapid and irregular heart chamber beatings. It affects three to six million people in the U.S. alone [91], among whom 85% to 90% are eligible for oral anticoagulation therapies [92, 93]. As well established in the literature, it is of great significance to properly choose among non-vitamin K antagonist oral anticoagulants (NOACs). Rivaroxaban and dabigatran were approved by the U.S. FDA in November 2011 and October 2010, respectively. They were the first and second NOACs marketed for preventing stroke among people with non-valvular AF, and have been widely used since marketing. A study shows that in 2014, they accounted for 74% NOAC prescription during AF office visits [94].

In general, to draw definitive and objective conclusions on a drug's treatment effects, well-

controlled randomized clinical trials are needed. There have been multiple trials comparing dabigatran with Warfarin (for example, the RE-LY trial [95]) and rivaroxaban with Warfarin (for example, the ROCKET-AF trial [96]). These trials show that both rivaroxaban and dabigatran are non-inferior to Warfarin in preventing stroke and reducing the risk of bleeding. Recent effort has also been made to compare rivaroxaban with dabigatran. Examples include the DARING-AF and DANNOAC-AF studies [97,98]. However, they have limited sample sizes and broadly consider anticoagulants without focusing on rivaroxaban against dabigatran directly.

With the high popularity of these two drugs, it is of high interest to conduct a direct and focused comparative study. However, with the potentially high cost and the already broad usage of both drugs, such a randomized clinical trial is not foreseeable in the near future. To tackle this problem, we resort to the analysis of existing observational data, whose cost is negligible compared to that of a clinical trial. For this specific comparison problem, observational data has additional advantages. For example, it has been noted that the administration and adherence patterns of rivaroxaban and dabigatran are different. In the existing trials, according to the drug instructions, the rivaroxaban group took pills once a day together with an evening meal, and the dabigatran group took pills twice daily. It has been suggested that such differences can lead to different adherence patterns, and these differences further differ between controlled clinical trials and real-world practice [99]. From a public health perspective, real-world clinical practice, which also accommodates the effects of behavioral factors, can be of more interest. In this perspective, observational data may more honestly reflect the real-world treatment effects. In addition, as to be shown below, the available observational data is also considerably larger than many clinical trials on related regimens.

With observational data, the most straightforward approach is to conduct a multivariate regression analysis, as has been pursued on this topic in several publications [100–102]. It has been well recognized that the associations observed in regression, even after controlling for confounding, may not reflect the desired causal treatment effects. Causal inference with observational data is a "classic" and broad field. The existing literature is vast and has been reviewed in multiple publications. For relevant discussions, we refer to Hernan and Robins [36] and references therein. It is noted that different approaches have different advantages and disadvantages, and that there is no dominating approach. In this study, we adopt the emulation approach [37], which has been developed through a series of publications on a variety of illness conditions and treatment strate-

gies [103–106]. Under emulation, clinical trials are explicitly assembled using observational data, and statistical techniques for randomized clinical trials can be directly applied. As such, compared to causal inference techniques such as TMLE [34] and BART [35], emulation can have much more lucid interpretations.

In this article, our goal is to conduct a direct comparison of rivaroxaban and dabigatran and draw more definitive conclusions on their relative treatment effects, which can potentially have high public health implications. Without an actual clinical trial, we analyze the observational data in Medicare. This database has been chosen because of its broad coverage (hence more power and a better reflection of the general population in the U.S.) and high data quality. In addition, AF and the utilization of rivaroxaban and dabigatran can be much more prevalent in the Medicare population than others. This study may advance the existing literature in multiple aspects. First, it focuses on rivaroxaban and dabigatran and targets at conducting a direct comparison. Second, it adopts the emulation approach to explicitly "assemble" a clinical trial using observational data. The conclusion so drawn can be more informative than those from regression analysis and more lucid/interpretable than those using some other causal inference techniques. Third, it showcases a new way of analyzing the Medicare data to better inform clinical practice. Last, it also marks a new application of the emulation technique. In addition, we also examine temporal variations, which have been largely neglected in published emulation studies. Overall, this study may provide a useful complement to the existing AF, clinical trial, emulation, and Medicare data analysis literature.

4.1.2 Data source and study population

Medicare is a national health insurance program in the U.S. The Medicare claims are bills for services provided to the Medicare beneficiaries: adults aged 65 years and above, certain younger people with disabilities, and people with end-stage renal diseases [2]. Data analyzed in this study is extracted from the Centers for Medicare & Medicaid Services (CMS) Chronic Conditions Data Warehouse (CCW). The CMS CCW offers a wide range of claims data that follows beneficiaries across multiple care settings. For our analysis particularly, the Part A inpatient data and Part D prescription drug event data are used to identify eligible study subjects (more details in Section 4.1.3). The Medicare beneficiary identification number is used to link beneficiaries between the two datasets.

Our study focuses on the subpopulation with multiple chronic conditions (MCCs). This subpopulation has also been studied for cardiometabolic conditions [107, 108] and diabetes [109, 110], among others. Health-wise, this subpopulation is more vulnerable. In particular, 80% of AF happens to patients 65 years old and above [91], and AF is often associated with MCCs such as diabetes, dementia, and other heart diseases [92]. We further limit our study period to 2012-2013. The consideration is that by early 2012, both drugs were already extensively used (as such, a large sample size/sufficient power and broad representativeness can be achieved). A two-year study period is "normal" for a real phase III clinical trial. In addition, sufficiently detailed drug utilization information after 2013 is unavailable to our study. As such, we are unable to determine drug continuation or switch afterward. Overall, the source population includes 35,758,327 subjects (before patient selection), among whom 179,510 had taken rivaroxaban or dabigatran during the study period (more details in Section 4.1.3).

4.1.3 Methods

As established in the literature [37], under the emulation analysis framework, the first step is to develop a randomized clinical trial protocol. The trial can be a real one, a "hypothetical" one, or a mixture of both. Then an observational data analysis protocol is developed to emulate this trial. Accordingly, in section *the targeted randomized clinical trial*, we first develop the trial protocol. As there is no existing trial that directly compares rivaroxaban and dabigatran, our design has been based on relevant trials particularly including the RELY and ROCKET-AF trials as well as observational data analysis [111, 112]. The emulated trial is designed in section *the emulated trial*, with its analysis approaches described in section *analysis of the emulated trial*.

The targeted randomized clinical trial

The goal of the randomized clinical trial is to compare effectiveness (prevention of stroke and other thromboembolic events) and safety (bleeding and mortality) between standard-dose rivaroxaban and dabigatran in patients with both non-valvular AF and MCCs.

The first step is to define the inclusion/exclusion criteria. The target trial enrolls participants 65 years old and above, who are diagnosed with non-valvular AF within three months before enrollment and have MCCs. A participant is excluded if one or more of the following criteria are met: 1) a

prescription of an anticoagulant within three months before study entry, which may increase the risk of bleeding; 2) a NOAC preference different from the tested drugs; 3) indications other than AF for the respective anticoagulant (mitral valve disease, heart valve repair or replacement, deep vein thrombosis, pulmonary embolism, or joint replacement) in the three months preceding study entry; 4) being in a skilled nursing facility, receiving hospice care, or being in hospitalization; and 5) having had a kidney transplant or undergoing dialysis.

The trial starts on 01/01/2012. A three-month washout period is adopted with the consideration that anticoagulants used in AF are usually prescribed in 30-day or 90-day cycles. This ensures that there would be no previous usage of anticoagulants that potentially influence our outcomes of interest. The trial closes on 12/31/2013. An overall two-year study period is normal for peer cardiovascular trials. To gain insights into potential temporal variations (which have been largely neglected in both real and emulated trials), we further dissect the 21-month study period into three equal intervals. As such, there are a total of three trials, each with a seven-month recruitment period. As the whole study terminates on 12/31/2013, the three trials have different lengths of follow-up, which may provide insights into the impact of follow-up length on clinical trial observations.

After recruitment, an eligible participant is randomly assigned to one of two treatment arms. One arm receives rivaroxaban 20 mg once daily, and the other arm receives dabigatran 150 mg twice daily. Each participant is followed until death, loss to follow-up, or end of the study (12/31/2013). A loss to follow-up is defined as a discontinuation of treatment (a gap in the assigned treatment for over 60 days) or switching NOAC or dosage, whichever occurs first.

The primary outcomes include time to ischemic stroke, other thromboembolic events (systemic embolism, transient ischemic attack, pulmonary embolism), major bleeding, all-cause death, and the first occurrence of any of the above. The secondary outcomes include time to any bleeding and acute myocardial infarction (AMI).

The emulated trial

The emulated trial has been designed largely in line with that described above. More specifically, we first identify Medicare beneficiaries who filed the first prescription for either rivaroxaban or dabigatran with the standard dosages between 04/01/2012 and 12/31/2013. We start from 04/01/2012

to ensure participants' three-month medical history is available to determine whether the washout period eligibility is satisfied. We define the first prescription date as the index date. We include only individuals who were 65 years old or above, had at least three months of continuous enrollment in Medicare, and had MCCs. We exclude a subject if one or more of the following criteria are met: 1) a prescription of an anticoagulant within three months before the index date (has not been washed-out); 2) a diagnosis indicating a potential alternative indication other than AF for anticoagulation within three months before the index date; 3) being in a skilled nursing facility, receiving hospice care, or being in hospitalization; and 4) having had a kidney transplant or undergoing dialysis. We also exclude individuals with incomplete data for potential confounders at the baseline (0.54% percentage, thus we go for complete case analysis). Information on the list of confounders is provided in Table 4.1. Additional information on patient selection is provided in the flowchart in Figure 4.1, and related drug names and ICD-9-CM codes are provided in Appendix Table C.1 and C.2.

Then the eligible subjects are classified into two treatment groups: rivaroxaban and dabigatran. Follow-up information is extracted for each subject (to death, loss to follow-up, or end of the study – 12/31/2013). To identify the primary outcome, we track the study subjects to their first Medicare inpatient claims with the primary diagnosis code indicating each of the following: ischemic stroke, other thromboembolic events, major bleeding, and all-cause death. For secondary outcomes, we also identify all occurrences of any bleeding event (including major bleeding) and AMI in the same manner. Detailed information on the ICD-9-CM codes for identifying these events is provided in Appendix Table C.3.

We then create three subsequent emulated trials to examine potential temporal variations. The same criteria and procedures are applied to the three trials to ensure comparability. All trials then have a seven-month recruitment period. However, they have different lengths of follow-up. Detailed information on the sizes of each trial is provided in Figure 4.1.

Analysis of the emulated trial

It has been recognized that an emulated trial may still differ from its real counterpart in multiple ways. We refer to Hernan and Robins [37] and Danaei et al [104] for discussions. As such, the analysis of an emulated trial, although similar to that for a real trial in many ways, also has

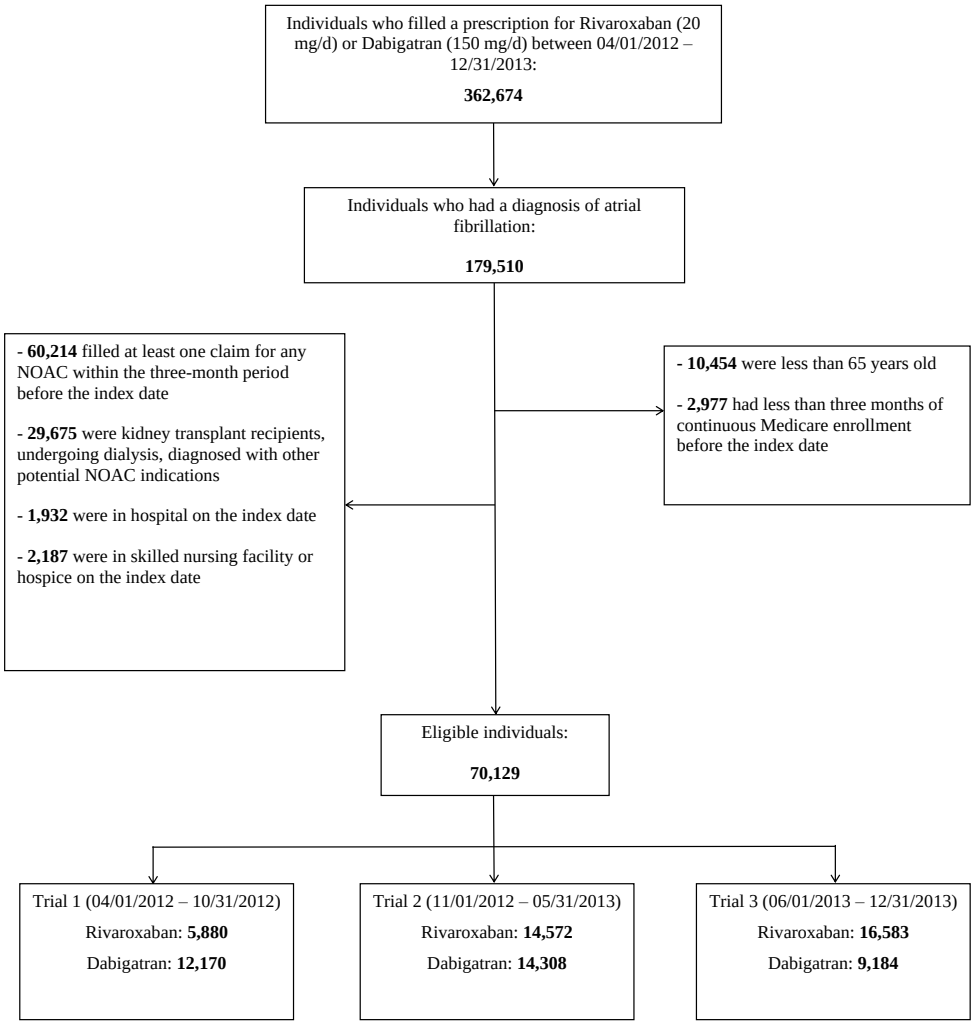


Figure 4.1: Flowchart of assembling clinical trials

differences.

First, to emulate the estimation of an intention-to-treat effect, which is common in real clinical trials, we use the observed therapy initiation (the first fill of rivaroxaban or dabigatran, the analog of random assignment in a clinical trial) as the treatment indicator. An observation is censored when a participant discontinued or switched therapy.

In a well-executed randomized clinical trial, different arms can be sufficiently balanced. As such, there may be no need for accounting for confounding – although it is noted that it is still commonly done out of caution. With emulated trials, since randomization does not really happen, there is a risk of imbalance. As such, as suggested in the literature [37], accounting for potential confounding is strongly recommended in emulation trial analysis. By reviewing the existing trials including RE-LY

and ROCKET-AF and published observational studies [111,112], and also taking into consideration data availability, we include the following variables as potential confounders: demographic variables (age, gender, race), medical conditions (AMI, congestive heart failure, previous stroke or transient ischemic attack, diabetes mellitus, hypertension, chronic kidney disease (CKD), history of bleeding, acquired hypothyroidism, and the number of other CMS comorbidities), and medication use (non-steroidal anti-inflammatory drug and antiplatelet). Information is also provided in Table 4.1.

For each emulated trial and each outcome, we fit a Cox proportional hazards model, including the treatment indicator and all baseline confounders as covariates. To further achieve balance as in a randomized clinical trial, we resort to the propensity score and inverse probability treatment (IPT) weighting approach. In particular, to calculate the propensity score, we estimate the probability of having a specific treatment using a logistic model and include the same set of baseline confounders as covariates [113]. Then the weight is calculated as the inverse of the propensity score for one treatment group and the inverse of one minus propensity score for the other group. By assigning such weights to study subjects, we create a pseudo-population, for which there exists no association between the baseline confounders and treatment.

With the three emulated trials, the above analysis can generate three sets of estimates. Beyond examining them individually, we also pooled the individual effect together and generated unified estimations. In particular, we adopt a random effect mixed model approach [114], which decomposes within-study and between-study variations and computes a weighted and combined effect.

Remarks

In some emulation studies [104,105,115], the sequential trial concept and corresponding techniques have been developed. Our setting has similarity with the sequential trials in that we also have three trials conducted in a consecutive manner. However, we note that the current setting can be somewhat simpler compared to that in Danaei et al [104], Caniglia et al [105], and some others. First, we study the direct comparison of two treatment drugs without having controls "waiting for" the switch to treatment, which creates significant overlaps between trials. Second, the number of subjects included in more than one trial is only 3.54%, dramatically smaller than many existing emulation studies. With such a small overlap, we are able to mostly ignore the potential correlations between trials. Another difference is that we allow temporal variations, as opposed to assuming

the same treatment effects. As such, we do not have the "inference with correlated trials" problem.

4.1.4 Results

Patient characteristics and unadjusted incidences

As shown in Figure 4.1, our analysis includes a total of 70,129 eligible participants, with 18,050, 28,880, and 25,767 in each of the three emulated trials, respectively. Table 4.1 shows the baseline characteristics by trial and treatment. It is observed that before IPT weighting, participants treated with rivaroxaban are slightly younger, more likely to be female, and with somewhat less cardiovascular disease history.

Variables	**Trial 1**		**Trial 2**		**Trial 3**	
	R	**D**	**R**	**D**	**R**	**D**
Demographics						
Age(years)						
65-74	2898 (49.29)	5850 (48.07)	6817 (46.78)	6553 (45.80)	8313 (50.13)	4433 (48.27)
75-84	2395 (40.73)	4935 (40.55)	6044 (41.48)	5947 (41.56)	6720 (40.52)	3745 (40.78)
≥ 85	587 (9.98)	1385 (11.38)	1711 (11.74)	1808 (12.64)	1550 (9.35)	1006 (10.95)
Sex (female)	2882 (49.01)	5910 (45.56)	7000 (48.04)	6540 (45.71)	7978 (48.11)	4155 (45.24)
Race						
White	5450 (93.19)	111016(90.94)	13478 (92.82)	13261 (93.11)	15185 (92.25)	8351 (91.46)
Black	192 (3.28)	533 (4.40)	563 (3.88)	512 (3.59)	678 (4.12)	428 (4.69)
Other	206 (3.52)	465 (4.66)	479 (3.30)	470 (3.30)	597 (3.63)	352 (3.85)
Medical history						
Diabetes mellitus	2327 (39.57)	5228 (42.96)	6071 (41.66)	6319 (44.16)	6984 (42.12)	4005 (43.61)
Hypertension	5382 (91.53)	11304 (92.88)	13496 (92.62)	13476 (94.19)	15287 (92.18)	8567 (93.28)
CKD	1359 (23.11)	3378 (27.76)	3659 (25.11)	3957 (27.66)	4339 (26.17)	2683 (29.21)
History of bleeding	128 (2.18)	290 (2.38)	370 (2.54)	402 (2.81)	416 (2.49)	270 (2.94)
Acquired HT	1670 (28.40)	3378 (27.76)	4210 (28.89)	4246 (29.68)	4840 (29.19)	2743 (29.87)
No. of other CMS comorbidities						
0-3	1621 (27.57)	3310 (27.20)	3672 (25.20)	3462 (24.20)	4499 (27.13)	2382 (25.94)
4-6	2836 (48.23)	5711 (46.93)	7106 (48.76)	7024 (49.09)	7930 (47.82)	4334 (47.19)
≤ 7	1423 (24.20)	2149 (25.88)	3794 (26.04)	3822 (26.71)	4154 (25.05)	2468 (26.87)
Cardiovascular disease						
AMI	429 (7.30)	940 (7.72)	1106 (7.59)	1092 (7.63)	1317 (7.94)	659 (7.18)
CHF	2501 (42.53)	5912 (48.58)	6618 (45.42)	7256 (50.71)	7227 (43.58)	4484 (48.82)
Stroke or TIA	1189 (20.22)	2853 (23.44)	3094 (21.23)	3351 (23.42)	3616 (21.81)	2163 (23.55)
Medication use						
NSAIDs	510 (8.67)	1054 (8.66)	1072 (7.36)	890 (6.22)	1355 (8.17)	616 (6.71)
Antiplatelet	760 (12.93)	1291 (10.61)	1577 (10.82)	1044 (7.30)	1977 (11.92)	711 (7.74)

R rivaroxaban, D dabigatran, CMS centers for Medicare & Medicaid Services,CKD chronic kidney disease, AMI acute myocardial infarction, CHF congestive heart failure, TIA transient ischemic attack, HT hypothyroidism, NSAIDs non-steroidal anti-inflammatory drug

Table 4.1: Baseline characteristics of the study cohorts, before IPT weighting

We then examine the temporal variations of the baseline characteristics across the three emulated trials. It is observed that for both the rivaroxaban and dabigatran arms, most of the baseline

characteristics do not change substantially over time. We do observe that the number of patients with CKD obviously rose during the two years. For rivaroxaban, the percentage of patients with CKD rose from 23.11% to 25.11% and then to 26.17%. For dabigatran, the percentages are 27.76%, 27.66%, and 29.21%, respectively. A similar increase is also observed in vulnerable patients with medical history (for example, having a history of bleeding, stroke, or acquired hypothyroidism) during the two-year period.

Table 4.2 shows the number of unadjusted incidence by trial and treatment. For the three trials, the numbers of participants that developed any of the primary outcomes are 1,717, 2,818, and 1,525, respectively; the numbers of participants lost to follow-up are 12,285, 15,333, and 8,850, respectively; and the numbers of participants that reached the end of follow-up without any primary outcome are 4,048, 10,729, and 15,392, respectively. For rivaroxaban, the average follow-ups are 273, 201, and 85 days for the three trials, respectively; For dabigatran, the average follow-ups are 255, 210, and 97 days, respectively.

