Using Machine Learning to Predict Super-Utilizers of Healthcare Services

by

Kevin P. Buchan Jr.

A Dissertation

Submitted to the University at Albany, State University of New York

In Partial Fulfillment of

the Requirements for the Degree of

Doctor of Philosophy

College of Emergency Preparedness, Homeland Security and Cybersecurity

Information Science Department

Spring 2021

ABSTRACT

In this dissertation, I aim to forecast high utilizers of emergency care and inpatient Medicare services (i.e., healthcare visits). Through a literature review, I demonstrate that accurate and reliable prediction of these future high utilizers will not only reduce healthcare costs but will also improve the overall quality of healthcare for patients. By identifying this population at risk before manifestation, I propose that there is still time to reverse undesirable healthcare trajectories (i.e., individuals whose clinical risk increases an excessive healthcare and treatment burden) through timely attention and proper care coordination. My dissertation culminates in the delivery of state-of-the-art predictive models that exploit well-researched clinical, behavioral, and social determinants associated with increased inpatient and emergency care utilization. I discuss my contributions to applied machine learning in healthcare herein, and further examine ethical concerns common to similar machine learning tasks. Finally, I conclude by reviewing how this research can be advanced through future work.

ACKNOWLEDGEMENTS

Allow me to start my acknowledgements where most conclude. If not for the love and support of my family, my dissertation work would not be possible. To my wife Cassie Buchan, who has helped me as a domain expert in her professional capacity as a physician assistant; to my mother Dr. Alida-Hayner Buchan, who has helped me as a domain expert in her professional capacity as a physician; to my father, Kevin Buchan Sr. who has been editing my work since my first book report; to my brothers, Thomas and Michael Buchan, who would drive the length of the country to bring me a pen: my actual primary aim is to make you proud.

To the brilliant Drs. Ozlem Uzuner and Feng Chen: thank you for taking a chance on me and teaching me everything that I know. To Dr. Luis-Luna Reyes, thank you for late-night meetings, well-timed jokes, and for steering the ship in always positive direction. To Dr. Archana Krishnan, thank you for challenging me to think outside my comfort zone and for your consistent thoughtful feedback. To William Kelly, I could fill a dissertation that explains the ways in which you have helped me develop as a researcher and data scientist, but I will follow your advice and try to be succinct: thank you.

To my great friends and colleagues, (soon to be Dr.) James McGaughan, Marcela Munoz, and Dr. Jason Piccone, thank you for stimulating conversation and unwavering friendship. To my fellow Uzunerds, Drs. Michele Filannino and Carson Tao, battling alongside you during our NLP challenges was some of the most fun I have ever had.

Finally, thank you to everyone at the State University of New York at Albany who has contributed to my research, including administrators and support staff. A special thanks to those at Nascate who always help me in countless ways, including Walt Mykins, Charles Lybrand, Sonny Ouyang, Colin Dirolf, Andrew Mondo, Brian Davey, and Juliet Zhao.

TABLE OF CONTENTS

CHAPTER 1.

INTRODUCTION

In this dissertation, I aim to forecast high utilizers of emergency care and inpatient Medicare services (i.e., healthcare visits). Through a literature review, I demonstrate that accurate and reliable prediction of these future high utilizers will not only reduce healthcare costs but will also improve the overall quality of healthcare for patients. By identifying this population at risk before manifestation, I propose that there is still time to reverse undesirable healthcare trajectories (i.e., individuals whose clinical risk increases an excessive healthcare and treatment burden) through timely attention and proper care coordination. . My dissertation culminates in the delivery of state-of-the-art predictive models that exploit well-researched clinical, behavioral, and social determinants associated with increased inpatient and emergency care utilization. I discuss my contributions to applied machine learning in healthcare herein, and further examine ethical concerns common to similar machine learning tasks. Finally, I conclude by reviewing how this research can be advanced through future work.

1.1 Background

1.1.1 The Pervasiveness of Analytics in Healthcare

In the last decade, the healthcare industry has seen an explosion in the amount of digital biological and medical data that is generated and stored by health plans, healthcare providers, and healthcare clearinghouses. To put this growth into perspective, the considerable quantity of healthcare, biomedical, and social research data collected by academic institutions, government agencies, insurance providers, and industry roughly doubles every year (Feinleib, 2014). In 2019, Americans used over 4.4 million gigabytes of internet data *every minute*, including an estimated 188 million emails, over 18 million text messages, and nearly 5 million Google searches; and in 2020, approximately 44 zettabytes

of digital data were produced (Albarracin, 2020). This immense volume of data—including the systems that capture, manage, and process vast datasets—is referred to as "big data" (Chen et al., 2014). The birth of big data has introduced numerous technical and ethical challenges, but it has also presented an abundance of scientific opportunity and technological advancement. Among the most important possibilities for advancement is discovery of an untapped wealth of knowledge through big data analytics i.e., the searching, mining, and analysis of big data (Kwon et al., 2014)).

Big data analytics will continue to permeate healthcare, as the domain is ripe with both data and potential for application development (Groves et al., 2016). Accordingly, the healthcare analytics market is expected to grow from $5.8 billion in 2015 to $84.2 billion in 2027 which translates to a *1,352% increase* in just 12 years. Still, most of the potential for value creation in healthcare analytics remains unrealized (Groves et al., 2016; Meticulous Market Research Pvt. Ltd, 2020). Applied big data analytics in healthcare shows promise in three analytic domains: (1) descriptive, (2) predictive, and (3) prescriptive analytics (Islam et al., 2018).

Descriptive analytics enables classification and categorization of both structured and unstructured data so that analysts can more easily consume extensive, multidimensional data sets (Raghupathi & Raghupathi, 2013). In healthcare, some widely used descriptive analytics technologies include:

- Health-related taxonomies (e.g., *the Centers for Medicare & Medicaid Services (CMS) Healthcare Provider Taxonomy Code Set* and the *National Center for Biotechnology Information (NCBI) taxonomy database*) (*Taxonomy - Centers for Medicare & Medicaid Services*, 2019), (Federhen, 2012)
- Thesauri (e.g., the *Unified Medical Language System (UMLS) Metathesaurus* and the *National Cancer Institute (NCI) Thesaurus*) (Tuttle et al., 1988), (Fragoso et al., 2004)

- Ontologies (e.g., the *Human Disease Ontology and* the *Gene Ontology Resource*) (Jupp et al., 2015), (Consortium & T. G. O. Consortium, 2001)
- Segmentation and clustering systems (e.g., the *Agency for Healthcare Research and Quality (AHRQ) Prevention Quality Indicators (PQIs)* and *AHRQ Clinical Classification Software*) (Indicators, 2001), (Elixhauser, 1996)
- Bio-surveillance software (e.g., the *Suite for Automated Global Electronic bioSurveillance (SAGES*)) (Lewis et al., 2011)
- Visualization tools (e.g., the *Commission on Social Determinants of Health visualization framework*) (Solar & Irwin, 2010)

The healthcare industry has invested heavily in predictive analytics, which is concerned with forecasting future events. This field has seen tremendous growth in the past decade, and this growth is likely to continue as big data in healthcare becomes increasingly available to the public in a de-identified form, and as the power to process is democratized by cloud platforms (e.g., Amazon Web Services, Google Cloud, Microsoft Azure, IBM Cloud, etc.). There are several products available for *risk scoring* of specific diseases (e.g., *3M Clinical Risk Groups* (Hughes et al., 2004), *Atherosclerotic Cardiovascular Disease (ASCVD) Risk Estimator Plus* (Rana et al., 2016), the *Mayo Clinic's Heart Disease Risk Calculator (Treeprasertsuk et al., 2012)*). Supplementing the development of clinical prediction systems is a recent influx of publications directed at clinical *prediction* tasks—including short-term hospital readmissions (Artetxe et al., 2018; Kansagara et al., 2011; Perera, 2018), mortality and frailty (Aoyama et al., 2018; Chang & Lin, 2015; Fahey et al., 2018; Kojima et al., 2018), and mental health deterioration (Alonso et al., 2018; Becker et al., 2018; Reece et al., 2017). There has been commercial and academic interest in tasks related to optimization of the healthcare system, such as the *forecasting* of appointment "no shows," (Goffman et al., 2017; Harvey et al., 2017; Topuz et al., 2018).

and (relevant to the focus of this dissertation) predicting healthcare utilization patterns (Sheets et al., 2017), (Yang et al., 2018).

Descriptive and prediction systems help to support health plan executives, practitioners in provider groups, and ultimately the beneficiaries who utilize healthcare services. However, decision makers are still faced with taking the best course of action considering categories and probabilities. Prescriptive analytics are meant to determine the optimal course of action when encountering certain choices. Development of precision medicine and new therapies through prescriptive analytics is an exciting frontier in healthcare. Several elements of clinical trial design—including the safety evaluation, dosage optimization, and adverse drug event detection—are supplemented by prescriptive analytics to advance precision medicine (Admes & Garets, 2014), (Islam et al., 2018). Prescriptive analytics are widely used by beneficiaries for several health-related tasks such as weight management optimization (Ross & Wing, 2016) and provider recommendation (Rogers et al., 2012).

In this dissertation, I employ descriptive analytics to inform a predictive analytic pipeline. I then discuss the role of prescriptive analytics to support the predicted population of interest: *patients who will significantly increase in inpatient and emergency care healthcare utilization*.

1.1.2 Research Goals

My primary aim is to develop a machine learning solution that will precisely identify future super-utilizers of emergency care (ER) and inpatient (IP) Medicare services. To support this aim, I refine the super-utilization definition through careful research and analysis, and through a comprehensive literature review, design member inclusion and exclusion criteria to focus on a particular subset of the super-utilizer population that is potentially amenable to intervention. Moreover, I examine the potential ethical implications of this machine learning solution by emphasizing interpretability and explainability of several models.

My secondary aim is to conduct a thorough investigation of different experimental designs to recommend the optimal setup for this task. This includes testing of different dataset architectures, class imbalance techniques, effects of hyperparameter tuning, and examination of diverse algorithms (*including to compare the respective drawbacks and benefits of a well-known, high-performing binary classification algorithm with a state-of-the-art deep learning algorithm*).

1.1.3 Significance

Despite the tremendous growth of analytics in healthcare, the still fledgling library of published research and publicly available tools focused on prediction of healthcare utilization is in its nascent form. Rapid progress in state-of-the-art machine learning creates a moving target for researchers and practitioners alike. Consider, for example, OpenAI's Generative Pre-Training (GPT) models.

When GPT was first released in 2018, its training set was a few thousand books (i.e., approximately 5GB of text). The pre-training step for GPT ("Improving Language Understanding with Unsupervised Learning," 2020) required one month of training on a single 8-GPU machine. GPT was tested on commonsense reasoning, reading comprehension, and sentiment analysis tasks. GPT-2 (Radford, Wu, Amodei, et al., 2019; Radford, Wu, Child, et al., 2019), released in 2019, was trained on 8 million web pages and contained 1.5 billion parameters, which is more than 10 times the amount of data and more than 10 times the number of parameters used during pre-training of GPT. GPT-2 has been evaluated on question answering, reading comprehension, interpretation, and description tasks. Released in 2020, GPT-3 (Radford, Wu, Child, et al., 2019) was trained on 175 billion parameters (i.e., three orders of magnitude larger than GPT-2). Even at the lowest 3-year reserved cloud pricing, GPT-3 (Li, 2020) is estimated to have cost $4.6M for a single training run, which would have taken approximately 355 GPU-years. Even without fine-tuning for a specific task, GPT-3 has demonstrated high utility in a wide and diverse array of applied natural language processing (NLP) tasks, including question

answering, reading comprehension, interpretation, description, machine translation, scripting poems and elementary mathematics. In just 3 years, GPT has evolved from a useful project with an ambitious vision to an AI behemoth that achieves state-of-the-art performance across an incredible suite of tasks.

In 2021, the Google Brain trained a model with 6 times the number of parameters as GTP-3; i.e., a model that scaled to a trillion parameters (Fedus et al., 2021). Thus, it is unsurprising that practitioners and researchers who apply machine learning in healthcare struggle to keep up; especially because as its function scales, as do concerns related to ethical principles in machine learning and AI.

Most of the research that applies machine learning to healthcare is predominantly directed at predicting clinical outcomes that cover mostly narrowly-defined populations, e.g., the prediction of specific conditions such as diabetes (Kandhasamy & Balamurali, 2015), hypertension (Nimmala et al., 2018), and coronary artery disease (CAD) (Buchan et al., 2017). Furthermore, unlike many clinical prediction tasks (including mortality and short-term readmission prediction) for which researchers have exploited large datasets (Purushotham et al., 2017) e.g., the MIMIC III dataset (Johnson et al., 2016), most forecasting of healthcare utilization has been tested on relatively small datasets of under 20,000 patients (Harris et al., 2016; Rinehart et al., 2018; R. L. Robinson et al., 2016; Sheets et al., 2017; Turbow et al., 2018). State-of-the-art machine learning methods rely on large datasets to maximize learning (Miotto et al., 2018), but it is arduous to obtain high-volume datasets in healthcare.

Difficulty obtaining publicly available health data has been the principal barrier to research targeted at the prediction of healthcare utilization. However, the Centers for Medicare & Medicaid Services (CMS) has recently prioritized data availability and accessibility, which is an important step to expanding machine learning research in healthcare. In 2014, CMS formed the Office of Enterprise Data and Analytics (OEDA), which was created to "*guide decision-making and develop frameworks promoting appropriate external access to and use of data to drive higher quality, patient-centered care at a lower cost*" ("CMS Creates New Chief Data Officer Post | CMS," 2014). Furthermore, CMS has

published Medicare Advantage encounter data for the first time ("CMS Is Releasing Medicare Advantage Encounter Data for Researchers to Analyze," 2019), and has even hosted a challenge that aims to estimate hospital inpatient utilization ("About the Challenge | Agency for Healthcare Research & Quality," 2019). As part of the challenge, CMS released customized analytic files that contain hospital inpatient discharges for six years. The purpose of CMS's advocacy of data analytics was to better understand the issues facing healthcare to address the Institute for Healthcare Improvement (IHI) triple aim (Berwick et al., 2008): (1) improving the patient experience of care (including quality and satisfaction); (2) improving the health of populations; and (3) reducing the per capita cost of healthcare ("About the Challenge | Agency for Healthcare Research & Quality," 2019 ; "CMS Is Releasing Medicare Advantage Encounter Data for Researchers to Analyze," 2019).

Since the IHI outlined the triple aim (Berwick et al., 2008) to improve the United States healthcare system, researchers have been especially interested in understanding, identifying, and ultimately predicting high utilizers of healthcare. In November 2016, a statistical briefing published by Emily Mitchell at the AHRQ further fueled this interest among researchers when she reported that "*only five percent of the population accounted for over half of healthcare spending*" in the 2014 Medical Expenditure Panel Survey (MEPS) File (*STATISTICAL BRIEF #497: Concentration of Health Expenditures in the U.S. Civilian Noninstitutionalized Population, 2014*, n.d.). Thereafter, this statistic justified the binary "high-utilizer" definition used by many researchers. Accordingly, this discrete definition states binarizes all individuals in the top 5% of healthcare expenditures as "high utilizers," and those outside the top 5% of healthcare expenditures as "not high utilizers." For example, Sheets et al. (2017) define high utilizers as those in the top 5% of healthcare expenditures and justify their definition by stating such high utilizers are a critical cohort to identify because only 5% of patients incur 50% of overall healthcare expenses.

Beyond definition, the purpose of identifying patients at risk for transition into the high utilizer population is to perhaps reverse the undesirable trajectory before it manifests. This means classifying people *who are headed toward super-utilization* due to certain identifiable clinical, behavioral, and social characteristics, but who are not yet super-utilizers; in addition, classifying people *who are already super-utilizers* and will continue this behavior in the near future. This is a strong use case for a predictive algorithm, as a functional model promotes improved quality of healthcare while saving a health plan considerable financial resource. However, as predictions are intended to be followed with interventions, practitioners demand sufficient model performance (*i.e., an acceptable measure of precision and recall*).

Care coordination improvement has been a recent focus by healthcare organizations, especially in Medicare. Medicare offers several programs that are designed to better coordinate care, including Accountable Care Organizations (ACOs), the Comprehensive Primary Care initiative, and the Oncology Care Model ("Coordinating Your Care | Medicare," 2019). Care coordination models have demonstrated the capacity to reduce healthcare visits and eliminate redundant tests while improving outcomes and lowering spending (Berkowitz et al., 2018), which aligns perfectly with the IHI's triple aim initiative. Such care coordination efforts are excellent candidates for supplementation by data analytics because they rely on finding the right patients who are most in need of care coordination. Identifying patients who are likely to become high utilizers in the future is a perfect example of people who would benefit from care coordination, as it allows time for the proper organization of activities between parties responsible for the patient's care.

The purpose of my research is to classify high utilizers of emergency care (ER) and inpatient (IP) Medicare services, because identifying members who are likely to overuse ER and IP services in the future allows time for healthcare professionals to better coordinate their care. Referring future super-utilizers to care coordination experts supports the IHI triple aim of improved care, better outcomes, and

reduced costs (Berwick et al., 2008). Specifically, patients who can be reliably identified before super-utilization manifests will benefit from improved ambulatory care to help them navigate the healthcare system with more direction, likely reducing IP (Harris et al., 2016) and ER utilization (Hasselman, 2013). My operational definition for "super-utilizer" is discussed further in "Section 2.1: "Super-Utilizer" Defined."

Finally, beneficiaries who super-utilize IP and ER services often suffer from complex physical, behavioral, and social needs. These needs, coupled with a lack of coordinated care, often materialize as avoidable utilization that is costly and suboptimal for a beneficiary's overall health (Hasselman, 2013). The predictive system I have developed identifies beneficiaries who frequently visit emergency departments and suffer inpatient admissions and readmissions. By predicting these beneficiaries in advance, it is possible to find a medical "home" that will better coordinate care and reduce ineffective utilization.

1.2 Data

I applied for and was granted access to a Medicare limited dataset (LDS) from the Center for Medicare and Medicaid Services (CMS), which contains deidentified beneficiary-level protected health information. Specifically, I received standard analytical files that contained claims for 5 of 6 claim types (*i.e., inpatient, outpatient, skilled nursing facility, home health agency, and hospice claims*). To evaluate the performance of models trained on longitudinal data, I requested 18 months of data, which is explained further in "Section 3.2: Technical Design of the Dataset." CMS offers LDS as quarterly analytical files in two quantities: (1) 5%, and (2) 100%. The 5% random sample is adequate to train and evaluate predictive models that will generalize to the national Medicare population (Mobley, 2013). The 5% random sample is sufficiently large (i.e., approximately 2.5 million beneficiaries) and removes

regional (*i.e.*, *geographic*) bias. The costs of these files are outlined in Table 1. Note that I did not request the Durable Medical Equipment File Quarterly, as it was not relevant to my research aims.

Table 1. 18 months of a 5% sample of Medicare data for 5 claim types

Name of file	Months of data	Data Duration	% Data	Cost data
Master Beneficiary Summary File Quarterly	18	20171Q – 20182Q	5%	$900
Carrier File Quarterly†	18	20171Q – 20182Q	5%	$6,450
Durable Medical Equipment File Quarterly†	18	20171Q – 20182Q	5%	$3,000
Home Health File Quarterly	18	20171Q – 20182Q	5%	$1,200
Hospice File Quarterly	18	20171Q – 20182Q	5%	$1,200
Inpatient File Quarterly	18	20171Q – 20182Q	5%	$1,500
Outpatient File Quarterly	18	20171Q – 20182Q	5%	$3,750
Skilled Nursing Facility File Quarterly	18	20171Q – 20182Q	5%	$1,200
			Total	$16,200

† Only 5% of this data is available

It is important to recognize that the data is missing two important elements: (1) complete expenditure information; and (2) medication information. Features engineered around cost and medication were consistently documented as important to identification and prediction of high utilizers in my review of the literature. The absence of this information is a limitation of my research.

Prior to my CMS data request, I received Institutional Review Board (IRB) approval for my dissertation research. The data was delivered by CMS on an encrypted physical drive 8 weeks upon receipt of request. To receive and process the data, I was required to complete a data use agreement (DUA) with CMS, which outlines the data integrity guidelines (See Supplementary Material Appendix A. Data Integrity Guidelines).

The limited data set (LDS) I procured from CMS contained Medicare entitlement and claims data in Standard Analytical Files (SAF) across a Master Beneficiary Summary File (MBSF) and six claim type sources: (1) carrier; (2) home health; (3) hospice; (4) inpatient; (5) outpatient; and (6) skilled nursing facility (SNF) data (See Supplementary Material Appendix B. Data Dictionaries to review each Data Dictionary per claim type in full).

The MBFS file describes beneficiary eligibility and demographic information, including (among other variables): (1) a de-identified unique beneficiary identifier that is used to track a beneficiary's claims across claim types; (2) sex; (3) race; (4) age; (5) county; (6) original/current reason for entitlement; (7) dual status; and (6) an HMO indicator. Table 2 shows the number of eligible members by months and year in the CMS dataset, and Table 3 shows the number of claims by year and month.

Table 2. Beneficiary Eligibility by Year and Month

Year	Month	Number of Members
2017	4	1,692,308
2017	5	1,694,001
2017	6	1,696,529
2017	7	1,706,392
2017	8	1,709,633
2017	9	1,712,673
2017	10	1,708,334
2017	11	1,709,970
2017	12	1,712,911
2018	1	1,682,100
2018	2	1,682,183
2018	3	1,683,682
2018	4	1,683,575
2018	5	1,684,495
2018	6	1,686,766
2018	7	1,692,277
2018	8	1,695,000

Table 3. Claims by Year and Month

Incurred Year	Incurred Month	Number of Claims
2017	4	4,351,765
2017	5	4,728,255
2017	6	4,541,137
2017	7	4,279,155
2017	8	4,751,814
2017	9	4,436,021
2017	10	5,020,524
2017	11	4,570,341
2017	12	4,179,168
2018	1	4,690,089
2018	2	4,218,033
2018	3	4,509,516
2018	4	4,568,470
2018	5	4,746,722
2018	6	4,395,753
2018	7	4,472,731
2018	8	4,746,741

2018	9	1,697,792	2018	9	4,335,763

Table 4 shows beneficiary eligibility totals from the Master Beneficiary Summary File, including the number of beneficiaries in the data who became eligible for Medicare by age, or through end-stage renal disease (ESRD) or disability. Table 4 also shows the number of beneficiaries who are dual eligible for Medicare and Medicaid in the data, and the counts of dual eligibility by type (*e.g., Qualified Medicare Beneficiary (QMB)-only, QMB and full Medicaid coverage, including prescription drugs, etc.*).

Table 4. Beneficiary Eligibility Statistics in the Master Beneficiary Summary File

Current Medicare Status	**Number of Members**
Aged without end-stage renal disease (ESRD)	1,691,555
Aged with ESRD†	13,184
Disabled without ESRD	335,707
Disabled with ESRD†	9,411
ESRD only†	6,137
None	512
Dual Indicator	**Number of Members**
Qualified Medicare Beneficiary (QMB)-only	74,145
QMB and full Medicaid coverage, including prescription drugs	250,949
Specified Low-Income Medicare Beneficiary (SLMB)-only	44,478
SLMB and full Medicaid coverage, including prescription drugs	20,722
Qualified Disabled Working Individual (QDWI)	1,298

Qualifying individuals (QI)	24,878
Other dual eligible (not QMB, SLMB, QWDI, or QI) with full Medicaid coverage, including Rx	130,234
Other dual eligible, but without Medicaid coverage	940
Unknown"	7,169
Non-Medicaid	1,684,148
Current Reason for Entitlement Annual	**Number of Members**
Old age and survivor's insurance (OASI)	1,702,752
Disability insurance benefits (DIB)	328,600
End-stage renal disease (ESRD)†	*6,064*
Both DIB and ESRD†	*1,556*
ESRD Indicator Annual	**Number of Members**
The beneficiary does not have ESRD	1,999,726

†Beneficiaries were excluded from the study in accordance with "Section 3.3.2: Exclusion Criteria."

1.3 Task Formulation

As described in "Section 1.1.2: Research Goals" (and reviewed further in "Section 2.1: 'Super-Utilizer' Defined"), the purpose of my research is to classify super-utilizers of emergency care (ER) and inpatient (IP) Medicare services. There are several important reasons that I operationalize the "super-utilizer" definition around measurements of ER and IP utilization. Principally, the data I obtained from CMS omits expenditure information, which means I could not reliably build a dependent variable using cost. Nevertheless, focusing on ER and IP utilization offers certain advantages that make it favorable for use in the dependent variable design.

When using cost, or even full utilization, to construct the dependent variable, risk adjustment is a complicated yet necessary process. Consider for example, a healthy patient for whom we see a first-ever maternity diagnosis in the data. The healthy patient, who saw a primary care physician (PCP) just once in the past 12 months, will now see an obstetrician 14 times over the next 9 months. Of course, this increase of 13 visits is substantial and yields an increase in healthcare expenditures. However, we do not think of this utilization as *avoidable*, nor should this patient fit a "super-utilizer" definition. Rather, this type of utilization is expected and perfectly appropriate given the patient's diagnosis. Another extreme example of this is a patient who is in cancer remission. Cancer treatment is costly, yet necessary and expected given the presence of a cancer diagnosis. So clearly it is important to ask, "*does this cost and/or utilization meet super-utilization criteria relative to the patient's diagnosis, and (perhaps) other non-clinical factors?*"

Risk adjustment is a method employed by healthcare insurers to estimate risk, typically using an individual's demographics. Using risk adjustment, we can determine if a patient represents a higher than estimated risk given the patient's age, sex, geography, and clinical diagnoses (e.g., chronic conditions). Risk-adjustment is often cost- or utilization-based, and complicating cost-based risk adjustment further is that two separate physicians may (for a variety of reasons) charge two different prices for similar healthcare services. If we do not risk-adjust and simply take the top *n*% of healthcare expenditures, we are likely to see patients with the costliest conditions e.g., cancer, heart disease, diabetes, Alzheimer's disease, etc. ("Health and Economic Costs of Chronic Diseases CDC," 2020). A reasonable degree of *expected* cost and utilization are implicit in these costly expenditures, like a maternity diagnosis. It is important to tease out avoidable or excessive cost and utilization from important utilization (e.g., treatment). Thus, if we want to ensure the utility of a predictive model built on cost and utilization thresholds, risk adjustment is compulsory.

An alternative approach, reviewed in "Section 2.1: 'Super Utilizer' Defined," is to focus on utilization that we know is excessive or avoidable given any condition. For example, one of the primary goals of the Affordable Care Act (ACA) was to decrease the number of emergency care visits by connecting patients with PCPs who could better manage non-emergent conditions (Createspace Independent Pub & Office of the Legislative Counsel, 2010). However, the requirement by the ACA to have health insurance took effect January 1, 2014, and a 2015 poll conducted by the American College of Emergency Physicians (of Emergency Physicians & Others, 2015) showed that 75% of physicians reported seeing an increase in the number of ED visits since the requirement was activated (Poole et al., 2016).

Substituting primary care for non-emergent ER utilization is an important goal and predicting future super-utilizers of ER services supports this goal. Outside of some truly unfortunate and improbable "bad luck," there are few sufficient reasons why a person should continuously experience emergent healthcare situations and visit an ER; however, a patient may be challenged by problems with PCP access, including issues surrounding social determinants of health (*e.g., socioeconomic status, household composition, minority status and language, and housing and transportation*). Identifying patients who face such obstacles is the first step in aligning them with the help they need.

Another similar example of excessive or avoidable utilization are repeated and/or avoidable hospital admissions and readmissions. In fact, the ACA required the Secretary of the Department of Health and Human Services to establish the Hospital Readmissions Reduction Program (HRRP) starting October 1, 2012. The HRRP is a "Medicare value-based purchasing program that encourages hospitals to improve communication and care coordination to better engage patients and caregivers in discharge plans and, in turn, reduce avoidable readmissions" ("Hospital Readmissions Reduction Program (HRRP)," 2020). Excess readmission ratios (ERRs) are used to assess hospital performance using predicted-to-expected readmissions rates for conditions and procedures in included in the program:

- Acute Myocardial Infarction (AMI)
- Chronic Obstructive Pulmonary Disease (COPD)
- Heart Failure (HF)
- Pneumonia
- Coronary Artery Bypass Graft (CABG) Surgery
- Elective Primary Total Hip Arthroplasty and/or Total Knee Arthroplasty (THA/TKA)

Similarly, the AHRQ has developed Prevention Quality Indicators (PQIs), which use hospital discharge data to identify admissions that might have been avoided through access to better ambulatory care ("Prevention Quality Indicators Overview," 2020). ER and IP super-utilization are similar in that both are hospital-based. Harris et al. (2016) submits that super-utilizers of hospital-based healthcare—*especially those with multiple chronic conditions (MCCs)*—are particularly amenable to care transition interventions. In their study, Harris et al. (2016) defines super-utilizers as patients with MCCs who experience multiple hospitalizations and emergency department (ED) visits in a 6-month period. Furthermore, following the method of Jencks et al. (2009), Harris et al. (2016) published exhaustive exclusion criteria, which included a principal diagnosis of cancer, pregnancy-related diagnosis, or a surgical procedure for an acute problem (See Supplementary Material Appendix C. Exclusion Criteria).

Following the method of Jencks et al. (2009) and Harris et al. (2017), I administered the comprehensive exclusion criteria to remove patients with a principal diagnosis such as cancer, pregnancy-related diagnosis, or a surgical procedure for an acute problem. I then conducted a thorough exploratory data analysis on the Medicare sample with exclusions applied to determine where members fell with respect to ER and IP utilization in the 95th percentile. I arrived at a "super-utilizer" definition like that of Harris et al. My operational definition for "super-utilizer" is any patient who has *at least two or more ER visit*s or *two or more IP visits* over a 6-month period. It is important to note that a patient who had one ER visit and one IP visit **would not qualify** as a "super-utilizer" (even after exclusion

criteria are applied), as this likely represents an inpatient admission via emergency department transfer—a common occurrence in the data. Just over 2.7% of Medicare beneficiaries fit the super-utilizer definition in the CMS data.

