THE UNIVERSITY OF CHICAGO

CONTINUOUS TEMPORAL SIGNALS AND ELECTRONIC HEALTH RECORDS FOR BROAD HEALTH STATES FORECASTING

A DISSERTATION SUBMITTED TO
THE FACULTY OF THE PRITZKER SCHOOL OF MOLECULAR ENGINEERING
IN CANDIDACY FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

BY

CHENGJIAN SHI

CHICAGO, ILLINOIS
AUGUST 2023

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

Reviewing my academic journey for the past five years, I would start by thanking my advisors, Professor. Andrey Rzhetsky and Professor. Robert L. Grossman. They put faith in me at my most difficult time and introduced me into data mining and statistical modeling, where I devoted to build a career for life. Andrey is a very patient and passionate advisor, who shows absolute support for me to explore different opportunities for my future careers. His trust and kindness motivate me to explore the unlimited possibility for my future. Robert is another wise advisor with rich experiences, both in academic and industrial fields, which I admire and constantly seeking suggestions from. I'm extremely fortunate to have them as my advisors through this journey, their characteristics and high standard in science would be something I will pursue forever.

Besides, I would also want to express appreciate to my collaborators, including Dr. Megan Huisingh-Sheetz, Dr. Rajesh Jain and Dr. Louis Philipson. They provided most professional knowledge and cautious attitude, which guaranteed the solid foundation for the project. Through the collaboration, I understood the University of Chicago core value on interdisciplinary study to learn from the top talents in multiple fields. Besides, I want to thank my committee members Professor. Sihong Wang and Professor. Allison Squires and my team's colleagues Dr. Hanxing Zhang, Dr. Chao Zhang, Dr. Gengjie Jia, Dr. Atif Khan and Dr. Yanan Long for their insightful advice.

Last but not least, I give my sincere appreciation to my parents, Yihong Shi and Haihua Hu, also my beloved girlfriend, Yawen Deng, for their unconditional support and understanding on my entire journey, without them I will never accomplish my goal and become who I'm today. Their companionship and encouragement provided the most pleasant harbor for me.

ABSTRACT

The modern medical data deluge accelerated when the vast amount of medical information gathered and stored by electronic sensors became widely available. Medical data are complex, heterogeneous, and continue to rapidly accumulate in electronic databases, therefore, data-driven statistical learning techniques have the potential to drastically improve clinical care by anticipating clinical complications and suggesting interventions.

This dissertation investigates the application of an assortment of statistical learning techniques to extract instructive patterns from raw medical data. Chapter 1 provides a brief overview of current statistical learning methods. We also examine both the limitations and the opportunities for state-of-the-art developments in medical forecasting. Chapter 2 introduces a project that began as a mere conjecture formulated by an endocrinologist but developed into a large-data analysis of linked pathogenesis, linking pancreatitis and type 2 diabetes mellitus. Chapter 3 describes a study in which we collaborated with a gerontologist interested in predicting cognitive decline in senior patients. In this study, we attempt such predictions by using accelerometry data collected from Chicago's south side community and implementing advanced machine learning methods for predicting patients' future clinical trajectories. In Chapter 4, we identify the novel, hip fracture risk factors and investigate whether statistical survival analysis could improve upon existing tools' accuracy. In Chapter 5, we constructed a state-of-art machine learning tool on fracture detection on patients’ broad prior disease history. Lastly, Chapter 6 summarizes the above projects and suggests future directions for our exploration of statistical learning from complex medical data. We also discuss our studies' potential importance for statistical learning from medical data and outline the problems that remain open in the field.

CHAPTER 1 INTRODUCTION

Continuous developments in the fields of statistics, biomedical, and computer science have become essential for every field of studies, including medicine. The explosive accumulation of diverse multi-modal data has made the use of computational tools both necessary and possible. In particular, medical data is intrinsically complex, heterogeneous, and normally requires a highly trained professional to interpret. Even professional physicians are unable to cope “manually” with this data avalanche. To extract hidden patterns from the raw data, researchers have had to develop highly-advanced statistical and computer science techniques [1, 2].

Long before the dawn of the “big data” epoch, researchers have attempted to apply machine learning algorithms to medical data analysis [3]. These earlier studies relied heavily on human expertise because data for automatic learning were scarce and often unavailable digitally. In addition, the computational infrastructure required for data-intensive machine learning, especially from massive image datasets, has only appeared relatively recently. Finally, recent advances in machine learning methodology have now made practical applications that were in the realm of science fictions just a decade ago. Such break-through applications include algorithms which exceed humans' ability to play games of strategy, such as checkers, chess, and Go. Modern deep learning generative models, such as OpenAI’s DALL-E, can generate stunning images from text inputs; others can provide automatic text summarization using long recordings of human discussions. The origin of these advanced methods can be traced back to the early 1990s, when researchers invented the basics of symbolic- [4], statistical- [5] and neural network-based learning [6].

The introduction of machine learning algorithms has produced promising results and have been applied in numerous medical areas, e.g. oncology [7], cardiology [8], gynecology [9], and psychiatry [10]. The wide range of machine learning applications in medical areas can bring multiple benefits to health professionals. Machine learning algorithms' pattern recognition capability in multi-modality medical data (X-rays or MRI scan) can be used to develop increasingly efficient and reliable tools to help doctors to diagnose quicker and more accurately. Besides improving diagnosis, deep learning-based methods allowed automation of drug design and discovery, speeding up development of new therapies. For instance, to better analyze clinical trial data to identify drug side-effects that, though previously unknown, could help improve patient care and safety during medical procedures. The positive impact of machine learning in health care is already noticeable, but there is enormous untapped potential at its current, early stage. As clinical data sets continue to grow, the significance of machine learning applications in health care will become crucial. Here, we will further discuss the availability of different data modalities and the current stage of machine learning algorithms that specialize in that domain.

Today, the increase of volume and diversity in electronically stored data's has combined with the invention of efficient graphical processing units (GPUs), exponentially increasing computer speed, and the array of processor availability to renders medical machine learning practical. However, medical data is recorded in many different modalities: tabular electronic medical records, medical images, biomedical analog and digital signals, and clinical unstructured text. Correspondingly, various machine learning algorithms have been designed to extract underlying patterns from every type of raw data input to accomplish a desired clinical outcome. While no dataset is perfect (due to missing data, biases, and noise), robust machine learning algorithms are nonetheless available today.

Structured electronic medical record data is collected by health care providers and insurance companies. Such data comprises patients’ demographic information, diagnoses, laboratory results, procedures, medications, and even family structure [11]. Stored in relational databases, structured data has been used for statistical analysis since electronic medical systems' inception. Some recent examples of medical machine learning are as follows: Zheng, et al. [12] used a support vector machine methodology to identify patients with type-2 diabetes mellitus; Maryam, et al. [13] proposed using an ensemble of survival machine learning models to predict heart failure from electronic medical records, and similarly; Wong, et al. [14] has studied delirium prediction. Though tabular data is the primary data storage method in the health industry, quite a few challenges remain. The most significant challenge when using tabular data is dimensionality; because tabular data commonly includes a large number of features, many of which may not be useful for machine learning algorithms' prediction task. Additionally, the rarity of some diseases in the general population creates a heavy class imbalance, making it difficult for machine learning algorithms to accurately classify diseases. This requires researchers to carefully build metrics to evaluate model performance.

Medical imaging is critical to clinicians' diagnosis routines in various fields -- including radiology, dermatology, and cardiology [15]. The use of feature-based computer vision methods in medical image analysis has been limited due to its labor-intensive design and less-than-stellar practical performance. However, breakthroughs in deep learning -- especially the last decade's introduction of convolutional neural networks, combined with a jump in GPU performance, opened the way for an explosion of successful applications in the field of medical imaging [16]. Popular applications of state-of-the-art computer vision models for medical images span a spectrum ranging from simpler technical tasks, such as image segmentation and feature detection, to more advanced tasks, such as anomaly detection and automated diagnostics. A few concrete examples of such applications are as follows: Dhungel, et al. [17] used convolutional neural networks with mammography images for low-cost and massive breast cancer screening; Zhen, et al. [18] showed promising results when applying deep learning to heart ventricular volume estimation without segmentation; Wang, et al. [19] applied a hierarchical model composed of convolutional neural network variants to accomplish high-performance results on vessel image segmentation, and; Gulsan, et al. [20] utilized a deep learning model to detect diabetic retinopathy and diabetic macular edema, trained with over 100,000 retinal images. Clearly, the number of successful applications in medical image analysis is going to increase dramatically in the near future.

Another important data modality is discrete time-stamped signals collected by electronic sensors, such as smart watches which monitor their owners' multiple vital functions. These discrete time-stamped datasets contain rich information that can be used for forecasting clinical trajectories for a wide range of diseases, such as cancer, Alzheimer’s, cardiovascular, and COVID-19. [21]. Though time-series data can be challenging to model, with its high noise-to-signal rate and varied time spanning, researchers have successfully developed robust applications dealing with such data. [22]. As in the previous paragraphs, we will give a few examples of computational applications using time sequence signals. Luca, et al. [23] performed an experiment that recruited participants’ movements, using six wearable devices, then proposed a machine learning algorithm to detect Parkinson’s disease. In another study, Yildirim, et al. [24] used a new architecture of neural networks, called "recurrent neural networks," to achieve state-of-the-art performance in multi-class arrhythmia classification. Finally, Prabhjot, et al. [25] used blood pressure management and heart rate data, collected by smartphones, to build hybrid models that combined Bayesian networks and heuristic techniques.

Clinical unstructured texts, such as discharge summaries, are another major data modality used to capture patients' health information. The text mining techniques researchers use on these texts originated in 1958, when Hans Peter Luhn [26] used word frequency statistics to categorize documents. Many significant breakthroughs in NLP technology have occurred since, and text analysis and natural language processing (NLP) have undergone major progress in just the last decade. Nowadays, a large language model, using a deep learning architecture such as Transformer, can successfully handle major practical tasks, such as text classification and machine translation [27]. Clinical texts associated with test results, discharge narratives, and descriptions of both the therapeutic and adverse effects of medications, inspired the development of revolutionary language models. In one study, Choi, et al. [28] used the popular word2vec [29] method to represent unstructured medical codes and to train a model to predict heart failure. In another study, Miotto, et al. [30] used auto-encoders to develop a patient-level representation that showed significantly better performance in forecasting chronic disease for the next year. Even though some characteristics of health domain text data (including lack of standardization, limited data availability, and privacy) makes it more challenging to utilize language models in this area, the recent development of large language models (ChatGPT etc) can empower researchers to investigate its potential on various tasks, such as patient chatbots that can understand patients' symptoms and provide appropriate medical advice or connect them with the appropriate health care professionals.

We have provided a cursory review of significant events in developing machine learning algorithms in medicine, covering diverse data modalities across clinical application areas. This dissertation is motivated by the philosophy that novel modeling paradigms and modern computational infrastructure allow us to maximize our imaginations when building models -- regardless of heterogeneous datasets. Yet, we must acknowledge that models alone are inadequate to guarantee successful projects -- both the availability of high-quality, large medical data sets and clinicians' empirical insights play an essential part as well. Collaboration between data scientists and clinicians is essential to the successful implementation of machine learning algorithms in health care settings. By leveraging machine learning algorithms' clinical expertise, we can improve patient outcomes and revolutionize the health care industry. In summary, this dissertation emphasizes the importance of combining machine learning algorithms with high-quality medical data and clinical expertise to achieve significant progress in the health industry. In the following chapters, we will introduce three studies which were inspired by problems suggested by clinicians. Our studies included disease association, disease forecasting, and survival analysis.

CHAPTER 2 COHORT ASSOCIATION STUDY OF ACUTE PANCREATITIS AND DIABETES MELLITUS THROUGH ELECTRONIC MEDICAL DATA

2.1 Introduction

The relationship between diabetes mellitus (DM) and pancreatitis is complex and bidirectional. While chronic pancreatitis (CP) has been well described as a risk factor for pancreatogenesis DM, studies that focused on the development of DM following acute pancreatitis (AP) have become popular. Numerous environmental and genetic factors, including consuming alcohol, smoking, co-morbid dyslipidemia and predisposing genetic variation, are known to put patients at risk for pancreatitis, which causes pain and morbidity from multiple complications. Also, the occurrence of DM after AP varies between 14 to 43 percent with a higher incidence with in those with severe AP in comparison to milder episodes [31-32]. Additionally, the risk for patients diagnosed with AP exposed to DM persists beyond the time frame for the first year, with patients retaining a greater than two-folded risk of developing DM even after five years [33]. Although the exact pathophysiology of the development of DM after AP remains unknown, as only a minority of these cases require insulin, which suggesting that the underlying pathophysiology may be similar to Type-2 DM (T2DM) [33]. However, the presence of DM is associated with poor outcomes in AP, DM shares common risk factors with AP as well as its treatment [34]. In this case, numerous studies have highlighted the health disparities that exist in screening and care of patients with T2DM [35]. Underserved communities with high-risk exposure to T2DM are less likely to receive optimal treatment and achieve glycemic control. For example, a recent study indicates that African Americans and Asian were more likely to have uncontrolled HbA1c values as compared to Whites [40]. Besides, as shown in a 2009 study, African American race was an independent risk factor for mortality related to AP [38]. Similar, the disparity in care situation can also be identified for patients with AP, such as the emergency room wait times have been found longer and rates of inpatient procedures or cholecystectomy for AP lower in non-White races [39]. The intrinsic factor such as gut microbiome may have contributed the development of complications related to DM, acute and chronic pancreatitis according to emerging studies discovery [36-37].

An estimated 220,000 Americans will have an episode of AP in yearly base. Based on the cited data above, many of these will develop DM because of AP or have pre-existing DM that may affect outcomes related to it. To further investigate this link and the reasons for disparities in underserved communities, this proposed project aims to perform retrospective and prospective longitudinal observational studies of our racially, ethnically, and geographically diverse population of subjects on the South Side of Chicago for the occurrence of diabetes before, during and after an episode of acute pancreatitis, and studying the outcomes of those patients with a focus on underserved communities.

The project is innovative both conceptually and methodologically. While the development of DM after AP has been described that pathophysiology remains unknown as well as lack of data on this relationship in underserved communities. This project has involved a multidisciplinary team comprising of medical pancreatology and gastroenterology, endocrinology and data science that’s equipped to answer the questions relevant to this project. We have facilitated identification of risk factors predict the development of diabetes and will leverage innovative and state-of-art approaches. In the meantime, we have compared these risk factors with national claims database that allow us to identify sequential appearance of AP and DM in individual patients, with individually measured genetic risk factors and environmental exposures. Overall, the goal of this study is to determine the incidence and risk factors for acute pancreatitis and diabetes in the population on the South Side of Chicago. Compare local estimates with national factors using insurance claims datasets.

2.2 Methods

2.2.1 Data and other materials

The use of massive data sets is a methodologically and conceptual innovative approach to the problem of acute pancreatitis (AP) – diabetes mellitus (DM). These data sets can be mined to be provide extraordinary insights into the frequency, location, chronology, and treatment of these diseases. It can also be analyzed in multiple ways to understand data integrity, coding errors and overlap, and then to provide insight into local prospective datasets.

Our access to electronic health record (EHR) to analysis including two resources. (1) The **University of Chicago Electronic Medical Record** contains patient data on nearly 2.5 million patients seen at University of Chicago Medicine in the last 10 years alone. Further data retrieval can also be requested through the Clinical Research Institute (CRI). This dataset will be used to collect internal data on patients seen within the University of Chicago health system. (2) **The MarketScan® database** (IBM Watson Health 2019) was originally complied by Truven Health from records of over a hundred large insurance companies in the US. The data set we intend to use contains 5 billion records of diagnosis for over 180 million unique patients in the US. For each entry of diagnosis, the database documents the date, age of patient, and a code in International Classification of Diseases, 9th or 10th Version's Clinical Modification (ICD-9-CM or ICD-10-CM). MarketScan data allows linking patients into families (through insurance policies), which allows us to disentangle genetic and environmental information.

2.2.2 Statistical Analyses

Taking advantage of well-formatted MarketScan diseases ICD groups, we're able to summarize the cohort association study on Acute Pancreatitis (AP) and Diabetes Mellitus (DM). There are multiple methods on identifying non-causal association measurements to quantify the statistical significance of two groups. In this we will utilize three types of widely used measurements on exploring the statistical significance between AP and DM in MarketScan electronic health records.

1. **Odds Ratio (OR):** The odds ratio is a measure of association used in case-control studies. It compares the odds of exposure among cases to the odds of exposure among controls. An OR greater than 1 indicates a positive association, meaning that the exposure is more common among cases than controls.
2. **Relative Risk (RR):** The relative risk is a measure of association used in cohort studies. It compares the risk of developing a disease or outcome among those exposed to a risk factor to the risk among those not exposed. An RR greater than 1 indicates a positive association, meaning that the risk of the outcome is higher among the exposed group.
3. **Logistic Regression (LR):** Logistic regression is a statistical model used to analyze the relationship between a binary dependent variable and one or more independent variables. It estimates the probability of the dependent variable given the values of the independent variables. The coefficients of the independent variables provide information about the strength and direction of the association between the variables.

The methodology for this project in order to investigate domain expertise inspiration on AP effect on following DM, we firstly extracted patients' electronic health records information based on AP and DM ICD codes filtered by domain experts. The electronic health records would contain detailed information include:

- Demographics information (gender, race, year of birth, etc)
- Time-stamped disease history (ICD 9/10 encoded)
- Time-stamped medical history
- Insurance enrollment history

The abundant patients’ information listed above enable us to query the desired data to create detailed table for case-control groups statistics computation. The case-control groups OR was firstly computed to identify the general association strength on multiple sub-categories of AP and DM occurrence. To further evaluate the degree of association between varies AP and DM subcategories pair, we applied the multiple hypothesis testing besides OR value to analyze the p-value with threshold that determine the statistically significant AP and DM subcategories pair. Besides, to avoid the numerically over-estimated False-Discovery-Rate (FDR) which is a common issue in multiple hypothesis testing with a fixed threshold, we have applied Benjamini-Hochberg (BH) method to carefully tune the threshold according to sample size. After filtered some desired pairs of AP and DM with statistically significant signals, we turned to RR as a measurement to quantify the risk of developing DM for patients exposed to AP in scale of years. We obtained all the RR values and estimate the effect of AP as risk factor in collaboration with domain expertise. In the end, to evaluate other possible risk factor for developing DM after exposed to AP, we have used the power of linear model to incorporate other indicator variables such as gender, race and age and built logistic regression model to investigate relationship between difference dependent variable (DM within 5 years) according to all the potential risk factors.

Variables definition

1. *D*: The disease code bag represented by its ICD-9-CM/ICD-10-CM code. Here, $D_{\text{AP}} \subseteq$ {K85, K89}, $D_{\text{DM}} \subseteq$ {249, 250, E08, E09, E11, E12, E13}.
2. *G*: The gender of patients. $G \subseteq$ {male, female}.
3. *Z*: Patients cluster divided by an age gap of 10 from 20 ~ 60 years old. $Z \subseteq$ {1, 2, 3, 4}.
4. *Pr*: Probability.
5. *p*: P-value to determine the likelihood of AP and DM association signal.
6. α*:* Statistically significant threshold.

The first step is to build case-control table through querying entire patients' electronic health records from MarketScan dataset. The data mining process was operated via the Structured Query Language (SQL) schema, a programming language that designed for managing and manipulating data fast and efficiently. Besides, during the data mining process, our team experts in the field of diabetes mellitus noticed that only a minority of patients with DM after AP require insulin as initial treatment, which suggesting that the underlying pathophysiology may be similar to Type-2 diabetes. Thus, we will maintain caution with data and attempt to validate coding of diseases by multiple test parameters, including analysis of overlapping codes, use of specific medications such as medications used for treatment of DM including insulin. The formatted case-control table was represented as below on Table 2.1. After that, we built contingency table (Table 2.2) based on queried case-control data to further calculate desired statistics introduced above such as OR, RR and p-value.

Table 2.1 Case-Control Representative Table

ICD	*K85*		*K89*	
	Case	**Control**	**Case**	**Control**
249	a_1	b_1	a_1	b_1
250	a_2	b_2	a_2	b_2
E08	a_3	b_3	a_3	b_3
….	…	…	…	…
Total	a_t	b_t	a_t	b_t

* Representative case-control table example. Case: patients with DM that exposed to AP; Control: patients with DM that not exposed to AP.

Table 2.2 Contingency Representative Table

	K85	
	Case	***Control***
$\boldsymbol{D_{DM} \subseteq \{249\}}$	a_1	b_1
$\boldsymbol{D_{DM} \not\subset \{249\}}$	$c_1 = a_t - a_1$	$d_1 = b_t - b_1$

Odds Ratio formula.

$$OR = \frac{a * d}{b * c}$$

Relative Risk formula.

$$RR = \frac{a \,/\, (a + b)}{c \,/\, (c + d)}$$

Hypothesis Testing. To build a model on hypothesis testing, we need first formulate null hypothesis (H_0) as well as alternate hypothesis (H_a), then compute both p-value through chi-square test and threshold α together to determine whether we should reject or accept the null hypothesis. In this project, the model is formulated as shown below:

$$H_0\text{: } D_{\mathrm{DM}} \text{ is independent from } D_{\mathrm{AP}}.$$

$$H_a\text{: } D_{\mathrm{DM}} \text{ is not independent from } D_{\mathrm{AP}}.$$

According to case-control table (Table 2.1) we can formulate chi-square test to calculate the p-value for each D_{DM} and D_{AP} pair. As discussed above, as we conducted multiple hypothesis testing due to varies combination of D_{DM} and D_{AP} pair, BH correction was chosen as regularization method to control FDR. To be specific, it works by ranking p-values obtained from multiple hypothesis tests in ascending order, then p-values are adjusted based on the rank and total number of tests conducted. The corrected p-values are being used in compared to the threshold α to determine if the null hypothesis H_0 should be rejected. If corrected p-value p if less than threshold value α then null hypothesis H_0 is rejected that DM is independent of AP and vice versa.

$$\{p\} \sim \chi_2(< D_{\mathrm{DM}}, D_{\mathrm{AP}} >)$$

$$\{p_{corr}\} = \mathrm{BH}(\{p\})$$

$$p < \alpha \rightarrow reject\ H_0 : D_{DM} \perp\!\!\!\perp \mathrm{D}_{\mathrm{AP}}$$

Logistic Regression. As we have acquired the statistically significant evidence on DM and AP association study via statistical tools discussed above, the next step is to build a logistic regression linear model. The main benefit of using logistic regression linear model to estimate odds ratio in contrast to traditional methods with contingency table is that logistic regression allows inclusion of multiple predictors in the same model thus we're capable to account for more potential confounding risk factor that take places in the following development of DM after patients exposed to AP. Besides, logistic regression provides estimates of the odds ratio along with confidence intervals, which indicates the precision of the estimates. Therefore, the MarketScan have provided us with rich patient's information that we can include the candidate factors such as age, gender etc that might have hidden effect on DM development. The computed coefficients in logistic regression can be used to derive OR on specific potential risk factors as shown in equations below. Besides, to build a logistic regression model based on enormous MarketScan dataset can enable us to obtain a robust tool that can be used to estimate the odds ratio with their confidence intervals of developing DM based on a set of predictor variables (risk factors) conditions.

$$z = \beta_0 + \beta_1 * \mathbf{1}(D_{AP}) + \beta_2 * \mathbf{1}(G) + \sum \beta_i * \mathbf{1}\,(Z_i)$$

$$Pr(D_{DM}) = \frac{1}{1 + exp(-z)}$$

$$OR_i = \exp(\beta_i)$$

$$Odds(D_{DM}) = \frac{Pr}{1 - Pr}$$

So far, we have proposed all statistical tools that being utilized in this project to fully investigate the DM that developed following patients' exposure to AP.

2.3 Results

Regarding the University of Chicago electronic health record that contains patient's diseases history information for local communities, we have examined and analyzed data on both the number and location of pancreatitis cases. The total number of diagnoses of "acute pancreatitis" between 2016 and 2020 was 474. The average age was about 56 years of age with an approximately a normal distribution. Based on this we were able to determine the number of patients who subsequently had ICD-9/10 codes of type 1 diabetes (84) type 2 diabetes (217), both (75) (suggestive of miscoding) or neither (173). We constructed a heat map of these patients (Figure 2.1), based on their zip code, which shows that the majority of these patients appear to be located on the South side of Chicago, thereby indicating that patients in this area may be especially vulnerable to this combination of diagnoses (AP and DM).

After analyzing the spatial distribution of DM development after patients' exposure to AP in Chicago communities. We further explore the electronic health records stored in MarketScan

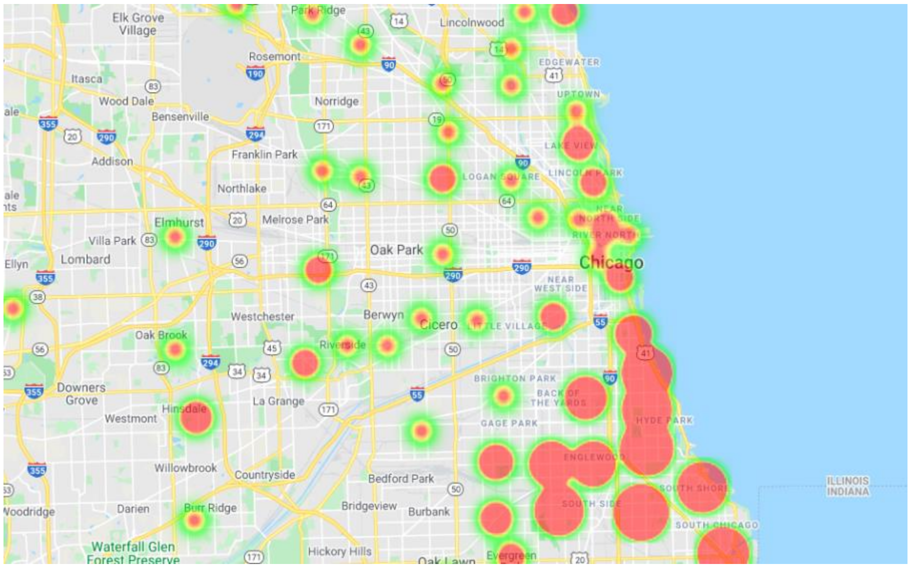


Figure 2.1. Based on information from the University of Chicago Medicine EMR (EPIC), this map of the Chicagoland area shows those who are diagnosed with acute pancreatitis and develop subsequent diagnosis of diabetes within our patient population based on patient zip code.

dataset for national wise patients' disease history to evaluate chronological separation of pancreatitis and diabetes as well as temporal variations with help of this massive dataset.

As a preliminary overview, we have sampled patients exposed to AP in their diseases history to analyze the chronical effect of developing DM afterwards. The distribution of time in developing DM in yearly scale is displayed in the pie plot we created (Figure 2.2). The total sample size in this experiment has 33,980 ppl that have AP occurrence before any DM identified, based on which we can observe ~25.9% patients would develop varies form of DM following occurrence of AP. Among that population, ~21.5% patients would have developed DM within next 5 years and only ~4.4% patients would develop DM after 5 years window.