Adjusted hazard ratio, temporal variation, and pooled results

As shown in Appendix Table C.4, after the IPT weighting, all sample characteristics are more balanced between the rivaroxaban and dabigatran groups. This is further "confirmed" by Appendix Figure C.1, which shows almost identical propensity score distributions for the weighted trial 1 cohort. Similar plots for the other two trials are omitted here.

Table 4.3 shows the adjusted hazard ratio and p-value for each outcome by trial and treatment. For any primary outcome, rivaroxaban is found to have a significantly higher hazard ratio in all three trials. However, the magnitudes of the estimated hazard ratio and p-values show considerable temporal variations. There are also significant temporal variations for the two primary effectiveness outcomes. Rivaroxaban is found to have a significant protective effect for ischemic stroke in trials 1 and 2, but non-significantly increases its risk in trial 3. To get more insight into this difference, we show in Figure 4.2 the Kaplan-Meier survival curves of ischemic stroke outcome as an example. We acknowledge the difference in follow-up periods for the three trials, but also note that the survival curves for the three trials differ significantly in the early days of follow-up. As such, the observed difference in treatment effect cannot be fully attributed to the follow-up difference. For other thromboembolic events, the two drugs are not significantly different in all three trials. However,

Trial	Study period	Primary outcome										Secondary outcome			
		Any primary events		Ischemic stroke		Other thromboembolic events		Major bleeding		All-cause death		Any bleeding		AMI	
		R	D	R	D	R	D	R	D	R	D	R	D	R	D
1	04/01/2012-	697	1344	143	372	63	151	417	683	117	243	567	827	48	98
	10/31/2012	(11.85)	(11.04)	(2.43)	(3.06)	(1.07)	(1.24)	(7.09)	(5.61)	(1.99)	(2.00)	(9.64)	(6.80)	(0.82)	(0.81)
2	11/01/2012-	1599	1297	323	379	134	122	898	618	289	259	1211	764	88	91
	05/31/2013	(10.70)	(9.06)	(2.22)	(2.65)	(0.92)	(0.85)	(6.16)	(4.32)	(1.98)	(1.81)	(8.31)	(5.34)	(4.60)	(0.64)
3	06/01/2013-	1060	488	257	145	106	53	570	206	179	103	762	261	64	41
	12/31/2013	(6.39)	(5.31)	(1.55)	(1.58)	(0.64)	(0.58)	(3.44)	(2.24)	(1.08)	(1.12)	(4.60)	(2.84)	(0.39)	(0.45)

Individuals may develop multiple primary outcomes (for example, first ischemic stroke and then death)
R rivaroxaban, D dabigatran, AMI acute myocardial infarction

Table 4.2: Number of events at the end of follow-up by treatment group in each trial

the estimated hazard ratio switches sign between trial 1 and the other two. For all the safety outcomes (all-cause mortality, major bleeding, and any bleeding), rivaroxaban is observed to be significantly inferior. Again, it is observed that the magnitudes of the estimated hazard ratios and p-values vary across trials. For AMI, there are no significant differences between the two drugs for all three trials.

Trial 1	**Hazard Ratio (HR)**	**p-value**
Primary outcome Any primary events	1.072	0.0403
Ischemic stroke	0.799	0.0004
Other thromboembolic events	0.935	0.4892
Major bleeding	1.119	0.1280
All-cause mortality	1.252	<.0001
Secondary outcome		
Any bleeding	1.413	<.0001
AMI	0.979	0.8549
Trial 2	**HR**	**P-value**
Primary outcome		
Any primary events	1.266	<.0001
Ischemic stroke	0.892	0.0333
Other thromboembolic events	1.133	0.1595
Major bleeding	1.237	0.0004
All-cause mortality	1.510	<.0001
Secondary outcome		
Any bleeding	1.661	<.0001
AMI	0.954	0.6517
Trial 3	**HR**	**P-value**
Primary outcome		
Any primary events	1.380	<.0001
Ischemic stroke	1.085	0.2490
Other thromboembolic events	1.228	0.0692
Major bleeding	1.183	0.0480
All-cause mortality	1.752	<.0001
Secondary outcome		
Any bleeding	1.867	<.0001
AMI	0.884	0.3699
Pooled Results	**HR**	**P-value**
Primary outcome		
Any primary events	1.232	0.0025
Ischemic stroke	0.915	0.2819
Other thromboembolic events	1.086	0.2934
Major bleeding	1.187	<.0001
All-cause mortality	1.488	<.0001
Secondary outcome		
Any bleeding	1.633	<.0001
AMI	0.944	0.3988

AMI acute myocardial infarction

Table 4.3: Adjusted hazard ratio and p-values for the primary and secondary outcomes

The pooled hazard ratio shows that, compared to dabigatran, rivaroxaban has a significantly increased risk of any primary outcome. It non-significantly reduces the risk of ischemic stroke and

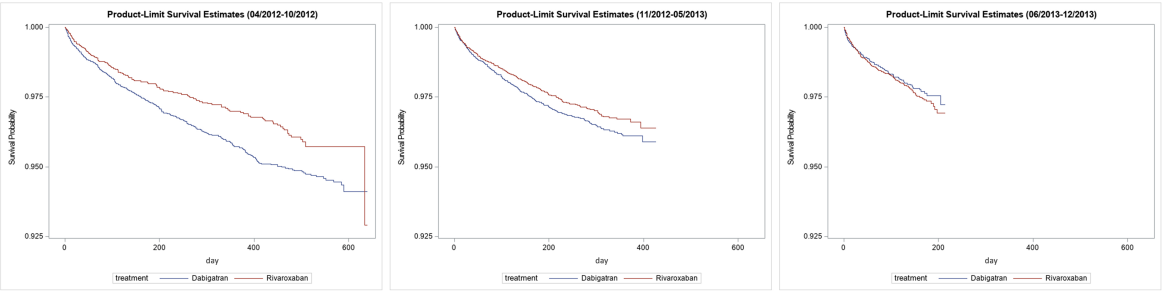


Figure 4.2: Survival curves of time to ischemic stroke

AMI, non-significantly increases the risk of other thromboembolic events, and significantly increases the risk of all the safety outcomes (all-cause mortality, major bleeding, and any bleeding).

4.1.5 Discussion

The effectiveness of rivaroxaban and dabigatran for AF has been well established. It is not our goal to challenge or re-establish this conclusion. Rather, considering that both have been extensively used in clinical practice, our goal is to conduct a direct comparison, which has been insufficiently pursued in the literature. Without being able to conduct a real trial, we have resorted to the emulation approach to "assemble" trials based on observational data to draw causal statements. We note that, in terms of statistical measures (accuracy, efficiency, etc.), there is no definitive conclusion on the performance of different causal inference techniques. The adopted emulation approach can be preferred with more intuitive interpretations. In addition, as partly shown in this article, it can also provide guidance to the development of a real trial (if it ever becomes feasible). The potential limitations of emulation have been well discussed in the literature [36, 37, 104, 115] and will not be repeated here.

We have analyzed the Medicare data, which is the most comprehensive and largest medical record database for seniors in the U.S. Nevertheless, it has limitations. Specifically, as a medical claims database, it does not include certain information [116]. As such, we are not able to verify if the two treatment groups are also balanced in other potentially relevant but unmeasured aspects. We note that this limitation is also shared by quite a few other emulation and other causal inference studies [103,104,106,115]. Medicare has an almost universal coverage in the U.S elderly population. As such, the findings can be applied to this population with reasonable confidence. Although there

is no indication that the treatment effects of rivaroxaban and dabigatran depend on race and other demographic factors, the applicability of our findings to other populations still needs further examination. Moreover, we have had limited data access, which has led to a not current dataset, a limited sample size, and a limited follow-up (we note that drug usage information is not available in the "generic" Medicare database). On the other hand, we also note that there has been no indication in the literature that the treatment effects of the two drugs differ between the study and other periods. The sample size in this study has already been considerably larger than its peers [101, 111, 117, 118]. In addition, there has been a considerable number of events, even in the third trial with the shortest follow-up.

Our analysis has led to the following main findings. The first is that, based on both the individual trials and pooled effects, dabigatran is found to be significantly superior to rivaroxaban in terms of reducing safety concerns (bleeding events and overall mortality). This finding is consistent with the existing observational data analyses [111,112,117,119]. As discussed in Graham [112], the higher bleeding risk and mortality of rivaroxaban may be the side effects of the once-daily regimen, leading to higher peak and lower trough serum concentrations than twice-daily administrated dabigatran. Another finding is that rivaroxaban and dabigatran are not significantly different in preventing ischemic stroke and other thromboembolic events. In some of the existing observational studies, results are non-conclusive regarding the preventive effects of rivaroxaban versus dabigatran [111, 112, 117–121]. This can be caused by differences in study populations and analysis techniques. As we do not have access to other data, it is impossible to exactly pin down the causes. An interesting finding, which has been largely neglected in the literature, is the temporal variations. First, it is observed that in the study period, the prescription of rivaroxaban was increasingly preferred over dabigatran. In trial 1, the rivaroxaban group accounts for 32.58% of the whole study sample, whereas in trial 3, it accounts for 64.36%. Southworth et al. suggested that this change in prescription pattern might be attributed to that post-marketing reports caused misconceived bleeding risks with dabigatran [122]. In addition, patients initiating rivaroxaban are slightly younger and healthier. Interestingly, some other studies have observed opposite prescription patterns. Because rivaroxaban has been studied in older and higher-risk populations, some physicians are more likely to prescribe rivaroxaban to high-risk elderlies [120]. Some other studies have also found that rivaroxaban patients are older and with more comorbidities. One plausible

explanation for the difference between this and other studies is that, our trial enrollment period includes the early stage after rivaroxaban's new approval, and physicians often tend to use familiar (and seemingly safer) drugs for sicker patients (it is noted that we have focused on patients with MCCs). This may also partly explain the increasing prescription of rivaroxaban over the two-year period.

This study may be the first to observe that some findings on treatment effects also have temporal variations. In the existing emulation studies, the study time periods are usually not strongly justified, with the underlying belief (or assumption) that the findings should not depend on time periods. As can be partly seen from the survival curves, observed differences are obvious. A deeper examination suggests that, although there are some differences between trials in terms of demographic and clinical characteristics, such differences are not sufficient to explain the treatment effect differences. This analysis may raise the alarm for other emulation studies, which may also face temporal variations.

4.1.6 Conclusion

This study has provided a direct and objective comparison of the treatment effects of rivaroxaban and dabigatran via conducting an emulation analysis of the Medicare data. The findings may assist clinicians making more informed decisions in the treatment of senior AF patients with MCCs. This study also provides another demonstration of applying the emulation technique to draw causal conclusions based on observational medical record data. Despite certain limitations, it is expected to be informative for various stakeholders, including clinicians, patients, and biomedical/statistical researchers.

4.2 Evaluation of survival outcomes of endovascular versus open aortic repair for abdominal aortic aneurysms with a big data approach

4.2.1 Introduction

Abdominal aortic aneurysm (AAA) is a balloon-like dilatation of the aorta that supplies blood to the body and happens below the chest. Each year, it is estimated that 200,000 people in the U.S. are diagnosed with AAA, and ruptured AAA (rAAA) poses significant clinical and public health challenges [123]. rAAA is associated with an overall mortality rate of over 80%, which causes more than 5,000 deaths in the country each year [124, 125]. Once rAAA occurs, repairing procedures need to be conducted immediately. In the current clinical practice, there are two main approaches: emergent open aortic repair (OAR) and endovascular aortic repair (EVAR). OAR has a relatively longer history and is still considered as the standard procedure for AAA repair, during which large incisions are unavoidable [126]. EVAR was first successfully conducted and reported in year 1994, and only small incisions in the groins are needed [127]. However, this circumvented procedure makes EVAR require more intense monitoring and probable reintervention [33]. Moreover, preoperative imaging and specific anatomic requirements make EVAR less well suitable for emergent rAAA. As suggested in multiple studies [128–132], the preferred minimum invasion but awaited long-term postoperative complications may account for the favorable 30-day mortality but similar or even inferior late survival of EVAR compared to OAR. With the criticalness of rAAA and prevalence of EVAR and OAR, it is of significant interest to objectively evaluate and directly compare their survival outcomes.

In general, to compare the effects of two treatments, the gold-standard approach is to conduct a randomized controlled clinical trial. However, most of the existing clinical trials have focused on patients who have elective/intact AAA (eAAA/iAAA) and excluded those who have rAAA and require emergent care (e.g., OVER [128], DREAM [129]). This is highly sensible as patients with rAAA cannot bear the prolonged process of eligibility examination, treatment assignment, and finally, surgical procedure, which are non-negligible steps in a clinical trial for bias control but unacceptable for saving lives in a real-world setting.

With the aforementioned concerns, researchers have focused on observational data and analysis to investigate the survival outcomes of the two procedures for rAAA patients. Our literature review suggests that quite a few of them have relied on large medical claims databases including Medicare [33, 131, 133, 134]. In these studies [33, 131], regression and other association analysis techniques have been the main tools. It is well recognized that such analyses, even after accounting for confounders, can only lead to conclusions on association, as opposed to the desired cause-and-effect relationship. To overcome such limitations, causal inference techniques [35, 36, 135] can be adopted. Here we note that, with extensive examinations and comparisons, no approach has been observed to dominate others – it is expected that such an approach may not exist, and different approaches have different pros and cons. In this article, we adopt the emulation approach, which is relatively new but has already been examined in many publications [38, 136, 137]. With this approach, a clinical trial is explicitly designed and assembled using observational data, and statistical analysis approaches designed for clinical trials can be then adopted, leading to causal statements. Comparatively, the biggest advantage of this approach may be its lucid interpretations.

Built on the emulation strategy, we take a big data analysis approach. Here "big data" is manifested in at least two perspectives. The first is that our effort is built on the Medicare data. The Medicare database is massive, covers the dominating majority of the U.S. senior population, and contains comprehensive information. Compared to for example hospital- and community-based data, Medicare data is advantageous with its unbiased sample selection and relatively uniform and detailed data collection. It has served as the basis of a large number of clinical and public health studies, including those that adopt causal inference analysis techniques [38, 138]. More details on the analyzed Medicare data are provided below in Section 4.2.2. The second big data perspective is that in analyzing the emulated trial, deep learning techniques are adopted. In "standard" emulation analysis (as well as most if not all analysis of real clinical trials) [38, 137], regression (e.g., logistic and Cox) techniques have been adopted. For diverse fields including engineering, business, social science, and others [139,140], the superiority of deep learning techniques in prediction has been well established through a myriad of published studies. Relatively recently, deep learning techniques have been applied to biomedical studies on cancer [141], fracture [142], chronic diseases [143], and cardiovascular diseases [144]. The studied outcomes/phenotypes include continuous [143], categorical [142], and, more recently, survival [145]. It is noted that the existing deep learning

analyses of biomedical data are mostly in the association analysis domain.

The overarching goal of this study is to directly compare EVAR versus OAR for rAAA patients and draw conclusions as close to causal as possible, so as to further inform clinical practice. This study may advance from the existing literature in multiple aspects. First, it strives to compare the treatment effects of EVAR and OAR under the clinical trial framework, as opposed to the commonly adopted observational data analysis framework. Second, it advances from the existing emulation analyses by investigating a new disease condition and treatments, which have critical clinical importance. In addition, deep learning techniques, as opposed to "simple" regressions, are adopted. This study may assist introducing deep learning to the emulation paradigm. Third, it may also foster deep learning research. More specifically, this is the first application of deep learning to the emulation analysis and study of rAAA. Built on the existing deep learning components, we assemble an analysis pipeline that mimics the "propensity score + inverse probability treatment – IPT weighting Cox regression" approach (which has been adopted in the existing emulation analyses [38, 137]).

Looking at a higher level, an "ordinary" clinical trial generates an information set (target), whose most notable characteristic is the balance in information between two treatment arms. In addition, it is usually assumed that such an information set can be sufficiently described using a (semi)parametric model. Information contained in observational data fundamentally differs from the target. As such, a central goal of the emulation approach is to properly carve a piece of information, as large as possible, that mimics the target. With the deep learning analysis approach, the (semi)parametric probabilistic structure can be significantly relaxed. Overall, this study falls into the intersection of information theory and machine learning.

4.2.2 Methods

Data source

As briefly mentioned above, we analyze the Medicare data in this study. Medicare is a federal health insurance program for adults aged 65 years and above, certain younger people with disabilities, and people with end-stage renal disease (permanent kidney failure requiring dialysis or a transplant). As the single largest payer of health care in the U.S., it covers 98% of adults who are over 65

years old, accounts for 99% of death in the elderly population, and generates a huge amount of medical claims data [146]. The centers for Medicare & Medicaid Services (CMS) offers a wide range of datasets that follow Medicare beneficiaries across multiple care settings. More specifically, it collects over two billion data points per year through reimbursement to hospital care (Medicare Part A), physician and outpatient services (Medicare Part B), drug prescription (Medicare Part D), and other health care claims. It also collects billions of other data points through enrollment information, beneficiary eligibility checks, quality metrics, and calls to 1-800-MEDICARE [4].

For our study, we first retrieve all inpatient claims between 01/01/2011 to 09/30/2015 from the Medicare provider utilization and payment data: hospital care (Part A), which contains detailed information on health services provided in 54 million inpatient episodes for 23 million Medicare beneficiaries. Information contained in each claim includes beneficiary demographics (e.g., age, sex, race), Medicare enrollment status, services provided (up to 25 diagnosis codes and up to 25 procedure codes), and beneficiary death information. More details on such information and how it is utilized in our analysis are provided below.

It is noted that for research purposes, the Medicare data can be viewed as publicly available. We only conduct secondary analysis of the existing deidentified data. As such, no IRB or other approvals are needed.

The target randomized clinical trial

Under the emulation analysis paradigm [37], one of the first and most important steps is the design of a target randomized clinical trial. For treating rAAA, there is a lack of real clinical trials. As such, similar to in some literature [38], we need to design a hypothetical target trial. The following design has been motivated by relevant observational studies [33, 131–134] and is clinically well grounded.

The target randomized clinical trial aims to compare the short- and long-term all-cause mortality of rAAA patients treated with EVAR and OAR. More specifically, we enroll participants who are diagnosed with rAAA within the enrollment period and exclude those who meet any of the following criteria: 1) the participant is under 65 years old at enrollment; 2) conversion between EVAR and OAR is necessary after randomization; 3) the participant has concurrent conditions of thoracic aneurysms, thoracoabdominal aneurysms, or aortic dissection; and 4) a repair of the thoracic aorta

or visceral or renal bypass is considered necessary for the participant. If a participant develops multiple cases of rAAA during the enrollment period, only the first is considered as the primary case and included in analysis. Such criteria have been motivated by observational studies [33, 132] and data availability, and have the same level of rigor as a real clinical trial.

The trial enrolls participants from 01/01/2011 to 09/30/2015. After enrollment, each eligible participant is randomized to receive either EVAR or OAR and followed until death, loss to follow-up, or end of the study (06/30/2019). Such decisions have been made with the considerations that both treatments have been extensively adopted in the study period, the enrollment is long enough to ensure a sufficient sample size, and the follow up is long enough to ensure a sufficient effective sample size.

To assess both short- and long-term mortality after EVAR and OAR, we define two primary outcomes: time from treatment to short-term perioperative mortality and time from treatment to long-term all-cause mortality. The short-term perioperative mortality is defined as death during the index hospitalization or within 30 days of discharge, for which all participants alive at 30 days after discharge are censored. For the long-term all-cause mortality, a subject is censored at loss to follow-up or end of the study (06/30/2019), whichever comes first. The two survival outcomes have different implications but are both critically important [132].

The emulated trial

To emulate the target randomized clinical trial described above, we develop an emulated trial using the Medicare claims data. The strategy closely follows that developed in the emulation literature [38]. First, we identify Medicare beneficiaries who were diagnosed with rAAA and underwent EVAR or OAR between 01/01/2011 and 09/30/2015. We exclude individuals that met any of the following criteria: 1) the individual was under 65 years old at diagnosis; 2) both EVAR and OAR were present in the same index hospitalization, which indicated conversion; 3) concurrent diagnosis codes of thoracic aneurysms, thoracoabdominal aneurysms, or aortic dissection; 4) concurrent procedure codes of repair of the thoracic aorta or visceral or renal bypass; and 5) less than 12 months of Medicare enrollment before the index hospitalization. If a beneficiary had multiple eligible claims, only the first was considered as the primary case and included in analysis. Additional information on patient selection is provided in the flowchart in Figure 4.3. The relevant International Classification

of Diseases, Ninth Revision, Clinical Modification (ICD-9-CM) codes are provided in Appendix Table C.5.

