A Longitudinal Study Based on Secondary Usage of Electronic Health Record for Identification of Erectile Dysfunction (ED) Risk Factors and Identification of Patients

Tianran Li

A thesis

submitted in partial fulfillment of the

requirements for the degree of

Master of Science

University of Washington

2024

Committee:

Peter Tarczy-Hornoch

Hunter Wessells

Program Authorized to Offer Degree:

Biomedical Health Informatics

University of Washington

Abstract

A Longitudinal Study Based on Secondary Usage of Electronic Health Record for Identification of Erectile Dysfunction (ED) Risk Factors and Identification of Patients

Tianran Li

Chair of the Supervisory Committee:

Peter Tarczy-Hornoch

Biomedical Informatics and Medical Education

ED affects one in five men in the United States, and its prevalence increases with age. Recognizing ED's chronic nature and the need for comprehensive, long-term healthcare documentation with multi-factor analysis, our study utilized integrated and enriched electronic health records (EHR) from the Electronic MEdical Records and Genomics (eMERGE) cohort 3 at the Kaiser Permanente/Washington University site. We developed a novel method for identifying ED cases from multiple sources to classify individuals with ED. We then conducted inferential analysis using logistic regression and Cox proportional hazard regression models, with longitudinal trajectory analysis for ED. Our study provides new insights into disease

pathogenesis, enables better clinical management, and ultimately aims to improve the quality of life and healthcare outcomes for affected individuals. The utilization of integrated EHR-based informatics holds great promise for accelerating the understanding and management of complex chronic conditions like ED.

Table of Contents

Chapter 1. Introduction *1*

1.1. ED Is a Public Health Challenge that Affects Millions of People Nationally and Globally 1
1.1.1. ED Definition, Categories, and Mechanism 1
1.1.2. ED's Prevalence and Epidemiology 2
1.1.3. ED and Its Risk Factors 3
1.1.4. The Economic Burdens of ED at Personal and Societal Levels 5

1.2. Electronic Health Record (EHR) Integrated by eMERGE Served as an Informative Source for Disease Phenotyping and Management 6
1.2.1. EHR's Advantage and Its Application 6
1.2.2. Integrated EHR via eMERGE 8

1.3. ED Exploration on EHR: Overcoming Prior Challenges 10
1.3.1. Barriers to ED Consultation Can Be Improved by Accessing EHRs 10
1.3.2. Previous Research on ED in EHR Systems 12
1.3.3. Challenges of EHR-based Research Lack ED Phenotyping 13
1.3.4. Our Study: A Different Method for ED Phenotyping, Longitudinal Tracking, and Analysis in EHR 15

Chapter 2. Methods *18*

2.1. Data Pre-Processing 18
2.1.1. Data Source and Data Integration 18
2.1.2. Data Quality, Missingness, and Imputation 19

2.2. Parametric and Non-parametric Modeling for Patients Case-Control Classification and Risk Factor Phenotyping 21
2.2.1. Methods of Phenotyping ED Patients 21
2.2.2. Method of Phenotyping Comorbidities and Coding Time-to-event Data 25
2.2.3. Statistical Modeling 28

Chapter 3. Results *31*

3.1 Cohort Characterization 31

3.2. Parametric and Non-parametric Estimation 36

3.3. Patients Trajectory over Time 56

Chapter 4. Discussion *61*

4.1. Observations and Insights from the Outputs 61
4.1.1. Correlation and Variable Selection 61
4.1.2. Insight from Regression Outputs 62

4.2. Limitations and Generalizability 65
4.2.1. Limitation of Data Properties 65
4.2.2. Limitation of Data Missingness 67
4.2.3. Generalizability 69

4.3. Suggested Future Work 71
4.3.1. Expansion of Current Analysis 71
4.3.2. Monitoring Younger Patients and Lifestyle Factors 73
4.3.3. COVID-19 as a Risk Factor for ED 74
Chapter 5. Conclusion *77*
References *79*
Appendix I *89*

List of Figures

Figure 1. Determination of case/control classification and time to event..... 22
Figure 2. Coding of ED Comorbidities..... 25
Figure 3. Distribution of frequency and type of encounters ordered by age at disease onset. Each row represents a patient and each dot was an encounter for that patient. 56
Figure 4. Kaplan-Meier estimation for ED cases entering eMERGE cohort 3 study. 58

List of Tables

Table 1. Data cleaning: quality control of vital measurement variables..... 20
Table 2. Integrated Variables for Modeling..... 27
Table 3. Characteristics of the integrated ED cohort data set..... 32
Table 4. The pairwise correlation of variables for the analysis of collinearity..... 34
Table 5 A summary of comorbidity-specific contingency tables 35
Table 6. The results of logistic regression on ED cohort, crude marginal univariable model..... 37
Table 7. The results of logistic regression on ED cohort, joint full model. CF=confounding 40
Table 8. The multi-variable logistic regression on ED cohort, removing diastolic blood pressure. CF=confounding 44
Table 9. The results of Cox regression on ED cohort, crude marginal univariable model..... 48
Table 10. The results of Cox regression on ED cohort, joint full model. CF=confounding 51
Table 11 . The multi-variable Cox regression on ED cohort, removing diastolic blood pressure. CF=confounding 54

Chapter 1. Introduction

1.1. ED Is a Public Health Challenge that Affects Millions of People Nationally and Globally

1.1.1. ED Definition, Categories, and Mechanism

ED is a chronic health problem characterized by the persistent or recurrent inability to attain or maintain an erection that is stable enough for sexual activity and satisfactory penetration in 75-90% of all attempts (NIH Consensus Conference, 1993; Burnett et al., 2018). Symptoms of ED can include occasional or chronic difficulty achieving or sustaining an erection during sexual activities and decreased libido (Bacon et al., 2002). ED is an organic chronic illness resulting from vascular, neurogenic, hormonal, and anatomical etiologies (Burnett et al., 2018; Muneer, Ralph & Minhas, 2014). A common vascular type of ED is linked to disorders involving the arteries or blood vessels, leading to an inability to obtain or maintain sufficient penile rigidity for adequate sexual performance (Burnett et al., 2018). Neurogenic ED is caused by diabetes, spinal cord injuries, and pelvic surgeries, which damage the nerves necessary for erection (Muneer et al., 2014). Hormonal ED originates from imbalances in testosterone or other sex hormones, while abnormal structures in the penis, e.g., morphological change by Peyronie's disease, will enhance the development of anatomical ED (Bacon et al., 2002). Psychogenic ED induced by stress, anxiety, and/or depression often exhibits premature ejaculation influenced by the physiological response with an acute onset (Aytac et al., 2000; Lue, 2000).

Obtaining and maintaining an erection involves serial mechanisms, beginning from sexual arousal triggering the cavernosal nerve terminals and endothelial cells in the penis to release neurotransmitters like nitric oxide (NO) (Burnett et al., 2018). Then, NO relaxes the smooth muscle cells at the corpora cavernosa by inducing the formation of cyclic guanosine monophosphate (cGMP) (Burnett et al., 2018). Increased cGMP levels relax the smooth muscle, resulting in vasodilation of penile arteries and increased blood flow into the corpora cavernosa (Lue, 2000). The expansion of the erectile tissue compresses the venous outflow, maintaining the erection (Burnett et al., 2018). Therefore, cGMP plays an important role in erection, and locally synthesized NO activates guanylate cyclase to synthesize a certain amount of intracellular cGMP, which is later cleaved by phosphodiesterase type 5 (PDE5), reducing its level within penile tissue. The balance between production and degradation of cGMP regulates the duration of an erection (Burnett et al., 2018). By prolonging erection, PDE5 inhibitors (PDE5i), such as sildenafil, inhibit the degradation of cGMP in the penis (Bacon et al., 2002). Dysregulation of these cellular pathways, such as decreased cavernosal NO production, increased oxidative stress, or other structural pathologies, leads to erectile dysfunction (Burnett et al., 2018). Conditions like diabetes, hypertension, and hyperlipidemia are commonly associated with endothelial dysfunction, resulting in vascular ED (Bacon et al., 2002).

1.1.2. ED's Prevalence and Epidemiology

ED is one of the most widespread male sexual disorders found to affect men from all across populations and ages worldwide. Although cultural and environmental factors vary, a review focusing on eight countries with high burdens of EDs, from the US to China and Brazil, also showed differences in prevalence among different environments and healthcare systems

(Kitaw et al., 2024). The 2004 Global Study of Sexual Attitudes and Behaviors (GSSAB) collected responses from over 27,000 men and women around the globe aged between 40-80 years old in more than 29 countries, reporting 12.9% impotence in Southern Europe and 20.6% among English-speaking populations (Laumann et al., 2004). The Massachusetts Male Aging Study (MMAS) followed 1290 men taking part in 1987, with a 35% prevalence of ED in men aged 40-70 (Feldman et al., 1994). Meanwhile, the National Health and Social Life Survey (NHSLS) reported a prevalence rate of 50% for ED in men 18-59 years across the US (Laumann et al., 1999). In the National Health and Nutritional Examination Survey (NHANES) data for 2001-2002, self-reported ED was found to have an 18.4% prevalence in men over 20 years old, rising to 77.5 % among men older than 75 -years (Saigal et al., 2006; Selvin et al., 2007). The overall prevalence of ED is nearly 30 million among men in the USA, while patients with comorbidities like diabetes mellitus type II, hypertension, and cardiovascular diseases are more prone to developing ED (Wessells et al.,2007).

The prevalence of ED was 33% among men aged 50-79 years in a rural New York State community-based study (Ansong et al., 2000). In one US cohort study at young adult ages, sexually active men reported erection difficulty (11.3% mild, 2.9% moderate-to-severe) (Calzo et al, 2021).

1.1.3. ED and Its Risk Factors

ED affects the quality of life, and its severity subtends commonly physical, physiological, and psychosocial components (Burnett et al., 2018; Lue, 2000; Feldman et al., 1994). The MMAS found that ED prevalence for 40-70-year-old men was 52% overall, while the prevalence of complete ED increased from 5% in the age 40 group to 15% in the age 70 group (Feldman et

al.,1994). Age-related increase in the prevalence of ED is driven by aetiological mechanisms such as vascular alterations, decreased testosterone levels or chronic comorbidities (e.g. diabetes, hypertension cardiovascular disease) (Selvin et al., 2007). Obesity is additionally related to ED and overweight men are at higher risk of suffering from cellular endothelial dysfunction or hormonal imbalance (Bacon et al., 2002). Meanwhile, race and ethnicity are other demographics that clearly impact the prevalence of ED (Laumann et al., 1999). A population-based cohort study found that Hispanic men had 1.89 times the odds of experiencing ED compared to non-Hispanic white elders, with diabetes, obesity, current smoker status, and hypertension being significant risk factors (Saigal et al., 2006).

Comorbidities such as diabetes, cardiovascular disease, and hypertension are closely associated with ED (Selvin et al., 2007). The true prevalence of ED among diabetic men is estimated at 51.3% (Selvin et al., 2007). According to Penson and Wessells (2004), studies have demonstrated that ED affects 20% to 71 % of men with diabetes (Penson & Wessells, 2004). ED pathophysiology in diabetic men usually argues vascular, neurological, and hormonal factors, rendering an explanation of the multifactorial complex nature of this condition that increases the risk for ED (Penson & Wessells. 2004). Compared to non-diabetic men, diabetic men are threefold more likely not to be able to have erections. (Bacon et al., 2002). Other crucial risk factors include hypertension and cardiovascular diseases, which cause impairment in blood flow (Wessells et al., 2007).

ED is associated with lower urinary tract symptoms (LUTS) concerning frequency, urgency, and nocturia (Rosen et al., 2003). The Multinational Survey of the Aging Male (MSAM-7) reported that men with moderate to severe LUTS had an ED rate of 72.2%, while men without LUTS had a 36.6% ED rate (Rosen et al., 2003). On the other hand, the amount of

testosterone serves as a main driving force for libido and penile erectile function, necessary in the management of not only sexual drive and erections but overall male health (Rosen et al., 2003). One of the most critical risk factors for ED is hypogonadism (Bacon et al., 2002) for the low-testosterone subpopulation. Hypogonadism was associated with age and ED (Muneer et al., 2014). Meanwhile, obesity, Type 2 diabetes, and metabolic syndrome are also concomitant conditions with low testosterone levels secondary to a higher risk of ED (Bacon et al., 2002). Men with spinal cord injuries or undergoing treatment for prostate cancer also have increased rates of ED (Rew & Heidelbaugh, 2016).

ED has huge psychological factors in its etiology; it is also related to stress, anxiety, and depression (Araujo et al., 2000). The conceptual model of an inter-relationship between ED and mental health disorders adapted from Araujo et al. (2000) created a cycle with changes in one adversely affecting the other, such that ED increases psychological distress, which worsens sexual function in return (Araujo et al., 2000). In the meantime, antihypertensives, antidepressants, and antipsychotics impact erectile features (Wessells et al., 2007).

1.1.4. The Economic Burdens of ED at Personal and Societal Levels

Patients with ED have high direct personal medical costs of treatment options. These therapies include oral medications, such as PDE5i, and more invasive penile venous occlusive procedures. The cost of surgery, such as penile prosthesis placement, can create large bills for patients (Tan, 2000). In addition, the economic burden of continuous drug treatment increases (Rezaee et al., 2020). In areas with less insurance coverage, the high out-of-pocket expenses create economic hardship and barriers to treatment access, adherence to healthcare solutions, and worse health outcomes (Rezaee et al., 2020; Elterman et al., 2021). Indirect personal costs are

mainly due to the impact of ED on work productivity and attendance. Meanwhile, sexual health is an integral component of well-being, and ED has retrievable impacts on the quality of life for patients as couples (Sonnentag et al., 2021). The ED-induced relationship difficulties and psychosocial consequences may further interfere with workplace performance.

ED also carries significant weight on a societal scale. ED treatment costs in the United States build up to billions of dollars yearly, with PDE5i drugs being a larger portion of this expenditure (Shabsigh, 2001). In contrast, employees with ED and resultant mental health disorders (e.g., depression and anxiety) suffer from more absenteeism at work but decreased workplace productivity, leading to lower worker supply in these workers. The higher rates of healthcare visits result in more work and pressure on health services concerning healthcare resources and financial implications (Elterman et al., 2021).

1.2. Electronic Health Record (EHR) Integrated by eMERGE Served as an Informative Source for Disease Phenotyping and Management

1.2.1. EHR's Advantage and Its Application

The traditional approach to evaluating the relationships between disease risk factors and intervention or treatment is randomized control trials (RCT), recruiting participants in cohort or case-control settings. They could reflect sufficient evidence to support a medical intervention (Nair, 2016). However, they may need help with the dilemma of the realistic cost of time and budget for recruitment, the study design limitation, and the selection bias of sampling (Nair,

2016). On the other hand, the RCT with a limited sample size may need to have larger statistical power or reduce the possibility of capturing statistical significance in risk factors (Verma, 2018). With the broad implementation of EHR systems all over the U.S., the limitations above could be overcome to an intermediate extent or above. EHR could capture much larger amounts and types of patient information than the pre-selected RCT variables, including yet not limited to social and behavioral factors (e.g., socio-economic determinants, tobacco, and alcohol consumption), with satisfactory sample sizes (Adler, 2015). Meanwhile, EHR could serve for secondary analysis by coding unstructured clinical narratives with aid from natural language processing algorithms and clinical standards such as International Classification of Diseases 9 (ICD-9), International Classification of Diseases-10 (ICD-10), and Health Level-7 (HL-7) (Roberts, 2015; Denny, 2013; Stanaway, 2019). The variables generated from EHR may contribute to the progression of diseases of interest and help scientists design interventions (Adler, 2015; Roberts, 2015).

EHR data has been used for surveillance of clinical disease risk factors (e.g., CVD and its risk factors such as body mass index, blood pressure, and cholesterol levels) for population health improvement initiatives (Sidebottom, 2014). EHR could also enrich the information necessary for clinical decision making by reminding higher-risk populations for regular preventive screening (e.g., cancer screening) (Goldie, 2005). Cancer development originates from the uncontrolled proliferation of healthy cells due to genetic mutation caused by UV exposure, and/or contact with physical or chemical carcinogens (Boytcheva, 2019). The expansion of cancer cells could possibly be prevented before the benign cells enter the stage of metastasis (Goldie, 2005). Although the prediction of cancer from EHR is not quite approachable, it was suggested that community-based cancer screening could hugely raise the

measurable cost-effectiveness of treatment or prevention for cancer progression (Goldie, 2005). If residents participate in regular screening for cancers, the healthcare providers could capture more true positives of cancer incidences at an earlier stage of cancer (Boytcheva, 2019). Carcinogenesis may proceed much more slowly with early intervention or treatment to prevent enormous costs for late stages (Boytcheva, 2019). The role of EHR in early detection and intervention for cancer prevention is crucial and will likely lower the cost of surgery and treatment, reassuring healthcare professionals and policymakers about the potential of EHR in improving patient outcomes (Boytcheva, 2019).

EHR has also been used for revealing and characterizing the disease factors among the linked open data for inference and prediction (Boytcheva, 2019). By analyzing the patient data on a large scale, we may find more cancer biomarkers and causal factors for optimizing screening protocols and excellent comprehension of carcinogenesis for treatment and prevention (Barnes, 2011; Norton, 2014; Broce, 2019). Observational epidemiology studies have found that CVD-/lifestyle-related risk factors are associated with dementia risk and targeting these modifiable Risk factors (e.g., diabetes, hypertension, obesity, physical inactivity) could present a viable dementia prevention strategy which prevented 35% of dementia (Barnes, 2011; Norton, 2014; Broce, 2019). The potential of EHR in revealing and characterizing disease factors is vast, and it is sure to intrigue healthcare professionals and researchers about the possibilities of EHR in healthcare research.

1.2.2. Integrated EHR via eMERGE

The EHR has been proven to enable a vision of "personalized medicine" in which patient information and genomic messages are incorporated into precision medicine (McCarty, 2011). Studies using clinical measurements from EHRs permit a long-term average of multiple independent clinical measurements from many different clinical visits, yielding reduced phenotypic variance (Ganesh, 2014). Accompanied by appropriate quality control and coding granularity during integration from various data providers, the scope of pooled EHR, ranging from tens of thousands to millions of samples, afford insight into treatment or disease effects that may be under-reported or missed in the underpowered RCTs (Nair, 2016). Many biobanks are beginning to link data from study participants' or patients' EHR with their genetic data to accelerate the study of genotype-phenotype associations while potentially facilitating clinical activities without genetic specialists (Lemke, 2010; Ayatollahi, 2019). The Electronic MEdical Records and GEnomics (eMERGE) Consortium was initiated in late 2007 by the National Human Genome Research Institute with a goal to develop, disseminate, and apply approaches to research that combine DNA biorepositories with EHR for large scale high-throughput genetic research (McCarty, 2011). The consortium has grown to cover many study sites including University of Washington and Group Health Cooperative in Washington, Fred Hutchinson Cancer Research Center, Washington State University, Mountain Part Health Center, Arizona State University, Mayo Clinic, Northwestern University, University of Chicago, Loyola University, Cincinnati Children's Hospital, University of Cincinnati, Vanderbilt University, University of Alabama at Birmingham, The Children's Hospital of Philadelphia, Mount Sinai /Columbia University, Massachusetts General Hospital/Harvard School of Public Health/Boston Children's Hospital, and Yale University (eMERGE Consortium, 2024). The consortium is still expanding and has reached its third phase, covering more than 130,000 cohort participants

(eMERGE Consortium, 2024). Beginning from its first phase, eMERGE has conducted studies that successfully used EHR data to classify cases or control, generate standardized disease-related phenotypes, and impute genotypes for relevant genome-wide association studies (GWAS) (Gottesman, 2013; Kho, 2011). The eMERGE allowed researchers to conduct influential association studies, requiring more than 10,000 samples, which were hard to reach during traditional RCTs (Park, 2010). The data mining on unstructured portions of EHR also suggests disease phenotypes (Bush, 2016). Therefore, our data provider, the KPUW site, has the vast potential to contribute to the phenotyping of ED with enriched EHR sources (eMERGE Consortium, 2024).