CHAPTER 2.

LITERATURE REVIEW

2.1 "Super-Utilizer" Defined

The term "super-utilizer," used interchangeably with the term "high utilizer" in the literature, is employed dichotomously in the literature, referring to either some threshold of healthcare expenditure, or some threshold of utilization of healthcare services. When defined by cost, it is typical to find references to the statistic published by Mitchell (Mitchell, 2017): "*only five percent of the population accounted for over half of healthcare spending*." In such studies, researchers generally define high utilization as some expenditure threshold, e.g., the top 5%, 10%, 15%, or 20% of a population by cost. Individuals who cost at or above the expenditure threshold are defined as "super-utilizer," and individuals who cost under the threshold are deemed "not super-utilizer."

Sheets et al. defined high utilization as the top 5% (Medicaid and Medicare members) with the highest health system charges (Sheets et al., 2017). Robinson et al. (R. L. Robinson et al., 2016), and Yang et al. (Yang et al., 2018) (who looked at Medicaid members) administered a threshold of the top 10% of healthcare expenditures. Sterling et al. classified high utilizers (Commercial, Medicaid, Medicare, and "other" coverage) as those in the top 20% ranked by dollars spent.

Researchers who defined "super-utilizers" by utilization measures generally established some threshold of the total number of visits in a specific setting over a given period. For example, Dworkis et al. focused on high utilizers of emergency care and defined high utilization as "individuals who utilize emergency care at higher-than-average rates" (Dworkis et al., 2016). Harris et al., 2016 researched emergency and inpatient settings and defined high utilizers as patients with three or more inpatient or observation status hospitalizations, or two or more hospitalizations and two or more prior emergency department visits in a 6-month period. Rinehart et al. (2018) and Turbow et al. (2018) investigated

utilization in the inpatient setting and defined high utilization three hospitalizations in a 12-month period. All healthcare services-based publications permitted readmissions as part of the operational definition for high utilization.

Hasselman (2013) reported that The Camden Coalition conducted a cluster analysis, which identified various subpopulations by categorizing patients based on their utilization history. Note that Rinehart et al. also conducted a cluster analysis on high utilizers (*i.e., using the definition of three hospitalizations in a 12-month period*) using Latent Class Analysis (LCA) to identify subgroups within the high utilization cohort (Rinehart et al., 2018).

For the purposes of my dissertation, I am focused on a particular subset of people *who do not suffer from* well-known clinical, social, or behavioral conditions that indicate the need for future inpatient admissions or emergency department visits, but who still experience several such events in the extreme. Thus, I operationalize the "super-utilizer" definition as dependent on high utilization as opposed to "super-cost."

2.2 Super-Utilizer Exclusions

As described in "Section 1.3: Task Formulation," administration of appropriate exclusion criteria (or proper risk-adjustment) is critical when measuring super-utilizers. In my research, I apply the exclusion criteria described by Jencks et al. (2009) and Harris et al. (2016) (See Supplementary Material Appendix C. Exclusion Criteria). In her publication detailing common themes from the "Super-Utilizer Summit," Hasselman (2013) describes how predictive models have been applied by multiple institutions to predict high utilizers. She specifies routine "rule out" criteria as utilization related to *pregnancy, oncology, trauma, a surgical procedure for an acute condition, and advanced age (e.g., patients over 80 years of age) and a dementia diagnosis*.

Without proper application of exclusion criteria or risk-adjustment, it is common to find misleading characteristics associated with super-utilization. For example, Yang et al. (2018) compares

the performance of four different predictive models, including ordinary least squares (OLS) linear regression (LR), regularized regression (LASSO), gradient boosting machine (GBM), and recurrent neural networks (RNN). The publication includes a figure illustrating the most important features for prediction of their top-performing model (i.e., RNN), which is shown in Figure 1.

Figure 1. Yang et al. illustrate the top performing features of their RNN model

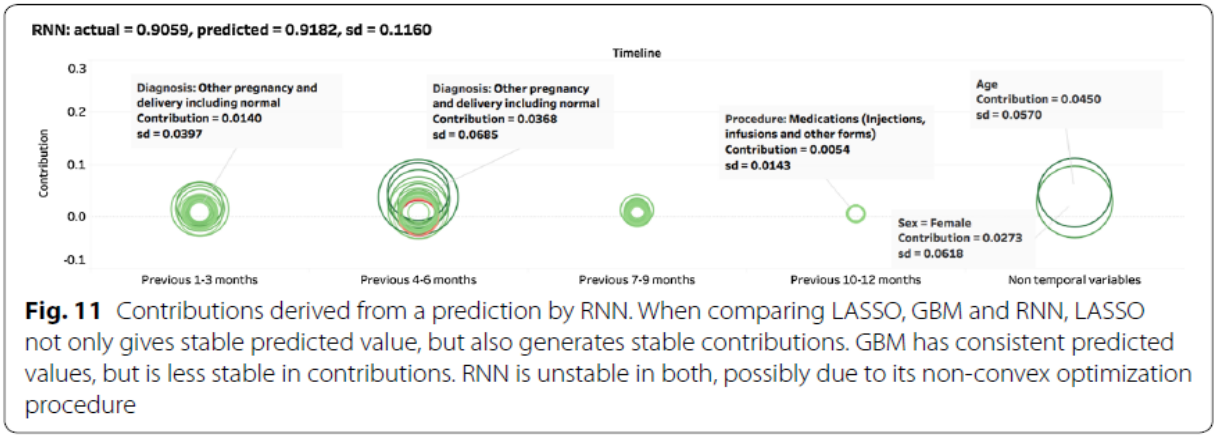


Fig. 11 Contributions derived from a prediction by RNN. When comparing LASSO, GBM and RNN, LASSO not only gives stable predicted value, but also generates stable contributions. GBM has consistent predicted values, but is less stable in contributions. RNN is unstable in both, possibly due to its non-convex optimization procedure

The figure indicates that among the most important features for prediction of high utilizers are multiple features related to pregnancy, which should not explain high utilization in practice. Although a maternity diagnosis is uncommon in Medicare data, it is still important to consider canonical exclusion criteria when tagging high utilizers to ensure that researchers properly label the population of interest.

2.3 Methods & Benchmark Performance

The utility of a model in production depends on if its predictive performance is acceptable to its end users (e.g., a healthcare institution). For the binary classification task of predicting high utilizers of healthcare, the most relevant performance metrics are precision and recall. These metrics are derived from the domain of information retrieval. In information retrieval, precision is defined as $precision = \frac{|\{relevant\ documents\} \cap \{retrieved\ documents\}|}{|\{retrieved\ documents\}|}$, where recall is $recall = \frac{|\{relevant\ documents\} \cap \{retrieved\ documents\}|}{|\{relevant\ documents\}|}$. For classifying super-utilizers, I use the same definitions

and treat true positive super-utilizer predictions as relevant documents; all predicted super-utilizers as retrieved documents; and all actual super-utilizers as relevant documents. Thus, precision is defined as $precision = \frac{true\ positive}{true\ positive + false\ positive}$, and recall as $recall = \frac{true\ positive}{true\ positive + false\ negative}$. Healthcare institutions have finite resources, so it is imperative that predictive models identify a focused set of patients at high risk for the outcome variable being modeled.

Yang et al. employ a linear regression and the tree based XGBoost algorithm (see "Section 3.4.1.2: XGBoost" for an introduction to the XGBoost algorithm) to analyze variations in Medicaid health care expenditures from health care utilization adjustment models (Yang et al., 2018). The dependent variable is represented by a beneficiary's per-member per-month (PMPM) dollar amount, which is the total Medicare expenditure divided by the number of months that a beneficiary is enrolled in Medicaid. The researchers analyze potentially preventable hospital readmissions (PPR) and potentially preventable emergency department visits (PPV) for a population with higher-than-expected residual values to examine variation in cost. The authors conclude that the high-utilizer population is associated with more potentially preventable events.

Ng et al. (2020) conducted a retrospective cohort study in Singapore to predict persistent high utilizers of healthcare in 683,160 cardiology patients. The authors formulate the prediction task as a binary classification and define high utilizers as patients who have expenditures in the 90th percentile for all inpatient and outpatient settings in one year. Support vector machine (SVM), penalized regression, and XGBoost were compared, and performance was evaluated using specificity, sensitivity, and area under the ROC curve (AUC) scores[1]. XGBoost was selected to improve predictive accuracy given the task's class imbalance, and it achieved top performance with an average precision of 37% at a sensitivity up to 40% for high-utilization prediction over 12 months.

[1] As compared with overall accuracy, ROC curve AUC exhibits several desirable properties as a single classification performance measure. (Bradley, 1997)

Sheets et al. (2017) formulated their super-utilizer prediction task as binary classification (i.e., they classified people as *super-utilizers* or *not super-utilizers*) that targeted the top 5% of healthcare expenditures in the overall population. The authors employed contrast mining to identify patterns frequently associated with high healthcare costs and then used the attributes in these patterns as input features to a logistic regression algorithm (See "Section 3.4.1.1: Logistic Regression" for an introduction to logistic regression). Their system reportedly achieved state-of-the-art results with an area under the ROC curve (AUC) score of 0.84, which they reported was "markedly higher than the ROC value of 0.7 reported in comparable models." However, the area under the ROC (AUC) curve—*commonly used to compare the performance of multiple classifiers in a single number*—is an inadequate measure of system performance in imbalanced classification tasks (Brabec & Machlica, 2018) because AUC does not take the class priors into account. The authors did report precision, which was 20.0%. However, the performance is justified as follows:

> *While the positive predictive value of 20% and negative predictive value of 98% appear low and high, respectively, they are reasonably useful given a population in which only 5% of patients are truly positive for high cost, and 95% of patients are negative. For example, a positive predictive value of 20% would result in five patients receiving the intervention of care management for every patient actually destined to incur high costs without intervention. This over-treatment penalty may be reasonable because care management is both extremely safe and relatively inexpensive, and because the 98% negative predictive value of the model would direct population health managers away from nearly all patients who will not incur the highest 5% of costs without the intervention.*

This research represents state-of-the-art performance, benchmarked at a precision of 20.0%, for a problem that is most like my task formulation. Like my research design, Sheets et al. (2017) formulated a binary classification task of "high utilizer" vs. "not high utilizer." However, my task is focused on "super utilizers" of ER and IP services, extending the work of Harris et al (2016). This is perhaps a more challenging task, as it is even more imbalanced (*i.e., a 2.7% positive to 97.3% negative class imbalance, as opposed to a 5.0% positive to 95.0% negative class imbalance*).

Neural network embeddings have become ubiquitous in support of different binary classification tasks. Embedding large and high-dimensional data into low-dimensional vector spaces through word embeddings has facilitated breakthroughs in natural language processing (NLP) and speech recognition and supports well-known applications such as Amazon Alexa and Google's search engine (Saravia, 2018). State-of-the-art embedding approaches like 'word2vec' (Goldberg & Levy, 2014) or 'node2vec' (Grover & Leskovec, 2016) are well established tools in this space. The idea of adapting embeddings to non-NLP problems is now mainstream. For example, image embeddings are now used in computer vision tasks (Kiela & Bottou, 2014), and recently embeddings have been leveraged to learn the multilevel structure of electronic healthcare records based on the encoded relationships between medical codes, which are used in downstream prediction tasks (Choi et al., 2018). In clinical NLP, embeddings have been used to learn clinical text (*e.g., characters, words, sentences, and documents*), concepts (*e.g., concept unique identifiers and medical codes*), and patients (Kalyan & Sangeetha, 2020).

2.4 Super-Utilizer Characteristics

In their study, Dworkis et al. examined the implications of multi-patient high-utilizer addresses (MPHUAs), e.g., homeless shelters and nursing homes, and how individuals representing these organizations likely require special interventions as compared with other groups of high utilizers. The researchers found that MPHUAs accounted for nearly 10% of emergency department visits in their sample. MPHUAs may indicate homelessness and correlate with mental health problems, which suggest individual and systemic health needs at these geographical locations. Abdominal pain and alcohol abuse were also factors consistent with MPHUAs (Dworkis et al., 2016).

Sterling et al. investigated the association of behavioral health factors and social determinants of health with high and persistently high healthcare costs. Interestingly, the authors found that age, gender, racial/ethnic distributions, education, employment status, income, or living conditions did not help to discriminate between high utilizers (i.e., cost) and patients who were not high utilizers. However, the

researchers found a correlation between high utilization and divorce, separation, or patients who were never married. Other behavioral health and social factors associated with high utilization include a psychiatric diagnosis and mental problems. Of chronic conditions examined, 75% were not correlated with high utilization, including diabetes, hypertension, congestive heart failure (CHF), ischemic heart disease, osteoporosis, epilepsy, and cancer. However, asthma, arthritis, chronic obstructive pulmonary disease (COPD), and chronic pain were all found to be associated with high utilizers. High utilizers also had a more substantial disease burden, including an increased prevalence of three or more chronic conditions, and more frequent combinations of both physical and behavioral health problems together. Lastly, high utilizers had an elevated prevalence of behavioral health conditions such as anxiety, depression, major psychosis, personality disorders and other psychiatric conditions (Sterling et al., 2018).

In another study exploring how depression affects utilization, Robinson et al. (2016) concluded that high utilizers of healthcare experience more symptoms of depression. These symptoms include fatigue, sleep issues, anxiety, pain, and substance abuse. However, the authors discriminated between the presence of these attributes in high utilizers and in those who become high utilizers and found that only some of the characteristics existed in both cohorts. Among these characteristics were obesity, cardiovascular disease, illness burden related to comorbidities, pain, and other psychiatric illnesses other than depression (Robinson et al., 2016).

In their study, Harris et al. (2016) reported that nearly 33% of emergency care patients with multiple chronic conditions were released to a nursing home or hospice or had major social risk factors that indicated home-based care transition programs centered on medication control would likely be ineffective. Remarkably, although the patients' conditions were considerably severe (i.e., all patients presented with multiple chronic conditions), 46% of these individuals did not retain a primary care

provider (PCP) (Harris et al., 2016). This overwhelming lack of PCP relationships signifies the importance of care coordination in high utilization populations.

Sheets et al. exploited traditional demographic variables and clinical attributes using billing codes from EHRs. The researchers used contrast mining to find frequent patterns among high utilizers. The most predictive features mined confirm well-known associations with high utilization, including expensive medical diagnoses, e.g., ischemic heart disease, depression, osteoarthritis, and hypertension; and prescriptions that treat these diseases, e.g., beta-adrenergic blocking agents, benzodiazepines, and respiratory agents (Sheets et al., 2017).

Although it is unfair to consider the predictive performance reported by Yang et al. (2018), it is relevant to note that the researchers examined diagnosis codes (i.e., ICD-9-CM) group into Clinical Classifications Software (CCS) diagnostic categories; procedure codes (i.e., current procedural technology (CPT) and Healthcare Common Procedure Coding System (HCPCS)) grouped into CCS procedural categories; pharmacy information represented by National Drug Codes (NDC) grouped by the U.S. Food and Drug Administration (FDA) NDC Directory pharmacy classes; and demographic information e.g., age, sex, race/ethnicity, and disabled status (Yang et al., 2018). These are conventional variables employed in healthcare utilization prediction tasks.

In her review of the Super-Utilizer Summit, Hasselman (2013) reported on the features different healthcare institutions employ to predict high utilizers. She published that in a 15-month timeframe in Washington State, 9 of 10 frequent utilizers of the emergency department had a substance abuse problem. Astonishingly, the paper states that 100% of patients had an indication of mental illness. Furthermore, 20% of high utilizers were homeless; 30% of individuals resided in (or recently lived in) a group care setting; and 10% utilized home health. Common clinical attributes included diagnoses of diabetes and cardiovascular disease. Interestingly, one transcendent theme in the otherwise homogenous population was childhood trauma (Hasselman, 2013). Also, like the work of Harris et al., Kronick et al.

(2007) described that 80% of high-cost Medicaid beneficiaries had three or more chronic conditions, and 60% had five or more chronic conditions.

In summary, researchers tend to focus on different clinical attributes (e.g., diagnosis, procedure, and drug information), characteristics of behavioral and mental health (e.g., psychiatric conditions related to sleep disorders, anxiety, and depression), and complex social challenges including joblessness, homelessness, substance abuse, and unstable or chaotic living conditions. In my dissertation, I intend to operationalize these features (and more), and plan to discuss the predictive power of each—including relationships of characteristics.

CHAPTER 3.

METHODS

3.1 Claims Processing

To process the CMS Medicare claims data, we followed the Medicare Claims Processing Manual published by CMS ("100-04," 2020). The data is loaded into a database (i.e., PostgreSQL) as separate raw claims files (i.e., *the Master Beneficiary Summary File Quarterly, the Carrier File Quarterly, the Durable Medical Equipment File Quarterly, the Home Health File Quarterly, the Hospice File Quarterly, the Inpatient File Quarterly, the Outpatient File Quarterly, and the Skilled Nursing Facility File Quarterly*). Each file contains a Limited Dataset Beneficiary Identifier, which has been deidentified by the Centers for Medicare & Medicaid Services, but makes it possible to link claims for members across the different files in accordance with the Medicare Claims Processing Manual documentation.

The data contains clinical information (such as diagnoses, procedures, and diagnosis related groups for inpatient admissions), demographic information (such as age, race, and gender), and benefit details (such as reasons for Medicare enrollment). Different publicly available clinical tools such as Clinical Classification Software (CCS), etc. are mapped on top of the clinical information (for more details, see "Section 5.3.1.1: Tabular Data" in the discussion section on feature and algorithm performance). A de-identified county-level address is provided, which is used to map different publicly available tools (e.g., Social Vulnerability Index (SVI), etc.; See "Section 5.3.1.1: Tabular Data").

3.2 Technical Design of Datasets for Model Training and Evaluation

To maximize the 18 months of retrospective data I received from CMS, I employ two different data partitioning schemes for model training and evaluation: (1) datasets as a decomposition of time; and (2) datasets as a decomposition of population (the details and implications of these experiments are discussed in "Section 5.1: Discussing Dataset Design"). The two different datasets are used to train two

distinct classifiers: (1) a classifier trained on 6 months of historical data to predict the subsequent 6 months of data, and (2) a classifier trained on 12 months of historical data to predict the subsequent 12 months of data. In general, the first data partitioning schema provides more training data, but the amount of historical data contains 50% fewer samples as compared with the second data partitioning scheme. Accordingly, the second data partitioning scheme provides fewer training samples, but the amount of historical data is 100% more than in the first partitioning scheme.

The splits of the datasets have an influence on model performance and can cause class imbalance problems or issues with sample representatives if not done carefully (Liu & Cocea, 2017). Stratified random sampling is employed to maintain a proper class balance and ensure sample representativeness at the cost of expensiveness to prepare (Acharya et al., 2013). Each dataset partition is further partitioned into a training set, a calibration set, and a holdout set. In the datasets as a decomposition of time experiment, the training set includes the first 6 months of historical data for features and the second 6 months of data to calculate labels, and the holdout set includes the second 6 months of historical data for features and the final 6 months of data to calculate labels. The calibration set, used to estimate hyperparameters, merely contains a 30% stratified random sample of the training set. After Bayesian Hyperparameter Optimization[2] is performed (which is introduced in "Section 3.4.3: Bayesian Hyperparameter Optimization"), the calibration set is padded back to the training set such that a final production model can be trained using the estimated hyperparameters on the full training set.

In the datasets as a decomposition of population experiment, a model development set is created using 12 months of historical data for feature engineering and the final 6 months of data is used to calculate labels. A 70% stratified random sample of the model development set is used for training, and the remaining 30% is used as a holdout (i.e., "test")[3] set. A calibration set is created from a 30%

[2] Hyperparameter optimization describes the process of selecting optimal hyperparameters for a learning algorithm because these model parameters cannot be estimated directly by the algorithm (Feurer & Hutter, 2019).

[3] Note that the terms "holdout set" and "test set" are used interchangeably.

stratified random sampling of the training set. After Bayesian Hyperparameter Optimization is performed (which is introduced in "Section 3.4.3: Bayesian Hyperparameter Optimization"), the calibration set is padded back to the training set such that a final production model can be trained using the estimated hyperparameters on the full training set.

After the models are fully trained, it is then evaluated on the holdout set and evaluation metrics (i.e., precision, recall, etc.) are reported for discussion.

3.2.1 Designing Datasets as a Decomposition of Time

I decompose the overall data into two separate datasets by *time*: (1) a model development set with a *6-month feature period* and *6-month label period* (See Table 5. Model Development Set Design with Decomposition of Time for the model development set longitudinal design); and (2) a test set with a *6-month feature period* and *6-month label period* (See Table 6. Test Set Design with Decomposition of Time for the test set longitudinal design) that are both six months ahead of the respective feature and label periods in the model development set.

Table 5. Model Development Set Design with Decomposition of Time

Model Development Set												
6-Month Feature Period						**6-Month Label Period**						
04-2017	05-2017	06-2017	07-2017	08-2017	09-2017	10-2017	11-2017	12-2017	01-2018	02-2018	03-2018

Table 6. Test Set Design with Decomposition of Time

Test Set												
6-Month Feature Period						**6-Month Label Period**						
10-2017	11-2017	12-2017	01-2018	02-2018	03-2018	04-2018	05-2018	06-2018	07-2018	08-2018	09-2018

I completely withhold the test set during the model development process, as it is used *only* for final evaluation against the production model as predictions generated in the wild. I further segment the model development set into training, calibration, and holdout sets (See Table 7. Number of Beneficiaries by Dataset for population sizes per dataset).

Table 7. Number of Beneficiaries by Dataset

	Model Development Set			
	Training Set	**Calibration Set**	**Holdout Set**	**Test Set**
# of beneficiaries	843,593	361,541	516,487	1,366,434

The training, calibration, holdout, and test sets are used independently or in combination depending on the phase of experiment, e.g., hyperparameter tuning, stratified k-fold cross-validation, model training, and model evaluation. The technical details of these experimental phases where data are decomposed d by time are described further in "Section 3.4" Experimental Design."

3.2.2 Designing Datasets as a Decomposition of Population

I decompose the entire duration of the longitudinal dataset into two partitions by *population*, i.e., a random 70%-30% split between the model development set and test set, respectively. Accordingly, both the model development and test sets contain *12-month feature period* and *6-month label period* (See Table 8. Model Development Set Design with Decomposition of Population for the longitudinal dataset design).

Table 8. Model Development Set Design with Decomposition of Population

Dataset (70% Model Development Set + 30% Test Set)											
12-Month Feature Period											
04-2017	05-2017	06-2017	07-2017	08-2017	09-2017	10-2017	11-2017	12-2017	01-2018	02-2018	03-2018
6-Month Label Period											
04-2018	05-2018	06-2018	07-2018	08-2018	09-2018						

Again, I completely withhold the test set during the model development process, as it is used *only* for final evaluation against the production model as predictions generated in the wild. I further segment the model development set into training, calibration, and holdout sets (See Table 9. Number of Beneficiaries by Dataset for population sizes per dataset).

Table 9. Number of Beneficiaries by Dataset

	Model Development Set			
	Training Set	**Calibration Set**	**Holdout Set**	**Test Set**
# of beneficiaries	260,191	111,511	159,301	227,345

Just as in my experimentation where the data are decomposed by time, here the training, calibration, holdout, and test sets are used independently or in combination depending on the phase of experiment, e.g., hyperparameter tuning, stratified k-fold cross-validation, model training, and model evaluation. The technical details of these experimental phases where data are decomposed by population are described further in "Section 3.4: Experimental Design."

3.3 Technical Design of the Dependent Variable

As described in Introduction "Section 1.3: Task Formulation," and discussed further in "Section 5.2: Discussing Dependent Variable Design," proper inclusion and exclusion criteria (detailed further in

"Section 3.3.1: Inclusion Criteria" and "Section 3.3.2: Exclusion Criteria", respectively) are essential to proper design of the dependent variable, which has been constructed to label future super-utilizers of ER and IP services for whom such high levels of utilization is unexpected given a beneficiary's demographic and diagnostic history.

3.3.1 Inclusion Criteria

Beneficiaries were included in the study if:

- Beneficiary's full claims history was represented in the data.
- Beneficiary was not enrolled through Medicare Advantage (i.e., an HMO)
- Beneficiary's entitlement reason was not end-stage renal disease (ESRD)
- Beneficiary did not have a valid "date of death" in their file (i.e., no known mortality)

To qualify for experimentation given the dataset design, a beneficiary was included if enrolled in Medicare throughout the full span of the data (i.e., the second quarter of 2017 to the third quarter of 2018). This inclusion is important for the decomposition of population experimentation because to ensure a 12-month claims history and 6-month prediction window, a beneficiary must have the full 18 months (i.e., 6 quarters) of data containing features a label (i.e., *"super-utilizer" vs. "not super-utilizer"*).

Medicare Advantage beneficiaries are not included in experimentation because CMS does not manage Medicare Advantage claims, i.e., it only has access to these beneficiary's membership information. Members who suffer from ESRD are much more likely to experience inpatient admissions associated with dialysis or kidney transplants (Tam-Tham et al., 2020), and ESRD is not known to be amenable to intervention. Thus, only members entitled to Medicare for reasons other than ESRD are included.

3.3.2 Exclusion Criteria

Motivations for exclusion criteria are reviewed in "Section 1.3: Task Formulation". Beneficiaries were excluded in the study if:

- Beneficiary's full claims history was not represented in the data.
- Beneficiary is enrolled in Medicare through Medicare Advantage (i.e., an HMO).
- Beneficiary's entitlement reason was end-stage renal disease (ESRD).
- Beneficiary had valid "date of death" in their file (i.e., known mortality).

Beneficiaries were excluded from model training if:

- Beneficiary had a diagnosis (i.e., DRG and/or ICD-10-CM) and/or procedure (i.e., ICD-10-PCS) code in the list of exclusions (See Supplementary Material Appendix C. Exclusion Criteria).

When a full claims history is not represented in the data, we cannot fairly compare utilization between beneficiaries. However, for nearly all the Fee-For-Service Medicare beneficiaries, a full claims history is provided by CMS. In some cases, a beneficiary had a valid "date of death" in their file indicating mortality. Beneficiaries who experienced mortality are not included in the study because a full 18-month claims history was not represented. In every instance, the models are trained with the diagnosis and procedure exclusions applied. These exclusion criteria are applied to focus classification on a subset of super-utilizers who are potentially amenable to intervention. See Supplementary Material Appendix C. Exclusion Criteria for the exhaustive set of exclusion criteria, and "Section 5.2: Discussing Dependent Variable Design" for more discussion on the design of the dependent variable.

3.3.3 Class Distributions

My operational definition of "super-utilizer" is any patient who has *at least two or more ER visit*s or *two or more IP visits* over a 6-month period *after* exclusions are applied (See "Section 1.3: Task Formulation" and "Section 2.1 'Super-Utilizer' Defined" for justifications of this definition), which represents about ~2.7% of the Medicare population in the study. Ultimately, when inclusion and

exclusion are applied across the experiments, we see consistent distributions between the majority and minority classes in the training and test sets. The majority and minority class distributions also indicate the difficulty of this task.

Figures 2, 3, 4, and 5 demonstrate a consistent majority to minority class distribution between the training and test sets across both experiments (*i.e., the 6-month and 12-month feature windows with 6-month label periods*). These consistent distributions between class labels across data sets and experiments are expected and provide face validity to the design of the dependent variable. Moreover, the consistent class distributions indicate that the super-utilization phenomenon of ER and IP services is similarly distributed across time, and that the exclusion logic is coherent.