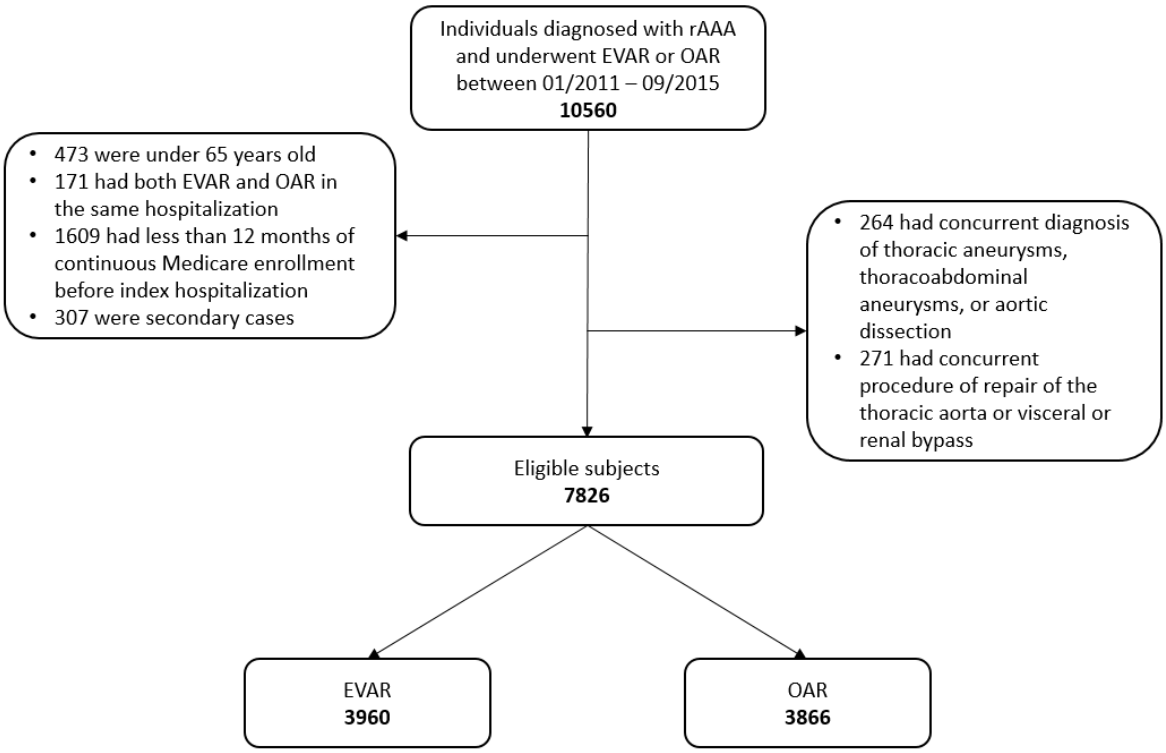


Figure 4.3: Flowchart of cohort definition

We then classify each eligible subject into one of the two treatment groups: EVAR and OAR, based on the procedure he/she actually received. Follow-up information is then extracted for each subject (to death, loss to follow-up, or end of the study which is 06/30/2019). A loss to follow-up is defined as discontinuation of Medicare enrollment. To identify the primary outcomes, we track each study subject from treatment to his/her documented death. We note that there are 5.35% study subjects for whom the date of treatment is missing. For these subjects, we use the date of admission to approximate the date of treatment, since rAAA is an emergent condition that needs immediate treatment, and the average lag time between admission and procedure is 0.53 days in our cohort.

Data analysis

This study has survival outcomes. If this were a real clinical trial, analysis could be conducted using a Cox model. Although balance is expected with proper randomization, to be cautious, in clinical trial analysis, potential confounders are still commonly adjusted. For an emulated trial with a survival outcome, published studies [38, 137] suggest the following main analysis steps: (a) conduct a propensity score analysis for treatment using the logistic regression approach, and (b) conduct a Cox regression analysis for survival with inverse probability treatment – IPT weighting.

As briefly mentioned in Section 4.2.1, deep learning has demonstrated promising performance with biomedical data. It is of significant interest to apply it to emulation. Equally importantly, the analysis presented in the Appendix C.2.2 shows that the Cox proportionality assumption is not satisfied. The deep learning approach described below, although has some connections with the Cox model, can be more flexible and less dependent on model assumptions, with its "built-in" flexibility. It consists of the following steps: (a) generate propensity scores for treatment using a single-layer neural network. This corresponds to the logistic regression mentioned above; (b) construct a multi-layer neural network for survival. Advancing from the "standard" deep learning survival, we incorporate weights generated in Step (a), which corresponds to the IPT weighted Cox regression mentioned above; and (c) advancing from the existing deep learning literature, we also conduct a bootstrap-type procedure to gain insights into the variation of the neural network weight estimation, which is analogous to the regression coefficient estimation and reveals the treatment effects.

Denote n as the number of independent subjects. For subject i, denote C_i as the censoring time and T_i as the event time. We observe the right-censored survival outcome $Y_i = min(T_i, C_i)$ and censoring indicator $d_i = I(T_i \leq C_i)$ with $I(\cdot)$ being the indicator function. Denote $X_i = (X_{i1}, X_{i2}, ..., X_{ip})$ as the baseline covariates and Z_i as the binary treatment assignment.

Step 1: We employ a single-layer neural network to estimate the propensity score, which is the probability of treatment assignment conditional on the baseline covariates. In particular, the input includes the covariates described in Table 4.4, with standardization for the continuous variables and coding for the categorical variables. The labels in the data are the binary treatment assignment variables. For the neural network architecture, we use Rectified Linear Units (ReLU)

	EVAR (N=3930)	**OAR** (N=3866)	P-value*
Demographic			
Age, mean(sd)	78.03 (7.52)	76.59 (6.90)	<.0001
Male	3023 (76.34)	2786 (72.06)	<.0001
Race			0.0030
White	3588 (90.77)	3550 (92.02)	
Black	249 (6.30)	178 (4.61)	
Other	116 (2.93)	130 (3.37)	
Medical conditions			
Congestive heart failure	464 (11.72)	299 (7.73)	<.0001
Cardiac arrhythmia	596 (15.05)	438 (11.33)	<.0001
Valvular disease	199 (5.03)	172 (4.45)	0.2304
Coronary disease	758 (19.14)	603 (15.60)	<.0001
Diabetes	329 (8.31)	256 (6.62)	0.0045
Hypertension	1250 (31.25)	1078 (27.88)	0.0004
Chronic obstructive pulmonary diseases	707 (17.85)	584 (15.11)	0.0011
Clinically significant lower extremity vascular diseases	26 (0.66)	27 (0.70)	0.8215
Renal atherosclerosis	20 (0.51)	27 (0.70)	0.2684
Vascular intestine disease	7 (0.18)	2 (0.05)	0.1027
Renal failure	493 (12.45)	358 (9.26)	<.0001
Other renal diseases	3 (0.08)	1 (0.03)	0.3289
Kidney transplant	4 (0.10)	3 (0.08)	0.7291
Liver disease	33 (0.83)	30 (0.78)	0.7766
Cerebrovascular diseases and paralysis	93 (2.35)	67 (1.73)	0.0544
Other neurological diseases	153 (3.86)	114 (2.95)	0.0258
Hyperlipidemia	817 (20.63)	687 (17.77)	0.0013
Cancer	132 (3.33)	87 (2.25)	0.0037
Rheumatoid arthritis	76 (1.92)	39 (1.01)	0.0008
Prior intact AAA diagnosis	511 (12.90)	440 (11.38)	0.0393
Other			
Year in which repair was performed			<.0001
2011	808 (20.40)	1013 (26.20)	
2012	869 (21.94)	913 (23.62)	
2013	819 (20.68)	785 (20.31)	
2014	837 (21.14)	701 (18.13)	
2015	627 (15.83)	454 (11.74)	
Outcome (followed until death, loss to follow-up, or 06/30/2019)			
All-cause mortality	2430 (61.36)	2542 (65.75)	<.0001
Perioperative mortality (in-hospital or 30 days after discharge)	1107 (27.95)	1704 (44.08)	<.0001

* P-values are based on t-tests for continuous variables and Chi-squared test for categorical variables

Table 4.4: Descriptive characteristics of the study cohort

as the activation function, sigmoid activation function to produce the probability output, and logarithmic loss function (binary cross-entropy). For optimization, a stochastic gradient descent algorithm with Nesterov momentum is used, and a grid search is conducted to tune the learning rate. For such tasks, we adopt the open-source python module keras (*https://keras.io*). With the outputted propensity score, we compute the IPT weight as its inverse for a subject in one treatment group and the inverse of one minus propensity score for a subject in the other group.

Step 2: Here we conduct the IPT weighted survival analysis. The input includes the same set

of covariates and treatment indicator as in Step 1, as well as the IPT weights computed above. For subject i, denote w_i as the IPT weight and $R_i = j : T_j > T_i$ as the at-risk set (at time T_i). We consider a neural network with two hidden layers and the number of nodes determined by tuning. Denote θ as the weights that characterize the network (note that they are not the IPT weights), and $g_\theta(X_i, Z_i)$ as the output for subject i. Partly motivated by the loss function under the Cox regression as well as recent deep learning studies, such as DeepSurv, we consider the objective function:

$$l(\theta) = -\frac{1}{\sum_i d_i} \sum_{i=1}^{n} d_i w_i [g_\theta(X_i, Z_i) - log \sum_{j \in R_i} exp(g_\theta(X_j, Z_j))].$$

For optimization, we adopt a gradient descent approach. ReLU is used as the activation function, and the adaptive moment estimation algorithm (Adam) for gradient descent optimization with a cyclical learning rate method is adopted. We perform a grid search for hyper-parameter tuning. The computational program is developed based on the open-source python module pycox (*https://github.com/havakv/pycox*).

Step 3: A procedure similar to the 0.632 bootstrap for regression analysis [147] is conducted. In particular, $0.632n$ samples are randomly selected from the original data without replacement. With the bootstrapped samples, the above analysis is conducted, and the neural network weight estimates are extracted. This is repeated multiple (e.g. 1,000) times to assess the variability of estimates. For regression, the 0.632 bootstrap is equivalent to the "n-out-of-n with replacement" bootstrap. By sampling without replacement, it can reduce ties and computational cost.

The above analysis can deliver the following. The first is a propensity score estimate for each subject. If needed, the weights of the neural network can be extracted to help assess the relative contributions of covariates. The second is the survival neural network. For a subject with a set of known confounder values and treatment assignment, it can generate the (relative) survival risk. Most of the existing deep learning studies have treated neural networks as black boxes. As we conduct a clinical trial analysis, the effect of the treatment is of the most essential interest. As such, we retrieve the estimated weights for the treatment indicator and confounders. With the presence of hidden layers, the weight matrices need to be multiplied across layers to obtain the overall contributions. The third product is that, for the (overall) weight of the treatment indicator, the bootstrap-type analysis can generate an evaluation of its variability. The same is also applicable

to the confounders.

Remarks For binary responses, the superiority of neural networks over logistic and other regressions has been demonstrated in a large number of publications [141,142]. Several recent publications, such as DeepSurv [145] and Cox-nnet [148] and others, seem to suggest similar superiority for survival data. As our goal is to take advantage of the recent deep learning developments, we choose not to "re-establish" the merit of deep learning. We also note that there are multiple "base techniques" for building neural networks. The adopted ones have been shown in recent studies as having competitive performance. To the best of our knowledge, there is still no study showing that certain techniques dominate the others.

4.2.3 Results

Patient characteristics and unadjusted incidences

Our analysis includes 7,826 eligible subjects, with 3,960 in the EVAR arm and 3,866 in the OAR arm. The summary statistics are shown in Table 4.4. It is observed that the study subjects were slightly younger in the OAR arm, and more likely to be white males in both arms. Participants in the OAR arm were healthier with lower percentages of almost all medical conditions (except for two rare conditions: clinically significant lower extremity vascular diseases and renal atherosclerosis). It is also observed that, as time passed by (from year 2011 to 2015), the rAAA patients were more and more likely to receive EVAR. Here we note that, without the IPT weighting, all demographic variables and most medical condition variables are significantly unbalanced between the two treatment arms, highlighting a significant difference between real clinical trials and observational data. Table 4.4 also shows the unadjusted incidence rates by treatment. The EVAR arm has a slightly lower unadjusted incidence rate for long-term all-cause mortality and a significantly lower unadjusted incidence rate for short-term perioperative mortality.

Analysis of the emulated trial

Prior to analysis, we delete 15 records with missing measurements (7 in the EVAR arm and 8 in the OAR arm). Analysis is conducted using the approach described in Section 4.2.2. For the propensity score analysis, the baseline covariates include age, gender, race, year in which repair

was performed (this variable has been considered in the published observational studies [33,131]; it is also motivated by the changing rates of EVAR and OAR), and 20 medical conditions, as shown in Table 4.4 (related ICD-9-CM codes in Appendix Table C.5). For survival analysis, the same baseline covariates and treatment indicator are included.

For both the propensity score and survival analysis, the obtained fully connected neural network architectures are available from the authors. For the propensity score analysis, the learning rate is tuned as 0.008. The distributions of propensity scores are shown in Figure 4.4. Minor differences between the two arms are observed.

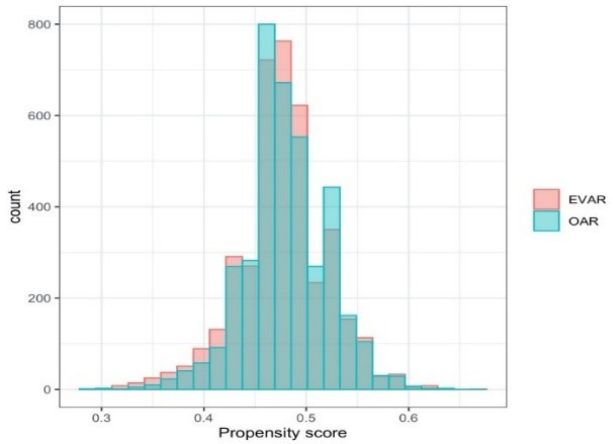


Figure 4.4: Distribution of propensity score

For the analysis of short-term survival, the learning rate for Adam optimizer is tuned as 0.016. The analysis results are summarized in Figure 4.5. The left panel shows the estimated survival curves, after accounting for IPT weights, for the two treatments separately. With the bootstrap procedure, we are also able to obtain the pointwise 90% confidence intervals. It is noted that this analysis mimics the "familiar" regression analysis and differs from most of the existing deep learning studies. EVAR is observed to have a modest survival advantage, with the lower bounds of its confidence intervals almost coinciding with the upper bounds of OAR's confidence intervals. Based on the estimated survival curves, we compute the expected survival under EVAR as 83.5 days, compared to 79.2 days under OAR. In the right panel of Figure 4.5, the forest plot, which shows the medians as well as the 25% and 75% quantile values of the overall estimated weights ("accumulated" over layers), again suggests the survival advantage of EVAR. The right panel of Figure 4.5 also contains weight information for confounders that demonstrate considerable and "persistent" effects

(across the bootstrapped datasets), including race and seven medical conditions.

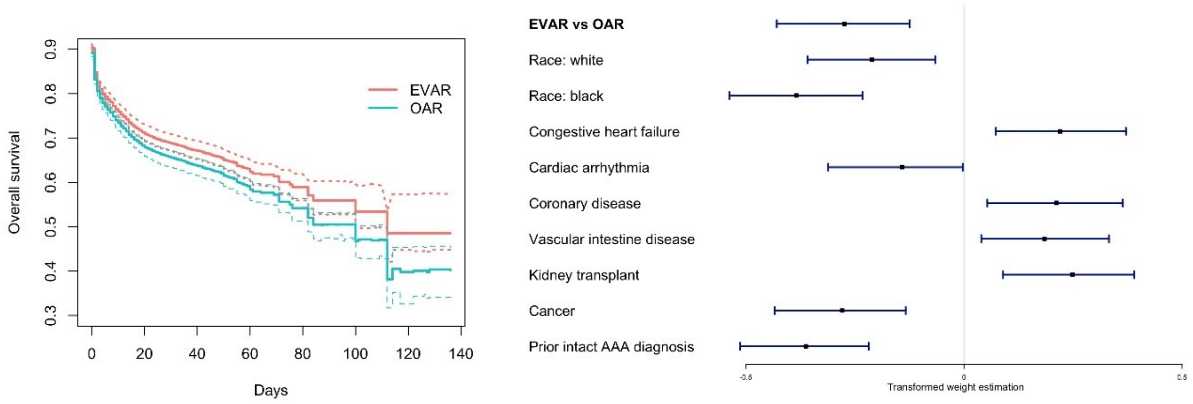


Figure 4.5: Analysis of short-term mortality (left: estimated survival curves with pointwise 90% confidence intervals; right: forest plot of the estimated weights)

For the analysis of long-term survival, the learning rate for Adam optimizer is tuned as 0.036. The analysis results are summarized in Figure 4.6, which are parallel to those in Figure 4.5. The findings are similar to those for short-term survival. Briefly, the left panel suggests some advantage of EVAR, but the pointwise confidence intervals overlap. We compute the expected survival as 1464.2 days under EVAR and 1348.0 days under OAR. The forest plot in the right panel shows that the advantage of EVAR is smaller than that for short-term survival. Confounders that demonstrate considerable and "persistent" effects include race, sex, and six medical conditions.

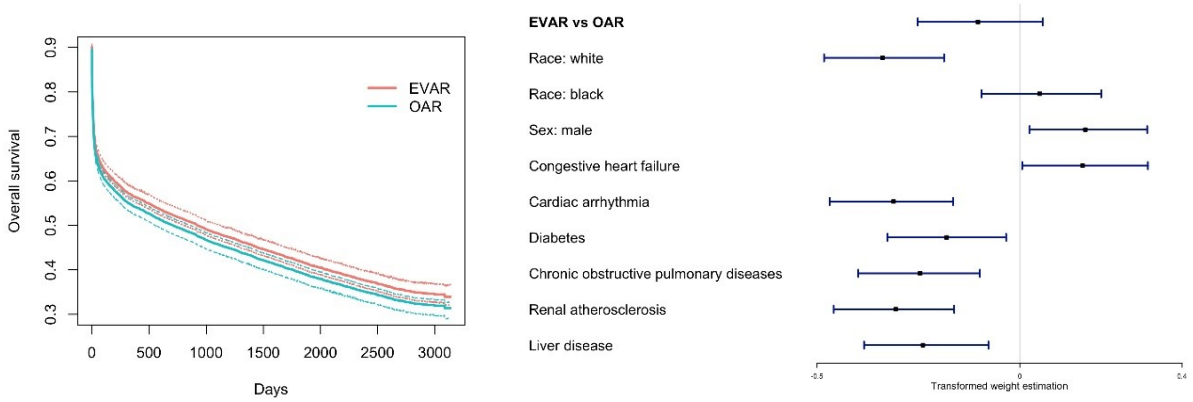


Figure 4.6: Analysis of long-term mortality (left: estimated survival curves with pointwise 90% confidence intervals; right: forest plot of the estimated weights)

Remarks For comprehensiveness, we also conduct regression-based analysis. The results are

presented in Appendix C.2.2. As the Cox model assumption is violated in both survival analyses, the results cannot be sensibly utilized.

4.2.4 Discussion

As fully discussed in the published literature, the Medicare data has multiple unique advantages. With its broad coverage of the U.S. elderly population, our findings can be applied to this population with high confidence. Although there is no evidence that the relative treatment effects of EVAR and OAR differ by age, sex, and race [132–134], application of the findings to the younger U.S. population and populations in other countries/regions should be conducted with cautions. On the other hand, it has also been recognized that the Medicare data has limitations [4,146]. For example, it does not contain certain information (e.g., over-the-counter drug use). As such, there may exist imbalance on unmeasured confounders. We note that this limitation is shared by other emulation studies and analysis of observational data. We have analyzed the Medicare inpatient data from 01/01/2011 to 06/30/2019. Both the enrollment and follow-up times are long enough, especially compared to peer studies [132–134]. Although there is no indication that the treatment effects have temporal variations, it may still be of interest to examine more extensive data (which is not pursued with data access limitations). Another data limitation is that we do not have access to data on other clinical settings (e.g., emergency room or outpatient). With the special nature of rAAA, inpatient claims should be able to catch the dominating majority of the cases.

The emulation strategy has been developed and adopted in quite a few studies. Its pros and cons have been well documented [36, 38, 137]. It is especially noted that, first, emulation trials, although resembling real clinical trials in multiple perspectives, still have notable limitations and cannot replace real clinical trials. Second, there is still a lack of objective comparison and definitive conclusion on its relative performance with respect to other causal inference approaches. Although important, this is beyond the scope of this study. The adopted deep learning methods have been based on certain well-developed components and software programs. Nevertheless, their "combination" and application to the emulation setting and rAAA treatment problem are new and novel. Our analysis has demonstrated how to "replace" regression using deep learning under settings more sophisticated than in the literature. As the "propensity score + survival analysis" strategy and individual components of the deep learning analysis have been more or less developed in the literature,

we choose not to methodologically further discuss or conduct more numerical investigations.

Our main finding is that EVAR has advantageous short- and long-term survival. Although the improvement in expected survival is modest, considering the severity of rAAA, it may still have important clinical implications. In the literature, the short-term survival advantage of EVAR has been suggested in multiple observational analyses [33, 131, 133]. However, there has been a lack of definitive conclusion on the long-term benefit. For example, Behrendt et al. [132] suggested early survival benefit of EVAR over OAR, which reversed at $\sim$2.5 years of follow-up, for iAAA and rAAA patients in Germany. Schermerborn et al. [33] observed similar survival of the two procedures after 3 years from initial surgery for iAAA patients in the Medicare population. And a 15-year follow-up resulted from the EVAR-1 trial indicated that EVAR had inferior late survival compared with OAR [130]. Multiple factors can contribute to the differences observed in the aforementioned and other studies. First, the studied populations have different characteristics. Second, the analysis strategies also differ, with our strategy closer to a controlled clinical trial. It is also noted that the study periods are different. Although there is still no indication of temporal variation in treatment effects, related confounders may change over time.