1.3. ED Exploration on EHR: Overcoming Prior Challenges

1.3.1. Barriers to ED Consultation Can Be Improved by Accessing EHRs

Despite the known prevalence and burden of ED symptoms, the reporting frequency and diagnosis rates have remained low (Rew, 2016). Embarrassment, lack of time, and privacy concerns often prevent men from seeking medical care for ED (Mark et al., 2024). Financial constraints also contribute to this reluctance, while the fear of embarrassment or humiliation when discussing sexual health issues can further deter men from seeking help and disclosing relevant information during medical consultations (Aguiar, 2024). Challenges to ED management include barriers involving patient and provider engagement (Wessells, 2007). The stigma surrounding sexual health issues and societal attitudes towards men with ED significantly contribute to the underreporting and undertreatment of the condition (Rew, 2016). Patients or

providers may not feel comfortable talking about sexual health with inadequate background or training in the diagnosis and treatment of ED such that opportunities for early detection during a routine medical encounter are often missed (Wessells et al., 2007).

Moreover, ED usually coexists with other chronic diseases such as diabetes mellitus and hypertension. In reality, in many cases, when a patient seeks out his physician for a review of sexual function, dysfunction is diagnosed simultaneously with other disease comorbidities (Bacon et al., 2002). Providers do not systemically screen for ED and only address this if the patient raises these concerns. In the absence of universal screening practices, many men with mild to moderate ED are undiagnosed until the conditions become severe or comorbidities related to cardiovascular disease have developed (Yafi et al., 2016).

Triggers in EHR systems can be built to remind healthcare professionals during routine office visits with at-risk populations for ED (e.g., having comorbidities like diabetes and hypertension) to evaluate and discuss sexual function regularly (Yafi et al., 2016).

Triggers in EHR systems can be built to remind healthcare professionals during routine office visits with at-risk populations for ED (e.g., having comorbidities like diabetes and hypertension) to evaluate and discuss sexual function regularly (Yafi et al., 2016). Integrating each patient's medical history through EHR systems can create a comprehensive profile that offers insight into the relevant health background gendering of comorbid conditions and current medication use required to better understand the health state (Bacon et al., 2002). Additionally, EHR systems may be used to identify patients with higher risk for ED according to their medical history, comorbidities, and lifestyle using patient data, which in turn would enable the primary care providers to help these patients by preventing further escalation. Although EHRs have great

potential to improve ED care and research, concerns surrounding the integrity of data captured in electronic form, patient angst concerning privacy and accuracy, as well changes related to information distortion need addressing before many practices can begin to experience the benefits (Selvin et al.,2007; Wessells et al., 2007). Addressing these challenges will help increase the accuracy and dependability of ED detection and treatment by healthcare professionals (Selvin et al., 2007; Wessells et al., 2007), ultimately improving patient outcomes.

1.3.2. Previous Research on ED in EHR Systems

Most previous research using EHR data on ED care patterns has been cross-sectional or short-term, observational, and often with limited follow-up of less than two years (Nunes et al., 2021a). This restriction prevents gaining insights into the temporal dynamics of ED and its risk factors. For example, a study on cardiovascular outcome risks in patients with ED-prescribed PDE5i and nitrates investigated acute outcomes of cardiovascular adverse events but was not able to focus on the long-term progression of ED or comorbidities within a large sample size (Nunes et al., 2021a). Goldstein et al. (2019) provided an epidemiological update on ED prevalence across eight countries and determined regional variations in the frequency of specific clinically identified EDs and relevant demographic variables associated with probing for cases (Goldstein et al., 2019). Mulhall et al. (2016) analyzed the relationship between age and increased ED diagnosis and treatment rates. However, the cross-sectional design of these studies restricted follow-up on ED progression (Goldstein, 2019; Mulhall, 2016).

The study by Nackeeran et al. (2021) exploring the effects of ED as a modifiable risk factor for major depressive disorder concluded that addressing ED might potentially reduce

depressive symptoms yet did not include other predictors for a multifactorial analysis (Nackeeran, 2021). Many studies do not sufficiently account for potential confounders and control only one or a few covariates in statistical modeling, leaving unexplored areas reflecting multifactorial associations of ED with other health conditions among EHR-based data (Nunes, 2021b; Chu, 2021; Hebert, 2023).

Meanwhile, the data source might affect ED prevalence and its relation to other health outcomes. For instance, in the study by Goldstein et al. (2019), self-reported data was used to assess ED prevalence in eight countries, which might not accurately reflect participants' status. Similarly, Mulhall et al. (2016) relied on self-reported ED diagnosis and treatment data in their study investigating the relationship between age and ED, while reluctance to answer sexual health-related questions might contribute to recall bias, and ED prevalence could have been inaccurately reported due to self-reported data (Mulhall, 2016).

Structured data studies present data fragmentation and a lack of temporal dynamics, while a consistent information format cannot compensate for non-longitudinal follow-up and clinical insight. Solving these problems could potentially alleviate the issues of ED undertreatment and underreporting significantly.

1.3.3. Challenges of EHR-based Research Lack ED Phenotyping

Numerous barriers still exist in the field of epidemiology and interventions for EDs. Some secondary limitations of EHRs (e.g., lack of standard data sets and incompleteness or sparsity) make quantitation through accurate methods even more challenging (Penson &

Wessells, 2004). Informative covariates, e.g., comorbidities, medication use, and lifestyle factors, are implicitly relevant as the lack of records can introduce bias in assessment for high-prevalence diseases like ED (Penson & Wessells, 2004; Selvin et al., 2007). Another issue that complicates the interpretation of prevalence rates calculated across studies and populations is the shortage of a uniform definition for EDs (Kessler et al., 2019).

Previous studies have identified comorbidities based on complex algorithms using EHR data, where misclassification bias and model performance optimization are present. Li et al. (2020) created a rule-based Computable Phenotype (CP) algorithm on structured EHR integrated with unstructured clinical narratives to identify sexual minorities in the population (Li et al., 2020). Moreover, Hanauer et al. (2014) found that structured billing codes and keyword searches of free-text clinical documentation captured more patients with disorders of sex development (DSD) than traditional methods in various institutional settings (Hanauer et al., 2014). In 2023, Woldemariam et al. studied a multicenter cohort of male patients using data-driven case-control definition techniques to analyze the association between comorbidities and male infertility (Woldemariam et al., 2023). Among sensitivity analyses, controls were matched by race or ethnicity, area deprivation index, and hospital utilization metrics to test individual associations for diagnosis with phenotype mappings using a chi-squared test with Bonferroni correction (Woldemariam et al., 2023).

However, every strategy has limitations. Li et al. (2020) suggested that using the existing standard of clinical quality documentation might underestimate cases without explicit EHR description. Woldemariam et al. (2023) faced challenges due to missing key variables like comorbid conditions and lifestyle factors resulting from incomplete and fragmented data.

Hanauer et al. (2014) experienced challenges in standardizing data collection procedures and heterogeneity in data quality across institutions (Hanauer et al., 2014). There has been little work on phenotyping ED or its comorbidities or risk factors, and this current situation leaves an opportunity for methodology improvements in future study designs.

1.3.4. Our Study: A Different Method for ED Phenotyping, Longitudinal Tracking, and Analysis in EHR

ED is a complex chronic disease with numerous risk factors, including mental health issues and other comorbidities (Wessells et al., 2007). Therefore, we designed a new method to identify ED patients and comorbidities of interest within a centralized EHR system. Our approach allows tracking longitudinal data, linking multiple EHR sources, and enables marginal and joint association studies between the risk of ED and important covariates. The risk factors we identified and assessed were obesity, race, age, cardiovascular disease, diabetes, hypogonadism, depression, hypertension, and LUTS. We collaborated with clinical experts and data analysts at Kaiser Permanente (KP) and the University of Washington Medical Center (UWMC) to develop and validate an interpretable and generalizable method for patient identification using longitudinal claims data and medical visits data to compensate for underreporting in clinical notes.

Insurance claims databases allow researchers to study more individuals, powered by larger samples that increase statistical power and higher precision that captures more rare events (Schneeweiss et al., 2001; Jensen et al., 2012). Claims data reflect real-world clinical practice

and patient outcomes among general populations compared to traditional prospective studies, such as controlled trials using nonclinical subjects (Garrison et al., 2007). Standardized coding systems like the International Classification of Diseases (ICD) and Current Procedural Terminology (CPT) improve accuracy, consistency, and completeness in data compared to other measures like diagnosis or brand drug name for use across providers and regions with generalizability (Schneeweiss et al., 2005). Claims data can complement clinical records by integrating EHR data, offering a more comprehensive view of patients' health conditions and healthcare utilization.

Compared to previous studies, our approach has a three-fold novelty:

1. Identification of ED patients in EHR data
2. Identification of comorbid conditions/symptoms of ED
3. Longitudinal tracking of this dataset for over 20 years with statistical analysis

Previous work has rarely addressed phenotyping or case identification methods to capture ED patients and relevant comorbid conditions in an EHR and follow-up more than ten years after their initial visits. This situation allows for identifying patients with ED concerning diagnoses and records on procedures or prescription drugs related to ED traces, thereby avoiding dependence solely on sporadic administrative contacts where the diagnosis is presented (Woldemariam et al., 2023). By phenotyping the ED comorbidities, we can perform association analyses controlling for other confounding factors in ED without losing disease information. Further, we intend to analyze patient EHR visits and encounters over the trajectory of their EHR from entry to exit, ensuring that the collection time difference and fragmentation of the EHR over decades will be manageable for statistical analysis.

Chapter 2. Methods

2.1. Data Pre-Processing

2.1.1. Data Source and Data Integration

A retrospective study was conducted utilizing the EHR extracted from the Kaiser Permanente/ University of Washington Medical Center (KPUW) site at the eMERGE cohort 3, totaling 3103 male patients treated at a regional insurer or health system from 1992-2020. Subjects who had EHR evidence of ED-based upon ICD-9/10 diagnosis codes (Dx), ICD/CPT procedure codes (Px), or prescriptions (Rx) were included. The datasets of lab measures, times entering EHR, and vital measures were also provided to analyze patients' evidence of relevant clinical factors. The subjects were continuously followed throughout their care within the EHR system.

The data were extracted from several datasets provided by KPUW: The Demographic dataset contained 7,794 records of all patients, with their eMERGE ID, year of birth, ethnicity (Hispanic or no), race, and gender. The original Dx dataset contained 3,709,630 records of the patients' diagnosis at KPUW, covering patients' 8-digit eMERGE ID, the ICD-9/10 diagnosis codes (variable of diagnosis code denoted as "dx") for each visit, and the age of patients' each corresponding diagnosis. Similarly, the original Px dataset contained 5,662,116 records of the patients' medical procedures utilized at KPUW, covering patients' eMERGE ID, ICD/CPT procedure codes (variable coded as "px") for each visit, and the age of patients' each corresponding procedure. The original Rx dataset contained 113,035 records with the patients'

eMERGE ID, 11-character National Drug Code (NDC) of the prescribed medication, NDC in Food and Drug Administration format (0000-0000-00 form), prescription date in day-month-year format, and the age of patients at the prescription. The laboratory dataset contains 268,222 records, coding patients' eMERGE ID, each lab test's type, corresponding local code and LOINC code, modifier code, result count, result unit, and patient age when he received the laboratory test. 7,793 records of patients' first entry into KPUW EHR and last exit of EHR were coded by the time-in-EHR dataset. And 391,573 records of patients' vital measures taken in each EHR visit, including raw body mass index (BMI), height (ht) in inches, weight (wt) in pounds, systolic blood pressure (bp), diastolic blood pressure (DBP), and age at vital measurement were analyzed. The patients' eMERGE ID indexed all the datasets above.

The Px, Dx, and Rx datasets were used to code patients' case or control identity (based on whether the patient has ED-related medical procedures, doctor diagnosis of ED, or picked up ED medications from the pharmacy) and the earliest time for the event of ED. Patients' times of entering and exiting EHR (from the time-in-EHR dataset) were used to determine the observational windows of each patient. The laboratory dataset performed a confirmation of comorbidities, and the vital measurement dataset provided precise BMI, systolic blood pressure, and diastolic blood pressure of each patient. The datasets above were then integrated into a meta-table of patient information, and the patients' eMERGE ID was the universal key to information matching among datasets.

2.1.2. Data Quality, Missingness, and Imputation

Variable	Quality Issue	Solving Method
Height (ht, in)	Missingness	Replaced with median ht
	Inputting error	Replaced with median ht
Weight (wt, lb)	Missingness	PMM*
	Inputting error	Eliminating Outliers
		PMM
Raw BMI	Missingness	PMM
	Inputting error	Retrieved from ht&wt
		Replaced with median BMI
SBP	Missingness	PMM
DBP	Missingness	PMM
Age	None	None

Table 1. Data cleaning: quality control of vital measurement variables

* Note: PMM is the abbreviation of predictive mean matching model, an imputation method minimizing predictive distance of neighboring measures

While coding the vital measurement datasets (n=391,573), the columns for raw BMI (193,732 missings), patient height (250,015 missings), patient weight (77,589 missings), systolic blood pressure (53,034 missings), and diastolic blood pressure (53,058 missings) exhibited varying degrees of missingness. Additionally, the data quality for BMI, height, and weight was

inconsistent, with outliers and erroneous inputs. Table 1 outlines the methods used to address the challenges of data quality and missingness within the vital measurement dataset.

Given that patient height remains relatively stable after the age of 18, median heights were employed to substitute missing or incorrectly inputted height values. However, patients' BMI and weight showed unpredictable changes over their tenure in the EHR, which typically spanned over ten years. After removing extreme outliers (e.g., 798 lbs., 923 lbs.), the missing and erroneously inputted weight values were normalized using the predictive mean matching (PMM) model. This model estimates the predicted value by minimizing the distance between neighboring measurements over time.

Raw BMI also contained several erroneous inputs that were inconsistent with the patient's actual BMI calculated from the height and weight of the corresponding visit. Therefore, all original BMI readings were proofread, and extreme outliers and significantly inconsistent inputs were replaced by BMIs calculated from weights and height or by the median of all BMIs from the same patient. Systolic and diastolic blood pressures had slightly less missingness than the other columns in the vital measurement dataset and did not contain extreme outliers; thus, only the missing values were imputed using PMM locally.

2.2. Parametric and Non-parametric Modeling for Patients Case-Control Classification and Risk Factor Phenotyping

2.2.1. Methods of Phenotyping ED Patients

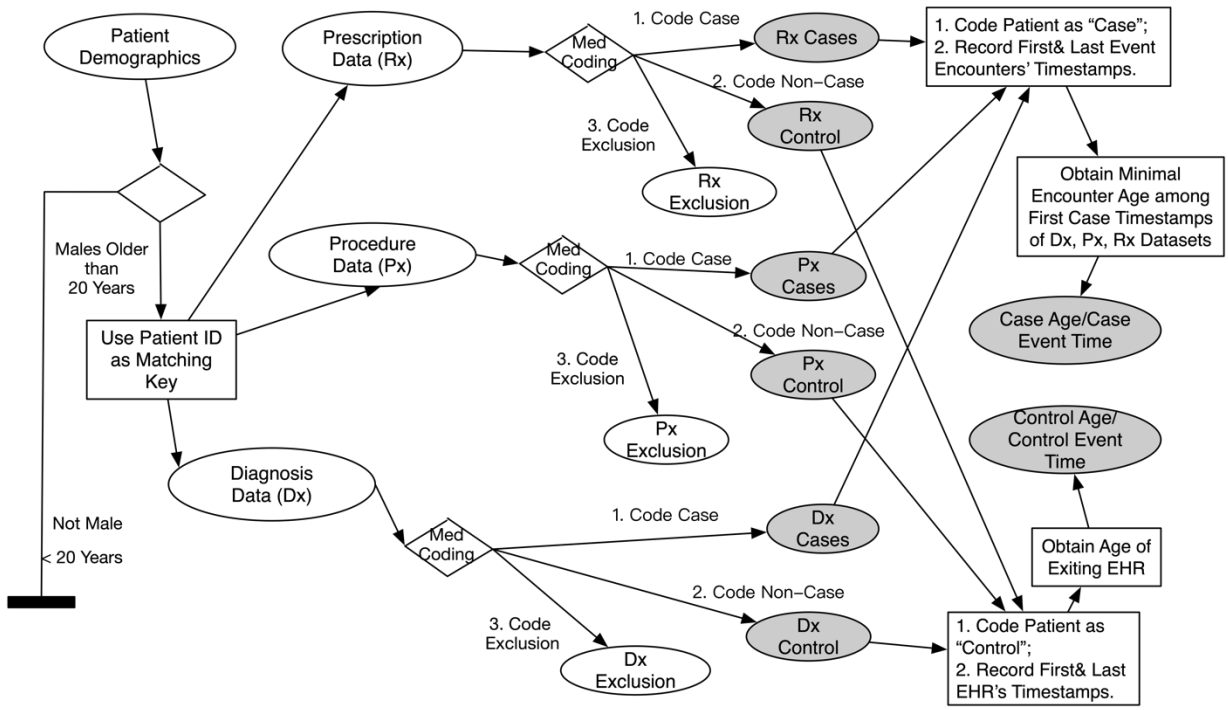

Figure 1. Determination of case/control classification and time to event

Due to the challenges of left-truncated and right-censored data, a comprehensive survival analysis was essential for extracting valuable insights from the non-uniform observational windows across each patient's records. A curated meta-table was developed to facilitate both parametric and non-parametric analyses. This table integrated key demographic risk factors, including age and race, and clinical risk factors, such as median BMI, systolic blood pressure, and diastolic blood pressure. Additionally, it encompassed binary indicators for comorbidities—cardiovascular disease ("has_CVD"), diabetes ("has_Diabetes"), hypogonadism ("has_Hypogonadism"), depression ("has_Depression"), lower urinary tract symptoms ("has_LUTS"), and hypertension ("has_hypertension").