Figure 2. Majority to Minority Class Samples in the Training Set for the 6-month feature window

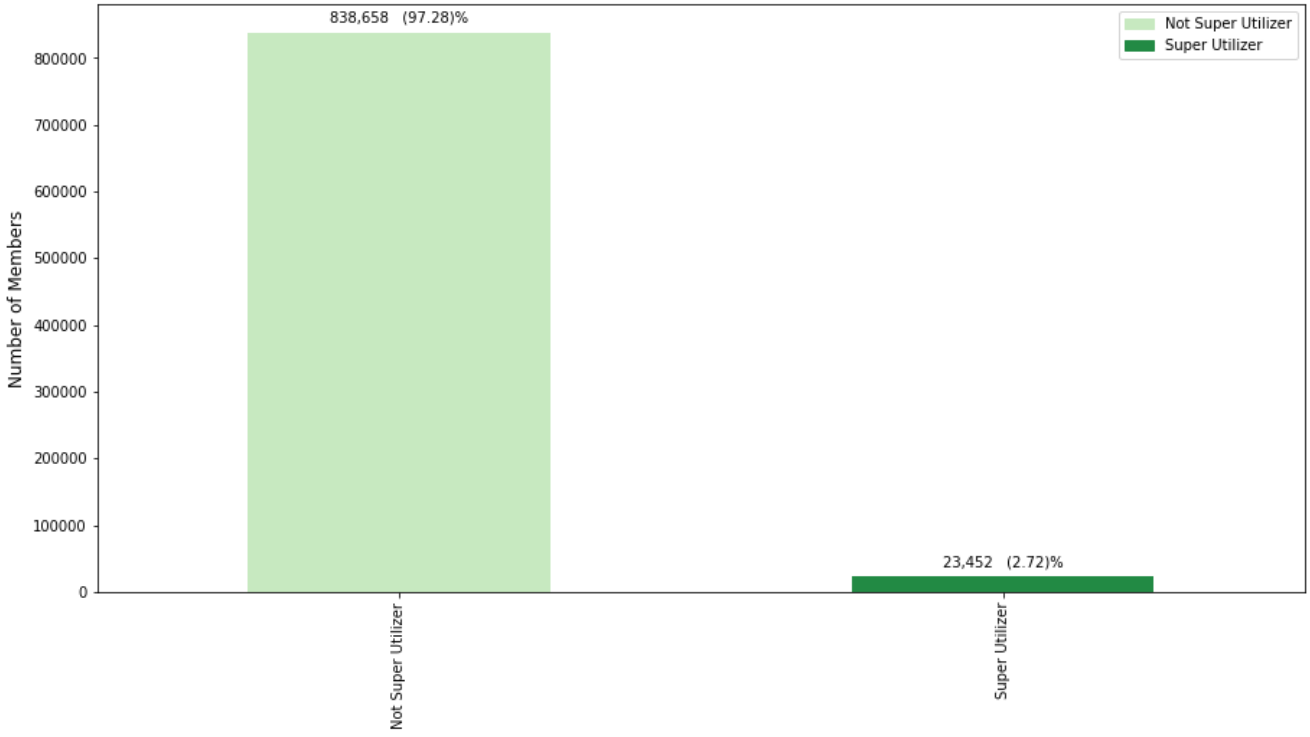


Figure 3. Majority to Minority Class Samples in the Test Set for the 6-month feature window

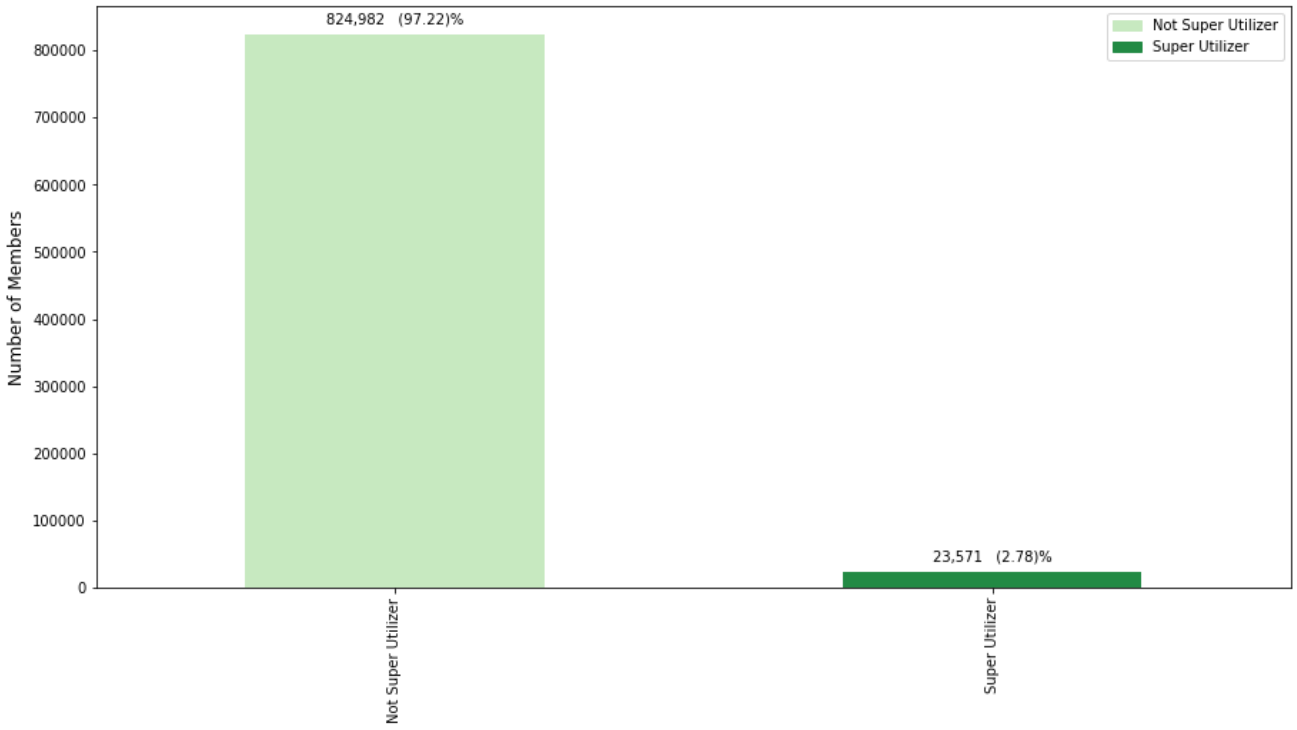


Figure 4. Majority to Minority Class Samples in the Training Set for the 12-month feature window

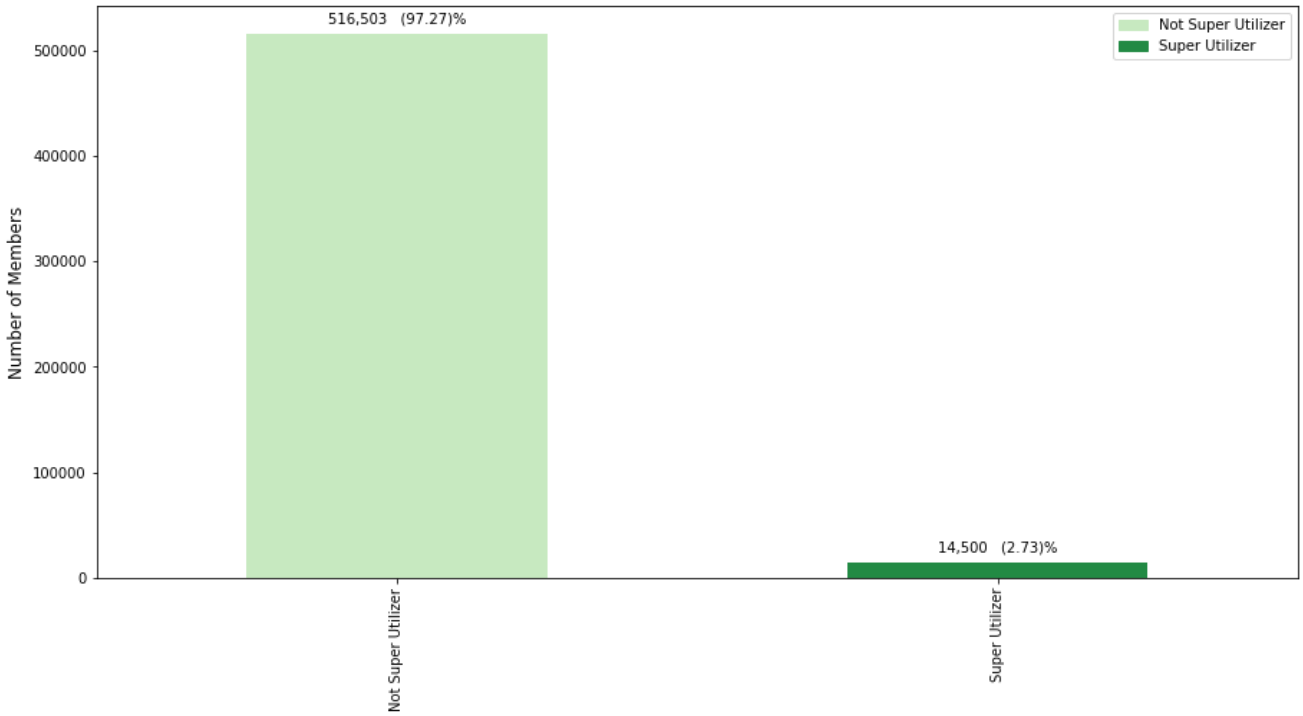
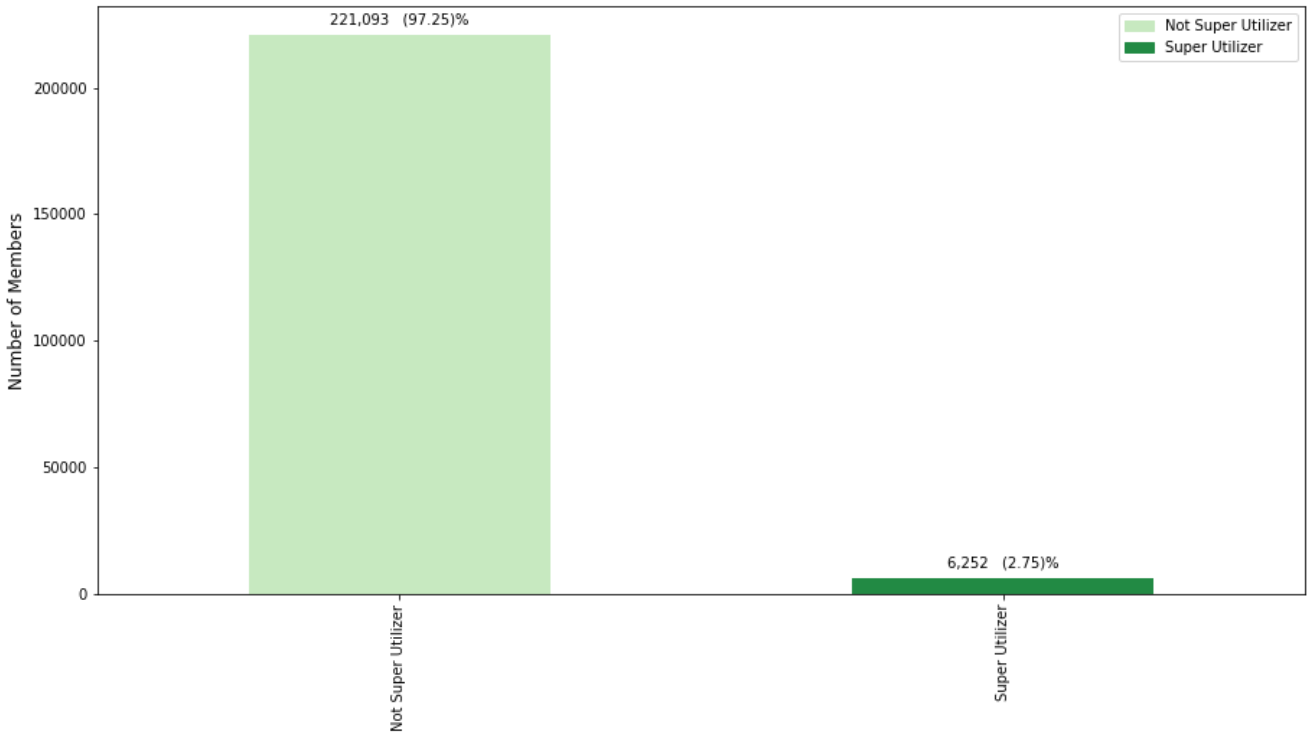


Figure 5. Majority to Minority Class Samples in the Test Set for the 12-month feature window



3.4 Experimental Design

As introduced in "Section 1.3: Task Formulation", to identify future super-utilizers, I formulate the problem as a binary classification task. Thus, each beneficiary is labeled as either a "super-utilizer" (i.e., *a positive sample "1"*) or "not super-utilizer" (i.e., *a negative sample "0"*). Further, each "instance" (i.e., a "beneficiary") is represented as a collection of features (i.e., the "feature space"), which are individual measurable characteristics that describe instances. I subsequently train and compare several models that learn to discriminate between "super utilizers" and "not super utilizers" so that I can evaluate different measures of model performance against *new data* that was deliberately never exposed to the model during training. This new data (i.e., the test set) was collected and set aside on the onset of experimentation to mimic data used in production (also commonly called "the wild").

Therefore, the final step of experimentation involves evaluating all models against the test set, but only after all other phases of experimentation are complete. No further tuning, training, or post-processing is appropriate once the models are exposed to the test set, as this would not adequately reflect data used in production.

I compare the performance of two state-of-the-art binary classification algorithms (i.e., XGBoost and Object2Vec) to two baseline classifiers (i.e., logistic regression and random forest) commonly applied in the literature (See Chapter 2. Related Work). The metrics I use to evaluate algorithm performance are precision (i.e., $precision = \frac{true\ positive}{true\ positive + false\ positive}$), and recall (i.e., $recall = \frac{true\ positive}{true\ positive + false\ negative}$). Aside from hyperparameter optimization, I treat the algorithms as black box classifiers. My aim is not to modify any of the algorithms I test; rather, my aim is to empirically evaluate which (if any) of the algorithms can be exploited to identify future super-utilizers.

All four of the algorithms that I investigated benefit from properly tuned hyperparameters, which—*unlike model parameters that can be estimated directly from the data*—are parameters external to the model whose values cannot be estimated directly from the data. Therefore, hyperparameters are generally set by a practitioner using either: (1) algorithm defaults, (2) values copied from similar problems in the literature, (3) heuristics, or (4) hyperparameter optimization.

Of the general methods for choosing hyperparameters, *hyperparameter optimization* is now common practice as it derives optimal hyperparameters from data used in a specific task and is thus less reliant on generalization. More specifically, the goal of hyperparameter optimization is to configure (i.e., "tune") hyperparameters to discover parameters of a model that result in predictions that are *most skillful* by some objective metric. Hyperparameter optimization should yield predictions that are *at least as skillful* as predictions generated using hyperparameters (1) set by algorithm defaults, (2) using hyperparameters referenced in literature, or (3) employing heuristics, but in most cases hyperparameter

optimization will yield predictions that are *more skillful*. In this study, I exploit a state-of-the-art Bayesian Hyperparameter Optimization algorithm, which is performed on all models.

The term "super-utilization" refers to an extraordinary level of utilization, which is a phenomenon that occurs infrequently. Thus, I need to mitigate the well-known "class imbalance" problem. The idea behind the class imbalance problem is that classifier performance is often hindered in the presence of highly imbalanced classes—i.e., when the number of samples in the majority class considerably outnumbers that in the minority class. Classification tasks become increasingly difficult as concepts reach higher degrees of complexity and as the level of imbalance in the dataset grows. As described in "Section 1.3: Task Formulation," the super-utilizer classification task is both highly complex and presents a challenging level of class imbalance (i.e., ~2.7% positive to 97% negative class representation). Therefore, I compare the results of a conventional class imbalance mitigation technique (i.e., cost-sensitive classification) to the results of classifiers where no class imbalance technique was employed during training.

To properly evaluate the effectiveness of each model, I compare the performances of all classifiers using (1) stratified 3-fold cross-validation, (2) a holdout set, and (3) on the test set. For strong reliability, I am looking for a model to have consistent performance (i.e., low variance) in each fold during cross-validation, and against the holdout set and test sets. I use the most common objective metrics for classification problems to evaluate classifier performance. Specifically, I am interested in understanding which classifier (if any) yields high precision while still returning an acceptably sized population of super-utilizers to target for intervention. Therefore, the performance measures that I interrogate most are primarily precision and recall. It is important to note that while the Receiver Operating Characteristics (ROC) is typical for evaluating classifier performance in the field of bioinformatics, it is an inappropriate metric to use for evaluation in this study as it is often misleading in the presence of imbalance data.

The implications of my experimental design are discussed further in "Section 5.3.1: Discussing Task Formulation."

3.4.1 Algorithms

3.4.1.1 Logistic Regression. Logistic regression (also known as logit regression, or maximum entropy classification) is a linear model for classification. Intuitively, the logistic regression classifier models the probabilities describing possible outcomes of a single trial using a logistic function. I employed a binary class ℓ_2 penalized (i.e., the default penalty) logistic regression optimization problem that minimizes the following cost function:

$$\min_{w,c} \frac{1}{2} w^T w + C \sum_{i=1}^{n} \log(\exp(-y_i(X_i^T w + c)) + 1)$$

I used the limited memory Broyden-Fletcher-Goldfarb-Shanno algorithm (L-BFGS) in the optimization problem, which is used as the default logistic regression algorithm in the Scikit Learn logistic regression implementation because of its robustness, *e.g., it is robust to unscaled datasets* Table 10. Solver Penalties and Behaviors shows the behaviors and different penalties supported by different solvers (*e.g., the Multinomial + L2 penalty is not supported by the liblinear solver but is supported by the lbfgs solver; and the liblinear solver penalizes the intercept, but the lbfgs does not*).

Table 10. Solver Penalties and Behaviors

Penalties	liblinear	lbfgs	newton-cg	sag	saga
Multinomial + L2 penalty	no	yes	yes	yes	yes
OVR + L2 penalty	yes	yes	yes	yes	yes
Multinomial + L1 penalty	no	no	no	no	yes

OVR + L1 penalty	yes	no	no	no	yes
Elastic-Net	no	no	no	no	yes
No penalty ('none')	no	yes	yes	yes	yes
Behaviors					
Penalize the intercept (bad)	yes	no	no	no	no
Faster for large datasets	no	no	no	yes	yes
Robust to unscaled datasets	yes	yes	yes	no	no

During each phase of experimentation, I experimented both with the *class weight* hyperparameter (1) set to *None* and alternatively (2) set to *balanced*. These experiments allow me to compare the effects of not addressing the class imbalance with the effects of addressing the class imbalance through cost-sensitive classification. When the *class weight* is set to *None*, the penalty for misclassifications in both classes is set to 1; however, when the class weight is set to *balanced*, the penalty for misclassifications of the positive class (i.e., super-utilizers) is set inversely proportional to its class frequencies in the input data. This commonly employed heuristic for setting class weights means that the penalty for a misclassification of a *super-utilizer* (i.e., the minority class of ~3%) is relatively much greater than the penalty for a misclassification of a *not super utilizer* (i.e., the majority class of ~97%). The results for both general classification and cost-sensitive logistic regression classification are presented in "Section 4.2: Cost-Sensitive Classification Results," and a comparison of these results is discussed further in "Section 5.3.2: Bayesian Hyperparameter Optimization & Discussing Cost-Sensitive Classification."

With respect to hyperparameter tuning of the logistics regression classifier, I performed Bayesian Hyperparameter Optimization on the *C* hyperparameter, which is the inverse regularization control

variable. The inverse regularization control variable (i.e., *C*) retains strength modification of regularization through its inverse positioning to the Lambda regulator, which controls regularization strength. Results from the Bayesian hyperparameter optimization experiments on logistic regression are presented in "Section 4.1: Bayesian Hyperparameter Optimization Results," and these results are discussed further in "Section 5.3.2: Bayesian Hyperparameter Optimization & Discussing Cost-Sensitive Classification."

3.4.1.2 Random Forest. A decision tree classifier is a supervised, non-parametric classifier (*i.e., a classifier does not rely on assumptions about an underlying distribution*) that learns to discriminate between classes using decision rules derived from features. A random forest classifier is a meta classifier (*i.e., a classifier that takes other classifiers as parameters*) that fits an ensemble of decision tree classifiers on numerous subsamples of the dataset. In general, ensemble models are based on the premise that the combination of numerous classifiers is more generalizable and robust than any single classifier. The intuition behind this specific ensemble method is that the combined meta classifier will perform better than any single classifier because of reduction in variance. The prediction of the random forest ensemble is generated by averaging the classifiers' probabilistic output to improve performance and reduce overfitting.

To construct each decision tree in the ensemble, samples can be drawn with replacement (often called "bootstrap" samples) or without replacement, which is controlled by the *bootstrap* hyperparameter. While building each decision tree, each node can be split using all input features or a random subset of size *n*, which is controlled by the *max_features* hyperparameter. The randomness inherent in sampling and node splitting serves to decrease the variance of the random forest classifier. One of the weaknesses of decision tree classifiers is that they tend to overfit and fail to generalize due to high variance. However, averaging the probabilistic output of each individual decision tree classifier often offsets prediction

errors in predictions. Furthermore, the ensemble works best when several diverse classifiers are combined to make a final prediction as the reduction in variance is often significant.

The number of individual classifiers to be combined by the meta classifier is controlled by the *n_estimators* hyperparameter. The nodes in each tree will expand until (1) all leaves are pure, or (2) all leaves contain less than the number of samples specified by the *min_samples_split* hyperparameter if the *max_depth* hyperparameter is not set; otherwise, the maximum depth of the tree can be controlled by the *max_depth* hyperparameter. The minimum number of samples required to be at a leaf node is controlled by the *min_samples_leaf* hyperparameter, which means that a split point at any depth must have at least *min_samples_leaf* in both the left and right branches to be considered. Specifying a minimum number of samples at a leaf node for a split point to be considered can have a smoothing effect for the model. For each random forest classifier, I experimented with Bayesian Hyperparameter Optimization to optimize the *max_features*, *n_estimators*, *min_samples_split*, *max_depth*, and *min_samples_leaf* hyperparameters. Results from the Bayesian hyperparameter optimization experiments on the random forest algorithm are introduced in "Section 4.1: Bayesian Hyperparameter Optimization Results," and these results are discussed further in "Section 5.3.2: Discussing Bayesian Hyperparameter Optimization."

Just as with experimentation of the logistic regression classifier, during each phase of experimentation I experimented both with the *class weight* hyperparameter (1) set to *None* and alternatively (2) set to *balanced* (See the explanation of the class weight hyperparameter for cost-sensitive classification in "Section 3.4.1.1: Logistic Regression"). The results for both general classification and cost-sensitive random forest classification are presented in "Section 4.2: Cost-Sensitive Classification Results," and these results are discussed further in "Section 5.3.2: Bayesian Hyperparameter Optimization & Discussing Cost-Sensitive Classification."

3.4.1.2 XGBoost. The "Extreme Gradient Boosting" (i.e., XGBoost) algorithm is another ensemble classifier built for generalizability and robustness. The motivation behind boosting methods is to reduce bias by combining numerous weak models to construct a powerful ensemble. The XGBoost algorithm is currently state-of-the-art, as it has demonstrated superior performance in several recent classification tasks on structured data. XGBoost is a gradient boosted decision tree (GBDT) variant that was specifically designed for both speed and performance. GBDT generalizes boosting to arbitrary differentiable loss functions. XGBoost is a special type of GBDT that includes a novel penalization of trees, proportional shrinking of leaf nodes, Newton Boosting (See Supplementary Material Appendix D. Pseudocode - Newton Boosting for pseudocode), and an extra randomization parameter.

XGBoost takes as input learning task parameters that specify both the learning task and corresponding learning objective. I specified the learning task as binary classification and the learning objective as logistic regression, which will output the probability that a sample belongs to the positive class (i.e., *super-utilizer*) and the negative class (i.e., *not super-utilizer*). The number of decision trees (or rounds) in the XGBoost ensemble is controlled by the *num_rounds* hyperparameter. The algorithm also accepts several hyperparameters for tree boosting.

It is possible to make the model more conservative by increasing the *alpha* hyperparameter, which controls ℓ_1 regularization. ℓ_1 regularization encourages sparsity (i.e., it encourages the weight to move toward *0*). Step size shrinkage used during update to prevent overfitting is controlled by the *eta* hyperparameter. The *eta* hyperparameter also makes the boosting process more conservative by directly retrieving the weights of new features after each boosting step and shrinking the feature weights. The *min_child_weight* hyperparameter controls splitting by setting a threshold for a certain degree of purity in a node, which the model can fit. The algorithm is more conservative as *min_child_weight* values increase. The *subsample* hyperparameter can reduce overfitting by subsampling a ratio of training instances prior to growing trees, which occurs once in every boosting iteration.

Just as in experimentation of the logistic regression and random forest classifiers, I compare general classification with cost-sensitive classification. The XGBoost algorithm has a *scale_pos_weight* hyperparameter that is used to scale the gradient for the positive class as a method of cost-sensitive classification. It is common practice to set the *scale_pos_weight* as inversely proportional to class frequencies in the input data. However, I employed Bayesian Hyperparameter Optimization hyperparameter to estimate this value for cost-sensitive classification. The results comparing general classification with cost-sensitive XGBoost classification are discussed in "Section 5.3.2: Bayesian Hyperparameter Optimization & Discussing Cost-Sensitive Classification."

For each XGBoost classifier, I experimented with Bayesian Hyperparameter Optimization to optimize the *num_rounds*, *alpha*, *eta*, *min_child_weight*, *subsample*, and *scale_pos_weight* hyperparameters. Results from the Bayesian hyperparameter optimization experiments on XGBoost are presented in "Section 4.1: Bayesian Hyperparameter Optimization Results," and these results are discussed further in "Section 5.3.2: Discussing Bayesian Hyperparameter Optimization."

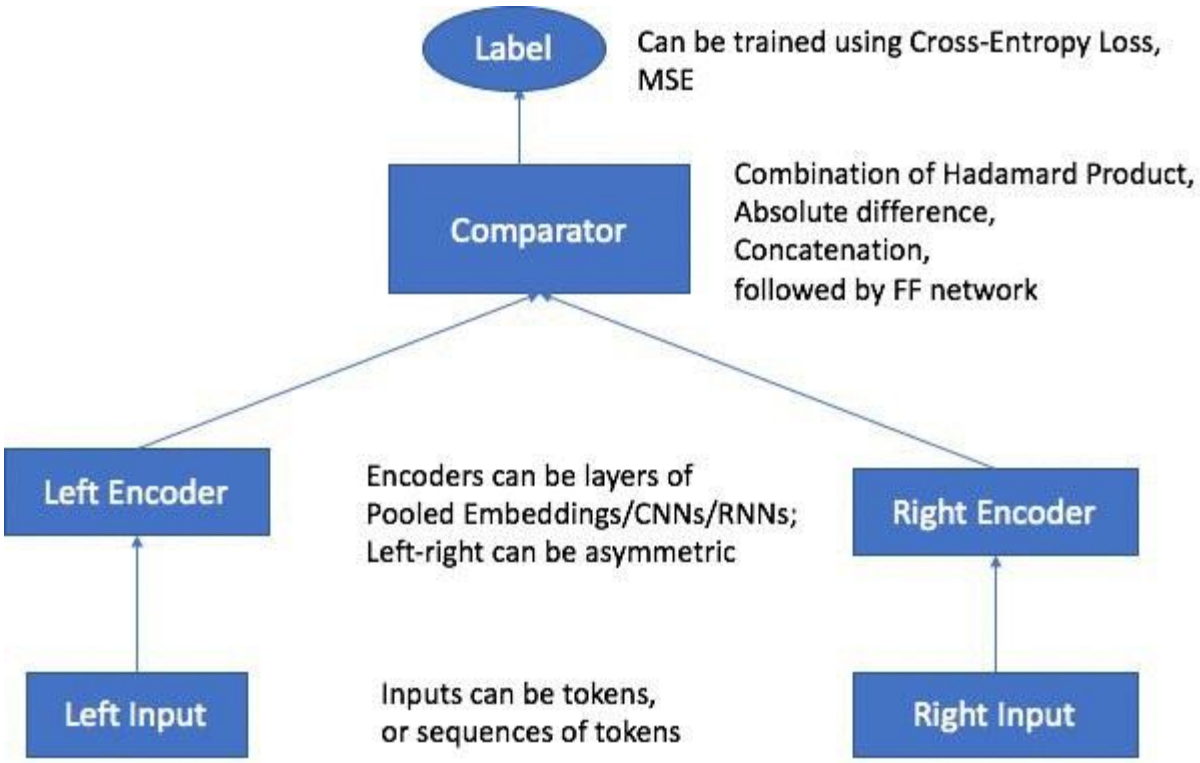
3.4.1.2 Object2Vec. Object2Vec is a state-of-the-art deep learning algorithm that can learn low-dimensional dense embeddings of high-dimensional objects. More specifically, it is a highly customizable, general purpose neural embedding algorithm in which the learned embeddings preserve the semantics of the relationship between objects in the original space and the embedding space. The model artifacts (i.e., the learned embeddings) can then be used for downstream unsupervised and/or supervised tasks. For example, to visualize natural clusters of related objects in the dense low-dimensional space, it is possible to perform cluster analysis by using the learned embeddings as input to an unsupervised clustering algorithm (e.g., nearest neighbors). It is also possible to use the pairs of object embeddings for a downstream supervised regression or classification task.

In this study, the system is architected to learn low-dimensional dense embeddings of pairs (*i.e., left, and right encoders*) of high-dimensional objects: (left encoder) relevant demographic, eligibility,

and utilization factors; and (right encoder) sequential diagnosis and procedure (i.e., ICD-10-CM and ICD-10-PCS) codes. These pairs of encoded embeddings are then qualified by a label (i.e., *super-utilizer* vs. *not super-utilizer)* so that a comparator can learn to quantify the relationship (i.e., similarity and dissimilarity) between pairs of objects relative to whether a beneficiary is a *super-utilizer*. The final output is provided by the comparator as a strength of the relationship score. In this study, the output is a similarity score of the relationship between pairs of *demographic, eligibility, and utilization* objects and *clinical* objects, where a positive relationship is defined as *super-utilization* and a negative relationship is defined as *not super-utilization*.

The architecture of this system, based on the Object2Vec deep learning algorithm (See Figure 6. Architectural Diagram of Object2Vec), has several advantages over a standard machine learning classifier (e.g., logistic regression, random forest, and XGBoost). As previously mentioned, model artifacts include two sets of embeddings, which are in and of themselves contributions that can be shared for other bioinformatics tasks (including downstream classification and regression tasks). What makes these object embeddings special and different from generic embeddings—which are learned representations of different objects (*e.g., words in text*) where objects have the same semantic meaning and similar representations—is that the comparator provides the context for similarity with respect to semantic meaning, as opposed to something more traditional like unsupervised patterns in a grammar. Through the comparator then, *similarity is user defined*. The user can define pairs of objects as similar or dissimilar through a discrete class label (i.e., classification), or as a continuous similarity score (i.e., regression).

Figure 6. Architectural Diagram of Object2Vec



Similarity is defined in this study as people who will (or who will not) experience super-utilization in the future. This means that in the embeddings, people who are *super-utilizers* should share common demographic, eligibility, utilization objects (e.g., age, gender, race, and levels of utilization); it also means that super-utilizers should share similar clinical concepts (i.e., ICD-10-CM and ICD-10-PCS codes). Most importantly, these embeddings are used by Object2Vec in the downstream classification task, which learns to use the relationships between *demographic, eligibility and utilization objects* and *clinical objects* to discriminate between *super-utilizers* and *not super-utilizers.* Furthermore, in addition to classifying whether a beneficiary will experience super-utilization, it is possible to cluster certain beneficiaries into groups based on similar contexts (e.g., beneficiaries with similar demographic, eligibility, and utilization objects and similar clinical objects with respect to super-utilization). However, cluster analysis was not in the scope of this research and is identified as future work (See Chapter 6 Conclusion).