Besides treatment, our analysis also suggests that race, gender, and certain medical conditions are associated with survival after EVAR and OAR among rAAA patients. While most observational studies that compare EVAR and OAR match study subjects or adjust for potential confounders, there is a lack of attention on how these variables may impact survival after rAAA. We have found that compared to other races, the white race is associated with lower short- and long-term mortality, and the black race is associated with lower short-term mortality. This race difference has been insufficiently studied in the literature. It can be caused by genetic effects (considering that genetic factors contribute to many cardiovascular diseases), lifestyle, cultural factors, access to care, and other factors that may confound survival. While Egorova et al. [133] observed significantly worse outcomes after EVAR and OAR for female patients, we have found no gender difference in short-term mortality and male associated with higher long-term mortality. One contributing factor is the difference in analysis technique: Egorova et al. [133] compared the observed survival with expected survival in a life table, while we have conducted a more comprehensive adjusted analysis. Lastly, we have identified certain medical conditions as associated with survival. What may seem counterintuitive is that some medical conditions are found as negatively associated with mortality.

For example, it is found that prior iAAA diagnosis decreases short-term mortality risk after rAAA, and the presence of cardiac arrhythmia increases both short- and long-term survival. One plausible explanation is that patients with related medical conditions are more likely to have regular hospital visits and more access to healthcare services, which may lead to more timely detection of emergent rAAA. For example, it is noted in Edwards et al. [131] that patients who had a prior diagnosis of iAAA were less commonly admitted through the emergency department, and were more commonly transferred between hospitals before treatment, which was associated with better survival. Dardik et al. [149] also found that the presence of hypertension, diabetes, and COPD were correlated with a statistically significant lower mortality rate, whereas the presences of smoking, heart disease, and renal disease were correlated with a statistically insignificant lower mortality rate after the diagnosis of rAAA.

4.2.5 Conclusion

This study has suggested certain short- and long-term survival advantage of EVAR over OAR for rAAA patients. It has also further advanced the emulation and deep learning techniques for analyzing data mined from large medical record databases. Both the medical findings and analytic developments can complement the existing literature and be of interest to stakeholders at multiple levels.

Chapter 5

Concluding Remarks

With the Medicare administrative data, this dissertation develops various statistical methods for biomedical research in older adults. Chapter 2 and Chapter 3 focus on examining disease interconnections in healthcare outcome measures using network analysis. To be more specific, in Chapter 2, we build a clinical treatment HDN on inpatient LOS. Considering that multiple outcomes are closely related to each other, In Chapter 3, we build a clinical treatment HDN that can incorporate multiple variables. We analyze LOS and readmission data in Chapter 3 and note that the proposed method can be easily expanded to incorporate more outcomes of different data types. In both chapters, to accommodate uniquely challenging data distributions (high-dimensionality and zero-inflation), novel modeling and estimation approaches have been developed. The methodological developments may have other applications and can foster complex network analysis. Based on the constructed networks, we analyze key network properties such as connectivity, module/hub, and temporal variation. The findings are found to be largely supported by existing biomedical evidence. A closer examination of analysis results also reveals novel findings that are less/not investigated in the individual-disease studies. These findings have clinical importance and can provide valuable information in guiding future analysis. The proposed clinical treatment HDNs complement existing molecular and phenotypic HDNs by being more clinically sensible and having more practical implications. They can be used to guide more efficient allocation of hospital beds and other resources, provide additional insights into disease interconnections from a treatment perspective, and define an alternative way of disease characterization and classification.

In Chapter 4, we adopt the emulation approach to conduct causal inferences on treatment/intervention

effects. With emulation analysis, we first compare rivaroxaban and dabigatran's effectiveness and safety outcomes for Medicare patients with atrial fibrillation. We analyze the emulated trial using propensity score and IPT weighting Cox proportional hazards regression. We find that dabigatran is superior in terms of time to any primary event (including ischemic stroke, other thromboembolic events, major bleeding, and death), major bleeding, and mortality. Differences between the two drugs in terms of stroke and other thromboembolic events are not significant. We also observe temporal variations, which have been largely neglected in other emulation studies. Considering that the regression-based statistical techniques generally have too strict data assumptions (e.g., Cox proportional model for survival data), we develop a novel deep learning strategy, which mimics the "propensity score + IPT weighting Cox regression" approach. We note that the proposed method can be easily generalized to other regression models and data types. We apply this approach to study survival outcomes of endovascular repair versus open aortic repair for Medicare patients with abdominal aortic aneurysms. We find that endovascular repair has survival advantages in both short- and long-term mortality. Overall, the developed "emulation + deep learning" approach provides an alternative and more flexible way of evaluating treatment/intervention effects using observational data. This study is the first application of deep learning to the emulation analysis, which can further analysis in both domains. It also showcases a new way of analyzing the scaled and comprehensive Medicare database. The emulated trials can guide the development of a real clinical trial (if it ever becomes feasible). The findings provide solid evidence for guiding better clinical decisions.

5.1 Limitation

This dissertation inevitably has limitations. Due to limited data accessibility, the analyses focus on Medicare inpatient claims. Inpatient treatment is only part of the healthcare system for Medicare beneficiaries. Many treatments are accomplished in other clinical settings, such as outpatient office visits and ED visits. Moreover, chronic diseases generally do not involve many hospital treatments and are often handled with drug refillings. Nevertheless, we note that inpatient treatment is the most serious type of treatment that consumes the most medical resources. Accordingly, analysis focus on inpatient treatment has its unique value. We have considered the whole Medicare popu-

lation. With the broad coverage of Medicare to the U.S. elderly population, our findings can be applied to this population with high confidence. It will be of interest to conduct stratified analysis to better accommodate heterogeneity. Although there is no evidence that the findings differ between race and other demographic factors, the generalization of our findings to other populations still needs further examination.

For Chapter 2 and Chapter 3, as in many other HDN analysis, our analysis can only infer associations (with undirected networks). Causal analysis would demand significant additional data. Besides, unadjusted outcomes are used to construct the proposed clinical treatment HDNs. It is noted that the proposed approach in Chapter 3 is regression-based and can be expanded to include covariants such as demographic and other risk factors. However, building HDNs on risk-adjusted outcomes is beyond the scope of this dissertation and will be postponed to future studies. In the analysis, we examine the temporal variations in a cross-sectional manner. Given that we have observed significant dynamics in network structures across time, it will be of interest to develop a time-varying network framework. This framework will require extensive new methodology development. Moreover, the presented graphical algorithms deal with a large sample size, a large number of unknown parameters, and a complex model structure. As a result, computation is expensive. Luckily, estimation for each node and under different tunings can be run parallel to reduce computer time. For larger data, further parallelization may be needed to make the analysis affordable. Our examination suggests that the findings are biomedically sensible to a large extent. However, we are unable to examine all findings considering a large number of diseases and interconnections. It will also be of interest to examine additional data and provide more interpretations of the findings.

For Chapter 4, it is noted that even with full access to different treatment settings, the Medicare database covers limited information [4, 146]. For example, it does not contain information on over-the-counter drug uses, patients' socioeconomic status, and adherence to medication regimen, all of which are essential confounders for treatment effects. It is unknown that whether treatment arms are imbalanced on these unmeasured confounders. In addition to data limitations, it is well documented in the literature that the emulation approach, while being lucidly interpretable, has notable limitations [36, 38, 137]. For example, it can only emulate target trials without blind assignment. There is still a lack of direct and objective comparison of this approach concerning other causal inference approaches. It will be of interest to conduct such a comparison, but it is

beyond the scope of this study. In the analysis, we develop an innovative deep learning approach to mimic the commonly used "propensity score + IPT weighting Cox proportional hazards model." Although we conduct a bootstrap-like procedure to gain insights into the variability of estimates, the rigorous statistical inference has not been pursued. Noting that statistical properties for neural networks are less investigated in the literature, it will be of future interest to research more along this direction.

5.2 Future study

With the possibility of more comprehensive data, we will expand the analysis scope to incorporate other treatment settings and other outcomes of different data distributions. This includes accommodating more outcome measures in constructing clinical treatment HDNs and investigating more medical conditional and treatments using the "emulation + deep learning" analysis approach. It is also of future interest to validate the findings of this dissertation by examining potential unmeasured risk factors. Regarding the analysis of clinical treatment HDNs, we plan to develop new modeling approaches, which could accommodate risk-adjusted outcomes, intervention/shock effects, and time-dependent variables. Regarding the emulation analysis, we will conduct a direct comparison for its relative performance concerning other causal inference techniques. We are also planning to expand the deep learning algorithm to different data types beyond survival analysis and derive corresponding theoretical properties.

Appendix A

Supplementary Materials for Chapter 2

A.1 Clinical Classifications Software disease categories

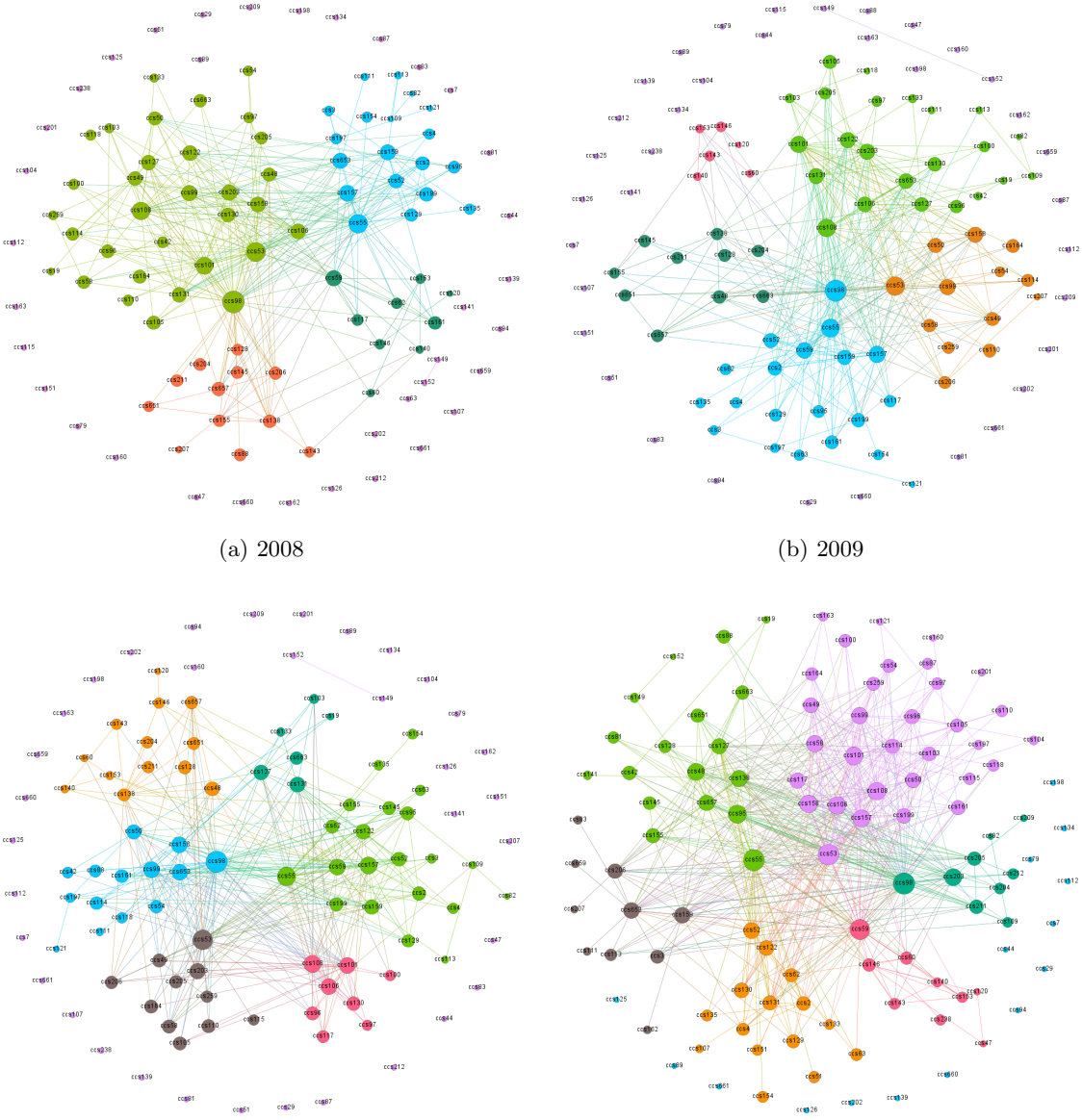
Table A.1: CCS disease categories

ccs100	Acute myocardial infarction
ccs101	Coronary atherosclerosis and other heart disease
ccs103	Pulmonary heart disease
ccs104	Other and ill-defined heart disease
ccs105	Conduction disorders
ccs106	Cardiac dysrhythmias
ccs107	Cardiac arrest and ventricular fibrillation
ccs108	Congestive heart failure; nonhypertensive
ccs109	Acute cerebrovascular disease
ccs110	Occlusion or stenosis of precerebral arteries
ccs111	Other and ill-defined cerebrovascular disease
ccs112	Transient cerebral ischemia
ccs113	Late effects of cerebrovascular disease
ccs114	Peripheral and visceral atherosclerosis
ccs115	Aortic; peripheral; and visceral artery aneurysms
ccs117	Other circulatory disease
ccs118	Phlebitis; thrombophlebitis and thromboembolism
ccs120	Hemorrhoids
ccs121	Other diseases of veins and lymphatics
ccs122	Pneumonia (except that caused by tuberculosis or sexually transmitted disease)
ccs125	Acute bronchitis
ccs126	Other upper respiratory infections

ccs127	Chronic obstructive pulmonary disease and bronchiectasis
ccs128	Asthma
ccs129	Aspiration pneumonitis; food/vomitus
ccs130	Pleurisy; pneumothorax; pulmonary collapse
ccs131	Respiratory failure; insufficiency; arrest (adult)
ccs133	Other lower respiratory disease
ccs134	Other upper respiratory disease
ccs135	Intestinal infection
ccs138	Esophageal disorders
ccs139	Gastroduodenal ulcer (except hemorrhage)
ccs140	Gastritis and duodenitis
ccs141	Other disorders of stomach and duodenum
ccs143	Abdominal hernia
ccs145	Intestinal obstruction without hernia
ccs146	Diverticulosis and diverticulitis
ccs149	Biliary tract disease
ccs151	Other liver diseases
ccs152	Pancreatic disorders (not diabetes)
ccs153	Gastrointestinal hemorrhage
ccs154	Noninfectious gastroenteritis
ccs155	Other gastrointestinal disorders
ccs157	Acute and unspecified renal failure
ccs158	Chronic kidney disease
ccs159	Urinary tract infections
ccs160	Calculus of urinary tract
ccs161	Other diseases of kidney and ureters
ccs162	Other diseases of bladder and urethra
ccs163	Genitourinary symptoms and ill-defined conditions
ccs164	Hyperplasia of prostate
ccs19	Cancer of bronchus; lung
ccs197	Skin and subcutaneous tissue infections
ccs198	Other inflammatory condition of skin
ccs199	Chronic ulcer of skin
ccs2	Septicemia (except in labor)
ccs201	Infective arthritis and osteomyelitis (except that caused by tuberculosis or sexually transmitted disease)
ccs202	Rheumatoid arthritis and related disease
ccs203	Osteoarthritis
ccs204	Other non-traumatic joint disorders
ccs205	Spondylosis; intervertebral disc disorders; other back problems
ccs206	Osteoporosis
ccs207	Pathological fracture
ccs209	Other acquired deformities
ccs211	Other connective tissue disease
ccs212	Other bone disease and musculoskeletal deformities

ccs238	Complications of surgical procedures or medical care
ccs259	Residual codes; unclassified
ccs29	Cancer of prostate
ccs3	Bacterial infection; unspecified site
ccs4	Mycoses
ccs42	Secondary malignancies
ccs44	Neoplasms of unspecified nature or uncertain behavior
ccs47	Other and unspecified benign neoplasm
ccs48	Thyroid disorders
ccs49	Diabetes mellitus without complication
ccs50	Diabetes mellitus with complications
ccs51	Other endocrine disorders
ccs52	Nutritional deficiencies
ccs53	Disorders of lipid metabolism
ccs54	Gout and other crystal arthropathies
ccs55	Fluid and electrolyte disorders
ccs58	Other nutritional; endocrine; and metabolic disorders
ccs59	Deficiency and other anemia
ccs60	Acute posthemorrhagic anemia
ccs62	Coagulation and hemorrhagic disorders
ccs63	Diseases of white blood cells
ccs651	Anxiety disorders
ccs653	Delirium dementia and amnestic and other cognitive disorders
ccs657	Mood disorders
ccs659	Schizophrenia and other psychotic disorders
ccs660	Alcohol-related disorders
ccs661	Substance-related disorders
ccs663	Screening and history of mental health and substance abuse codes
ccs7	Viral infection
ccs79	Parkinson‘s disease
ccs81	Other hereditary and degenerative nervous system conditions
ccs82	Paralysis
ccs83	Epilepsy; convulsions
ccs87	Retinal detachments; defects; vascular occlusion; and retinopathy
ccs88	Glaucoma
ccs89	Blindness and vision defects
ccs94	Other ear and sense organ disorders
ccs95	Other nervous system disorders
ccs96	Heart valve disorders
ccs97	Peri-; endo-; and myocarditis; cardiomyopathy (except that caused by tuberculosis or sexually transmitted disease)
ccs98	Essential hypertension
ccs99	Hypertension with complications and secondary hypertension

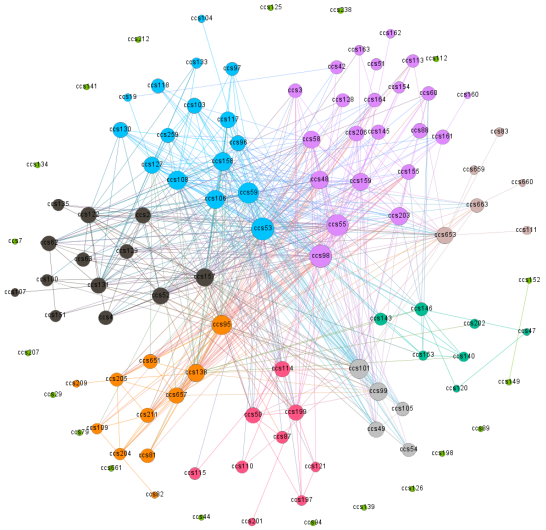
A.2 Year-specific LOS HDNs



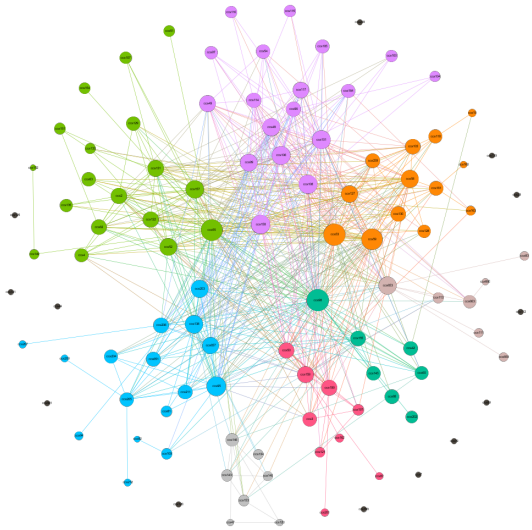
(a) 2008 (b) 2009

(c) 2010 (d) 2011

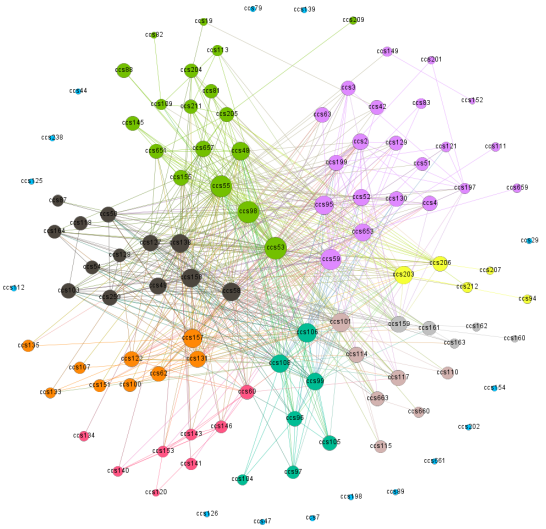
Figure A.1: Year-specific LOS HDNs



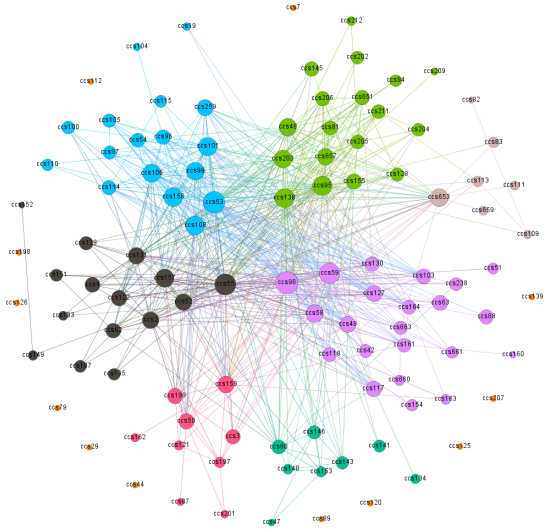
(e) 2012



(f) 2013

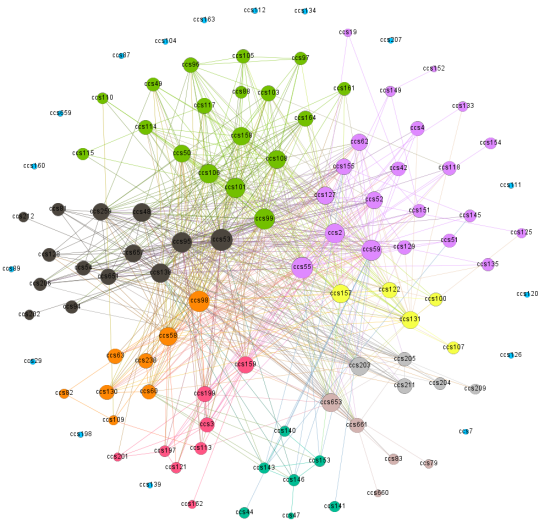


(g) 2014

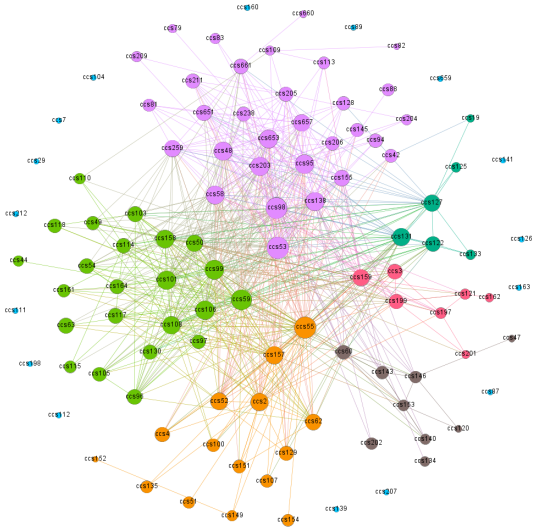


(h) 2015

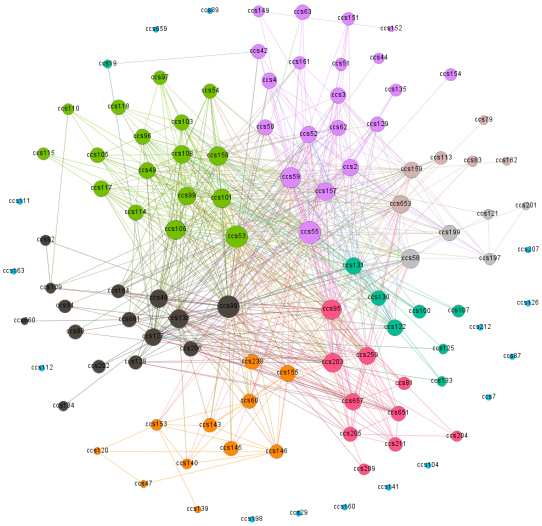
Figure A.1: Year-specific LOS HDNs



(i) 2016



(j) 2017



(k) 2018

Figure A.1: Year-specific LOS HDNs

A.3 Alternative analysis

As briefly discussed in Chapter 2, the interconnections in LOS among diseases differ from those obtained based on genetic and phenotypic information. To more firmly establish such differences and gain additional insights into the proposed analysis, we further analyze and compare with the molecular HDN (as developed in Goh et al. [24]) and phenotypic HDN (as developed in Hidalgo et al. [27]).