Based on the pie chart shown in Figure 2.2, we can gain a general understanding of the proportion distribution of patients who were exposed to AP prior to being diagnosed with DM in 5 years cutoff. The results of this experiment are consistent with the initial hypothesis of domain experts, who predicted that approximately 20% of patients would develop DM following AP. However, in order to gain a more detailed understanding of the development of DM after an AP

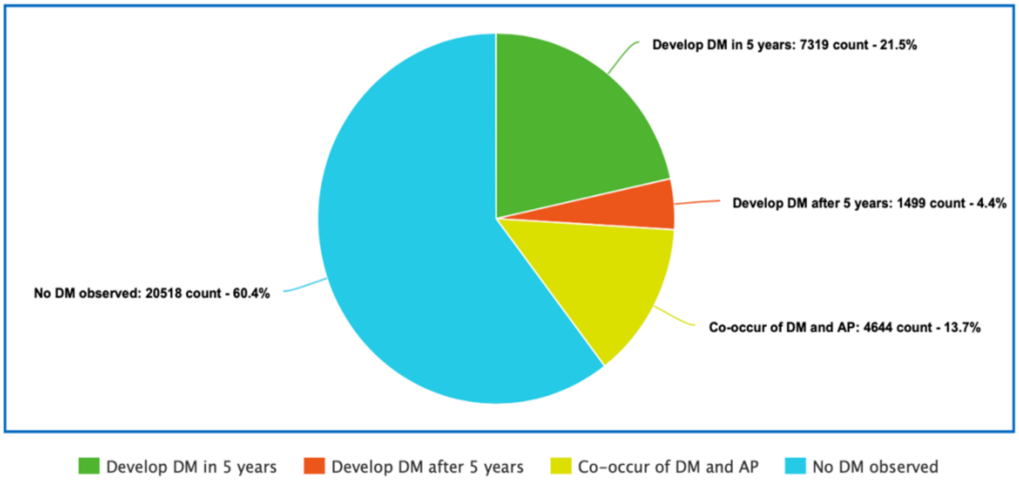


Figure 2.2. Based on information from MarketScan, this pie plot shows the distribution of time to develop diabetes mellitus for those patients who have diagnosed with acute pancreatitis ahead of existence of diabetes mellitus in disease history record. The total sample size is 33,980 ppl.

diagnosis, we must gather more specific information on the observation of DM. This information should not only be collected on a yearly basis, but should also include details about the type of DM that developed (i.e. primary Type-1 diabetes and Type-2 diabetes). Therefore, we have created a cumulative distribution plot (Figure 2.3) on DM development in yearly basis beginning with AP exposure and end with 15 years after that. Here, we focused specifically on the primary type of DM development, as shown in Figure 2.3, ~78.23% DM (including Type-1 and Type-2) can be observed in the first 5 years after occurrence of AP. Specifically, the majority of DM developed is Type-2 DM (~82.5%), among which ~76.94% would be observed in the first 5 years. For Type-1 diabetes, ~84.36% cases would occur in the first 5 years following AP diagnosed.

To control the falsely encoded ICD codes for diabetes in MarketScan dataset that might cause mistakes on our analysis conclusion. We have attempted to extract patients' medication history apart from encoded ICD code for diabetes. Dr. Louis Philipson, a co-investigator in this

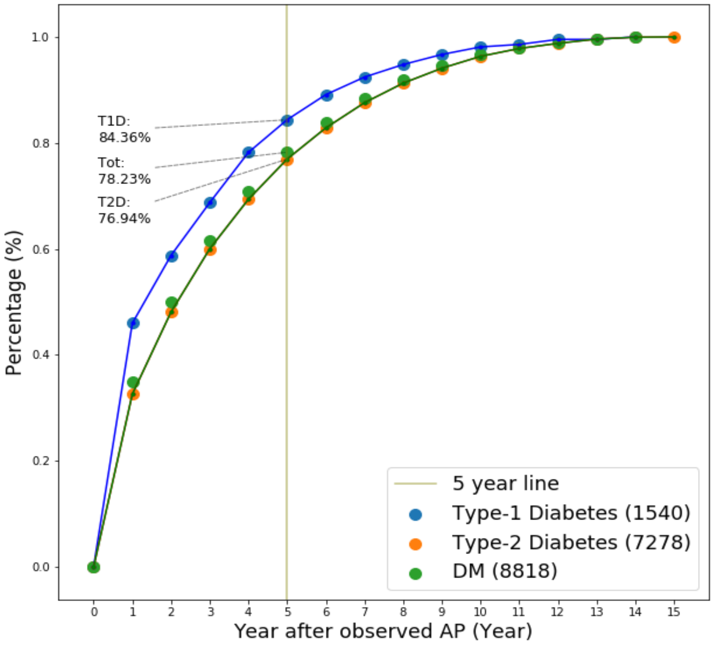
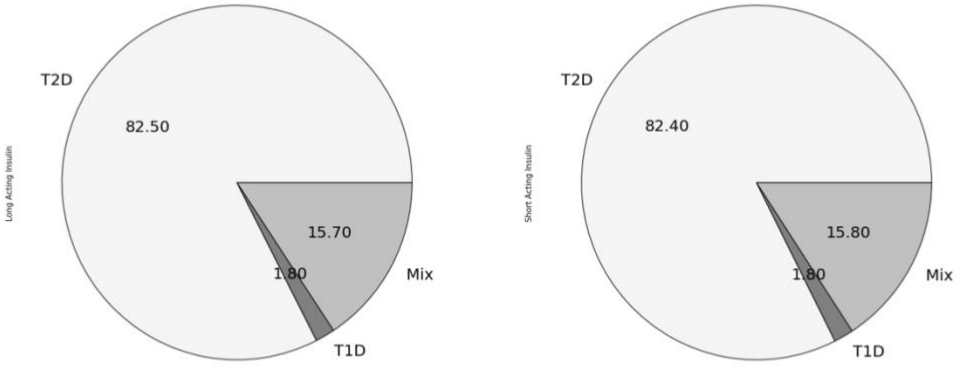


Figure 2.3. Cumulative distribution plot of patient's specific diabetes mellitus development after diagnosed with acute pancreatitis in year 0. Only primary type of diabetes mellitus (blue dot: type-1 diabetes; orange dot: type-2 diabetes) is included with total sample size of 8818.



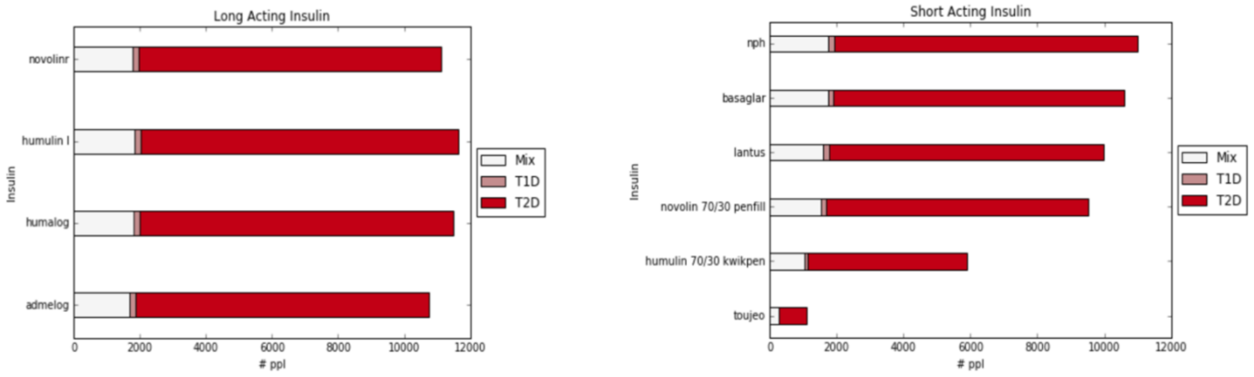
(a) LA Insulins 11640/11680 ppl

(b) SA Insulins 11673/11680 ppl

Figure 2.4. (a) Long acting (LA) and (b) short acting (SA) insulin distribution for patients diagnosed with three types of primary diabetes mellitus (T1D: Type-1 diabetes; T2D: Type-2 diabetes; Mix: T1D and T2D both observed). The entire sample population is 13,592 patients with diabetes mellitus, among which 11,680 (~86%) population have medication history prescription of insulin usage.

project who is an endocrinology and leading world authority on DM, recommended to identify the patient’s prescription on insulin as indicator for varies types of DM. Here we presented an experiment to compare the usage of long/short acting (LA/SA) insulin for DM patients that we’ve separated in three categories: Type-1 diabetes, Type-2 diabetes and Mix diabetes (both type-1 and type-2 diabetes observed in disease history), the result is presented on Figure 2.4.

Dr. Louis Philips provides a list of prescriptions for long acting and short acting insulin after extracting a patient's medication history dataset. For instance, short acting insulin category includes keywords such as "Humulin", "Humalog" and "Admelog" whereas "Basaglar", "Lantus" and "Levemir" are part of the long acting insulin category. Despite this classification, our analysis of Figure 2.4, which sampled 13,592 patients with a history of DM, did not reveal any significant difference between long and short acting insulin prescriptions for the diabetes types listed above. In order to investigate the effects of long and short acting insulin on primary DM, we analyzed the usage of specific insulin prescription types for each category, and the results are presented in



(a) Long Acting Insulin

(b) Short Acting

Figure 2.5. (a) Long acting (LA) and (b) short acting (SA) specific insulin prescription distribution for patients diagnosed with three types of primary diabetes mellitus (T1D: Type-1 diabetes; T2D: Type-2 diabetes; Mix: T1D and T2D both observed). The Y-axis list specific insulin prescription name for LA/SA insulin categories.

Figure 2.5. Our findings suggest that there is no single insulin prescription that dominates usage for any specific type of diabetes.

Table 2.3 Risk Of Diabetes Following Pancreatitis Using Marketscan Database.

Diabetes Type	**Case (AP)**	**Control (No AP)**	**Odds Ratio**	**p-value**	**Relative Risk**
Type 1 DM	**167**	**566**	**55.290**	**<1e-10**	**47.17**
Type 2 DM	**466**	**4,925**	**17.731**	**<1e-10**	**12.21**
Other specified DM	13	237	10.279	1.56x 10^{-9}	/
DM due to underlying condition	8	122	12.288	5.95x 10^{-7}	/
Drug or Chemical Induced DM	0	18	0	1	/
No DM	949	177,833	/	/	/

*As compared to age-, sex- and county-matched patients with no prior history of AP, patients with AP had 55 times the odds of having a diagnosis code of T1DM following AP and 18 times of having a diagnosis code of T2DM following AP.

So far, we have explored the spatial distribution of AP followed DM in Chicago local communities and shows the significant high risk for underserved communities' vulnerability to DM after exposed to AP. Also, we have thoroughly analyzed the national wise MarketScan dataset to justify that over 20% of patients who diagnosed with AP would develop DM, among which, ~80% of DM would be observed in the next 5 years for varies diabetes. We have explored the possibility to utilize the MarketScan medication history record to cross-validate the diagnosed diabetes type by accommodating with their prescription of long/short acting insulin. Next, In Table 2.3 below, we show the power of using the MarketScan data for asking important questions about the fundamental data related to this project, and the possible questions that those answer generate in turn. This analysis shows an impressive increase in risk of 55 times for those with a diagnosis

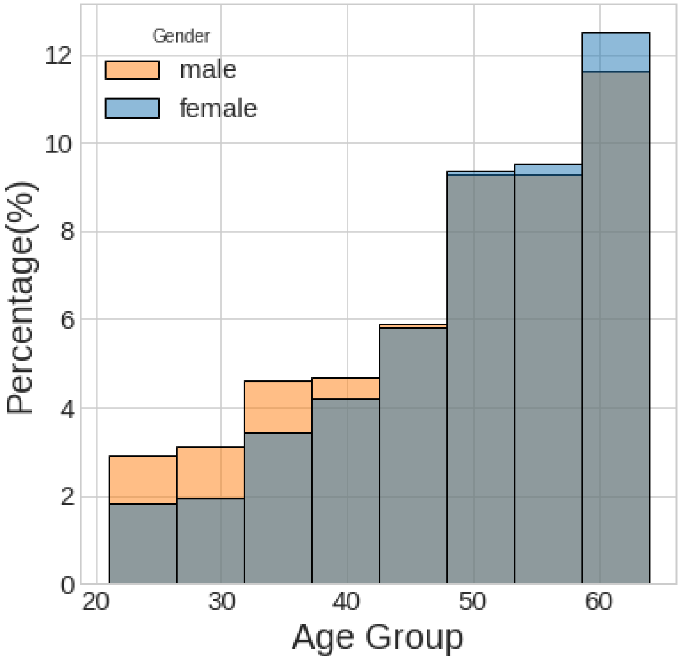


Figure 2.6. Age and gender distribution sampled from MarketScan dataset to build logistic regression model. The total sample size is 183,957.

of acute pancreatitis to then have a diagnosis coded as Type-1 diabetes, and 18 times more likely to have a diagnosis coded as Type-2 diabetes than non-pancreatitis peers in the insurance claims. Verification of this data is critical in the face of possible coding errors is currently underway, but these results are highly supported statistically with astronomically small *p*-values. The relative risk regarding Type-1 and Type-2 DM conditioned on AP are computed as well, which indicated 47.17 risk exposed to Type-1 diabetes for people exposed to AP and 12.21 for Type-2 diabetes development.

Finally, we have built a logistic regression model that consider the effect of other potential demographic risk factors to DM development apart from AP. Here we consider the effect gender and age (between 20 ~ 65 years), the age and gender distribution of sampled population is shown in bin plot Figure 2.6. As we can observe that the population age group has tendency to the right with approximately similar gender distributed in each age group. Then we trained a logistic regression model and estimated the odds ratio along with 95% confidence intervals regarding each predictor variable in the model. The plot is shown in Figure 2.7 with dashed vertical line highlighted the neutral effect odds ratio value. The Figure 2.7 shows that AP has significant contribution to DM outcome compared to potential demographic risk factor such as gender and age in this model. The estimated value for AP odds ratio is ~18 which is close to Type-2 DM odds ratio computed through contingency table method, the reason can be Type-2 DM patients dominated number in contract to other types of DM. Moreover, the age group 5 (patients age between 58-65) also presented positive effect on DM development in comparison to other age groups, and the gender has neutral effect to DM outcome.

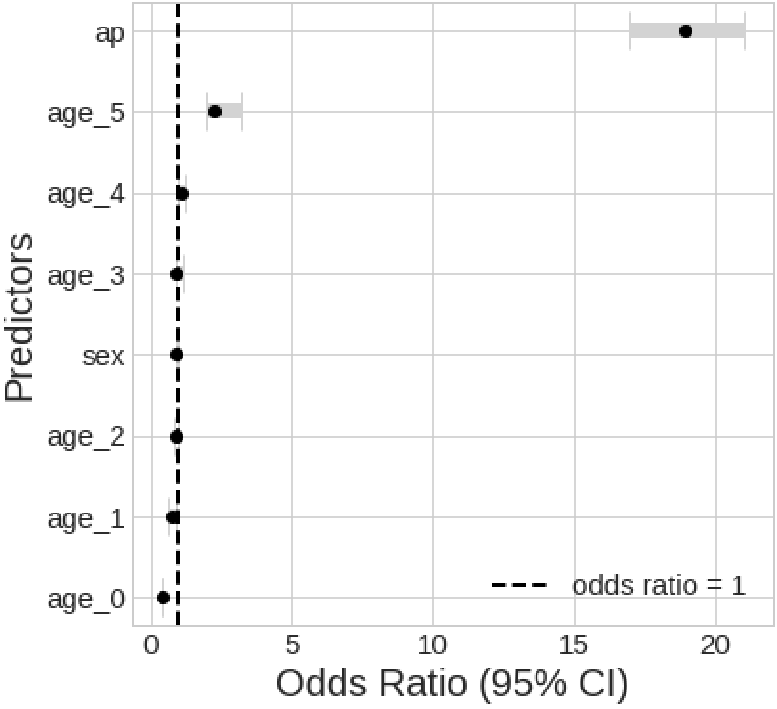


Figure 2.7. Odds ratios point plot for each included risk factors in logistic regression model, the black dot represent the estimated value for odds ratio and grey bar represented 95% lower and upper bound confidence intervals. Y-axis ticks: ap: acute pancreatitis; age between 20-65 is divided into 6 groups represented by number after underscore sign; sex is the gender factor (0: male, 1: female)

2.4 Discussion

In the first part of study, we analyzed the spatial distribution of patients who diagnosed AP and later developed DM in the subsequent years. The result represented in form of heat map in Figure 2.1 that heavily concentration of cases in south side communities of the Chicago area, which predominately occupied by African American. This finding, extracted from University of Chicago Medicine electronic health record, confirms our proposed hypothesis that both DM and AP are further amplified in underserved communities. Even though the underlying causal relationship between environmental factors in underserved communities and the amplified development of DM after AP is not yet clear, we plan to explore these disparities including any factors beyond socioeconomic status, by studying the microbiome in patients with AP and DM in the future study.

Next, we studied the temporal factor of AP on DM development using a rich nationwide insurance claims MarketScan dataset. We initially set a 5-year window to study the proportion of patients who were diagonalized with AP and subsequently developed DM during that period, the 5-year window time period was based on empirical observation in clinics and supported by our teams' domain expertise. The results, displayed in Figure 2.2, confirm our initial observation, suggesting that ~20% of patients who diagnosed with AP would develop DM in the next 5 years. To explore the distribution of various types of primary DM among patients exposed to AP, as well as the cumulative percentage of DM development on a yearly scale, we constructed a cumulative percentage plot in Figure 2.3. From Figure 2.3, we can observe that approximately 80% of both Type-1 and Type-2 DM cases would occur within the first 5 years after AP. Type-1 DM has a higher percentage of occurrence (~85%) in that time window compared to Type-2 DM (~77%), although Type-2 DM is more commonly observed among the total sampled patients diagnosed with AP (~82%).

Since the patients' disease history is encoded using ICD-9/ICD-10 codes stored in the MarketScan dataset, which may have the risk of miscoding, and there may be co-occurrence of encoded Type-1 and Type-2 diabetes ICD codes in the "Mix" category if we only query the disease history through MarketScan. Therefore, based on advice from Dr. Louis Philipson, a leading world expert in diabetes, we minimized the risk of miscoding by cross-checking the patients' medication history in addition to the ICD-coded disease history, in order to better distinguish between Type-1 and Type-2 diabetes from the "Mix" category. To be more specific, Dr. Louis Philipson provided a list of popular insulin prescription names that are commonly used for patients diagnosed with DM, which can be categorized into short-acting and long-acting insulin with different effects on different types of DM. The exploratory results obtained from a sample of 11,680 patients with DM are shown in Figure 2.4 and Figure 2.5. Based on Figure 2.4, we can observe that the difference in the distribution of primary diabetes usage between short-acting and long-acting insulins is insignificant (~82% for T2D, ~16% for Mix, and ~2% for T1D). Further study on specific insulin prescriptions listed in Figure 2.5 did not reveal any meaningful differences in their usage for each type of diabetes.

We then constructed a case-control table, similar to Table 2.1, based on MarketScan electronic health records for patients with no prior history of acute pancreatitis (AP), matched by demographic features such as age group and sex. Subsequently, we built a contingency table, as shown in Table 2.3, for various types of diabetes outcomes, with the count of patients in each cell extracted from the MarketScan dataset based on specific ICD-9/ICD-10 codes for diabetes. Using the contingency table results, we computed the odds ratio (OR) with associated p-values and relative risk to estimate the risk factor of AP in contributing to DM development.

The computed odds ratios for each type of diabetes indicated a positive effect of AP as a risk factor for future development of diabetes. Specifically, patients diagnosed with AP were approximately 55 times more likely to develop Type-1 diabetes and 18 times more likely to develop Type-2 diabetes compared to patients without a prior history of AP, as supported by the extremely small p-values. We also computed the relative risks for primary Type-1 and Type-2 diabetes, which showed that patients exposed to AP as a risk factor were approximately 47 times more likely to develop Type-1 diabetes and 12 times more likely to develop Type-2 diabetes compared to unexposed patients. Relative risk is preferred when the outcome is not a rare disease, such as Type-1 and Type-2 diabetes in our case, as it provides a more intuitive interpretation of the strength of association between the case and control groups. However, odds ratio is more commonly used when the outcome is rare, as it is less biased than relative risk in such situations.

Furthermore, we applied a logistic regression model to consider potential confounding factors apart from AP existence in DM development. The sampled demographic distribution is shown in Figure 2.6, with the training dataset being the same population sampled from the MarketScan dataset to construct contingency table 2.3. The demographic predictors included in the logistic regression model, as shown in Figure 2.6, indicated that the sampled population had approximately equal percentages of gender, but the age distribution was skewed towards the 50-65 age group, which is common in the MarketScan dataset. The logistic regression model was built upon this population, and the results displayed in Figure 2.7 indicated that the estimated odds ratio for AP had a significant effect on DM development compared to age and gender factors. The confidence intervals of the estimated odds ratios in Figure 2.7 showed that only age group 5 and AP had a significant positive impact on DM development, while other factors had a neutral effect, as their 95% confidence intervals included the neutral odds ratio value (dashed vertical line). This experiment emphasized that AP has a significant impact on DM development even after considering potential confounding predictors.

In conclusion, this project thoroughly analyzed both the spatial and temporal distribution of patients diagnosed with AP and their subsequent development of DM. The experiments results support the hypothesis inspired by domain expertise empirical observations in clinics. This study contributes to the growing body of literature on the association between AP and DM development and underscores the importance of addressing health disparities in underserved communities. It’s important to support further research in this area for early identification and management of subsequent DM development in patients with prior history of AP, especially in vulnerable elderly populations.

2.5 Limitations

This project suggests the AP significant positive effect on subsequent DM development initiated by clinician's empirical observation, however, there are two main caveats accompany this assertion.

First, it's important to be aware that diseases are not truly independent of each other, the existent of one disease can facilitate the development of another, and other factors such as environmental, demographic and medication history may also lead to the exacerbation of chronic disease. Therefore, though we have attempted to use logistic regression model to take demographic factors (gender and age) into consideration when compute strength of association between prior AP and subsequent DM, the size of included potential confounding variables are notably limited. For instance, as we have observed significant biased spatial distribution of DM following AP diagnosed in local Chicago underserved communities, the possible confounding factors such as zip code or district air pollution conditions can have significant contribution to trigger DM development apart from AP. Note that low-p values and narrow confidence intervals are driven by large datasets. Moreover, our analysis cannot distinguish the causal relationship between AP and DM, this part will require research on future study.

Second, as noted in the discussion, it's possible that diseases coding errors could influence our odds ratio estimates. The large proportions of "Mix" category indicates the confusion to distinguish between Type-1 and Type-2 diabetes through ICD-9/ICD-10 codes in MarketScan. This limitation of accuracy on disease history recording would impair our capability to study the difference of AP effect on specific DM category. Though we have attempted to solve this issue via cross-validate with patients' medication history on their insulin usage, the results didn't show any significant signal that we can use to distinguish between Type-1 / Type-2 diabetes. However, as we have estimated that the miscoding rate at the MarketScan is around 0.52 percent, thus the error rate is positive but small in comparison to the derived association effect size.

CHAPTER 3 FREE-LIVING WRIST AND HIP ACCELEROMETRY FORECAST COGNITIVE DECLINE AMONG OLDER ADULTS WITHOUT DEMENTIA OVER ONE- OR FIVE-YEARS IN TWO DISTINCT OBSERVATIONAL COHORTS

This chapter is adapted from the manuscript “**Free living wrist and hip accelerometry forecast cognitive decline among older adults without dementia over 1- or 5-years in two distinct observation cohorts**” authored by Chengjian Shi, Niser Babiker, Jacek K. Urbanak, Robert L. Grossman, Megan Huisingh-Sheetz and Andrey Rzhetsky.

3.1 Introduction

Alzheimer’s disease and related major neurocognitive disorders (ADRD) affect over 50 million people worldwide, with an increase of 10 million new cases per year. The ADRD disease burden is expected to increase as the world population ages [41-43]. ADRD disproportionately affects socioeconomically disadvantaged groups and minorities [44] and is associated with lower quality of life, increased mortality, care dependence, and institutionalization. Preservation of cognitive abilities and a positive mindset may maintain quality of life in later years [45]. Few US Food and Drug Administration (FDA) approved treatment options exist at this time; therefore, the mainstay of current management remains on the prevention side6. Cognitive trajectories vary widely among older adults, with recent studies showing that different races experience varying rates of decline [47-48]. Finding sensitive forecasters of early decline could trigger more frequent monitoring and aggressive preventative interventions, advance care planning, and even ADRD research study eligibility [49].

There is an acute need for easily deployed, noninvasive, clinical tools to identify cognitively intact older adults most at risk of subsequent cognitive decline. Certain clinical and environmental factors including age, gender, education, body mass index, neighborhood socioeconomic status, and history of stroke or diabetes are easy to gather clinically during a visit or even during a telephone screen [50]. In a meta-analysis, structural and functional aspects of one's social environment (including network size, social activity, and loneliness) are also predictive of cognitive decline among older adults [51]. Genetic susceptibilities, such as APOE carrier status, can improve forecast models but are more invasive for patients to collect [52].

Wearable sensors have been gaining attention for their ability to remotely collect free-living activity and sleep patterns and the association of these patterns with other important age-related conditions: frailty [53-55], disability [56], social disengagement [57], and death [58]. The relationship between free-living activity and cognitive performance has been less studied. In cross-section, greater activity volume (highest and middle tertiles of active minutes/day) was associated with better processing speed among cognitively intact adults at risk of mobility disability [59] and steps/day were associated with better executive functioning in healthy older adults [60]. Longitudinally, cognitively intact older adults with a higher percentage of moderate to vigorous physical activity (MVPA) per week had a lower risk of cognitive impairment and better maintenance of executive function and memory over an average of 3 years [61]. However, these findings were not consistent across racial/ethnic groups. A higher percent of MVPA predicted maintenance of only memory and not executive function in African American/Black adults, as compared to White adults.

Few prior studies have leveraged the high-resolution nature of accelerometer data in analyses to maximize unique pattern recognition that may differentiate health risk across individuals, a concept familiar to those studying precision medicine. While accelerometry is not currently used in routine clinical care, it has been increasingly used in major research studies to remotely assess older adult health and poses significant advantages in the era of telehealth [62-68]. Translation of accelerometry in clinical practice has been challenged by the lack of accelerometry tools with clear clinical applications and the inability to apply research findings across device body locations and manufacturers.

The objective of this study was to significantly advance the prior work on forecasting early cognitive decline among older adults without dementia by discovering prognostic, free-living accelerometry patterns using 24-h data. We considered 98 accelerometry measures, the most comprehensive set of movement-related measures in a study of its kind to date. With a screening clinical application in mind, we chose a simple, binary clinical outcome that is most relevant to triggering clinical or research decision making: any cognitive decline versus stable or improving cognition. We further probed into the generalizability of the developed methodology, by applying it to data from two studies that gathered data from two different accelerometers worn at different body locations and with different wear protocols.

3.2 Methods

3.2.1 Data and other materials

Study population. To evaluate the robustness of our proposed methodology, we used information about two non-overlapping cohorts of community-dwelling older adults, one cohort equipped with hip-based and another with wrist-based accelerometers.

(1) **Hip accelerometry cohort: frailty, aging, body composition and energy expenditure in aging (FACE aging) study.** Study participants ($n = 151$) were recruited from the community around the primary geriatrics practice site for the University of Chicago located on the south side of Chicago. The sample was limited to community-dwelling (not living in residential care) older adults, 65 or older. Exclusion criteria included hospitalization, surgery, or procedure within 2 months of participating in the study; addition or change in dose of the thyroid (e.g, levothyroxine) or a diuretic (e.g, furosemide, hydrochlorothiazide, or spironolactone) medication within 2 months of participating in the study; use of oral steroids; use of beta-blockers (e.g., metoprolol, atenolol, or carvedilol); persistent hyperglycemia greater than 250; life expectancy less than 1 year; and history of moderate or advanced dementia or Montreal Cognitive Assessment (MoCA) less than or equal to 18. Hospital, surgery, medication, and hyperglycemia exclusion criteria were required to optimize resting metabolic rate testing at baseline (data not used in this analysis). Data collection occurred over multiple evaluations: (1) baseline survey and physical exam in the clinic, (2) a 7-day free-living hip accelerometry protocol immediately following the exam, (3) fasting resting metabolic rate measurement with indirect calorimetry and DEXA scan for body composition within 2 weeks of baseline assessment, (4) a 1-year follow-up survey

and physical exam in the clinic. We restricted the study sample to participants with complete clinical data and one or more valid (≥10 daytime hours) accelerometer-wear days, which left us with 115 participants eligible for our classifier development.

(2) **Wrist accelerometry cohort: the national social life, health, and aging project.** We used wrist accelerometry data generated by the National Social Life, Health, and Aging Project (NSHAP) as the sample. NSHAP is a nationally-representative, longitudinal survey study that collects extensive information on physical, mental, cognitive, and social health in United Study, community-dwelling older adults [71]. The first wave of NSHAP was in 2005–6 which included a nationally, statistically representative sample of community-dwelling adults born between 1920–47 (aged 57–85) and over-sampled for African-Americans, Hispanics, and males; 3377 respondents participated (weighted response rate = 75.5%). Five years later (2010–11), respondents were re-interviewed as were their cohabiting spouse or partner, for a total *n* = 3377. Interviews were conducted in the homes of each respondent by professional interviewers from NORC at the University of Chicago. A random subset of the 2010–11 respondents were invited to participate in a wrist accelerometry protocol, the data used in the current analysis.