Next, patients were categorized as excluded, case, or control within the cohort. Males under 20 years, females, and patients diagnosed with prostate cancer, spinal cord injury, radical pelvic surgery, and hyperactive sexual desire disorder were excluded from the analysis. Patients with EHR evidence of organic ED, psychosexual dysfunction, ED-related procedures (e.g., penile venous occlusive procedure), and ED-relevant medications were classified as ED cases, as specified in Appendix I. Patients with no ED diagnoses, no ED medication, or no ED procedure were determined to be controls of the cohort. The classification of case/control/exclusion was based on the Rx, Px, Dx, and Demographics datasets and underwent two rounds of iteration.

In the first iteration, patients in the Demographic dataset with age <20 and female gender had their eMERGE ID marked, and all their records in the Rx, Px, and Dx datasets were excluded. For the remaining subjects, all EHR encounters in the Px, Dx, and Rx datasets were classified, based on the inclusion criteria specified in Appendix I, coarsely as case, control, or exclusion. Each visit was temporarily coded as “case”, “control”, or “exclusion”. The patients’ first and second times of ED diagnosis, procedure, or medication were recorded based on the age at the recorded EHR visit. Each coding procedure iterated on the Px, Dx, and Rx datasets, creating temporary variables for patients’ dataset-specific recording of minimal age and second time of Px, Dx, or Rx visits. After iterating over all EHR datasets, the times in units of patient age) of the first and second “case” visits were collected as values or N/A. If the patient had an “exclusion” visit before the first “case” visit, he was defined as excluded from the cohort study. If there were no “case” or “exclusion” events throughout the patient’s EHR history, he was defined as a control within the cohort.

Age at ICD/CPT record was the time-to-event metric used in analyses, coded as "PatientAge." Both left and right censoring were implemented in this analysis. A subject's first ICD/CPT record was used as the left censor age (age_start). A minimum age of 20 years old was utilized for left truncation and used as the minimum age for left censoring (age_start). The right censor was the last ICD/CPT record for controls (age_stop). Cases had the age at the first ICD/CPT record as the age at the event (age_stop). In later survival analysis for cases, the "time to event" variable was determined by measuring the interval from an initial point, such as the beginning of the study or the time of first exposure, to the occurrence of the event, typically an ED EHR record. If the event did not occur by the study's end or the last follow-up, the data was considered censored (and this individual was a control), indicating the event did not happen during the observed period. Therefore, the event was determined to be either from the entry of the study until ED onset for ED cases or from the entry of the study until the exit of the study for ED controls. Patients' observed survival times were coded as the variable "time".

Prior to coding comorbidity covariates and case/control status, patient information including birth year (demog_birth_year), race (race1.factor), age at eMERGE EHR entry (Age_Entry_in_EHR), age at eMERGE EHR exit (exitAge), calculated time spent in the EHR system from entry to exit (observed_Years_in_EHR), median standardized BMI (median.BMI), median standardized systolic blood pressure (median.SBP), median standardized diastolic blood pressure (median.DBP), time/age at the second diagnosis encounter (Dx.2nd.age), minimal age of the ED diagnosis at the first Dx encounter (Dx.min.age), and minimal age of an ED medication at the first Rx encounter (Rx.min.age) were recorded. Once the minimal ages of the first encounter in Dx, Px, and Rx were recorded for each patient, the category and time of the first ED encounter in EHR were recorded as "first.case.class" and "min.age.case".

2.2.2. Method of Phenotyping Comorbidities and Coding Time-to-event Data

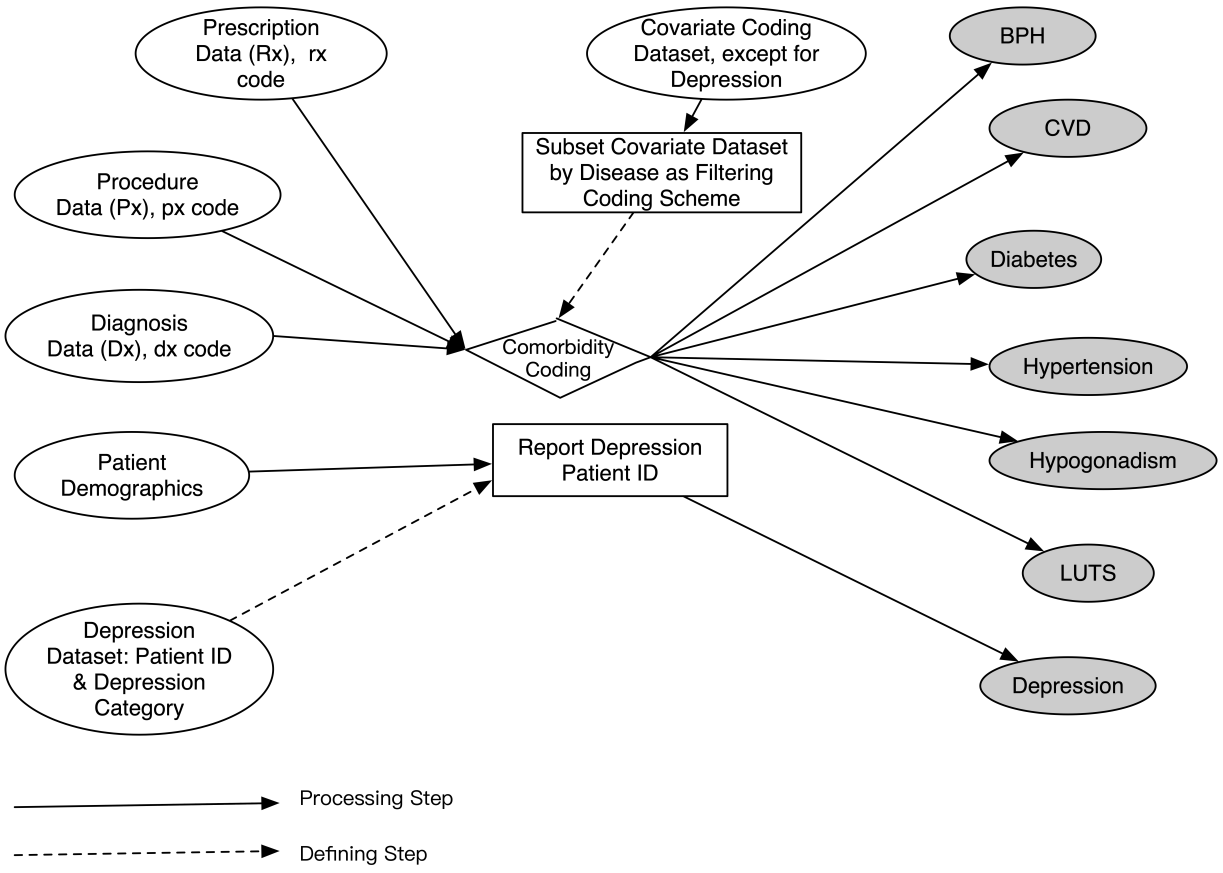

Figure 2. Coding of ED Comorbidities

Other than the status of patients' ED-related diagnoses, procedures, and medications, we were also interested in ED's relevant comorbidities, as listed in aforementioned past publications, to analyze whether the ED outcome was associated with clinical risk factors.

The ED comorbidities of interest, selected with approval from clinicians and researchers, included BPH, CVD, diabetes, hypertension, hypogonadism, LUTS, and depression. All comorbidities were coded as binary variables (0/1), indicating whether the patients had an indicator of the disease in the EHR. The coding of patients' depression status was independent of other coding procedures, extracting patients' eMERGE ID from the Demographic dataset and matching it with the processed depression dataset provided by the eMERGE consortium using an algorithm that reported patients' eMERGE ID. If the patient's ID appeared in the eMERGE depression dataset, he was coded as having depression. Other comorbidities, including BPH, CVD, diabetes, hypertension, hypogonadism, and LUTS, were determined by the patients' EHR. Each disease of interest had a coding dataset containing the relevant ICD-9 and ICD-10 codes. All patients' procedures (Px), diagnoses (Dx), and prescriptions (Rx) were filtered by the coding dataset, and an indicator of the patient's disease status was reported if records of comorbidity were found. The iterative coding of comorbidity ceased once all comorbidities were tested in all patients' EHRs.

All variables relevant to regression and survival analysis, including comorbidity indicators and patient demographics (birth year, race, age at eMERGE EHR entry, and duration in EHR), were integrated into a unified meta-dataset as a comprehensive disease history document, as shown in Table 2 below.

Variable Name	**Variable Type**	**Variable Source**	**Note**
EMERGE ID	Numeric	Demographical Data	
Case/Control	Binary	Secondary Coding	Refer to Fig.1
Age Entry in EHR	Numeric	EHR History Data	

Age Exiting EHR	Numeric	EHR History Data	Validate exit event time
Time-to-event	Numeric	Secondary Coding	Refer to Fig. 1
Min Age Rx Case	Numeric	Rx Data	
Min Age Dx Case	Numeric	Dx Data	
Min Age Px Case	Numeric	Px Data	
Patient Age	Numeric	Secondary Coding	
Birth Year	Numeric	Demographical Data	
Race	Categorical	Demographical Data	
Median DBP	Numeric	Vital Data	
Median SBP	Numeric	Vital Data	
Median BMI	Numeric	Vital Data	
has_BPH	Binary	Secondary Coding	Refer to Fig. 2
has_CVD	Binary	Secondary Coding	
has_hypertention	Binary	Secondary Coding	
has_Diabetes	Binary	Secondary Coding	
has_depression	Binary	Secondary Coding	
has_hypogonadism	Binary	Secondary Coding	
has_LUTS	Binary	Secondary Coding	

Table 2. Integrated Variables for Modeling

2.2.3. Statistical Modeling

Descriptive statistics and all statistical models were conducted in RStudio (version 1.4.1103, by Posit PBC, formerly RStudio PBC/ RStudio Inc) utilizing R (version 3.6.2, by the R Foundation for Statistical Computing).

To assess the relationship between the odds or hazard of the ED and the covariates covering all risk factors of interest extracted from EHR integrated EHR from the KPUW subpopulation of eMERGE cohort 3 were followed from the time of the entry of EHR until the exit of EHR and was thoroughly coded. The 11 covariates of interest are: patient age (in years), race (white, Asian, Black, Hispanic, American Indian Other, Unknown), median of patient BMI, median of patient DBP, median of patient SBP, indicator of cardiovascular disease (CVD=1, non-CVD=0), indicator of diabetes (diabetes=1, non-diabetes=0), indicator of hypogonadism (hypogonadism=1, non-hypogonadism=0), indicator of depression (depression=1, non-depression=0), indicator of hypertension (hypertension=1, non-hypertension=0), and indicator of LUTS (LUTS=1, non-LUTS=0). The response variable for the logistic regression is the indicator of ED status (1=ED positive case, 0=ED-negative control).

To evaluate the association between ED and covariates of interest marginally and jointly, we compared the odds of patient groups differing by one unit of covariate with Huber-White robust standard errors. We fitted a retrospective logistic regression model using an indicator of ED case (case=1, control=0) as the response and all aforementioned 11 covariates as the predictors. We tested a null hypothesis of equal odds of ED between patient groups differing by one unit of covariate readings marginally (for jointly multivariable models, patient groups are

homogenous in all other variables) against the alternative of unequal odds using a Wald test with robust standard errors and defined significance as $p < 0.05$. We reported a point estimate and 95% Wald interval calculated with robust standard errors on the log odds scale for the odds ratio (exponentiated) comparing ED odds between patient groups differing by one unit of covariate readings marginally (for jointly multivariable models, patient groups are homogenous in all other variables).

Meanwhile, we utilized semiparametric Cox proportional hazards (PH) models depending on the assumption that hazard functions in subgroups are proportional to one another. The time of disease onset was recorded for those who developed ED by plotting the Kaplan-Meier curve. Both marginal and joint modeling with adjustment were performed. We used crude, univariate Cox PH models to analyze the marginal associations between 11 covariates and hazards of ED. Later, the full joint Cox models were adjusted for in a multivariate Cox PH model (covering all 11 covariates) to assess joint associations. It compared the hazards of ED between patient groups, differing by one unit of covariate readings, holding all other variables constant, producing HRs and 95% CIs for the associations observed. Exponentiated model coefficients yield hazard ratios, or the relative hazards of an outcome, when comparing two patient groups. Inferential analyses rely on an alpha 0.05 to minimize the probability of type I errors in findings.

In most modeling analyses adjusted for covariates, we did not utilize stratification or interaction to preserve data richness with an intermediate sample size range no larger than 5000.

Chapter 3. Results

3.1 Cohort Characterization

	Control (N=1794)	Case (N=1041)	Cohort(N=2835)
Age	71.73 (IQR 63.55-81.58)	66.00 (IQR 57.19-73.03)	69.10 (IQR 61.00-78.44)
Race: Asian	18% (316)	9% (90)	14% (406)
Race: Black	2% (39)	3% (32)	3% (71)
Race: Hispanic	1% (12)	0.3% (4)	1% (16)
Race: American Indian	1% (10)	1% (10)	1% (20)
Race: Other	0.4% (6)	1% (11)	1% (17)
Race: Unknown	3% (59)	4% (46)	4% (105)
Race: White	75% (1352)	81% (848)	78% (2200)
Median BMI	26.7 (IQR 24.8-28. 8)	27.3 (IQR 25.1-29.9)	26.9 (IQR 24.9-29.3)
Median SBP	127.0 (IQR 120.0-132.0)	129.5 (IQR 122.0-135.0)	128.0 (IQR 120.0-133.0)
Median DBP	74.0	72.0	74.0

	(IQR 70.0-79.0)	(IQR 69.0-78.0)	(IQR 70.0-78.5)
Cardiovascular Disease	46% (819)	59% (617)	51% (1436)
Diabetes	20% (366)	29% (302)	24% (668)
Hypogonadism	2% (39)	8% (86)	4% (125)
depression	15% (271)	25% (258)	19% (529)
LUTS	2% (39)	8% (86)	4% (125)

Table 3. Characteristics of the integrated ED cohort data set.

The study analyzed a cohort of 2835 individuals, focusing on their demographic and clinical characteristics. The median age at entry into the EHR system was 51.5 years (IQR 40.5-60.7 years), and the median age at erectile dysfunction (ED) diagnosis was 65.0 years (IQR 56-72.6 years). This situation indicates a predominantly middle-aged to elderly cohort, with the case group having a lower average age (median 66.00 years, IQR 57.19-73.03 years) than the control group.

The racial distribution revealed that the majority of participants were White (78%), followed by Asians (14%), Blacks (3%), Hispanics (1%), American Indians (1%), and others (1%). Additionally, 4% of the participants' race was unknown, reflecting a significant predominance of white individuals within the cohort.

Health metrics showed a median BMI of 26.9 (IQR 24.9-29.3), with cases having a slightly higher median BMI (27.3) compared to controls (26.7). The median systolic blood pressure (SBP) was 128.0 mmHg (IQR 120.0-133.0), and the median diastolic blood pressure (DBP) was 74.0 mmHg (IQR 70.0-78.5), with a slightly higher median systolic blood pressure in cases compared to controls (129.5 vs. 127.0 mmHg), revealing variations in patient health metrics. The prevalence of comorbid conditions was relatively high. CVD was present in 51% of the cohort, diabetes in 24%, hypogonadism in 4%, depression in 19%, and LUTS in 4%, highlighting the significant burden of comorbidities within the cohort. Cardiovascular diseases were observed in 59% of cases compared to 46% of controls, indicating a higher prevalence in the ED group. Similarly, diabetes was more prevalent in ED cases at 29% versus 20% in controls among the cohort. The analysis also highlighted higher rates of depression (25% in cases vs. 15% in controls), hypogonadism (8% in cases vs. 2% in controls), and LUTS (8% in cases vs. 2% in controls) among the ED group.

These findings indicate a significantly higher burden of comorbidities among individuals with ED. These conditions were confirmed through diagnostic codes and analyzed in relation to ED, demonstrating the comprehensive scope of health challenges faced by the cohort over an average EHR involvement spanning 26.4 years (IQR 18.3-30.4 years).

	BMI	**SBP**	**DBP**	**CVD**	**Diabetes**	**Hypogo-nadism**	**LUTS**	**Depre-ssion**	**Age**	**Hyper-tension**
BMI	1	0.15	0.2	0.06	0.23	0.09	0.01	0.08	-0.13	0.2
SBP	0.15	1	0.37	0.12	0.09	0.004	0.07	0.02	0.17	0.34

DBP	0.2	0.37	1	-0.26	-0.12	-0.04	-0.18	-0.01	-0.32	0.04
CVD	0.06	0.12	-0.26	1	0.19	0.04	0.28	0.07	0.42	0.38
Diabetes	0.23	0.09	-0.12	0.19	1	0.05	0.07	0.09	0.05	0.27
Hypogo-nadism	0.09	0	-0.04	0.04	0.05	1	0.06	0.11	-0.09	0.05
LUTS	0.01	0.07	-0.18	0.28	0.07	0.06	1	0.08	0.28	0.25
Depress-ion	0.08	0.02	-0.01	0.07	0.09	0.11	0.08	1	-0.05	0.12
Age	-0.13	0.17	-0.32	0.42	0.05	-0.09	0.28	-0.05	1	0.21
Hyper-tension	0.2	0.34	0.04	0.38	0.27	0.05	-0.02	0.12	0.21	1

Table 4. The pairwise correlation of variables for the analysis of collinearity

Table 4 demonstrates the analysis of pairwise correlation between covariates of interest. BMI, an indicator of diabetes, hypogonadism, LUTS, and depression, have a lower than 0.28 correlation with the rest of the covariates, while systolic blood pressure held an intermediate correlation with diastolic blood pressure (0.37), and with an indicator of hypertension (0.34). The CVD indicator had an even higher correlation with age (0.42) and hypertension (0.38). This correlation analysis revealed that there might exist a linear correlation of variables among systolic blood pressure, diastolic blood pressure, age, and hypertension indicator. The aforementioned variables could potentially cause confounding effects.

Comorbidity	Patient Type	# Disease	# No Disease
CVD	# Case	617	424
	# Control	819	975
Depression	# Case	258	783
	# Control	271	1523
Diabetes	# Case	302	739
	# Control	366	1428
Hyperten-sion	# Case	780	261
	# Control	1066	728
Hypogo-nadism	# Case	86	955
	# Control	39	1755
LUTS	# Case	641	400
	# Control	693	1101

Table 5 A summary of comorbidity-specific contingency tables

After coding the comorbidities, a series of contingency tables for the comorbidities were generated to form Table 5, detailing the prevalence rates of key comorbidities in both the control group and the case group diagnosed with ED. Analyzing the association between comorbidity

and ED by the Fisher's test is equivalent to univariable logistic regression and thus was merged with later analysis.