A.3.1 The molecular HDN

This HDN is gene-centric, under which two diseases are interconnected if they share common genetic risk factors. We note that the disease list investigated in Goh et al. [24] differs from that in this study. To make the results more directly comparable, we need to re-do the analysis using the approach in Goh et al. [24] but with a different list of diseases. More specifically, we first extract data from *https://phewascatalog.org/phewas* [150], which contains information on shared genes between the electronic health record driven Phenome-wide association studies codes (PheCode). Like the CCS used in our study, PheCode is also a disease classification software that groups the International Classification of Diseases (ICD) codes. Based on common ICD codes, we convert the 1,723 PheCodes into 204 CCS codes. We further focus on the network for the same 108 CCS codes as in our analysis.

A.3.2 The phenotypic HDN

Hidalgo et al. [27] introduced two comorbidity measures to quantify the correlation between two diseases. Here we adopt the ϕ-correlation, which is Pearson's correlation for binary variables. Specifically, the ϕ-correlation between disease i and j is defined as:

$$\phi_{ij} = \frac{C_{ij} - P_i P_j}{\sqrt{P_i P_j (N - P_i)(N - P_j)}},$$

where N is the total number of study subjects, C_ij is the number of subjects diagnosed with both diseases, and P_i and P_j are numbers of subjects diagnosed with disease i and disease j, respectively. Then using the same inpatient Medicare data from January 2008 to December 2018, we construct

the phenotypic HDN for the same 108 CCS diseases.

A.3.3 Results

Similar to the proposed LOS HDN, the molecular HDN and phenotypic HDN constructed here are also unweighted and undirected. The obtained network structures as shown in Figure A.2, which also includes the one constructed in Chapter 2 for comparison. The molecular HDN has 1,344 edges and 5 modules, and the phenotypic HDN has 1,059 edges and 9 modules. In comparison, the one constructed in Chapter 2 has 1,049 edges and 11 modules. A closer examination confirms that the three networks are significantly different.

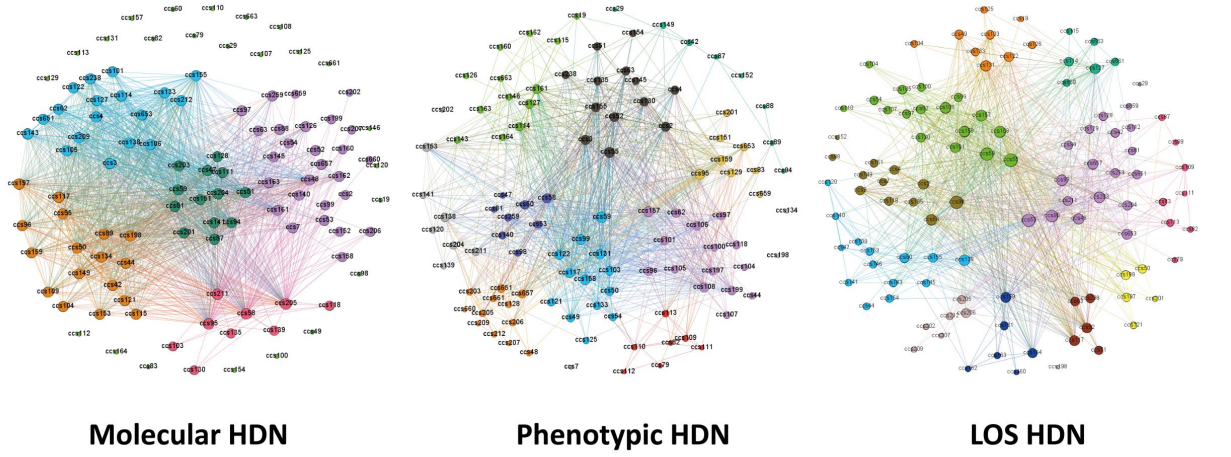


Figure A.2: HDNs constructed using three different approaches

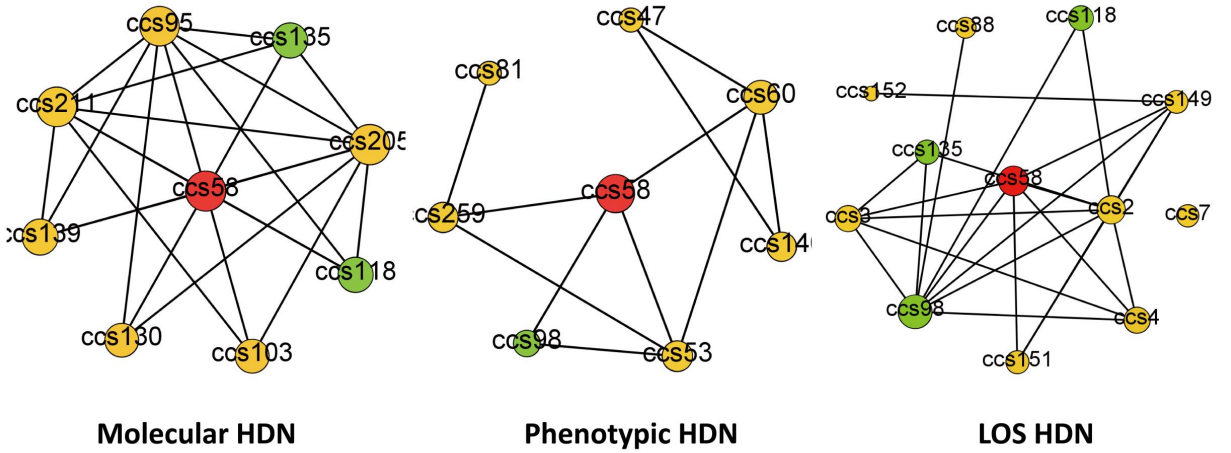
Connectivity Table A.2 shows the top 10 diseases with the highest connectivity values. It is observed that CCS58 and CCS95 are the only two diseases that are in the top 10 list for all three networks. For the molecular HDN, the top 10 list is enriched with endocrine, metabolic, and nervous diseases. The phenotypic HDN and LOS HDN are more similar and share seven common diseases (CCS55, CCS59, CCS95, CCS108, CCS157, CCS58, and CCS99). Considering that molecular information is "farther away" from clinics, the observed results are sensible and further establish the necessity of additional analysis beyond the molecular HDN.

Module We further examine differences in module structures. Specifically, if two diseases are in the same module under the LOS HDN but not under an alternative network, we say that their disease module relationship is different (across networks). Comparing the LOS HDN with the molecular

HDN	CCS	Disease	Connectivity
Molecular	ccs205	Spondylosis; intervertebral disc disorders; other back problems	82
	ccs58	Other nutritional; endocrine; and metabolic disorders	82
	ccs211	Other connective tissue disease	77
	ccs95	Other nervous system disorders	75
	ccs47	Other and unspecified benign neoplasm	73
	ccs51	Other endocrine disorders	67
	ccs204	Other non-traumatic joint disorders	66
	ccs87	Retinal detachments; defects; vascular occlusion; and retinopathy	66
	ccs59	Deficiency and other anemia	65
	ccs198	Other inflammatory condition of skin	64
Phenotypic	ccs59	Deficiency and other anemia	66
	ccs55	Fluid and electrolyte disorders	58
	ccs95	Other nervous system disorders	57
	ccs157	Acute and unspecified renal failure	54
	ccs99	Hypertension with complications and secondary hypertension	54
	ccs159	Urinary tract infections	51
	ccs131	Respiratory failure; insufficiency; arrest (adult)	50
	ccs58	Other nutritional; endocrine; and metabolic disorders	49
	ccs3	Bacterial infection; unspecified site	47
	ccs108	Congestive heart failure; nonhypertensive	46
LOS	ccs55	Fluid and electrolyte disorders	75
	ccs59	Deficiency and other anemia	67
	ccs98	Essential hypertension	63
	ccs53	Disorders of lipid metabolism	63
	ccs95	Other nervous system disorders	56
	ccs108	Congestive heart failure; nonhypertensive	48
	ccs157	Acute and unspecified renal failure	46
	ccs58	Other nutritional; endocrine; and metabolic disorders	45
	ccs138	Esophageal disorders	45
	ccs99	Hypertension with complications and secondary hypertension	42

Table A.2: Top 10 diseases with the highest connectivity

HDN shows that 23.5% of the 5,778 pairwise module relationships are different. Comparing the LOS HDN with the phenotypic HDN shows that 14.36% of the pairwise module relationships are different. As a representative example, Figure A.3 shows the modules that other nutritional; endocrine; and metabolic disorders (CCS58) belongs to in the three networks. It is noted that CCS58 has relatively higher significance as it is one of the two diseases that appear in the top 10 connectivity list for all three networks. Figure A.3 shows that the three modules are highly different. The modules under the molecular HDN and phenotypic HDN do not share any common diseases other than CCS58. The modules under the molecular HDN and LOS HDN share two common diseases (CCS118 and CCS135) other than CCS58. The modules under the phenotypic HDN and LOS HDN share one common disease (CCS98) other than CCS58. In Figure A.3, shared diseases are colored in green, and different diseases are colored in orange.



ccs58	Other nutritional; endocrine; and metabolic disorders	ccs149	Biliary tract disease	ccs4	Mycoses
ccs98	Essential hypertension	ccs151	Other liver diseases	ccs47	Other and unspecified benign neoplasm
ccs118	Phlebitis; thrombophlebitis and thromboembolism	ccs152	Pancreatic disorders (not diabetes)	ccs53	Disorders of lipid metabolism
ccs135	Intestinal infection	ccs2	Septicemia (except in labor)	ccs60	Acute posthemorrhagic anemia
ccs103	Pulmonary heart disease	ccs205	Spondylosis; intervertebral disc disorders; other back problems	ccs7	Viral infection
ccs130	Pleurisy; pneumothorax; pulmonary collapse	ccs211	Other connective tissue disease	ccs81	Other hereditary and degenerative nervous system conditions
ccs139	Gastroduodenal ulcer (except hemorrhage)	ccs259	Residual codes; unclassified	ccs88	Glaucoma
ccs140	Gastritis and duodenitis	ccs3	Bacterial infection; unspecified site	ccs95	Other nervous system disorders

Figure A.3: Modules that contain CCS58

A.3.4 Remarks

This analysis further supports our argument that the proposed disease interconnection analysis is warranted beyond the existing molecular and phenotypic HDNs. It can be of interest to further explore differences and consolidate the three networks. However, that demands significant additional research and will be postponed to the future.

Appendix B

Supplementary Materials for Chapter 3

B.1 Clinical Classifications Software disease categories

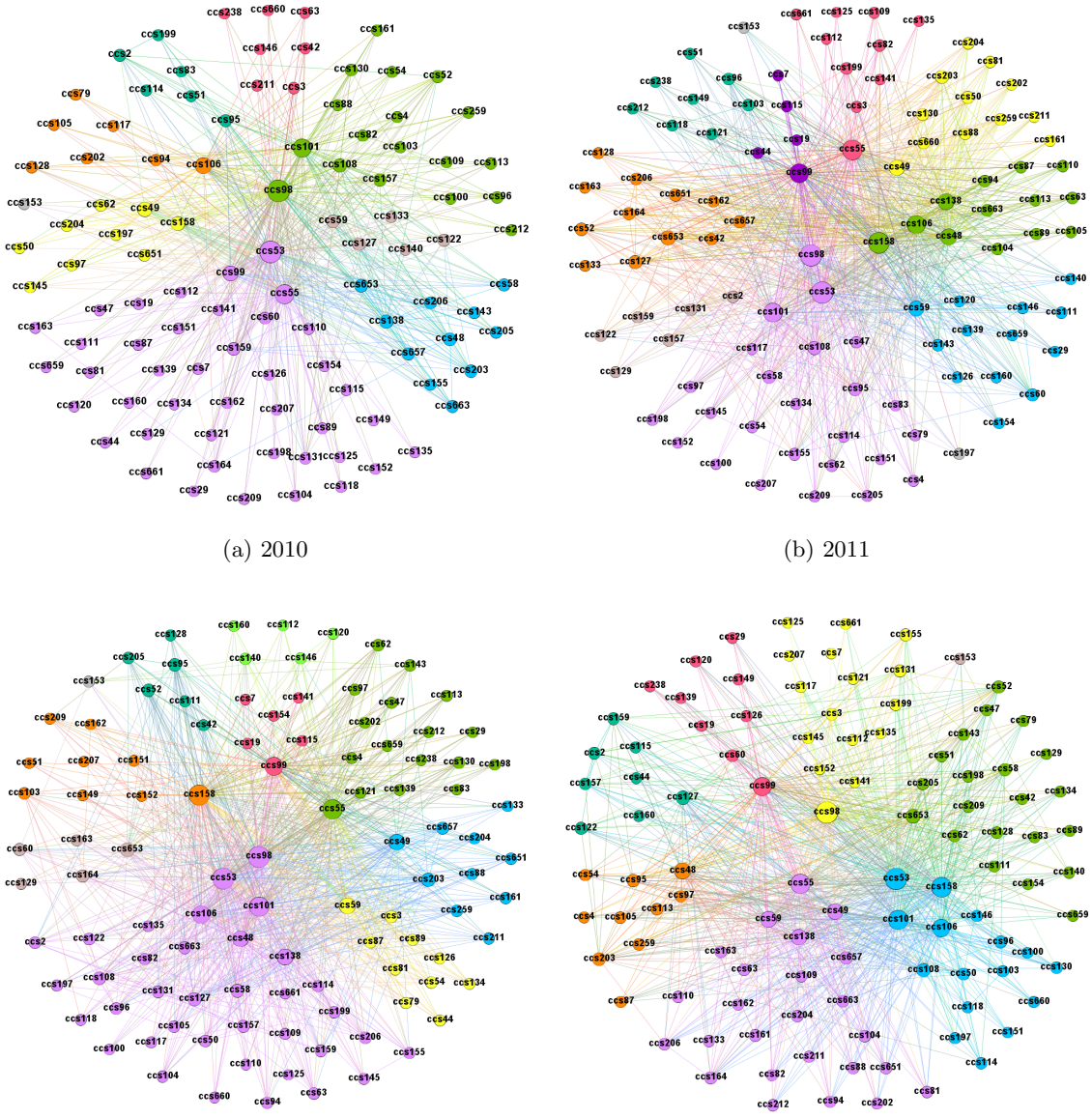
Table B.1: CCS disease categories

ccs100	Acute myocardial infarction
ccs101	Coronary atherosclerosis and other heart disease
ccs103	Pulmonary heart disease
ccs104	Other and ill-defined heart disease
ccs105	Conduction disorders
ccs106	Cardiac dysrhythmias
ccs107	Cardiac arrest and ventricular fibrillation
ccs108	Congestive heart failure; nonhypertensive
ccs110	Occlusion or stenosis of precerebral arteries
ccs111	Other and ill-defined cerebrovascular disease
ccs112	Transient cerebral ischemia
ccs113	Late effects of cerebrovascular disease
ccs114	Peripheral and visceral atherosclerosis
ccs115	Aortic; peripheral; and visceral artery aneurysms
ccs117	Other circulatory disease
ccs118	Phlebitis; thrombophlebitis and thromboembolism
ccs120	Hemorrhoids
ccs121	Other diseases of veins and lymphatics
ccs122	Pneumonia (except that caused by tuberculosis or sexually transmitted disease)
ccs125	Acute bronchitis
ccs126	Other upper respiratory infections
ccs127	Chronic obstructive pulmonary disease and bronchiectasis

ccs128	Asthma
ccs129	Aspiration pneumonitis; food/vomitus
ccs130	Pleurisy; pneumothorax; pulmonary collapse
ccs131	Respiratory failure; insufficiency; arrest (adult)
ccs133	Other lower respiratory disease
ccs134	Other upper respiratory disease
ccs135	Intestinal infection
ccs138	Esophageal disorders
ccs139	Gastroduodenal ulcer (except hemorrhage)
ccs140	Gastritis and duodenitis
ccs141	Other disorders of stomach and duodenum
ccs143	Abdominal hernia
ccs145	Intestinal obstruction without hernia
ccs146	Diverticulosis and diverticulitis
ccs149	Biliary tract disease
ccs151	Other liver diseases
ccs152	Pancreatic disorders (not diabetes)
ccs153	Gastrointestinal hemorrhage
ccs154	Noninfectious gastroenteritis
ccs155	Other gastrointestinal disorders
ccs157	Acute and unspecified renal failure
ccs158	Chronic kidney disease
ccs159	Urinary tract infections
ccs160	Calculus of urinary tract
ccs161	Other diseases of kidney and ureters
ccs162	Other diseases of bladder and urethra
ccs163	Genitourinary symptoms and ill-defined conditions
ccs164	Hyperplasia of prostate
ccs19	Cancer of bronchus; lung
ccs197	Skin and subcutaneous tissue infections
ccs198	Other inflammatory condition of skin
ccs199	Chronic ulcer of skin
ccs2	Septicemia (except in labor)
ccs202	Rheumatoid arthritis and related disease
ccs203	Osteoarthritis
ccs204	Other non-traumatic joint disorders
ccs205	Spondylosis; intervertebral disc disorders; other back problems
ccs206	Osteoporosis
ccs207	Pathological fracture
ccs209	Other acquired deformities
ccs211	Other connective tissue disease
ccs212	Other bone disease and musculoskeletal deformities
ccs238	Complications of surgical procedures or medical care
ccs259	Residual codes; unclassified

ccs29	Cancer of prostate
ccs3	Bacterial infection; unspecified site
ccs4	Mycoses
ccs42	Secondary malignancies
ccs44	Neoplasms of unspecified nature or uncertain behavior
ccs47	Other and unspecified benign neoplasm
ccs48	Thyroid disorders
ccs49	Diabetes mellitus without complication
ccs50	Diabetes mellitus with complications
ccs51	Other endocrine disorders
ccs52	Nutritional deficiencies
ccs53	Disorders of lipid metabolism
ccs54	Gout and other crystal arthropathies
ccs55	Fluid and electrolyte disorders
ccs58	Other nutritional; endocrine; and metabolic disorders
ccs59	Deficiency and other anemia
ccs60	Acute posthemorrhagic anemia
ccs62	Coagulation and hemorrhagic disorders
ccs63	Diseases of white blood cells
ccs651	Anxiety disorders
ccs653	Delirium dementia and amnestic and other cognitive disorders
ccs657	Mood disorders
ccs659	Schizophrenia and other psychotic disorders
ccs660	Alcohol-related disorders
ccs661	Substance-related disorders
ccs663	Screening and history of mental health and substance abuse codes
ccs7	Viral infection
ccs79	Parkinson‘s disease
ccs81	Other hereditary and degenerative nervous system conditions
ccs82	Paralysis
ccs83	Epilepsy; convulsions
ccs87	Retinal detachments; defects; vascular occlusion; and retinopathy
ccs88	Glaucoma
ccs89	Blindness and vision defects
ccs94	Other ear and sense organ disorders
ccs95	Other nervous system disorders
ccs96	Heart valve disorders
ccs97	Peri-; endo-; and myocarditis; cardiomyopathy (except that caused by tuberculosis or sexually transmitted disease)
ccs98	Essential hypertension
ccs99	Hypertension with complications and secondary hypertension