Hip accelerometer protocol. Hip accelerometry data were collected from all participants at baseline. Following the baseline survey and physical exam, an Actigraph wGT3X+ hip accelerometer was placed over the participant’s mid, anterior right hip and secured with an elastic belt. Study participants were asked to keep the device on their hip continuously for 7 full days (including during bathing or showering). The accelerometers recorded data at a frequency of 30 Hz. The subsecond-level data were extracted from the devices using the ActiLife software (version 6.0). The low-frequency extension filter was NOT applied.

Wrist accelerometry sub-study protocol. Wrist accelerometry data were collected from a randomly selected subset of 793 respondents in the 2010–2011 data collection wave. The 2010–2011 accelerometry protocol has been previously described in detail [68]. Briefly, randomly selected respondents in the 2010–2011 data collection wave were asked to wear an ActiWatch Spectrum® on their non-dominant wrist continuously for 72 consecutive hours (including during bathing or swimming activities). The accelerometers recorded data at a frequency of 32 Hz. Upon receiving returned devices, data were downloaded from the device and then pre-processed using the Actiware® software [72]. The maximum absolute value was computed for each second; the sum of these absolute values was then computed for every 15-s epoch. The ActiWatch has a galvanic heat sensor that identifies when a device is on the wrist. All non-wear periods were excluded (only 0.17% of epochs across all wake data were classified as non-wear). Days with at least 10 h of daytime recording were considered "valid"; days with less than 10 h of daytime recording were excluded. The 24-h time interval was used to generate the wrist accelerometry features for this analysis. The study sample was restricted to participants with complete clinical data and ≥1 valid accelerometry wear day which left 584 participants eligible for our classifier development.

Data availability. The NSHAP data are publicly available and can be obtained from the National Archive of Computerized Data on Aging after completing a Data Use Agreement. The FACE Aging study data are available from one of the corresponding authors (M.H.S.) upon reasonable request and after completion of a Data Use Agreement and Institutional Review Board assessment.

3.2.2 Statistical Analyses

Clinical Measures. The cognitive function, the target inference in our project, refers to patients' ability to process incoming information. In this section we will introduce the measurement methods to identify cognitive decline among the cohort as well as derivation of covariates that being used as distinguish health signals to predict cognitive decline outcome.

Cognitive Function

Hip accelerometry cohort. The Montreal Cognitive Assessment (MoCA) was used to determine cognitive function at baseline and 1-year follow-up for the hip accelerometry training sample. The MoCA evaluates seven domains of cognitive function. The scale ranges from 0 to 30 with higher scores indicating better function. Because education was included as a covariate and our primary focus was on change in cognition, we did not add an additional point to the MoCA score for education levels below 12 years as is clinically done [73].

Wrist accelerometry cohort. In the wrist accelerometry sample, cognitive function was assessed in 2010–11 and 2015–16 using the survey-adapted Montreal Cognitive Assessment (MoCA-SA) as previously described in detail [72]. MoCA scores (range 0–30) are estimated from the 18-item MoCA-SA using a linear prediction model [74,75]. The wrist accelerometer data were collected in 2010–11 along with a baseline MoCA. The MoCA was repeated in 2015–16. In both cohorts, we calculated cognitive change as a difference in MoCA scores between the baseline and follow-up assessments (1 year for the hip accelerometry cohort and 5 years for the wrist accelerometry cohort):

$$\Delta = \mathrm{MoCA}_{\mathrm{follow-up}} - MoCA_{baseline}$$

Patients with deteriorating MoCA scores ($\Delta < 0$) were assigned to the cognitively declined group, denoted as Δ_{-}. The remaining patients were assigned to the group with a lack of cognitive decline, denoted as Δ_{+}. The ratio of Δ_{+}/Δ_{-} was 67/48 in hip-worn- and 279/296 in wrist-worn-accelerometer cohorts. The range of 1-year cognitive change (hip) was −8 to 6. The range of 5-year cognitive change (wrist) was −14.9 to 14.9.

Demographic Covariates

(1) **Hip accelerometry cohort.** In the hip accelerometry cohort, age, race, gender (female vs. male), education (high school≥ vs. < high school graduate), and monthly income category ($0 < 2000, $2000–3999, $4000–5999, and $6000+) were recorded through self-reported measures. Options for the race included Black or African American and Other (White, American Indian or Alaska Native, Asian Indian, Chinese, Filipino, Japanese, Korean, Vietnamese, Other Asian, Native Hawaiian, Guamanian or Chamorro, Samoan, Other Pacific Islander, or Other). No participants reported Hispanic ethnicity. Information on previously diagnosed comorbidities (self-reported and chart review) was recorded and scored using the Charlson Comorbidity Index and included heart attack, asthma, emphysema, chronic bronchitis, a chronic obstructive pulmonary disorder, peripheral vascular disease, liver disease, diabetes, and cancer (continuous, range 0–30) [76].

(2) **Wrist accelerometry cohort.** In the wrist accelerometry cohort, age (centered, continuous) was calculated using the reported date of birth and interview date. Gender (female versus male), race (White/Caucasian, Black/African American, other), and Hispanic ethnicity [77]. A modified Charlson Comorbidity Index (range 0–16, continuous) was constructed using self-reported comorbidity data in the 2010–2011 data collection wave. Respondents were asked whether they had ever been told by a doctor that they had any of the following

conditions (number of points given in parentheses): congestive heart failure (1), heart attack (1), coronary procedure (1), stroke (1), diabetes (1), rheumatoid arthritis (1), asthma, emphysema, chronic obstructive pulmonary disease, or chronic bronchitis (1), dementia (1), non-metastatic cancer excluding skin cancer (2), or metastatic cancer excluding skin cancer (6) [78].

Accelerometer data preparation. Data were restricted to enrollees with at least one valid day. We calculated the Euclidean norm minus one (ENMO), counts per minute (CPM), and vector magnitude count (VMC) for each participant using the hip and wrist data. To calculate these metrics, the accelerometry data needed to be in the form of the vector magnitude/Euclidean norm. The subsecond-level wrist accelerometry data were already converted to the vector magnitude/Euclidean norm by the manufacturer’s software, 1 data point for every 15-s epoch, where $N = 24$ h per day × 60 min per hour × 4 samples per minute = 5760 samples per day for wrist-worn accelerometer data.

The hip accelerometer data were in the form ($x(t)$, $y(t)$, z(t)), where $x(t)$, $y(t)$, and $z(t)$ are dimensionless data provided by the accelerometry device, which are approximately proportional to the (x-, y-, and z-axis) directional acceleration [79]. Time t is discrete, which for each day t runs from 1 to N, where $N = 24$ h per day × 60 min per hour × 60 s per minute × 30 samples per second = 2,592,000 samples per day for hip-worn accelerometer data. The vector magnitude/Euclidean norm *r(t)* was computed in the hip accelerometry data as follows:

$$\boldsymbol{r(t) = \sqrt{x(t)^2 + y(t)^2 + z(t)^2}}$$

To normalize the vector magnitude/Euclidean norm *r(t)* to a consistent length across both the wrist and hip accelerometry cohorts, the vector magnitude/Euclidean norm *r(t)* was reshaped to a $D \times T$ matrix $\boldsymbol{R} = \boldsymbol{R}_{\mathbf{dt}}$ where D represents the total number of wear days and T represents collected samples per day. The average, normalized vector magnitude/Euclidean norm $\bar{r}(t)$ is computed as follows:

$$\bar{r}(t) = \frac{1}{D}\sum_{d=1}^{D} R_{dt}$$

We then used non-overlapping 1-minute, sliding windows to extract the Euclidean norm minus one (ENMO), the counts per minute (CPM), and the vector magnitude count (VMC), both formally defined below. The ENMO was used to remove noise and gravitation effects from subminute and subsecond-level data. Letting H denote the number of time measurements in a one-minute sliding window, we can write ENMO as:

$$ENMO(t) = \frac{1}{H}\sum_{h=1}^{H-1} \max[\bar{r}(\mathrm{t}+\mathrm{h}) - 1\,, 0]$$

The feature CPM was further derived as:

$$CPM(t) = H * ENMO(t)$$

Note that $H = 60$ s per sliding window × 30 samples per second = 1800 samples per sliding window for hip accelerometer data and $H = 60$ s per sliding window × 4 samples per minute = 4 samples per sliding window for wrist accelerometer data.

The VMC was used to evaluate the mean amplitude deviation in the sliding window period with size H, defined as:

$$VMC(t) = \frac{1}{H}\sum_{h=0}^{H-1} |\bar{r}(t+h) - \bar{r}(t)|$$

where t now varies over the minutes each day, from 1 to N, with $N = 24$ h per day × 60 min per hour = 1,440 min per day.

Accelerometry activity level measures (C4 and V4). Two categorical activity measures that we call C4 and V4 were computed. After extracting CPM and VMC measures from the accelerometer data, we generated the 75th percentile for CPM and VMC data points for each participant denoted as CPM_{75} and VMC_{75}. The sample-based distribution of the CPM_{75} and the VMC_{75} were then categorized into four levels at each quartile to create a C4 and V4, respectively. Figure 3.1 shows the cohort-specific-based quartiles for CPM_{75} and VMC_{75} labeled as: inactive [0–25%], moderately active [25–50%], active [50–75%], and extremely active [75–100%].

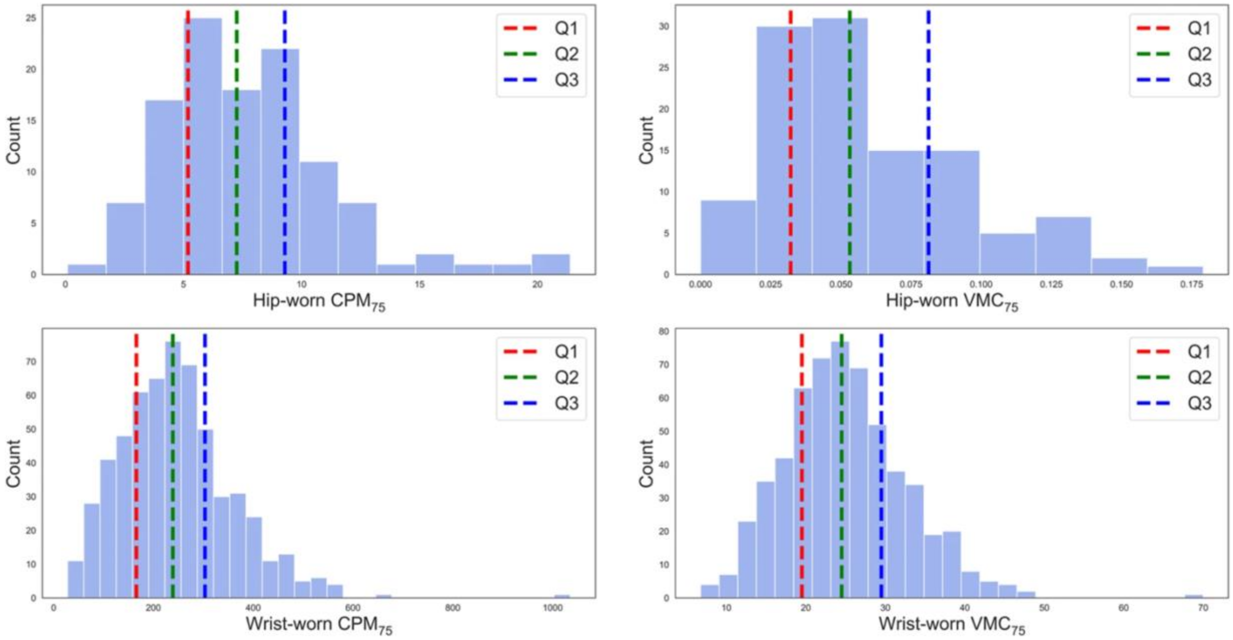


Figure 3.1. Hip/Wrist accelerometry cohorts distribution of $\text{CPM}_{75}/\text{VMC}_{75}$. Red dashed line represents Q1 quantiles on inactive: [0, 25%]; Green dashed line represents Q2 quantiles on moderately active [25–50%]; Blue dashed line represents Q3 quantiles on active [50–75%], and above Q3 represents extremely active [75–100%].

Accelerometry pattern measure. After obtaining minute-level ENMO(t) and VMC(t), we then computed 98 statistical and harmonic features. This resulted in 105 features for those wearing the hip accelerometers (study population $N = 115$) and 104 features for those wearing the wrist accelerometers (study population $N = 575$). The number of features for the wrist- and hip-worn devices differed because the income and ethnicity/race categories in the two datasets were not identical. The specific features for ENMO(t) and VMC(t) are listed in Table 3.1. We illustrated the meaning of the individual harmonic features in Fig. 3.2. While we computed a relatively large number of harmonic features, the features belong to just a few categories: differential entropy (flatness of a distribution), fast Fourier transform (revealing periodicity in activity), and statistics describing shapes of a distribution, such as mean, variance, skewness, and kurtosis.

Table 3.1. Statistic And Harmonic Features Extracted From CPM (T) And VMC (T).

Statistical Features
Mean and median
Standard deviation
Minimum and maximum
25th and 75th percentile
Skewness and kurtosis
Entropy
Beta distribution shape (α,β)
Harmonic features
Top 15 FFT[a] Coefficients (frequency/signal)
FFT[a] entropy
Periodogram frequency mean, Standard deviation
RMS[b] amplitude
Periodogram frequency Kurtosis and Skewness

[a] Fast fourier transformation (FFT).

[b] Root mean squared (RMS).

Statistical analysis. First, we computed characteristics of the two cohorts: means (±standard error, SE) for continuous measures and proportions (±SE) for categorical variables. Second, to analyze the statistical significance of the covariates for distinguishing between the two classes (Δ_-, Δ_+), we evaluated the predictive importance of each of the demographic, comorbidity, and accelerometry measure in both cohorts. Third, we built a binary classifier called CDPred to distinguish Δ_+ from Δ_- in the two cohorts using XGBoost (Extreme Gradient Boosting). To evaluate the performance of the model in each cohort, we randomly chose 10% of the hip accelerometry cohort and 15% of the sample from the wrist accelerometry cohort as a hold-out sample. The CDPred hyperparameters were fine-tuned using 5-fold cross-validation, to maximize the area under the curve (AUC) score. We then reported the performance of each model in terms of predicted accuracy and AUC on a hold-out sample. The feature importance of distinguishing Δ_+ and Δ_- were then listed in descending order of importance for the best-performing model in each dataset.

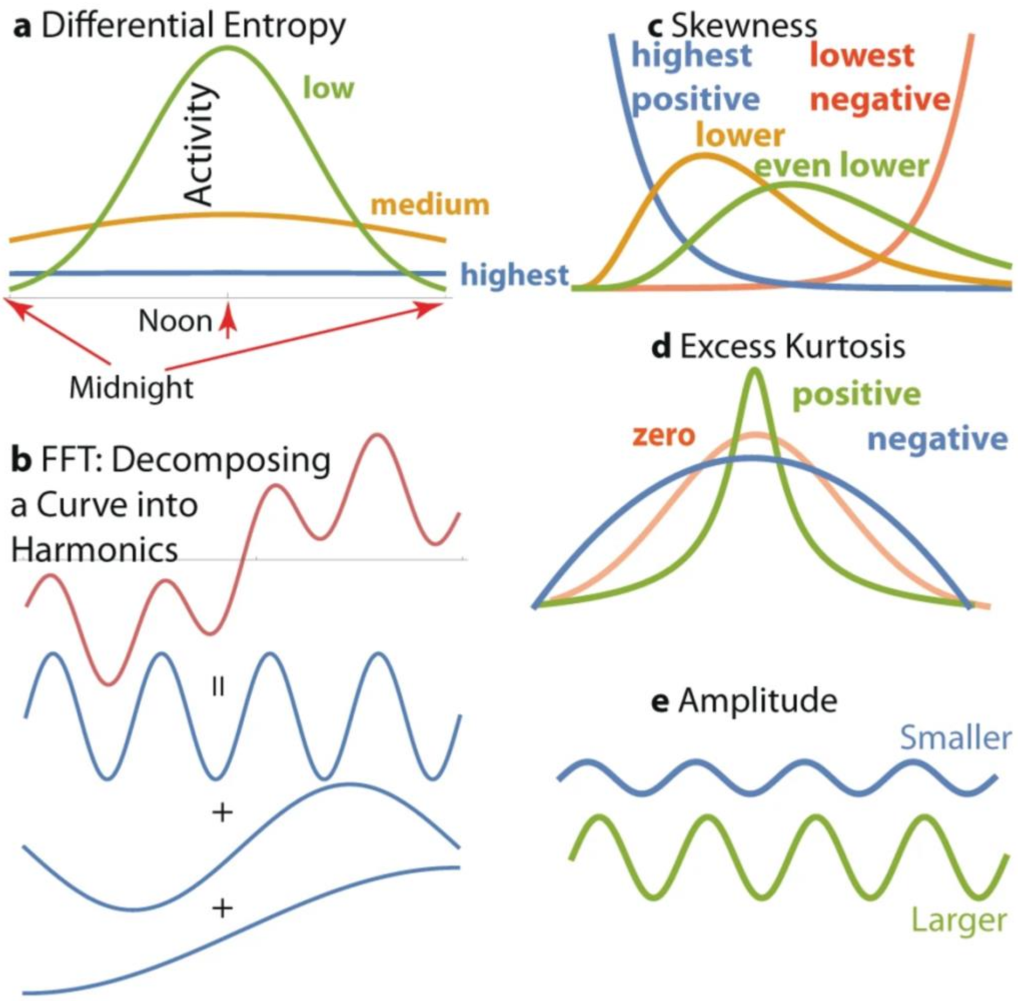


Figure 3.2. **a** Differential entropy: Differential entropy is the highest for a uniform distribution of activity (for example, when a person stays inactive 24 h a day, so there are no bursts of activity). When a person is more active through the day and inactive at night, the entropy of activity drops, because the daytime activity exceeds the night-time activity average. **b** Fast Fourier transform (FFT): The fast Fourier transform refers to the number of harmonics that can be used to describe a curve. Any curve can be decomposed into a spectrum of harmonics. In this case, the hypothetical activity curve shown in red is the sum of 3 harmonics with nonzero amplitude: one with four cycles a day, one with a single full cycle a day, and one with a two-day cycle. In real accelerometry data, the number of accelerometry harmonics composing a 24-h circadian pattern is typically over 15 harmonics. **c** Skewness is a statistic characterizing the asymmetry of the distribution of activity; it can be applied to entire device wear time or to smaller intervals of accelerometry readings. **d** Excess kurtosis is a statistic indicating deviation of a distribution from a normal distribution. Kurtosis is zero for a normal distribution, positive for distributions with heavier (than normal) tails, such as *t*-distribution, and negative for distributions that have lighter tails, such as Beta with parameters (2,2). **e** Amplitude: The amplitude of each harmonic in an FFT reflects the distance between minimum and maximum activity values. For non-essential (noise-level) harmonics in FFT, the amplitude is close to zero.

3.3 Results

Cohorts characteristics. The characteristics of the two study cohorts are shown in Table 3.2. The hip accelerometry cohort was older (mean age 73.2), had a slightly higher baseline Montreal Cognitive Assessment (MoCA) score (mean 25.4), and included a larger proportion of females (80.9%) and those self-identifying as African American (81.7%) than the wrist accelerometry cohort (mean age 70.0, mean MoCA 23.4, proportion female 59.1%, proportion African American/Black 11.3%).

Table 3.2: Demographic Composition Of The Two Accelerometry Cohorts.

	Mean (SD) or *N* (%)	
Characteristics	**Hip accelerometry cohort (*N*=115)**	**Wrist accelerometry cohort (*N*=575)**
Age (year)	73.2 (5.9)	67.0 (7.9)
Gender (female)	93 (80.9 %)	340 (59.1 %)
Race		
African-American / Black	94 (81.7 %)	65 (11.3 %)
White	21 (18.3 %)	423 (73.6 %)
Hispanic	None	67 (11.7 %)
Other	None	20 (3.5 %)
Education		
Some college or junior college	45 (31.9 %)	211 (36.7 %)
Post-graduate	28 (24.4 %)	137 (23.8 %)
College graduate	27 (23.5 %)	
High School graduate or GED (grade 12)	11 (9.6 %)	135 (23.5 %)
Some high school (grades 9-11)	4 (3.5 %)	92 (16.0 %)
Income ($/month)		
< 2000	65 (50.0 %)	-
2000 ~ 3999	40 (30.8 %)	-
4000 ~ 5999	16 (12.3 %)	-
≥ 6000	9 (6.9%)	-
Charlson Comorbidity Index Score	1 (1.3)	0.9 (1.2)
MoCA (baseline)	25.4 (2.6)	23.4 (4.0)
MoCA (Hip: 1 year; Wrist: 5 years)	25.6 (3.0)	22.6 (4.4)

Demographic and clinical predictors of cognitive decline. As we observe in Table 3.3, the clinical characteristics had somewhat limited capability to distinguish between those with stable/improving cognition versus those with declining cognition at 1 and 5 years in the local and national cohorts, respectively. We provided a full dictionary of features in Supplemental Table 3.1.

Table 3.3: Effect Size of Predictors of 1- and 5-Year Cognitive Function

Characteristics	Δ_+ **Vs** Δ_- **Hip-worn (N = 115)**	Δ_+ **Vs** Δ_- **Wrist-worn (N = 575)**
	Cohen's D/Odds Ratio (95% Confidence Intervals)	
Age (year)[a]	0.457 (0.103, 0.801)	0.284 (0.119, 0.449)
Gender (female)[b]	3.533 (1.223, 10.207)	0.865 (0.620, 1.207)
Race[b]		
African American / Black	1.124 (0.472, 2.680)	0.897 (0.741, 1.086)
White	-	1.014 (0.795, 1.293)
Hispanic	-	1.143 (0.873, 1.499)
Other	-	1.344 (0.769,2.345)
Education[b]		
Some college or junior college	2.265 (1.092, 4.697)	0.850 (0.605, 1.195)
Post-graduate	0.973 (0.443, 2.134)	0.909 (0.619, 1.335)
College graduate	0.532 (0.233, 1.215)	-
High School graduate or GED	0.437 (0.110, 1.726)	1.018 (0.652, 1.591)
Some high school (grades 9-11)	0.821 (0.132, 5.088)	1.337 (0.908, 1.969)
Income ($/month)[b]		
< 2000	1.283 (0.641, 2.566)	-
2000 ~ 3999	0.656 (0.306, 1.405)	-
4000 ~ 5999	0.715 (0.244, 2.101)	-
≥ 6000	2.653 (0.634, 11.110)	-
Charlson Comorbidity Index Score[a]	0.135 (-0.213, 0.485)	-0.008 (-0.172, 0.156)
MoCA (baseline)[a]	0.572 (0.216, 0.928)	0.418 (0.252, 0.584)

[a]Effect size computed by Cohen's D method; [b]Effect size computed by odds ratio.

Combining demographic, clinical, and accelerometry predictors of cognitive decline. To investigate the importance of the accelerometry activity measures and harmonic features beyond that of the demographic and clinical characteristics on cognitive degradation forecasting, we trained CDPred on three different sets of measures: (1) the CDPred basic model using demographic and clinical characteristics; (2) the CDPred-4 model using demographic and clinical characteristics with C4 and V4; (3) the CDPred-4+ model using demographic and clinical characteristics, C4 and V4, plus the harmonic features derived from accelerometry. The number of features in each model are listed in Table 3.4. To summarize, we compared three models: CDPred, CDPred-4, and CDPred-4+. CDPred includes the baseline demographic and clinical features. CDPred-4 model uses the baseline demographic and clinical features and two baseline accelerometry metrics (C4 and V4). CDPred-4+ models use the full gamut of information: the baseline demographic and clinical features, the two baseline accelerometry metrics, and all extracted 98 accelerometry harmonic features.

Table 3.4. Number Of Predictors In The 3 Hip And Wrist Accelerometry Models.

Model	**Number of predictors**	
	Hip-worn	Wrist-worn
CDPred	7	6
CDPred-4	9	8
CDPred-4+	105	104

Performance of the models. The model performance metrics on the hold-out samples are shown in Table 3.4. The CDPred-4+ model including all measures predicted cognitive decline 1 year later with an accuracy of over 85% (hip accelerometry cohort) and predicted cognitive decline 5 years later with nearly 70% accuracy (wrist accelerometry cohort). The hip-worn accelerometry confusion matrix and ROC-AUC for the CDPred-4+ model in the hold-out sample is shown in Fig. 3.3. Figure 3.4 shows predictors sorted by relative importance, from the highest to lowest, excluding features with zero importance. Similarly, we show the confusion matrix and ROC-AUC for the CDPred-4+ model in the wrist-worn accelerometer data in Fig. 3.5, and nonzero predictor relative importance sorted in descending order in Fig. 3.6.

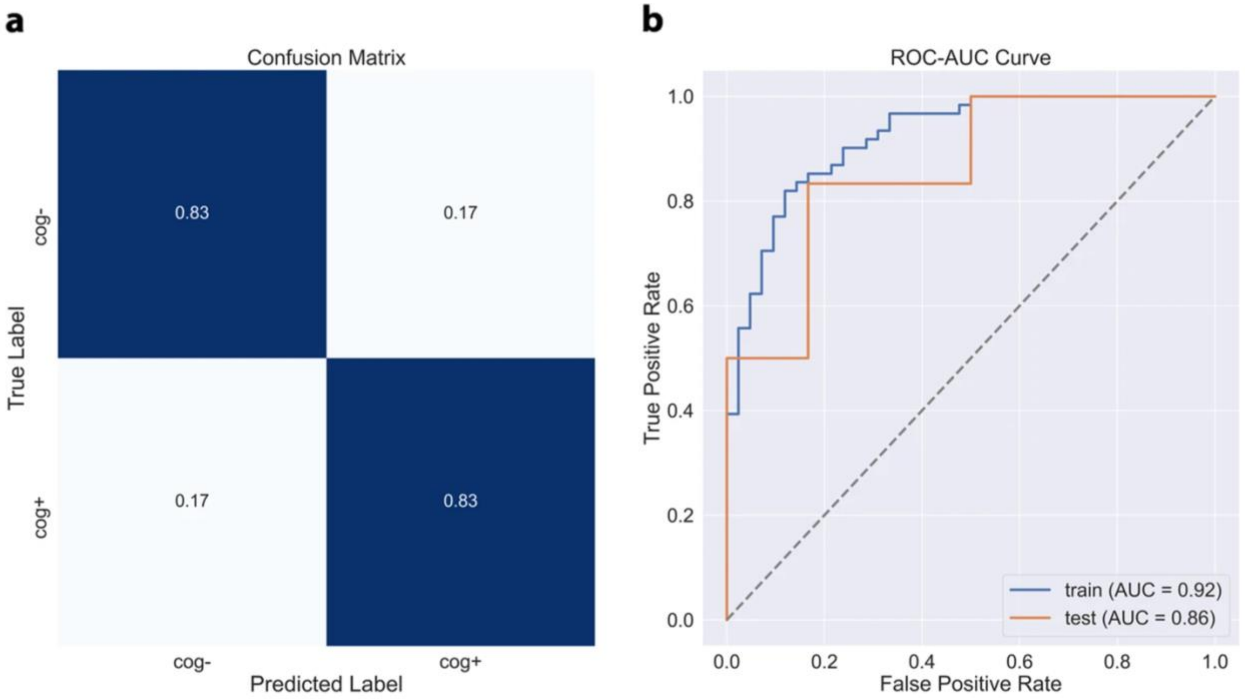


Figure 3.3. The figure shows a. confusion matrix, b. ROC-AUC curve for the held out sample.