3.2. Parametric and Non-parametric Estimation

		OR(95% CI)	p<0.05
Patient Age		0.96 (0.96 ,0.97)	*
Race	**White**	Baseline	
	Asian	0.46(0.36 ,0.58)	*
	Black	1.31(0.82 ,2.12)	
	Hispanic	0.57(0.19 ,1.75)	
	American Indian	1.59(0.66 ,3.85)	
	Other	2.82(1.04 ,7.62)	*
	Unknown	1.25(0.84 ,1.85)	
Median BMI		1.00(1.00, 1.00)	*
Median DBP		0.99(0.99 ,1.00)	*
median SBP		1.01(1.00 ,1.01)	*
10 median SBP		1.09(1.05 ,1.13)	*
Cardiovascular Disease		1.73(1.48 ,2.02)	*

Diabetes	1.59(1.34 ,1.90)	*
Hypogonadism	4.02(2.73 ,5.92)	*
Depression	1.85(1.53 ,2.24)	*
Hypertension	2.04(1.72 ,2.41)	*
LUTS	4.02(2.73 ,5.92)	*

Table 6. The results of logistic regression on ED cohort, crude marginal univariable model

The univariable marginal logistic regression models were performed individually on all 11 covariates of interest, containing patient age, race, median patient BMI, median patient DBP, median patient SBP, an indicator of cardiovascular disease, an indicator of diabetes, an indicator of hypogonadism, an indicator of depression, an indicator of hypertension, and indicator of LUTS. The response variable is the indicator of the ED case (case=1, control=0).

Based on marginal univariable logistic regression of ED disease status on age, race, BMI, SBP, DBP, or indicators of comorbidities (CVD, diabetes, hypogonadism, depression, hypertension, LUTS), we estimated that the odds ratio (OR) for getting ED between Asian male and White male (baseline) groups is 0.46, with 95% robust Wald confidence interval (95% CI: 0.36 - 0.58). It indicated that Asian males in the eMERGE KPUW cohort have 54% lower odds of ED than white males. A Wald test using robust standard errors of the null hypothesis of equal ED odds in patient groups over all races returned p-values smaller than 0.05. We find this significant association at the 0.05 level and reject the null of no association between ED disease status and race. The odds ratio for getting ED between one unit of BMI larger than the reference BMI group and the reference BMI group is 1.00, with a 95% robust Wald confidence interval

(95% CI: 1.00- 1.00), indicating that males with one unit higher of BMI in the eMERGE KPUW cohort have 0.01% higher odds of ED than males of the reference group, with an association of significance at the 0.05 level.

Both SBP and DBP are included in the logistic regression analysis. The results show an intricate relationship between blood pressure measurements and ED. The odds ratio for ED with one unit increase in SBP is 1.01 (95% CI: 1.00 - 1.01), indicating that males with one unit higher SBP have 0.85% higher odds of ED than those with the reference SBP. Conversely, the odds ratio for ED with one unit increase in DBP is 0.99 (95% CI: 0.98 – 1.00), indicating that males with one unit higher DBP have 1.04% lower odds of ED than those with the reference DBP.

Other covariates of interest, including indicators of comorbidities (CVD, diabetes, hypogonadism, depression, hypertension, LUTS), are all associated with ED, significant at the 0.05 level, which confirmed the observations of comorbidity-related contingency tables in Table 5. The comorbidity-specific positive groups (indicator value =1) of patients with CVD, diabetes, hypogonadism, depression, hypertension, or LUTS, have higher odds of ED than the comorbidity-specific negative groups (indicator value =0), in different extents of odds ratios (cardiovascular disease OR 1.73, 95% CI: 1.48 - 2.02, diabetes OR 1.59, 95% CI: 1.34 - 1.90, hypogonadism OR 4.02, 95% CI: 2.73 - 5.92, depression OR 1.85, 95% CI: 1.53 - 2.24, hypertension OR 2.04, 95% CI: 1.72 - 2.41, LUTS OR 4.02, 95% CI: 2.73 - 5.92).

One exponentiated coefficient of interest for patient age was smaller than 1 (0.96, 95% CI: 0.96- 0.97). This situation indicates that males of the 70-year-old group in the eMERGE KPUW cohort have 4.2% lower odds of ED than the 69-year-old subgroup males.

Table 6 presents the logistic regression analysis of the study, focusing on the relationship between various clinical, demographic, and comorbidity indicators and the prevalence of ED

among the cohort. The analysis reveals statistically significant associations, indicating that certain factors are correlated with the likelihood of being diagnosed with ED. For instance, the presence of cardiovascular disease, diabetes, and hypertension are correlated with the odds of ED, as evidenced by their respective odds ratios, which are all above 1 and statistically significant.

Logistic Regression		**Adjusted Full Model**		
		OR(95% CI)	**p<0.05**	**CF**
Patient Age		0.89(0.88, 0.90)	*	
Race	**White**	Baseline		
	Asian	0.20(0.15, 0.28)	*	
	Black	1.02(0.59, 1.75)		
	Hispanic	0.18(0.05, 0.64)	*	
	American Indian	1.67(0.62, 4.50)		
	Other	2.35(0.72, 7.63)		
	Unknown	1.57(0.99, 2.48)		
Median BMI		0.97(0.95, 0.99)	*	
Median DBP		0.92(0.90, 0.93)	*	*
median SBP		1.05(1.04, 1.06)	*	
10 median SBP		1.62(1.44, 1.81)	*	

Cardiovascular Disease	2.07(1.66, 2.57)	*	
Diabetes	1.12(0.90, 1.41)		*
Hypogonadism	1.89(1.19, 3.01)	*	
Depression	1.22(0.98, 1.53)		*
Hypertension	1.64(1.30, 2.08)	*	
LUTS	3.18(2.60, 3.89)	*	

Table 7. The results of logistic regression on ED cohort, joint full model. CF=confounding

Table 7 builds upon the insights gleaned from Table 6 by employing a full model logistic regression analysis that simultaneously adjusts for multiple risk factors. This approach allows for a more nuanced understanding of how each variable interacts within the broader context of multiple risk factors. The full model contains all 11 covariates of interest, including patient age, race, median BMI, median DBP, median SBP, indicators of cardiovascular disease, diabetes, hypogonadism, depression, hypertension, and LUTS, with the response variable being the indicator of ED case (case=1, control=0).

Based on the joint full logistic regression of ED disease status on age, race, BMI, SBP, DBP, and indicators of comorbidities (CVD, diabetes, hypogonadism, depression, hypertension, LUTS), we estimated that the odds ratio for getting ED between Asian male and White male (baseline) groups, homogeneous in age, BMI, SBP, DBP, and indicators of comorbidities (CVD, diabetes, hypogonadism, depression, hypertension, LUTS), is 0.20, with 95% robust Wald

confidence interval (95% CI: 0.15 - 0.28). A Wald test using robust standard errors of the null hypothesis of equal ED odds in patient groups differing by one unit of predictor value, homogeneous in age, BMI, SBP, DBP, and indicators of comorbidities, against the general alternative, returned a p-value smaller than 0.05, so we find this association is significant at the 0.05 level and reject the null hypothesis of no association between ED case and race while adjusting for other covariate factors. The sample sizes for Hispanic, American Indian, or other race categories are too small for a robust analysis and interpretation is neglected.

The odds ratio for ED between older and younger individuals, holding other variables constant, is 0.89 (95% CI: 0.88 - 0.90). It indicates that males who are one year older have 11% lower odds of ED than those who are one year younger, with an association of significance at the 0.05 level. The odds ratio for ED between males with one unit higher BMI and the reference BMI group, holding other variables constant, is 0.97 (95% CI: 0.95-0.99), indicating that males with one unit higher BMI have 2.9% lower odds of ED than those in the reference BMI group, with an association of significance at the 0.05 level. The odds ratio for ED between males with one unit higher DBP and the reference DBP group, holding other variables constant, is 0.92 (95% CI: 0.90 - 0.93), indicating that males with one unit higher DBP have 8.3% lower odds of ED than those in the reference DBP group, with an association of significance at the 0.05 level. The odds ratio for ED between males with one unit higher SBP and the reference SBP group, holding other variables constant, is 1.05 (95% CI: 1.04 - 1.06), indicating that males with one unit higher SBP have 5% higher odds of ED than those in the reference SBP group, with an association of significance at the 0.05 level.

The odds ratio for ED between males with CVD and those without CVD, holding other variables constant, is 2.07 (95% CI: 1.66-2.57), indicating that males with CVD have 107% higher odds of ED than those without CVD, with an association of significance at the 0.05 level. The odds ratio for ED between males with diabetes and those without diabetes, holding other variables constant, is 1.12 (95% CI: 0.90-1.41), indicating no statistically significant association. The odds ratio for ED between males with hypogonadism and those without hypogonadism, holding other variables constant, is 1.89 (95% CI: 1.19 - 3.01), indicating that males with hypogonadism have 89% higher odds of ED than those without hypogonadism, with an association of significance at the 0.05 level. The odds ratio for ED between males with depression and those without depression, holding other variables constant, is 1.22 (95% CI: 0.98 - 1.53), indicating no statistically significant association. The odds ratio for ED between males with hypertension and those without hypertension, holding other variables constant, is 1.64 (95% CI: 1.30 - 2.08), indicating that males with hypertension have 64% higher odds of ED than those without hypertension, with an association of significance at the 0.05 level. The odds ratio for ED between males with LUTS and those without LUTS, holding other variables constant, is 3.18 (95% CI: 2.60-3.89), indicating that males with LUTS have 218% higher odds of ED than those without LUTS, with an association of significance at the 0.05 level.

The results in Table 7 demonstrate that when adjusting for other variables, certain factors such as cardiovascular disease and diabetes retain their significant association, with comorbidity-bearing patient groups having a higher odd of ED while comparing with the non-comorbidity patient group, suggesting robust links irrespective of other conditions. Conversely, the significance of some factors observed in Table 6, like hypertension, might diminish when controlled against a more comprehensive array of variables. This indicates that their initial

associations could be influenced by confounders that are not accounted for in the simpler models. The adjusted partial multiple logistic regression models (full model included all 11 covariates, and adjusted models exclude one covariate per model) excluding median DBP, indicator of diabetes, or indicator of depression changes the coefficients and/or significance of other covariates, indicating that they are the confounders of ED in our eMERGE KPUW cohort. This refined analysis emphasizes the complexity of ED's etiology. It highlights the necessity of considering a multifactorial approach in research and clinical assessment to identify and manage risk factors associated with ED.

Logistic Regression		**Adjusted Multiple Model**		
		OR(95% CI)	**p<0.05**	**CF**
Patient Age		0.90(0.89, 0.91)	*	
Race	**White**	Baseline		
	Asian	0.21(0.15, 0.29)	*	
	Black	0.92(0.54, 1.58)		
	Hispanic	0.18(0.053, 0.64)	*	
	American Indian	1.78(0.67, 4.72)		
	Other	2.16(0.67, 6.95)		
	Unknown	1.50(0.96, 2.36)		
Median BMI		0.96(0.94, 0.98)	*	

Median DBP	NA		
median SBP	1.02(1.01, 1.03)	*	
10 median SBP	1.21(1.10, 1.33)	*	
Cardiovascular Disease	2.40(1.94, 2.98)	*	
Diabetes	1.33(1.07, 1.66)	*	
Hypogonadism	2.25(1.44, 3.53)	*	
Depression	1.25(1, 1.56)	edgy, 0.05	
Hypertension	1.54(1.22, 1.93)	*	
LUTS	3.39(2.78, 4.13)	*	

Table 8. The multi-variable logistic regression on ED cohort, removing diastolic blood pressure. CF=confounding

Table 8 refines the logistic regression model discussed in Table 7 by excluding one specific variable, the median of patient DBP, to assess its impact on the model's outcomes for the progression of ED. This deliberate exclusion helps isolate the effects of the remaining variables, providing a clearer picture of their individual and collective influence on ED's onset. At the same time, it only serves as a reference for further studies. The comparison of Table 8 with Table 7, where the full model includes all variables, allows for an evaluation of how significant the excluded variable is in the context of ED. The adjusted model contains 10 covariates of interest, including patient age, race, median of patient BMI, median of patient SBP, indicator of cardiovascular disease, indicator of diabetes, indicator of hypogonadism, indicator of

depression, indicator of hypertension, and indicator of LUTS, with the response variable being the indicator of ED case (case=1, control=0).

Based on the adjusted multiple logistic regression of ED disease status on age, race, BMI, SBP, and indicators of comorbidities (CVD, diabetes, hypogonadism, depression, hypertension, LUTS), we estimated that the odds ratio for getting ED between Asian male and White male groups, homogeneous in age, BMI, SBP, and indicators of comorbidities (CVD, diabetes, hypogonadism, depression, hypertension, LUTS), is 0.21, with a 95% robust Wald confidence interval (0.15 - 0.29). A Wald test using robust standard errors of the null hypothesis of equal ED odds in patient groups differing by one unit of covariate, homogeneous in all other variables, against the general alternative returned a p-value smaller than 0.05. Thus, we find this association significant at the 0.05 level and reject the null hypothesis of no association between ED and race after adjusting for the other predictors.

The odds ratio for getting ED between older and younger individuals, holding other variables constant, is 0.90 (95% CI: 0.89 - 0.91), indicating that males who are one year older have 9.8% lower odds of ED than those who are one year younger, with an association significant at the 0.05 level. The odds ratio for getting ED between males with one unit higher BMI and the reference BMI group, holding other variables constant, is 0.96 (95% CI: 0.94 - 0.98), indicating that males with one unit higher BMI have 3.9% lower odds of ED than those in the reference BMI group, with an association significant at the 0.05 level. The odds ratio for getting ED between males with one unit higher systolic blood pressure (SBP) and the reference SBP group, holding other variables constant, is 1.05 (95% CI: 1.04 - 1.06), indicating that males

with one unit higher SBP have 4.9% higher odds of ED than those in the reference SBP group, with an association significant at the 0.05 level.

The odds ratio for getting ED between males with CVD and those without CVD, holding other variables constant, is 2.40 (95% CI: 1.94 - 2.98), indicating that males with CVD have 140% higher odds of ED than those without CVD, with an association significant at the 0.05 level. The odds ratio for getting ED between males with diabetes and those without diabetes, holding other variables constant, is 1.33 (95% CI: 1.07 - 1.66), indicating that males with diabetes have 33% higher odds of ED than those without diabetes, with an association significant at the 0.05 level. The odds ratio for getting ED between males with hypogonadism and those without hypogonadism, holding other variables constant, is 2.25 (95% CI: 1.44 - 3.53), indicating that males with hypogonadism have 125% higher odds of ED than those without hypogonadism, with an association significant at the 0.05 level. The odds ratio for getting ED between males with depression and those without depression, holding other variables constant, is 1.25 (95% CI: 1 - 1.56), indicating that males with depression have 25% higher odds of ED than those without depression, with an association significant at the 0.05 level. The odds ratio for getting ED between males with hypertension and those without hypertension, holding other variables constant, is 1.54 (95% CI: 1.22 - 1.93), indicating that males with hypertension have 54% higher odds of ED than those without hypertension, with an association significant at the 0.05 level. The odds ratio for getting ED between males with lower urinary tract symptoms (LUTS) and those without LUTS, holding other variables constant, is 3.39 (95% CI: 2.78 - 4.13), indicating that males with LUTS have 239% higher odds of ED than those without LUTS, with an association significant at the 0.05 level.

The effect size and the significance of the levels of race do not change much in the trimmed model compared to the full model. However, we see that the effect sizes of comorbidity covariates (CVD, diabetes, hypogonadism, depression, hypertension, and LUTS) all increased to different extents. Diabetes is associated with ED in the trimmed model at an alpha level of 0.05 (it was not associated with ED at the 0.05 level in the full model), and the p-value of depression decreased to close to 0.05. The explanatory power of DBP is distributed to other variables, confirming its confounding effect.

Conversely, the significance of some factors observed in Table 6, like hypertension, might diminish when controlled for a wider array of variables, indicating that their initial associations could be influenced by confounders not accounted for in the simpler models. The adjusted partial multiple logistic regression models (full model included all 11 covariates, and adjusted models exclude one covariate per model) excluding median DBP, indicator of diabetes, or indicator of depression changes the coefficients and/or significance of other covariates, indicating that they are the confounders of ED in our eMERGE KPUW cohort.

Cox Regression		**Crude**	
		HR(95% CI)	**p<0.05**
Patient Age		0.97(0.96, 0.97)	*
Race	**White**	Baseline	
	Asian	0.57(0.46, 0.71)	*
	Black	1.50(1.05, 2.13)	*

	Hispanic	0.83(0.31, 2.21)	
	American Indian	1.06(0.57, 1.97)	
	Other	1.49(0.82, 2.70)	
	Unknown	1.31(0.97, 1.77)	
Median BMI		1.02(1.01, 1.03)	*
Median DBP		0.98(0.97, 0.99)	*
median SBP		1.01(1.01, 1.02)	*
Cardiovascular Disease		1.65(1.45, 1.86)	*
Diabetes		1.44(1.26, 1.65)	*
Hypogonadism		2.52(2.02, 3.14)	*
Depression		1.45(1.26, 1.67)	*
Hypertension		1.64(1.42, 1.88)	*
LUTS		2.52(2.02, 3.14)	*

Table 9. The results of Cox regression on ED cohort, crude marginal univariable model

Table 9 utilized a Cox regression survival analysis model to assess the impact of various risk factors on the time to onset of ED. Statistical outputs such as hazard ratios, confidence intervals, and p-values provided quantifiable measures of these effects, underscoring the critical temporal dynamics involved in the progression of ED. This approach demonstrated how the

duration before ED manifestation can vary significantly based on individual health profiles and risk factors.

The univariable model covered all the variables of interest, including patient age, race, median patient BMI, median patient DBP, median patient SBP, indicator of cardiovascular disease, indicator of diabetes, indicator of hypogonadism, indicator of depression, indicator of hypertension, and indicator of LUTS, with the response variable being the time-to-event variable.

Findings indicated significant independent associations between all 11 covariates and the risk of ED. All 11 covariates are significantly associated with the hazard of ED at the alpha level of 0.05. The exponentiated slope of the crude marginal univariable Cox PH model is the hazard ratio (HR) of ED risk between two patient groups differing by one unit of covariate. Based on Table 9, it is estimated that the hazard ratio for getting ED between Asian male and White male groups is 0.57, with a 95% confidence interval (0.46, 0.71), indicating that Asian males have a 43% lower hazard of having ED than white males in the eMERGE cohort 3. Conversely, Black males have a hazard ratio of 1.50 (95% CI: 1.05 - 2.13), suggesting they have a 50% higher hazard of ED than white males.

The effects of BMI (HR 1.02, 95% CI 1.01 - 1.03), DBP (HR 0.98, 95% CI 0.97 - 0.99), SBP (HR 1.01, 95% CI 1.01 - 1.02), and age (HR 0.97, 95% CI 0.96 - 0.97) are significant. Yet, the effect sizes are close to 1, indicating a relatively small difference in ED hazards between groups differing by one unit. The effects of CVD (HR 1.65, 95% CI 1.45 - 1.86), Diabetes (HR 1.44, 95% CI 1.26 - 1.65), Hypogonadism (HR 2.52, 95% CI 2.02 - 3.14), Depression (HR 1.45, 95% CI 1.26 - 1.67), Hypertension (HR 1.64, 95% CI 1.42 - 1.88), and LUTS (HR 2.52, 95% CI 2.02 - 3.14) are also significant at the level of 0.05, with a much larger hazard of ED in disease

groups than in non-disease groups (CVD 64.5% higher ED hazard than non-CVD, Diabetes 44.4% higher ED hazard than non-Diabetes, Hypogonadism 65.2% higher ED hazard than non-Hypogonadism, Depression 44.9% higher ED hazard than non-Depression, Hypertension 63.6% higher ED hazard than non-Hypertension, and LUTS 151.8% higher ED hazard than non-LUTS).