B.2 Year-specific LOS and readmission HDNs



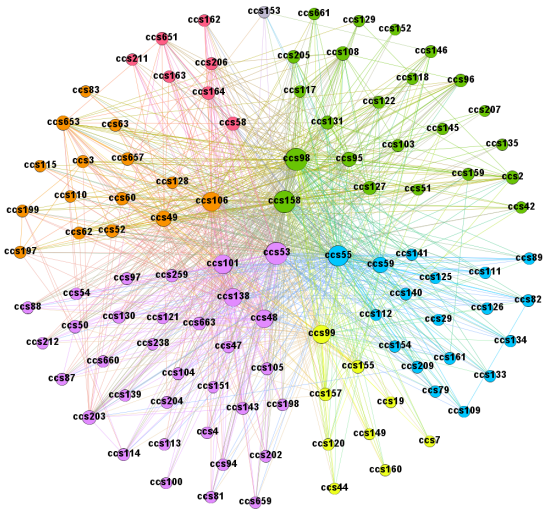
(a) 2010

(b) 2011

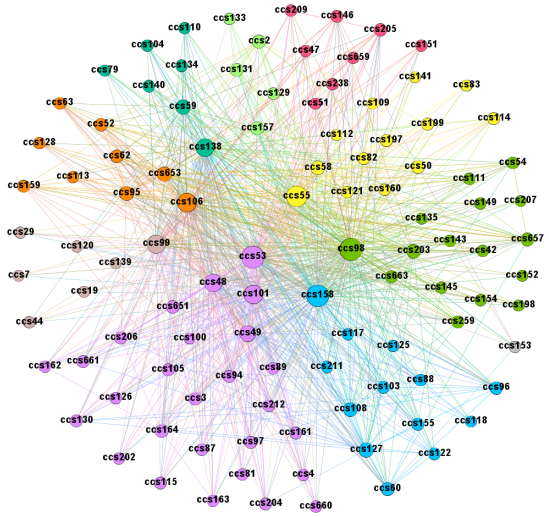
(c) 2012

(d) 2013

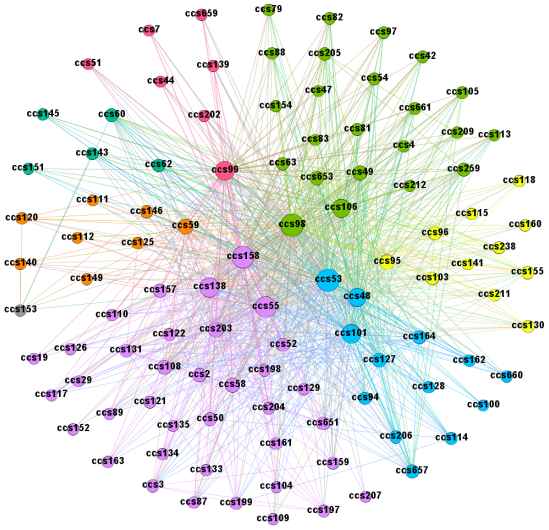
Figure B.1: Year-specific LOS and readmission HDNs



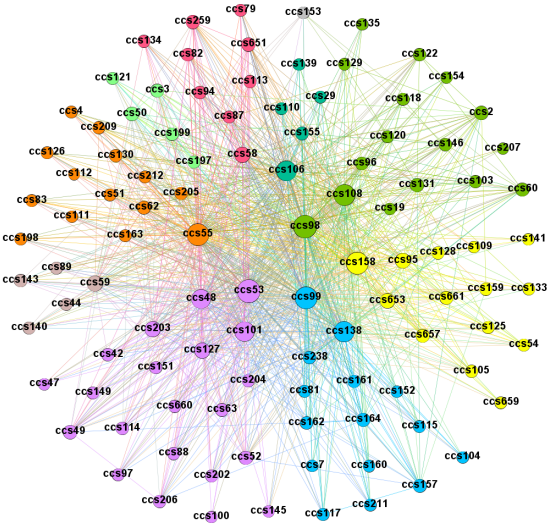
(e) 2014



(f) 2015

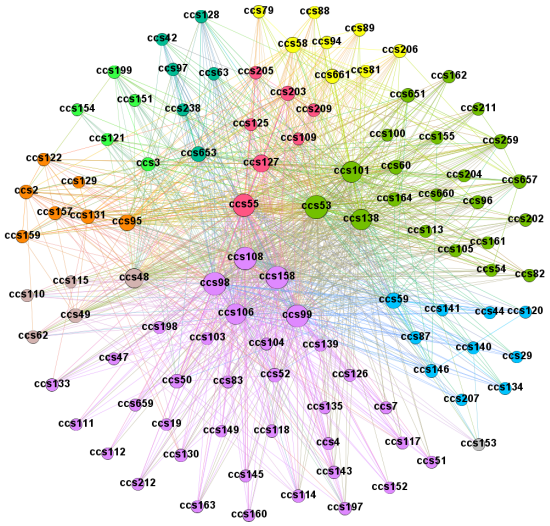


(g) 2016



(h) 2017

Figure B.1: Year-specific LOS and readmission HDNs



(i) 2018

Figure B.1: Year-specific LOS and readmission HDNs

Appendix C

Supplementary Materials for Chapter 4

C.1 Rivaroxaban versus dabigatran for atrial fibrillation

Table C.1: ICD-9-CM codes for identifying eligible individuals

Eligibility criteria	**Generic drug name / ICD-9-CM Code***
Inclusion	
Atrial fibrillation	427.31
First prescription for	
Dabigatran-150mg	dabigatran etexilate mesylate
Rivaroxaban-20mg	rivaroxaban
Exclusion (three months of history)	
Received prior treatment with	
Warfarin	warfarin, warfarin sodium
Eliquis	apixaban
Savaysa	edoxaban tosylate
Kidney transplant	V42.0
Renal dialysis	V45.11
A potential alternative indication for anticoagulation	
Mitral valve disease Heart valve repair or replacement	394.x-397.x, 398.9, 42.4x, V42.4, V43.3
Venous thromboembolism	451.1, 451.2, 451.81,451.9, 453.1, 453.2, 453.8, 453.9, 671.3, 671.4
Phlebitis or thrombophlebitis	451.x
Pulmonary embolism	415.1x
Joint replacement	V43.60-V43.66, V43.69

* Primary or any secondary diagnosis code

Table C.2: Generic drug names, ICD-9-CM codes, and Medicare MCC indicators for defining confounders

Variable	Generic drug name / ICD-9-CM Code* / Medicare MCC indicator
Medical History	
Diabetes mellitus	DIABETES_EVER
Hypertension	HYPERT_EVER
CKD	CHRONICKIDNEY_EVER
History of bleeding (3m preceding)	Intracranial bleeding 430, 431, 432; Hemoperitoneum 568.81; Hematuria 599.7; GI hemorrhage 530.7, 531.0, 531.2, 531.4, 531.6, 532.0, 532.2, 532.4, 532.6, 533.0, 533.2, 533.4, 533.6, 534.0, 534.2, 534.4, 534.6,569.3, 535.01, 535.11, 535.21, 535.31, 535.41, 535.51, 535.61, 535.71, 537.83, 537.84 , 562.02 ,562.03, 562.12, 562.13, 569.85, 578; Epistaxis 784.7; Hemoptysis 786.3; Vaginal hemorrhage 623.8, 626.2; Hemarthrosis 719.1, 719.2; NOS hemorrhage 459
Acquired HT	HYPOTH_EVER
No. of other CMS comorbidities	
Alzheimer's disease	ANEMIA_EVE
Related disorders or senile dementia	ALZH_DEMEN_EVER
Anemia	ANEMIA_EVER
Asthma	ASTHMA_EVER
Benign prostatic hyperplasia	HYPERP_EVER
Cataract	CATARACT_EVER
COPD	COPD_EVER
Ischemic heart disease	ISCHEMICHEART_EVER
Hip or pelvic fracture	HIP_FRACTURE_EVER
Glaucoma	GLAUCOMA_EVER
Hyperlipidemia	HYPERL_EVER
Osteoporosis	OSTEOPOROSIS_EVER
Rheumatoid arthritis or osteoarthritis	RA_OA_EVER
Breast cancer	CANCER_BREAST_EVER
Colorectal cancer	CANCER_COLORECTAL_EVER
Prostate cancer	CANCER_PROSTATE_EVER
Lung cancer	CANCER_LUNG_EVER
Endometrial cancer	CANCER_ENDOMETRIAL_EVER
Cardiovascular Disease	
AMI	AMI_EVER
CHF	CHF_EVER
Previous stroke or TIA	STROKE_TIA_EVER
Medication Use	
NSAIDs	Filling a prescription for diclofenac, ibuprofen, naparofen, ketoprofen, fenoprofen, flurbiprofen, piroxicasm, meloxicam, mefenamic acid, or indomethacin after the index date
Antiplatelet	Filling a prescription for aspirin, clopidogrel, prasugrel, dipyridamole, ticlopidine, or ticagrelor after the index date

* Primary or any secondary diagnosis code

MCC multiple chronical condition, CMS centers for Medicare & Medicaid services, CKD chronic kidney disease, HT hypothyroidism, COPD chronic obstructive pulmonary disease, AMI acute myocardial infarction, CHF congestive heart failure, TIA transient ischemic attack, NSAIDs non-steroidal anti-inflammatory drug

Table C.3: ICD-9-CM codes used for defining study outcomes

Outcome	ICD-9-CM Codes*
Primary outcome	
Ischemic stroke	433, 434, 436
Other thromboembolic events	Systemic embolism (444), TIA (435), Pulmonary embolism (415.1)
Major bleeding	Intracranial bleeding (430, 431, 432), Hemoperitoneum (568.81), Hematuria (599.7), GI hemorrhage (530.7, 531.0, 531.2, 531.4, 531.6, 532.0, 532.2, 532.4, 532.6, 533.0, 533.2, 533.4, 533.6, 534.0, 534.2, 534.4, 534.6,569.3, 535.01, 535.11, 535.21, 535.31, 535.41, 535.51, 535.61, 535.71, 537.83, 537.84 , 562.02 ,562.03, 562.12, 562.13, 569.85, 578)
All-cause mortality	N/A
Secondary outcome	
Any bleeding event	Major bleeding; Epistaxis (784.7), Hemoptysis (786.3), Vaginal hemorrhage (623.8, 626.2), Hemarthrosis (719.1, 719.2), NOS hemorrhage (459)
Acute myocardial infarction	410

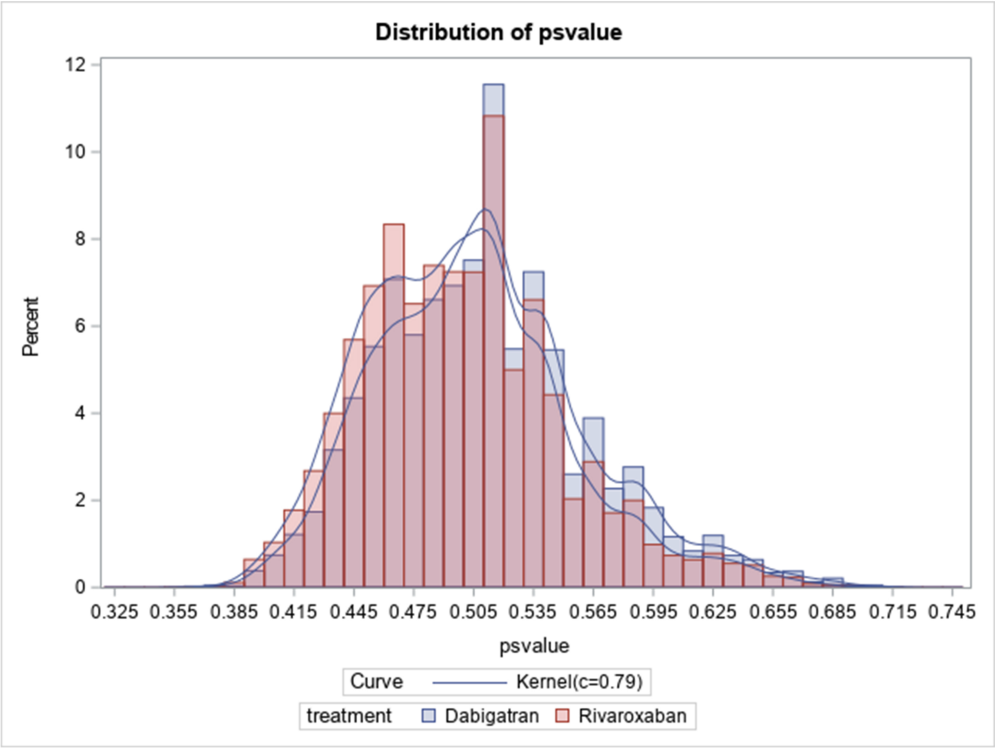
* Primary or any secondary diagnosis code

Table C.4: Baseline characteristics of the study cohorts, after IPT weighting

Variables	Trial 1		Trial 2		Trial 3	
	R	D	R	D	R	D
Demographics						
Age(years)						
65-74	8743 (48.43)	8749 (48.47)	13384 (46.35)	13391 (46.36)	12747 (49.47)	12768 (49.53)
75-84	7314 (40.52)	7327 (40.59)	11983 (41.50)	11985 (41.49)	10467 (40.63)	10469 (40.61)
$\geq$ 85	1995 (11.05)	1975 (10.94)	3509 (12.15)	3510 (12.15)	2551 (9.90)	2541 (9.86)
Sex (female)	8804 (48.77)	8794 (48.72)	13536 (46.87)	13540 (46.87)	12130 (47.08)	12136 (47.08)
Race						
White	16460 (91.65)	16465 (91.67)	26734 (92.97)	26743 (92.96)	23531 (91.96)	23537 (91.94)
Black	727 (4.05)	725 (4.04)	1073 (3.73)	1075 (3.74)	1107 (4.33)	1112 (4.34)
Other	774 (4.31)	770 (4.29)	949 (3.30)	950 (3.30)	950 (3.71)	951 (3.72)
Medical history						
Diabetes mellitus	7547 (41.81)	7552 (41.84)	12400 (42.94)	12404 (42.94)	10999 (42.69)	11026 (42.77)
Hypertension	16688 (92.44)	16686 (92.44)	26968 (93.39)	26979 (93.40)	23856 (92.59)	23880 (92.64)
CKD	4680 (25.93)	4673 (25.89)	7605 (26.34)	7608 (26.34)	7013 (27.22)	7003 (27.17)
History of bleeding	422 (2.34)	419 (2.32)	774 (2.68)	774 (2.68)	684 (2.66)	688 (2.67)
Acquired HT	5037 (27.90)	5045 (27.95)	8453 (29.28)	8458 (29.28)	7578 (29.41)	7573 (29.38)
No. of other CMS comorbidities						
0-3	4945 (27.39)	4935 (27.34)	7142 (24.73)	7144 (24.73)	6878 (26.69)	6871 (26.65)
4-6	8536 (47.29)	8543 (47.33)	14130 (48.93)	14137 (48.94)	12271 (47.63)	12303 (47.73)
$\leq$ 7	4571 (25.32)	4571 (25.33)	7605 (26.34)	7604 (26.33)	6616 (25.68)	6604 (25.62)
Cardiovascular disease						
AMI	1384 (7.66)	1372 (7.60)	2202 (7.63)	2206 (7.64)	1981 (7.69)	2007 (7.79)
CHF	8406 (46.57)	8411 (46.60)	13857 (47.99)	13856 (47.97)	11705 (45.43)	11693 (45.36)
Stroke or TIA	4040 (22.38)	4041 (22.39)	6432 (22.28)	6433 (22.27)	5775 (22.42)	5770 (22.38)
Medication use						
NSAIDs	1562 (8.65)	1565 (8.67)	1967 (6.81)	1970 (6.82)	1971 (7.65)	1971 (7.65)
Antiplatelet	2050 (11.35)	2051 (11.37)	2627 (9.10)	2636 (9.13)	2692 (10.45)	2719 (10.55)

R rivaroxaban, D dabigatran, CMS centers for Medicare & Medicaid Services,CKD chronic kidney disease, AMI acute myocardial infarction, CHF congestive heart failure, TIA transient ischemic attack, HT hypothyroidism, NSAIDs non-steroidal anti-inflammatory drug

Figure C.1: Histograms of the after-weighting propensity scores for trial 1



C.2 Endovascular versus open aortic repair for abdominal aortic aneurysms

C.2.1 ICD-9-CM codes for defining study cohort and confounders

Table C.5: ICD-9-CM codes for defining study cohort and confounders (one year of medical history)

Variable	ICD-9-CM Code*
Inclusion	
Ruptured abdominal aortic aneurysm	441.3
Endovascular aortic repair	39.71
Open aortic repair	38.44, 39.25, 39.52, 38.34, 38.64, 38.40, 38.60
Exclusion	
Thoracic aneurysms	441.1, 441.2
Thoracoabdominal aneurysms	441.6, 441.7
Aortic dissection	441.00-441.03
Repair of the thoracic aorta	38.35, 38.45, 39.73
Visceral or renal bypass	38.46, 39.24, 39.26
Medical history	
Congestive heart failure	398.91, 402.01, 402.11, 402.91, 404.01, 404.03, 404.11, 404.91, 404.13, 404.93, 425.4, 425.5, 425.7, 425.8, 425.9, 428.0, 428.1, 428.20, 428.22, 428.30, 428.32, 428.40, 428.42, 428.9
Cardiac arrhythmia	426.0, 426.10, 426.11, 426.12, 426.13, 426.7, 426.9, 427.0, 427.1, 427.2, 427.3, 427.9, V45.0, V53.3
Valvular disease	093.2, 394, 395, 396, 397, 424, V42.2, V43.3
Coronary disease	412, 413, 414, 429.2
Diabetes	250
Hypertension	401, 402, 403, 404, 405
Chronic Obstructive Pulmonary diseases	416, 417.9, 490, 491, 492, 493, 494, 495.0, 495.1, 495.2, 495.3, 495.4, 495.5, 495.6, 495.8, 495.9, 496, 500, 501, 502, 503, 504, 505, 506.0, 506.2, 506.4, 506.9, 508.1, 508.8, 508.9
Clinically significant lower extremity vascular diseases	440.22, 440.23, 440.24, 440.3, 444.22, V43.4
Renal atherosclerosis	440.1
Vascular intestine disease	557.1
Renal failure w dialysis	V45.1, V56.0, V56.1, V56.2, V56.3, V56.8, 585.6, 39.95 (w/o 586)
Renal failure without dialysis	403.01, 403.11, 403.91, 404.02, 404.03, 404.12, 404.13, 404.92, 404.93, 585 (w/o 585.6), 588.0
Other renal diseases	582, 583.0, 583.1, 583.2, 583.4
Kidney transplant	V420
Liver disease	070.22, 070.23, 070.32, 070.33, 070.44, 070.54, 070.9, 456.0, 456.1, 571, 572.1, 572.2, 572.3, 572.4, 572.8, 573.0, 573.1, 573.8, 573.9
Cerebrovascular diseases and paralysis	342, 344.1, 344.3, 344.4, 344.5, 344.9, 437.0, 438
Other neurological diseases	330, 331, 332, 333, 334.0, 334.1, 334.2, 334.4, 334.8, 335.0, 335.1, 335.2, 335.8, 335.9, 336.0, 336.2, 343, 344.0, 348.1, 348.3, 344.2, 344.6, 345, 437.3, 437.4, 437.5, 437.6, 437.7
Hyperlipidemia	272
Cancer	140, 141, 142, 143, 144,145, 146, 147, 148, 149, 150, 151, 152, 153, 154, 155, 156, 157, 158,159, 160, 161, 162, 163, 164, 165, 170, 171,172, 174, 175, 176, 179, 180, 181, 182, 183, 184, 185, 186, 187, 188, 189, 190, 191, 192, 193, 194, 195, 196, 197, 198, 199, 200, 201, 202, 203.0, 238.6
Rheumatoid arthritis	446, 701.0, 710.0, 710.1, 710.2, 710.3, 710.4, 710.8, 710.9, 711.2, 719.3, 714,720, 725, 728.5, 728.89
Prior intact AAA diagnosis	441.4, without mention 441.3

* Primary or any secondary diagnosis/procedure code

C.2.2 Regression analysis

In "standard" clinical trial analyses as well as emulated trial analyses [131, 151], regression techniques have generally been adopted. Here to complement the analysis presented in the main text, we also conduct regression-based analysis. The overall strategy is similar to that presented in Section 4.2.2. The first step is to estimate the propensity score using a logistic regression. The same set of variables as presented in Table 4.4 is included. Then we conduct the IPT weighted Cox regression analysis. Both steps can be realized using existing R functions. For inference, we directly use the p-values generated by Cox regression. There is hence no need for bootstrap.

The estimated propensity scores are shown in Figure C.2.

In the analysis of short-term survival, the estimated coefficient of EVAR is -0.611, with a standard error of 0.027 and a p-value less than 0.001. Other significant variables include age, hypertension, chronic obstructive pulmonary diseases, and prior intact AAA diagnosis. More detailed estimation and inference results are available from the authors. We test the proportional hazards assumption, and the global Chi-squared test returns a p-value less than 0.001, suggesting a violation of model assumptions. We further plot the scaled Schoenfeld residuals for treatment in the left panel of Figure C.3 and observe that the residuals are correlated with time. The plot of deviance in the right panel of Figure C.3 also suggests a violation of model assumptions.