The specific architecture of the Object2Vec system includes 2 input channels, 2 encoders, and a comparator (See Figure 6. Architectural Diagram of Object2Vec). The 2 input channels are highly customizable, as they take different types of inputs (e.g., sequence pairs, token pairs, sequence, and token pairs, etc.). The data from these input channels are then passed to the 2 encoders, which convert objects into fixed length embedding vectors. Note that the assignment of data into the left and right encoder is defined by the user, but arbitrary (*i.e., there is no reason that an object should be placed in the left encoder versus the right encoder, or the right encoder versus the left encoder*). Finally, the fixed length embedding vectors are passed to the comparator, which learns to quantify the strength of the relationship between objects using a discrete (i.e., classification) or continuous (i.e., regression) label provided by the user. As the deep learning model is trained, the loss function minimizes the differences between relationships predicted by the model and those assigned by the user.

Choice of encoder network is important in the design of the neural embedding algorithm. For the purposes of this dissertation, the architectures of encoder networks are considered "black box," but the general intuitions of each are briefly introduced in this section. Object2Vec offers a hierarchical convolutional neural network (HCNN), a bidirectional long short-term memory network (BiLSTM), and pooled embedding. HCNNs are a type of feed-forward artificial neural network (ANN) that is generally suitable for tasks like image, video, and spatial data processing. In the BiLSTM, a signal propagates backward and forward in time, which makes it an appropriate recurrent neural network (RNN) architecture for supervised classification. In Object2Vec, HCNN is beneficial for faster training speed because of parallelization. However, if using sequential input, BiLSTM generally yields better predictive performance. Pooled embedding is specifically designed for super-efficient training (i.e., high speed), with the tradeoff of some loss in predictive accuracy. I compare the results of these different networks in "Section 4.3: Deep Learning Results" and discuss these results (including limitations of the architecture) in "Section 5.3.4: Discussing Algorithm Performance."

The Object2Vec algorithm takes numerous hyperparameters (See "Section 4.1: Bayesian Hyperparameter Optimization Results" for a list of hyperparameters), but the main hyperparameters include (among the encoder network): *optimizer*, *token embedding dimension*, *encoding dimension*, *early stopping tolerance*, *and early stopping patience*. I found that it was critical also to tune the *learning rate* and the *weight decay* hyperparameters. I also decided to tune the number of *epochs* for efficiency and to reduce overfitting, and to tune the number of *mlp layers* in the network to improve predictive performance.

The optimizer is the algorithm used by Object2Vec in the optimization problem that minimizes the loss function between predictions and user-specified ground truth. I chose to use the default adaptive moment estimation algorithm (ADAM) as opposed to empirically testing different optimizers (e.g., per-dimension learning rate method for gradient descent (ADADELTA); adaptive gradient descent algorithm (ADAGRAD); stochastic gradient descent (SGD); root mean square propagation (RMSPROP)). Here, because I was using a baseline approach with respect to the optimizer due to budgetary constraints, I simply chose the default. However, it would be interesting future work to empirically test the efficacy of different optimizers for this task, which is discussed further in "Section 5.3.4: Discussing Algorithm Performance."

Tuning the *enc0_token_embedding_dimension* and *enc0_token_embedding_dimension1* hyperparameters optimizes the left and right encoder token embedding dimensions, which are the output dimensions of the encoder embedding layer. Accordingly, the *enc_dim* hyperparameter controls the dimension of the output of the embedding layer. To prevent overfitting, the *dropout* hyperparameter serves as a form of regularization in neural networks by trimming codependent neurons. Also related to overfitting and generalization, the *early_stopping_patience* hyperparameter is a threshold that controls the number of consecutive epochs allowed for training before improvement is achieved. Accordingly, *improvement* is defined by the *early_stopping_tolerance* hyperparameter, which measures a certain reduction in the loss function that the algorithm must achieve to avoid early termination (i.e., termination specified by the

number of consecutive epochs in the *early_stopping_patience* hyperparameter). The *epochs* hyperparameter then indicates how many passes the algorithm can make over the data as an upper bound (i.e., if early stopping does not occur). Finally, the number of multilayer perceptron (MLP) layers in the network is defined by the *mlp_layers* hyperparameter and tuning the learning rate for training (*which controls how much the change the model in reposed to the estimated error from the loss function each time model weights are updated*) is regulated by the *learning_rate* hyperparameter.

Results from the Bayesian hyperparameter optimization experiments on Object2Vec are presented in "Section 4.1: Bayesian Hyperparameter Optimization Results," and these results are discussed further in "Section 5.3.2: Discussing Bayesian Hyperparameter Optimization."

3.4.2 Feature Engineering

3.4.2.1 Classifiers. The feature space engineered for use with the logistic regression, random forest, and XGBoost classifiers are standard, structured tabular data that include one-hot encoded and summary statistic features. In the decomposition of data by time experiments, features are summarized over a period of 6 months; in the decomposition of data by population experiments, features are summarized over a period of 12 months. Features were not scaled because the algorithms selected are robust to unscaled feature sets, and the features are more interpretable for analysis when unscaled. While it is possible to transform the features to and from a scaled state, for the purposes of this study it was unnecessary to manage the transform (*i.e., it would require an additional processing step for little perceived benefit*). However, it would be interesting to examine the effect of scaling the feature spaces on the predictive performance of the models over time, which I have identified as future work (See Chapter 6 Conclusion).

One-hot encoding transforms categorical data into binary features. For example, in the CMS Medicare data, gender is a single field where values are represented by 'M' for male or 'F' for female.

One-hot encoding transforms this single gender variable into two binary variables: (1) ‘gender=male’, (with binary values of 0 or 1); or (2) ‘gender=female’ (with binary values of 0 or 1). The one-hot encoded features included in the study are engineered around eligibility data—e.g., gender, race, and information about Medicare enrollment such as original enrollment reason, current entitlement reason, Medicare status, and dual eligibility—and clinical variables (i.e., ICD-10-CM codes). The ICD-10-CM codes are aggregated up their three-letter prefixes so that, e.g., ICD-10-CM code *H353231* is represented as H35. This preprocessing of ICD-10-CM codes reduces granularity (and thus, variance) in the feature space so that patterns can be more easily discovered over the sparsely represented one-hot encodings.

Summary statistic features include sums and percentiles of sums of events by event type or facility (e.g., office, inpatient, home, hospice, outpatient, emergency, and skilled nursing facility events). Also included are sums, cumulative sums, averages, and maximums of (1) diagnoses, (2) procedures, (3) diagnoses affecting certain body systems; (2) chronic conditions; (3) clinical classification software categories by diagnoses and procedures; (4) substance abuse and mental health diagnoses; (5) hierarchical condition category groupings; (6) provider types (e.g., nurse practitioners, physician assistants, primary care providers, and specialists); (7) surgical flags (i.e., invasive vs. non-invasive); (8) hospital readmission; and (9) prevention quality indicators. In total, 2,851 features are included in the feature spaces for the classifiers (See Supplementary Material Appendix B. Data Dictionaries - Features for a full list of features).

3.4.2.2 Deep Learning Classifier. As described in "Section 3.4.1.2: Object2Vec," Object2Vec was designed with 2 input channels that accept and feed data to 2 encoders (i.e., the left and right encoder).

3.4.2.2.1 Left Encoder. I impose an artificial grammar on the demographic, eligibility, and utilization features in the left encoder by adding a fixed structure to the way data are represented in the input channel. First, I choose densely populated features with high variance; e.g., demographic features such as (1) age, (2) race, and (3) gender; eligibility features such as (4) original reason for entitlement and dual status; and utilization features such as (5) sums of ER events, (6) sums of IP events, (7) quartiles of ER events, (8) quartiles of IP events, (9) quartiles of total events, (10) averages of distinct providers, (11) maximums of distinct providers, (12) averages of ICD10 codes, (13) cumulative sums of distinct chronic conditions, and (14) maximums of distinct chronic conditions.

These 14 features with a total of 60 possible unique values are only a small subset of the demographic, eligibility, and utilization features used in the classifiers, as deep learning algorithms generally require much longer training times and run on more expensive hardware. Deep learning algorithms also generally require more training data in the presence of large numbers of features to avoid overfitting (in addition to the standard techniques to mitigate overfitting presented in "Section 3.4.1.2: Object2Vec"). Thus, although the number of features is fewer in comparison to the feature space used with the classifiers, it is designed with purpose. The features selected here are rich in that they are both highly populated (i.e., dense) and contain a lot of information (i.e., high variance). Moreover, the fixed structure imposed on the features, which are treated as concepts that approximate tokens in a word embeddings paradigm, acts as an artificial grammar to assist the algorithm in discovering useful patterns.

In language, a "grammar" is a common set of rules that define the structure of how people communicate. For example, people communicate using different language structures on Twitter (with its

280-character limit) than structures used in doctor's notes. The character limit on Twitter may discourage a user from following a strict, formal structure leading to shorter, more ambiguous phrasing; whereas a doctor must be extremely descriptive and precise when documenting a patient visit leading to a more formal language structure. Whatever the case, the rules of the grammar are specific to the domain in which it is applied. Without the grammar, it would be difficult to parse and interpret the semantic meanings embedded in communication within a particular domain. Just as the grammar aids the person consuming information in interpreting semantics, it also helps an algorithm to discover patterns in language. For example, word embeddings capitalize on grammar to learn representations of different tokens (i.e., words) in text where tokens have the same semantic meaning and similar representations.

To impose structure, I align the 14 features such that each feature is always in the same position, e.g., *feature 1* is always at *index 0*; *feature 2* is always at *index 1*, etc. Then, I express the value of the feature in its appropriate index as a "token" (*i.e., a numerical value that can be mapped back to the actual value of the feature*). In this way, we can think of the features as a sequence of tokens that follow a well-designed grammar, which should assist the algorithm in quantifying relationships between objects.

3.4.2.2.2 Right Encoder. The right encoder similarly treats diagnosis and procedure objects (i.e., ICD-10-CM and ICD-10-PCS codes) as sequential tokens. Each code is represented by a numerical value that can be mapped back to the actual value of the feature. However, the structure imposed in the right encoder is not artificial; rather, it is structured temporally in the order that the diagnosis or procedure object appeared in the data. In this way, the temporal nature of the diagnosis and procedure codes is preserved. We can think of this temporal sequence as a beneficiary's disease progression, or the evolution of their clinical conditions. In the experiment that decomposes the sample by time, the evolution is 6 months. In the experiment that decomposes the sample by population, the evolution is 12 months.

Just as in the feature engineering with the classifiers, the features are preprocessed to reduce granularity (and thus, variance). Here, ICD-10-CM and ICD-10-PCS codes are aggregated up their respective four-letter prefixes, which was chosen empirically again to reduce variance and assist the algorithm in finding patterns (See "Section 5.3.1.1: Object2Vec Encoders" for a discussion of this analysis). To keep maximize the efficiency of the encoders, only distinct ICD-10-CM and ICD-10-PCS codes are added to the encoder. So, if a beneficiary's first three codes in the data are *H353231*, *Z0000*, and then *H353231* again, the data is represented in the input channel as *H353231*, *Z0000*. The second instance of *H353231* (i.e., the third code in the example) is redundant and adds noise to the sequence. By removing redundant codes, the vocabulary size for the right encoder shrinks, which means less training time and lower expenses; yet the natural temporal sequence of codes is still preserved. I have identified different representations of sequences of codes as interesting future work (See Chapter 6 Conclusion).

3.4.3 Bayesian Hyperparameter Optimization

As reviewed in the introduction to "Section 3.4: Experimental Design," hyperparameter optimization is a critical step in the classification pipeline to maximize the predictive performance of a model. Traditionally, hyperparameters have been set using values found in literature, algorithm defaults,

or heuristics. However, it is now common practice to perform hyperparameter optimization, which tunes hyperparameters facilitate skillful predictions by the model. Hyperparameter optimization can be performed manually, which means that a user can try values found in the literature, using heuristics, or based on intuition, and then increase or decrease the hyperparameters to adjust its values and find an optimal setting. However, this method is crude as it is time consuming and error prone. Think of the case of Object2Vec, which has 25 hyperparameters and sometimes takes days to run. A person could spend months of full-time work to manually tune the 25 hyperparameters of an Object2Vec model, which means testing different values for each hyperparameter.

The most common automated hyperparameter tuning methods are (1) grid search, (2) random search, and (3) Bayesian hyperparameter optimization. In both random and grid search, the domain of hyperparameters examined is contained in a grid. In grid search, every possible combination of hyperparameters in the grid is evaluated. Whichever combination of hyperparameters yields the lowest error term (e.g., minimized root mean squared error) or highest predictive strength (e.g., maximized accuracy) is selected as the optimal setting. However, testing all combinations is expensive—both computationally and with respect to time. So, in random search, a random subset of combinations in the grid is tested until the number of searched iterations meets some predefined threshold based on run time or computational resources exhausted. Compared to grid search, random search is limited in that it is possible that the optimal combination of hyperparameters in the grid was never evaluated. Both random and grid search are limited in that the optimal setting must exist in the domain of the grid, but the optimal setting likely exists somewhere outside of the grid (although, ideally the grid contains hyperparameters that approximate some optimal value).

Neither grid search nor random search learn from previous iterations, meaning that if (for example) as hyperparameter values increase throughout the grid, the error term continues to increase as well, both methods will continue to exhaust combinations of higher hyperparameter values even though

it is clear higher values yield increased error (i.e., poorer performance). This is because neither method considers the history of previous evaluations of hyperparameters as evidence for evaluating more sensible values in the future. However, Bayesian hyperparameter optimization builds a probability model (i.e., a surrogate model) of the objective function and uses it to select the best hyperparameters for evaluation in the true objective function. This means that the Bayesian hyperparameter optimization algorithm can update the surrogate model using the history consisting of hyperparameter values and corresponding scores. As the history of scores and hyperparameters builds, the surrogate model of the object function improves. This method reduces the computational expense (i.e., runtime and computational resource) compared to grid search and random search and produces better hyperparameters for more skillful predictions on the test set.

3.4.4 Interpretability and Explainability

Note, that due to budget constraints (discussed in more detail in "Section 5.3.1.2: Object2Vec Encoders"), Shapley values were not produced for the object2vec models; however, Shapley values can be calculated for deep learning models, which is mentioned as future work (Ancona et al., 2019; S. Lundberg, n.d.-b; S. Lundberg & Lee, 2017). However, a thorough disaggregation analysis was performed on the object2vec models.

3.4.4.1 Feature Importances. To analyze feature importance, I report 'gain' because I want to examine the relative importance of each feature (Friedman et al., 2001). Gain is the principal reference measure of feature importance in XGBoost tree branches (Zheng et al., 2017), and the best among available methods (*e.g., including 'weight' and 'coverage'*) for comparing feature importances (Abu-Rmileh, 2019). In a single tree, the importance of each feature X_ℓ is represented by

$$w_\ell^2(T) = \sum_{t=1}^{J-1} \hat{\tau}_t^2$$

(Breiman et al., 1984). For all $J-$ 1nodes, the tree splits by feature X_ℓ at each node *t.* The feature that yields the highest estimated improvement $\hat{\tau}_t^2$ in the squared error risk is selected. Where selected as the splitting feature, the squared importance of feature X_ℓ is the sum of squared improvement over all $J-1$ nodes in the tree. This formula is then used to compute importance then over *M* additive trees

$$w_\ell^2(T) = \frac{1}{M}\sum_{m=1}^{M} \hat{\tau}_t^2(T_m)$$

. Intuitively, a feature's importance relies on the change in predictive performance when it is replaced with random noise (Zheng et al., 2017). The results from this analysis are presented in "Section 4.4.1: Feature Importances" and discussed in "Section 5.4.3: Discussing Feature Importances & Disaggregation Analysis."

3.4.4.2 Shapley Values. To quantify the contribution of each feature with respect to a given prediction, I produce Shapley values. Shapley Additive exPlanations (SHAP) is a game theoretic approach to interpreting a target model. The basic intuition is that all features in a target model are "contributors," trying to predict the task is the "game," and the "reward" is the difference between the actual prediction and the result from the explanation model. Instead of comparing a prediction to the average prediction of the entire dataset, Shapley values make it possible to contrast explanations for a subset (i.e., including a single data point). The Shapley value works for both regression and classification (i.e., in the form of probabilities). Represented as an additive feature attribution method (i.e., a linear model), SHAP specifies the explanation as:

$$g(z') = \phi_0 + \sum_{j=1}^{M} \phi_j z_j'$$

where $g$is the explanation model, $z' \in \{0,1\}^M$is the simplified features vector, *M* is the maximum size of the simplified features vector and $\phi_j \in \mathbb{R}$ is the feature attribution for a feature *j*, the Shapley values (Molnar, 2021). For this specific task (i.e., calculating Shapley values for a "Super-

Utilizer" classifier), TreeSHAP is used with XGBoost models (S. M. Lundberg et al., 2018). The results from this analysis are presented in "Section 4.4.2: Shapley Values" and discussed in "Section 5.4.3: Discussing Shapley Values."

3.4.4.3 Disaggregation Analysis. I conduct a thorough analysis of disaggregated demographic variables (Sevo & Chubin, 2010) from the Master Beneficiary Summary File, which is the same process followed by the Agency for Healthcare Research and Quality (AHRQ) (*2019 National Healthcare Quality and Disparities Report Executive Summary*, 2020). This analysis is conducted on the general population, the ground truth "Super-Utilizer" population, and the predicted "Super-Utilizer" population. The results from this analysis are presented in "Section 4.4.2: Disaggregation Analysis" and discussed in "Section 5.4.3: Discussing Feature Importances & Disaggregation Analysis."

CHAPTER 4.

RESULTS

As described in "Section 2.3: Methods & Benchmark Performance," when we are most concerned in precisely identifying the positive class in an imbalanced binary classification task, a precision-recall (PR) curve is the preferred evaluation metric. The PR curve clearly displays the number of correct (i.e., true positive) predictions. Precision, defined as $\frac{TP}{FP+TP}$, cannot be directly inferred from an area under the ROC (AUC) curve. Moreover, the PR curve also conveys recall, defined as $\frac{TP}{TP+FN}$. Note, I **do not perform** down sampling of the majority class (i.e., "not super-utilizers") in the training set, and the distribution of majority to minority samples in the training set is consistent with the test set. Thus, the true imbalance ratio is not violated when the model is deployed (Brabec & Machlica, 2018).

In addition to PR curves, I also include tabular precision, recall, and F-measure results for the top-performing models. I also present the area under the ROC (AUC) curve in some cases to demonstrate that by not considering class priors, it can be misleading. A complete set of results for each algorithm, on each dataset, across all experiments is available in Supplementary Material Appendix F Performance Tables. These results are summarized in this chapter (i.e., Chapter 4. Results), and are organized by method (e.g., Bayesian hyperparameter optimization, cost-sensitive classification, and algorithm performance).

4.1 Bayesian Hyperparameter Optimization Results

Performing Bayesian hyperparameter optimization (BHO), described in detail in "Section 3.4.3: Bayesian Hyperparameter Optimization," is an expensive and time-consuming process. I performed BHO using Amazon SageMaker (Liberty et al., 2020) and Sagify (Mtsoulis, 2020), a command-line utility to train and deploy machine learning models to SageMaker. Amazon describes SageMaker on its website as a "fully managed service that provides every developer and data scientist with the ability to

build, train, and deploy machine learning (ML) models quickly" ("Amazon SageMaker," 2020). I used cloud-based virtual machines (See Table 11 for a list of specifications), often in parallel and over the course of multiple days, to train and deploy models and generate predictions.

Table 11. Virtual Machine Specifications

p2.8xlarge	
vCPU	32
GPU	8
Mem (GiB)	61
GPU Mem (GiB)	488
Network Performance	10 Gbps
Price	$1.075
ml.p3.2xlarge	
vCPU	8
GPU	1xV100
Mem (GiB)	61
GPU Mem (GiB)	16
Network Performance	Up to 10 Gbps
Price	$4.284
ml.m5.4xlarge	
vCPU	16
GPU	-
Mem (GiB)	64
GPU Mem (GiB)	-
Network Performance	High
Price	$1.075

As described in "Section 3.4.3: Bayesian Hyperparameter Optimization," hyperparameters are often set by values found in the literature. For this reason, I have published a complete set of hyperparameters tuned for each model in Supplementary Material Appendix E. Hyperparameters.

4.2 Cost-Sensitive Classification Results

As described in "Section 3.4.1: Algorithms," I examined the benefits of cost-sensitive classification with the logistic regression, random forest, and XGBoost algorithms. The effects of cost-sensitive classification can be seen by comparing algorithm results of (1) unset class weight parameters versus setting balanced class weights for the logistic regression and random forest classifiers; or (2) the scale_pos_weight hyperparameter set to the default value of 1.0 such that the positive class and negative class are weighted equally versus using Bayesian Hyperparameter Optimization to estimate the value in the case of the XGBoost classifier.

To determine the effects of cost-sensitive classification through class weights on performance, it is more straightforward to compare tabular precision results than to compare PR curves. To develop tabular precision scores, I compare the top 2.7% of predicted probabilities of super-utilization (which matches the distribution of super-utilizers in production) across the 6-month and 12-month feature period experiments for all classifiers. It is common to discretize predictions into positive and negative classes by probability threshold (DeCaprio et al., 2020). While I could "cherry pick" a threshold that optimizes precision for presentation purposes, it is more reproducible to start with a threshold that matches the distribution of positive samples in the wild. Furthermore, I break the predictions into quintiles to report precision so that the upper quintile represents the highest probability positive predictions, and the lower quintile represents the lowest probability positive predictions.

With respect to cost-sensitive classification using class weights, using a default class weight consistently outperformed balanced class weights. Tables 12-23 show classifier performance set by precision on the test (*i.e., with and without cost-sensitive classification*). The primary purpose of these models is to precisely identify beneficiaries who will become super-utilizers in the subsequent period. Super-utilization is extremely rare in the dataset (*i.e., only 2.7% of beneficiaries will become super-utilizers*). Thus, we are interested in confidently identifying a small cohort of beneficiaries who are

likely to become super-utilizers. More specifically, the cost of a false positive is more expensive than the cost of a false negative for this task. If we can precisely predict a subset of beneficiaries who will become super-utilizers, it is possible to refer these members to care coordinators (Harris et al., 2016). Therefore, I recommend comparing precision between classifiers in the upper quintile.

In Table 12., where no class weights were used with the logistic regression classifier on the 6-month feature window, we see that 7,516 beneficiaries are predicted in the upper quintile with a precision of 47.18%; however, in Table 13. (where balanced class weights were used), precision is 46.43%.

Table 12. Logistic Regression On 6-Month Feature Window (No Class Weight)

# of Beneficiaries	Prediction Quintile	Avg. Precision	Avg. Recall	Avg. F-Measure
7,516	Upper Quintile	47.18%	4.15%	7.63%
7,515	Upper-Middle Quintile	32.61%	9.51%	14.73%
7,515	Middle Quintile	27.03%	13.24%	17.78%
7,515	Lower-Middle Quintile	23.65%	16.26%	19.27%
7,515	Lower Quintile	21.39%	18.93%	20.09%

Table 13. Logistic Regression On 6-Month Feature Window (Balanced Class Weight)

# of Beneficiaries	Prediction Quintile	Avg. Precision	Avg. Recall	Avg. F-Measure
7,516	Upper Quintile	46.43%	4.10%	7.54%
7,515	Upper-Middle Quintile	32.61%	9.53%	14.75%
7,515	Middle Quintile	27.71%	13.59%	18.24%
7,515	Lower-Middle Quintile	24.74%	17.01%	20.16%
7,515	Lower Quintile	22.51%	19.92%	21.14%

In Table 14, where no class weights were used with the random forest classifier on the 6-month feature window, we see that 7,516 beneficiaries are predicted in the upper quintile with a precision of 35.36%; however, in Table 15 (where balanced class weights were used), precision is 31.52%.

Table 14. Random Forest On 6-Month Feature Window (No Class Weight)

# of Beneficiaries	Prediction Quintile	Avg. Precision	Avg. Recall	Avg. F-Measure
7,516	Upper Quintile	35.36%	3.15%	5.78%
7,515	Upper-Middle Quintile	25.95%	7.61%	11.76%
7,515	Middle Quintile	22.84%	11.20%	15.03%
7,515	Lower-Middle Quintile	20.56%	14.15%	16.76%
7,515	Lower Quintile	18.97%	16.80%	17.82%

Table 15. Random Forest On 6-Month Feature Window (Balanced Weight)

# of Beneficiaries	Prediction Quintile	Avg. Precision	Avg. Recall	Avg. F-Measure
7,516	Upper Quintile	31.52%	2.91%	5.33%
7,515	Upper-Middle Quintile	25.16%	7.38%	11.41%
7,515	Middle Quintile	22.22%	10.91%	14.63%
7,515	Lower-Middle Quintile	20.31%	13.97%	16.55%
7,515	Lower Quintile	18.67%	16.52%	17.53%

In Table 16, where the scale_pos_weight hyperparameter of the XGBoost classifier was set to the default value of 1.0 such that the positive and negative class were weighted equally, 7,516 beneficiaries were predicted in the upper quintile with a precision of 47.54% on the 6-month feature window. However, in Table 17 where the scale_pos_weight hyperparameter was set to the value estimated by Bayesian Hyperparameter Optimization, top performance was achieved to predict the same number of beneficiaries with a precision of 54.36% on the 6-month feature window.

Table 16. XGBoost on 6-Month Feature Window (Default pos_scale_weight† = 1.0)

# of Beneficiaries	Prediction Quintile	Avg. Precision	Avg. Recall	Avg. F-Measure
7,516	Upper Quintile	47.54%	4.15%	7.64%
7,515	Upper-Middle Quintile	31.12%	9.06%	14.03%
7,515	Middle Quintile	24.93%	12.21%	16.39%
7,515	Lower-Middle Quintile	21.37%	14.69%	17.41%
7,515	Lower Quintile	19.02%	16.83%	17.86%

† The scale_pos_weight hyperparameter is set to 1.0, which means that the positive and negative class are weighted equally.

Table 17. XGBoost on 6-Month Feature Window (pos_scale_weight† = optimized value)

# of Beneficiaries	Prediction Quintile	Avg. Precision	Avg. Recall	Avg. F-Measure
7,516	Upper Quintile	54.36%	4.78%	8.79%
7,515	Upper-Middle Quintile	37.91%	11.06%	17.13%
7,515	Middle Quintile	31.46%	15.42%	20.69%
7,515	Lower-Middle Quintile	27.51%	18.92%	22.42%
7,515	Lower Quintile	24.75%	21.90%	23.24%

† The scale_pos_weight hyperparameter is set to the value estimated by Bayesian Hyperparameter Optimization.

In Table 18, where no class weights were used with the logistic regression classifier on the 12-month feature window, we see that 1,251 beneficiaries are predicted in the upper quintile with a precision of 51.46%; however, in Table 19 (where balanced class weights were used), precision is 50.24%.

Table 18. Logistic Regression On 12-Month Feature Window (No Class Weight)

# of Beneficiaries	Prediction Quintile	Avg. Precision	Avg. Recall	Avg. F-Measure
1,251	Upper Quintile	51.46%	4.70%	8.61%
1,250	Upper-Middle Quintile	36.56%	10.84%	16.72%
1,250	Middle Quintile	30.40%	15.12%	20.19%
1,250	Lower-Middle Quintile	26.59%	18.56%	21.86%
1,250	Lower Quintile	24.15%	21.70%	22.86%

Table 19. Logistic Regression On 12-Month Feature Window (Balanced Class Weight)

# of Beneficiaries	Prediction Quintile	Avg. Precision	Avg. Recall	Avg. F-Measure
1,251	Upper Quintile	50.24%	4.54%	8.33%
1,250	Upper-Middle Quintile	34.76%	10.32%	15.92%
1,250	Middle Quintile	29.95%	14.92%	19.92%
1,250	Lower-Middle Quintile	26.58%	18.55%	21.85%
1,250	Lower Quintile	24.05%	21.60%	22.76%

In Table 20, where no class weights were used with the random forest classifier on the 12-month feature window, we see that 1,251 beneficiaries are predicted in the upper quintile with a precision of 32.75%; however, in Table 21 (where balanced class weights were used), precision is 34.47%.

Table 20. Random Forest On 12-Month Feature Window (No Class Weight)

# of Beneficiaries	Prediction Quintile	Avg. Precision	Avg. Recall	Avg. F-Measure
1,251	Upper Quintile	32.75%	3.18%	5.81%
1,250	Upper-Middle Quintile	28.11%	8.37%	12.90%
1,250	Middle Quintile	25.16%	12.53%	16.73%
1,250	Lower-Middle Quintile	21.98%	15.34%	18.07%
1,250	Lower Quintile	19.85%	17.84%	18.79%

Table 21. Random Forest On 12-Month Feature Window (Balanced Weight)

# of Beneficiaries	Prediction Quintile	Avg. Precision	Avg. Recall	Avg. F-Measure
1,251	Upper Quintile	34.47%	3.19%	5.85%
1,250	Upper-Middle Quintile	26.80%	7.97%	12.29%
1,250	Middle Quintile	23.20%	11.55%	15.43%
1,250	Lower-Middle Quintile	20.88%	14.59%	17.18%
1,250	Lower Quintile	19.20%	17.25%	18.17%

In Table 22, where the scale_pos_weight hyperparameter of the XGBoost classifier was set to the default value of 1.0 such that the positive and negative class were weighted equally, top performance was achieved to predict 1,251 beneficiaries in the upper quintile with a precision of 55.94% on the 12-month feature window. In Table 23, where the scale_pos_weight hyperparameter was set value estimated by Bayesian Hyperparameter Optimization, the same number of beneficiaries was predicted with a slightly lower precision of 54.64% on the 12-month feature window.

Table 22. XGBoost on 12-Month Feature Window (Default pos_scale_weight† = 1.0)

# of Beneficiaries	Prediction Quintile	Avg. Precision	Avg. Recall	Avg. F-Measure
1,251	Upper Quintile	55.94%	5.29%	9.67%
1,250	Upper-Middle Quintile	40.70%	12.05%	18.60%
1,250	Middle Quintile	33.88%	16.86%	22.51%
1,250	Lower-Middle Quintile	29.70%	20.73%	24.42%
1,250	Lower Quintile	26.76%	24.04%	25.33%

† The scale_pos_weight hyperparameter is set to 1.0, which means that the positive and negative class are weighted equally.