Cox Regression		**Adjusted Full Model**		
		HR(95% CI)	**p<0.05**	**CF**
Patient Age		0.92(0.92, 0.93)	*	
Race	**White**	Baseline		
	Asian	0.44(0.35, 0.56)	*	
	Black	1.44(1.01, 2.05)	*	
	Hispanic	0.45(0.17, 1.22)		
	American Indian	1.08(0.58, 2.02)		
	Other	1.22(0.67, 2.22)		
	Unknown	1.30(0.97, 1.76)		
Median BMI		0.97(0.96, 0.99)	*	
Median DBP		0.94(0.93, 0.95)	*	
median SBP		1.04(1.03, 1.04)	*	

Cardiovascular Disease	1.67(1.44, 1.93)	*	
Diabetes	1.10(0.94, 1.28)		*
Hypogonadism	1.39(1.11, 1.75)	*	
Depression	1.09(0.94, 1.26)		*
Hypertension	1.40(1.18, 1.67)	*	
LUTS	2.11(1.84, 2.42)	*	

Table 10. The results of Cox regression on ED cohort, joint full model. CF=confounding

Table 10 extends the analysis presented in Table 9 by refining the Cox regression survival model through additional adjustments for confounding factors, thereby enhancing the robustness of the findings regarding the time to onset of ED. This table specifically assesses how combinations of factors like age, cardiovascular disease, diabetes, and lifestyle choices simultaneously impact ED progression. The adjusted model in Table 10 provides more nuanced insights into the effects of comorbid conditions on the progression of ED when other factors are included. The full model contains all 11 covariates of interest, including patient age, race, median of patient BMI, median of patient DBP, median of patient SBP, indicator of cardiovascular disease, indicator of diabetes, indicator of hypogonadism, indicator of depression, indicator of hypertension, and indicator of LUTS. The response variable is the time-to-event variable.

Based on joint full Cox PH regression of ED disease status on age, race, BMI, SBP, DBP, and indicators of comorbidities (CVD, diabetes, hypogonadism, depression, hypertension,

LUTS), we estimated that the hazard ratio for getting ED between Asian male and white male (baseline) groups, homogenous in all other variables, is 0.44, with a 95% robust confidence interval (95% CI: 0.35-0.56), indicating a 66% lower hazard of ED. The null hypothesis is that there is no association between the covariate and the hazard of the event of ED controlling for all other variables, stating equal ED hazards in patient groups differing by one unit of covariate value, homogenous in age, BMI, SBP, DBP, and indicators of comorbidities, against the general alternative returned a p-value smaller than 0.05. We found this association is significant at the 0.05 level and rejected the null hypothesis of no association between ED case and race while adjusting for all other covariate factors. Black males have a hazard ratio of 1.44(95% CI: 1.01-2.05), indicating they have a 44% higher hazard of ED compared to white males, holding other variables constant.

Age (HR 0.92, 95% CI 0.92, 0.93), BMI (HR 0.97, 95% CI 0.96 - 0.99), DBP (HR 0.94, 95% CI 0.93 - 0.95), and SBP (HR 1.04, 95% CI 1.03 - 1.04) are associated with risk of ED holding other variables constant at a significance level of 0.05, with HRs within the interval of (0.92, 1.04), indicating a small effect size on hazard estimation. Cardiovascular disease (HR 1.67, 95% CI 1.44- 1.93), hypogonadism (HR 1.39, 95% CI 1.11 - 1.75), hypertension (HR 1.40, 95% CI 1.18 - 1.67), and LUTS (HR 2.11, 95% CI 1.84 - 2.42) present larger effect sizes of hazard in ED, indicating a much higher hazard of ED in the disease group than the non-disease group, with significance at an alpha level of 0.05.

Although diabetes (HR 1.10, 95% CI 0.94 - 1.28) and depression (HR 1.09, 95% CI 0.94 - 1.26) have exponentiated coefficients larger than 1, indicating a slightly higher hazard of ED in the disease group than the non-disease group, the associations are not significant at an alpha level of 0.05. Meanwhile, the inclusion of DBP, diabetes, and depression variables changes the

significance and/or effect size of other variables in comparison to adjusted multivariable Cox PH regression models without those variables in a stepwise manner, indicating that DBP, diabetes, and depression are confounders. However, diabetes and depression are important comorbidities of interest for clinicians. Thus, their presence is recommended to remain in future models, yet stratification is recommended for later modeling.

Cox Regression		**Adjusted Multiple Model**		
		HR(95% CI)	**p<0.05**	**CF**
Patient Age		0.93(0.93, 0.94)	*	
Race	**White**	Baseline		
	Asian	0.44(0.34, 0.56)	*	
	Black	1.19(0.84, 1.7)		
	Hispanic	0.48(0.18, 1.27)		
	American Indian	1.24(0.66, 2.32)		
	Other	1.11(0.61, 2.02)		
	Unknown	1.30(0.96, 1.75)		
Median BMI		0.96(0.95, 0.97)	*	
median SBP		1.01(1.00, 1.02)	*	
Cardiovascular Disease		1.90(1.65, 2.19)	*	

Diabetes	1.32(1.14, 1.52)	*	
Hypogonadism	1.54(1.23, 1.94)	*	
Depression	1.12(0.97, 1.29)		*
Hypertension	1.31(1.10, 1.55)	*	
LUTS	2.20(1.92, 2.53)	*	

Table 11 . The multi-variable Cox regression on ED cohort, removing diastolic blood pressure. CF=confounding

Table 11 further refines the Cox regression survival model used in Table 10 by potentially excluding certain variables, such as diastolic blood pressure, to test their impact on the model's accuracy and the strength of associations. This adjustment allows for a clearer understanding of which factors are most predictive of the time to onset of ED, while it only serves as a reference for further studies. Comparatively, while Table 10 provides a comprehensive view incorporating all known confounders, Table 11 might show stronger or weakened associations for remaining variables, indicating the direct impact of the excluded factors on ED progression. This side-by-side analysis between the tables enhances the understanding of each variable's role and offers a more precise tool for estimating ED's hazard in clinical assessments.

The adjusted model contains 10 covariates of interest, including patient age, race, median of patient BMI, median of patient SBP, indicator of cardiovascular disease, indicator of diabetes, indicator of hypogonadism, indicator of depression, indicator of hypertension, and indicator of LUTS. The response variable is the time-to-event variable.

Based on the adjusted multiple Cox PH regression of ED risk on age, race, BMI, SBP, and indicators of comorbidities (CVD, diabetes, hypogonadism, depression, hypertension, LUTS), we estimated that the hazard ratio for getting ED between Asian male and White male groups, homogenous in age, BMI, SBP, DBP, and indicators of comorbidities (CVD, diabetes, hypogonadism, depression, hypertension, LUTS), is 0.44, with a 95% confidence interval (0.35, 0.56), indicating a 56.12% lower hazard of ED. The null hypothesis is that there is no association between the covariate and the hazard of the ED event, controlling for all other variables. Stating equal ED hazards in patient groups differing by one unit of covariate value, homogenous in age, BMI, SBP, DBP, and indicators of comorbidities, against the general alternative returned a p-value smaller than 0.05. So, we find this association is significant at the 0.05 level and reject the null hypothesis of no association between ED risk and race while adjusting for all other covariate factors. Black males have an HR of 1.19 (95% CI: 0.84-1.70), indicating no statistically significant difference in the hazard of ED compared to white males, holding other variables constant.

The effects of other variables are as follows: Age (HR 0.93, 95% CI: 0.93-0.94), BMI (HR 0.96, 95% CI: 0.95-0.97), and SBP (HR 1.01, 95% CI: 1.00-1.02) are associated with the risk of ED, holding other variables constant at a significance level of 0.05, indicating a small effect size on hazard estimation. Cardiovascular disease (HR 1.90, 95% CI: 1.65-2.19), diabetes (HR 1.32, 95% CI: 1.14-1.52), hypogonadism (HR 1.54, 95% CI: 1.23-1.94), hypertension (HR 1.31, 95% CI: 1.10-1.55), and LUTS (HR 2.20, 95% CI: 1.92-2.53) present larger effect sizes of hazard in ED, indicating a much higher hazard of ED in the disease group than the non-disease group, with significance at an alpha level of 0.05. However, depression (HR 1.12, 95% CI: 0.97-

1.29) is not significantly associated with ED at the 0.05 level, even though its hazard ratio is slightly above 1.

The effect size and the significance of the levels of race, age, BMI, and SBP do not change much in the trimmed model in comparison to the full model. However, we see that the effect sizes of comorbidity covariates (CVD, diabetes, hypogonadism, depression, hypertension, and LUTS) all increased to different extents. Diabetes is associated with ED in the trimmed model when alpha is 0.05 (it was not associated with ED at the 0.05 level in the full model). The explanatory power of DBP is distributed to other variables, confirming its confounding effect.

3.3. Patients Trajectory over Time

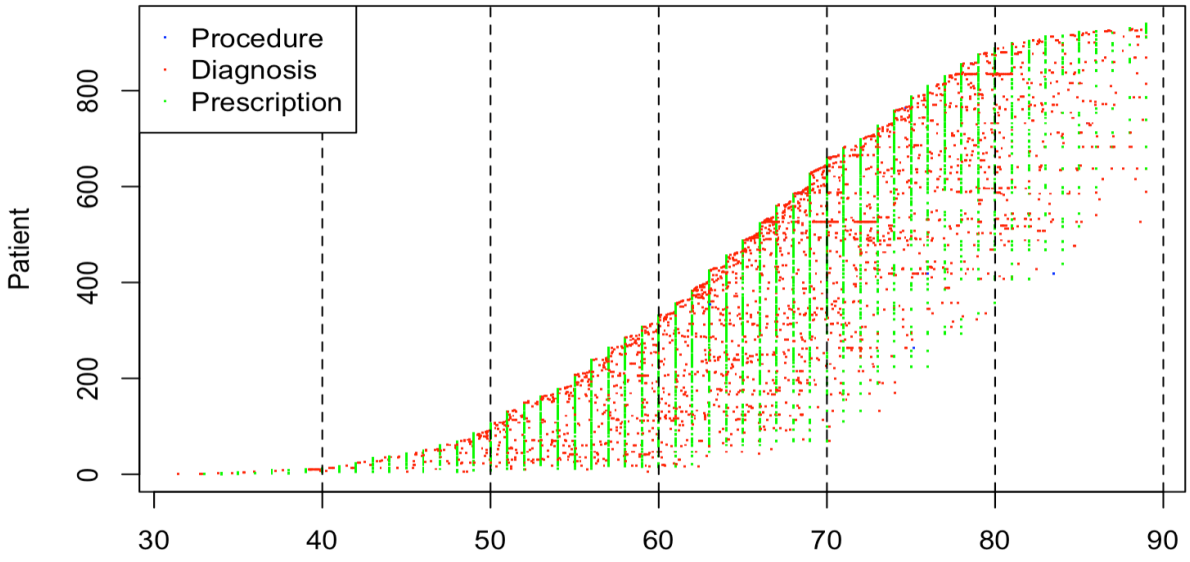

Figure 3. Distribution of frequency and type of encounters ordered by age at disease onset. Each row represents a patient and each dot was an encounter for that patient.

All the case subjects were included with a median of 26.4 years of care (IQR 18.3-30.4 years). The mean age at ED diagnosis was 65 years (± 10.9 years), and subjects had an average of 3 encounters (IQR 1-4) to address ED. The median duration of clinical care for ED was 4 years (IQR 0-9 years).

The composition of ED-related encounters included an average of 8 prescriptions (Rx) visits (IQR 2-22), 3 procedure (Px) visits (IQR 1-4), and 3 diagnosis-only (Dx) visits (IQR 1-4). The indication for ED procedures was much less frequent, and most EHR encounters were diagnosis and prescription visits. However, given that the severity of ED could be high for patients undergoing ED-related procedures, the validity of procedure visits is of significant importance.

The visualization of the EHR encounters indicates that the granularity of Rx visits was lower than that of Dx visits. Rx visits were timed in years, while Dx visits were timed in days. The EHR footprint of ED is heterogeneous, with initial diagnosis occurring across the adult lifespan and significant variation evident in the number of disease-related encounters. The average frequency and duration of ED care were limited, with an average of 3 encounters per subject and a median duration of 4 years from the initial diagnosis to the last ED encounter. Only a minority of patients had 5 or more medication-related encounters, potentially representing long-term prescription renewals.

Further investigation in EHR is needed to understand how healthcare utilization differs among specific subgroups of patients with ED.

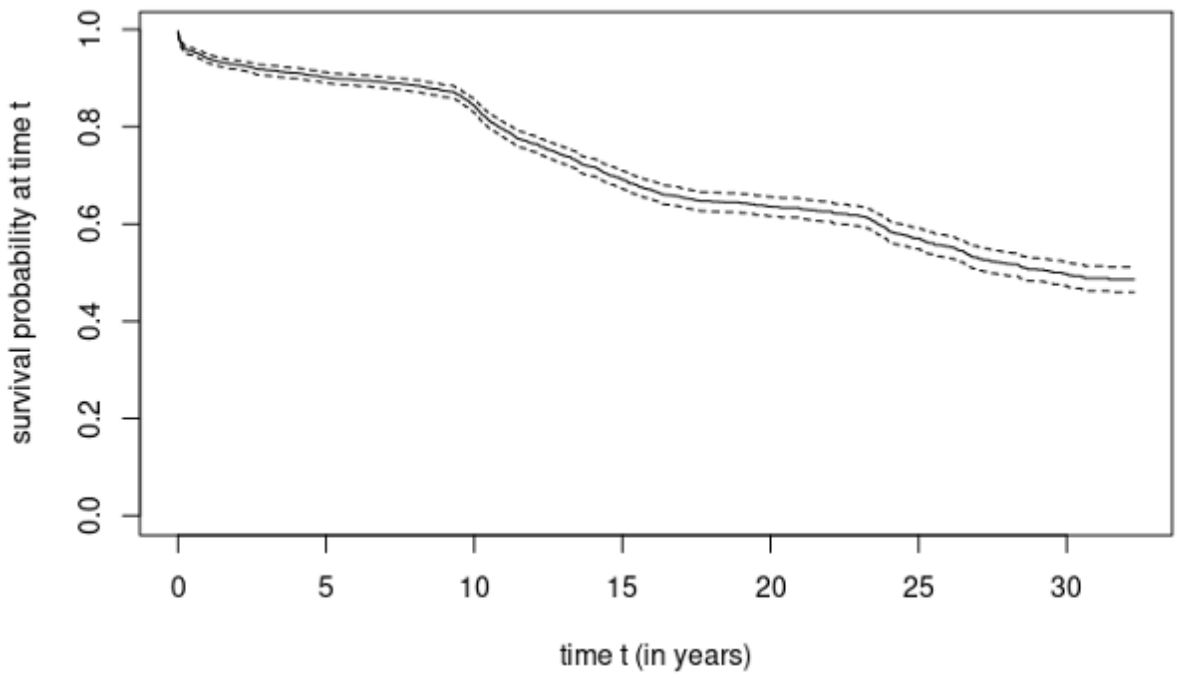

Figure 4. Kaplan-Meier estimation for ED cases entering eMERGE cohort 3 study.

We also used the Kaplan-Meier survival plot to estimate the probability that no event of ED or exit of study will occur by the time of EHR tracking among the cases, with a 95% confidence interval reported. The Kaplan-Meier curve is a step function that decreases at each event occurrence. Each step down indicates an event (ED diagnosis) happening to one or more individuals. The x-axis represents the time in years since cohort entry, while the y-axis indicates the probability of not developing ED (survival probability). The median survival time is when the survival probability reaches 0.5 (or 50%), while half of the subjects have experienced the event (ED diagnosis/ study exit), and the other half have not. Tick marks along the curve denote censored data points, indicating subjects who were either lost to follow-up or remained free of ED by the end of the study period.

It shows an initial survival probability of 1 at the start of the study. Here, we estimated that by the year 10 after entry of EHR, the estimated probability that no ED onset or exit of EHR will occur is 84% for cases, indicating that 16% of the subjects were diagnosed with ED within the 10 years after entering EHR. We can see that the years 10 and 23 after EHR entry have a steep curving down, indicating that a higher event rate of ED happens at around year 10 and year 23. Year 10 to year 15 indicated a steep decline, i.e., a rapid increase in ED encounters. Between year 15 and year 23, the curve plateaus, suggesting a lower rate of ED diagnoses or stable survival among the remaining subjects.

Chapter 4. Discussion

4.1. Observations and Insights from the Outputs

4.1.1. Correlation and Variable Selection

The correlation matrix reveals pairs of variables correlated with each other. When two highly correlated predictor variables (e.g., with correlation $r > 0.7$) contain much-shared information, the possibility of collinearity challenges the model. It potentially results in unstable estimates of the regression coefficient(s) while complicating the efforts to evaluate the individual impact of each predictor variable. The strategies to mitigate this problem include eliminating or merging collinear variables. Meanwhile, regularization techniques such as Ridge regression (which applies an L2 penalty equal to the square of the magnitudes of coefficients) and Lasso regression (which penalizes the absolute magnitudes of coefficients with variable selection) may support more stable and consistent models while allowing for more accurate estimation.

In our results, moderate correlations were observed between several variables, suggesting potential collinearity issues, particularly involving SBP, DBP, CVD, hypertension, and age. There was a moderate to high level of correlation between SBP and hypertension (SBP-hypertension $r = 0.34$) or DBP (SBP-DBP $r = 0.37$). Additionally, correlations between CVD and age (CVD-age, $r = 0.42$) or hypertension (CVD-hypertension, $r = 0.38$) were also observed. For future studies, these correlations might suggest the need for variable selection or adjustment for correlated variables when modeling. In our current study, we still included all the variables to preserve the explanation of variance.

4.1.2. Insight from Regression Outputs

Both logistic regression and Cox proportional hazards models are powerful tools for analyzing relationships between predictor variables and outcomes. Logistic regression models the probability of a binary outcome (ED presence/absence) as a function of predictor variables, providing odds ratios that indicate how the odds of ED change by comparing groups with a one-unit difference in the predictor variable, holding all other variables constant. For instance, in our analysis, the adjusted OR for CVD from the adjusted full multivariable logistic regression model was 2.07, meaning that patients with CVD had more than twice the odds of having ED compared to those without CVD, controlling for other variables. However, it does not account for the timing of the event or the feasibility of analyzing censored data.

On the other hand, the Cox proportional hazards model analyzes the time until the occurrence of an event (ED onset/exit of study), assessing the impact of predictor variables on specific times. It provides hazard ratios (HR) that indicate how the event's hazard (or risk) changes by comparing groups with a one-unit difference in the predictor variable, holding all other variables constant. For instance, the adjusted HR for CVD from the adjusted Cox model was 1.67, indicating that patients with CVD had a 67% higher risk of developing than those without CVD while homogeneous on other covariates. The assumption of proportional hazards may not hold, and Cox regression outputs are more complex to interpret than those of logistic regression.