In the analysis of long-term survival, the estimated coefficient of EVAR is -0.317, with a standard error of 0.002 and a p-value less than 0.001. Other significant variables include age, hypertension, chronic obstructive pulmonary diseases, and prior intact AAA diagnosis. The global Chi-squared test returns a p-value less than 0.001, which, along with Figure C.4, suggests that the Cox model assumptions are not satisfied.

Overall, we present the regression-based results for completeness but note that, with the violations of model assumptions, their results should not be utilized.

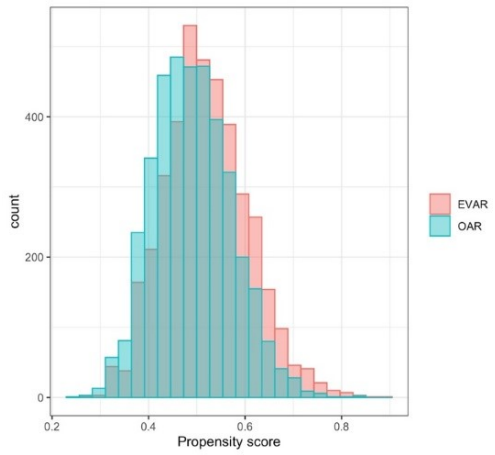


Figure C.2: Distribution of propensity score using logistic regression

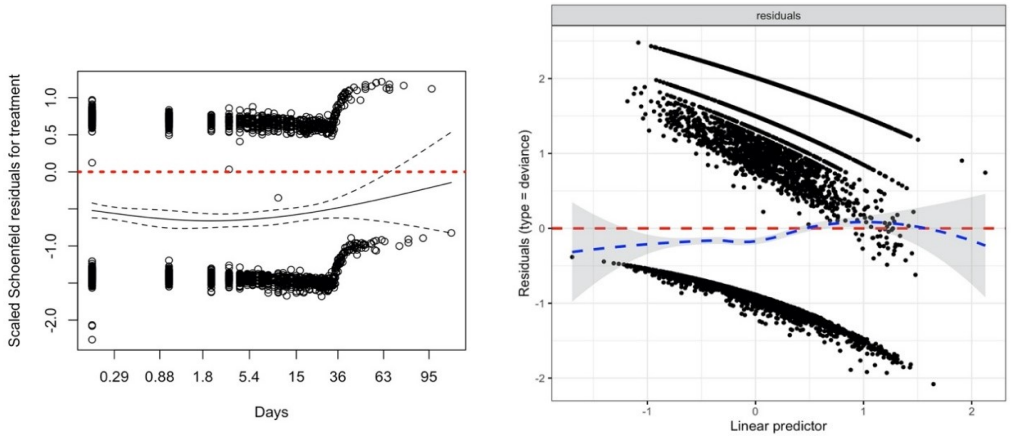


Figure C.3: Analysis of short-term survival using Cox regression (left: Scaled Schoenfeld residuals for treatment; right: Deviance residuals)

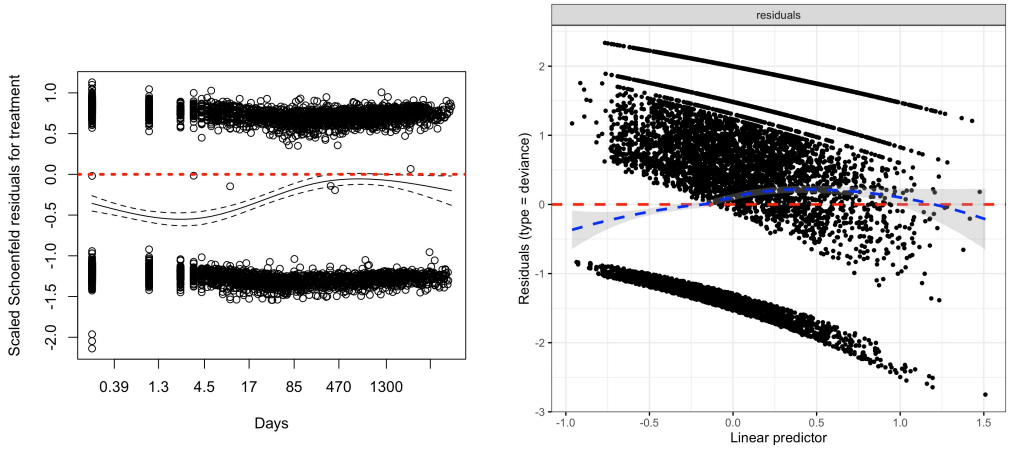


Figure C.4: Analysis of long-term survival using Cox regression (left: Scaled Schoenfeld residuals for treatment; right: Deviance residuals)

Bibliography

[1] J. Henry, Y. Pylypchuk, T. Searcy, and V. Patel. Adoption of electronic health record systems among us non-federal acute care hospitals: 2008–2015. *ONC data brief*, 35:1–9, 2016.

[2] A. B. Rosenkrantz, D. R. Hughes, and R. Duszak Jr. Medicare claims data resources: a primer for policy-focused radiology health services researchers. *Journal of the American College of Radiology*, 14(12):1538–1544, 2017.

[3] J. A. Baron, G. Lu-Yao, J. Barrett, D. McLerran, and E. S. Fisher. Internal validation of medicare claims data. *Epidemiology*, 5(5):541–544, 1994.

[4] N. Brennan, A. Oelschlaeger, C. Cox, and M. Tavenner. Leveraging the big-data revolution: Cms is expanding capabilities to spur health system transformation. *Health Affairs*, 33(7):1195–1202, 2014.

[5] M. E. Porter et al. What is value in health care. *N Engl J Med*, 363(26):2477–2481, 2010.

[6] A. MacLaurin and H. McConnell. Utilizing quality improvement methods to prevent falls and injury from falls: Enhancing resident safety in long-term care. *Journal of safety research*, 42(6):525–535, 2011.

[7] Piedmont Healthcare. Evidence-based care standardization reduces pneumonia mortality rates and los. *Health Catalyst*, 2018.

[8] UTMB Healthcare. Care transitions improvements reduces 30-day all-cause readmissions saving nearly $2 million. *Health Catalyst*, 2018.

[9] M. Bo, G. Fonte, F. Pivaro, M. Bonetto, C. Comi, V. Giorgis, L. Marchese, G. Isaia, G. Maggiani, and E. Furno. Prevalence of and factors associated with prolonged length of stay in

older hospitalized medical patients. *Geriatrics & gerontology international*, 16(3):314–321, 2016.

[10] J. E. Myers and M. C. Harper. Evidence-based effective practices with older adults. *Journal of Counseling & Development*, 82(2):207–218, 2004.

[11] C. X. Feng and L. Li. Modeling zero inflation and overdispersion in the length of hospital stay for patients with ischaemic heart disease. *Advanced Statistical Methods in Data Science*, pages 35–53, 2016.

[12] L. Xiang, A. H. Lee, K. K. Yau, and G. J. McLachlan. A score test for overdispersion in zero-inflated poisson mixed regression model. *Statistics in medicine*, 26(7):1608–1622, 2007.

[13] A. Gupta, L. A. Allen, D. L. Bhatt, M. Cox, A. D. DeVore, P. A. Heidenreich, A. F. Hernandez, E. D. Peterson, R. A. Matsouaka, C. W. Yancy, et al. Association of the hospital readmissions reduction program implementation with readmission and mortality outcomes in heart failure. *JAMA cardiology*, 3(1):44–53, 2018.

[14] S. T. Rinne, M. C. Graves, L. A. Bastian, P. K. Lindenauer, E. S. Wong, P. L. Hebert, and C.-F. Liu. Association between length of stay and readmission for copd. *The American journal of managed care*, 23(8):e253, 2017.

[15] L. A. Petersen, S.-L. T. Normand, J. Daley, and B. J. McNeil. Outcome of myocardial infarction in veterans health administration patients as compared with medicare patients. *New England journal of medicine*, 343(26):1934–1941, 2000.

[16] J. D. Huling, M. A. Smith, and G. Chen. A two-part framework for estimating individualized treatment rules from semicontinuous outcomes. *Journal of the American Statistical Association*, pages 1–23, 2020.

[17] S. Chatterjee, S. Chowdhury, H. Mallick, P. Banerjee, and B. Garai. Group regularization for zero-inflated negative binomial regression models with an application to health care demand in germany. *Statistics in medicine*, 37(20):3012–3026, 2018.

[18] D. H. Jung, E. DuGoff, M. Smith, M. Palta, A. Gilmore-Bykovskyi, and J. Mullahy. Likelihood of hospital readmission in medicare advantage and fee-for-service within same hospital. *Health Services Research*, 55(4):587–595, 2020.

[19] T. M. Dall, P. D. Gallo, R. Chakrabarti, T. West, A. P. Semilla, and M. V. Storm. An aging population and growing disease burden will require alarge and specialized health care workforce by 2025. *Health affairs*, 32(11):2013–2020, 2013.

[20] R. E. Hall and J. V. Tu. Hospitalization rates and length of stay for cardiovascular conditions in canada, 1994 to 1999. *The Canadian journal of cardiology*, 19(10):1123–1131, 2003.

[21] D. Gupta, P. G. Vashi, C. A. Lammersfeld, and D. P. Braun. Role of nutritional status in predicting the length of stay in cancer: a systematic review of the epidemiological literature. *Annals of Nutrition and Metabolism*, 59(2-4):96–106, 2011.

[22] S. Berki, M. L. Ashcraft, and W. C. Newbrander. Length-of-stay variations within ICDA-8 diagnosis-related groups. *Medical Care*, 22(2):126–142, 1984.

[23] S. T. Rinne, J. Castaneda, P. K. Lindenauer, P. D. Cleary, H. L. Paz, and J. L. Gomez. Chronic obstructive pulmonary disease readmissions and other measures of hospital quality. *American journal of respiratory and critical care medicine*, 196(1):47–55, 2017.

[24] K.-I. Goh, M. E. Cusick, D. Valle, B. Childs, M. Vidal, and A.-L. Barabási. The human disease network. *Proceedings of the National Academy of Sciences*, 104(21):8685–8690, 2007.

[25] D.-S. Lee, J. Park, K. Kay, N. A. Christakis, Z. N. Oltvai, and A.-L. Barabási. The implications of human metabolic network topology for disease comorbidity. *Proceedings of the National Academy of Sciences*, 105(29):9880–9885, 2008.

[26] Y. Li and P. Agarwal. A pathway-based view of human diseases and disease relationships. *PloS one*, 4(2):e4346, 2009.

[27] C. A. Hidalgo, N. Blumm, A.-L. Barabási, and N. A. Christakis. A dynamic network approach for the study of human phenotypes. *PLoS computational biology*, 5(4):e1000353, 2009.

[28] Y. Jiang, S. Ma, B.-C. Shia, and T.-S. Lee. An epidemiological human disease network derived from disease co-occurrence in taiwan. *Scientific reports*, 8(1):1–12, 2018.

[29] F. S. Roque, P. B. Jensen, H. Schmock, M. Dalgaard, M. Andreatta, T. Hansen, K. Søeby, S. Bredkjær, A. Juul, and T. Werge. Using electronic patient records to discover disease correlations and stratify patient cohorts. *PLoS computational biology*, 7(8):e1002141, 2011.

[30] A. Halu, M. De Domenico, A. Arenas, and A. Sharma. The multiplex network of human diseases. *NPJ systems biology and applications*, 5(1):1–12, 2019.

[31] C. Ma, Y. Li, B. Shia, and S. Ma. Human disease cost network analysis. *Statistics in Medicine*, 39(9):1237–1249, 2020.

[32] D. J. Graham, M. E. Reichman, M. Wernecke, R. Zhang, M. R. Southworth, M. Levenson, T.-C. Sheu, K. Mott, M. R. Goulding, M. Houstoun, et al. Cardiovascular, bleeding, and mortality risks in elderly medicare patients treated with dabigatran or warfarin for nonvalvular atrial fibrillation. *Circulation*, 131(2):157–164, 2015.

[33] M. L. Schermerhorn, D. B. Buck, A. J. O'Malley, T. Curran, J. C. McCallum, J. Darling, and B. E. Landon. Long-term outcomes of abdominal aortic aneurysm in the medicare population. *New england journal of medicine*, 373(4):328–338, 2015.

[34] M. J. Van Der Laan and D. Rubin. Targeted maximum likelihood learning. *The international journal of biostatistics*, 2(1), 2006.

[35] H. A. Chipman, E. I. George, R. E. McCulloch, et al. Bart: Bayesian additive regression trees. *The Annals of Applied Statistics*, 4(1):266–298, 2010.

[36] M. A. Hernán and J. M. Robins. Causal inference: what if, 2020.

[37] M. A. Hernán and J. M. Robins. Using big data to emulate a target trial when a randomized trial is not available. *American journal of epidemiology*, 183(8):758–764, 2016.

[38] L. C. Petito, X. García-Albéniz, R. W. Logan, N. Howlader, A. B. Mariotto, I. J. Dahabreh, and M. A. Hernán. Estimates of overall survival in patients with cancer receiving different treatment regimens: emulating hypothetical target trials in the surveillance, epidemiology,

and end results (seer)–medicare linked database. *JAMA network open*, 3(3):e200452–e200452, 2020.

[39] X. García-Albéniz, J. Hsu, M. Bretthauer, and M. A. Hernán. Effectiveness of screening colonoscopy to prevent colorectal cancer among medicare beneficiaries aged 70 to 79 years: a prospective observational study. *Annals of internal medicine*, 166(1):18–26, 2017.

[40] A. K. Venkatesh, H. Mei, K. E. Kocher, M. Granovsky, Z. Obermeyer, E. S. Spatz, C. Rothenberg, H. M. Krumholz, and Z. Lin. Identification of emergency department visits in medicare administrative claims: approaches and implications. *Academic Emergency Medicine*, 24(4):422–431, 2017.

[41] A. K. Venkatesh, H. Mei, L. Shuling, G. D'Onofrio, C. Rothenberg, Z. Lin, and H. M. Krumholz. Cross-sectional analysis of emergency department and acute care utilization among medicare beneficiaries. *Academic Emergency Medicine*, 27(7):570–579, 2020.

[42] A. K. Venkatesh, C. J. Gettel, H. Mei, S.-C. Chou, C. Rothenberg, S.-L. Liu, G. D'Onofrio, Z. Lin, and H. M. Krumholz. Where skilled nursing facility residents get acute care: Is the emergency department the medical home? *Journal of Applied Gerontology*, page 0733464820950125, 2020.

[43] L. Kelley, R. Capp, J. F. Carmona, G. D'Onofrio, H. Mei, D. Cobbs-Lomax, and P. Ellis. Patient navigation to reduce emergency department (ed) utilization among medicaid insured, frequent ed users: A randomized controlled trial. *The Journal of emergency medicine*, 58(6):967–977, 2020.

[44] S.-X. Li, Y. Wang, S. D. Lama, J. Schwartz, J. Herrin, H. Mei, Z. Lin, S. M. Bernheim, S. Spivack, H. M. Krumholz, et al. Timely estimation of national admission, readmission, and observation-stay rates in medicare patients with acute myocardial infarction, heart failure, or pneumonia using near real-time claims data. *BMC health services research*, 20(1):1–9, 2020.

[45] A. T. Janke, H. Mei, C. Rothenberg, R. D. Becher, Z. Lin, and A. K. Venkatesh. Analysis of hospital resource availability and covid-19 mortality across the united states. *Journal of Hospital Medicine*, 2021.

[46] C. McDermott and G. N. Stock. Hospital operations and length of stay performance. *International Journal of Operations & Production Management*, 27(9):1020–1042, 2007.

[47] N. Hicks and R. Kammerling. The relationship between a severity of illness indicator and mortality and length-of-stay. *Health trends*, 25(2):65–68, 1993.

[48] A.-L. Barabási, N. Gulbahce, and J. Loscalzo. Network medicine: a network-based approach to human disease. *Nature reviews genetics*, 12(1):56–68, 2011.

[49] S. Epskamp. Brief report on estimating regularized gaussian networks from continuous and ordinal data. *arXiv preprint arXiv:1606.05771*, 2016.

[50] C. D. Van Borkulo, D. Borsboom, S. Epskamp, T. F. Blanken, L. Boschloo, R. A. Schoevers, and L. J. Waldorp. A new method for constructing networks from binary data. *Scientific reports*, 4(1):1–10, 2014.

[51] M. E. Tinetti, G. J. McAvay, T. E. Murphy, C. P. Gross, H. Lin, and H. G. Allore. Contribution of individual diseases to death in older adults with multiple diseases. *Journal of the American Geriatrics Society*, 60(8):1448–1456, 2012.

[52] J. C. Robinson and H. S. Luft. The impact of hospital market structure on patient volume, average length of stay, and the cost of care. *Journal of Health economics*, 4(4):333–356, 1985.

[53] A. Elixhauser, C. Steiner, and L. Palmer. Clinical classifications software (ccs) for icd-9-cm. *Databases and Related Tools from the Healthcare Cost and Utilization Project (HCUP). Agency for Healthcare Research and Quality*, 2012.

[54] H. Bueno, J. S. Ross, Y. Wang, J. Chen, M. T. Vidán, S.-L. T. Normand, J. P. Curtis, E. E. Drye, J. H. Lichtman, and P. S. Keenan. Trends in length of stay and short-term outcomes among medicare patients hospitalized for heart failure, 1993-2006. *JAMA*, 303(21):2141–2147, 2010.

[55] S. F. Jencks, D. K. Williams, and T. L. Kay. Assessing hospital-associated deaths from discharge data: the role of length of stay and comorbidities. *JAMA*, 260(15):2240–2246, 1988.

[56] E. D. Kolaczyk. Statistical analysis of network data: methods and models. *Springer: New York, N.Y.*, 2009.

[57] E. Yang, P. Ravikumar, G. I. Allen, and Z. Liu. Graphical models via univariate exponential family distributions. *The Journal of Machine Learning Research*, 16(1):3813–3847, 2015.

[58] A. Voorman, A. Shojaie, and D. Witten. Graph estimation with joint additive models. *Biometrika*, 101(1):85–101, 2014.

[59] B. Fellinghauer, P. Bühlmann, M. Ryffel, M. Von Rhein, and J. D. Reinhardt. Stable graphical model estimation with random forests for discrete, continuous, and mixed variables. *Computational Statistics & Data Analysis*, 64:132–152, 2013.

[60] H. Liu, F. Han, M. Yuan, J. Lafferty, and L. Wasserman. High-dimensional semiparametric gaussian copula graphical models. *The Annals of Statistics*, 40(4):2293–2326, 2012.

[61] A. McDavid, R. Gottardo, N. Simon, and M. Drton. Graphical models for zero-inflated single cell gene expression. *The annals of applied statistics*, 13(2):848, 2019.

[62] Z. Wang, S. Ma, M. Zappitelli, C. Parikh, C.-Y. Wang, and P. Devarajan. Penalized count data regression with application to hospital stay after pediatric cardiac surgery. *Statistical methods in medical research*, 25(6):2685–2703, 2016.

[63] S. Horvath. Weighted network analysis: applications in genomics and systems biology. *Springer Science & Business Media*, 2011.

[64] A. Marengoni, B. Winblad, A. Karp, and L. Fratiglioni. Prevalence of chronic diseases and multimorbidity among the elderly population in sweden. *American journal of public health*, 98(7):1198–1200, 2008.

[65] C. McCreary and R. N. Ríordáin. Systemic diseases and the elderly. *Dental update*, 37(9):604–607, 2010.

[66] C. Langston. Managing fluid and electrolyte disorders in renal failure. *Veterinary Clinics of North America: Small Animal Practice*, 38(3):677–697, 2008.

[67] B. T. Bateman, P. Bansil, S. Hernandez-Diaz, J. M. Mhyre, W. M. Callaghan, and E. V. Kuklina. Prevalence, trends, and outcomes of chronic hypertension: a nationwide sample of delivery admissions. *American journal of obstetrics and gynecology*, 206(2):134–e1, 2012.

[68] N. K. Arden, S. Crozier, H. Smith, F. Anderson, C. Edwards, H. Raphael, and C. Cooper. Knee pain, knee osteoarthritis, and the risk of fracture. *Arthritis Care & Research: Official Journal of the American College of Rheumatology*, 55(4):610–615, 2006.

[69] B. J. Barchiesi, R. H. Eckel, and P. P. Ellis. The cornea and disorders of lipid metabolism. *Survey of ophthalmology*, 36(1):1–22, 1991.

[70] L. Mody and M. Juthani-Mehta. Urinary tract infections in older women: a clinical review. *JAMA*, 311(8):844–854, 2014.

[71] D. Ye, F. Dong, X. Lu, Z. Zhang, Y. Feng, and C. Li. Analysis of various etiologies of hypertension in patients hospitalized in the endocrinology division. *Endocrine*, 42(1):174–181, 2012.

[72] B. R. Klarenbeek, N. de Korte, D. L. van der Peet, and M. A. Cuesta. Review of current classifications for diverticular disease and a translation into clinical practice. *International journal of colorectal disease*, 27(2):207–214, 2012.

[73] Y. C. Lee, N. J. Nassikas, and D. J. Clauw. The role of the central nervous system in the generation and maintenance of chronic pain in rheumatoid arthritis, osteoarthritis and fibromyalgia. *Arthritis research & therapy*, 13(2):211, 2011.

[74] M. E. Schneider. Cms reminds physicians of hipaa 5010 deadline. *Journal of Medicine*, 360:1740–1748, 2009.