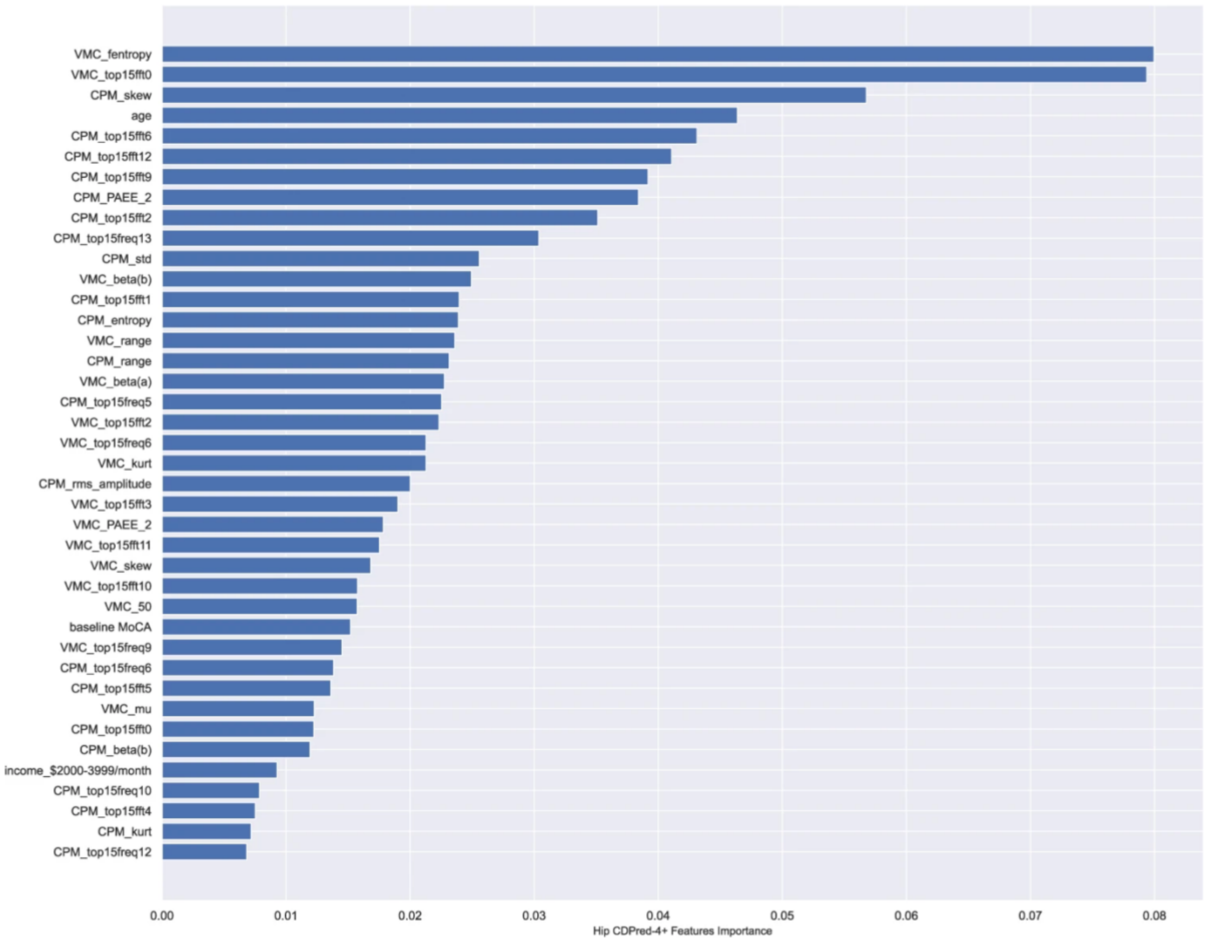


Figure 3.4. The features are listed in order of decreasing importance, from top to bottom of the graph.

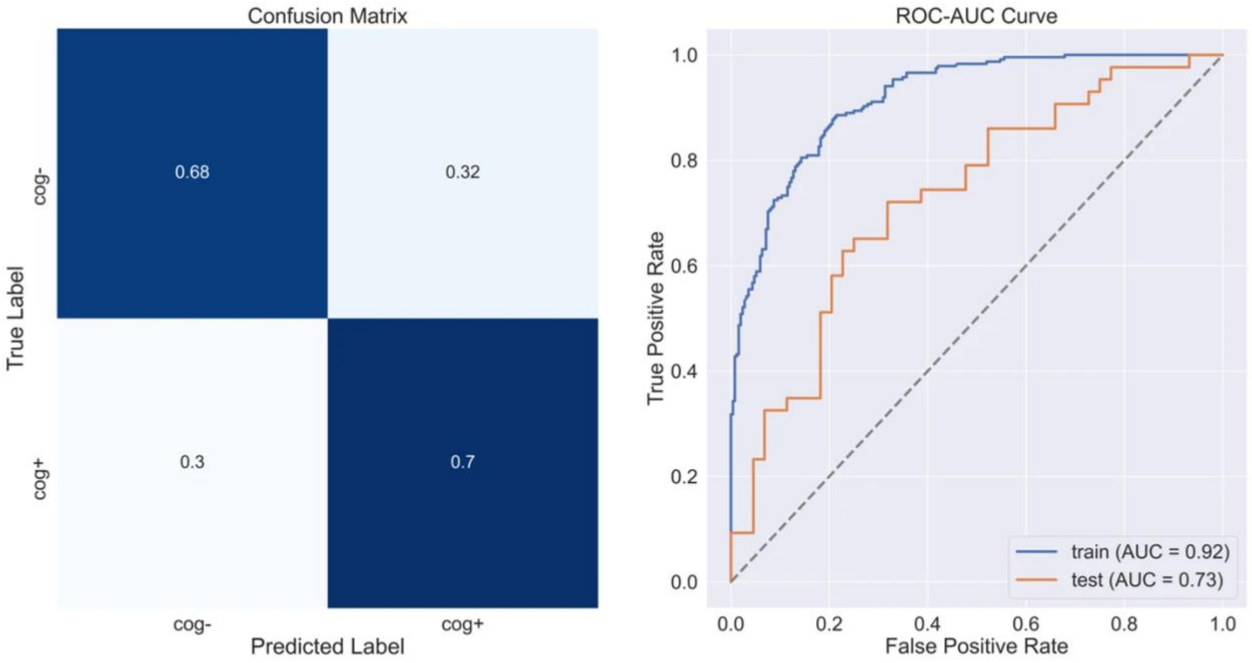


Figure 3.5. The features are listed in order of decreasing importance, from top to bottom of the graph.

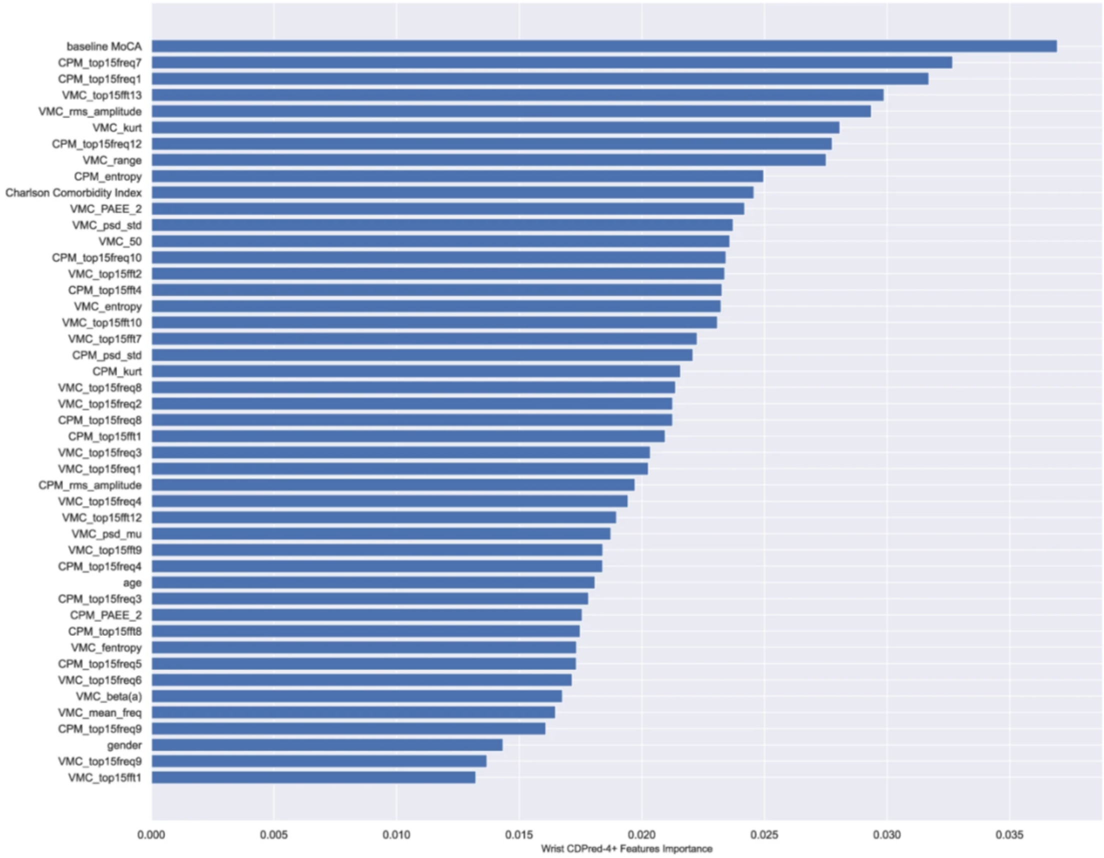


Figure 3.6. The features are listed in order of decreasing importance, from top to bottom of the graph.

3.4 Discussion

Our model significantly expands work previously published in this space. Casanova et al. (2020) similarly used a Random Forest Classifier to distinguish cognitive trajectories [52]. Three classes, low-, medium-, and high-risk trajectories were created using a combination of baseline and repeated cognitive performance scores. This study found that age, gender, education, BMI, stroke, diabetes, neighborhood socioeconomic status, and APOE carrier status were among the top predictors of cognitive trajectories. They did not include accelerometry assessments. We found that the accelerometry pattern features outperformed many demographic and clinical characteristics in predicting cognitive decline in a community-dwelling cohort, suggesting the potential value of noninvasive and remote accelerometry in augmenting the clinical evaluation.

Our analyses have shown that, compared to simpler, clinical models predicting cognitive decline (e.g., using only demographic and clinical characteristics), our accelerometry-based classifier model performs significantly better. This model uniquely identifies preclinical cognitive decline among older adults without a diagnosis of dementia over short (1-year) and longer-term (5-year) follow-up. The model was robust to varying wear protocols (7 days versus 72 h), device location (hip versus wrist), and device manufacturer. In both models, many accelerometry features were rated more ‘important’ in distinguishing those who experienced any decline in cognition than many demographic and clinical characteristics including age. We are hopeful that the current level of model performance may be useful to flag older adults most vulnerable to subsequent cognitive decline. We note and emphasize that accelerometry currently has no diagnostic capacity for any clinical diseases; its role in the current study is restricted to an assessment of day-to-day movement (accelerations and decelerations) which seems to reflect some level of health, here cognitive risk.

3.5 Limitations

This study has several limitations worth mentioning. It is not technically possible to guarantee (or test) that there was no overlap between the two cohorts used in this study. The NSHAP dataset did collect zip code information on participants, but the FACE Aging dataset did not collect any address information. Since the FACE Aging dataset is composed of study participants residing in the few neighborhoods surrounding the University of Chicago and NSHAP sampled across the nation using a complex sampling design based on census tracts, if overlap occurred, it would have been a very small number of participants.

Another limitation of the current study is, despite the importance of understanding ADRD for socio-demographically disadvantaged groups, the datasets for this study were not sufficient in size for understanding the relative predictive power of the models for different sociodemographic groups. We did include effect size measures for different race/ethnicities in Table 3.3 and the effect size of accelerometer features on race/ethnicity in Supplemental Figure 3.1 as the first step in this direction. We show ranked effect sizes of top 20 individual features in Supplemental Fig. 3.2.

Our forecast model was only 70–80% accurate leaving room for improvement. It is likely that our forecast model could be enhanced in future work to reach higher and more consistent accuracy. This can be achieved by including additional metrics derived from accelerometry data, possibly using additional physiologic sensors such as heart rate monitoring to capture richer data, and incorporating additional clinical data, such as blood or genetic markers, family history of dementia, and current medications.

The wrist accelerometry model did not perform as well as the hip accelerometry model. The weaker performance of the wrist accelerometry location might be due to the shorter wear protocols, increased motion “noise” related to the position, and longer follow-up cognitive assessments. Future work comparing more similar wear protocols and devices, even if worn at different body locations, would be of value.

Our experiments show that this predictive model can forecast preclinical cognitive decline using data from dissimilar accelerometry device locations, wear protocols, follow-up times, and unique cohorts. Hip- and wrist-worn accelerometers are subject to unique patterns of movements in space, yet data from both accelerometry devices improved the predictive capacity of the respective models. The somewhat inferior performance of the wrist accelerometry, among other factors, may be related to the shorter wear protocol (72 h versus 7 days versus “noisy” data at the wrist) and longer follow-up (5 years). A major challenge to accelerometry research and clinical translation has been the reliance on a particular device location, protocol duration, and/or proprietary data processing software for generating accelerometry measures [69, 70]. These limitations have stimulated movement toward using open-source programs or approaches for generating accelerometry metrics, as we have done in this study, and identifying methodologic approaches applicable across multiple devices and varying wear protocols.

CHAPTER 4 SURVIVAL ANALYSIS ON OSTEOPOROTIC FRACTURE RISK FORECAST WITH ELECTRONIC HEALTH RECORDS

4.1 Introduction

Osteoporosis is a systemic skeletal disease characterized by reduced bone strength and pre-disposition to fracture [80]. Osteoporosis leas to two million fracture per year in the US, resulting in pain and disability and 54 million people in the US have low bone density or osteoporosis [81]. Despite this morbidity, osteoporosis is usually not diagnosed or treated and, even after a hip fracture, only 3.3% of patients received osteoporosis therapy [82]. Many patients never receive bone density screening or treatment, despite a high risk of fracture. Fracture risk is multi-facetted and includes risk factors including age, gender, prior fracture, co-morbid conditions, medication use, etc [83]. There are several fracture risk calculators available such as Fracture Risk Assessment Tool (FRAX) and QFracture.

Unfortunately, osteoporosis and fracture risk are often not considered in busy primary care practices, leading to large numbers of osteoporotic fractures that could have been prevented with appropriate diagnosis and treatment [80]. The electronic health records provide a unique opportunity to automate the identification of patients who are at risk of fracture within a primary care provider's workflow. However, there is relatively little data about the ability to the electronic health records to identify patients at risk of fracture using currently available risk calculations. FRAX was derived from prospective, international cohorts using demographics, body mass index or bone density, and questionnaire, while QFracture was derived in the UK primary for White people [85-87]. Given that FRAX and QFracture were developed and validated predominately outside of the US, it is important to examine their performance in a US population where there is greater racial and ethnic diversity. Therefore, it’s important to develop a fracture risk forecast tool that target primary on the US population with electronic health records, which include more race and ethnic disparity.

The use of race and ethnicity in medical calculators has come under scrutiny. Indeed, many of the differences seen in medicine that are attributed to race are related to historical and current social inequalities and institutional racism. However, osteoporosis and fragility fractures stand out from many other areas of medicine. The rate of fracture varies widely throughout the world [88]. Moreover, there are very few studies validating fracture calculators in diverse US multi-ethnic cohorts and none, to our knowledge, using the electronic health records. Therefore, it’s clear that further study of fracture risk is calculators is limited especially in racial and ethnic minorities in the US. If fracture calculators could be authorized into US. It would be helpful in both reducing the osteoporosis treatment gap overall and reducing racial disparities in osteoporosis screening and treatment, which lags behind in Black women [89] using US based electronic health records developed tool. Besides, we sought to first validate the ability of electronic health records derived survival analysis models to identify patients with high risk of fracture and to determine whether the current adjustments for race/ethnicity accurately capture the difference in fracture risk. We have sampled the data from primary care population in an urban, tertiary care medical center with large portions of Black and Hispanic Patients.

4.2 Methods

4.2.1 Data and other materials

Population Datasets: The routine clinical data is from the electronic health record (EHR) of Temple University Hospital in Philadelphia, PA obtained from encounters between October 1, 2010 and December 1, 2018. Subjects were required to have at least 2 full years of follow-up. Of these two years, the first year was used for baseline data collection for fracture risk factors and subsequent time was used for observation of the outcome (i.e., fractures). Furthermore, because the target clinician population for risk calculators is the primary care setting, we required common health maintenance measures as a proxy for subjects receiving their primary care at the institution. This furthermore helped to signify that some of the visits were routine (i.e., not problem visits), that health maintenance could be addressed, and subjects were willing participants in routine health maintenance.

For women, our inclusion criteria were at least one measurement of LDL and at least one diagnosis code for both mammogram (ICD-10 Z12.31, ICD-9 Z76.12) and vaccination (ICD-9 V03-V06, ICD-10Z23). For men, our inclusion criteria were at least one measurement of both PSA and LDL and at least one diagnosis code for vaccination (ICD-9 V03-V06, ICD-10 Z23). Of note, the screening tests could have occurred at any time and the actual screen or test (e.g. mammogram) did not have to be complete in order to be included in the study. Subjects with missing demographic data (age, race, or gender) or body mass index (BMI) data were excluded, given the use of these variables in fracture risk calculators. We also excluded subjects with a prescription for an osteoporosis medication at the time of entry into the study. If patient was later put on a medication for osteoporosis after entry, they were censored at the time of the prescription, but earlier data was used. The study was approved by the Temple University Institutional Review Board (IRB).

Determination of Fractures: The presence and characteristics of fractures were determined by the presence of physician-billing codes. Major osteoporotic fractures (MOFs, a composite of fractures of the wrist, humerus and vertebrae) and hip fractures were analyzed separately. As the determination of vertebral fractures were through physician diagnosis codes, only clinical vertebral fractures, not morphometric vertebral fractures, were determined. Fractures temporally associated within 30 days of trauma codes were excluded, and we continued to follow these subjects to observe for non-traumatic fracture. To prevent double counting, subjects with prior fracture at any individual site (e.g., wrist or lumbar spine) were not counted as having an incident fracture at the same site. A subset of 140 random subjects with fracture codes were examined with 86% accuracy (121/140) for identifying MOF based on chart review of imaging studies and physician notes. The most common reasons for erroneous codes were incorrect site (e.g., hand fracture, instead of wrist fracture) or pain at the site with a normal imaging study.

4.2.2 Statistical Analyses

The population detailed information is shown in Table 4.1 which include demographic information such as race, age, BMI as well as specific major fracture outcome measurements such as years-of-follow-up, primary fracture type, etc. Those rich health record enable us to build complementary survival analysis for major fracture outcome risk progress in scale of years. First, we have defined some basic concepts to quantify the measurements and desired outcome.

1. Survival time T is a random non-negative variable, the duration from the start of treatment to desired outcome. In our project, T is the discrete time for patients from registered to major fracture observation. The density function f for discrete survival time distribution can be defined as:

$$T \in \{0,1,2,3,\ldots\},\;\; f_i = \; P(T = i)$$

2. Survival function/curve $S(t)$ is the survival function defined as the probability that at time t, the probability that case outcome (major fracture) would occur after that time point. The equation of this concept can be expressed as

$$S(t) = P(T > t)$$

The discrete definition of $P(T > t)$ can be further computed as cumulative percentage of cases among all observations, expressed as

$$P(T > t) = \sum\nolimits_{j>i} f_i$$

where N is the size of sampled population, c_i is the survival time for each case.

3. Hazard rate/function $h(t)$ is defined as:

$$h(t) = \frac{f_t}{S(t-1)}$$

4. Accumulative hazard function $H(t)$ is defined as

$$H(t) = \sum\nolimits_{j \le t} h(j)$$

In our dataset with denoted major fracture occurrence time as $T_1, T_2, T_3, \dots, T_n$. However, in some cases, censoring can occur due to cases that when the study ends, some individuals have not had observed major fracture yet. Therefore, we need to introduce sample's censoring time as $C_1, C_2, C_3, \dots, C_n$. Then what we can actually observe for each sample are $Y_i = \min\,(T_i, C_i)$ and an indicator of whether censoring occurs:

$$\delta_i = \begin{cases} 0, if\ T_i \le C_i (observed\ death) \\ \quad 1, Otherwise \end{cases}$$

When each sample also has its covariate, what we observe can be denoted as (Y_i, X_i, δ_i) for each sample. In this project, we only consider non-informative censoring, which is basically requiring that

$$T_i \ \perp C_i \mid X_i$$

Which means that the censoring time is not associated with the survival time, at least conditioning on other known covariates X_i.

For now, we have clarified some important concepts on major fracture survival analysis, to estimate the survival function, we have utilized both non-parametric and parametric approaches. In both methods we have considered the scenario that there is no observed covariates X_i and survival time T_i are i.i.d distributed.

First, we've utilized popular Kaplan-Meier estimator to estimate the survival function $S(t)$, also we will explore the conditioned survival function given varies demographic features such as gender and race. For our project the discrete survival time T for major fracture observations, we can discretize the survival time into bins. For each bin i or discrete survival time T_i, assume we observe r_i samples that are still free of major fracture at the beginning of this time bin with d_i major fracture observations during this time bin and c_i censored cases at the end of time bins. Now we can assume a Bernoulli distribution of major fracture occurrence at this time point as:

$$d_i \sim Bernouli(r_i, h_i)$$

The unbiased estimator of h_i is

$$\widehat{h_\iota} = \frac{d_i}{r_i}$$

Which gives us Kaplan-Meier estimator of survival function $S(t)$ as:

$$\hat{S}(t) = \prod_{j \leq i} \left(1 - \widehat{h_\iota}\right) = \prod_{j:\tau_j \leq i} \frac{r_j - d_j}{r_j}$$

Where $\{\tau_1, \tau_2, \tau_3, \ldots, \tau_K\}$ is the set of K distinct uncensored major fracture observed in the sample, d_j is the number of major fracture occurrence at τ_j and r_j is the total number of patients who are at risk right before τ_j.

So far we have discussed non-parametric method on estimating major fracture survival function, the Kaplan-Meier method is driven by survival time distribution and build best estimator with prior assumption of Bernoulli distribution on major fracture outcome. However, this method couldn't take associated covariate into account and is vulnerable when prior distribution assumption is not valid. Hence, we will further explore the parametric methods to estimate the survival function, after including possible information such as demographics and disease history.

To include covariates into survival mode, we have utilized proportional hazards linear regression model as benchmark model, which was proposed by David Cox. For each sample s, we have observed (y_s, X_s, d_s) where X_s is the covariate introduced to the major fracture datasets. The proportional hazards (PH) model assumes that

$$h_s(t) = e^{X_s^T \beta} h_0(t)$$

Here we have no assumption regarding $h_0(t)$, so it's a semi-parametric model and X_s does not include the intercept for identifiability. Next, we can identify proportional hazard as:

$$log\left\{\frac{h_s(t)}{h_0(t)}\right\} = X_s\beta$$

The benefit of building the model on the hazard rate instead of survival function is that the survival function need to be less than 1, while the hazard rate does not have that constraint. The benefit of having a proportional model is that there is no constraint on the range of β to have the hazard rate positive.

Now we have defined the proportional hazard rate and its linear regression model to other covariates X_s. Then we can denote the risk set as $R(t) = \{s: y_s \geq t\}$, which is the set of people that is at risk at time t. For each y_s with $d_s = 1$ where the event is observed, there are $R(t)$ individuals that are at risk, conditional on the fact that there is exactly one person have major fracture observation, the likelihood that individual s is chosen is then

$$L_s = \frac{h_s(y_s)}{\sum_{l \in R(y_s)} h_l(y_s)} = \frac{e^{X_s^T \beta}}{\sum_{l \in R(y_s)} e^{X_l^T \beta}}$$

We can further construct the full likelihood equation, for each sample s, assume we observe (y_s, δ_s). We build a likelihood for each sample conditional on C_s such that if

- If $\delta_s = 0$, then we observe $T_s = y_s$, the likelihood is $L_s = f(y_s) = S(y_s)h(y_s)$
- If $\delta_s = 1$, then we only observe $T_s \geq y_s$, the likelihood is $L_s = S(y_s)$.

Thus the full likelihood can be expressed as

$$L = \prod_{S} L(s) = \prod_{s=1}^{n} S(y_s) h(y_s)^{\delta_s}$$

After obtain the full likelihood of survival function, we can estimate and inference with the likelihood equation with Maximum-Likelihood-Estimation (MLE) method by taking the log-likelihood equation as:

$$l = \log(L) = \sum_{s=1}^{n}(1-\delta_s)\left[X_s^T\beta - log\left\{\sum_{t\in R(y_s)} e^{X_s^T\beta}\right\}\right]$$

With coefficient estimator $\hat{\beta}$ can be solved with MLE that $\dot{l}(\beta) = 0$ and it's been proved that β has asymptotic distribution that

$$\hat{\beta} \sim N\left(\beta, \ddot{l}(\hat{\beta})^{-1}\right)$$

We can further generalize the learner function $f(x)$ beyond the constraints of linear regression model as $f(x) = X_s^T\beta$ to broader machine learning algorithms to construct base learner functions. To expand the spectrum of base learner, we can use alternate method to MLE on coefficients estimation, the popular methodology is to use gradient descent method to find minimum point on pre-defined loss function. Depending on the loss function to be minimized and base learner used, different model arises, we can have selections of tree-structured model such as random forest and gradient boosting or more complicated neural networks. Here we can re-write the Cox's proportional hazards log-likelihood as target loss function that to be minimize given base learner $f(x)$.

$$f(x) = argmin\sum_{s=1}^{n}(\delta_s - 1)\left[f(x_s) - log\left\{\sum_{t\in R(y_s)} e^{f(x_t)}\right\}\right]$$

In this project, we will train survival analysis models with varies base learners, to compare the performances of those base learners, we use the concordance index as the metric. The concordance index is defined as the proportion of all comparable pairs in which the predictions and outcome are concordant. More specifically, two samples are concordant if the one with higher estimated risk score has short survival time or major fracture observed time in our project.

4.3 Results

The cohort characteristics for the group is shown in Table 4.1 as discussed above. As we can observe from the characteristics of the study population by race, the white patients were older, had a higher proportion of man and had significantly fewer visits overall. In general, the average BMI was in the obese range. Years of follow-up varied from 1 to 7 years with average to nearly 4 years. Besides, the Charlson comorbidity score was significantly higher in Hispanics and Black subjects compared to Whites. Fracture rates, by race and sex are shown in Figure 4.1. As expected, White women and mem had the highest fracture rates (7.6 and 4.9 per 1,000 person-years, respectively) and Black women and men had the lowest fracture rates (3.9 and 3.2 per 1,000 person-years, respectively).

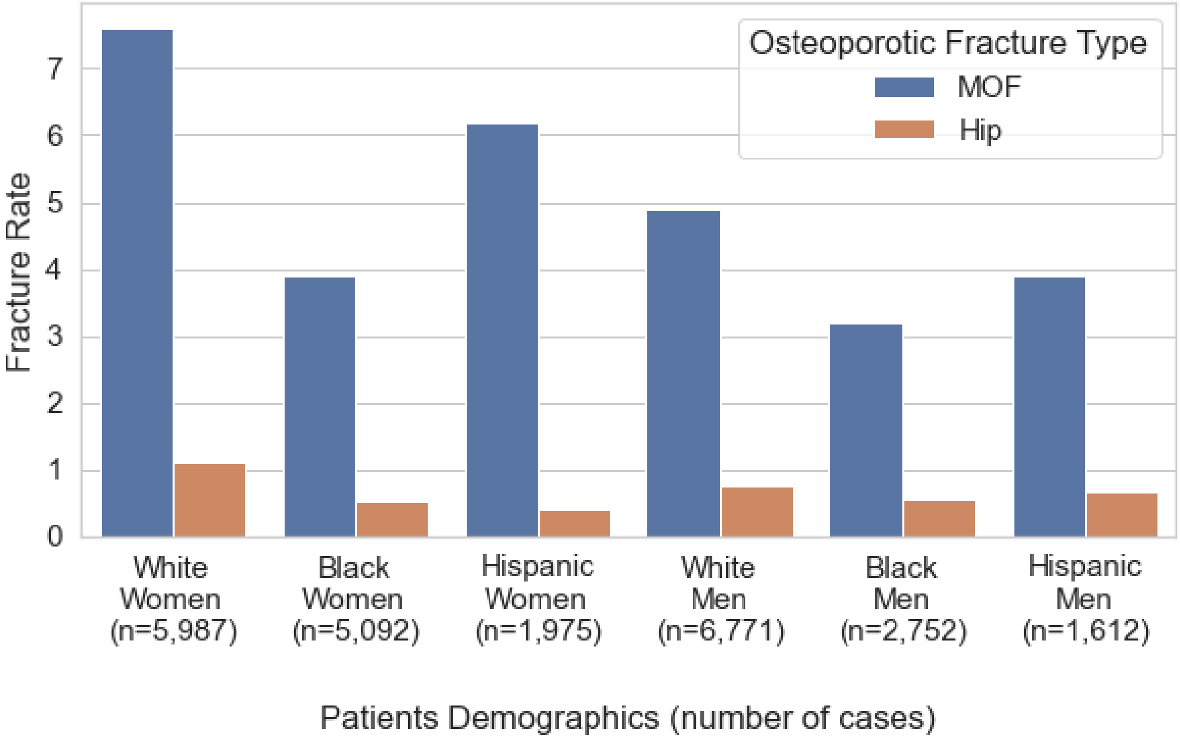


Figure 4.1. Major fracture occurrence rate across different gender and race with the number of cases noted.