Table 23. XGBoost on 12-Month Feature Window (pos_scale_weight† = optimized value)

# of Beneficiaries	Prediction Quintile	Avg. Precision	Avg. Recall	Avg. F-Measure
1,251	Upper Quintile	54.64%	5.06%	9.27%
1,250	Upper-Middle Quintile	39.27%	11.64%	17.96%
1,250	Middle Quintile	33.29%	16.57%	22.13%
1,250	Lower-Middle Quintile	29.01%	20.24%	23.85%
1,250	Lower Quintile	26.20%	23.55%	24.80%

† The scale_pos_weight hyperparameter is set to the value estimated by Bayesian Hyperparameter Optimization.

Cost-sensitive classification results are discussed further in "Section 5.3.2: Discussing Bayesian Hyperparameter Optimization & Cost-Sensitive Classification."

4.3 Deep Learning Results

I present results similar in structure to those contained in "Section 4.2: Cost-Sensitive Classification Results" for the Object2vec algorithms. On the 6-month feature window, Table 24 shows that for the Object2Vec algorithm with the Pooled Embeddings networks, the upper quintile is 37.25% precise on 7,516 beneficiaries.

Table 24. Object2Vec 6-Month Feature Window (Pooled Embeddings Networks)

# of Beneficiaries	Prediction Quintile	Avg. Precision	Avg. Recall	Avg. F-Measure
7,516	Upper Quintile	37.25%	2.35%	4.41%
7,515	Upper-Middle Quintile	28.72%	5.95%	9.86%
7,515	Middle Quintile	26.07%	9.04%	13.43%
7,515	Lower-Middle Quintile	24.20%	11.76%	15.83%
7,515	Lower Quintile	22.66%	14.16%	17.43%

Table 25 shows that for the Object2Vec algorithm with the HCNN networks on the 6-month feature window, the upper quintile is only 10.53% precise on 7,516 beneficiaries.

Table 25. Object2Vec 6-Month Feature Window (HCNN Networks)

# of Beneficiaries	Prediction Quintile	Avg. Precision	Avg. Recall	Avg. F-Measure
7,516	Upper Quintile	10.53%	0.82%	1.52%
7,515	Upper-Middle Quintile	14.31%	3.00%	4.97%
7,515	Middle Quintile	14.72%	5.11%	7.59%
7,515	Lower-Middle Quintile	14.47%	7.04%	9.47%
7,515	Lower Quintile	14.37%	8.99%	11.06%

Table 26 shows that for the Object2Vec algorithm with the BiLSTM networks on the 6-month feature window, the upper quintile is 40.54% precise on 7,516 beneficiaries, which is the highest performing Object2Vec model for this experiment.

Table 26. Object2Vec 6-Month Feature Window (BiLSTM Networks)

# of Beneficiaries	Prediction Quintile	Avg. Precision	Avg. Recall	Avg. F-Measure
7,516	Upper Quintile	40.54%	2.61%	4.91%
7,515	Upper-Middle Quintile	32.47%	6.72%	11.14%
7,515	Middle Quintile	29.02%	10.05%	14.93%
7,515	Lower-Middle Quintile	26.49%	12.87%	17.33%
7,515	Lower Quintile	24.54%	15.34%	18.88%

Table 27 shows that for the Object2Vec algorithm with the Pooled Embeddings networks on the 12-month feature window, the upper quintile is only 34.90% precise on 1,251 beneficiaries.

Table 27. Object2Vec 12-Month Feature Window (Pooled Embeddings Network)

# of Beneficiaries	Prediction Quintile	Avg. Precision	Avg. Recall	Avg. F-Measure
1,251	Upper Quintile	34.90%	3.32%	6.06%
1,250	Upper-Middle Quintile	29.19%	8.86%	13.59%
1,250	Middle Quintile	26.16%	13.26%	17.60%
1,250	Lower-Middle Quintile	23.85%	16.96%	19.82%
1,250	Lower Quintile	22.33%	20.42%	21.34%

Table 28 shows that for the Object2Vec algorithm with the HCNN networks on the 12-month feature window, the upper quintile is only 31.29% precise on 1,251 beneficiaries.

Table 28. Object2Vec 12-Month Feature Window (HCNN Network)

# of Beneficiaries	Prediction Quintile	Avg. Precision	Avg. Recall	Avg. F-Measure
1,251	Upper Quintile	31.29%	2.99%	5.45%
1,250	Upper-Middle Quintile	23.64%	7.15%	10.98%
1,250	Middle Quintile	20.73%	10.50%	13.94%
1,250	Lower-Middle Quintile	18.64%	13.24%	15.48%
1,250	Lower Quintile	17.17%	15.69%	16.40%

Table 29 shows that for the Object2Vec algorithm with the BiLSTM networks on the 12-month feature window, the upper quintile is 41.02% precise on 1,251 beneficiaries, which is the highest performing Object2Vec model for this experiment.

Table 29. Object2Vec 12-Month Feature Window (BiLSTM Network)

# of Beneficiaries	Prediction Quintile	Avg. Precision	Avg. Recall	Avg. F-Measure
1,251	Upper Quintile	41.02%	3.95%	7.21%
1,250	Upper-Middle Quintile	31.68%	9.58%	14.71%
1,250	Middle Quintile	27.73%	14.06%	18.66%
1,250	Lower-Middle Quintile	25.03%	17.77%	20.78%
1,250	Lower Quintile	23.07%	21.09%	22.04%

Cost-sensitive classification and deep learning results are explored further and summarized in "Section 4.4: Algorithm Results."

4.4 Algorithm Results

As demonstrated in "Section 4.2: Cost-Sensitive Classification Results," the XGBoost classifier using Bayesian Hyperparameter Optimization to set the pos_scale_weight hyperparameter produced the best results on the 6-month feature window test set with 54.36% precision on 7,516 beneficiaries in the upper quintile. The XGBoost classifier where no method of cost-sensitive classification was applied (i.e., where the positive and negative class were weighted equally through the default pos_scale_weight hyperparameter setting) produced the best results on the 12-month feature window test set with 55.94% precision on 1,251 beneficiaries in the upper quintile.

Of course, we can examine these results further by investigating precision within the upper quintile. The purpose of this algorithm performance evaluation is to see how the algorithm performs on predictions for which it is most it is most confident. Result tables are included so that readers can compare the performance on predictions for which the algorithm is most confident where no method of

cost-sensitive classification was applied (i.e., the results in Table 30 for the 6-month feature window experiment and Table 32 for the 12-month feature window experiment, respectively) against Bayesian Hyperparameter Optimization of the scale_pos_weight hyperparameter for cost-sensitive classification (i.e., the results in Table 31 for the 6-month feature window experiment and Table 33 for the 12-month feature window experiment, respectively). In both experiments, top performance for the most confident predictions is achieved when cost-sensitive classification is applied through Bayesian Hyperparameter Optimization of the scale_pos_weight hyperparameter (i.e., 72.97% precision in the top 20% of the upper quintile for the 6-month feature window experiment, and 66.09% precision in the top 20% of the upper quintile for the 12-month feature window experiment).

Table 30. XGBoost Upper Quintile On 6-Month Feature Window (Default pos_scale_weight† = 1.0)

# of Beneficiaries	Prediction Quintile	Avg. Precision	Avg. Recall	Avg. F-Measure
1,504	Top 20%	64.03%	1.19%	2.33%
1,503	Top 40%	51.08%	3.00%	5.66%
1,503	Top 60%	44.69%	4.39%	7.99%

† The scale_pos_weight hyperparameter is set to 1.0, which means that the positive and negative class are weighted equally.

Table 31. XGBoost Upper Quintile On 6-Month Feature Window (pos_scale_weight† = optimized value)

# of Beneficiaries	Prediction Quintile	Avg. Precision	Avg. Recall	Avg. F-Measure
1,504	Top 20%	72.97%	1.35%	2.65%
1,503	Top 40%	57.69%	3.38%	6.39%
1,503	Top 60%	50.40%	4.95%	9.01%

† The scale_pos_weight hyperparameter is set to the value estimated by Bayesian Hyperparameter Optimization.

Table 32. XGBoost Upper Quintile On 12-Month Feature Window (Default pos_scale_weight† = 1.0)

# of Beneficiaries	Prediction Quintile	Avg. Precision	Avg. Recall	Avg. F-Measure
251	Top 20%	62.10%	1.30%	2.56%
250	Top 40%	61.64%	3.70%	6.97%
250	Top 60%	56.55%	5.64%	10.26%

† The scale_pos_weight hyperparameter is set to 1.0, which means that the positive and negative class are weighted equally.

Table 33. XGBoost Upper Quintile On 12-Month Feature Window (pos_scale_weight† = optimized value)

# of Beneficiaries	Prediction Quintile	Avg. Precision	Avg. Recall	Avg. F-Measure
251	Top 20%	66.09%	1.33%	2.61%
250	Top 40%	58.67%	3.52%	6.64%
250	Top 60%	53.23%	5.32%	9.68%

† The scale_pos_weight hyperparameter is set to the value estimated by Bayesian Hyperparameter Optimization.

As demonstrated in "Section 4.3: Deep Learning Results," the Object2Vec BiLSTM algorithm achieved top performance with 41.02% precision on 1,251 beneficiaries in the upper quintile of the 6-month feature window experiment, and 40.54% precision on 7,516 beneficiaries in the upper quintile of the 12-month feature window experiment. To further examine these results, I evaluated precision within the upper quintile, which quantifies performance when the model is most confident. Table 34 shows that for the Object2Vec algorithm with the BiLSTM network on the 6-month feature window, the top 20% of the upper quintile yields 49.70% precision on 1,504 beneficiaries. Table 35 shows that for the Object2Vec algorithm with the BiLSTM network on the 12-month feature window (i.e., the top-performing Object2Vec model), the top 20% of the upper quintile yields 46.69% precision on 251 beneficiaries.

Table 34. Object2Vec 6-Month Feature Window (BiLSTM Network)

# of Beneficiaries	Prediction Quintile	Avg. Precision	Avg. Recall	Avg. F-Measure
251	Top 20%	49.70%	0.64%	1.26%
250	Top 40%	42.42%	1.76%	3.38%
250	Top 60%	38.69%	2.68%	5.02%

Table 35. Object2Vec 12-Month Feature Window (BiLSTM Network)

# of Beneficiaries	Prediction Quintile	Avg. Precision	Avg. Recall	Avg. F-Measure
251	Top 20%	46.69%	0.94%	1.85%
250	Top 40%	43.15%	2.64%	4.98%
250	Top 60%	41.03%	4.17%	7.57%

The algorithm results are discussed further in "Section 5.4.2: Discussing Algorithm Performance."

4.4 Feature Results

4.4.1 Feature Importances

As Introduced as part of the interpretability and explanation analysis in "Section 3.4.4.1: Feature Importances," I report 'gain' as the primary metric for feature importance analysis because it is the principal reference measure and most appropriate among available methods for XGBoost, which is the top-performing Super Utilizer classifier (Friedman et al., 2001), (Zheng et al., 2017), (Abu-Rmileh, 2019). Figure 7 shows the top 50 most important features of the best performing production model (i.e., XGBoost) for the data decomposition of time experiments.

Figure 7. XGBoost Top 50 Most Important Features for the Data Decomposition of Time Experiment

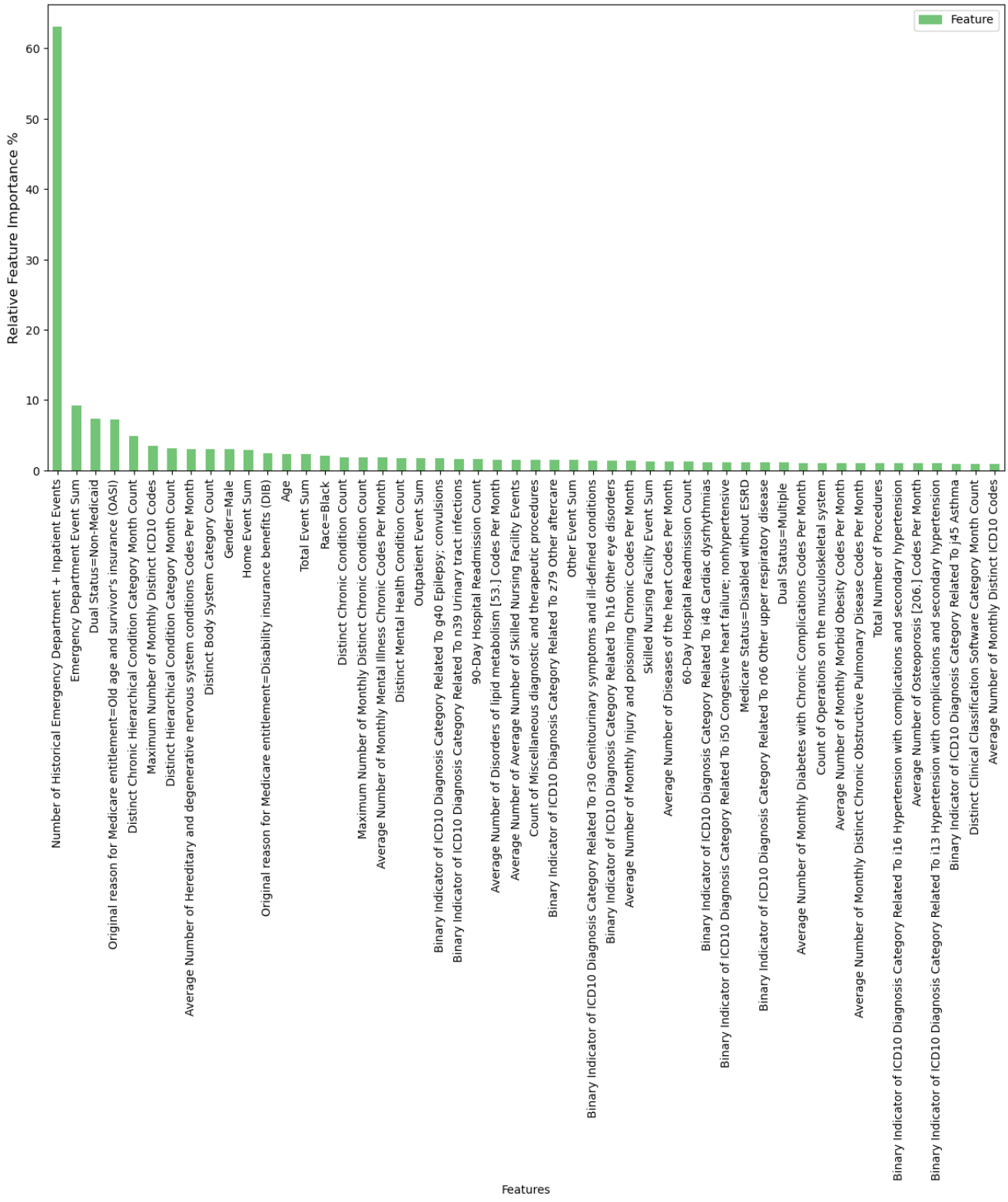
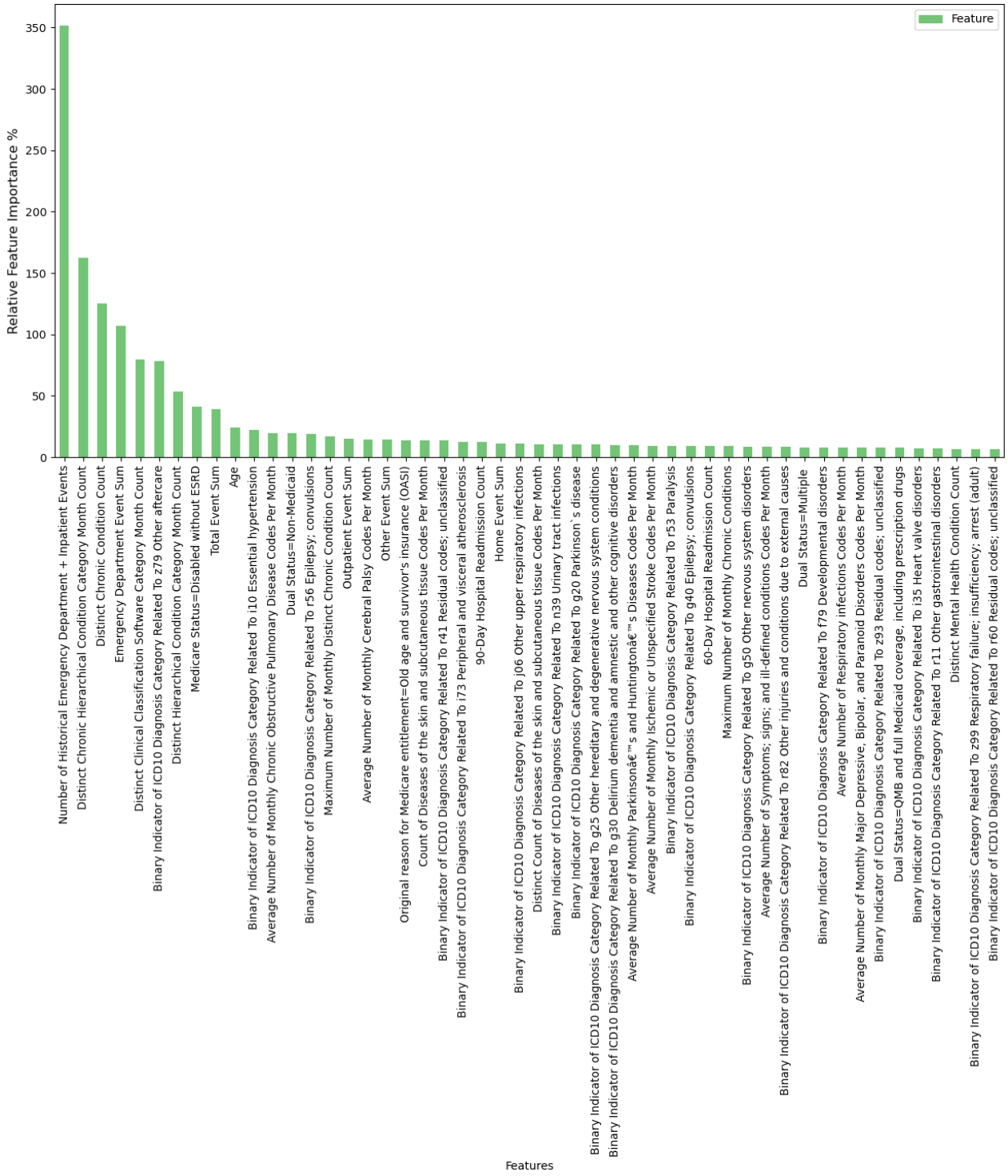


Figure 8 shows the top 50 most important features of the best performing production model (i.e., XGBoost) for the data decomposition of population experiments.

Figure 8. XGBoost Top 50 Most Important Features for the Data Decomposition of Population Experiment



To examine feature importances, I analyze the most important features for the top-performing final production models in both experiments, which is XGBoost where the data is a decomposition of time (*i.e., 54.36% precision in the upper quintile; 37.91% precision in the upper-middle quintile; 31.46% in the middle quintile; 27.51% in the lower-middle quintile; and 24.75% precision in the lower quintile*) and XGBoost where the data is a decomposition of population (*i.e., 54.64% precision in the upper quintile; 39.27% precision in the upper-middle quintile; 33.29% in the middle quintile; 29.01% in the lower-middle quintile; and 26.20% precision in the lower quintile*). Remember that the precision performance in each quintile of both experiments outperform the state-of-the-art performance of the benchmark system, which was 20% precision.

For a complete review of all feature importances of the XGBoost models for both experiments, see Tables G1 and G2 in Supplementary Material - Feature Importances. Feature importance results are discussed further in "Section 5.4.3: Discussing Feature Importances & Disaggregation Analysis."

4.4.2 Shapley Values

As described in "Section 3.4.4.2: Shapley Values," Shapley Additive exPlanations (SHAP) is an approach to interpreting a target model. Figure 9 shows the average of the SHAP value magnitudes across the dataset and is plotted as a bar chart for the data as a decomposition of time experiment. The top 10 highest average SHAP value magnitudes across the dataset include: (1) age; (2) the number of historical emergency department an inpatient visits; (3) gender is male; (4) the outpatient event sum; (5) the average number of monthly distinct ICD10 codes; (6) the total sum of events; (7) the original reason for Medicare entitle is old age and survivor's insurance; (8) dual status is non-Medicaid; (9) the emergency department event sum; and (10) the distinct count of codes in distinct body systems.

Figure 9. Bar chart of mean Shapley value importance for the Data Decomposition of Time Experiment

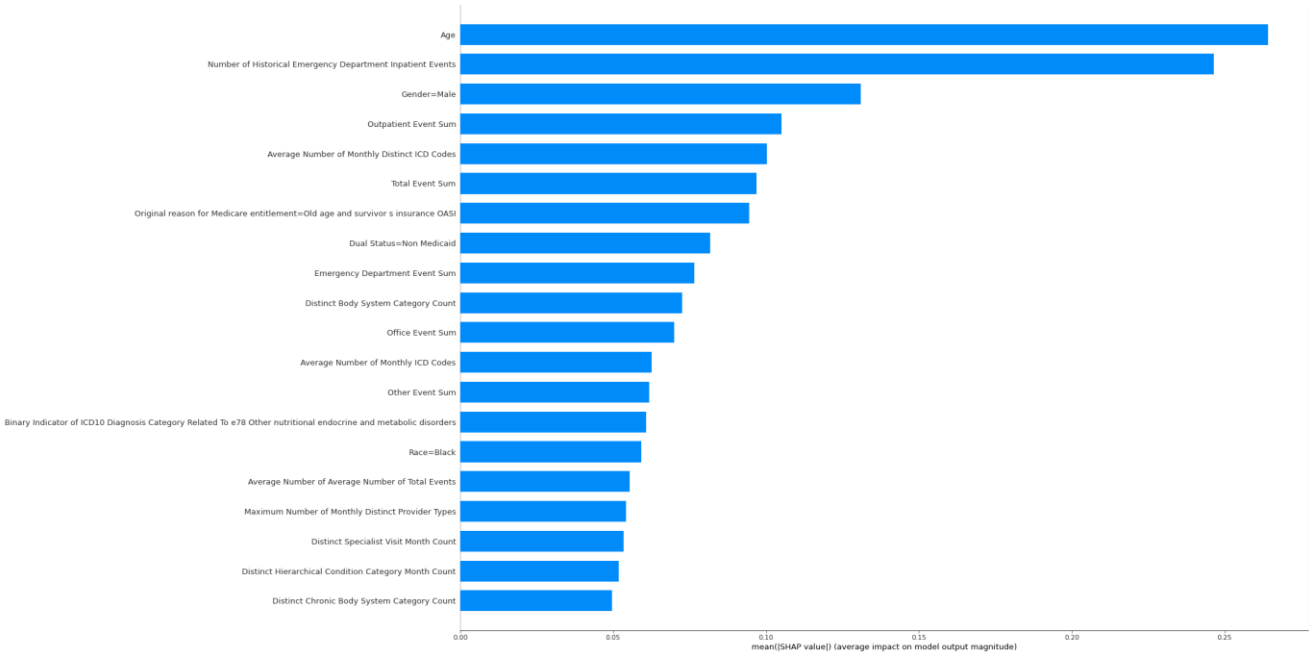


Figure 10 shows the average of the SHAP value magnitudes across the dataset for several of the most important features plotted as a bar chart for the data as a decomposition of population experiment. The top 10 highest average SHAP value magnitudes across the dataset include: (1) the number of historical emergency department an inpatient visits; (2) age; (3) the distinct chronic condition count; (4) the outpatient event sum; (5) the emergency department event sum; (6) the distinct chronic hierarchical condition category month count; (7) the original reason for Medicare entitle is old age and survivor's insurance; (8) the distinct hierarchical condition category month count; (9) the maximum number of monthly distinct chronic conditions; and (10) the total event sum.

Figure 10. Bar chart of mean Shapley value importance for the Data Decomposition of Population Experiment

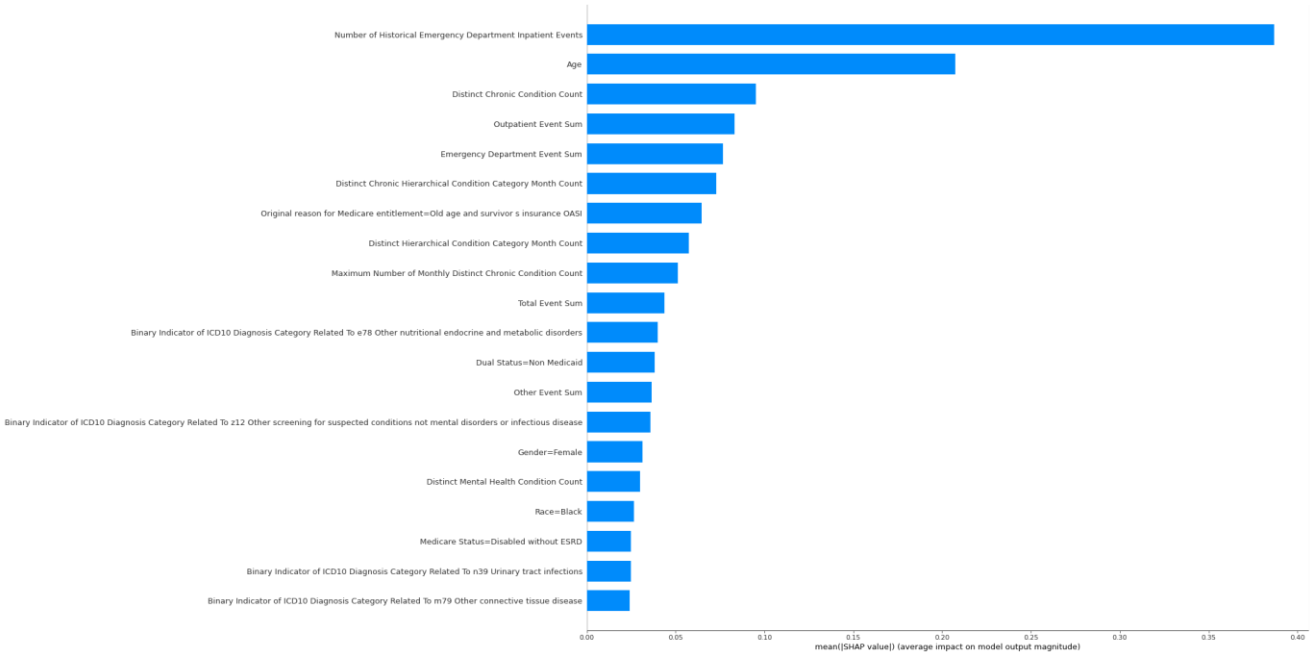


Figure 11 shows a density scatter plot of SHAP values for several of the most important features to quantify the impact each feature has on a prediction for beneficiaries in the holdout dataset for the data as a decomposition of time experiment. Features have been sorted by the sum of the SHAP value magnitudes for all beneficiaries. It is interesting to note here that age affects all predictions by a small amount, but the number of historical emergency department inpatient events affects a few predictions by a large amount.

Figure 11. SHAP Summary Plot for the Data Decomposition of Time Experiment

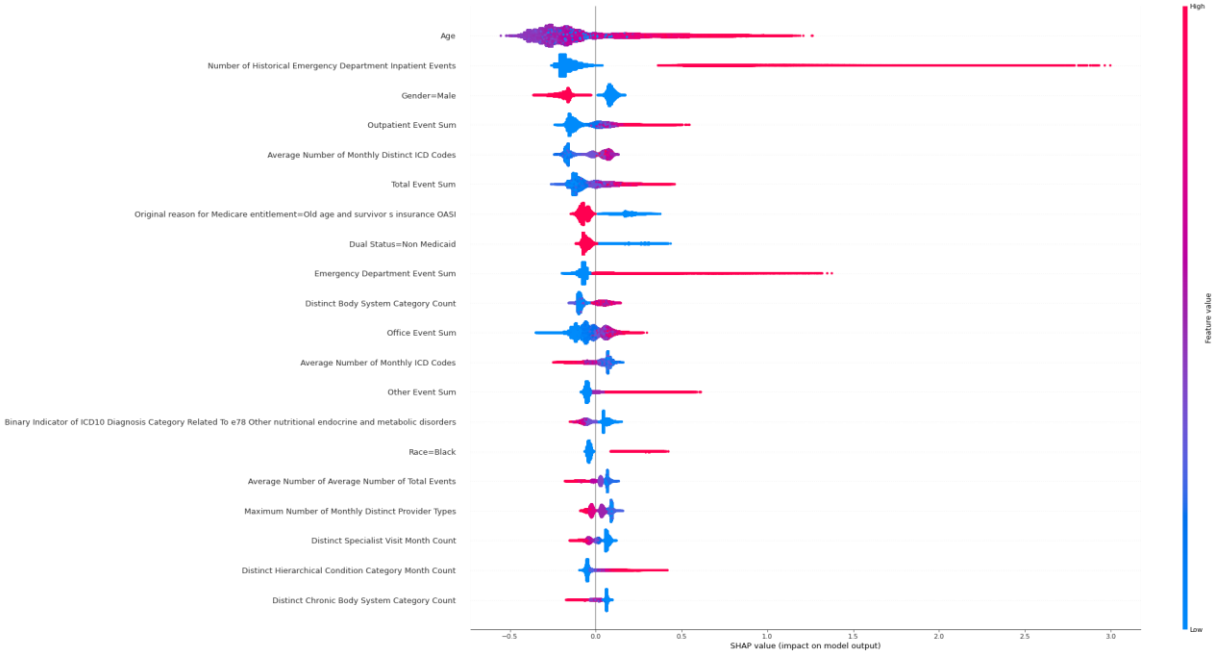


Figure 12 shows a density scatter plot of SHAP values for several of the most important features to quantify the impact each feature has on a prediction for beneficiaries in the holdout dataset for the data as a decomposition of population experiment. Age, age affects all predictions by a small amount, but the number of historical emergency department inpatient events affects a few predictions by a large amount.