Both logistic regression and Cox regression models indicate that the one-year older age group showed a lower risk of ED than the reference patient group, with a significant association

at alpha level 0.05. This could be explained by survivor bias of the cohort where healthier older adults remained in the study, highlighting that early monitoring is suggested in relatively younger patients. Logistic regression showed that patient groups older by one year have lower odds of ED (OR = 0.89, Adjusted Model) than the one-year younger patient groups, while homogeneous in other variables. The Cox model similarly indicated that patient groups older by one year had a reduced hazard of ED onset over time (HR = 0.92, Adjusted Model) than the one-year younger patient groups, suggesting that relatively younger patients might experience ED encounters in EHR earlier while holding other variables constant. Meanwhile, the conservative effect size of BMI, which is close to one, serves as a reference in our study, suggesting the inclusion of BMI in future regression models to investigate its mechanism.

Univariable logistic regression and Cox regression models both indicated an association between each of the 11 covariates (age, race, median patient BMI, median patient DBP, median patient SBP, indicator of cardiovascular disease, indicator of diabetes, indicator of hypogonadism, indicator of depression, indicator of hypertension, and indicator of LUTS), and risk of ED. In full multivariable regression models containing all 11 covariates, strong associations of CVD, hypertension, LUTS, and hypogonadism with both odds and hazard of ED emphasize the importance of managing these conditions to prevent ED, supporting integrated care approaches that address multiple risk factors simultaneously. In the multivariable analysis, DBP, diabetes, and depression showed confounding effects. Subsequently, through the integration of aspects from logistic regression and Cox regression models, a multifactorial understanding of ED risk and progression translates into clinical applicability with respect to tailored interventions and preventive strategies.

There lied possibility that comorbidities of ED such as diabetes acts as a confounder or a mediator in the relationship between other comorbidities like cardiovascular disease, hypertension, and depression. There might exist interconnection or directional causal/association pathways or graphs among comorbidities, and this analysis require testing for confounding and interaction in regression models. In our analysis, we tested the confounding effect of each comorbidity. For instance, diabetes was initially included as one of the 11 covariates in the full model. To determine its role as a confounder, we compared the full model (with diabetes included) to a reduced model (where diabetes was excluded). We observed the changes in the effect sizes and significance levels of the other 10 covariates when diabetes was removed. If excluding diabetes led to a shift in these parameters, it indicated that diabetes had a confounding effect, impacting the relationships between the other covariates and ED. The fact that removing diabetes altered the effect sizes or significance of some covariates suggests that diabetes does indeed interfere with or confound the relationships between those covariates and the outcome. In such cases, the standard approach is either to adjust for diabetes in the analysis or explore alternative models that might handle its confounding effect differently. As for using interaction terms in regression models for detecting mediators, it's a valid approach to explore the combined effect of variables, such as diabetes and hypertension, on ED. However, creating interaction terms would require a larger sample size to ensure enough statistical power for the analysis, especially given the stratification that would result from combining multiple variables. With our dataset of approximately 1,000 cases and 11 covariates, adding interaction terms could have led to a significant reduction in the available sample size for each stratified group, making it challenging to conduct a robust analysis. Thus, while interaction terms could offer deeper insights, the sample size constraints in this study limited their feasibility.

Another important aspect to consider is whether the comorbidities were analyzed on a time-to-event scale. In our study, we coded each patient's comorbidity status as a binary outcome—indicating whether the patient had a particular comorbidity at any point during the observation period. We did not specifically track the timing of the comorbidity in relation to the onset of ED. However, incorporating a time-to-event analysis could indeed add more precision to our findings. For instance, instead of coding comorbidities across the entire observation window, we could record the occurrence of each comorbidity in relation to the closest onset of ED. This approach could provide a more accurate understanding of how the timing of comorbidities influences the development of ED. However, implementing such an analysis would require high-quality data and a sufficiently large sample size to ensure robust results.

4.2. Limitations and Generalizability

4.2.1. Limitation of Data Properties

The retrospective nature of our study, while valuable for its broad scope, comes with inherent limitations. These include potential biases in data recording and challenges related to handling sparse, left-truncated, and right-censored data. The secondary use of EHR data for clinical research has shown data quality issues, including incompleteness, inconsistency, and inaccuracy, especially in the vital measurement traces (e.g., SBP, DBP, BMI). Biased results can arise from incorrect data entry, incomplete documentation, and differences in practice, all of which can lead to inaccurate information, skewing findings and reducing research reliability (Kitchen et al., 2022; Botsis et al., 2010). Incomplete data may result from patients visiting

multiple providers, missing follow-up appointments, and providers differing protocols for data entry (Blaisure & Ceusters, 2017; Zhao & Tsubota, 2023). Discrepancies in data recording processes among institutions and electronic health record (EHR) systems can result in inconsistencies when integrating multiple source information (Nobles et al., 2015). A lack of standardization in coding procedures between providers and insurers introduces bias to research results that utilize diagnostic and procedural codes, which depend on reliable coding practices (Wells et al., 2013; Kharrazi et al., 2014). Challenges are introduced in collecting data from multiple sources, infrastructures, and systems or using different formats by coding systems with software platforms for comprehensive analysis (Sandhu et al., 2012). One example is fragmented mental health data, often reflecting insurance claims that contain only partial information about a person's mental health due to privacy restrictions and our compartmentalization of medical care between the general (where physical issues were managed) and the psychological/psychiatric domain.

Moreover, many insurance claims might not capture the detailed clinical information required for a thorough understanding of ED and its comorbidities. Essential factors like lab results, socioeconomic status, lifestyle aspects, or even complete medical history could be lacking, causing underreporting of hormonal and psychological profiles for accurate diagnosis and treatment (Madden et al., 2016).

Insurance data provide more in-depth billing and administrative information but less about medication regimens, treatment completion, or symptom severity (Wells et al., 2013; Madden et al., 2016). When data collected for reimbursement coding does not capture the complete clinical complexity of ED, then the patient information may be biased by excluding records that lack a billable claim (Madden et al., 2016).

4.2.2. Limitation of Data Missingness

One challenge in using retrospective EHR to study ED is data missingness. Among the cohort of 2,835 patients in the study, 462 patients had one or fewer vital measure readings, with the reasons for missingness unknown and untraceable. We hypothesized that the original handling of EHR data input caused this missingness. For example, a nurse station might occasionally miss recording patients' blood pressure due to distraction or error. This scenario is classified as Missing Completely at Random (MCAR) (Little & Rubin, 2002). Similarly, if a software glitch randomly causes data loss independent of patient characteristics, it is also considered MCAR. However, if errors correlate with specific shifts, patient types, or conditions, the missingness may be classified as Missing at Random (MAR) or Missing Not at Random (MNAR) (Zhang et al., 2021). Given these scenarios, MCAR is the appropriate categorization.

For patients who were out-of-network and not followed up for a period, the missing data associated with this period could be classified as MNAR. It is because the missingness is related to the reason the data is missing; the patient's absence from the network likely affects the type and amount of data collected. EHR data may lack comprehensive information on ED diagnoses and treatments because patients receive care outside the recorded system. This incomplete capture of their health status can result in an underestimation of ED prevalence and skew the association between ED and comorbidities like cardiovascular disease and diabetes. Such missingness complicates analysis since it systematically relates to the patient's circumstances, such as moving away, financial status, or satisfaction with care.

To address the missing data issue, it is recommended that the workflow for the corresponding period of EHR be traced, considering that the missingness occurred over different years and decades, possibly due to variations in nurse training and software updates. However, extracting or compensating for these patterns is challenging due to the retrospective nature of EHRs and the extensive tracking period (over 20 years). The missingness could be due to the aforementioned limitations of secondary EHR and insurance data. Haneuse and Daniels (2016) argued that modeling the mechanism for how an observation of EHR data provenance is more likely to lead to correct specification and valid analysis than specifying why data are missing (Haneuse et al., 2016). However, the effect of missingness is still complex to capture or troubleshoot, let alone quantify for patterns. In addition, the documentation of ED symptoms may vary from one healthcare provider to another, potentially introducing the misclassification or bias of the ED cases.

The vital measurement dataset presented massive data missingness, especially for BMI, SBP, and DBP readings. We replaced the missing BMI values with the median and used PMM imputation for SBP and DBP values that were not recorded. Although the median BMI helps to address missing data, it does introduce some bias as it may not fully capture the precise health status of patients at the time of ED onset. Nevertheless, this approach provides a helpful way of handling missingness, especially when a patient is treated elsewhere outside the primary insurance system. We used median BMI as a surrogate when early ED traces are available but no corresponding BMI readings at timepoint. Using the median BMI also ensures consistency in the model analysis.

Meanwhile, using PMM to impute missing blood pressure readings may also lead to some regression-based bias, especially when the proportion of missing data is substantial. The extent and direction of these biases have yet to be quantified in our study, and further validation is necessary to determine whether these methods might lead to overestimation or underestimation of the impact of parameters like SBP and DBP. Currently, SBP and DBP are being used as proxies for the overall health of the cardiovascular system, but the actual effect sizes of these variables in our regression models will require further validation to confirm their accuracy and reliability.

4.2.3. Generalizability

The Cox PH regression model is a commonly used model for time-to-event data. This approach of modeling could be generalized for other databases as well. However, including too many covariates results in overfitting of the model (Vinzamuri & Reddy, 2013). Regularizations and modifications to the simple Cox PH regression model can be applied to mitigate this shortcoming, allowing it for better generalization. In 2013, Vinzamuri and Reddy proposed a Kernel Elastic Net Regularized Cox Regression (KEN-COX) with a novel kernel elastic net penalty term as a regularization factor. They assessed this model versus the Graph-based Octagonal Shrinkage and Clustering Algorithm for Regression (OSCAR) regularized Cox regression method, calculated the corresponding mean squared errors, and compared with LASSO-regularized Cox regression and classical Cox PH models on five EHR datasets and five synthetic datasets with sample size larger than 8000 (Vinzamuri & Reddy, 2013). In 2022, Wen et al. compared the performances of several advanced models, including DeepSurv, a deep feed-

forward neural network (Katzman et al., 2018), Cox-CC, a case-control approximation for Cox loss function estimation (Kvamme et al., 2019), random survival forest, an extension of random forest for censored survival data (Ishwaran et al., 2008), DeepHit, a deep neural network learning the distribution of survival times (Lee et al., 2018), and Multi-Task Logistic Regression, which combines a series of logistic regression models (Yu et al., 2011). They utilized EHR data extracted from the University of Texas Medical Branch at Galveston (UTMB) for a cohort of 805 COVID-19 patients with 59 features, followed between March 2020 and November 2020, to predict the length of hospital stay (Wen et al., 2022). Interestingly, when comparing performance metrics, the simpler Cox model and support vector regression (SVR) models, adapted from support vector machine methods with linear or radial basis functions, outperformed the more complex deep learning methods (Wen et al., 2022). The results above indicated that comparing to other pioneering methods, Cox PH regression model has low estimation error and high interpretability. Therefore, the Cox PH regression model might be generalizable if moderate number of variables are assessed to avoid overfitting.

However, the integration of patient information across different care settings is compounded by a lack of interoperability between EHR systems and insurance databases which combine to produce fragmented data that many research studies struggle with (Madden et al., 2016; Kharrazi et al., 2014). The setting of the Kaiser Permanente Washington (KPUW) site, with collaboration from different departments and a highly centralized concentration of a single insurance provider, may limit the generalizability of the findings. It is anticipated that datasets from smaller hospitals or personal clinics will have more scattered and cross-sectional patient histories than those from KPUW, reducing the richness and completeness of patients' longitudinal EHR data.

Currently, Epic, the major EHR vendor in the Pacific Northwest (PNW) region, has initiated efforts to unify and integrate patient profiles across different panels from collaborating practitioners and hospitals. However, this implementation's regional coverage and progression have not been quantified. The expansion of EHR implementation holds the potential for interoperable EHR analysis. By 2016, in the Medicare EHR Incentive Program, 186 certified health IT developers supplied certified health IT to 4,520 non-federal acute care hospitals (including Critical Access hospitals), with 96% using 2014 certified edition technology. Cerner, MEDITECH, Epic Systems, CPSI and its subsidiaries, McKesson, and MEDHOST provided 2014-certified technology to 92% of these hospitals (Office of the National Coordinator for Health Information Technology, 2017). The generalizability of our study is expected to be better in EHRs collected by the Epic system.

Overall, we utilized only 11 covariates of interest, and the Cox PH regression model proved to be a powerful tool for inference in our cohort EHR study. Given its performance with our dataset, we anticipate that the Cox PH model will maintain its effectiveness when expanded to other data sources. However, the generalizability of our findings may be limited to a collaborative and centralized data provider scenario due to the specific setting of the KPUW site and the centralized concentration of a single insurance provider.

4.3. Suggested Future Work

4.3.1. Expansion of Current Analysis

4.3.1.1. Modification and Addition to Regression Models

Additional research testing how combinations of risk factors influence ED via introducing interaction terms (e.g., CVD: age interaction) in regression models if sample size allows, could explore the inter-comorbidity relationships. Another suggestion is to consider time-dependent Cox models to handle the adjustment of predictors that may change over a period of follow-up. We may also compute the comparison between logistic and Cox models with reduction or adjustment of factors with high collinearity risk to reduce potential confounding effect. On the other hand, we may also expand the survival analysis by calculating the Nelson-Aalen cumulative hazard estimates to visualize the change of hazards over time. Meanwhile, the comparison between case groups vs. control groups or other stratified groups of patients can reveal group-specific survival curves. Moreover, comparing Cox regression simulated survival graphs and the Kaplan-Meier plots on full/multiple regression models and variable-specified stratified curves could provide additional model validation.

If data completeness allows, another analysis of consideration would be evaluating the distributions of vital measures and relevant regression models using fully PMM vs. partial estimation by neighboring values (height, weight, BMI) and partial imputation by PMM (SBP, DBP) to reveal the bias of full PMM method on imputation. Meanwhile, replacing numerical age values with periodic ranges or groups (e.g., 30-35-year-old men, 35-40-year-old men) may enhance the robustness and interpretability of regression models.

4.3.1.2. Areas for Further Improvement

Given that our cohort was lacking in Black, Hispanic, American Indian, and other minority groups, comparisons of it with population distribution within Washington state would reveal insight into potential sampling bias and underrepresentation. Meanwhile, the quality of the data can influence results and interpretations heavily. Increasing the date entries to a higher granularity level (e.g., retrieving days of EHR encounters instead of visit years) would increase accuracy in predicting survival rates. Survival analysis methods, such as the Cox model, require the proportional hazards assumption to be further validated. Furthermore, incorporating diverse patient profiles could improve personalized ED management. For instance, integrating patients' gene sequencing reports, using natural language processing for extracting narratives from unstructured clinical notes, and recording patients' social determinants (e.g., income, educational levels) and behavioral factors (e.g., cigarette alcohol consumption) can build up a foundation of phenotyping for more comorbidities on risk estimation and ED management.

4.3.2. Monitoring Younger Patients and Lifestyle Factors

Although ED is often described as a geriatric concern, one study by Araujo et al. (2000) found that depression interfered with sexual health in younger adult men (Araujo et al., 2000). In 2021, Calzo et al. reported that the estimated prevalence of ED in young adult sexually active men aged 18-31 years was approximately 11.3% mild and about 2.9% moderate-to-severe (Calzo et al., 2021). Therefore, we recommend expanding the monitoring of ED to younger populations too.

Meanwhile, lifestyle choices such as smoking, obesity, and physical inactivity contribute to the risk of ED (McVary et al., 2001). Obese men are riskier to endothelial dysfunction and hormonal dysregulation, while physical inactivity is associated with cardiovascular disease and diabetes (Feldman et al., 1994; Mannino et al., 1994; Bacon et al., 2002; Salem et al., 2009). All factors above increase the risk of ED development. Information on comorbid conditions, medications, and lifestyle factors is recommended to be included in the medical history for a comprehensive analysis and clinical decision support (Lue, 2000).

4.3.3. COVID-19 as a Risk Factor for ED

Recent research reported a high prevalence of ED in patients with COVID-19, indicating that ED can be an index to measure the severity of COVID-19 (Sansone et al., 2022). The vascular damage caused by the viral attacks and the inflammatory response induces endothelial dysfunction of the penis, injuring normal blood vessels and creating an obstruction for a fulfilling erection (Hsieh et al., 2022). At the same time, long COVID reactions, including chronic inflammation, ongoing fatigue, and cognitive symptoms, damage the nervous system chronically (Sansone et al., 2022). Psychological stressors such as anxiety and depression during the pandemic have an adverse effect on sexual health (Varma et al., 2021). Social isolation causes reduced physical activity, increased sedentary behavior, and alcohol intake, while all the physical, neural, and mental burdens related to COVID-19 contributed to an increasing incidence of ED during the pandemic (Peçanha et al., 2020).

COVID-19 has created limited access to regular healthcare services and face-to-face consultations during lockdowns with widespread adoption of telemedicine (Salonia et al.,

2021). The dynamic of ED progression, diagnosis, and management have changed massively after the pandemic of COVID-19. Managing modifiable lifestyle factors, including physical exercise promotion, reduction of alcohol consumption, and stress management with mental health support, will be important when attempting to address overall sexual health during and after the pandemic context(Peçanha et al., 2020; Varma et al., 2021).

Chapter 5. Conclusion

Our study successfully phenotyped the ED patients and its comorbidities in EHR provided by eMERGE KPUW cohort 3. And we were able to integrated the multi-sourced datasets for a comprehensive statistical analysis using logistic regression, Cox PH regression, and survival analysis methods. Univariable logistic regression and Cox regression models both indicated an association between each of the 11 covariates (age, race, median patient BMI, median patient DBP, median patient SBP, indicator of cardiovascular disease, indicator of diabetes, indicator of hypogonadism, indicator of depression, indicator of hypertension, and indicator of LUTS), and risk of ED. In full multivariable regression models containing all 11 covariates, strong associations of CVD, hypertension, LUTS, and hypogonadism with both odds and hazard of ED emphasize the importance of managing these conditions to prevent ED, supporting integrated care approaches that address multiple risk factors simultaneously. In the multivariable analysis, DBP, diabetes, and depression showed confounding effects.

Our study has implications for leveraging temporal relationships within integrated EHR-based informatics to allow insights into complex chronic diseases such as ED and contribute to the personalized care of patients with diabetes. Advanced informatics and statistical techniques provide a larger-scale vision of the multivariate aspects revolving around ED disease pathogenesis, enabling improved formulation/solutions. The goal is to ultimately improve the quality of life and healthcare outcomes for ED patients. These results also emphasize the importance of comorbidities in ED management, which can massively affect disease progression and outcomes. These principal methods of management may reduce the burden associated with ED and have a significant impact on patient satisfaction as well as health in general.