Table 4.1. Osteoporotic Fracture Survival Analysis Cohort Demographics and Follow-Up Time

	White (12,758)	Black (7,844)	Hispanic (3,587)	Other (1748)
Age	64.5 ± 9.9	61.2 ± 9.3	60.2 ± 8.7	63.3 ± 7.8
Male	53.1%	35.1%	44.9%	47.7%
BMI	30.3 ± 6.4	32.3 ± 7.6	30.9 ± 6.4	31.4 ± 6.3
Average visits per year	4.8 ± 4.1	7.2 ± 6.2	7.0 ± 5.2	7.2 ± 6.7
Average years of follow-up	3.7 ± 1.8	3.9 ± 1.8	3.7 ± 1.8	3.7 ± 1.8
Prior fracture (any)	5.0%	4.0%	4.9%	2.4%
Chronic obstructive pulmonary disease	9.2%	10.3%	6.3%	4.9%
Steroids,3 months of use	0.5%	0.9%	0.5%	0.3%
Rheumatoid arthritis	1.3%	1.6%	1.6%	1.5%
Secondary osteoporosis	18.6%	22.1%	20.2%	2.4%
Charlson Comorbidity Score	0.6 ± 1.0	1.1 ± 1.6	0.9 ± 1.4	0.6 ± 1.1

Following the introduction of the osteoporotic fracture rate with respect to various demographic characteristics of patients and the covariates listed in Table 4.1, we proceeded to conduct an analysis on the development of osteoporotic fracture risk using the survival analysis method. The distribution of osteoporotic fracture occurrence time (in years) is depicted in Figure 4.2, encompassing both cases of major osteoporotic fractures (MOF) and hip fractures. Figure 4.2 illustrates that the occurrence of fractures spans a duration ranging from 1 to 7 years. Although the

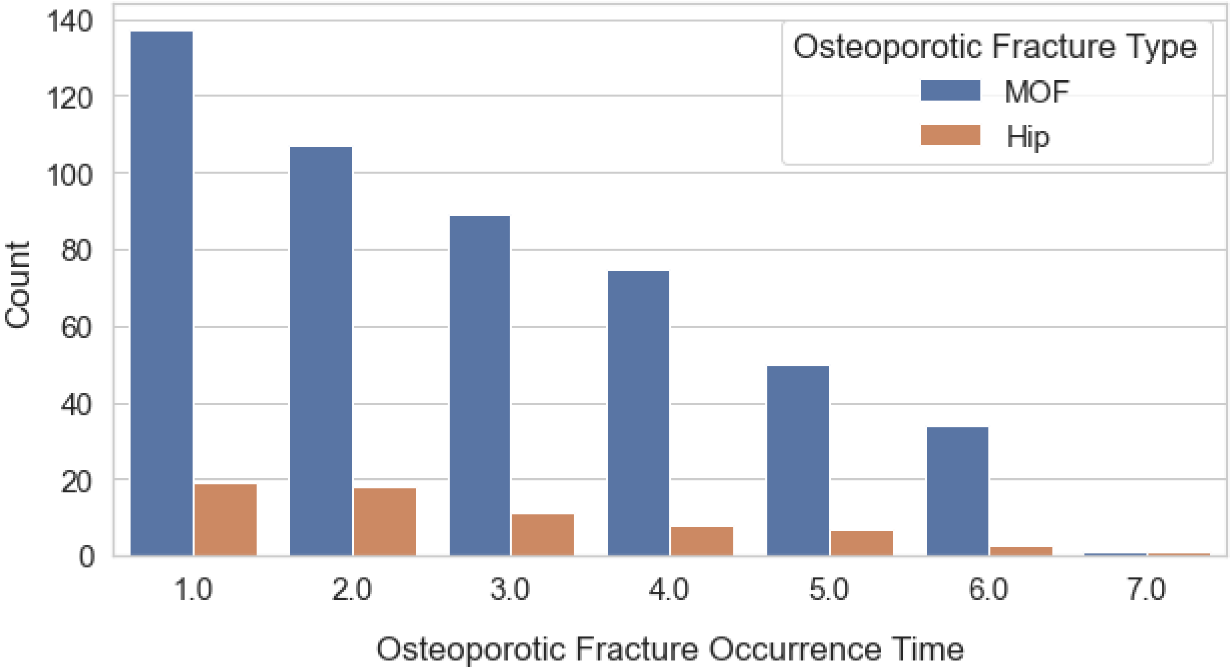


Figure 4.2. Osteoporotic fracture occurrence time (year) count plot with respect to MOF and Hip fracture type.

occurrence rate of hip fractures was considerably lower compared to MOF, in accordance with the findings presented in Figure 4.1, both types of fractures exhibited a similar pattern. The majority of osteoporotic fractures were observed within the initial year of the study, with a subsequent decrease in occurrence rate as the observation time increased.

As demonstrated in the preceding section, the survival function $S(t)$ was introduced to quantify the probability of patients experiencing osteoporotic fractures beyond a specified time point, denoted as t. To estimate $S(t)$ along with a 95% confidence interval, we employed the non-parametric Kaplan-Meier method, considering a time range spanning from 1 to 7 years. Additionally, we conducted an analysis of $S(t)$ in relation to various demographic characteristics, such as gender and race. The Kaplan-Meier estimates of $S(t)$ were separately presented for MOF

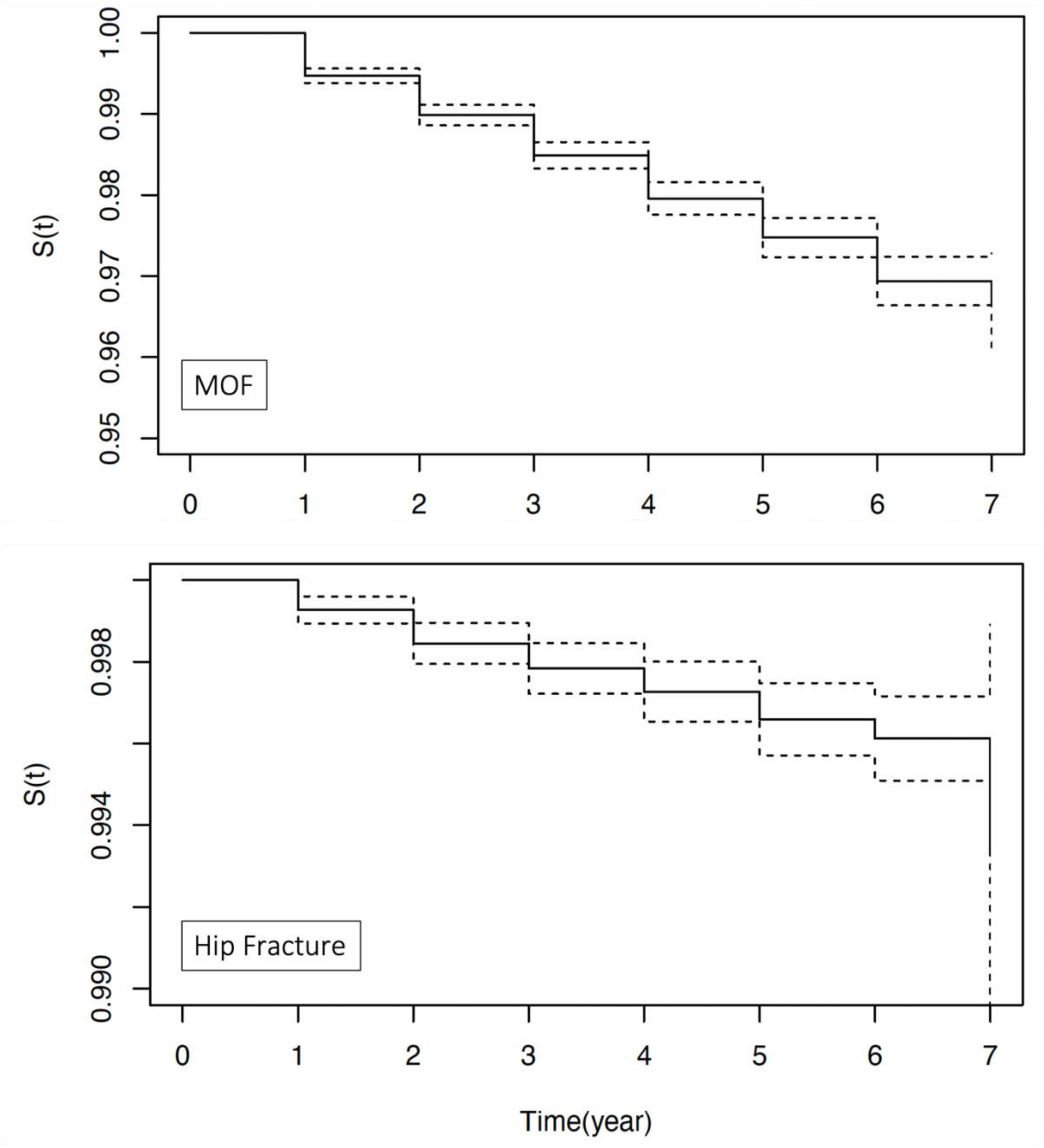


Figure 4.3. MOF and Hip fracture estimated survival function $\hat{S}(t)$ estimated from Kaplan-Meier (KM) model with 95% confidence intervals (dashed line).

and hip fractures in the subsequent figures. The survival function estimator $\hat{S}_{MOF}(t)$ and $\hat{S}_{Hip}(t)$ was shown in Figure 4.3. As we can observe that given fixed time t, $\hat{S}_{Hip}(t)$ was higher than $\hat{S}_{MOF}(t)$ due to less Hip fracture occurrence rate in compared to MOF observation. However, the confidence intervals range for $\hat{S}_{Hip}(t)$ has higher band width as well since the number of cases for Hip fracture was much smaller compared to MOF cases. In the subsequent analysis, we will investigate the survival function regarding distinct demographic characteristics, specifically race and gender. The estimated $\hat{S}_{MOF}(t)$ conditioned on demographic features has been depicted in Figure 4.4. Notably, the sub-figure displaying $\hat{S}_{MOF}(t)$ conditioned on gender reveals that, for a given fixed time t, female patients exhibit a lower probability of experiencing MOF beyond time t. In other words, female patients demonstrate a higher susceptibility to MOF risk, which escalates as the duration increases. Regarding the impact of race on the survival function, as illustrated in Figure 4.4, notable observations emerge. Among all racial groups, Caucasians exhibit the highest susceptibility to MOF, while Blacks demonstrate the highest resistance. A similar influence of demographic factors on the estimated survival function $\hat{S}_{Hip}(t)$ for hip fractures can be observed, as depicted in Figure 4.5. However, the disparity between males and females is less pronounced compared to $\hat{S}_{MOF}(t)$. In Figure 4.5, the risk of hip fractures appears to converge as the duration increases for both males and females. Concerning the impact of race, Caucasians experience the highest probability of hip fracture, whereas Blacks exhibit the lowest risk at any given fixed time t. It is worth noting that the flat green line observed after the first year is a result of missing tracking or a small sample size for cases involving other racial groups.

The influence of various demographic factors on the estimated survival function $\hat{S}(t)$ for both MOF and hip fractures has been observed significantly, as evident from Figure 4.4 and Figure 4.5. Consequently, to estimate the survival function $\hat{S}(t)$ considering these factors, we developed parametric models, as outlined in Table 4.1. Firstly, we randomly partitioned the dataset into separate 80% of training and 20% testing sets, ensuring no data leakage. The training dataset was utilized to construct multiple supervised models, while the testing dataset was employed to evaluate the performance of each model using a pre-defined evaluation metric from the previous section, specifically the concordance-index (C-index). The results of this evaluation are presented in Table 4.2.

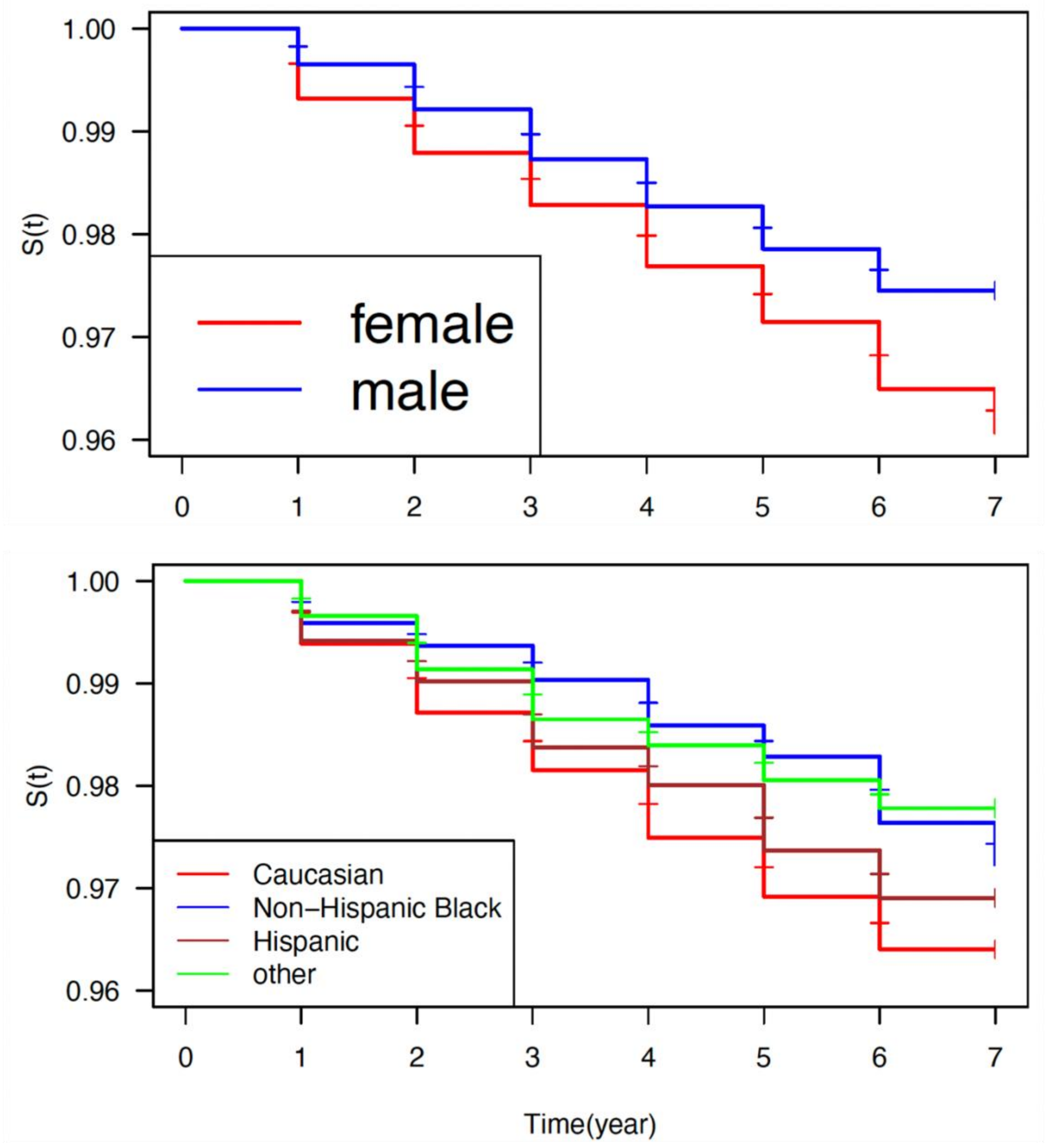


Figure 4.4. MOF estimated survival function $\hat{S}_{MOF}(t)$ estimated from Kaplan-Meier (KM) model conditioned to gender (upper Figure) and race (lower Figure).

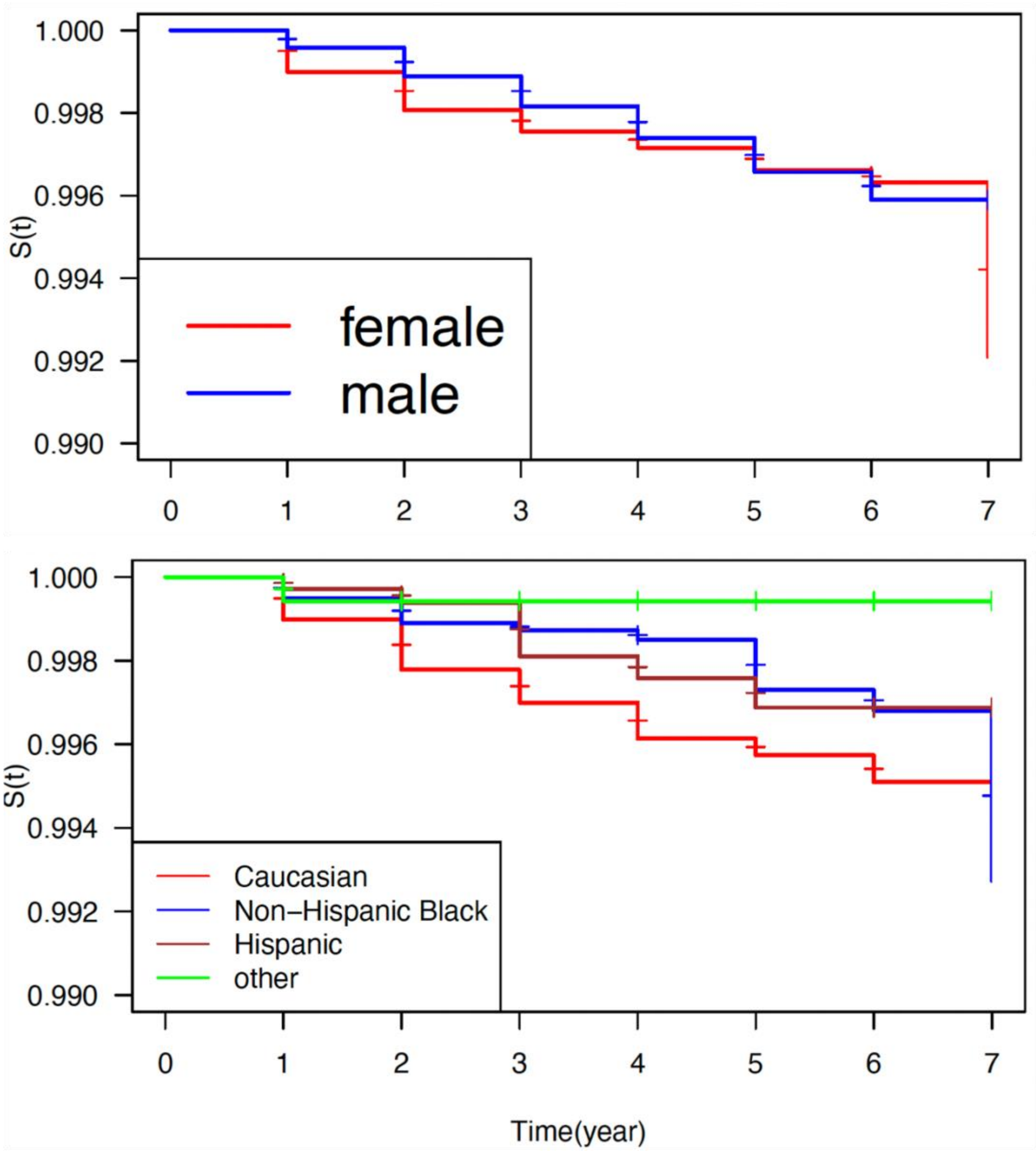


Figure 4.5. Hip estimated survival function $\hat{S}_{Hip}(t)$ estimated from Kaplan-Meier (KM) model conditioned to gender (upper Figure) and race (lower Figure).

Table 4.2. Models Concordance-Index Score to Test Dataset

Model	**Osteoporotic Fracture Type**	
	MOF	**Hip**
Cox Proportional Hazard	0.688	0.791
Random Forest	0.630	0.784
Gradient Boosting	**0.695**	**0.809**

As presented in Table 4.2, three distinct machine learning models were trained, and their performance was evaluated on the testing dataset using the concordance-index score. The Cox-Proportional-Hazard model (CoxPH), a widely employed linear regression structure statistical model in survival analysis, was utilized as the benchmark in this study. Random Forest (RF) and Gradient Boosting (GB), both tree-based machine learning approaches with Cox's proportional hazards log-likelihood serving as the target loss function $f(x)$ (as explained in the preceding section), were also employed. Observing the results, it becomes apparent that GB achieved the highest concordance-index score among the three models for both major osteoporotic fractures (MOF) and hip fractures. This outcome suggests that the GB model is most effective in assessing the risk of an individual being observed with an osteoporotic fracture outcome at a given time. Interestingly, it is worth noting that the models performed better in the case of hip fracture datasets compared to MOF datasets. This finding contrasts with the observations depicted in Figure 4.4 and Figure 4.5, where demographic factors exhibited a more significant influence on MOF exposure risk. These results underscore the distinct discriminatory ability of hip fractures based on different factors, highlighting their divergence from the patterns observed in the mentioned figures before.

To further study the gradient boosting model predictive performance regarding varies demographic features. We stratified the sample population into multiple subgroups according to different combination of race and gender and then measured the performance of gradient boosting model on each subgroup by their concordance-index score with 95% confidence intervals. The model's predictive performances on MOF were shown in Figure 4.6 and on hip fracture were shown in Figure 4.7. As shown in both Figures, we have observed that model delivers better results on female group in comparison to male group in both MOF and hip fracture outcome, which is consistent with our Kaplan-Meier estimators that female group experience higher risk at compared to man at the same time. Besides, different race types also influence the model's performances significantly for both MOF and hip fracture outcome. The model obtains higher score on White people, while model suffers worst predictive power on Other races. This result was consistent with Kaplan-Meier estimators as well since White people endure the most risk in compared to other races at given time, however, the Other race have flat survival function after first few years which maybe resulted from censorship. Moreover, the model's performance on non-Hispanic Black and Black people shows different results in MOF and hip fracture outcome, as on MOF outcome model gives better results on non-Hispanic Black but vise-versa on hip fracture outcome.

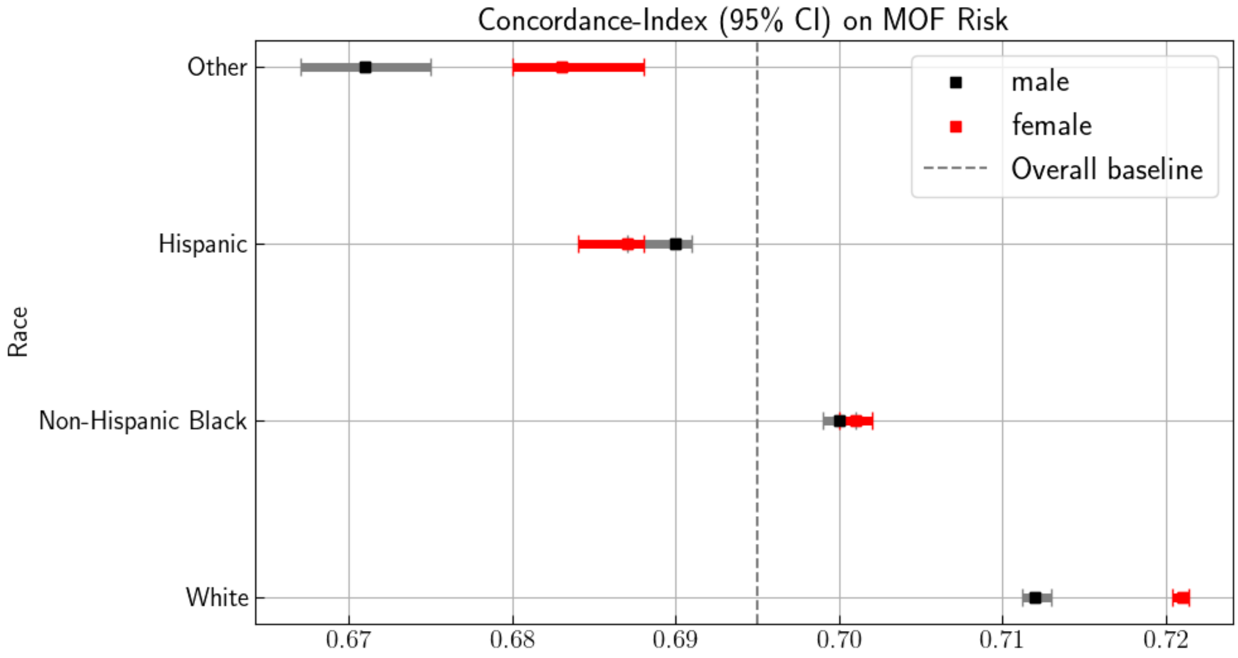


Figure 4.6. Gradient Boosting model predictive performance on MOF risk development measured by concordance-index with 95% confidence intervals across multiple demographic subgroups.

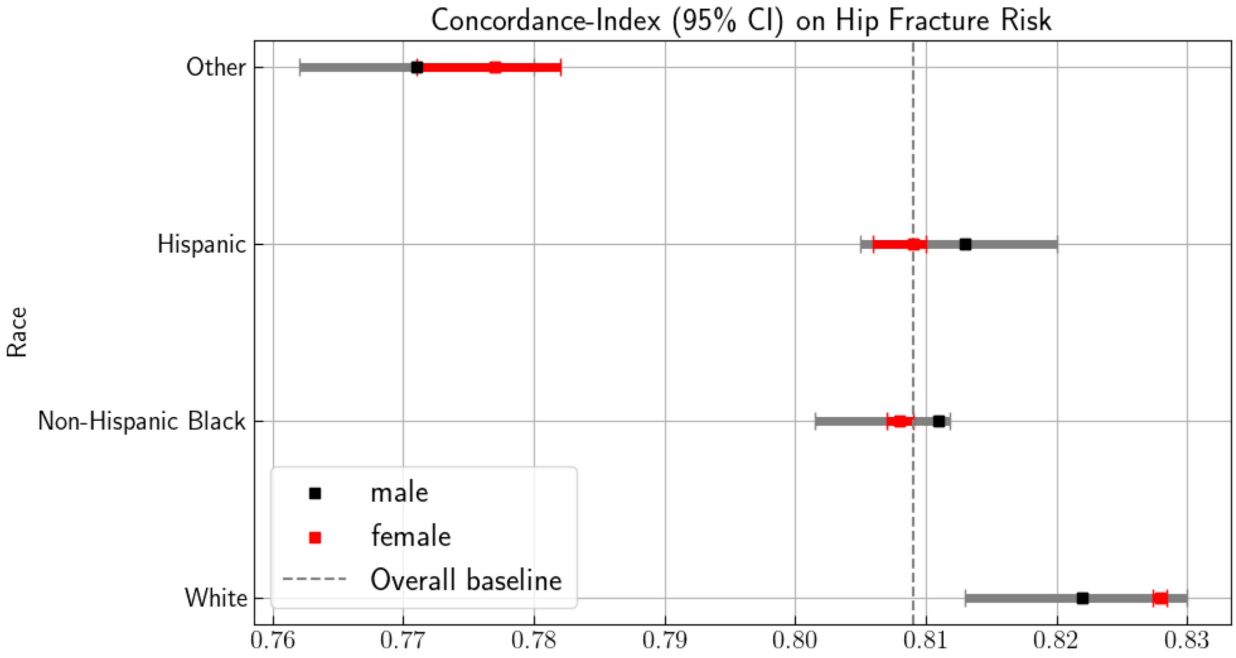


Figure 4.7. Gradient Boosting model predictive performance on Hip risk development measured by concordance-index with 95% confidence intervals across multiple demographic subgroups.

4.4 Discussion

This study demonstrates that training machine learning model derived from the electronic health record of a large urban medical center can accurately discriminate between high and low fracture risk subjects receiving primary care in the US. Furthermore, we have achieved state-of-art concordance-index using gradient boosting model, our study includes Black and Hispanic subjects and men which have usually been under-represented in osteoporosis risk investigation. These findings support that the use of electronic health record in population management approaches can be powerful to facilitate care of patients at high risk of fracture in more generous race group. Even for Hip fracture occurrences which is generally a rarer fracture outcome as shown in Figure 4.1 and Figure 4.2, our models have shown superior predictive power to patients across different races.