Figure 12. SHAP Summary Plot for the Data Decomposition of Population Experiment

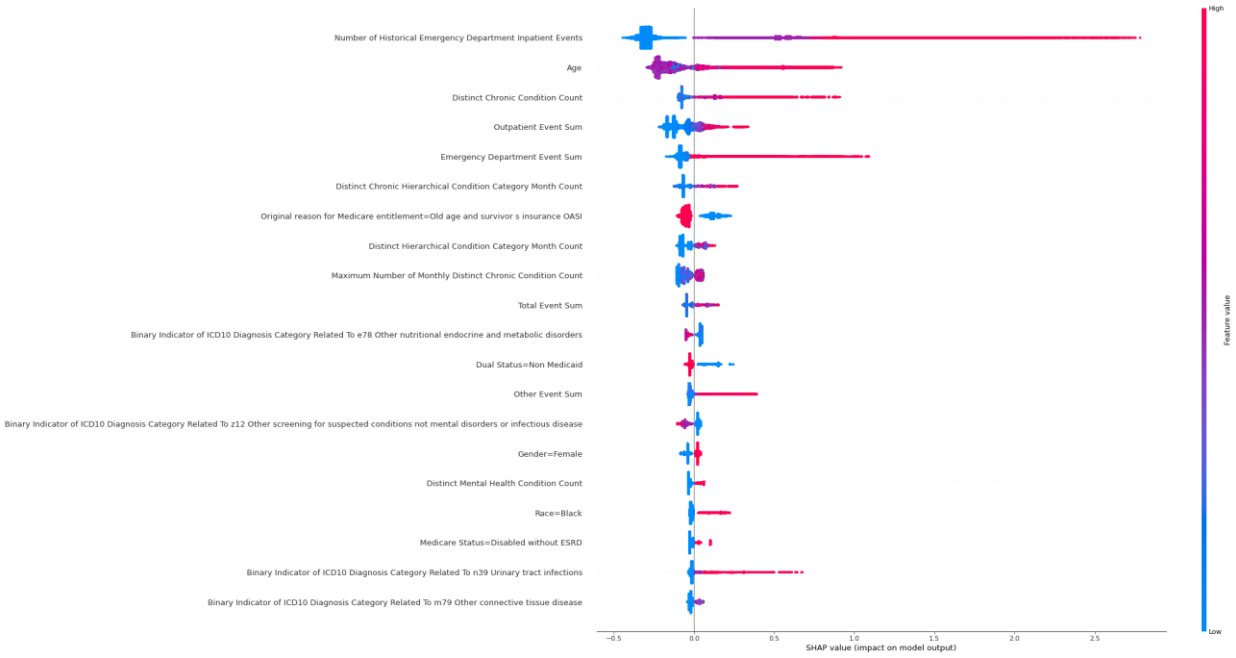


Figure 13 shows a SHAP dependence plot for the age feature in the data as a decomposition of time experiment, which demonstrates the effect of age across the whole dataset. This dependence plot shows the value of age versus its SHAP value of across numerous samples. It is analogous to a partial dependence plot; however, it accounts for interaction effects between age and other features in the dataset. Also, the SHAP dependence plot is only defined in regions of the input space where it is supported by data. Note that the vertical dispersion of SHAP values for age here are driven by interaction effects, and the number of distinct hierarchical condition category month count is chosen to highlight possible interactions (S. Lundberg, n.d.-a). Dependence plots are produced for every across both experiments (See Supplementary Material - Dependence Plots for more information).

Figure 13. SHAP Dependence Plot for Age in the Data Decomposition of Time Experiment

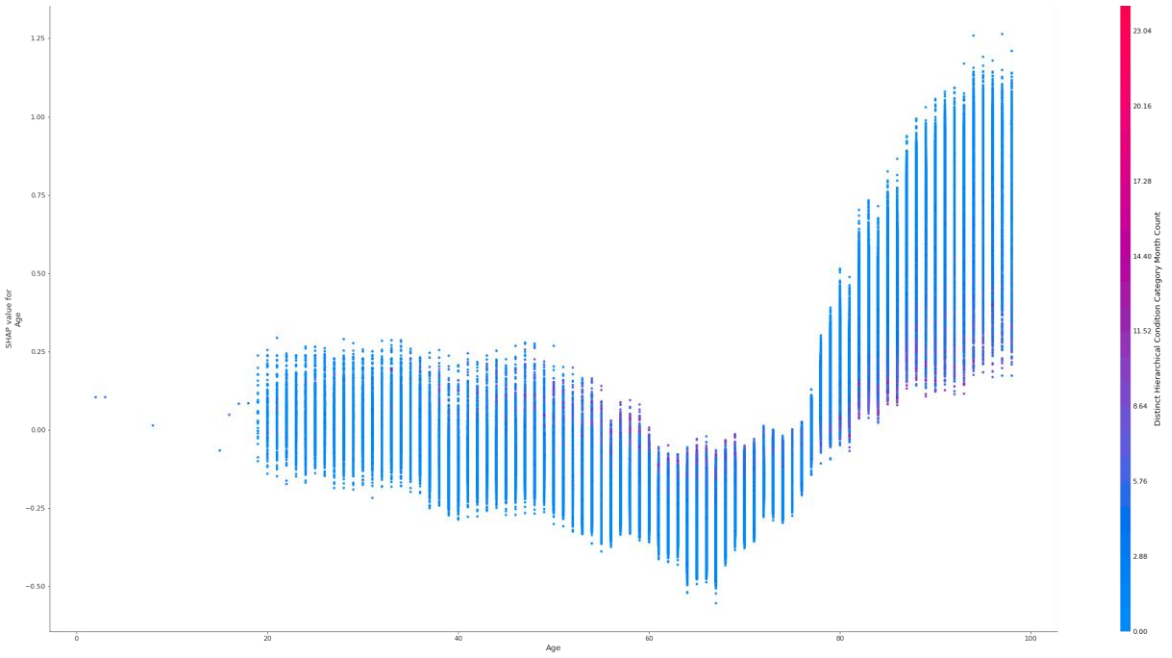


Figure 14 shows a SHAP dependence plot for the age feature in the data as a decomposition of population experiment. The distinct chronic count is chosen to highlight possible interactions, shown by the vertical dispersion of SHAP values for age.

Figure 14. SHAP Dependence Plot for Age in the Data Decomposition of Population Experiment

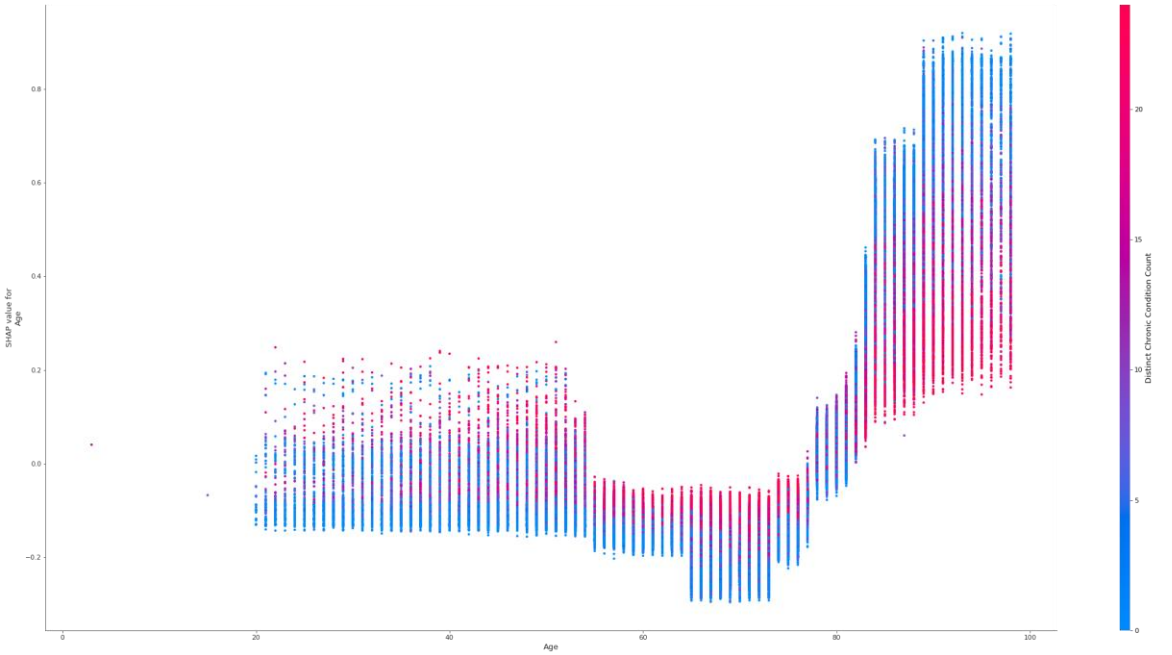


Figure 15 shows the Tree SHAP implementation integrated into XGBoost to explain the top 1,000 beneficiaries with respect to predicted probability of super-utilization for the data as a decomposition of time experiment. Note that all positively predicted beneficiaries are explained in partitions of 1,000 descending by probability of super-utilization (See Supplementary Material - Visualizations of 1,000 Predictions for the Data as a Decomposition of Time Experiment for more information).

Figure 15. SHAP Visualization Many Predictions for the Data Decomposition of Time Experiment

Figure 16 explains the top 1,000 beneficiaries with respect to predicted probability of super-utilization for the data as a decomposition of population experiment (for the exhaustive collection of plots, see Supplementary Material - Visualizations of 1,000 Predictions for the Data as a Decomposition of Population Experiment for more information).

Figure 16. SHAP Visualization Many Predictions for the Data Decomposition of Population Experiment

Figure 17 contains a force plot that explains a single prediction, which (in this case) is the beneficiary with the highest predicted probability of super-utilization for the data as a decomposition of time experiment. For this beneficiary, the historical emergency department event sum of 6.0 and the number of historical emergency department and inpatient events sum of 8.0 were the most important

reasons for the positive prediction. Note that a force plot is produced that explains every positively predicted super-utilizer (*i.e., over 38,00 plots for the data as a decomposition of time experiment, and over 7,000 plots for the data as a decomposition of population experiment*; See Supplementary Material - Visualizations of Single Predictions for the Data as a Decomposition of Time Experiment for more information).

Figure 17. SHAP Visualization of a Single Prediction for the Data Decomposition of Time Experiment

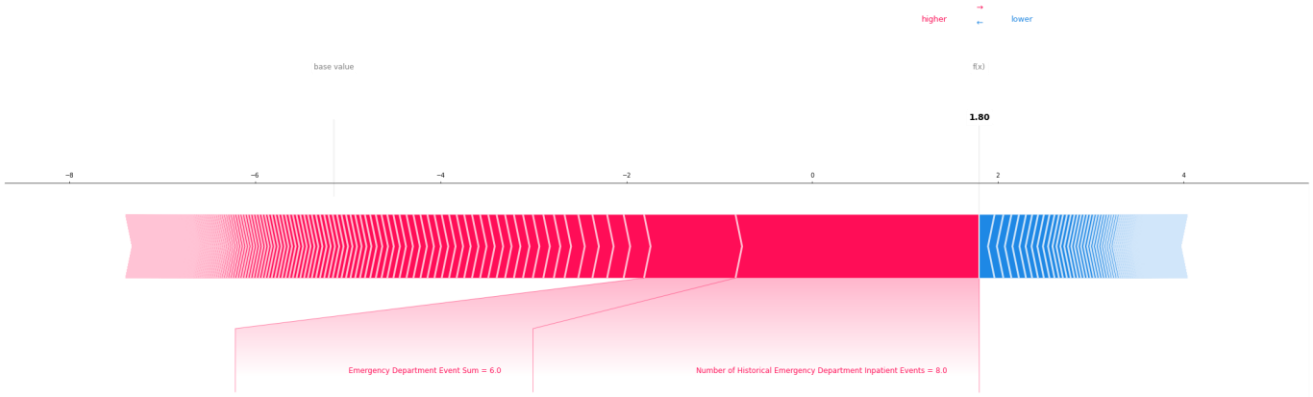
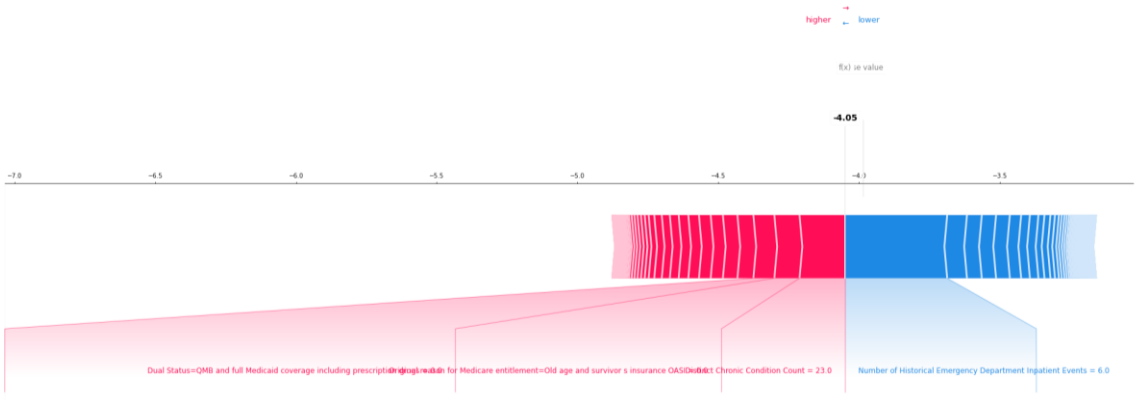


Figure 18 contains a force plot that explains the prediction for the beneficiary with the highest predicted probability of super-utilization for the data as a decomposition of population experiment. For this beneficiary, Qualified Medicare Beneficiary (QMB) program with Medicaid dual status enrollment, an old age and survivor's insurance (OASI) reason for Medicare entitlement, and a distinct chronic condition count of 23.0 were the most important reasons for the positive prediction (for the exhaustive collection of plots, see Supplementary Material - Visualizations of Single Predictions for the Data as a Decomposition of Population Experiment).

Figure 18. SHAP Visualization of a Single Prediction for the Data Decomposition of Population Experiment



Due to budget and time constraints, Shapley values were only generated for the XGBoost models. However, Shapley values can also be produced for deep learning models (Lund, n.d.). SHAP value results are discussed further in "Section 5.4.4: Discussion Shapley Values."

4.4.3 Disaggregation Analysis

To understand the effects of predictions on specific demographics, predictions are disaggregated by all relevant demographic variables and how the variables are distributed in the predicted populations is compared to how the variables are distributed in actual super-utilizers. In addition to average age, the frequency distributions of several important demographic variables are included in the tables in the disaggregation analysis, including the number of members across different **age cohorts** (i.e., Age=0-18, Age=19-44, Age=45-64, Age=65-84, and Age=85+); **gender** (i.e., Gender=Male and Gender=female); **race** (i.e., Race=Unknown, Race=White, Race=Black, Race=Other, Race=Asian, Race=Hispanic, Race=Native American, and Race=Multiple); **Medicare Enrollment Status** (i.e., Medicare Status=Aged No End Stage Renal Disease (ESRD), Medicare Status=Aged End Stage Renal Disease (ESRD), Medicare Status=Medicare Status Disabled No End Stage Renal Disease (ESRD), Medicare Status=Disabled End Stage Renal Disease (ESRD), Medicare Status=End Stage Renal Disease (ESRD), and Medicare Status=Multiple); and **Dual Eligibility for Medicare and Medicaid Status** (i.e., Dual

Status=Non-Medicaid, Dual Status=No Medicare, Dual Status=Qualified Medicare Beneficiary (QMB), Dual Status=Medicaid & Drugs, Dual Status=Specified Low-Income Medicare Beneficiary (SLMB) and full Medicaid Coverage (including prescription drugs), Dual Status=Specified Low-Income Medicare Beneficiary (SLMB), Dual Status=Qualified Disabled Working Individual (QDWI), Dual Status=Qualifying individuals (QI), Dual Status=Other Dual Medicaid & Drugs, Dual Status=Other Drug Eligibility Without Medicaid, Dual Status=Unknown, and Dual Status=Multiple).

The purpose of the tables included in the disaggregation analysis is to show any imbalance for a given variable between the overall dataset (which contains all beneficiaries, including those who do not super-utilize), the actual super-utilizer population (called "ground truth super-utilizers" in the analysis tables), the predicted super-utilizer population, and a population of super-utilizers predicted in the extreme (i.e., beneficiaries whose probabilities of super-utilization is in the 80^{th} percentile—or upper quintile—of all positively predicted super-utilizers). For example, Table 36—which compares the distributions of numerous demographic variables in actual super-utilizers to the population of super-utilizers predicted by the XGBoost algorithm for the data as a decomposition of time experiment—shows that there are 30,083 beneficiaries aged 19-44 in the overall dataset (i.e., 3.57%); 1,395 beneficiaries aged 19-44 in the actual super-utilizer population (i.e., 6.08%); 3,109 beneficiaries aged 19-44 in the predicted super-utilizer population (i.e., 8.27%); and 1,027 beneficiaries aged 19-44 predicted in the extreme (i.e., 13.66%). This means that beneficiaries aged 19-44 are nearly 2 times more likely to appear in the actual super-utilizer population as compared to the baseline (i.e., the overall dataset), and (as such) are more likely to appear in the predicted super-utilizer population (i.e., 8.27%) and the super-utilizer population predicted in the extreme (i.e., 13.66%).

Table 36. Overall vs. Ground Truth Super-Utilizer vs. Predicted XGBoost Super-Utilizer Demographic Analysis for the Data Decomposition of Time Experiment

	Overall		Ground Truth Super-Utilizers		Predicted Super-Utilizers		Predicted Upper Quintile Super-Utilizers	
Demographic Variable	**Feature Overall**	**Feature Distribution Overall**	**Feature**	**Feature Distribution Total**	**Feature**	**Feature Distribution Total**	**Feature**	**Feature Distribution Total**
Average Age	71.08		72.4		72.14		67.13	
# Age 0-18	7	0.00%	1	0.00%	4	0.01%	2	0.03%
# Age 19-44	30,083	3.57%	1,395	6.08%	3,109	8.27%	1,027	13.66%
# Age 45-64	93,622	11.10%	3,450	15.03%	7,345	19.55%	1,939	25.80%
# Age 65-84	631,900	74.91%	13,156	57.32%	16,186	43.08%	2,931	39.00%
# Age 85+	87,981	10.43%	4,948	21.56%	10,932	29.09%	1,617	21.51%
# Gender Male	355,623	42.16%	7,099	30.93%	8,875	23.62%	1,948	25.92%
# Gender Female	487,976	57.84%	15,851	69.07%	28,701	76.38%	5,568	74.08%
# Gender Multiple	6	0.00%	0	0.00%	0	0.00%	0	0.00%
# Race Unknown	15,695	1.86%	206	0.90%	308	0.82%	64	0.85%
# Race White	688,446	81.61%	17,812	77.61%	27,932	74.33%	5,250	69.85%
# Race Black	76,941	9.12%	3,254	14.18%	6,606	17.58%	1,604	21.34%
# Race Other	17,095	2.03%	321	1.40%	359	0.96%	67	0.89%
# Race Asian	20,693	2.45%	415	1.81%	553	1.47%	104	1.38%
# Race Hispanic	19,739	2.34%	762	3.32%	1,487	3.96%	351	4.67%
# Race Native American	4,984	0.59%	180	0.78%	331	0.88%	76	1.01%
# Race Multiple	0	0.00%	0	0.00%	0	0.00%	0	0.00%
# Medicare Status Aged No End Stage Renal Disease (ESRD)	729,404	86.46%	18,294	79.71%	27,382	72.87%	4,597	61.16%
# Medicare Status Aged End Stage Renal Disease (ESRD)	55	0.01%	2	0.01%	5	0.01%	2	0.03%
# Medicare Status Disabled No End Stage Renal Disease (ESRD)	116,468	13.81%	4,745	20.68%	10,345	27.53%	2,944	39.17%
# Medicare Status Disabled End Stage Renal Disease (ESRD)	63	0.01%	5	0.02%	9	0.02%	0	0.00%
# Medicare Status End Stage Renal Disease (ESRD)	77	0.01%	2	0.01%	18	0.05%	4	0.05%

# Medicare Status Multiple	2,500	0.30%	98	0.43%	183	0.49%	31	0.41%
# Dual Status Non-Medicaid	709,624	84.12%	16,140	70.33%	21,947	58.41%	3,526	46.91%
# Dual Status No Medicare	0	0.00%	0	0.00%	0	0.00%	0	0.00%
# Dual Status Qualified Medicare Beneficiary (QMB)	20,985	2.49%	960	4.18%	2,090	5.56%	525	6.99%
# Dual Status Medicaid & Drugs	81,627	9.68%	4,242	18.48%	9,810	26.11%	2,576	34.27%
# Dual Status Specified Low-Income Medicare Beneficiary (SLMB)	12,239	1.45%	591	2.58%	1,114	2.96%	254	3.38%
# Dual Status Specified Low-Income Medicare Beneficiary (SLMB) and full Medicaid Coverage (including prescription drugs)	4,621	0.55%	316	1.38%	787	2.09%	212	2.82%
# Dual Status Qualified Disabled Working Individual (QDWI)	458	0.05%	27	0.12%	56	0.15%	16	0.21%
# Dual Status Qualifying individuals (QI)	6,303	0.75%	263	1.15%	557	1.48%	125	1.66%
# Dual Status Other Dual Medicaid & Drugs	24,622	2.92%	1,458	6.35%	3,632	9.67%	908	12.08%
# Dual Status Other Drug Eligibility Without Medicaid	193	0.02%	7	0.03%	27	0.07%	6	0.08%
# Dual Status Unknown	350	0.04%	11	0.05%	30	0.08%	4	0.05%
# Dual Status Multiple	15,967	1.89%	967	4.21%	2,262	6.02%	583	7.76%

The data in Table 37 shows any imbalance for the included demographic variables between the overall dataset, the actual super-utilizer population, the predicted super-utilizer population, and a population of super-utilizers predicted in the extreme. Table 37 compares the distributions of demographic variables in actual super-utilizers to the population of super-utilizers predicted by the Object2Vec BiLSTM algorithm for the data as a decomposition of time experiment.

Table 37. Overall vs. Ground Truth Super-Utilizer vs. Predicted o2v BiLSTM Super-Utilizer Demographic Analysis for the Data Decomposition of Time Experiment

	Overall		Ground Truth Super-Utilizers		Predicted Super-Utilizers		Predicted Upper Quintile Super-Utilizers	
Demographic Variable	Feature Overall	Feature Distribution Overall	Feature	Feature Distribution Total	Feature	Feature Distribution Total	Feature	Feature Distribution Total
Average Age	71.08		72.4		70.87		70.75	
# Age 0-18	7	0.00%	1	0.00%	2	0.01%	2	0.03%
# Age 19-44	30,083	3.57%	1,395	6.08%	1,318	3.51%	269	3.58%
# Age 45-64	93,622	11.10%	3,450	15.03%	4,358	11.60%	839	11.16%
# Age 65-84	631,900	74.91%	13,156	57.32%	27,008	71.88%	5,446	72.47%
# Age 85+	87,981	10.43%	4,948	21.56%	4,889	13.01%	959	12.76%
# Gender Male	355,623	42.16%	7,099	30.93%	15,901	42.32%	3,173	42.22%
# Gender Female	487,976	57.84%	15,851	69.07%	21,674	57.68%	4,342	57.78%
# Gender Multiple	6	0.00%	0	0.00%	0	0.00%	0	0.00%
# Race Unknown	15,695	1.86%	206	0.90%	877	2.33%	192	2.55%
# Race White	688,446	81.61%	17,812	77.61%	30,473	81.10%	6,127	81.53%
# Race Black	76,941	9.12%	3,254	14.18%	3,478	9.26%	638	8.49%
# Race Other	17,095	2.03%	321	1.40%	664	1.77%	130	1.73%
# Race Asian	20,693	2.45%	415	1.81%	959	2.55%	196	2.61%
# Race Hispanic	19,739	2.34%	762	3.32%	881	2.34%	183	2.44%
# Race Native American	4,984	0.59%	180	0.78%	243	0.65%	49	0.65%
# Race Multiple	0	0.00%	0	0.00%	0	0.00%	0	0.00%
# Medicare Status Aged No End Stage Renal Disease (ESRD)	729,404	86.46%	18,294	79.71%	32,417	86.27%	6,510	86.63%
# Medicare Status Aged End Stage Renal Disease (ESRD)	55	0.01%	2	0.01%	2	0.01%	0	0.00%
# Medicare Status Disabled No End Stage Renal Disease (ESRD)	116,468	13.81%	4,745	20.68%	5,269	14.02%	1,029	13.69%
# Medicare Status Disabled End Stage Renal Disease (ESRD)	63	0.01%	5	0.02%	4	0.01%	2	0.03%
# Medicare Status End Stage Renal Disease (ESRD)	77	0.01%	2	0.01%	4	0.01%	3	0.04%
# Medicare Status Multiple	2,500	0.30%	98	0.43%	123	0.33%	29	0.39%

# Dual Status Non-Medicaid	709,624	84.12%	16,140	70.33%	31,448	83.69%	6,327	84.19%
# Dual Status No Medicare	0	0.00%	0	0.00%	0	0.00%	0	0.00%
# Dual Status Qualified Medicare Beneficiary (QMB)	20,985	2.49%	960	4.18%	976	2.60%	176	2.34%
# Dual Status Medicaid & Drugs	81,627	9.68%	4,242	18.48%	3,625	9.65%	727	9.67%
# Dual Status Specified Low-Income Medicare Beneficiary (SLMB)	12,239	1.45%	591	2.58%	576	1.53%	115	1.53%
# Dual Status Specified Low-Income Medicare Beneficiary (SLMB) and full Medicaid Coverage (including prescription drugs	4,621	0.55%	316	1.38%	216	0.57%	48	0.64%
# Dual Status Qualified Disabled Working Individual (QDWI)	458	0.05%	27	0.12%	21	0.06%	4	0.05%
# Dual Status Qualifying individuals (QI)	6,303	0.75%	263	1.15%	263	0.70%	37	0.49%
# Dual Status Other Dual Medicaid & Drugs	24,622	2.92%	1,458	6.35%	1,252	3.33%	230	3.06%
# Dual Status Other Drug Eligibility Without Medicaid	193	0.02%	7	0.03%	7	0.02%	3	0.04%
# Dual Status Unknown	350	0.04%	11	0.05%	10	0.03%	5	0.07%
# Dual Status Multiple	15,967	1.89%	967	4.21%	766	2.04%	142	1.89%

The data in Table 38 shows any imbalance for the included demographic variables between the overall dataset, the actual super-utilizer population, the predicted super-utilizer population, and a population of super-utilizers predicted in the extreme. Table 38 compares the distributions of demographic variables in actual super-utilizers to the population of super-utilizers predicted by the XGBoost algorithm for the data as a decomposition of population experiment.

Table 38. Overall vs. Ground Truth Super-Utilizer vs. Predicted XGBoost Super-Utilizer Demographic Analysis for the Data Decomposition of Population Experiment

	Overall		Ground Truth Super-Utilizers		Predicted Super-Utilizers		Predicted Upper Quintile Super-Utilizers	
Demographic Variable	**Feature Overall**	**Feature Distribution Overall**	**Feature**	**Feature Distribution Total**	**Feature**	**Feature Distribution Total**	**Feature**	**Feature Distribution Total**
Average Age	71.08		72.4		71.82		63.68	
# Age 0-18	7	0.00%	1	0.00%	2	0.03%	2	0.16%
# Age 19-44	30,083	3.57%	1,395	6.08%	510	8.16%	238	19.02%
# Age 45-64	93,622	11.10%	3,450	15.03%	1,250	20.00%	355	28.38%
# Age 65-84	631,900	74.91%	13,156	57.32%	2,861	45.77%	442	35.33%
# Age 85+	87,981	10.43%	4,948	21.56%	1,628	26.04%	214	17.11%
# Gender Male	355,623	42.16%	7,099	30.93%	1,523	24.36%	333	26.62%
# Gender Female	487,976	57.84%	15,851	69.07%	4,728	75.64%	918	73.38%
# Gender Multiple	6	0.00%	0	0.00%	0	0.00%	0	0.00%
# Race Unknown	15,695	1.86%	206	0.90%	40	0.64%	13	1.04%
# Race White	688,446	81.61%	17,812	77.61%	4,765	76.23%	866	69.22%
# Race Black	76,941	9.12%	3,254	14.18%	993	15.89%	268	21.42%
# Race Other	17,095	2.03%	321	1.40%	64	1.02%	8	0.64%
# Race Asian	20,693	2.45%	415	1.81%	104	1.66%	20	1.60%
# Race Hispanic	19,739	2.34%	762	3.32%	237	3.79%	61	4.88%
# Race Native American	4,984	0.59%	180	0.78%	48	0.77%	15	1.20%
# Race Multiple	0	0.00%	0	0.00%	0	0.00%	0	0.00%
# Medicare Status Aged No End Stage Renal Disease (ESRD)	729,404	86.46%	18,294	79.71%	4,503	72.04%	657	52.52%
# Medicare Status Aged End Stage Renal Disease (ESRD)	55	0.01%	2	0.01%	1	0.02%	0	0.00%
# Medicare Status Disabled No End Stage Renal Disease (ESRD)	116,468	13.81%	4,745	20.68%	1,811	28.97%	601	48.04%
# Medicare Status Disabled End Stage Renal Disease (ESRD)	63	0.01%	5	0.02%	1	0.02%	0	0.00%
# Medicare Status End Stage Renal Disease (ESRD)	77	0.01%	2	0.01%	4	0.06%	3	0.24%
# Medicare Status Multiple	2,500	0.30%	98	0.43%	69	1.10%	10	0.80%

# Dual Status Non-Medicaid	709,624	84.12%	16,140	70.33%	3,882	62.10%	605	48.36%
# Dual Status No Medicare	0	0.00%	0	0.00%	0	0.00%	0	0.00%
# Dual Status Qualified Medicare Beneficiary (QMB)	20,985	2.49%	960	4.18%	341	5.46%	95	7.59%
# Dual Status Medicaid & Drugs	81,627	9.68%	4,242	18.48%	1,633	26.12%	455	36.37%
# Dual Status Specified Low-Income Medicare Beneficiary (SLMB)	12,239	1.45%	591	2.58%	195	3.12%	55	4.40%
# Dual Status Specified Low-Income Medicare Beneficiary (SLMB) and full Medicaid Coverage (including prescription drugs	4,621	0.55%	316	1.38%	144	2.30%	44	3.52%
# Dual Status Qualified Disabled Working Individual (QDWI)	458	0.05%	27	0.12%	13	0.21%	4	0.32%
# Dual Status Qualifying individuals (QI)	6,303	0.75%	263	1.15%	125	2.00%	35	2.80%
# Dual Status Other Dual Medicaid & Drugs	24,622	2.92%	1,458	6.35%	690	11.04%	172	13.75%
# Dual Status Other Drug Eligibility Without Medicaid	193	0.02%	7	0.03%	6	0.10%	3	0.24%
# Dual Status Unknown	350	0.04%	11	0.05%	10	0.16%	2	0.16%
# Dual Status Multiple	15,967	1.89%	967	4.21%	668	10.69%	184	14.71%

The data in Table 39 shows any imbalance for the included demographic variables between the overall dataset, the actual super-utilizer population, the predicted super-utilizer population, and a population of super-utilizers predicted in the extreme. Table 39 compares the distributions of demographic variables in actual super-utilizers to the population of super-utilizers predicted by the Object2Vec algorithm for the data as a decomposition of population experiment.