References

Adler, N. E., & Stead, W. W. (2015). Patients in context — EHR capture of social and behavioral determinants of health. *New England Journal of Medicine*, *372*(8), 698–701. doi:10.1056/nejmp1413945

Aguiar, J. A., Greenberg, D. R., Brannigan, R. E., Halpern, J. A., & Dubin, J. M. (2024). Beyond the prescription: Trends and challenges in erectile dysfunction medications among young adult men. *International Journal of Impotence Research*. doi:10.1038/s41443-024-00902-w

Ansong, K. S., Lewis, C., Jenkins, P., & Bell, J. (1999). Epidemiology of erectile dysfunction. *The Journal of Urology*, 222. doi:10.1097/00005392-199904010-00890

Araujo, A. B. (2000). Relation between psychosocial risk factors and incident erectile dysfunction: Prospective results from the Massachusetts Male Aging Study. *American Journal of Epidemiology*, *152*(6), 533–541. doi:10.1093/aje/152.6.533

Araujo, Andre B., Durante, R., Feldman, H. A., Goldstein, I., & McKinlay, J. B. (1998). The relationship between depressive symptoms and male erectile dysfunction. *Psychosomatic Medicine*, *60*(4), 458–465. doi:10.1097/00006842-199807000-00011

Atlantis, E., & Sullivan, T. (2012). Bidirectional association between depression and sexual dysfunction: A systematic review and meta-analysis. *The Journal of Sexual Medicine*, *9*(6), 1497–1507. doi:10.1111/j.1743-6109.2012.02709.x

Ayatollahi, H., Hosseini, S. F., & Hemmat, M. (2019). Integrating genetic data into electronic health records: Medical Geneticists' perspectives. *Healthcare Informatics Research*, *25*(4), 289. doi:10.4258/hir.2019.25.4.289

Aytaç, I. A., Araujo, A. B., Johannes, C. B., Kleinman, K. P., & McKinlay, J. B. (2000). Socioeconomic factors and incidence of erectile dysfunction: Findings of the longitudinal Massachussetts Male Aging Study. *Social Science & Medicine*, *51*(5), 771–778. doi:10.1016/s0277-9536(00)00022-8

Bacon, C. G., Mittleman, M. A., Kawachi, I., Giovannucci, E., Glasser, D. B., & Rimm, E. B. (2003). Sexual function in men older than 50 years of age: Results from the health professionals follow-up study. *Annals of Internal Medicine*, *139*(3), 161. doi:10.7326/0003-4819-139-3-200308050-00005

Barnes, D. E., & Yaffe, K. (2011). The projected effect of risk factor reduction on alzheimer's disease prevalence. *The Lancet Neurology*, *10*(9), 819–828. doi:10.1016/s1474-4422(11)70072-2

Blaisure, J., & Ceusters, W. (2017). Business Rules to Improve Secondary Data Use of Electronic Healthcare Systems. *Studies in health technology and informatics*, 235, 303–307.

Botsis, T., Hartvigsen, G., Chen, F., & Weng, C. (2010). Secondary Use of EHR: Data Quality Issues and Informatics Opportunities. *Summit on translational bioinformatics*, 2010, 1–5.

Boytcheva, S., Angelova, G., Angelov, Z., Tcharaktchiev, D., & Vodenicharov, V. (2019). Enrichment of EHR with linked open data for risk factors identification. *Proceedings of*

the 20th International Conference on Computer Systems and Technologies. doi:10.1145/3345252.3345290

Broce, I. J., Tan, C. H., Fan, C. C., Jansen, I., Savage, J. E., Witoelar, A., … Desikan, R. S. (2018). Dissecting the genetic relationship between cardiovascular risk factors and alzheimer's disease. *Acta Neuropathologica*, *137*(2), 209–226. doi:10.1007/s00401-018-1928-6

Burnett, A. L., Nehra, A., Breau, R. H., Culkin, D. J., Faraday, M. M., Hakim, L. S., … Shindel, A. W. (2018). Erectile dysfunction: AUA Guideline. *Journal of Urology*, *200*(3), 633–641. doi:10.1016/j.juro.2018.05.004

Bush, W. S., Oetjens, M. T., & Crawford, D. C. (2016). Unravelling the human genome–phenome relationship using phenome-wide association studies. *Nature Reviews Genetics*, *17*(3), 129–145. doi:10.1038/nrg.2015.36

Calzo, J. P., Austin, S. B., Charlton, B. M., Missmer, S. A., Kathrins, M., Gaskins, A. J., & Chavarro, J. E. (2021a). Erectile dysfunction in a sample of sexually active young adult men from a U.S. cohort: Demographic, metabolic and mental health correlates. *Journal of Urology*, *205*(2), 539–544. doi:10.1097/ju.0000000000001367

Calzo, J. P., Austin, S. B., Charlton, B. M., Missmer, S. A., Kathrins, M., Gaskins, A. J., & Chavarro, J. E. (2021b). Erectile dysfunction in a sample of sexually active young adult men from a U.S. cohort: Demographic, metabolic and mental health correlates. *Journal of Urology*, *205*(2), 539–544. doi:10.1097/ju.0000000000001367

Chen, X., Pan, W., Kwok, J. T., & Carbonell, J. G. (2009). Accelerated gradient method for multi-task sparse learning problem. *2009 Ninth IEEE International Conference on Data Mining*. doi:10.1109/icdm.2009.128

Chu, K. Y., Nackeeran, S., Horodyski, L., Masterson, T. A., & Ramasamy, R. (2021). Covid-19 infection is associated with new onset erectile dysfunction: Insights from a national registry. *Sexual Medicine*, *10*(1), 100478–1. doi:10.1016/j.esxm.2021.100478

Cummings, J. R., Allen, L., Clennon, J., Ji, X., & Druss, B. G. (2017). Geographic access to specialty mental health care across high- and low-income US communities. *JAMA Psychiatry*, *74*(5), 476. doi:10.1001/jamapsychiatry.2017.0303

Denny, J. C., Bastarache, L., Ritchie, M. D., Carroll, R. J., Zink, R., Mosley, J. D., … Roden, D. M. (2013). Systematic comparison of phenome-wide association study of Electronic Medical Record Data and genome-wide association study data. *Nature Biotechnology*, *31*(12), 1102–1111. doi:10.1038/nbt.2749

Elterman, D. S., Bhattacharyya, S. K., Mafilios, M., Woodward, E., Nitschelm, K., & Burnett, A. L. (2021). The quality of life and economic burden of erectile dysfunction. *Research and Reports in Urology*, *Volume 13*, 79–86. doi:10.2147/rru.s283097

eMERGE Consortium. (n.d.). Retrieved from https://emerge-network.org/emerge-sites/

Feldman, H. A., Goldstein, I., Hatzichristou, D. G., Krane, R. J., & McKinlay, J. B. (1994). Impotence and its medical and psychosocial correlates: Results of the Massachusetts Male Aging Study. *Journal of Urology*, *151*(1), 54–61. doi:10.1016/s0022-5347(17)34871-1

Ganesh, S. K., Chasman, D. I., Larson, M. G., Guo, X., Verwoert, G., Bis, J. C., … Munroe, P. B. (2014). Effects of long-term averaging of quantitative blood pressure traits on the detection of genetic associations. *The American Journal of Human Genetics*, *95*(1), 49–65. doi:10.1016/j.ajhg.2014.06.002

Garrison, L. P., Neumann, P. J., Erickson, P., Marshall, D., & Mullins, C. D. (2007). Using real-world data for coverage and payment decisions: The ISPOR Real-World Data task force report. *Value in Health*, *10*(5), 326–335. doi:10.1111/j.1524-4733.2007.00186.x

Goldie, S. J., Gaffikin, L., Goldhaber-Fiebert, J. D., Gordillo-Tobar, A., Levin, C., Mahé, C., & Wright, T. C. (2005). Cost-effectiveness of cervical-cancer screening in five developing countries. *New England Journal of Medicine*, *353*(20), 2158–2168. doi:10.1056/nejmsa044278

Goldstein, I., Goren, A., Li, V. W., Tang, W. Y., & Hassan, T. A. (2019). Epidemiology update of erectile dysfunction in eight countries with high burden. *Sexual Medicine Reviews*, *8*(1), 48–58. doi:10.1016/j.sxmr.2019.06.008

Gottesman, O., Kuivaniemi, H., Tromp, G., Faucett, W. A., Li, R., Manolio, T. A., … Williams, M. S. (2013). The electronic medical records and genomics (EMERGE) network: Past, present, and future. *Genetics in Medicine*, *15*(10), 761–771. doi:10.1038/gim.2013.72

Hagar, Y., Albers, D., Pivovarov, R., Chase, H., Dukic, V., & Elhadad, N. (2014a). Survival analysis with electronic health record data: Experiments with chronic kidney disease. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, *7*(5), 385–403. doi:10.1002/sam.11236

Hagar, Y., Albers, D., Pivovarov, R., Chase, H., Dukic, V., & Elhadad, N. (2014b). Survival analysis with electronic health record data: Experiments with chronic kidney disease. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, *7*(5), 385–403. doi:10.1002/sam.11236

Hanauer, D. A., Gardner, M., & Sandberg, D. E. (2014). Unbiased identification of patients with disorders of sex development. *PLoS ONE*, *9*(9). doi:10.1371/journal.pone.0108702

Haneuse, S., & Daniels, M. (2016). A general framework for considering selection bias in EHR-based studies: What data are observed and why? *eGEMs (Generating Evidence & Methods to Improve Patient Outcomes)*, *4*(1), 16. doi:10.13063/2327-9214.1203

Hebert, K. J., Matta, R., Horns, J. J., Paudel, N., Das, R., McCormick, B. J., … Hotaling, J. M. (2023). Prior covid-19 infection associated with increased risk of newly diagnosed erectile dysfunction. *International Journal of Impotence Research*, *36*(5), 521–525. doi:10.1038/s41443-023-00687-4

Holmes, J. H., Beinlich, J., Boland, M. R., Bowles, K. H., Chen, Y., Cook, T. S., … Moore, J. H. (2021). Why is the electronic health record so challenging for research and clinical care? *Methods of Information in Medicine*, *60*(01/02), 032–048. doi:10.1055/s-0041-1731784

Hsieh, T.-C., Edwards, N. C., Bhattacharyya, S. K., Nitschelm, K. D., & Burnett, A. L. (2022). The epidemic of covid-19-related erectile dysfunction: A scoping review and Health Care Perspective. *Sexual Medicine Reviews*, *10*(2), 286–310. doi:10.1016/j.sxmr.2021.09.002

Ishwaran, H., Kogalur, U. B., Blackstone, E. H., & Lauer, M. S. (2008). Random survival forests. *The Annals of Applied Statistics*, *2*(3). doi:10.1214/08-aoas169

Jensen, P. B., Jensen, L. J., & Brunak, S. (2012). Mining Electronic Health Records: Towards Better Research Applications and clinical care. *Nature Reviews Genetics*, *13*(6), 395–405. doi:10.1038/nrg3208

Jorgenson, E., Matharu, N., Palmer, M. R., Yin, J., Shan, J., Hoffmann, T. J., … Van Den Eeden, S. K. (2018). Genetic variation in the *sim1* locus is associated with erectile dysfunction. *Proceedings of the National Academy of Sciences*, *115*(43), 11018–11023. doi:10.1073/pnas.1809872115

Katzman, J. L., Shaham, U., Cloninger, A., Bates, J., Jiang, T., & Kluger, Y. (2018). Deepsurv: Personalized Treatment Recommender system using a Cox proportional hazards deep neural network. *BMC Medical Research Methodology*, *18*(1). doi:10.1186/s12874-018-0482-1

Kern, L. M., & Kaushal, R. (2013). Electronic Health Records and Ambulatory Quality. *Journal of General Internal Medicine*, *28*(9), 1133–1133. doi:10.1007/s11606-013-2475-4

Kessler, A., Sollie, S., Challacombe, B., Briggs, K., & Van Hemelrijck, M. (2019). The global prevalence of erectile dysfunction: A Review. *BJU International*, *124*(4), 587–599. doi:10.1111/bju.14813

Kharrazi, H., Wang, C., & Scharfstein, D. (2014). Prospective EHR-based clinical trials: The Challenge of Missing Data. *Journal of General Internal Medicine*, *29*(7), 976–978. doi:10.1007/s11606-014-2883-0

Kho, A. N., Pacheco, J. A., Peissig, P. L., Rasmussen, L., Newton, K. M., Weston, N., … Denny, J. C. (2011). Electronic medical records for genetic research: Results of the emerge consortium. *Science Translational Medicine*, *3*(79). doi:10.1126/scitranslmed.3001807

Kitaw, T. A., Abate, B. B., Tilahun, B. D., Yilak, G., & Haile, R. N. (2024). Umbrella Review protocol: Global burden and risk factors of erectile dysfunction in diabetic population. *Health Science Reports*, *7*(6). doi:10.1002/hsr2.2159

Kitchen, C. A., Chang, H.-Y., Bishop, M. A., Shermock, K. M., Kharrazi, H., & Weiner, J. P. (2022). Comparing and validating medication complexity from insurance claims against Electronic Health Records. *Journal of Managed Care & Specialty Pharmacy*, *28*(4), 473–484. doi:10.18553/jmcp.2022.28.4.473

Kvamme, H., Borgan, Ø., & Scheel, I. (2019). Time-to-event prediction with neural networks and cox regression. *Journal of Machine Learning Research*, 20(129), 1–30.

Laumann, E O, Nicolosi, A., Glasser, D. B., Paik, A., Gingell, C., Moreira, E., & Wang, T. (2004). Sexual problems among women and men aged 40–80 y: Prevalence and correlates identified in the global study of sexual attitudes and behaviors. *International Journal of Impotence Research*, *17*(1), 39–57. doi:10.1038/sj.ijir.3901250

Laumann, Edward O., Paik, A., & Rosen, R. C. (1999). Sexual dysfunction in the United States. *JAMA*, *281*(6), 537. doi:10.1001/jama.281.6.537

Lee, C., Zame, W., Yoon, J., & Van der Schaar, M. (2018). DeepHit: A deep learning approach to survival analysis with competing risks. *Proceedings of the AAAI Conference on Artificial Intelligence*, *32*(1). doi:10.1609/aaai.v32i1.11842

Lemke, A. A., Wu, J. T., Waudby, C., Pulley, J., Somkin, C. P., & Trinidad, S. B. (2010). Community engagement in biobanking: Experiences from the Emerge Network. *Genomics, Society and Policy*, *6*(3). doi:10.1186/1746-5354-6-3-50

Li, Y., He, X., Wheldon, C., Wu, Y., Prosperi, M., Shenkman, E. A., Jaffee, M. S., Guo, J., Wang, F., Guo, Y., & Bian, J. (2024). A Computable Phenotype for the Identification of Sexual and Gender Minorities in Electronic Health Records. AMIA ... Annual Symposium proceedings. AMIA Symposium, 2023, 1057–1066.

Little, R. J. A., & Rubin, D. B. (2020). *Statistical analysis with missing data*. Hoboken, NJ: Wiley.

Liu, J., Ji, S., & Ye, J. (2009). Multi-task feature learning via efficient l 2, 1-norm minimization. *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, 339–348.

Liu, Q., Zhang, Y., Wang, J., Li, S., Cheng, Y., Guo, J., … Zhu, Z. (2018). Erectile dysfunction and depression: A systematic review and meta-analysis. *The Journal of Sexual Medicine*, *15*(8), 1073–1082. doi:10.1016/j.jsxm.2018.05.016

Lue, T. F. (2000). Erectile dysfunction. *New England Journal of Medicine*, *342*(24), 1802–1813. doi:10.1056/nejm200006153422407

Madden, J. M., Lakoma, M. D., Rusinak, D., Lu, C. Y., & Soumerai, S. B. (2016). Missing clinical and behavioral health data in a large electronic health record (EHR) system. *Journal of the American Medical Informatics Association*, *23*(6), 1143–1149. doi:10.1093/jamia/ocw021

Malavige, L. S., & Levy, J. C. (2009). Erectile dysfunction in diabetes mellitus. *The Journal of Sexual Medicine*, *6*(5), 1232–1247. doi:10.1111/j.1743-6109.2008.01168.x

Mannino, D. M., Klevens, R. M., & Flanders, W. D. (1994). Cigarette smoking: An independent risk factor for impotence? *American Journal of Epidemiology*, *140*(11), 1003–1008. doi:10.1093/oxfordjournals.aje.a117189

Mark, K. P., Arenella, K., Girard, A., Herbenick, D., Fu, J., & Coleman, E. (2024). Erectile dysfunction prevalence in the United States: Report from the 2021 National Survey of Sexual Wellbeing. *The Journal of Sexual Medicine*, *21*(4), 296–303. doi:10.1093/jsxmed/qdae008

McCarty, C. A., Chisholm, R. L., Chute, C. G., Kullo, I. J., Jarvik, G. P., Larson, E. B., Li, R., Masys, D. R., Ritchie, M. D., Roden, D. M., & eMERGE Network. (2011). The eMERGE Network: A consortium of biorepositories linked to electronic medical records data for conducting genomic studies. *BMC Medical Genomics*, 4(13). https://doi.org/10.1186/1755-8794-4-13

McVary, K. T., Carrier, S., & Wessells, H. (2001). Smoking and erectile dysfunction: Evidence based analysis. *Journal of Urology*, *166*(5), 1624–1632. doi:10.1016/s0022-5347(05)65641-8

Motheral, B. R., & Fairman, K. A. (1997). The use of claims databases for outcomes research: Rationale, challenges, and strategies. *Clinical Therapeutics*, *19*(2), 346–366. doi:10.1016/s0149-2918(97)80122-1

Mulhall, J. P., Luo, X., Zou, K. H., Stecher, V., & Galaznik, A. (2016). Relationship between age and erectile dysfunction diagnosis or treatment using real-world observational data in the USA. *International Journal of Clinical Practice*, *70*(12), 1012–1018. doi:10.1111/ijcp.12908

Muneer, A., Kalsi, J., Nazareth, I., & Arya, M. (2014). Erectile dysfunction. *BMJ*, *348*(jan27 7). doi:10.1136/bmj.g129

Nackeeran, S., Havanur, A., Ory, J., Althof, S., & Ramasamy, R. (2021). Erectile dysfunction is a modifiable risk factor for major depressive disorder: Analysis of a Federated Research Network. *The Journal of Sexual Medicine*, *18*(12), 2005–2011. doi:10.1016/j.jsxm.2021.09.016

Nair, S., Hsu, D., & Celi, L. A. (2016). Challenges and opportunities in secondary analyses of electronic health record data. *Secondary Analysis of Electronic Health Records*, 17–26. doi:10.1007/978-3-319-43742-2_3

Nelson, C. J., Mulhall, J. P., & Roth, A. J. (2011). The association between erectile dysfunction and depressive symptoms in men treated for prostate cancer. *The Journal of Sexual Medicine*, *8*(2), 560–566. doi:10.1111/j.1743-6109.2010.02127.x

NIH Consensus Conference. (1993). Impotence. *JAMA*, *270*(1), 83. doi:10.1001/jama.1993.03510010089036

Nobles, A. L., Vilankar, K., Wu, H., & Barnes, L. E. (2015). Evaluation of data quality of multisite electronic health record data for secondary analysis. *2015 IEEE International Conference on Big Data (Big Data)*. doi:10.1109/bigdata.2015.7364060

Norton, S., Matthews, F. E., Barnes, D. E., Yaffe, K., & Brayne, C. (2014). Potential for primary prevention of alzheimer's disease: An analysis of population-based data. *The Lancet Neurology*, *13*(8), 788–794. doi:10.1016/s1474-4422(14)70136-x

Nunes, A. P., Seeger, J. D., Stewart, A., Gupta, A., & McGraw, T. (2021a). Cardiovascular outcome risks in patients with erectile dysfunction co-prescribed a phosphodiesterase type 5 inhibitor (PDE5I) and a nitrate: A retrospective observational study using electronic health record data in the United States. *The Journal of Sexual Medicine*, *18*(9), 1511–1523. doi:10.1016/j.jsxm.2021.06.010

Nunes, A. P., Seeger, J. D., Stewart, A., Gupta, A., & McGraw, T. (2021b). Retrospective observational real-world outcome study to evaluate safety among patients with erectile dysfunction (ED) with co-possession of Tadalafil and anti-hypertensive medications (anti-HTN). *The Journal of Sexual Medicine*, *19*(1), 74–82. doi:10.1016/j.jsxm.2021.10.012

Office of the National Coordinator for Health Information Technology (Ed.). (2017). Certified Health IT Developers and Editions Reported by Hospitals Participating in the Medicare EHR Incentive Program. Retrieved from https://www.healthit.gov/data/quickstats/hospital-health-it-developers Health IT Quick-Stat #29

Park, J.-H., Wacholder, S., Gail, M. H., Peters, U., Jacobs, K. B., Chanock, S. J., & Chatterjee, N. (2010). Estimation of effect size distribution from genome-wide association studies and implications for future discoveries. *Nature Genetics*, *42*(7), 570–575. doi:10.1038/ng.610

Penson, D. F., & Wessells, H. (2004). Erectile dysfunction in diabetic patients. *Diabetes Spectrum*, *17*(4), 225–230. doi:10.2337/diaspect.17.4.225

Peçanha, T., Goessler, K. F., Roschel, H., & Gualano, B. (2020). Social isolation during the COVID-19 pandemic can increase physical inactivity and the global burden of cardiovascular disease. *American Journal of Physiology-Heart and Circulatory Physiology*, *318*(6). doi:10.1152/ajpheart.00268.2020

Rew, K. T., & Heidelbaugh, J. J. (2016). Erectile Dysfunction. *American family physician*, 94(10), 820–827.