Our findings show evidence to use patients' electronic health record as predictors to build survival analysis model on future fracture risk forecasting, which enable automated fracture risk calculation which could improve osteoporosis screening in several ways. For instance, fracture risk calculations could be directly integrated into electronic health record, which allows primary care physicians to be alerted at the point of care for patients at high risk of fracture. Since the datasets in this study is one of the largest validations in the US minority population, where most of existed studies population weighs heavily on Whites, our research addresses the need to evaluate the fracture algorithm performance in multiethnic population. As shown in our estimated survival function using Kaplan-Meier methods, it indicates the significant difference of fracture risk development across different race which is commonly underestimated by research before. Besides, our model's risk analysis performance on stratified population subgroups in different demographic conditions supports the arguement that race and gender are critical factors to be included when training a model, otherwise, the biased dataset would result in biased predictive results for minority groups.

Overall, this study validates the use of electronic health record generated fracture predictions in the US for the first time and adds to the evidence for the use of race or ethnicity to train survival analysis model with state-of-art concordance-index score. The result supports that the electronic health record inputs allow automate fracture risk predictions that provide trustworthy discrimination over several years follow-up without patient or provider effort. The trained model can perform in clinical population with high rates of comorbid disease and with substantial racial-ethnic validation in fracture rates. At last, the result of this project demonstrates that the inclusion of race improved fracture prediction and could help target those in need of osteoporosis screening or treatment.

CHAPTER 5 PREDICTING HIP FRACTURE IN PATIENTS WITH DIABETES FROM A BROAD RANGE OF PRIOR DIAGNOSIS

This chapter is adapted from the manuscript draft in preparation for publication "**Predicting Hip Fracture in Patients with Diabetes from a Broad Range of Prior Diagnosis**" authored by Chengjian Shi, Atif Khan, Robert L. Grossman, Rajesh Jain and Andrey Rzhetsky.

5.1 Introduction

There are 37.3 million people in the United States with type 2 diabetes (T2DM). In addition to well-known complications of diabetes, such as retinopathy or neuropathy, people with T2DM also have an increased risk of fragility, or low trauma, fracture. Studies suggest 25-60% higher risk of fragility fracture and 200%-500% higher risk of hip fracture as compared to those without T2DM [90-94]. With long-term follow up, approximately 50% of patients with T2DM will experience a fracture [95]. Proposed mechanisms for the increased fracture risk includes the higher risk of falls, reduced bone turnover, and worse mechanical properties of bone, though some uncertainty remains [96-98].

Fragility fractures are a significant public health problem, occur 9 million times per year worldwide [99], lead to significant disability and excessive mortality [99-100], and cost $19 billion per year in the United States alone [101]. Indeed, not only do people with T2DM have a higher rate of fracture, but they also experience higher morbidity from fractures that do occur, including poor wound healing, higher rates of delayed union, and worse recovery of function [102-105]. However, proper identification of people with T2DM who are at high risk of fracture is a clinical challenge.

Fracture risk is assessed using fracture risk calculators, with the Fracture Risk Assessment Tool, or FRAX, being the most widely used. FRAX combines demographics, BMI, and clinical risk factors for fracture to estimate the 10-year risk of fracture, but it does not include diabetes as a risk factor. The tool has generally been well-validated; however, in cohorts with T2DM, FRAX underestimated fracture risk [90, 92].

If the traditional tools do not work well in people with T2DM, a specific tool for people with T2DM will be needed. Uncertainty remains about the reasons for increased fracture risk in T2DM, and there may be novel, yet-to-be identified risk factors for fracture in people with T2DM. The use of a large, national database may be key in studying diabetes-related fracture risk since the number of fracture events gives the statistical power to see rare, but important, predictors. Furthermore, a machine learning tool created from a large commercial claims database may improve upon the accuracy of existing tools. Machine learning methods are now being utilized in medicine with promising results, such as the identification of early sepsis or the automated identification of pneumonia on chest X-ray [106-107]. Because these methods make fewer assumptions about the relationships between variables, they have often shown superior predictive ability than traditional statistical methods.

Because of the disability related to fractures in those with diabetes and the lack of available methods to identify those at risk, our study proposes to identify novel risk factors for fracture in people with diabetes and develop a diabetes-specific tool using machine learning. We will do so using data from IBM Health MarketScan database, a large U.S. claims database.

5.2 Methods

5.2.1 Data and other materials

Population Datasets: The study population is from the electronic health record (EHR) of MarketScan dataset. The MarketScan dataset was originally complied by Truven Health from records of over a hundred large insurance companies in the US. The data set we intend to use contains 5 billion records of diagnosis for over 180 million unique patients in the US. For each entry of diagnosis, the database documents the date, age of patient, and a code in International Classification of Diseases, 9th or 10th Version's Clinical Modification (ICD-9-CM or ICD-10-CM). MarketScan data allows linking patients into families (through insurance policies), which allows us to disentangle genetic and environmental information.

The screening process to obtain the target patients group was designed to guarantee the continuous patients' enrollment for at least one year from MarketScan dataset, then we set the age threshold for at least 20 years before or during the enrollment period. The diabetes diagnosis was determined with specific ICD-9/ICD-10 code on at least 2 days of service and the less than 2 years apart of clinical visiting. The target diabetes mellitus was limited to type-2 diabetes for chronic diseases control, which means we've excluded primarily type-1 cases ICD codes when screen the MarketScan dataset. Besides, the type-2 patients were further filtered to the group with age threshold for at least 20 years. This sampled group with type-2 diabetes with population size over 10,000,000 cases, we applied the principle to determine the occurrence of hip fractures to further divide this sampled group into hip fracture positive outcome patients sub-group with sample size of 101,017 and negative outcome patients sub-group with sample size of 10,846,673, which indicates less than 1% of patients with type-2 diabetes have presence of hip fracture across the entire target population. The data flow for the screening process is presented in Figure 5.1.

Determination of Fractures: The hip fracture outcome for this population was determined by ICD-9/ICD-10 code recorded in MarketScan EHR. The target hip fracture outcome was hand-picked by Dr. Rajesh Jain that shows diagnosis codes in the principal position on the inpatient anchor hospitalization claim indicates specific hip fracture. The details ICD coding algorithms for fracture diagnosis codes and excluding trauma codes were selected based on multiple fracture subcategories.

5.2.2 Statistical Analyses

In this project we focused on predicting hip fracture outcome for patients conditioned with history of diabetes. Besides, instead of using selected features computed by domain expertise like we conducted the survival analysis project before, here we took advantage of well-structured diseases record as covariates candidates and enormous data size from MarketScan to build powerful prediction model. As discussed before, the step-by-step data screening process was displayed in next section with annotated population number in each step and detailed demographic information for filtered targeted groups was shown in Table 5.1.

The numerous recorded diseases ICD-9/ICD-10 codes were carefully categorized into 567 disease bags that represented essential diseases record with common characteristics that could be used as candidate predictors to major hip fracture outcome. In addition, we have removed hip fracture identity ICD codes related disease bags to avoid data leakage in further model modeling section. Similar to the cohort association study we have done in Chapter 2 for acute pancreatitis and diabetes mellitus, we studied the association between each disease group and major hip fracture outcome regarding diabetes mellitus patients' population only. Odds ratio (OR) was selected as the measurement of association strength since we've compared case-control study for each disease group and major hip fracture outcome with the same demographic features such as age group and gender, the odds ratio was computed with 95% confidence interval to better illustrate the association strength distribution for each disease group to major fracture outcome. To specify the computation procedure, we have defined some important variables as:

1. Disease bag b_i is a category of ICD-9/ICD-10 codes that share common characteristics selected by domain expertise.
2. Disease group D_i is a collection of patients with MarketScan record of corresponding disease bag b_i. Relatively, $\overline{D_i}$ is the collection of patients without record of specific disease bag ICD-9/ICD-10 codes.
3. Hip fracture outcome H is a collection of patients with specific collection of major hip fracture ICD-9/ICD-10 codes that satisfy the determination of fracture criteria. Relatively, $\overline{H}$ is the collection of patients didn't satisfy the determination of fracture criteria.
4. The gender of patients as $G \subseteq$ {male, female}.
5. Patients cluster divided by an age group Z from 20 ~ 60 years old.
6. Level of two-sided confidence interval α. In this project we used α as 5% to achieve 95% confidence intervals.
7. Odds ratio θ_i for the association strength between each disease group D_i and major hip fracture outcome H.

Within the entire MarketScan diabetes mellitus population, we have built contingency table represented by Table 5.1 regarding each disease group D_i and major hip fracture outcome H for counted patients matched with same gender G and age group Z.

Table 5.1. Contingency table for disease group D_i and hip fracture outcome H

	$\boldsymbol{H}$	$\overline{\boldsymbol{H}}$
$\boldsymbol{D_i}$	n_{11}	n_{12}
$\overline{\boldsymbol{D}}_{\boldsymbol{i}}$	n_{21}	n_{22}

* The counted patients $(n_{ij}, \forall i, j)$ in each cell are matched with same gender G and age group Z

Furthermore, the odds ratio θ_i along with its 95% confidence intervals range were computed shown below as:

$$\theta_i = \frac{n_{11} * n_{22}}{n_{12} * n_{21}}$$

$$Var(\log(\theta_i)) = \sum\nolimits_{i,j} \left(n_{ij}\right)^{-1}$$

Then we can construct the approximate standard error and 95% confidence intervals derived for the OR as:

$$SE(\log(\theta_i)) = \left(\sum\nolimits_{i,j} \left(n_{ij}\right)^{-1}\right)^{\frac{1}{2}}$$

$$CI = OR \pm z_{1-\frac{\alpha}{2}} \times SE(\log(\theta_i))$$

The computed odds ratio collections $\{\theta_i\}$ for the entire disease groups $\{D_i\}$ was further ranked and compared by domain expertise to analysis underlying clinical insights.

Once we acquired the comprehensive association strength for each disease bag and major hip fracture outcome, the next step was to construct a predictive model capable of forecasting the major hip fracture outcome H. This model utilized a combination of disease bags $\{b_i\}$ for each patient's MarketScan medical history as predictive features, along with demographic information such as age group $\{Z\}$ and gender $\{G\}$. Besides, it's challenging to build a predictive machine learning model due to low occurrence rate of disease bag $\{b_i\}$ in each patients' health record, leading to difficulties in handling high-dimensional sparse features, which indicates that most of the features have a value of zero or missing data, resulting in lack of information for training the model effectively. Therefore, we have utilized some strategies to control the overfitting problem when training the model, such as regularization to control predictive model complexity, varies optimization methods to avoid loss function trapped in local minimum and weighted sampling methods to penalize false prediction on under-sampled cases.

The model training started with train/test split on the available dataset into two distinct subsets, due to the enormous sample size we have filtered from MarketScan electronic health record as discussed above, we split 2/3 of dataset as train dataset which would be used to develop the model and 1/3 of dataset as test dataset which be used to validate the model performance given unseen dataset. To evaluate the model performance, we chose ROC-AUC (Receiver Operating Characteristic – Area Under Curve) score to measure the model classification ability. There were some reasons to choose ROC-AUC score that quantify the models' discrimination ability regarding our dataset characteristics. Firstly, the ROC-AUC score is robust to evaluate highly imbalanced class distribution for major hip fracture outcome, the positive case percentage is around 0.92% among the entire sample population. Secondly, the ROC-AUC score is insensitive to decision threshold, which provides an aggregated measure of its overall classification ability. Last, we can evaluate the models' probabilistic predictions rather than binary outcome which allows for analyzing the quality of those probability estimators by considering the trade-off between true-positive-rate and false-positive-rate. Apart from using ROC-AUC score as general metric to evaluate model performance, we have also used true-positive-rate (TPR) to analyze model's ability to correctly identify positive instances to from the total actual positive instances.

Next, we trained two commonly used classification models for tabular dataset, stochastic gradient descent linear (SGD) model and classification-and-regression-tree extreme gradient boosting (XGB) model. Both model families have superior characteristics to clinical problem such as high interpretability and well-developed communities. For the estimators implements SGD learning, two important hyper-parameters were carefully explored to find the best combination to fit the dataset: (1) the loss function was arbitrary chosen to utilize varies linear models including SVM, perceptron and logistic regressor, (2) the penalized regularizer was added to loss function to shrink the model complexity in different style. The XGB estimators were more versatile and have high capacity in comparison to SGD estimators, which making them capable to capture intricate non-linear relationship and interactions in the data and more expensive to tune the complicate hyper-parameters as well. Both models' hyper-parameters were fine-tuned using 5-fold cross-validation to maximize their ROC-AUC score within train dataset.

5.3 Results

The study cohort for this research project specifically focused on individuals with a documented history of diabetes. The process of filtering and selecting qualified patients from the MarketScan dataset is illustrated in Figure 5.1. The MarketScan dataset encompassed a total population of over 200 million patients enrolled between 2003 and 2020. To minimize the influence of outliers, we initially screened patients with a minimum enrollment duration of one year and an age of 20 years or older. This resulted in approximately 111 million potential candidates for further screening. Subsequently, we identified patients with a documented medical history of Type-2 diabetes mellitus (T2DM) using a set of predefined criteria to ensure accurate diagnosis based on the MarketScan electronic medical records. Consequently, we identified approximately 10 million eligible patients with T2DM. As the primary objective of this project was to predict the risk of hip fractures in patients with a history of T2DM, we utilized the hip fracture International Classification of Diseases (ICD) codes provided by Dr. Rajesh Jain to discern patients who were likely to develop hip fractures in the future. As illustrated in Figure 5.1, our analysis revealed that approximately 101,000 patients with T2DM were projected to develop hip fractures, representing less than 1% of the initially screened patients with T2DM. The occurred time of hip fracture after observing T2DM was shown in Figure 5.2, as shown that almost 50% of hip fracture could be observed within the first 5 years and then the risk dropped significantly to less than 3% year by year until maximum range of 17 years.

The demographic factors we've included from MarketScan dataset in this project were participants' age and gender, the detailed demographic information for patients conditioned with T2DM were shown in Table 5.1. As we can observe from the demographic statistics that eligible patients with T2DM have different distribution of gender and age with respect to whether they would develop hip fracture. The hip fracture group have larger percentage of male cases and higher average age value in comparison to group without occurrence of hip fracture.

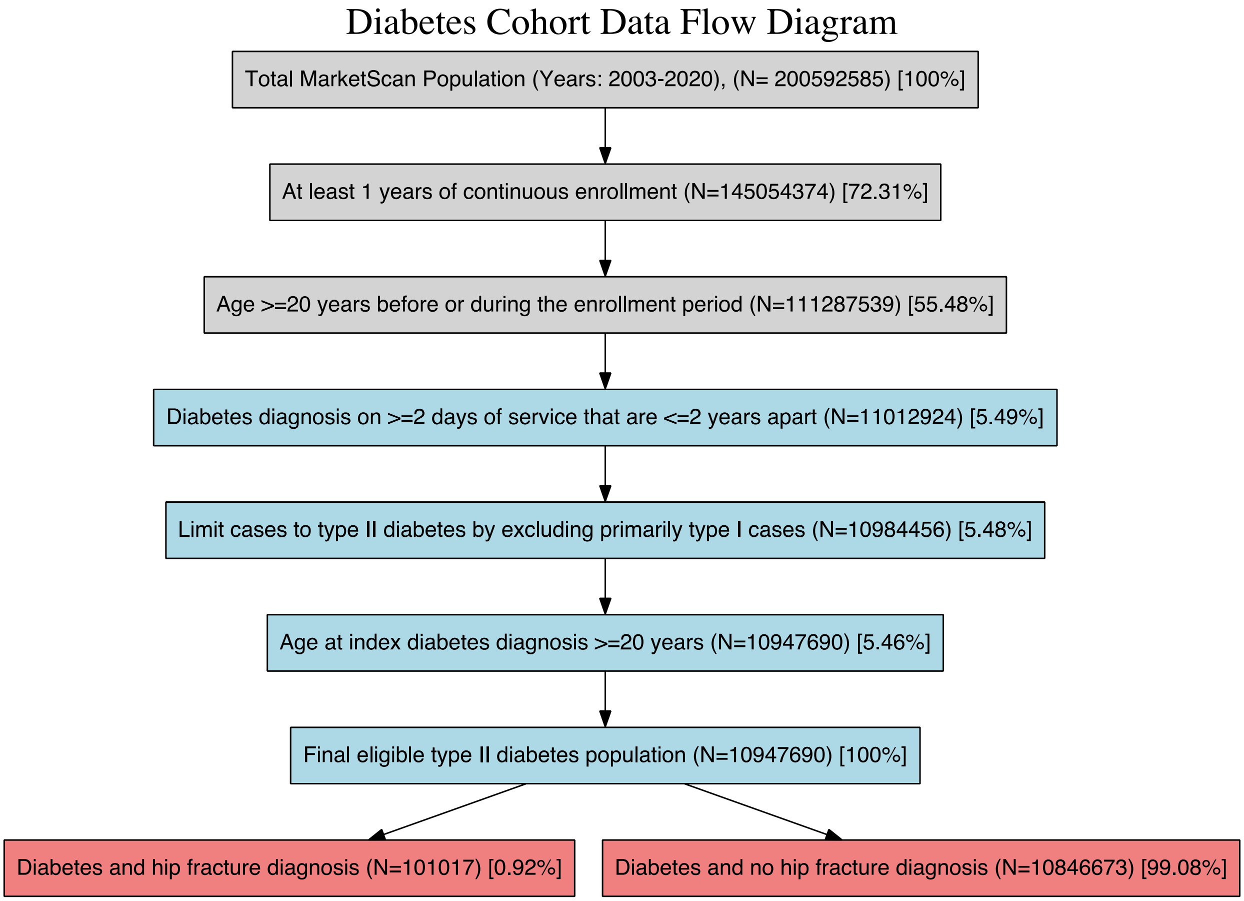

Diabetes Cohort Data Flow Diagram
Total MarketScan Population (Years: 2003-2020), (N= 200592585) [100%]
At least 1 years of continuous enrollment (N=145054374) [72.31%]
Age >=20 years before or during the enrollment period (N=111287539) [55.48%]
Diabetes diagnosis on >=2 days of service that are <=2 years apart (N=11012924) [5.49%]
Limit cases to type II diabetes by excluding primarily type I cases (N=10984456) [5.48%]
Age at index diabetes diagnosis >=20 years (N=10947690) [5.46%]
Final eligible type II diabetes population (N=10947690) [100%]
Diabetes and hip fracture diagnosis (N=101017) [0.92%]
Diabetes and no hip fracture diagnosis (N=10846673) [99.08%]

Figure 5.1. Data screen process with criteria to discriminate qualified patients with type-2 diabetes mellitus that would develop hip fracture in the future.

Table 5.2. Patients with Type-2 Diabetes Mellitus (T2DM) Cohort Demographics

	Eligible Patients with T2DM		
	Hip Fracture	**Non-Hip Fracture**	**Overall**
Male	61.9%	48.1%	48.4%
Age	63.9 ± 11.9	54.1 ± 13.9	54.3 ± 14.1

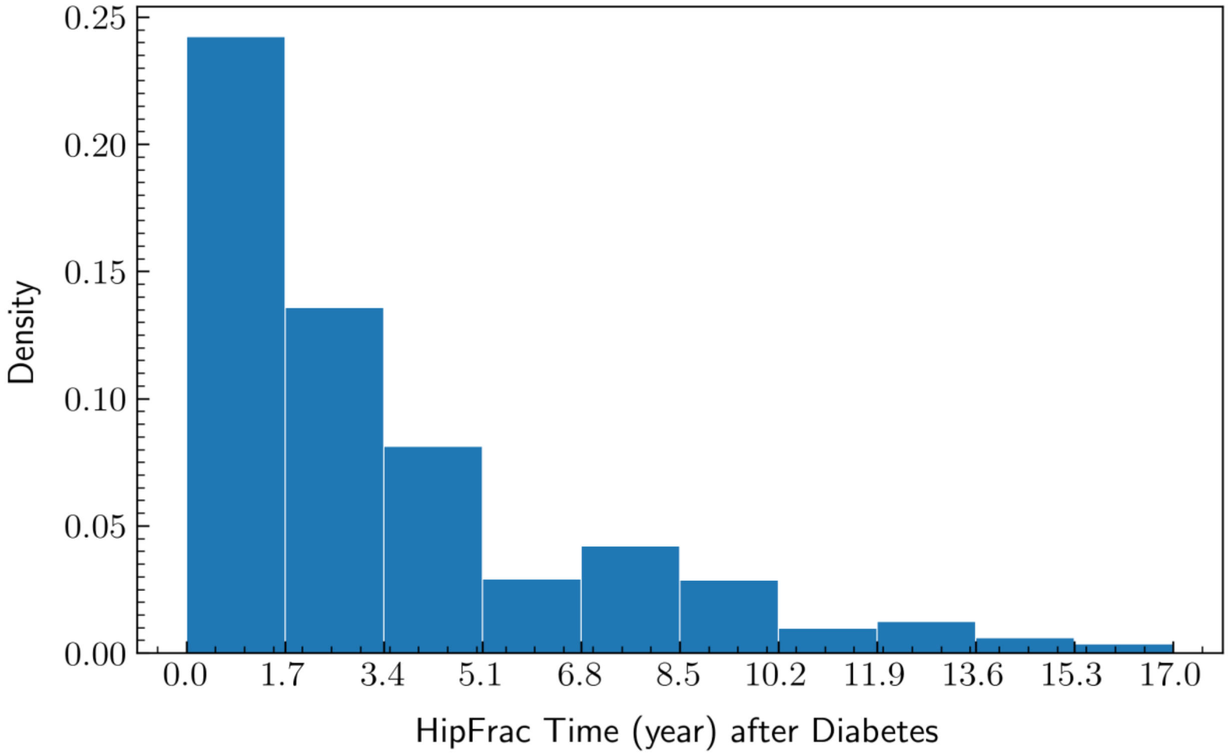


Figure 5.2. Density of sample size observe Hip fracture after diagnosed with type-2 diabetes.

The primary objective of this research project was to utilize the electronic disease records of patients with Type-2 diabetes mellitus (T2DM) to predict future hip fracture occurrences. To achieve this, we employed a categorization approach wherein thousands of ICD-9/ICD-10 codes were grouped into 566 distinct disease bags, as defined in the previous section as $\{b_j\}$. Subsequently, we binary encoded the presence or absence of disease bags within each patient's electronic disease record as predictors for forecasting hip fracture outcomes, alongside gender and age information. To begin, we conducted a cohort association study to examine the relationship between each disease bag b_j and the occurrence of hip fractures. We measured the association using odds ratios, accompanied by 95% confidence intervals, computed through case-control cohorts with matched age groups and genders. Figure 5.3 displays the top 50 ranked odds ratios for each disease bag.

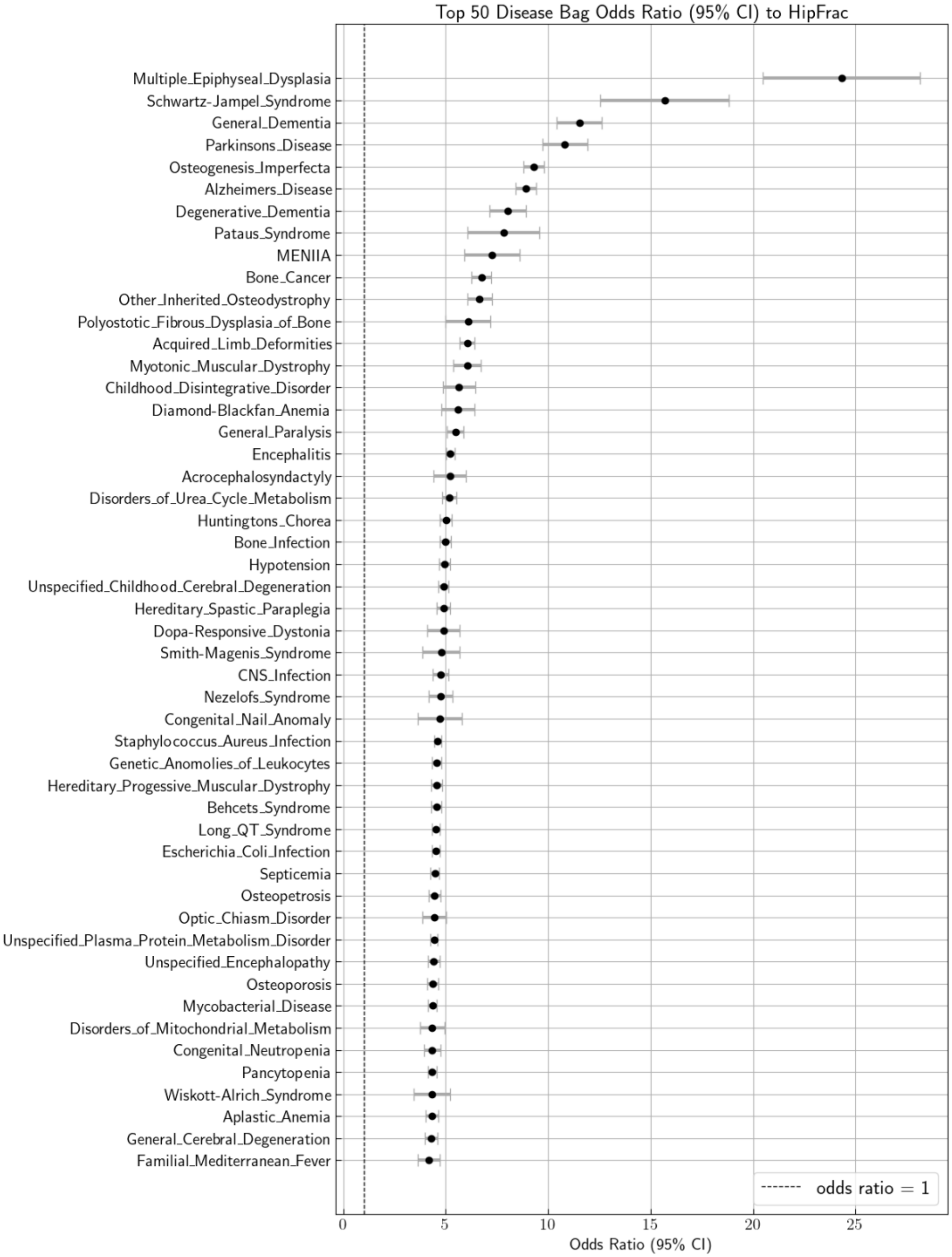


Figure 5.3. The top 50 ranked odds ratio value for each disease bag to Hip fracture with 95% confidence intervals.

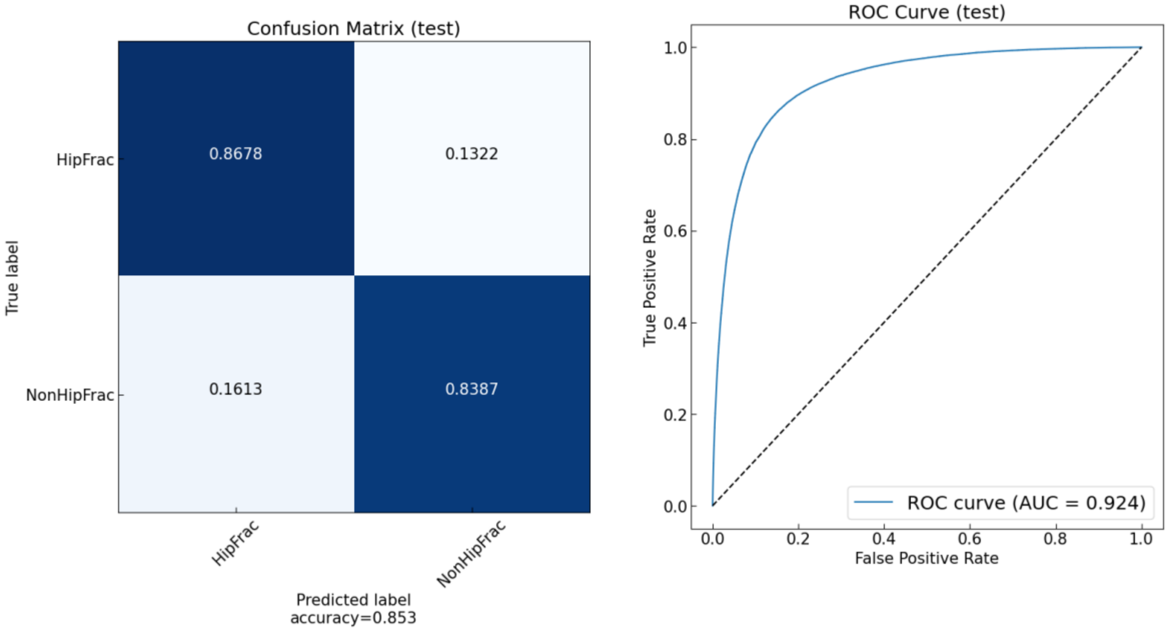


Figure 5.4. Confusion matrix and ROC-AUC curve for XGB predictive performance on testing dataset.