Table 39. Overall vs. Ground Truth Super-Utilizer vs. Predicted o2v BiLSTM Super-Utilizer Demographic Analysis for the Data Decomposition of Population Experiment

	Overall		Ground Truth Super-Utilizers		Predicted Super-Utilizers		Predicted Upper Quintile Super-Utilizers	
Demographic Variable	Feature Overall	Feature Distribution Overall	Feature	Feature Distribution Total	Feature	Feature Distribution Total	Feature	Feature Distribution Total
Average Age	71.08		72.4		78.39		78.94	
# Age 0-18	7	0.00%	1	0.00%	0	0.00%	0	0.00%
# Age 19-44	30,083	3.57%	1,395	6.08%	132	2.11%	16	1.28%
# Age 45-64	93,622	11.10%	3,450	15.03%	589	9.42%	111	8.87%
# Age 65-84	631,900	74.91%	13,156	57.32%	4,158	66.52%	848	67.79%
# Age 85+	87,981	10.43%	4,948	21.56%	1,372	21.95%	276	22.06%
# Gender Male	355,623	42.16%	7,099	30.93%	2,459	39.34%	490	39.17%
# Gender Female	487,976	57.84%	15,851	69.07%	3,793	60.68%	762	60.91%
# Gender Multiple	6	0.00%	0	0.00%	1	0.02%	1	0.08%
# Race Unknown	15,695	1.86%	206	0.90%	4	0.06%	0	0.00%
# Race White	688,446	81.61%	17,812	77.61%	5,228	83.63%	1,022	81.69%
# Race Black	76,941	9.12%	3,254	14.18%	455	7.28%	102	8.15%
# Race Other	17,095	2.03%	321	1.40%	158	2.53%	39	3.12%
# Race Asian	20,693	2.45%	415	1.81%	177	2.83%	33	2.64%
# Race Hispanic	19,739	2.34%	762	3.32%	185	2.96%	49	3.92%
# Race Native American	4,984	0.59%	180	0.78%	44	0.70%	6	0.48%
# Race Multiple	0	0.00%	0	0.00%	0	0.00%	0	0.00%
# Medicare Status Aged No End Stage Renal Disease (ESRD)	729,404	86.46%	18,294	79.71%	5,549	88.77%	1,130	90.33%
# Medicare Status Aged End Stage Renal Disease (ESRD)	55	0.01%	2	0.01%	0	0.00%	0	0.00%
# Medicare Status Disabled No End Stage Renal Disease (ESRD)	116,468	13.81%	4,745	20.68%	722	11.55%	128	10.23%
# Medicare Status Disabled End Stage Renal Disease (ESRD)	63	0.01%	5	0.02%	0	0.00%	0	0.00%
# Medicare Status End Stage Renal Disease (ESRD)	77	0.01%	2	0.01%	0	0.00%	0	0.00%

# Medicare Status Multiple	2,500	0.30%	98	0.43%	20	0.32%	7	0.56%
# Dual Status Non-Medicaid	709,624	84.12%	16,140	70.33%	5,001	80.00%	1,001	80.02%
# Dual Status No Medicare	0	0.00%	0	0.00%	0	0.00%	0	0.00%
# Dual Status Qualified Medicare Beneficiary (QMB)	20,985	2.49%	960	4.18%	145	2.32%	26	2.08%
# Dual Status Medicaid & Drugs	81,627	9.68%	4,242	18.48%	797	12.75%	163	13.03%
# Dual Status Specified Low-Income Medicare Beneficiary (SLMB)	12,239	1.45%	591	2.58%	83	1.33%	18	1.44%
# Dual Status Specified Low-Income Medicare Beneficiary (SLMB) and full Medicaid Coverage (including prescription drugs	4,621	0.55%	316	1.38%	65	1.04%	10	0.80%
# Dual Status Qualified Disabled Working Individual (QDWI)	458	0.05%	27	0.12%	6	0.10%	0	0.00%
# Dual Status Qualifying individuals (QI)	6,303	0.75%	263	1.15%	55	0.88%	9	0.72%
# Dual Status Other Dual Medicaid & Drugs	24,622	2.92%	1,458	6.35%	235	3.76%	47	3.76%
# Dual Status Other Drug Eligibility Without Medicaid	193	0.02%	7	0.03%	1	0.02%	1	0.08%
# Dual Status Unknown	350	0.04%	11	0.05%	4	0.06%	1	0.08%
# Dual Status Multiple	15,967	1.89%	967	4.21%	129	2.06%	23	1.84%

The results from this disaggregation analysis (including imbalances discovered between the overall dataset, the actual super-utilizer population, the predicted super-utilizer population, and the super-utilizer population predicted in the extreme) are discussed further in “Section 5.4.3: Discussing Feature Importances & Disaggregation Analysis.” For an exhaustive review of all disaggregation analyses, see (See Supplementary Material - Disaggregation Analysis for more information).

CHAPTER 5.

DISCUSSION

5.1 Discussing Dataset Design

To maximize the data, I was granted for research, I broke the experimental design of my datasets into two separate streams: (1) datasets as a decomposition of time; and (2) datasets as a decomposition of population. To ensure a proper class balance and sample representativeness across datasets, stratified random sampling was applied (Acharya et al., 2013). It is worthwhile to note that many of the limitations I face in my dataset design come from the fact that I have only 18 months of data on which to train and evaluate models. Healthcare researchers and practitioners generally have access to several years of historical data, which of course mitigates these limitations. Nonetheless, it is indeed possible to create useful models even with only 18 months of retrospective data, especially given that I have access to a considerably large, randomized dataset.

5.1.1 Discussing Datasets as a Decomposition of Time

When the datasets are decomposed by time, each dataset (*i.e., those used for cross-validation, training, and evaluation*) is broken into a 6-month feature period and a 6-month label period. Decomposing the data by time is good practice as I can test models trained during the model development period on new (future) data, which are withheld from the models until production for evaluation. Although both the model development and test sets contain data for the same beneficiaries, I am unconcerned about information leakage because ER and IP super-utilization in one period (i.e., the model development period) does not guarantee ER and IP super-utilization in the next period (i.e., the label period), which is evident by the member overlap analysis of the dependent variable in Table 34.

Table 34. Member Overlap Analysis for Datasets Decomposed by Time

# of Beneficiaries	Train Label	Test Label
721,595	Not Super-Utilizer	Not Super-Utilizer
16,704	Not Super-Utilizer	Super-Utilizer
16,001	Super-Utilizer	Not Super-Utilizer
4,048	Super-Utilizer	Super-Utilizer

Table 34 shows the training and test label distributions for the 6-month feature window with 6-month label period experiments. This is a useful analysis, and it shows longitudinal behaviors of ER and IP utilization for beneficiaries in two contiguous 6-month periods (See Table 5 and Table 6, which show the design that decomposes data by time). The first label period (i.e., the training set label period) contains beneficiaries who had at least 2 ER or 2 IP visits between 10/01/2017 and 03/31/2018; the second label period (i.e., the test set label period) contains beneficiaries who had at least 2 ER or 2 IP visits between 04/01/2017 and 09/30/2018.

Table 34 contains only beneficiaries who exist in both the training and test sets. It shows that only 16,704 of 738,299 (i.e., 2.26%) beneficiaries had at least 2 ER or 2 IP visits in label period 1 (i.e., the training period), and of these 16,704 beneficiaries, only 4,048 (i.e., 24.23%) also had at least 2 ER or 2 IP visits in label period 2 (i.e., the test period). This means that the overwhelming majority (i.e., 75.77%) of beneficiaries were not super-utilizers in the prior label period. Note that we cannot repeat this analysis for the 12-month feature window and 6-month label period, as this data is decomposed by population, which prevents labels from longitudinal analysis.

Decomposing the CMS Medicare dataset by time presents several limitations. First, I have only six months of historical data from which a model can mine patterns indicative of future healthcare utilization behaviors, and generally 12 months of historical data is preferred. In this way, I am restricted to mining data immediately leading up to ER and IP super-utilization, and stronger signals of super-utilization occur over longer historical durations.

It is important to note that there typically exists a 90-day adjustment period for medical claims, which means that if a user decides to wait for adjustments (which my models are trained on), then it can take an additional three months to wait for properly adjusted data to generate predictions. Thus, a model user has three options: (1) run the model on unadjusted data after collecting six months of data; (2) gather three months of data over six months (i.e., collect three months of data and wait an additional three months to be sure that the data is correct); or, (3) gather six months of history over nine months (i.e., collect six months of data and wait an additional three months to be sure that the data is correct). Each of these three options has important limitations which I elaborate on in the discussion section.

The first option (i.e., running the model on unadjusted claims) is perhaps the least desirable of the three options available to a user when the data are decomposed into sets with 6-month feature periods and 6-month label periods. Although the user can run the model as intended, many important features on which the model was trained are likely missing or inaccurate. Typically, the most complex claims (e.g., inpatient events) take the longest to adjudicate. The simple nature of these complex claims indicate that the information contained therein is both highly predictive and concentrated in an important population. Therefore, when the user runs the model, it is fair to say that some of the most important information is missing or incorrect for some of the most important people with respect to the task of predicting super-utilizers of ER and IP services. Further complicating this problem is that it is difficult to reliably impute this missing information (*let alone identify the correct beneficiaries who need imputation*).

The adjudication process will have different effects depending on the design of the dependent variable. For example, the implications of adjudication are likely more significant when developing a cost-based model than a utilization-based model. This is because it is (perhaps) easier to identify that some event has occurred (even if the details are unclear) than it is to understand how much the event will eventually cost the beneficiary or payer. However, I cannot test this hypothesis because the

retrospective CMS data that I have access to has already been completely adjudicated. I have identified this as an interesting future research direction, although it is only tangentially related to the task at hand.

Through the second option (i.e., gathering six months of history over nine months), the user runs a model on six full months of historical data, but the user must wait an additional three months for the data to be adjudicated for accuracy. This is undesirable because the model was designed to generate a prediction that projects six months into the future based on six months of historical data. If the user waits nine months to generate a prediction, then the prediction is three months old already by the time a user can act on the prediction, which means that the user is three months late to intervene on a phenomenon set to occur in the subsequent six months. One of the primary aims of my research is to identify populations at risk with time for intervention, but this option undermines the capacity for prompt intervention. Table 35 shows that after collecting data for 6 months and waiting 3 months for the data to be adjudicated, we can only start to intervene on super-utilizer predictions in month 4, which is near the end of the prediction period (i.e., 3 months into the 6-month period).

Table 35. 6-Month Data Collection Over 9 Months with Adjudication

Feature Period						**Super-Utilizer Prediction Period**					
Month 1	Month 2	Month 3	Month 4	Month 5	Month 6	AP† 1 Month 1	AP 2 Month 2	AP 3 Month 3	Month 4	Month 5	Month 6

† **AP = Adjudication Period**

The final option (i.e., gathering three months of data over six months) means that the user must run a model on partial historical data (i.e., *only three months of data*), but at least the data has been adjudicated (i.e., *it has converged to maximum accuracy*). This is undesirable because even though the data is much more reliable, it is difficult for the model to find patterns with only three months of longitudinal history. Compared to 180 days for example, 90 days is a short duration and provides a mere

snapshot of the beneficiary's activity. Table 36 shows that we can only collect for 3 months to allow for proper adjudication in the 6-month feature window. While this allows for a full 6-month super-utilizer prediction window (which is ideal), 3 months is not enough longitudinal history to reliably predict the 6-month outcome.

Table 36. 3-Month Data Collection Over 6 Months with Adjudication

Feature Period			**Adjudication Period**			**Super-Utilizer Prediction Period**					
Month 1	Month 2	Month 3	Month 1	Month 2	Month 3	Month 1	Month 2	Month 3	Month 4	Month 5	Month 6

† **AP = Adjudication Period**

In the context of using a model trained on only 6 months of historical data to predict a 6-month future outcome, I recommend gathering six months of history over nine months because the underlying features are accurate and *there is still time to act*, even if not much. However, given more historical data (e.g., 12 months), I recommend option three because it is possible to run the model on nine months of reliable features with an additional three months of data that has not been fully adjudicated.

5.1.2 Discussing Datasets as a Decomposition of Population

As opposed to partitioning the data by time, decomposing the data by population is a much simpler proposition. To decompose the data by population, we use the full 18 months of data for every sample (i.e., the first 12 months for features, and the last 6 months to develop labels). We then perform stratified random sampling to split the data into a 70% training set and 30% test set. Stratifying the sample ensures that the distribution of super-utilizers to non-super-utilizers in the training and test sets is consistent and reflects the actual distribution of the outcome variable in the wild.

We still need to reconcile adjudication, however with the 12-month feature period, adjudication presents far less of an issue in preparing the data. Table 37 shows that even with the adjudication period, we can collect a full 9 months of longitudinal historical data to inform a super-utilizer prediction while

preserving the full 6-month label period. However, when decomposing the data by population, we lose most of the temporal nature of the data. For example, we cannot analyze multiple super-utilizer label periods to analyze what super-utilizers looked like in the previous period, or what they will look like in the subsequent period. This more longitudinal view of looking at beneficiary's super-utilization progression as a type of evolution is useful in trying to understand long-term patterns. Thus, I recommend that researchers explore both data decomposed by time and population as a multidimensional approach to examining different experimental strengths and weaknesses when faced with limited data.

Table 37. 9-Month Data Collection Over 12 Months with Adjudication

12-Month Feature Period									**Adjudication Period**		
1	2	3	4	5	6	7	8	9	1	2	3
6-Month Label Period											
1	2	3	4	5	6						

5.2 Discussing Dependent Variable Design

As described in "Section 1.3: Task Formulation," I carefully construct the dependent variable by performing several deliberate preprocessing steps to maximize the utility of the predictive models. The technical implementation of the dependent variable is detailed in "Section 3.3: Technical Design of the Independent Variable." I implement specific inclusion and exclusion criteria to maximize the utility of the models so that the models are focused on identifying beneficiaries who are (1) at risk of ER and IP super-utilization, and who have (2) no clear indicators in the data (i.e., diagnostic, or behavioral) that would explain superuser of ER and IP services. Through these carefully designed procedures, I attempt to ensure that the models will automatically discover patterns of behavior that produce *unexpected* levels

of emergency department (ED) and inpatient (IP) utilization. By removing ER and IP events from consideration that can be expected given a beneficiary's clinical history, and by further excluding beneficiaries who exhibit common characteristics associated with ER and IP super-utilization (e.g., issues such as mental health and substance abuse), I focus model training on a novel population of ER & IP "super-utilizers." Thus, my aim is to uncover a novel population of beneficiaries for whom it remains possible to alter trajectories of super-utilization through timely intervention.

5.3 Discussing Results

Harris et al. concluded in their study "Characteristics of Hospital and Emergency Care Super-utilizers with Multiple Chronic Conditions," that there is a **subgroup of super-utilizers** of ER and IP services that merit attention because they are particularly amenable to intervention due to patient characteristics, future utilization patterns, and health outcomes. The researchers suggest that this group is important to identify because "*care transitions programs delivered too broadly are less likely to be cost-effective and sustainable*." Their research culminates in an algorithm that can successfully identify historical and concurrent patients who are amenable to a highly targeted care transitions program (Harris et al., 2016). Through this dissertation, I extend the work of Harris et al. to develop a model that precisely *predicts* future super-utilizers of ER and IP services.

Through my literature review, I discovered a wide range of useful features that are not available in the limited data set used to conduct this dissertation research. As discussed further in "Section 5.3.1: Feature Engineer & Algorithm Performance," due to missing data I am unable to utilize common features that have repeatedly demonstrated to be highly predictive in a variety of health outcome prediction tasks. These absent features include *any* (even basic) drug information, social determinants of health, and cost. Furthermore, although I have a considerably large Medicare dataset with respect to population, I have extremely limited historical data for these beneficiaries. As discussed in "Section 5.1: Discussing Dataset Design," I have only 6 or 12 months (depending on the experiment) of historical data

to learn from for model training. Also, of this 6 and 12 months, only 3 and 9 months (respectively) of the data is fully adjudicated. This means that in both experiments, 3 months of the data (i.e., 50% and 25%, respectively) contains only partial data.

To maximize the utility of the model, I concentrate on a predicted population that will complement known super-utilizers. Hence, I am careful to remove ER and IP visits that are known to be correlated with super-utilization, including certain types of neoplasms, surgeries, and mental health issues. The parameters of this task create a truly difficult problem where I have limited data (both in duration and completeness) to predict a subgroup of super-utilizers who should not obviously super-utilize in the near future based on known clinical and behavioral information. Thus, this prediction task is only enabled by beneficiary-based features such as demographic information, basic utilization patterns, and advanced algorithms. Yet, despite these challenges, I was able to develop a state-of-the-art model that reliably predicts future super-utilizers of ER and IP services. Harris et al. suggest that beneficiaries with multiple chronic conditions, and certain utilization patterns diagnoses are especially amenable to intervention. Because these same categories of features demonstrated importance in my models, I propose that I can predict the **subgroup of super-utilizers** that Harris et al. targeted (*those amenable to intervention*) with high precision (See 5.4 Research Implications).

Each of the sets of predictions I generate are qualified by some degree of model confidence in the form of a probability quintile. The prediction quintiles range from 1 (*the upper quintile, which means "most confident positive prediction"*) to 5 (*the lower quintile, which means "least confident positive prediction"*). The performance in each prediction quintile outperforms benchmark performance, as the top-performing model ranges from nearly *55% precision on over 5% recall* to nearly *24% precision on* nearly *24% recall*. When the model is most certain, it captures over 5% for over 5-in-10 super-utilizers predicted. This is extremely useful, as only 2.7% of beneficiaries belong to the positive class, which is at

the level of anomaly detection. When the model's positive prediction is least certain, it captures many more super-utilizers (i.e., almost 25%) for close to 1-in-4 predicted.

For each algorithm, model performance was remarkably consistent from cross-validation to prediction on the holdout set, to prediction on the production set (See Chapter 4 Results). As expected, XGBoost (which is state-of-the-art in the literature for classifiers built on tabular data) consistently outperformed the other classifiers in all experiments. Logistic regression also consistently outperformed random forest, although the three classifiers yielded similar results overall. The Object2Vec deep learning algorithm did not outperform XGBoost for this task, but likely would have if the research budget allowed for more experimentation on a more sophisticated configuration (this is explained in more detail in "Section 5.3.1: Feature Engineering & Algorithm Performance").

5.3.1 Feature Engineering & Algorithm Performance

5.3.1.1 Tabular Data. In the tabular data, I engineered numerous utilization-based features discussed in the literature, including Clinical Classifications Software (CCS) diagnostic categories (used by (Yang et al., 2018) and Hierarchical Condition Categories (HCCs). CCS diagnosis categories were introduced by The Healthcare Cost and Utilization Project (HCUP), which is a suite of healthcare databases and other products sponsored by the Agency for Healthcare Research and Quality (AHRQ) (Cost et al., 2016). CCS allows for classification of diagnoses and procedures into a limited number of categories by aggregating individual ICD-10 codes into broad diagnosis and procedure groups. The HCC model was implemented in 2004 by CMS to "adjust Medicare capitation payments to private health care plans for the health expenditure risk of their enrollees (Pope et al., 2004)." The HCC model stratifies clinical diagnoses into risk groupings, which are scored alongside demographics to perform risk adjustment. I incorporated the HCC risk groupings and similar demographic features in the development of my models.

As discussed in Feature Engineering "Section 3.4.2.1: Classifiers" and "Section 3.4.2.2: Deep Learning Classifiers," I included ICD-10-CM categories represented by the first three digits of an ICD-10-CM code, which describes the general type of the injury or disease without the specific subcategory. I did not include the full ICD-10-CM codes as features because there are almost 70,000 codes and most were too specific to provide adequate coverage. I included procedure information in the form of CCS procedure codes, which I aggregated up into total procedure counts. I engineered numerous features around different statistics, including counts, averages, and maximums. Some of the more important features are demographic features, discussed further in "Section 5.4.3: Research Implications."

One of the most important limitations of this research is the lack of meaningful person-centric features. The "Limited Data Set" I procured follows strict protocols to remove potentially identifiable protected health information (PHI), which means that the most granular geographic indicator I have is at

the county level. Unfortunately, I was unable to include many open-source resources at the block-group or census-tract level, which significantly increase variance and reduce bias. These resources are often crucial to health prediction systems, especially those mentioned in the literature that inform individuals' social determinants of health (SDOH) (Ancker et al., 2018; J. H. Chen & Asch, 2017; Vest & Ben-Assuli, 2019). Some of these resources include the area deprivation index (ADI), which was developed by Dr. Amy Kind, MD, PhD, and her team at the University of Wisconsin (University of Wisconsin School of Medicine Public Health, n.d.); and the Center for Disease Control's (CDC) social vulnerability index (SVI) (*CDC Social Vulnerability Index*, 2014).

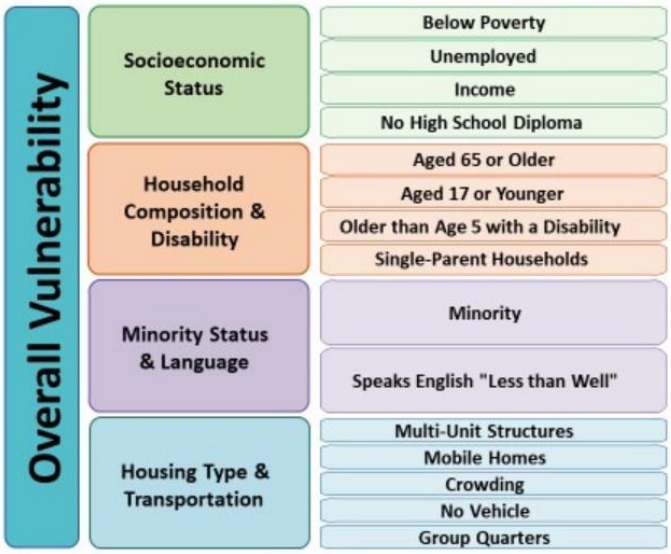
The ADI is available at the block-group level. It quantifies important SDOH, such as factors related to income, education, employment, and housing quality. This permits rankings of neighborhoods by socioeconomic status disadvantage. Typically, the ADI is used to inform health delivery and policy, especially for the most disadvantaged neighborhood groups.

The SVI, available at the census-tract level, quantifies SDOH into 5 different themes, including:

1. Socioeconomic status
2. Household composition and disability
3. Minority state and language
4. Housing type and transportation
5. Composite Vulnerability

See Figure 9. SVI - Overall Vulnerability for a more detailed classification of SVI themes (*CDC SVI Documentation*, 2018).

Figure 9. SVI - Overall Vulnerability



Drug and pharmacologic features are also repeatedly discussed in the literature as being highly predictive in medical informatics systems (Buchan et al., 2017). Medicare Part D data was not included in my Limited Data Set, so I am missing all information related pharmacologic treatment, including whether a beneficiary's condition is managed through medication. Based on my experience as an applied data scientist in healthcare, I am certain that the availability of these resources would substantially improve the predictive performance of my models.

Finally, despite our best efforts, our research team could not locate a consistent and reliable method to adjudicate cost, which is widely cited in the literature as both an important predictor and useful for validation. Payment information was sometimes missing and other times contradictory, so we are unsure if there was a data extraction issue or if there is a more comprehensive foreign protocol to process payment information. Because the dependent variable in my model was utilization-based, the lack of cost data did not prevent my research from proceeding. However, stable payment information would have likely improved predictive performance and validation efforts.

In the absence of these useful person-centric descriptors and pharmacologic features, I engineered various utilization-based features, many of which were indeed predictive. A full list of

features is available in Supplementary Material Appendix B. Data Dictionaries Features, and the relative importances of these feature importances is discussed further in "Section 5.4.3: Research Implications."

5.3.1.2 Object2Vec Encoders. The full object embeddings design is discussed in featuring engineering "Section 3.4.2.2: Deep Learning Classifiers." The left encoder contained relevant demographic, eligibility, and utilization factors, while the right encoder consisted of sequential diagnosis and procedure (i.e., ICD-10-CM and ICD-10-PCS) codes. These pairs of encoded embeddings were modeled using the task label such that the comparator learned to represent the distance of different demographic objects using the similarity in sequential diagnoses and procedures for *super-utilizers* and *not super-utilizers*; and it also learned to represent the distance of different sequential diagnoses and procedures objects using the similarity in demographics for *super-utilizers* and *not super-utilizers.* The model artifacts, which are the learned embeddings, were then used for a downstream classification task (i.e., classifying super-utilizers).

The embeddings were indeed predictive, as the top performing Object2Vec deep learning model trained with BiLSTM networks outperformed benchmark state-of-the-art performance (*i.e., 20% precision*) **on a more difficult task** in each of the predicted quintiles for both experiments. (*i.e., 40.54% in the upper quintile; 32.47% in the upper-middle quintile; 29.02% in the middle quintile; 26.49% in the lower-middle quintile; 24.54% in the lower quintile for the Decomposition of Time Experiment; and 41.02% in the upper quintile; 31.68% in the upper-middle quintile; 27.73% in the middle quintile; 25.03% in the lower-middle quintile; 23.07% in the lower quintile for the Decomposition of Population Experiment*).

The Object2Vec deep learning models did not outperform the XGBoost models, which were much more precise (*i.e., 54.36% precision in the upper quintile; 37.91% precision in the upper-middle quintile; 31.46% in the middle quintile; 27.51% in the lower-middle quintile; and 24.75% precision in the lower quintile for the Decomposition of Time Experiment; and 54.64% precision in the upper*

quintile; 39.27% precision in the upper-middle quintile; 33.29% in the middle quintile; 29.01% in the lower-middle quintile; and 26.20% precision in the lower quintile for the Decomposition of Population Experiment).

There were several reasons as to why the Object2Vec deep learning models did not outperform the XGBoost models, but most are related to lack of time and budget limitations. To conduct this research, all experimentation was required to take place in a PHI-compliant environment, including data migration, data processing, exploratory data analysis, modeling, and exploration of results. After the purchase of the data, I operated within a $4,000 budget. Given the compliance requirements and resource limitations, we decided to utilize Amazon Web Services (AWS) and its vertically integrated machine learning platform Amazon SageMaker. With support from a small research team, we were able to store the data in an RDS relational database, analyze the data in a Jupyter Notebook on an EC2 instance, transfer data to s3 buckets for storage, and transfer the data via API to conduct rapid experimentation using Amazon SageMaker. To employ different machine learning methods not native to SageMaker, I developed homegrown solutions using the Sagify API in Python (*Sagify*, n.d.).

Given the budget constraints and cost of this environment, I limited the number of features that we fed to the encoders. For example, the deep learning classifiers did not benefit from any of the CCS, HCC, and other critical summary statistic features, many of which demonstrated to be among the most important features in the XGBoost experiments (See "Section 5.4.3: Research Implications"). Again, due to cost, I also only tuned a partial set of hyperparameters for the deep learning experiments (e.g., I used the default 'adam' optimizer for all experiments, which is reported in Supplementary Material Appendix E. Hyperparameters), while we were able to perform full hyperparameter tuning on the other classifiers.

Finally, during our experimentation, we only compared pairs of networks (e.g., we compared the different performance of left and right-side encoders as Pooled Embedding, HCNN, and BiLSTM

networks). However, we should have tested the encoders using combinations of different networks. For example, the left-side encoder was composed of demographic features, which are categorical. It makes sense to evaluate the Pooled Embedding and HCNN networks for these non-sequential features. The right-side encoder was composed of sequential features, so it should have been evaluated with the BiLSTM network. Although we performed each of these experiments independently, we never tested the combination of the Pooled Embedding or HCNN network for the left-side encoder and the BiLSTM encoder for the right-side encoder.

The impact of not mixing networks for the encoders is evident in the performance of the Object2Vec model trained using hierarchical convolutional neural networks (*i.e., 10.53% precision in the upper quintile; 14.31% precision in the upper-middle quintile; 14.72% in the middle quintile; 14.47% in the lower-middle quintile; and 14.37% precision in the lower quintile for the Decomposition of Time Experiment; and 31.29% precision in the upper quintile; 23.64% precision in the upper-middle quintile; 20.73% in the middle quintile; 18.64% in the lower-middle quintile; and 17.17% precision in the lower quintile* for the Decomposition of Population Experiment). The considerably low performance of the models is due to the sequential nature of the right-side encoder. Here, we would expect performance to significantly improve if we implemented the HCNN for the left-side encoder but used a recurrent neural network (RNN) architecture such as BiLSTM for the sequential right-side encoder.