Rezaee, M. E., Ward, C. E., Brandes, E. R., Munarriz, R. M., & Gross, M. S. (2020). A review of economic evaluations of erectile dysfunction therapies. *Sexual Medicine Reviews*, *8*(3), 497–503. doi:10.1016/j.sxmr.2019.06.001

Roberts, K., Shooshan, S. E., Rodriguez, L., Abhyankar, S., Kilicoglu, H., & Demner-Fushman, D. (2015). The role of fine-grained annotations in supervised recognition of risk factors for heart disease from ehrs. *Journal of Biomedical Informatics*, *58*. doi:10.1016/j.jbi.2015.06.010

Rosen, R. C., Riley, A., Wagner, G., Osterloh, I. H., Kirkpatrick, J., & Mishra, A. (1997). The international index of erectile function (IIEF): A multidimensional scale for assessment of erectile dysfunction. *Urology*, *49*(6), 822–830. doi:10.1016/s0090-4295(97)00238-0

Rosen, R., Altwein, J., Boyle, P., Kirby, R. S., Lukacs, B., Meuleman, E., … Giuliano, F. (2003). Lower urinary tract symptoms and male sexual dysfunction: The multinational survey of the aging male (MSAM-7). *European Urology*, *44*(6), 637–649. doi:10.1016/j.eururo.2003.08.015

Saigal, C. S. (2006). Predictors and prevalence of erectile dysfunction in a racially diverse population. *Archives of Internal Medicine*, *166*(2), 207. doi:10.1001/archinte.166.2.207

Salem, S., Abdi, S., Mehrsai, A., Saboury, B., Saraji, A., Shokohideh, V., & Pourmand, G. (2009). Erectile dysfunction severity as a risk predictor for coronary artery disease. *The Journal of Sexual Medicine*, *6*(12), 3425–3432. doi:10.1111/j.1743-6109.2009.01515.x

Salonia, A., Pontillo, M., Capogrosso, P., Gregori, S., Tassara, M., Boeri, L., … Montorsi, F. (2021). Severely low testosterone in males with COVID-19: A case-control study. *Andrology*, *9*(4), 1043–1052. doi:10.1111/andr.12993

Sandhu, E., Weinstein, S., McKethan, A., & Jain, S. H. (2012). Secondary uses of electronic health record data: Benefits and barriers. *The Joint Commission Journal on Quality and Patient Safety*, *38*(1), 34–40. doi:10.1016/s1553-7250(12)38005-7

Sansone, A., Mollaioli, D., Ciocca, G., Colonnello, E., Limoncin, E., Balercia, G., & Jannini, E. A. (2022). “mask up to keep it up”: Preliminary evidence of the association between erectile dysfunction and covid-19. *The Journal of Sexual Medicine*, *19*(Supplement_4). doi:10.1016/j.jsxm.2022.08.093

Schneeweiss, S, Gagne, J. J., Glynn, R. J., Ruhl, M., & Rassen, J. A. (2011). Assessing the comparative effectiveness of newly marketed medications: Methodological challenges and implications for drug development. *Clinical Pharmacology & Therapeutics*, *90*(6), 777–790. doi:10.1038/clpt.2011.235

Schneeweiss, S., Seeger, J. D., Maclure, M., Wang, P. S., Avorn, J., & Glynn, R. J. (2001). Performance of comorbidity scores to control for confounding in epidemiologic studies using claims data. American journal of epidemiology, 154(9), 854–864. https://doi.org/10.1093/aje/154.9.854

Schneeweiss, Sebastian, & Avorn, J. (2005). A review of uses of health care utilization databases for epidemiologic research on Therapeutics. *Journal of Clinical Epidemiology*, *58*(4), 323–337. doi:10.1016/j.jclinepi.2004.10.012

Selvin, E., Burnett, A. L., & Platz, E. A. (2007). Prevalence and risk factors for erectile dysfunction in the US. *The American Journal of Medicine*, *120*(2), 151–157. doi:10.1016/j.amjmed.2006.06.010

Shabsigh, R. (2001). Socioeconomic considerations in erectile dysfunction treatment. *Urologic Clinics of North America*, *28*(2), 417–422. doi:10.1016/s0094-0143(05)70149-x

Sidebottom, A. C., Johnson, P. J., VanWormer, J. J., Sillah, A., Winden, T. J., & Boucher, J. L. (2015). Exploring electronic health records as a population health surveillance tool of cardiovascular disease risk factors. *Population Health Management*, *18*(2), 79–85. doi:10.1089/pop.2014.0058

Stanaway, I. B., Hall, T. O., Rosenthal, E. A., Palmer, M., Naranbhai, V., Knevel, R., Namjou-Khales, B., Carroll, R. J., Kiryluk, K., Gordon, A. S., Linder, J., Howell, K. M., Mapes, B. M., Lin, F., Joo, Y. Y., Hayes, M. G., Gharavi, A. G., Pendergrass, S. A., Ritchie, M. D., de Andrade, M., ... Crosslin, D. R. (2019). The eMERGE genotype set of 83,717 subjects imputed to ~40 million variants genome wide and association with the herpes zoster medical record phenotype. *Genetic epidemiology*, 43(1), 63–81. https://doi.org/10.1002/gepi.22167

Tan, H. L. (2000). Economic cost of male erectile dysfunction using a decision analytic model. *PharmacoEconomics*, *17*(1), 77–107. doi:10.2165/00019053-200017010-00006

Varma, P., Junge, M., Meaklim, H., & Jackson, M. L. (2021). Younger people are more vulnerable to stress, anxiety and depression during COVID-19 pandemic: A Global Cross-sectional survey. *Progress in Neuro-Psychopharmacology and Biological Psychiatry*, *109*, 110236. doi:10.1016/j.pnpbp.2020.110236

Verma, A., Bradford, Y., Dudek, S., Lucas, A. M., Verma, S. S., Pendergrass, S. A., & Ritchie, M. D. (2018). A simulation study investigating power estimates in phenome-wide association studies. *BMC Bioinformatics*, *19*(1). doi:10.1186/s12859-018-2135-0

Vinzamuri, B., & Reddy, C. K. (2013). Cox regression with correlation based regularization for Electronic Health Records. *2013 IEEE 13th International Conference on Data Mining*. doi:10.1109/icdm.2013.89

Wells, B. J., Nowacki, A. S., Chagin, K., & Kattan, M. W. (2013). Strategies for handling missing data in electronic health record derived data. *eGEMs (Generating Evidence & Methods to Improve Patient Outcomes)*, *1*(3), 7. doi:10.13063/2327-9214.1035

Wen, Y., Rahman, M. F., Zhuang, Y., Pokojovy, M., Xu, H., McCaffrey, P., … Tseng, T.-L. (Bill). (2022). Time-to-event modeling for hospital length of stay prediction for covid-19 patients. *Machine Learning with Applications*, *9*, 100365. doi:10.1016/j.mlwa.2022.100365

Wessells, H., Joyce, G. F., Wise, M., & Wilt, T. J. (2007). Erectile dysfunction. *Journal of Urology*, *177*(5), 1675–1681. doi:10.1016/j.juro.2007.01.057

Woldemariam, S., Xie, F., Roldan, A., Roger, J., Tang, A., Oskotsky, T., … Sirota, M. (2023). Leveraging electronic medical records reveals comorbidities significantly associated with male infertility. *Fertility and Sterility*, *120*(4). doi:10.1016/j.fertnstert.2023.08.173

Yafi, F. A., Jenkins, L., Albersen, M., Corona, G., Isidori, A. M., Goldfarb, S., … Hellstrom, W. J. (2016). Erectile dysfunction. *Nature Reviews Disease Primers*, *2*(1). doi:10.1038/nrdp.2016.3

Yang, S., Varghese, P., Stephenson, E., Tu, K., & Gronsbell, J. (2022). Machine learning approaches for electronic health records phenotyping: A methodical review. *Journal of the American Medical Informatics Association*, *30*(2), 367–381. doi:10.1093/jamia/ocac216

Yu C.-N., Greiner R., Lin H.-C., Baracos V. Learning patient-specific cancer survival distributions as a sequence of dependent regressors. *Advances in Neural Information Processing Systems*. 2011;24:1845–1853.

Zhang, C., Maroufy, V., Chen, B., & Wu, H. (2020). Missing data issues in Ehr. *Statistics and Machine Learning Methods for EHR Data*, 149–173. doi:10.1201/9781003030003-6

Zhao, Y., & Tsubota, T. (2023). The current status of secondary use of claims, electronic medical records, and Electronic Health Records in epidemiology in Japan: Narrative literature review. *JMIR Medical Informatics*, *11*. doi:10.2196/39876

Appendix I

Erectile Dysfunction (ED)
Phenotype Algorithm
July 2020

Authors: Hunter Wessells wessells@uw.edu, Melody Palmer, David Crosslin davidcr@uw.edu, Ian Stanaway bard@uw.edu, Alex Skokan ajskokan@uw.edu, Lester Kirchner hlkirchner@geisinger.edu, Anne Justice aejustice1@geisinger.edu

1. Data requirements
- Patient demographics
- Diagnosis codes (ICD9 vs 10)
- Procedure codes
- Medication records

2. Inclusion criteria
- Male
- Age greater than or equal to 20 years
- Exclusions:
    - Prostate cancer
        - ICD-9 code: 185
        - ICD-10 code: C61, N52.31 – N52.37
    - Spinal cord injury
        - ICD-9 codes: 952 (all subcodes), 344.0 (all subcodes), 344.1
        - ICD-10 codes: S14.1 (all subcodes), S24.1 (all subcodes), S34.1 (all subcodes), G82 (all subcodes)
    - Radical pelvic surgery
    - Hyperactive sexual desire disorder: ICD-10 F52.8

A csv file of the Exclusion Codes, including all subcodes in the ICD tree is available on the web here:

https://droog.gs.washington.edu/~bard/emerge/ED/exclusion_codes.csv

- Minimum contact:

  Contact on two (2) or more occasions with at least 5 years between first and last ICD and/or CPT code.

3. ED Case criteria
- ICD-9 for organic ED: 607.84
- ICD-9 for psychosexual dysfunction: 302.70, 302.72
    - Plan to look at initial Dx of psychogenic ED & subsequent onset of "organic ED"

- ICD-10 for organic ED: N52.0-N52.03, N52.1, N52.8, N52.9
- ICD-10 for psychosexual dysfunction: F52.21, R37

and/or

- ICD-9 procedure codes: 64.94, 64.95, 64.96, 64.97
- ICD-10 procedure codes: 0VUSXJZ, 0VUS0JZ, 0VUS4JZ, 0VUS0KZ, 0VUS4KZ, 0VUSXKZ, 0VPS0JZ, 0VPS3JZ, 0VPS4JZ, 0VPS7JZ, 0VPS8JZ, 0VWS0JZ, 0VWS3JZ, 0VWS4JZ, 0VWS8JZ, 0VWSXJZ
- CPT codes: 37788, 37790, 54115, 54230, 54231, 54235, 54240, 54250, 54400, 54401, 54402, 54405, 54406, 54407, 54408, 54409, 54410, 54411, 54415, 54416, 54417
- Medications: Aphrodyne, Caverject, Cialis*, Dayto-Himbin, Edex, Erex, Levitra*, Mederek, Muse, Staxyn, Supina Nr, Testomar, Viagra, Viril-Lam, Viritab, Yocon, Yohimar, Yohimbine Hcl, Yohimex, Yoman, Yovital, Alprostadil, Sildenafil Citrate*, Tadalafil*, Vardenafil Hcl*, Yohimbine Hcl, Yohimbine Hcl/Zinc Sulfate
  (Note: eMERGE does not have access to these medication records)

A csv file of the Case Definition Codes, including all subcodes in the ICD tree is available on the web here:

https://droog.gs.washington.edu/~bard/emerge/ED/case_codes.csv

*Exclude patients taking drug with diagnosis of pulmonary hypertension unless they have additional CPT or ICD-9 of ED.

4. ED Control criteria

- No ED diagnoses
- No ED medication
- No ED procedure
- Exclusions: all same as cases

5. Covariates

- Ancestry
- Sites
- Patient demographics
- Age at diagnosis or Rx, or most recent record for controls
- Smoking status – hard to get from EMERGE or Geisinger
- Diabetes diagnoses
  - ICD-9 codes: 249 (all subcodes), 250 (all subcodes)
  - ICD-10 codes: E08 (all subcodes), E09 (all subcodes), E10 (all subcodes), E11 (all subcodes), E13 (all subcodes)
- Hypertension diagnosis – code diagnosis
  - ICD-9 codes: 401 (all subcodes), 402 (all subcodes), 403 (all subcodes), 404 (all subcodes), 405 (all subcodes)
  - ICD-10 codes: I10, I11, I12, I13, I15, I16
- BMI – dirty phenotype in EMERGE; coding errors metric vs. English.

- Depression diagnosis
  - ICD-9 codes: 296 (all subcodes), 311
  - ICD-10 codes: F30 – F34 (all subcodes)
- BPH diagnosis
  - ICD-9 codes: 594.1, 599.6, 600.00-600.01, 600.20-600.21, 600.90-600.91
    - Previously: 594.1, 599.6, 600.0 – 600.91
  - ICD-10 codes: N21.0, N40.0 – N40.3
  - These codes have been removed from the BPH covariate definition:
    BPH,ICD10,N40.2,Nodular_prostate_without_lower_urinary_tract_symptoms
    BPH,ICD9,600.10,Nodular_prostate_without_urinary_obstruction
    BPH,ICD9,600.3,Cyst_of_prostate
- LUTS diagnosis
  - ICD-9 codes: 788.2 (all subcodes), 788.3 (all subcodes), 788.4 (all subcodes), 788.6 (all subcodes), 788.9 (all subcodes)
  - ICD-10 codes: R32, R33.0, R33.8, R35.0, R35.1, R35.8 R39.1 (all subcodes), R39.8, R39.81, R39.82, R39.89, R39.9
- Cardiovascular disease
  - ICD-9 codes: 402 (all subcodes), 404 (all subcodes), 410-414 (all subcodes), 425 (all subcodes), 428 (all subcodes), 429.2, 433 (all subcodes), 434 (all subcodes), 440.0-440.29, 440.4, 440.8
  - ICD-10 codes: I11, I13, I20-I25, I42, I43, I50, I63, I65, I67, I70.0-I70.299
- Hypogonadism
  - ICD-9 codes: 257.2, 257.8, 257.9
  - ICD-10 codes: E29.1, E29.8, E29.9

A csv file of the Covariate Definition Codes, including all subcodes in the ICD tree is available on the web here:

https://droog.gs.washington.edu/~bard/emerge/ED/covariates_codes.csv

6. Time to event criteria.

Age at ICD/CPT record is the time to event metric used in analyses. Both left and right censoring are implemented in this analysis. A subject's first ICD/CPT record is used as the left censor age (age_start). One year is subtracted from this value to prevent ties where the first record is also the event of interest (i.e. ED). A minimum age of 20 years old is utilized for left truncation and is used as the minimum age for left censoring (age_start). The right censor is the last ICD/CPT record for controls (age_stop). Cases have the age at first ICD/CPT records as the age at event (age_stop).

7. R Markdown Scripts that implement the phenotyping algorithm are available here. Since this is eMERGE data we are not including the medications. The algorithm uses ICD and CPT codes.

There are two versions of the phenotype code:

7.1 The minimum code count 1 (mcc1) ED phenotype version that requires the cases to have one ICD/CPT code supporting their ED phenotype:

https://droog.gs.washington.edu/~bard/emerge/ED/ED_mcc1_phenotype_w_covariates_20200608.Rmd

This is the summary phenotyping output:

https://droog.gs.washington.edu/~bard/emerge/ED/ED_mcc1_phenotype_w_covariates_20200608.html
An example analysis, model fit testing and plotting of the Kaplan Meier Curves are here:

https://droog.gs.washington.edu/~bard/emerge/ED/ed_mcc1_survival_model_left_censor_20200513.Rmd

Emple Output:

https://droog.gs.washington.edu/~bard/emerge/ED/ed_mcc1_survival_model_left_censor_20200513.html

7.2 The minimum code count 2 (mcc2) ED phenotype version that requires the cases to have two ICD/CPT codes supporting their ED phenotype:

https://droog.gs.washington.edu/~bard/emerge/ED/ED_mcc2_phenotype_w_covariates_20200512.Rmd

This is the summary phenotyping output:

https://droog.gs.washington.edu/~bard/emerge/ED/ED_mcc2_phenotype_w_covariates_20200512.html

An example analysis, model fit testing and plotting of the Kaplan Meier Curves are here:

https://droog.gs.washington.edu/~bard/emerge/ED/ed_mcc2_survival_model_left_censor_20200807.Rmd

Example Output:

https://droog.gs.washington.edu/~bard/emerge/ED/ed_mcc2_survival_model_left_censor_20200807.html