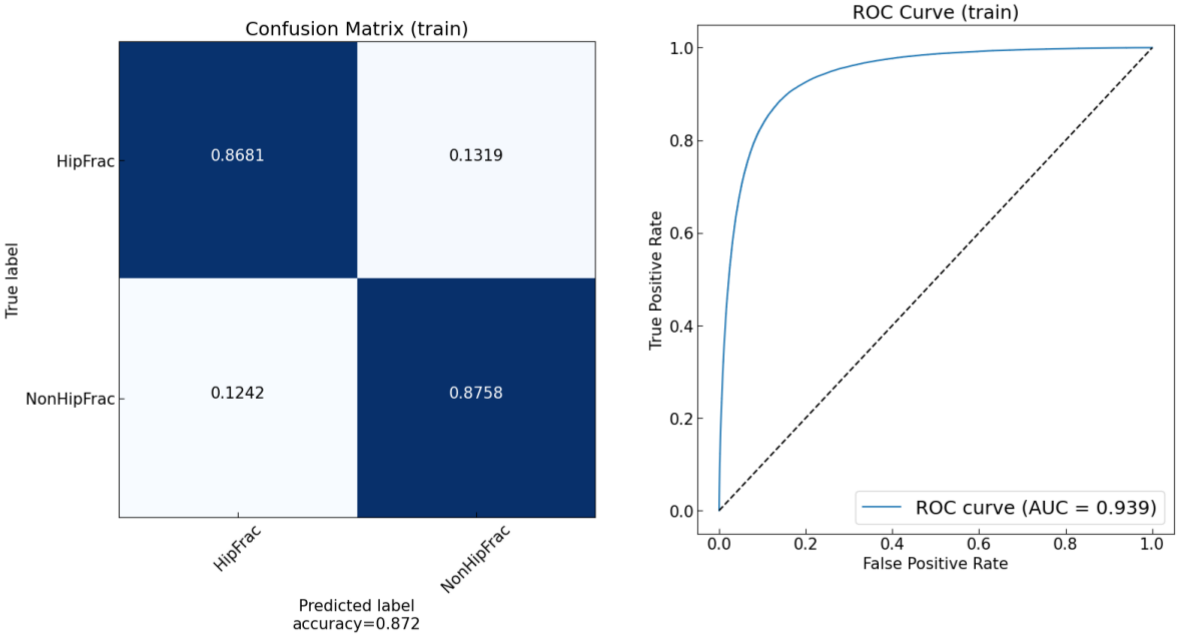


Figure 5.5. Confusion matrix and ROC-AUC curve for XGB predictive performance on training dataset.

Notably, nearly all disease bags $\{b_j\}$ demonstrated a positive correlation with hip fractures. The ranked disease bags tended to converge towards the dashed line (odds ratio equal to 1), indicating no association as the ranking decreased. Among the top 50 ranked disease bags, we observed a diverse range of diseases, encompassing physical conditions like bone cancer and osteogenesis imperfecta, as well as mental illnesses such as general dementia and Alzheimer's disease. While domain experts could intuitively explain the positive effects of certain disease bags on hip fracture development based on empirical observations, there were other confounding disease bags that could potentially contribute to hip fracture occurrence but have not been studied yet. This study provided valuable insights into the confounding associations between each disease bag and hip fracture outcomes. Significant positive associations were observed through the case control method. As a result, we will proceed to apply machine learning tools to leverage the entire collection of disease bags $\{b_j\}$ to predict hip fracture outcomes.

Table 5.3. ROC-AUC Score for SGD and XGB Predictions on Eligible Patients with T2DM

	ROC-AUC Score	
	Train	**Test**
SGD	0.843	0.834
XGB	**0.939**	**0.924**

In accordance with the methods described in the previous section, we trained two primary models: the stochastic gradient descent linear model (SGD) and the extreme gradient boosting model (XGB). The predictors employed in these models encompassed the binary encoded collection of disease bags, as well as demographic factors like gender and age. Additionally, we

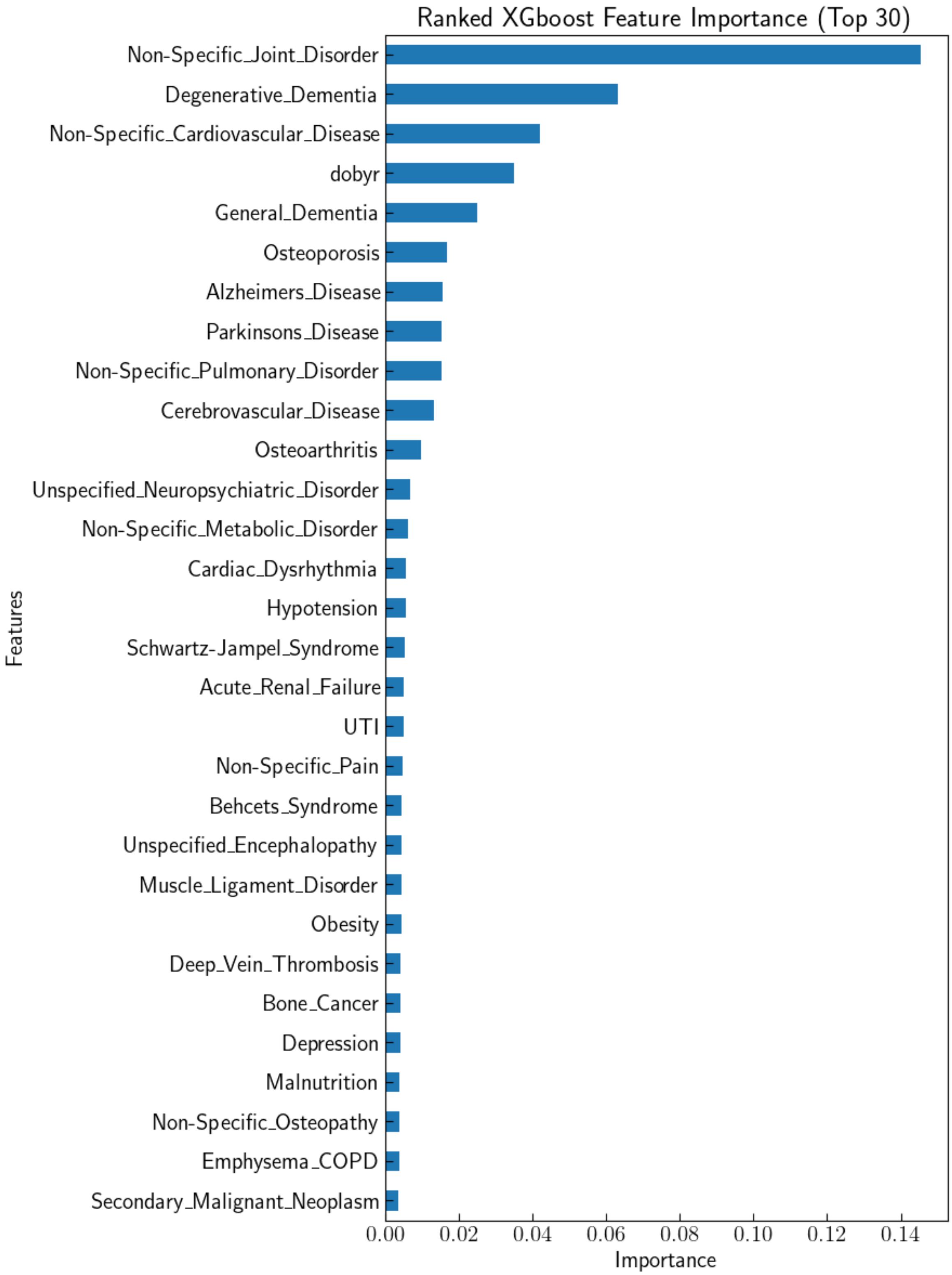


Figure 5.6. Top 30 relative importance of predictive features in XGB experiments.

stratified patients' age into four groups, with a 20-year gap between each group. Table 5.3 presents the optimized performance results of both models on the training and testing datasets, measured in terms of the ROC-AUC score. It is evident that the XGB model outperformed the SGD model in patients with Type-2 diabetes mellitus (T2DM). To gain further insight into the performance of the XGB model on both datasets, we depicted the confusion matrix and ROC-AUC curve in Figure 5.4 and Figure 5.5, respectively. The XGB model demonstrated a prediction accuracy of 0.872 for the training dataset and 0.853 for the testing dataset. Figure 5.6 displays the top 30 predictors sorted by their relative importance, from highest to lowest, as determined by the XGB model. It is noteworthy that many of the highly important predictors derived from the XGB model's encoded disease bags also appeared among the disease bags that exhibited significant associations with hip fractures, such as degenerative dementia, Parkinson's disease, and obesity, among others. A comprehensive list of disease bags that appeared in the top ranks of both the cohort association study and the XGB model's feature relative importance is presented in Table 5.4.

Table 5.4. Top Ranked Diseases Bags with Specific Ranking Number in Both XGB Relative Feature Importance and Cohort Association Odds Ratio Rank

Disease Bag	**XGB Relative Feature Importance Rank**	**Cohort Association Odds Ratio Rank**
Alzheimer’s Disease	6	5
Behcets Syndrome	19	33
Bone Cancer	24	9
Degenerative Dementia	4	2
General Paralysis	38	16
Hypotension	14	22
Osteoporosis	5	41
Parkinson’s Disease	7	3
Schwartz-Jampel Syndrome	15	1
Septicemia	36	36
Staphylococcus Aureus Infection	48	30
Unspecified Encephalopathy	20	40
Unspecified Plasma Protein Metabolism Disorder	47	39

Thus far, we have achieved a state-of-the-art ROC-AUC score using the complete collection of disease bags as predictors to forecast patients' future hip fracture outcome. Moreover, the optimized extreme gradient boosting (XGB) model has identified important features that align with our cohort association study. In addition, we sought to analyze the XGB model's predictive performance across various subgroups within the sample population. To accomplish this, we stratified the population based on different conditions, including gender, age group, and the presence of specific disease bags. To evaluate the model's performance in these subgroups, we employed the true positive rate (TPR), also known as sensitivity, which is a critical measure in clinical screening. TPR was chosen as the measurement in this study for several reasons: (1) it enables early detection of diseases by effectively identifying individuals with the condition or disease; (2) it helps reduce false negatives, which can lead to missed diagnoses and delayed treatment; and (3) it optimizes resource allocation by efficiently allocating healthcare resources.

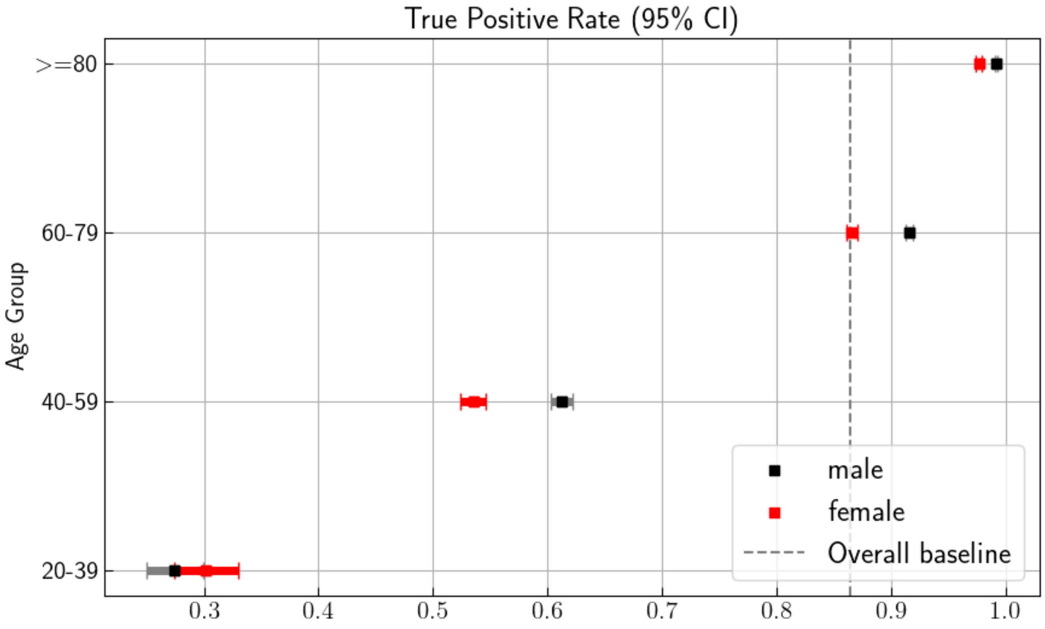


Figure 5.7. True Positive Rate (TPR) predictive value with 95% confidence intervals of the demographic factors specific results across patients' subgroups.

The XGB model's predictive performance, measured by TPR with 95% confidence intervals, was assessed within stratified demographic subgroups and subgroups based on the top-ranked disease bag features. The results are presented in Figure 5.7 and Figure 5.8, respectively. In Figure 5.7, it is evident that the model's predictive power increases as the age group progresses from 20-39 to >=80, with the 60-79 and >=80 age groups performing above average (indicated by the dashed line), while the 20-39 to 40-59 age groups perform worse than average. Additionally, the model exhibits better performance in the male subgroup compared to the female subgroup. When considering the XGB performance among disease bag subgroups, we observe that the model demonstrates superior TPR performance in subgroups where positive cases were identified for each specific disease bag, except for obesity. Significant heterogeneity exists among all subgroups, with lower TPR observed among patients without the majority of the top-ranked disease bags and among younger female age groups. Detailed information, including the number of patients labeled as true positive and false positive cases by the XGB model, is provided in Table 5.5.

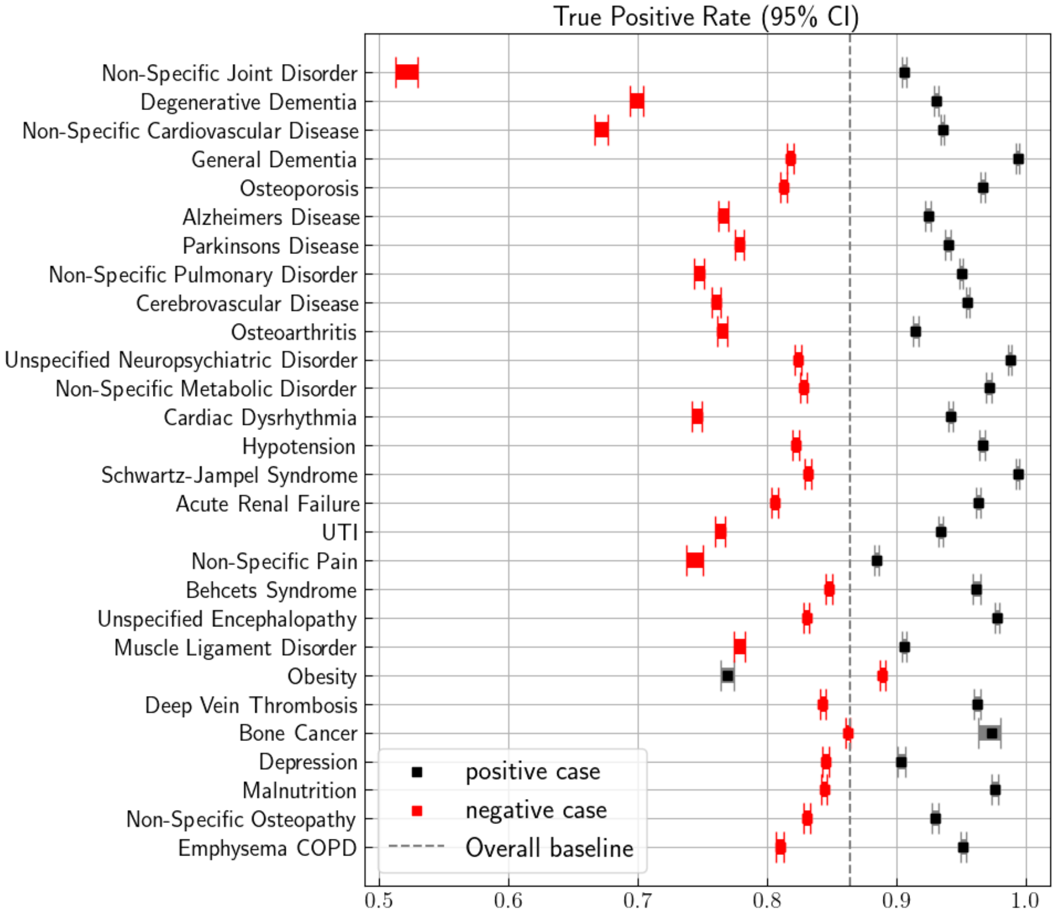


Figure 5.7. True Positive Rate (TPR) predictive value with 95% confidence intervals of the top 30 ranked XGB relative important disease bag specific results across patients' subgroups.

Table 5.5. XGB Model True Positive Rate (TPR) Predictive value with 95% Confidence Intervals with Detailed True Positive and False Positive Cases for Demographic and Top 30 Relative Important Features

Subgroups	Case	True Positive	False Positive	True Positive Rate (95% CI)
Sex	Male	66,757	741,841	88.8 (88.6-89.1)
	Female	37,287	562,193	82.2 (81.9-82.6)
Age	>=65	2,906	48,999	37.4 (36.4-38.5)
	<65	101,138	1,255,035	89.7 (89.5-89.9)
Non-Specific Joint Disorder	Yes	97,178	1,137,166	90.6 (90.4-90.7)
	No	6,866	166,868	52.1 (51.3-53.0)
Degenerative Dementia	Yes	79,810	882,731	93.0 (92.8-93.2)
	No	24,234	421,303	69.9 (69.4-70.4)
Non-Specific Cardiovascular Disease	Yes	82,070	912,013	93.5 (93.4-93.7)
	No	21,974	392,021	67.1 (66.6-67.7)
General Dementia	Yes	31,352	202,312	99.3 (99.2-99.4)
	No	72,692	1,101,722	81.8 (81.5-82.0)
Osteoporosis	Yes	38,698	324,206	96.6 (96.5-96.8)
	No	65,346	979,828	81.2 (81.0-81.5)
Alzheimers Disease	Yes	68,765	762,223	92.4 (92.2-92.6)
	No	35,279	541,811	76.6 (76.2-77.0)
Parkinsons Disease	Yes	59,860	644,292	93.9 (93.8-94.1)
	No	44,184	659,742	77.8 (77.5-78.2)
Non-Specific Pulmonary Disorder	Yes	65,729	702,547	95.0 (94.8-95.1)
	No	38,315	601,487	74.7 (74.4-75.1)
Cerebrovascular Disease	Yes	61,052	629,684	95.5 (95.3-95.6)
	No	42,992	674,350	76.0 (75.7-76.4)
Osteoarthritis	Yes	72,583	845,377	91.5 (91.3-91.6)
	No	31,461	458,657	76.5 (76.1-76.9)
Overall	N/A	104,044	1,304,034	86.4 (86.2-86.5)

5.4 Discussion

In patients presenting with diabetes, we developed and validated machine learning tools with broad prior electronic disease history to inform the probability of hip fracture development in the future. This is the first study that utilize the advantages of enormous electronic health records from MarketScan datasets which contains more than 200 million records across the US, which is so far the largest training and validating dataset that explore the cohort association between broad prior diseases and fracture outcome. Furthermore, this is the first study to demonstrate the potential of using prior disease history as significant factors to forecast future hip fracture outcome with the best ROC-AUC score achieved for eligible patients diagnosed with diabetes.

Our thorough studies on broad diseases association with fracture outcome via case-control method has proved consistent findings by comparing with their relative importance weight contributing to develop machine learning tool on future fracture detection. The critical predictors involved in the model development are diagnosis of cognitive diseases and physical dysfunction, majority of those diseases also have odds ratio greater than 5 which indicates eligible patients who have observed with cognitive disease and physical dysfunction have over 5 times risk to develop hip fracture outcome in the near future in comparison to others without them. Our researches on exploring confounding association between prior disease and fracture have directly benefited from data mining the rich electronic disease history records from MarketScan. Moreover, with access to patients' demographic factors and broad prior disease, we are capable to identify the significant heterogeneity in all subgroups categorized based on their age group, sex or occurrence of specific disease. Our studies emphasis model's heterogeneity predictive performance measured by true-positive-rate across these subgroups, which is the first study that incorporate such broad disease predictors and it’s critical to improve hip fracture screening process for primary care physicians to be alerted on patients that model has best true-positive-rate predictive performance.

In conclusion, we have constructed and validated a powerful fracture detection tool for eligible patients with diabetes to develop fracture in near future, by using the electronic health records from MarketScan across the US. This model achieved superior fracture detection performance measured by ROC-AUC score especially for patients conditioned with cognitive disease and physical dysfunction, we hope can be adopted in the practice for primary care physician to screen eligible patients on fracture detection and improve early treatment, with benefit for both patients and health care providers.

CHAPTER 6 CONCLUSION

We have covered multiple topics on continuous temporal signals and electronic health records for broad health state forecasting, given dataset modalities and desired task purpose we have utilized varies statistical analysis methods and machine learning tools accordingly. Studies of cohort association uncover the confounding relationships between diseases that patients' prior disease history lead to potential causality to the production of another severe disease in the future. Understanding cohort association between diseases has, therefore, paramount importance for preventive medical care. It's undoubtedly that development of statistical analysis and machine learning technologies have improve researchers' capability to utilize varies health data modalities information on a broader range of disease forecasting. Due to the availability of modern computational tools and extensive observational data collected from more complicated and precise body signal collect equipment, we are able to combine the electronic health data with temporal continuous signals to predict elderly cognitive decline. Moreover, the classical statistical fields benefited from the "Big Data Era" and improvement computation power, osteoporosis is a complex disease that previous osteoporosis risk calculation tools are limited with race/ethnic diversity and constraint access to electronic health record. Our studies explored the model improvement on fracture detection and risk calculation that could help target those in need of osteoporosis screening or treatment.

We started our analysis on cohort association study in MarketScan dataset in Chapter 2 that investigated the spatial and temporal distribution of patients diagnosed with acute pancreatitis (AP) and their subsequent development of varies diabetes mellitus (DM). The spatial analysis uncovered that a high concentration of AP and DM in Chicagoland underserved communities, particularly in South and West side of Chicago area predominately occupied by African American.

In the temporal analysis, the national wide electronic health record was used to confirm that ~20% of patients diagnosed with AP would develop DM in the next 5 years. This finding was further validated with measurement of multiple association signals such as odds ratio and relative risk using case-control method, which shows that patients with AP were significantly more likely to develop DM. To exclude other potential confounding factors such as gender and age, logistic regression model was applied to confirm the significant impact of AP on DM development independent of other factors.

In Chapter 3, we explored the accelerometry temporal signal effects on the cognitive decline prediction with electronic health records, we expanded our model features and studied the harmonic features contribution. We built an extreme-gradient-boosting (XGB) models that achieve state-of-art ROC-AUC score on patients' cognitive decline in the near future. This is the first study demonstrates that the overall collection of harmonic and statistical features derived from accelerometry temporal signals collected from hip and wrist sensors that delivered superior predictive power on elderly's cognitive deterioration status. In addition, our study shows that combination of electronic health record and accelerometry features would allow automated cognitive decline forecast that provides good discrimination among elderly population.

Chapter 4 and Chapter 5 focused on critical osteoporosis tasks, we utilized state-of-art machine learning techniques and comprehensive health data to investigate novel application in osteoporosis risk calculations and fracture detection. In Chapter 4, we built the gradient-boosting survival model from electronic health records collected from race/ethnicity diversified US based population, which was superior in comparison to traditional osteoporosis risk calculations tools such as FRAX and QFracture that lacks consideration to race/ethnicity contributions. We have thoroughly investigated the heterogeneity on varies race subgroups, and our study demonstrates that the inclusion of race improved fracture prediction. In Chapter 5, we explored the broad range of prior disease from electronic health records to hip fracture outcome on eligible patients who diagnosed with diabetes. We have investigated the top associated diseases through case-control study and relative important features in gradient-boosting model which shows the significant contribution of both cognitive disease and physical dysfunction to fracture outcome. This study uncovered for the first time that use thorough prior diseases history to future fracture detection and achieved superior ROC-AUC score.

As the central topic of this dissertation, we have explored the power of advanced statistical tools on varies health data modalities to solve different tasks. The projects investigated from this dissertation were proposed by domain expertise from their empirical experiences, then we have cooperated closely to utilize available datasets to validate their hypothesis and construct predictive tools on diseases screening. As impactful factors, electronic health records are essential tools to not only discover confounding association between diseases, but also significant contribution to develop disease forecasting AI/machine learning models in the fields of disease detection, risk calculation and diseases association study. Besides, we have also innovated methodology on extracting harmonic features from temporal signals that empower the tabular electronic health records on building state-of-art models. In the meantime, there are still lots of challenges and opportunities to push the limit of AI/machine learning tools and data mining techniques to varies health problem as the exponentially growing volume and novel modalities of stored health datasets worldwide.

REFERENCES

1. Islam, M. Ataharul, and Abdullah Al-Shiha. *Foundations of biostatistics*. Singapore: Springer, 2018.
2. Z. Obermeyer, \Predicting the Future | Big Data, Machine Learning, and Clinical Medicine Ziad," Physiology & behavior, vol. 176, no. 1, pp. 139{148, 2017.
3. 3. Kononenko, Igor. "Machine learning for medical diagnosis: history, state of the art and perspective." *Artificial Intelligence in medicine* 23, no. 1 (2001): 89-109.
4. Hunt, Earl B., Janet Marin, and Philip J. Stone. "Experiments in induction." (1966).
5. Nilsson, N.J., 1965. Learning machines McGraw-Hill. *New York, 19652*.
6. ëRosenblatt62ë Rosenblatt, F. "í Principles of Neurodynamics. í Washington DC." (1962).
7. Elomaa, T., and N. Holsti. "An experimental comparison of inducing decision trees and decision lists in noisy domains." In *Proceedings of the Fourth European Working Session on Learning, Montpelier, France, 4-6 December 1989*, pp. 59-69. 1989.
8. Bratko, Ivan, Igor Mozetič, and Nada Lavrač. *KARDIO: a study in deep and qualitative knowledge for expert systems*. Mit Press, 1990.
9. Wielinga, Bob, John Boose, Brian Gaines, and Maarten Van Someren, eds. *Current trends in knowledge acquisition*. Vol. 8. IOS Press, 1990
10. Muggleton, Stephen H. "Inductive acquisition of expert knowledge." (1989): 0468-0468.
11. Xiao, Cao, Edward Choi, and Jimeng Sun. "Opportunities and challenges in developing deep learning models using electronic health records data: a systematic review." *Journal of the American Medical Informatics Association* 25, no. 10 (2018): 1419-1428.

12. Zheng T, Xie W, Xu L, He X, Zhang Y, You M, et al. A machine learning-based framework to identify type 2 diabetes through electronic health records. Int J Med Inform. 2017;97:120-127.
13. Panahiazar M, Taslimitehrani V, Pereira N, Pathak J. Using EHRs and machine learning for heart failure survival analysis. Stud Health Technol Inform. 2015; 216:240.
14. Wong A, Young AT, Liang AS, Gonzales R, Douglas VC, Hadley D. Development and validation of an electronic health record–based machine learning model to estimate delirium risk in newly hospitalized patients without known cognitive impairment. JAMA Netw Open. 2018;1(4):e181018.
15. Suzuki, Kenji. "Overview of deep learning in medical imaging." *Radiological physics and technology* 10, no. 3 (2017): 257-273
16. Lawrence, Steve, C. Lee Giles, Ah Chung Tsoi, and Andrew D. Back. "Face recognition: A convolutional neural-network approach." *IEEE transactions on neural networks* 8, no. 1 (1997): 98-113.
17. Dhungel, Neeraj, Gustavo Carneiro, and Andrew P. Bradley. "Deep learning and structured prediction for the segmentation of mass in mammograms." In *International Conference on Medical image computing and computer-assisted intervention*, pp. 605-612. Springer, Cham, 2015.
18. Zhen X, Wang Z, Islam A, Bhaduri M, Chan I, Li S (2016) Multi-scale deep networks and regression forests for direct bi-ventricular volume estimation. Med Image Anal 30:120–129.