[75] W. Alexander. The checkpoint immunotherapy revolution: what started as a trickle has become a flood, despite some daunting adverse effects; new drugs, indications, and combinations continue to emerge. *Pharmacy and Therapeutics*, 41(3):185, 2016.

[76] C. F. Friedman, T. A. Proverbs-Singh, and M. A. Postow. Treatment of the immune-related adverse effects of immune checkpoint inhibitors: a review. *JAMA Oncology*, 2(10):1346–1353, 2016.

[77] H. Mei, R. Jia, G. Qiao, Z. Lin, and S. Ma. Human disease clinical treatment network for the elderly: The analysis of medicare inpatient length of stay data. *Statistics in Medicine*, 2021.

[78] J. Donzé, D. Aujesky, D. Williams, and J. L. Schnipper. Potentially avoidable 30-day hospital readmissions in medical patients: derivation and validation of a prediction model. *JAMA internal medicine*, 173(8):632–638, 2013.

[79] R. Burton. Improving care transitions. *Health Affairs (https://www. healthaffairs. org/do/10.1377/hpb20120913. 327236/full/)*, 13, 2012.

[80] L. I. Horwitz, C. Partovian, Z. Lin, J. N. Grady, J. Herrin, M. Conover, J. Montague, C. Dillaway, K. Bartczak, L. G. Suter, et al. Development and use of an administrative claims measure for profiling hospital-wide performance on 30-day unplanned readmission. *Annals of internal medicine*, 161(10_Supplement):S66–S75, 2014.

[81] N. Meinshausen, P. Bühlmann, et al. High-dimensional graphs and variable selection with the lasso. *Annals of statistics*, 34(3):1436–1462, 2006.

[82] P. Ravikumar, M. J. Wainwright, J. D. Lafferty, et al. High-dimensional ising model selection using 1-regularized logistic regression. *The Annals of Statistics*, 38(3):1287–1319, 2010.

[83] G. I. Allen and Z. Liu. A log-linear graphical model for inferring genetic networks from high-throughput sequencing data. In *2012 IEEE International Conference on Bioinformatics and Biomedicine*, pages 1–6. IEEE, 2012.

[84] P. J. Curran and A. M. Hussong. Integrative data analysis: the simultaneous analysis of multiple data sets. *Psychological methods*, 14(2):81, 2009.

[85] J. G. Fox and J. A. Bell. Diseases of the genitourinary system. *Biology and diseases of the ferret*, pages 335–361, 2014.

[86] R. C. Langan. Benign prostatic hyperplasia. *Primary Care: Clinics in Office Practice*, 46(2):223–232, 2019.

[87] R. Khera, K. Dharmarajan, Y. Wang, Z. Lin, S. M. Bernheim, Y. Wang, S.-L. T. Normand, and H. M. Krumholz. Association of the hospital readmissions reduction program

with mortality during and after hospitalization for acute myocardial infarction, heart failure, and pneumonia. *JAMA network open*, 1(5):e182777–e182777, 2018.

[88] C. D. M. M. C. R. R. . K. C. Heckerman, D. Dependency networks for collaborative filtering and data visualization. In *Conference on uncertainty in artificial intelligenceUAI-2000*, pages 264–273, 2013.

[89] C. J. Edwards, J. Campbell, T. van Staa, and N. K. Arden. Regional and temporal variation in the treatment of rheumatoid arthritis across the uk: a descriptive register-based cohort study. *BMJ open*, 2(6), 2012.

[90] L. A. Hatfield, D. B. Kramer, R. Volya, M. R. Reynolds, and S.-L. T. Normand. Geographic and temporal variation in cardiac implanted electric devices to treat heart failure. *Journal of the American Heart Association*, 5(8):e003532, 2016.

[91] A. S. Go, E. M. Hylek, K. A. Phillips, Y. Chang, L. E. Henault, J. V. Selby, and D. E. Singer. Prevalence of diagnosed atrial fibrillation in adults: national implications for rhythm management and stroke prevention: the anticoagulation and risk factors in atrial fibrillation (atria) study. *JAMA*, 285(18):2370–2375, 2001.

[92] C. T. January, L. S. Wann, J. S. Alpert, H. Calkins, J. E. Cigarroa, J. Cleveland, J. C., J. B. Conti, P. T. Ellinor, M. D. Ezekowitz, M. E. Field, K. T. Murray, R. L. Sacco, W. G. Stevenson, P. J. Tchou, C. M. Tracy, C. W. Yancy, and A. A. T. F. Members. 2014 aha/acc/hrs guideline for the management of patients with atrial fibrillation: a report of the american college of cardiology/american heart association task force on practice guidelines and the heart rhythm society. *Circulation*, 130(23):e199–267, 2014.

[93] J. V. Freeman, D. N. Simon, A. S. Go, J. Spertus, G. C. Fonarow, B. J. Gersh, E. M. Hylek, P. R. Kowey, K. W. Mahaffey, L. E. Thomas, P. Chang, E. D. Peterson, J. P. Piccini, I. Outcomes Registry for Better Informed Treatment of Atrial Fibrillation, and Patients. Association between atrial fibrillation symptoms, quality of life, and patient outcomes: Results from the outcomes registry for better informed treatment of atrial fibrillation (orbit-af). *Circ Cardiovasc Qual Outcomes*, 8(4):393–402, 2015.

[94] G. D. Barnes, E. Lucas, G. C. Alexander, and Z. D. Goldberger. National trends in ambulatory oral anticoagulant use. *Am J Med*, 128(12):1300–5 e2, 2015.

[95] L. Wallentin, S. Yusuf, M. D. Ezekowitz, M. Alings, M. Flather, M. G. Franzosi, P. Pais, A. Dans, J. Eikelboom, J. Oldgren, J. Pogue, P. A. Reilly, S. Yang, and S. J. Connolly. Efficacy and safety of dabigatran compared with warfarin at different levels of international normalised ratio control for stroke prevention in atrial fibrillation: an analysis of the re-ly trial. *The Lancet*, 376(9745):975–983, 2010.

[96] R. Becker, S. D. Berkowitz, G. Breithardt, R. M. Califf, K. Fox, W. Hacke, J. Halperin, G. Hankey, K. Mahaffey, C. Nessel, D. Singer, D. Ardissin, and A. Avezum. Rivaroxaban-once daily, oral, direct factor xa inhibition compared with vitamin k antagonism for prevention of stroke and embolism trial in atrial fibrillation: rationale and design of the rocket af study. *Am Heart J*, 159(3):340–347 e1, 2010.

[97] ClinicalTrials.gov. the danish non-vitamin k antagonist oral anticoagulant study in patients with atrial fibrillation (DANNOAC-AF), (2017). identifier NCT03129490.

[98] ClinicalTrials.gov. comparison of efficacy and safety among dabigatran, rivaroxaban, and apixaban in non-valvular atrial fibrillation (DARING-AF), (2016). identifier NCT02666157.

[99] C. A. McHorney, C. Crivera, F. Laliberté, W. W. Nelson, G. Germain, B. Bookhart, S. Martin, J. Schein, P. Lefebvre, and S. Deitelzweig. Adherence to non-vitamin-k-antagonist oral anticoagulant medications based on the pharmacy quality alliance measure. *Current medical research and opinion*, 31(12):2167–2173, 2015.

[100] M. Sherid, H. Sifuentes, S. Sulaiman, S. Samo, H. Husein, R. Tupper, D. Thiruvaiyaru, C. Spurr, and S. Sridhar. Risk of gastrointestinal bleeding with dabigatran: a head-to-head comparative study with rivaroxaban. *Digestion*, 90(2):137–46, 2014.

[101] F. Al-Khalili, C. Lindstrom, and L. Benson. The safety and persistence of non-vitamin-k-antagonist oral anticoagulants in atrial fibrillation patients treated in a well structured atrial fibrillation clinic. *Curr Med Res Opin*, 32(4):779–85, 2016.

[102] W. H. Li, D. Huang, C. E. Chiang, C. P. Lau, H. F. Tse, E. W. Chan, I. C. K. Wong, G. Y. H. Lip, P. H. Chan, and C. W. Siu. Efficacy and safety of dabigatran, rivaroxaban, and warfarin for stroke prevention in chinese patients with atrial fibrillation: the hong kong atrial fibrillation project. *Clin Cardiol*, 40(4):222–229, 2017.

[103] E. C. Caniglia, J. M. Robins, L. E. Cain, C. Sabin, R. Logan, S. Abgrall, M. J. Mugavero, S. Hernandez-Diaz, L. Meyer, R. Seng, D. R. Drozd, G. R. Seage Iii, F. Bonnet, F. Le Marec, R. D. Moore, P. Reiss, A. van Sighem, W. C. Mathews, I. Jarrin, B. Alejos, S. G. Deeks, R. Muga, S. L. Boswell, E. Ferrer, J. J. Eron, J. Gill, A. Pacheco, B. Grinsztejn, S. Napravnik, S. Jose, A. Phillips, A. Justice, J. Tate, H. C. Bucher, M. Egger, H. Furrer, J. M. Miro, J. Casabona, K. Porter, G. Touloumi, H. Crane, D. Costagliola, M. Saag, and M. A. Hernan. Emulating a trial of joint dynamic strategies: An application to monitoring and treatment of hiv-positive individuals. *Stat Med*, 38(13):2428–2446, 2019.

[104] G. Danaei, L. A. Garcia Rodriguez, O. F. Cantero, R. W. Logan, and M. A. Hernan. Electronic medical records can be used to emulate target trials of sustained treatment strategies. *J Clin Epidemiol*, 96:12–22, 2018.

[105] E. C. Caniglia, L. P. Rojas-Saunero, S. Hilal, S. Licher, R. Logan, B. Stricker, M. A. Ikram, and S. A. Swanson. Emulating a target trial of statin use and risk of dementia using cohort data. *Neurology*, 2020.

[106] A. Atkinson, M. Zwahlen, D. Barger, A. d'Arminio Monforte, S. De Wit, J. Ghosn, E. Girardi, V. Svedhem-Johansson, P. Morlat, C. Mussini, A. Noguera-Julian, C. Stephan, G. Touloumi, O. Kirk, A. Mocroft, P. Reiss, J. M. Miro, J. R. Carpenter, H. Furrer, and H. I. V. E. R. E. i. E. Opportunistic Infections Project Working Group of the Collaboration of Observational. Withholding primary pcp prophylaxis in virologically suppressed hiv patients: An emulation of a pragmatic trial in cohere. *Clin Infect Dis*, 2020.

[107] M. E. Tinetti, L. Han, G. J. McAvay, D. S. Lee, P. Peduzzi, J. A. Dodson, C. P. Gross, B. Zhou, and H. Lin. Anti-hypertensive medications and cardiovascular events in older adults with multiple chronic conditions. *PLoS One*, 9(3):e90733, 2014.

[108] M. L. Maciejewski, B. G. Hammill, C. I. Voils, L. Ding, E. A. Bayliss, L. H. Curtis, and V. Wang. Prescriber continuity and medication availability in older adults with cardiometabolic conditions. *SAGE Open Medicine*, 6:12–10, 2018.

[109] J. M. Clements, B. T. West, Z. Yaker, B. Lauinger, D. McCullers, J. Haubert, M. A. Tahboub, and G. J. Everett. Disparities in diabetes-related multiple chronic conditions and mortality: The influence of race. *Diabetes Res Clin Pract*, 159:107984, 2020.

[110] J. M. Clements, M. Rosca, C. Cavallin, S. Falkenhagen, T. Ittoop, C. K. Jung, M. Mazzella, J. A. Reed, M. Schluentz, and C. VanDyke. Type 2 diabetes and chronic conditions disparities in medicare beneficiaries in the state of michigan. *Am J Med Sci*, 359(4):218–225, 2020.

[111] I. Hernandez and Y. Zhang. Comparing stroke and bleeding with rivaroxaban and dabigatran in atrial fibrillation: Analysis of the us medicare part d data. *Am J Cardiovasc Drugs*, 17(1):37–47, 2017.

[112] D. J. Graham, M. E. Reichman, M. Wernecke, Y. H. Hsueh, R. Izem, M. R. Southworth, Y. Wei, J. Liao, M. R. Goulding, K. Mott, Y. Chillarige, T. E. MaCurdy, C. Worrall, and J. A. Kelman. Stroke, bleeding, and mortality risks in elderly medicare beneficiaries treated with dabigatran or rivaroxaban for nonvalvular atrial fibrillation. *JAMA Intern Med*, 176(11):1662–1671, 2016.

[113] R. B. D'agostino. Propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. *Statist. Med.*, 17:2265–2281, 1998.

[114] L. V. Hedges and J. L. Vevea. Fixed- and random-effects models in meta-analysis. *Psychol. Methods*, 3(4):486–504, 1998.

[115] G. Danaei, L. A. Rodriguez, O. F. Cantero, R. Logan, and M. A. Hernan. Observational data for comparative effectiveness research: an emulation of randomised trials of statins and primary prevention of coronary heart disease. *Stat Methods Med Res*, 22(1):70–96, 2013.

[116] E. R. Haut, P. J. Pronovost, and E. B. Schneider. Limitations of administrative databases. *JAMA*, 307(24):2589–90, 2012.

[117] A. Gorst-Rasmussen, G. Y. H. Lip, and T. Bjerregaard Larsen. Rivaroxaban versus warfarin and dabigatran in atrial fibrillation: comparative effectiveness and safety in danish routine care. *Pharmacoepidemiol Drug Saf*, 25(11):1236–1244, 2016.

[118] T. C. Villines, A. Ahmad, M. Petrini, W. Tang, A. Evans, T. Rush, D. Thompson, K. Oh, and E. Schwartzman. Comparative safety and effectiveness of dabigatran vs. rivaroxaban and apixaban in patients with non-valvular atrial fibrillation: a retrospective study from a large healthcare system. *Eur Heart J Cardiovasc Pharmacother*, 5(2):80–90, 2019.

[119] A. Mentias, G. Shantha, P. Chaudhury, and M. S. Vaughan Sarrazin. Assessment of outcomes of treatment with oral anticoagulants in patients with atrial fibrillation and multiple chronic conditions: A comparative effectiveness analysis. *JAMA Netw Open*, 1(5):e182870, 2018.

[120] P. A. Noseworthy, X. Yao, N. S. Abraham, L. R. Sangaralingham, R. D. McBane, and N. D. Shah. Direct comparison of dabigatran, rivaroxaban, and apixaban for effectiveness and safety in nonvalvular atrial fibrillation. *Chest*, 150(6):1302–1312, 2016.

[121] G. Y. H. Lip, A. Keshishian, X. Li, M. Hamilton, C. Masseria, K. Gupta, X. Luo, J. Mardekian, K. Friend, A. Nadkarni, X. Pan, O. Baser, and S. Deitelzweig. Effectiveness and safety of oral anticoagulants among nonvalvular atrial fibrillation patients. *Stroke*, 49(12):2933–2944, 2018.

[122] M. R. Southworth, M. E. Reichman, and E. F. Unger. Dabigatran and postmarketing reports of bleeding. *N Engl J Med*, 368(14):1272–4, 2013.

[123] K. C. Kent, R. M. Zwolak, N. N. Egorova, T. S. Riles, A. Manganaro, A. J. Moskowitz, A. C. Gelijns, and G. Greco. Analysis of risk factors for abdominal aortic aneurysm in a cohort of more than 3 million individuals. 52(3):539–548.

[124] M. Heikkinen, J.-P. Salenius, and O. Auvinen. Ruptured abdominal aortic aneurysm in a well-defined geographic area. 36(2):291–296.

[125] A. M. Minino, S. H. Murphy, J. Xu, and K. D. Kochanek. Division of vital statistics. deaths: Final data for 2009. centers for disease control and prevention. 59(10).

[126] N. Sakalihasan, R. Limet, and O. D. Defawe. Abdominal aortic aneurysm. 365:1577–89.

[127] S. MacSweeney, M. Ellis, R. M. Greenhalgh, J. T. Powell, and J. T. Powell. Smoking and growth rate of small abdominal aortic aneurysms. 344(8923):651–652.

[128] F. A. Lederle, J. A. Freischlag, T. C. Kyriakides, F. T. Padberg Jr, J. S. Matsumura, Ted R. Kohler, Peter H. Lin, Jessie M. Jean-Claude, D. F. Cikrit, Kathleen M. Swanson, and Peter N. Peduzzi. Outcomes following endovascular vs open repair of abdominal aortic aneurysm: A randomized trial. 302(14):1535–1542.

[129] J. L. D. Bruin, Annette F. Baas, Jaap Buth, Monique Prinssen, Eric LG Verhoeven, Philippe WM Cuypers, Marc RHM van Sambeek, Ron Balm, Diederick E. Grobbee, and Jan D. Blankensteijn. Long-term outcome of open or endovascular repair of abdominal aortic aneurysm. 362(20):1881–1889.

[130] R. Patel. Endovascular versus open repair of abdominal aortic aneurysm in 15-years' follow-up of the UK endovascular aneurysm repair trial 1 (EVAR trial 1): a randomised controlled trial. 388(10058):2366–2374.

[131] S. T. Edwards, Marc L. Schermerhorn, A. James O'Malley, Rodney P. Bensley, Rob Hurks, Philip Cotterill, and Bruce E. Landon. Comparative effectiveness of endovascular versus open repair of ruptured abdominal aortic aneurysm in the medicare population. 59(3):575–582.

[132] C.-A. Behrendt, A. Sedrakyan, H. C. Rieß, F. Heidemann, T. Kölbel, and E. S. Debus. Short-term and long-term results of endovascular and open repair of abdominal aortic aneurysms in germany. 66(6):1704–1711.

[133] N. N. Egorova, A. G. Vouyouka, J. F. McKinsey, P. L. Faries, K. C. Kent, A. J. Moskowitz, and A. Gelijns. Effect of gender on long-term survival after abdominal aortic aneurysm repair based on results from the medicare national database. 54(1):1–12.

[134] R. S. Jackson, D. C. Chang, and J. A. Freischlag. Comparison of long-term survival after open vs endovascular repair of intact abdominal aortic aneurysm among medicare beneficiaries. 307(15):1621–1628. Publisher: American Medical Association.

[135] M. J. v. d. Laan and S. Rose. *Targeted Learning: Causal Inference for Observational and Experimental Data.* Springer Science & Business Media. Google-Books-ID: RGnSX5aCAgQC.

[136] M. A. Hernán, A. Alonso, R. Logan, F. Grodstein, M. J. Stampfer, W. C. Willett, J. E. Manson, and M. Robins. Observational studies analyzed like randomized experiments: an application to postmenopausal hormone therapy and coronary heart disease. 19(6):766–779.

[137] B. A. Dickerman, X. García-Albéniz, R. W. Logan, S. Denaxas, and M. A. Hernán. Avoidable flaws in observational analyses: an application to statins and cancer. 25:1601–1606.

[138] C. M. Zigler, Chanmin Kim, Christine Choirat, John Barrett Hansen, Yun Wang, Lauren Hund, Jonathan Samet, Gary King, and Francesca Dominici. Causal inference methods for estimating long-term health effects of air quality regulations. - abstract - europe PMC. (187):5–49.

[139] L. Deng and D. Yu. Deep learning: Methods and applications. 7(3):197–387. Publisher: Now Publishers, Inc.

[140] Y. Lv, Y. Duan, W. Kang, Z. Li, and F.-Y. Wang. Traffic flow prediction with big data: A deep learning approach - IEEE journals & magazine. 16(2):865 – 873.

[141] D. Wang, A. Khosla, R. Gargeya, H. Irshad, and A. H. Beck. Deep learning for identifying metastatic breast cancer.

[142] M. A. Badgeley, John R. Zech, Luke Oakden-Rayner, Benjamin S. Glicksberg, Manway Liu, William Gale, Michael V. McConnell, Bethany Percha, Thomas M. Snyder, and Joel T. Dudley. Deep learning predicts hip fracture using confounding patient and healthcare variables. 2(1):1–10.

[143] K. Chung, H. Yoo, and D. Choe. Ambient context-based modeling for health risk assessment using deep neural network. 11:1387–1395.

[144] Han C.W. Hsiao, Sean H.F. Chen, and Jeffrey J.P. Tsai. Deep learning for risk analysis of specific cardiovascular diseases using environmental data and outpatient records - IEEE conference publication.

[145] J. L. Katzman, U. Shaham, A. Cloninger, J. Bates, T. Jiang, and Y. Kluger. DeepSurv: personalized treatment recommender system using a cox proportional hazards deep neural network. 18(1):24.

[146] K. E. Mues, A. Liede, J. Liu, J. B. Wetmore, R. Zaha, B. D. Bradbury, A. J. Collins, and D. T. Gilbertson. Use of the medicare database in epidemiologic and health services research: a valuable source of real-world evidence on the older and disabled populations in the US. 9:267. Publisher: Dove Press.

[147] W. Jiang and R. Simon. A comparison of bootstrap methods and an adjusted bootstrap approach for estimating the prediction error in microarray classification. 26:5320–5334.

[148] T. Ching, X. Zhu, and L. X. Garmire. Cox-nnet: An artificial neural network method for prognosis prediction of high-throughput omics data. 14(4):e1006076.

[149] A. Dardik, G. P. Burleyson, H. Bowman, T. A. Gordon, G. M. Williams, T. H. Webb, and B. A. Perler. Surgical repair of ruptured abdominal aortic aneurysms in the state of maryland: Factors influencing outcome among 527 recent cases. 28(3):413–421.

[150] PheWas of GWAS catalog of SNPs. https://phewascatalog.org/phewas. Accessed: 2020-10-10.

[151] M. L. Schermerhorn and P. Cotterill. Endovascular vs. open repair of abdominal aortic aneurysms in the medicare population. 358(5):464–474.