19. Wang S, Yin Y, Cao G, Wei B, Zheng Y, Yang G (2015) Hierarchical retinal blood vessel segmentation based on feature and ensemble learning. Neruocomputing 149:708–717
20. Gulshan, Varun, Lily Peng, Marc Coram, Martin C. Stumpe, Derek Wu, Arunachalam Narayanaswamy, Subhashini Venugopalan et al. "Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs." *Jama* 316, no. 22 (2016): 2402-2410.
21. Bock, Christian, Michael Moor, Catherine R. Jutzeler, and Karsten Borgwardt. "Machine learning for biomedical time series classification: from shapelets to deep learning." *Artificial Neural Networks* (2021): 33-71.
22. Kaushik, Shruti, Abhinav Choudhury, Pankaj Kumar Sheron, Nataraj Dasgupta, Sayee Natarajan, Larry A. Pickett, and Varun Dutt. "AI in healthcare: time-series forecasting using statistical, neural, and ensemble architectures." *Frontiers in big data* 3 (2020): 4.
23. Lonini L, Dai A, Shawen N, Simuni T, Poon C, Shimanovich L, et al. Wearable sensors for Parkinson's disease: which data are worth collecting for training symptom detection models. Npj Digit Med. 2018;1:64.
24. Yildirim, Özal. "A novel wavelet sequence based on deep bidirectional LSTM network model for ECG signal classification." *Computers in biology and medicine* 96 (2018): 189-202.
25. Kaur P, Malhotra S. Improved SLReduct Framework for Stress Detection Using Mobile Phone-Sensing Mechanism in Wireless Sensor Network. Advanced Computing and Intelligent Engineering. 2019;499-507.
26. Luhn, Hans Peter. "A business intelligence system." *IBM Journal of research and development* 2, no. 4 (1958): 314-319.
27. Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. "Attention is all you need." *Advances in neural information processing systems* 30 (2017).
28. Choi, Edward, Andy Schuetz, Walter F. Stewart, and Jimeng Sun. "Medical concept representation learning from electronic health records and its application on heart failure prediction." *arXiv preprint arXiv:1602.03686* (2016).
29. Church, Kenneth Ward. "Word2Vec." *Natural Language Engineering* 23, no. 1 (2017): 155-162.
30. Miotto, Riccardo, Li Li, Brian A. Kidd, and Joel T. Dudley. "Deep patient: an unsupervised representation to predict the future of patients from the electronic health records." *Scientific reports* 6, no. 1 (2016): 1-10.
31. Appelros, S., S. Lindgren, and A. Borgstrom, *Short and long term outcome of severe acute pancreatitis.* Eur J Surg, 2001. **167**(4): p. 281-6.
32. Boreham, B. and B.J. Ammori, *A prospective evaluation of pancreatic exocrine function in patients with acute pancreatitis: correlation with extent of necrosis and pancreatic endocrine insufficiency.* Pancreatology, 2003. **3**(4): p. 303-8.
33. Das, S.L., et al., *Newly diagnosed diabetes mellitus after acute pancreatitis: a systematic review and meta-analysis.* Gut, 2014. **63**(5): p. 818-31.

34. Huang, D.B. and P. Raskin, *Diabetic hypertriglyceridemia-induced acute pancreatitis masquerading as biliary pancreatitis.* Journal of Diabetes and its Complications, 2002. **16**(2): p. 180-182.

35. Martínez, J., et al., *Obesity is a definitive risk factor of severity and mortality in acute pancreatitis: An updated meta-analysis.* Pancreatology, 2006. **6**(3): p. 206-209.

36. Chen, X. and S. Devaraj, *Gut Microbiome in Obesity, Metabolic Syndrome, and Diabetes.* Curr Diab Rep, 2018. **18**(12): p. 129.

37. Vangipurapu, J., et al., *Microbiota-Related Metabolites and the Risk of Type 2 Diabetes.* Diabetes Care, 2020.

38. Buscaglia, J.M., et al., *Disparities in demographics among patients with pancreatitis-related mortality.* Jop, 2009. **10**(2): p. 174-80.

39. Chiang, A.L., P.A. Banks, and J. McNabb-Baltar, *Racial and Insurance Status Disparities in ERCP Utilization in Acute Pancreatitis: 83.* American Journal of Gastroenterology, 2015. **110**: p. S35.

40. Tan, X., et al., *Sociodemographic disparities in the management of type 2 diabetes in the United States.* Curr Med Res Opin, 2020: p. 1.

41. Prince, M. in *An Analysis of Prevalence, Incidence, Cost, and Trends* (ed. Wimo, A) (Alzheimer's Disease International, 2015).
42. Institute of Medicine of the National Academies (n.d.) Institute of Medicine of the National Academies. Accessed November 13, 2008. http://www.iom.edu/ Millenson, M. L. Evidence of a need for change. *Miller-McCune*, **1**, 34–39 (2008).
43. Federal Interagency Forum on Aging-related Statistics. (U.S. Government Printing Office, 2016).
44. Babulal, G. M. et al. Perspectives on ethnic and racial disparities in Alzheimer's disease and related dementias: Update and areas of immediate need. *Alzheimers Dement.* **15**, 292–312 (2019).
45. Hoe, J. et al. Changes in the quality of life of people with dementia living in care homes. *Alzheimer Dis. Assoc. Disord.* **23**, 285–290 (2009).
46. U.S. Preventive Services Task Force. Cognitive impairment in older adults screening https://www.uspreventiveservicestaskforce.org/uspstf/recommendation/cognitive-impairment-in-older-adults-screening (2020).
47. Amariglio, R. E. et al. Examining cognitive decline across Black and White participants in the Harvard aging brain study. *J. Alzheimers Dis.* **75**, 1437–1446 (2020).

48. Lee, H. B. et al. Race and cognitive decline among community-dwelling elders with mild cognitive impairment: findings from the Memory and Medical Care Study. *Aging Ment. Health* **16**, 372–377 (2012).

49. National Institute of Aging. Strategic directions for research, 2020–2025. https://www.nia.nih.gov (2020).

50. Castanho, T. C. et al. Telephone-based screening tools for mild cognitive impairment and dementia in aging studies: a review of validated instruments. *Front. Aging Neurosci.* **6**, 16 (2014).

51. Kuiper, J. S. et al. Social relationships and cognitive decline: a systematic review and meta-analysis of longitudinal cohort studies. *Int. J. Epidemiol.* **45**, 1169–1206 (2016).

52. Casanova, R. et al. Investigating predictors of cognitive decline using machine learning. *J. Gerontol. B Psychol. Sci. Soc. Sci.* **75**, 733–742 (2020).

53. Huisingh-Scheetz, M. et al. Wrist accelerometry in the health, functional, and social assessment of older adults. *J. Am. Geriatr. Soc.* **64**, 889–891 (2016).

54. Huisingh-Scheetz, M. et al. The relationship between physical activity and frailty among U.S. older adults based on hourly accelerometry data. *J. Gerontol. A. Biol. Sci. Med. Sci.* **73**, 622–629 (2018).

55. Huisingh-Scheetz, M. et al. *International Conference on Frailty and Sarcopenia Research* (2020).

56. Huisingh-Scheetz, M. J. et al. Relating wrist accelerometry measures to disability in older adults. *Arch. Gerontol. Geriatr.* **62**, 68–74 (2016).

57. Ho, E. C., Hawkley, L., Dale, W., Waite, L. & Huisingh-Scheetz, M. Social capital predicts accelerometry-measured physical activity among older adults in the U.S.: a cross-sectional study in the National Social Life, Health, and Aging Project. *BMC Public Health* **18**, 804 (2018).

58. Saint-Maurice, P. F. et al. Association of daily step count and step intensity with mortality among US adults. *JAMA* **323**, 1151–1160 (2020).

59. Wanigatunga, A. A. et al. Community-based activity and sedentary patterns are associated with cognitive performance in mobility-limited older adults. *Front. Aging Neurosci.* **10**, 341 (2018).

60. Spartano, N. L. et al. Accelerometer-determined physical activity and cognitive function in middle-aged and older adults from two generations of the Framingham Heart Study. *Alzheimers Dement.* **5**, 618–626 (2019).

61. Zhu, W. et al. Objectively measured physical activity and cognitive function in older adults. *Med. Sci. Sports Exerc.* **49**, 47–53 (2017).

62. Committee on Technology of the National Science and Technology Council. Executive office of the President of the United States (2019).

63. Gjoreski, H., Rashkovska, A., Kozina, S., Lustrek, M. & Gams. M. "Telehealth using ECG sensor and accelerometer," 2014 37th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), pp. 270–274, https://doi.org/10.1109/MIPRO.2014.6859575 (2014).

64. Sudlow, C. et al. UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med.* **12**, e1001779 (2015).

65. Ferrucci, L. The Baltimore Longitudinal Study of Aging (BLSA): a 50-year-long journey and plans for the future. *J. Gerontol. A. Biol. Sci. Med. Sci.* **63**, 1416–1419 (2008).

66. Michos, E. D. et al. Rationale and design of the Study To Understand Fall Reduction and Vitamin D in You (STURDY): A randomized clinical trial of Vitamin D supplement

doses for the prevention of falls in older adults. *Contemp. Clin. Trials* **73**, 111–122 (2018).

67. Waite, L. et al. Inter-university consortium for political and social research [distributor]. (2019).
68. Huisingh-Scheetz, M. et al. Geriatric syndromes and functional status in NSHAP: rationale, measurement, and preliminary findings. *J. Gerontol. B Psychol. Sci. Soc. Sci.* **69**, S177–S190 (2014).
69. Smirnova, E. et al. The predictive performance of objective measures of physical activity derived from accelerometry data for 5-year all-cause mortality in older adults: National Health and Nutritional Examination Survey 2003–2006. *J. Gerontol. Series A* **75**, 1779–1785 (2020).
70. Shiroma, E. J., Schrack, J. A. & Harris, T. B. Accelerating accelerometer research in aging. *J. Gerontol. Series A* **73**, 619–621 (2018).
71. O'Muircheartaigh, C., English, N., Pedlow, S. & Kwok, P. K. Sample design, sample augmentation, and estimation for wave 2 of the NSHAP. *J. Gerontol. B Psychol. Sci. Soc. Sci.* **69**, S15–S26 (2014).
72. PhilipsRespironics.Actiwatch. http://www.healthcare.philips.com/main/homehealth/sleep/actiwatch/default.wpd#&&/wEXAQUOY3VycmVudFRhYlBhdGgFCUVkdWNhdGlvbrs7D4d8dwFrxbRmM0TsUP60b3xr (2013).
73. Nasreddine, Z. S. et al. The Montreal cognitive assessment, MoCA: a brief screening tool for mild cognitive impairment. *J. Am. Geriatr. Soc.* **53**, 695–699 (2005).
74. Kotwal, A. A. et al. Evaluation of a brief survey instrument for assessing subtle differences in cognitive function among older adults. *Alzheimer. Dis. Assoc. Disord.* **29**, 317–324 (2015).
75. Shega, J. W. et al. Measuring cognition: the chicago cognitive function measure in the national social life, health and aging project, wave 2. *J. Gerontol. B Psychol. Sci. Soc. Sci.* **69**, S166–S176 (2014).
76. Charlson, M. E., Pompei, P., Ales, K. L. & MacKenzie, C. R. A new method of classifying prognostic comorbidity in longitudinal studies: development and validation. *J. Chronic Dis.* **40**, 373–383 (1987).
77. World Health Organization. WHO: global database on body mass index. http://apps.who.int/bmi/index.jsp?introPage=intro_3.html (2022).
78. Vasilopoulos, T. et al. Comorbidity and chronic conditions in the National Social Life, Health and Aging Project (NSHAP), wave 2. *J. Gerontol. B Psychol. Sci. Soc. Sci.* **69**, S154–S165 (2014).
79. Karas, M. et al. Estimation of free-living walking cadence from wrist-worn sensor accelerometry data and its association with SF-36 quality of life scores. *Physiol. Meas.* https://doi.org/10.1088/1361-6579/ac067b (2021).
80. Siu, Albert, Heather Allore, Darryl Brown, Susan T. Charles, and Matthew Lohman. "National Institutes of Health Pathways to Prevention Workshop: research gaps for long-

term drug therapies for osteoporotic fracture prevention." *Annals of internal medicine* 171, no. 1 (2019): 51-57.

81. DeSalvo, Demi. "Educational intervention to improve orthopedic splinting techniques in the emergency department setting." PhD diss., Boston University, 2022.
82. Desai, Rishi J., Mufaddal Mahesri, Younathan Abdia, Julie Barberio, Angela Tong, Dongmu Zhang, Panagiotis Mavros, Seoyoung C. Kim, and Jessica M. Franklin. "Association of osteoporosis medication use after hip fracture with prevention of subsequent nonvertebral fractures: an instrumental variable analysis." *JAMA network open* 1, no. 3 (2018): e180826-e180826.
83. Chandran, Manju. "Fracture risk assessment in clinical practice: why do it?" Journal of Clinical Densitometry 20. No.3 (2017): 274-279
84. Beaudoin C, Moore L, Gagné M, Bessette L, Ste-Marie LG, Brown JP, et al. Performance of predictive tools to identify individuals at risk of non-traumatic fracture: a systematic review, metaanalysis, and meta-regression. Osteoporos Int. 2019;1–20.
85. Kanis JA, Johnell O, Odén A, Johansson H, McCloskey E. FRAXTM and the assessment of fracture probability in men and women from the UK. Osteoporos Int. 2008;19(4):385–97.
86. Hippisley-Cox J, Coupland C. Derivation and validation of updated QFracture algorithm to predict risk of osteoporotic fracture in primary care in the United Kingdom: prospective open cohort study. Bmj. 2012;344:e3427.
87. Dagan N, Cohen-Stavi C, Leventer-Roberts M, Balicer RD. External validation and comparison of three prediction tools for risk of osteoporotic fractures using data from population based electronic health records: retrospective cohort study. bmj. 2017;356:i6755.
88. Kanis JA, Odén A, McCloskey EV, Johansson H, Wahl DA, Cooper C, et al. A systematic review of hip fracture incidence and probability of fracture worldwide. Osteoporos Int. 2012;23(9):2239–56.
89. Cauley JA, Wu L, Wampler NS, Barnhart JM, Allison M, Chen Z, et al. Clinical Risk Factors for Fractures in Multi-Ethnic Women: The Women's Health Initiative. J Bone Miner Res. 2007;22(11):1816–26.
90. Schwartz AV, Sellmeyer DE, Ensrud KE, Cauley JA, Tabor HK, Schreiner PJ, et al. Older women with diabetes have an increased risk of fracture: a prospective study. J Clin Endocrinol Metab. 2001;86(1):32–8.
91. Vestergaard P, Rejnmark L, Mosekilde L. Relative fracture risk in patients with diabetes mellitus, and the impact of insulin and oral antidiabetic medication on relative fracture risk. Diabetologia. 2005;48(7):1292–9.
92. Giangregorio LM, Leslie WD, Lix LM, Johansson H, Oden A, McCloskey E, et al. FRAX underestimates fracture risk in patients with diabetes. J Bone Miner Res. 2012;27(2):301–8.

93. Sellmeyer DE, Civitelli R, Hofbauer LC, Khosla S, Lecka-Czernik B, Schwartz AV. Skeletal metabolism, fracture risk, and fracture outcomes in type 1 and type 2 diabetes. Diabetes. 2016;65(7):1757–66.

94. Janghorbani M, Feskanich D, Willett WC, Hu F. Prospective study of diabetes and risk of hip fracture the nurses' health study. Diabetes Care. 2006;29(7):1573–8.

95. Melton LJ, Leibson CL, Achenbach SJ, Therneau TM, Khosla S. Fracture risk in type 2 diabetes: update of a population-based study. J Bone Miner Res. 2008;23(8):1334–42.

96. Pijpers E, Ferreira I, de Jongh RT, Deeg DJ, Lips P, Stehouwer CD, et al. Older individuals with diabetes have an increased risk of recurrent falls: analysis of potential mediating factors: the Longitudinal Ageing Study Amsterdam. Age Ageing. 2012;41(3):358–65.

97. Hygum K, Starup-Linde J, Harsløf T, Vestergaard P, Langdahl BL. Mechanisms in endocrinology: diabetes mellitus, a state of low bone turnover–a systematic review and meta-analysis. Eur J Endocrinol. 2017;176(3):R137–57.

98. Furst JR, Bandeira LC, Fan WW, Agarwal S, Nishiyama KK, McMahon DJ, et al. Advanced glycation endproducts and bone material strength in type 2 diabetes. J Clin Endocrinol Metab. 2016;101(6):2502–10.

99. Johnell O, Kanis JA. An estimate of the worldwide prevalence and disability associated with osteoporotic fractures. Osteoporos Int. 2006;17(12):1726–33.

100. Johnell O, Kanis JA. An estimate of the worldwide prevalence and disability associated with osteoporotic fractures. Osteoporos Int. 2006;17(12):1726–33.

101. Burge R, Dawson-Hughes B, Solomon DH, Wong JB, King A, Tosteson A. Incidence and economic burden of osteoporosis-related fractures in the United States, 2005–2025. J Bone Miner Res. 2007;22(3):465–75.

102. LODER RT. The influence of diabetes mellitus on the healing of closed fractures. Clin Orthop Relat Res 1976-2007. 1988;232:210–6.

103. Folk JW, Starr AJ, Early JS. Early wound complications of operative treatment of calcaneus fractures: analysis of 190 fractures. J Orthop Trauma. 1999;13(5):369–72.

104. Retzepi M, Donos N. The effect of diabetes mellitus on osseous healing. Clin Oral Implants Res. 2010;21(7):673–81.

105. Norris R, Parker M. Diabetes mellitus and hip fracture: a study of 5966 cases. Injury. 2011;42(11):1313–6.

106. Churpek MM, Yuen TC, Winslow C, Meltzer DO, Kattan MW, Edelson DP. Multicenter comparison of machine learning methods and conventional regression for predicting clinical deterioration on the wards. Crit Care Med. 2016;44(2):368.

107. Kermany DS, Goldbaum M, Cai W, Valentim CC, Liang H, Baxter SL, et al. Identifying medical diagnoses and treatable diseases by image-based deep learning. Cell. 2018;172(5):1122–31.

108. Jain RK, Weiner MG, Polley E, Iwamaye A, Huang E, Vokes T. Adjustments for Black Race and Hispanic Ethnicity in FRAX and QFracture Do Not Correctly Predict Fracture Risk. Poster Presentation presented at: American Society of Bone and Mineral Research 2022; 2022 Sep 11; Austin, TX.

109. Cheng YJ, Kanaya AM, Araneta MRG, Saydah SH, Kahn HS, Gregg EW, et al. Prevalence of diabetes by race and ethnicity in the United States, 2011-2016. Jama. 2019;322(24):2389–98.

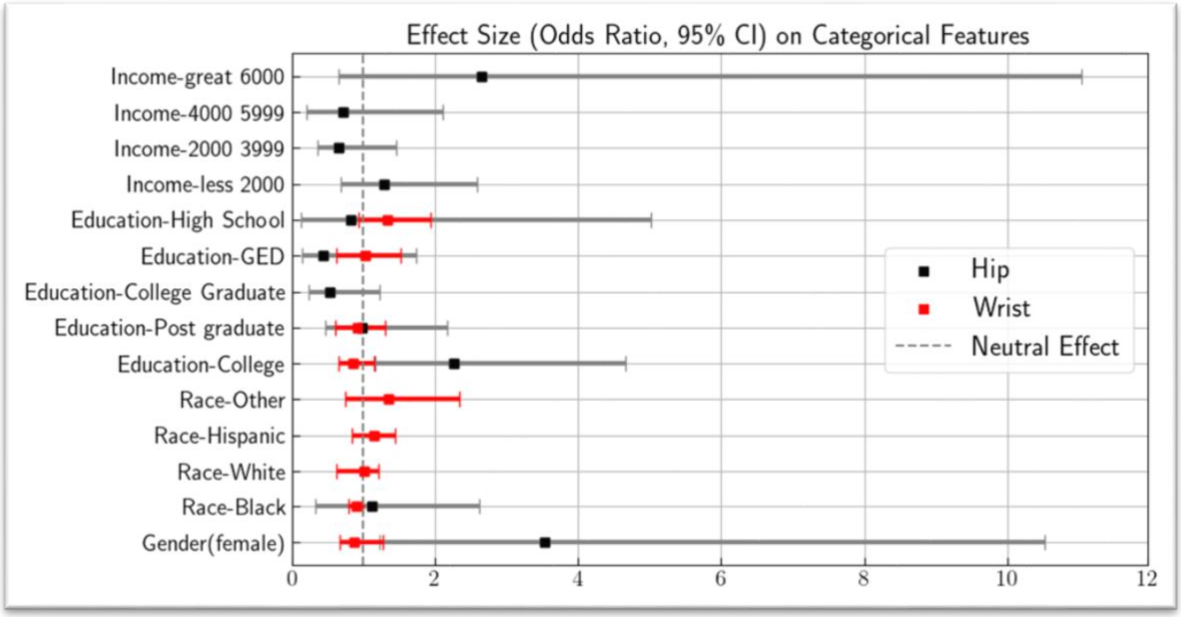
110. Ferrari SL, Abrahamsen B, Napoli N, Akesson K, Chandran M, Eastell R, et al. Diagnosis and management of bone fragility in diabetes: an emerging challenge. Osteoporos Int. 2018;29(12):2585–96.

111. McBean AM, Yu X. The underuse of screening services among elderly women with diabetes. Diabetes Care. 2007;30(6):1466–72.

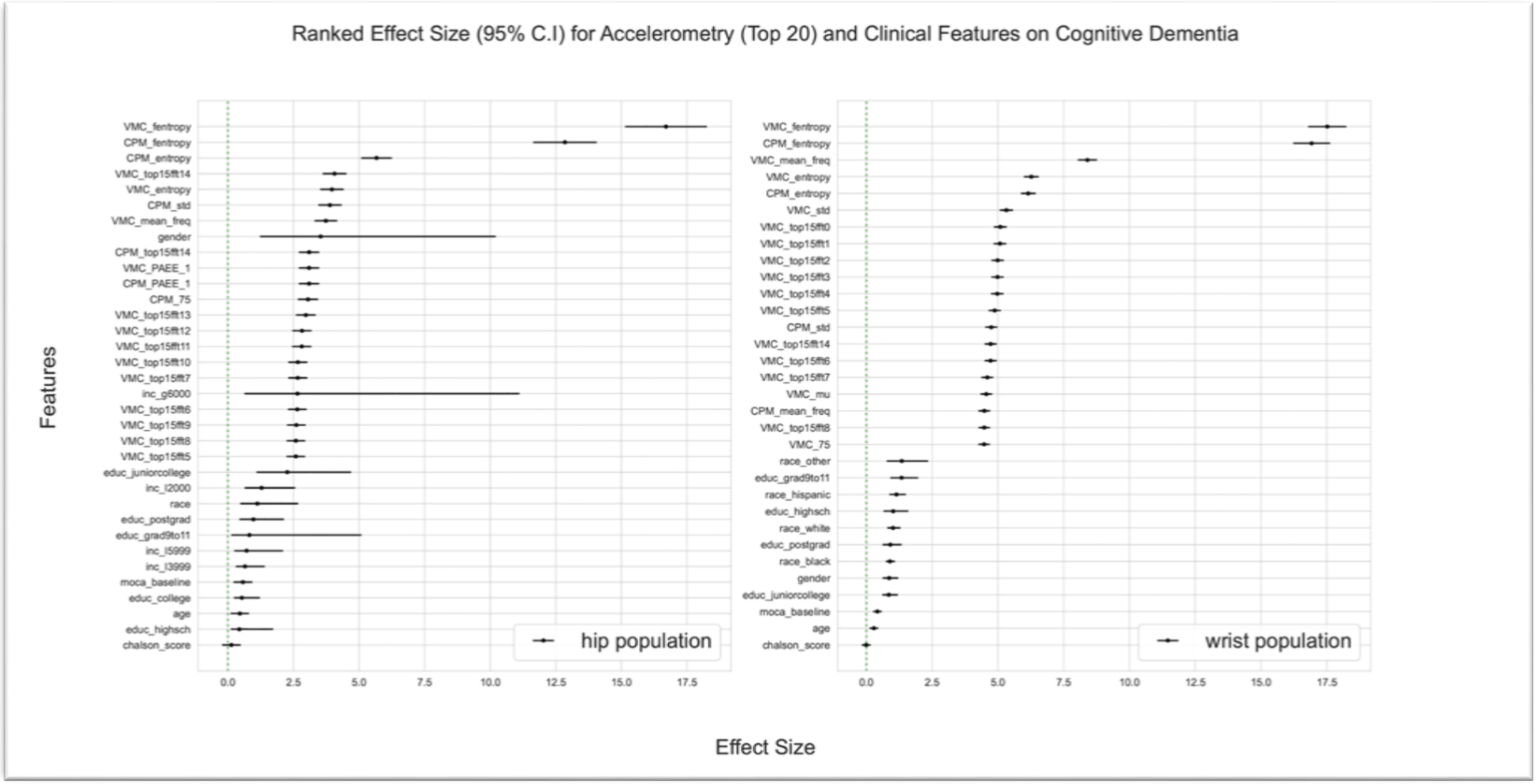
SUPPLEMENTAL MATERIALS

Supplemental Table 3.1. Full Features Acronym For Accelerometry Cohorts Prediction

gender age, ethgrp, baseline, charlson_score, CPM_mu, VMC_mu, CPM_std, VMC_std, CPM_0, VMC_0, CPM_25, VMC_25, CPM_50, VMC_50, CPM_75, VMC_75, CPM_range, VMC_range, CPM_skew, VMC_skew, CPM_kurt, VMC_kurt, CPM_beta(a), VMC_beta(a), CPM_beta(b), VMC_beta(b), CPM_entropy, VMC_entropy, CPM_PAEE, VMC_PAEE, CPM_top15fft0, VMC_top15fft0CPM_top15fft1, VMC_top15fft1, CPM_top15fft2, VMC_top15fft2, CPM_top15fft3, VMC_top15fft3, CPM_top15fft4, VMC_top15fft4, CPM_top15fft5, VMC_top15fft5, CPM_top15fft6, VMC_top15fft6, CPM_top15fft7, VMC_top15fft7, CPM_top15fft8, VMC_top15fft8, CPM_top15fft9, VMC_top15fft9, CPM_top15fft10, VMC_top15fft10, CPM_top15fft11, VMC_top15fft11, CPM_top15fft12, VMC_top15fft12, CPM_top15fft13, VMC_top15fft13, CPM_top15fft14, VMC_top15fft14, CPM_top15freq0, VMC_top15freq0, CPM_top15freq1, VMC_top15freq1, CPM_top15freq2, VMC_top15freq2CPM_top15freq3, VMC_top15freq3, CPM_top15freq4, VMC_top15freq4, CPM_top15freq5, VMC_top15freq5, CPM_top15freq6, VMC_top15freq6, CPM_top15freq7, VMC_top15freq7, CPM_top15freq8, VMC_top15freq8, CPM_top15freq9, VMC_top15freq9, CPM_top15freq10, VMC_top15freq10, CPM_top15freq11, VMC_top15freq11, CPM_top15freq12, VMC_top15freq12, CPM_top15freq13, VMC_top15freq13, CPM_top15freq14, VMC_top15freq14, CPM_fentropy, VMC_fentropy, CPM_psd_mu, VMC_psd_mu, CPM_psd_std, VMC_psd_std, CPM_rms_amplitude, VMC_rms_amplitude, CPM_mean_freq, VMC_mean_freqCPM_median_freq, VMC_median_freq, educ_hs/equiv, educ_voc, cert/some, college/assoc, educ_less_hs, educ_bachelors, or, more, income_<=$2000/month, income_$2000-3999/month, income_$4000-5999/month, income_>=$6000/month



Supplemental Figure 3.1. The effect size of demographic features on accelerometry cohorts with cognitive decline and without cognitive decline.



Supplemental Figure 3.2. The effect size of top 20 features on accelerometry cohorts with cognitive decline and without cognitive decline.