Given that the deep learning experiments outperformed the benchmark, the Object2Vec method shows great promise for this task. With additional funding, I would like to train a model that includes additional features, perform full hyperparameter tuning, and test different combinations of networks. Furthermore, the model artifacts (i.e., the learned embeddings) can then be used for downstream unsupervised and/or supervised tasks. For example, to visualize natural clusters of related objects in the dense low-dimensional space, it is possible to perform cluster analysis by using the learned embeddings

as input to an unsupervised clustering algorithm (e.g., nearest neighbors). As future work, I would like to explore the utility of clustering the embeddings to supplement this and other tasks.

5.3.2 Discussing Bayesian Hyperparameter Optimization & Cost-Sensitive Classification

When tuning the Object2Vec deep learning classifier, if the decay term or learning rate is too large, all weights are essentially zeroed out so that the Object2Vec model will yield an identical prediction regardless of input data.

In my experiments, implementing a class weight to mitigate the class imbalance problem for the logistic regression and random forest classifiers did not produce better results. However, tuning the XGBoost pos_scale_weight parameter, which is designed to mitigate the class imbalance problem, was critical to better performance (See "Section 4.2: Cost Sensitive Classification Results").

5.4 Research Implications

5.4.1 Why Should You Trust Machine Learning?

> *"**Why should you trust machine learning**? We have a particular problem as we move towards the AI applications, we are beginning to automate the ineffable. That means that we are solving things where we could not come up with the recipe. I do not know about you; I take a pretty positive view of my species. And I imagine that we are fairly smart creatures, and if the recipe were simple and there was enough reward in it, I imagine just through sheer bloody-minded stubbornness we would figure out a way to handcraft that solution. So, if we are forced to rely on something like deep learning to solve it, and we can solve it no other way, I imagine that maybe that is because the underlying recipe is now so complicated that it is too much for us to read and think about. There is a memory capacity limitation in here and so if that thing is really complicated, you cannot really expect to open it up and read the recipe and go, 'Aha! now that's how it's figuring out what is a cat and what's not a cat.' If the pixel in the top right-hand corner is blue, then it is a cat. It is not going to be that simple. You are not going to be able to read it and understand it. So, you are not going to be able to base your trust of it on the recipe: you are stuck. **Now what I propose is that that's not a great basis for trust anyway.**"*
>
> - **Cassie Kozyrkov**, Chief Decision Scientist at Google (Kozyrkov, 2021a)

In the epilogue to this section, titled “Why Should You Trust Machine Learning?” Cassie Kozyrkov (2021a), who is Chief Decision Scientist at Google, introduces the topic of trust (or perhaps *mistrust*) in machine learning. Here, Cassie is focusing on an aspect of trust that is related to the technical complexity of deep learning. However, issues of trust in machine learning span from underlying technical complexities that are uninterpretable by humans, to the use of data that is biased (*i.e., discriminatory instead of discriminative*), misinformation, privacy, and confidentiality, and how (and why) models are deployed in the real world to make peoples’ lives better (or worse). While Kozyrkov (2021a) proposes a solution to promote trust in machine learning (which is discussed later in this section), we should first examine why users have justifiable cause for concern and if it is possible to strike a balance between rapid innovation in artificial intelligence and the thorough consideration of ethical implications that should take precedence.

In the introduction section of the dissertation, I discuss how this research is made possible by advancements in big data technology. Well before we could train models with one trillion parameters, researchers were contemplating issues related to privacy and forecasting how these problems would become more pervasive as big data applications became ubiquitous. In a 2013 article titled “Big Data for All: Privacy and User Control in the Age of Analytics,” Tene and Polonetsky (2012) claimed that protecting privacy was becoming a serious challenge as data became easier to aggregate and share. The privacy issues extended to “profiling, tracking, discrimination, exclusion, government surveillance, and loss of control” (Solove, 2005), and automated decision making built on a foundation of hidden (*i.e., uninterpretable, or unexplainable*) features (D. Robinson et al., 2014).

In a 2016 report published by the Council for Big Data, Ethics, and Society, Metcalf and Crawford (2016) discussed the ethical concern of informed consent in research. The authors described how consent is front-loaded in the research process, i.e., subjects agree to data collection well before the data is used for research. In this situation, it is difficult to assess if a subject truly understands the

potential risks or benefits associated with their consent. This concern of "temporal stretching" is greatly intensified in the context of a big data framework because as data becomes increasingly easier to collect, store and analyze, consumer awareness of what personal information an institution can extract and then *infer* from an individual is made more ambiguous. In a qualitative study of big data and the opioid epidemic, Evans et al. suggest that big data users fear potential privacy infringements, which include "increased profiling and surveillance capabilities, limitless lifespan, and lack of explicit informed consent… including the inability of affected groups to control how big data are used, the potential of big data to increase stigmatization and discrimination of those affected despite data anonymization" (Evans et al., 2020).

One (now infamous) myth of personal information gathering, and unexpected inference is the story of how Target purportedly tracked store purchases to predict if a customer was pregnant (*How Target Gets the Most out of Its Guest Data to Improve Marketing ROI « Machine Learning Times*, 2010). Supposedly, a father was then surprised to learn of his daughter's pregnancy through a series of targeted coupons (Duhigg, 2012; Hill, 2012). Although this unsubstantiated tale is likely either the result of some coincidence or embellishment (Piatetsky, n.d.), such controversial marketing lists have certainly stirred controversy. For example, in a 2014 report on social technology and justice, Robinson et al. (2014) investigated several unethical consequences of big data applications in production at various institutions. The report detailed how data brokers enabled targeting of financially vulnerable communities. Specifically, the Federal Trade Commission and Senate Commerce Committee released reports that identified marketing lists with categories such as "'*Rural and Barely Making It*,' '*Ethnic Second-City Strugglers*,' and '*Retiring on empty: Singles*'," among others.

Of course, challenges to big data privacy extend well beyond invasive acts of inference for various forms of targeted marketing. The pervasiveness of big data and the politicization of vital intelligence has catalyzed rapid spread of misinformation. For example, early Coronavirus reports

greatly exaggerated death tolls to over 10,000 fatalities in Wuhan alone when the estimated death toll was still under 100 cases. “Magical mineral solutions” (also referred to as MMS or 20-20-20), which consists of a bleaching agent, were marketed on social media to effectively prevent Coronavirus infections (Santos Rutschman, 2020). Information authenticity is now a major problem with the advent of omnipresent machine learning models. The purpose of generative adversarial networks (or GANs), which are responsible for “deepfake” technology, is often to create information that is artificial. In doing so, applications of GANs have demonstrated the capacity to cause harm to vulnerable populations. For example, an app called “DeepNude" was developed to produce undressed photos of women (Hao, 2019).

Even when we can rely on data that we know to be mostly authentic, historical data often contain hidden bias—and not just statistical bias that helps to discriminate between choices like “Super-Utilizer” and “Not Super-Utilizer.” For healthcare tasks, the data that we learn from almost always contains differences in behavior between certain groups of individuals, and while much of this information will demonstrate utility with respect to some predictive end, other information may serve to trap individuals in a negative reinforcement loop. For example, information related to sexual orientation is usually not available to data science practitioners during exploratory data analysis, development, deployment, and beyond (Tomasev et al., 2021). In 2019, *Science* published a study that described how a hospital based its patient recommendations to a care management program on an algorithm that relied heavily on predicted cost as an important feature. In using the algorithm, the hospital inadvertently recommended more white than black patients even though the patient populations carried similar clinical risk. Despite comparable illness burdens, the authors hypothesized that black patients were more likely to suffer from structural issues like lack of healthcare access (Obermeyer et al., 2019).

When the data is known to be authentic and fully representative, we still face problems related to the amount of predictive information carried by a feature and how these features are distributed between certain groups of individuals. I touched on this phenomenon previously in “Section 1.3: Task

Formulation" when first discussing exclusion criteria. Imagine, for example, that maternity diagnoses were not excluded from the dependent variable used in this study. The presence of a maternity diagnosis should almost perfectly correlate with an inpatient visit because most child births occur in inpatient facilities. Thus, a maternity diagnosis carries strong predictive information with respect to predicting inpatient visits, and if we examine the tail where the model is most confident in its predictions, we should not be surprised to see an overrepresentation of women (i.e., women with maternity diagnoses) as compared with men as distributed in both the actual general patient population and the actual patient population of "Super-Utilizers." This is because there typically exists a small portion of the population for whom it is easy to predict, and this condition persists across "sexual orientation, race, health status, location and even your intention to leave your job" (Siegel, 2020).

5.4.2 Machine Learning Should Earn Your Trust

> *Think about going on a new kind of airplane. An entirely new kind of airplane—a spacecraft, let us do that. A new kind of spacecraft. And you have got two options for the spacecraft you are going to sit in.*
>
> *(1) Spacecraft number one: all the physics is written out exactly for how it works. All of it is available and you can read that thick stack of documents.*
>
> *(2) Spacecraft number two: no information on how it works, but lots and lots of information about all the previous flights that it completed successfully. all the trials, all the testing conditions.*
>
> ***Which of the two would you like to trust yourself to?*** *The one you know exactly how it works, but it has never been tested? Or the one that you do not understand how it works, but it has been tested thoroughly. Number 2, I have a strong preference for number 2. For a lot of our medication, we have no idea how they work—we just know that it does work. That is good enough for us. As long as we can be sure it works. I submit to you that that is a much better basis for trust—checking that it does work. Knowing how it works… that is a pleasant extra thing.* ***But checking that it does work, that is what you should be basing your trust on.***
>
> - **Cassie Kozyrkov**, Chief Decision Scientist at Google (Kozyrkov, 2021b)

In the epilogue to this section, titled "Machine Learning Should Earn Your Trust," Kozyrkov (2021b) now delivers the hook: the best way to trust machine learning algorithms is by checking that

they work. This concept is well articulated by Maya Krishnan (2020) in "Against Interpretability: A Critical Examination of the Interpretability Problem in Machine Learning," where Krishnan writes:

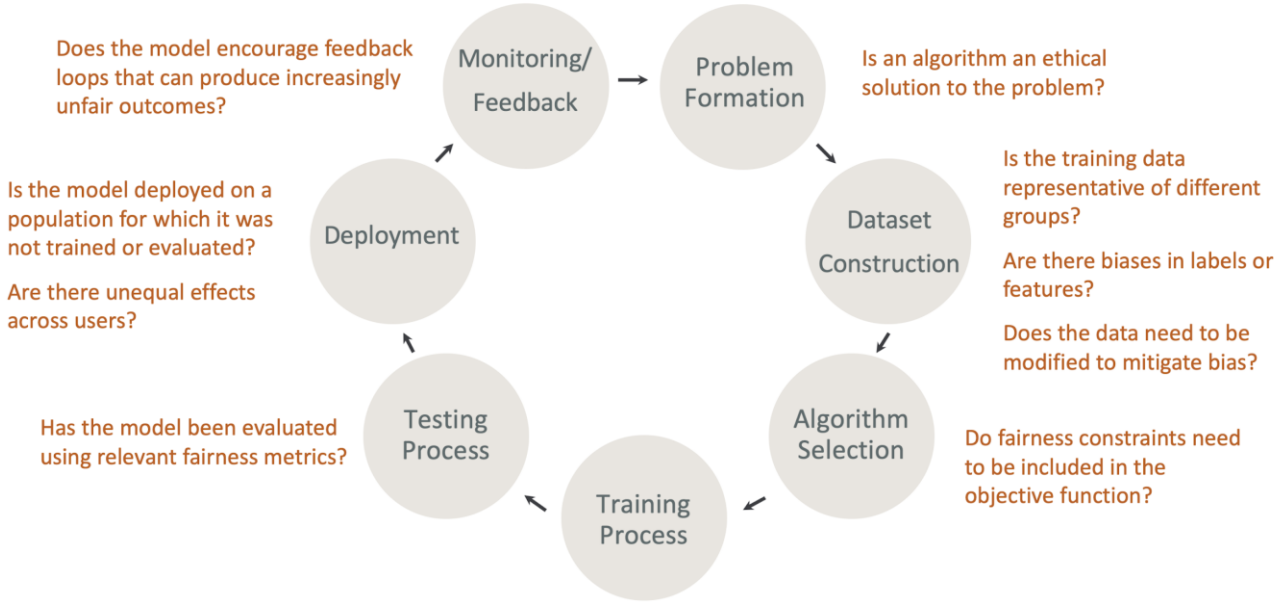
> *An interpretation or an explanation of an algorithm is supposed to provide insight into how it works. If we have an interpretation of an algorithm, we should be in a position to know why, when that algorithm generates a particular output, it produces the output that it does. While the last section highlighted the difficulty of determining what exactly an interpretation is, this section questions how the desired function of an interpretation—to provide an account of why the algorithm provides the outputs that it does—can be made relevant to ultimate ends of interpretation.* ***It is not clear how or why knowing about the process that leads an ML algorithm to produce particular outputs leads to knowledge about the basis or justification for the output itself****.*

As my intention is to build a system that successfully predicts "Super-Utilizers" across all groups of individuals (which should tangibly contribute to the elimination of disparities in health systems), I must do my part to build users' trust (Simpson & Jain, 2021; Wesson et al., 2019). So how do I build users' trust? By emphasizing interpretability, explainability, and transparency in the context of ***checking that the system does work*** *(Kozyrkov, 2021b; Krishnan, 2020)*.

To "check that the system works," I apply techniques designed to address the issue of explainability in machine learning, which help to identify and mitigate bias (Silberg & Manyika, 2019). For example, I produce Shapley values that explain why a particular prediction was reached, which features in the data led to the prediction, and precisely how much these features contributed to the prediction. I also produce feature importances using 'gain', which help define the relative feature importance of a single feature (for more information, see "Section 3.4.4: Interpretability and Explainability"). Finally, I perform an exhaustive disaggregation analysis that compares ground truth feature distributions (e.g., variables such as age, sex, and race) in the general population to the "Super-Utilizer" population to the predicted "Super-Utilizer" population. These explainability techniques serve to diagnose if a prediction reflects bias and will enable accountability in the pilot phase (Silberg & Manyika, 2019).

My approach to addressing interpretability and explainability is in accordance with Amazon Clarify's best practices for evaluating fairness and explainability in the ML lifecycle (See Figure 19. Amazon SageMaker Clarify - Best Practices for Evaluating Fairness and Explainability in the ML Lifecycle). Amazon Clarify was developed to "provide machine learning developers with greater visibility into their training data and models so they can identify and limit bias and explain predictions (*Amazon SageMaker Clarify*, n.d.)."

Figure 19. Amazon SageMaker Clarify - Best Practices for Evaluating Fairness and Explainability in the ML Lifecycle



Furthermore, the exhaustive disaggregation analysis conducted follows the example set forth by the Agency for Healthcare Research and Quality (AHRQ) in response to a mandate by Congress to provide a comprehensive overview of national healthcare disparities in care experienced by different racial and socioeconomic groups. In response to the mandate, AHRQ has published an annual report on healthcare quality and disparities for the past 17 years. The 2019 National Healthcare Quality and Disparities Report was published in December 2020, and its purpose is to assess performance of our national healthcare system by identifying strengths and weaknesses for access to quality healthcare, including

disparities.

The 2019 report indicates that although some disparities were minimized from 2000 to 2018, other disparities continued to persist through 2019 while others (especially those related to poor and underinsured populations in priority areas) worsened. For example, for approximately 40% of quality measures reviewed in the report, Blacks and American Indians and Alaska Natives received worse care than Whites; in over 33% of quality measures, Hispanics received worse care than Whites; for nearly 30% of quality measure, Asians received worse care than Whites, but received better care than Whites for almost 33% of quality measures. These disparities varied by geographic region, as residents of large central metropolitan areas received worse care than residents of large fringe metropolitan areas in nearly 25% of quality measures; residents of micropolitan and noncore areas received worse care than residents of large fringe metropolitan areas in 33% of quality measures; and medium and small metropolitan residents received worse care than residents of large fringe metropolitan areas for almost 20% of quality measures. AHRQ discloses the methods it uses during the review, which is published in the 2019 National Healthcare Quality and Disparities document Report Introduction and Methods document (*2019 National Healthcare Quality and Disparities Report Executive Summary*, 2020).

Efforts like the transparency in healthcare outcomes and disparities outlined in the 2019 National Healthcare Quality and Disparities document are becoming prevalent in healthcare. There have been analogous calls for the elimination of healthcare disparities in the areas of benefit design (Bruce Sherman, 2020), and Coronavirus vaccine distribution (Persad et al., 2020; Samantha Artiga, 2020). Also in 2021, the American Public Health Association (APHA) sponsored the 2021 Strengthening Public Health and Improving Equity Through Policy and Data Innovation Conference (supported in part by the U.S. Department of Health and Human Services (HHS)), which featured three days of discussions focused on topics such as “Addressing Health Equity,” “Implicit Bias in Health Care,” “The Ongoing COVID Response and Next Pandemic,” and “What Does Racism Have to Do with Data and Health

Policy? (*Strengthening Public Health and Improving Equity through Policy and Data Innovation*, 2021)"

5.4.3 Discussing Feature Importances & Disaggregation Analysis & Disaggregation Analysis

As introduced in "Section 3.4.4.1: Feature Importances" and presented in "Section 4.4.1: Feature Importances," 'gain' is used to calculate feature importances for the XGBoost models to examine the relative importance of each feature. For the data as a decomposition of time experiment, 605 of 2,853 features (i.e., 21.2%) register at least some measure of predictive importance. However, for the data as a decomposition of population experiment, only 158 of 2,853 features (i.e., 5.5%) indicate some level of feature importance. This is expected, as many the one-hot encoded clinical features—*e.g., the presence of a specific and rare ICD10 code such as the truncated ICD-10-CM code A04.9 for "bacterial intestinal infection, unspecified," which is only coded for one beneficiary*—are sparse and do not provide sufficient "cover" (i.e., the relative quantity of observations for a feature). Features that do not indicate importance should be trimmed to reduce the dimensionality of the model and improve its generalizability.

For the data as a decomposition of time experiment, the most important features are the "Number of Historical Emergency Department + Inpatient Events" and the "Emergency Department Event Sum," which account for 18.68% of the model's importance. This gives face validity to the model, as both are historical measures of the dependent variable and should exhibit strong importance. Note that the "Emergency Department Event Sum" feature ranks third in importance for the data as a decomposition of population experiment as well, and the number of 90-day readmissions was the 23rd most important feature in both experiments.

Similar utilization-based features that are representative of the dependent variable include the "Total Event Sum" feature (which ranks 14th in importance for the data as a decomposition of time experiment and 8th for the data as a decomposition of population experiment); the "Home Event Sum"

feature (which ranks 11th in importance for the data as a decomposition of time experiment and 23rd for the data as a decomposition of population experiment); and the "Outpatient Event Sum" feature (which ranks 20th in importance for the data as a decomposition of time experiment and 15th for the data as a decomposition of population experiment).

The top 20 most important features from both experiments provide more face validity, as most were specified as highly important in the literature and support the claim by Harris et al. that certain super-utilizers are particularly amenable to intervention—*especially those with multiple chronic conditions* (See "Section 2.4: Super-Utilizer Characteristics"). Thus, it is promising that the "'Distinct Chronic Condition Count" features were the 16th most important feature in the data as a decomposition of time experiment and the **2nd most important feature** in the data as a decomposition of population experiment. The "Distinct Chronic Hierarchical Condition Category Month Count" (*which is like a distinct chronic condition count as it is a hierarchical classification of certain chronic diseases that are most relevant to Medicare for the purposes reimbursement*), was the 5th most important feature for the data as a decomposition of time experiment and the **most important feature** in the data as a decomposition of population experiment. Similar clinical features that were engineered as different measures of chronic illness (such as the "Maximum Number of Monthly Distinct ICD10 Codes", "Distinct Body System Category Count", and the "Distinct Clinical Classification Software Category Month Count" features) were in the top 10 for either or both experiments.

Beneficiaries who suffer from hypertension (the 10th most important feature in the data as a decomposition of population experiment) and chronic obstructive pulmonary disease (the 11th most important feature in data as a decomposition of population experiment), are precisely the population that Harris et al. suggest are likely to be amenable to intervention through more directed ambulatory care. Beneficiaries who are likely to respond to intervention also include those who experience super-utilization due to urinary tract infections (the 22nd most important feature in the data as a decomposition

of time experiment and the 26th most important feature in the data as a decomposition of population experiment).

As for important features engineered around demographics and social determinants of health, age was the 13th ranked feature in the data as a decomposition of time experiment and the 9th most important feature in the data as a decomposition of population experiment; the distinct mental health condition count was the 19th-ranked and 47th-ranked in the data as a decomposition of time and population experiments, respectively; and seven different features engineered around representations of substance abuse codes demonstrated importance, ranging in rankings from 133 to 578. Age, mental health, and substance abuse are widely cited as highly predictive of super-utilization in the literature.

In the data as a decomposition of time experiment, the "gender=male" feature ranked 10th in feature importance. To properly contextualize these features for analysis, it is necessary to conduct a thorough disaggregation analysis that compares the actual distribution of certain demographic features that indicate importance to the distribution of these variables in actual and predicted super-utilizers. The dataset contains 58.47% "gender=female" compared to 42.16% "gender=male" population (See Table 36. Overall vs. Ground Truth Super-Utilizer vs. Predicted XGBoost Super-Utilizer Demographic Analysis for the Data Decomposition of Time Experiment), a gender distribution that is historically consistent with traditional Medicare data (*Distribution of Medicare Beneficiaries by Sex*, 2020). Furthermore, 69.07% of females fit the super-utilization (as operationalized in this dissertation), as compared to 30.93% of super-utilization by males. Accordingly, 76.38% of the super-utilizers predicted with the highest probabilities are females, as compared to 23.63% males. Thus, "gender=male" is predictive (and subsequently indicates feature importance) because of the significant difference in gender distribution among actual super-utilizers.

In the data as a decomposition of time experiment, the "Race=Black" feature ranked 15th in importance. The dataset contains a baseline race=black" population of 9.16% (See Table 36. Overall vs.

Ground Truth Super-Utilizer vs. Predicted XGBoost Super-Utilizer Demographic Analysis for the Data Decomposition of Time Experiment), a distribution that is historically consistent with traditional Medicare data (*Distribution of Medicare Beneficiaries by Race/ethnicity*, 2020). Furthermore, 14.58% of the "race=black" population fit the super-utilization, which is a 5% increase over the baseline distribution. As such, 17.58% of the super-utilizers predicted with the highest probabilities will contain a "race=black" feature.

To expand on the disaggregation analysis provided in my dissertation, a follow-up study that aims to understand the causal mechanisms that drive differences in super-utilization between different demographic variables is meaningful future work. Unfortunately, the Fee-For-Service Medicare claims dataset does not contain the information necessary to conduct such a study, so supplemental data is required.

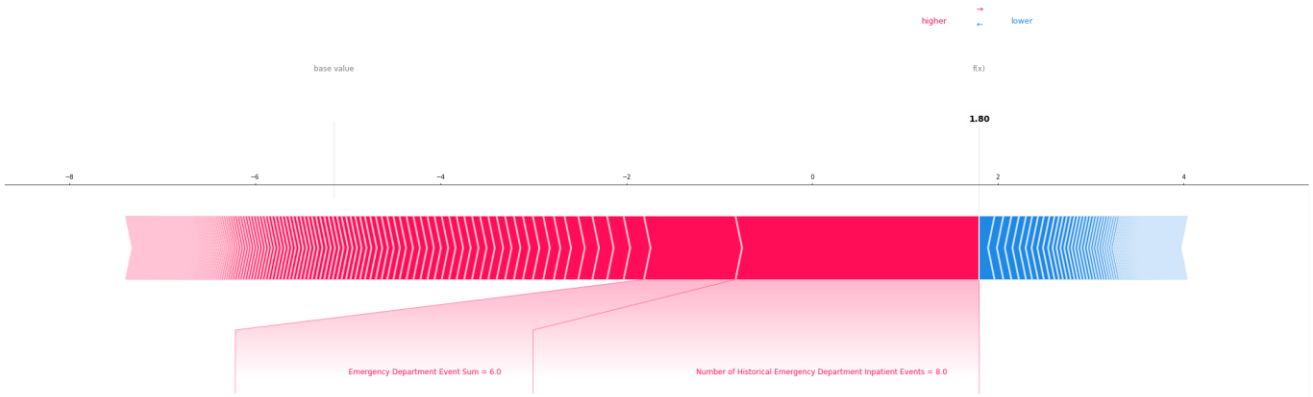
For an exhaustive list of feature importances across both experiments, see Tables G1 and G2 in Supplementary Material - Feature Importances; for an exhaustive review of all disaggregation analysis, see Supplementary Material - Disaggregation Analysis.

5.4.4 Discussing Shapley Values

Although feature importances and disaggregation analysis disclose high-level patterns across the entire dataset (as discussed in the previous section), neither explains why a model arrived at a particular individual decision for a specific beneficiary. Thus, although we know that the chronic condition count feature is important for the purposes of classifying super-utilization relative to numerous other features, exactly how important it is to predict super-utilization remains unknown (especially in the context of certain other features). As introduced in "Section 3.4.4.2: Shapley Values" and presented in "Section 4.4.2: Shapley Values," Shapley values quantify precisely how much information is carried by a feature with respect to an individual prediction.

Consider again Figure 17, which contains a force plot explaining the prediction for a specific beneficiary. The historical emergency department event sum of 6.0 and the number of historical emergency department and inpatient events sum of 8.0 explain exactly why this beneficiary was classified as a future super-utilizer.

Figure 17. SHAP Visualization of a Single Prediction for the Data Decomposition of Time Experiment



To "earn your trust," machine learning solutions should be interpretable and explainable. Feature importances and disaggregation analysis inform decisions on whether an algorithm is an ethical solution to the problem; if the training data is representative of different groups; if there are biases in the labels or features; if the data needs to be modified to mitigate bias; if the model has been evaluated using relevant fairness metrics; and if there are unequal effects across users.

Based on a thorough examination of results from feature importances and disaggregation analysis, the results from prediction I have presented are representative of ground truth super-utilizers, and the models outperform known baseline systems. The algorithm is an ethical solution to the problem. In producing and sharing Shapley values for every positively predicted super-utilizer, effects across

users are completely transparent. Even at scale (i.e., for nearly 50,000 beneficiaries), every decision is explained and interpreted such that it is possible to share a summary of why a beneficiary is at risk for super-utilization along with a recommendation for intervention.

CHAPTER 6.

CONCLUSION

In my dissertation, I operationalize "super-utilization" to match the definition proposed by Harris et al. (2016), which focuses on super-utilizers of emergency departments and inpatient services who are likely to be amenable to intervention. In their study "Characteristics of Hospital and Emergency Care super-utilizers with Multiple Chronic Conditions," Harris et al. (2016) demonstrated that super-utilizers who exhibit certain utilization patterns and diagnoses and multiple chronic conditions are particularly responsive to mediation. Moreover, the authors submit that the identification of this subgroup is significant because care transition programs that are administered too generally are less sustainable and are unlikely to be cost effective. My research culminates in multiple models that precisely predict this class of super-utilizers with state-of-the-art precision and recall, ranging from nearly 55% precision on over 5% recall to nearly 24% precision on nearly 24% recall.

My thorough examination of features important to the prediction task are promising, as results indicate that the predicted super-utilizer population is likely to respond to care management. For example, most beneficiaries identified by my models suffer from multiple chronic conditions (such as hypertension and obstructive chronic pulmonary disease). For predicted super-utilizers, many actual emergency department visits were from primary diagnoses such as urinary tract infections (i.e., 3.47% of all emergency department visits for predicted super-utilizers for the model that decomposes time and 3.19% for the model that decomposes population, respectively); headache (i.e., 2.35% and 2.28%); essential hypertension (i.e., 1.78% and 1.79%, respectively); etc. (See Appendix Table A1 and Table A2 for an exhaustive list of emergency department primary diagnosis codes for predicted super-utilizers) across both models. Accordingly, many inpatient events included diagnosis related group (DRG) codes for diagnoses that include urinary tract infections without major complication or comorbidity (i.e., 2.12% and 2.06%, respectively); and for issues related to fluids and electrolytes without major

complication or comorbidity (i.e., 1.50% and 1.24%, respectively); etc. (See Appendix Table A3 and Table A4 for an exhaustive list of inpatient DRG codes for predicted super-utilizers across both models). that would be better managed by directed ambulatory care, as opposed to emergency department visits and inpatient admissions, which are costly and suboptimal for the overall health of beneficiaries. Numerous features associated with super-utilization in the literature demonstrated strong importance with respect to the prediction task (as presented in "Chapter 5: Discussion"), which supports face validity of the system.

I deliberated the ethical implications of my system by discussing feature importances and Shapley values, and by presenting a comprehensive disaggregation analysis in support of this task. As I reported, further research is needed to explain certain imbalances in important demographic variables in the ground truth super-utilization population. For example, when controlled for population, women super-utilizer more than men by 11.23%. To study the causal mechanisms that produce this imbalance (and other imbalances presented in "Chapter 5: Discussion"), additional data sources beyond claims files are needed. However, all imbalances among important demographic variables that exists in the data are declared in "Section 4.4.2: Disaggregation Analysis."

My analyses demonstrate that the super-utilizer prediction models are both interpretable and explainable. I disclose the exact reasons for positive prediction (which I share for all beneficiaries who are predicted as super-utilizers), including explanations for nearly 45,000 beneficiaries. This is a meaningful step toward full transparency and accountability, which also enhances the utility of the system in the wild. The output of my system includes not just the "**who**" (*and "how likely"*), but the "**why**." For each beneficiary that is run through the model, the system outputs the probability that the user will super-utilize emergency department and/or inpatient services in the next six or twelve months (i.e., depending on which model is run). It also quantifies precisely why the probability was predicted by the model for each individual beneficiary. This level of personalization and detail promotes fairness in

that the beneficiary is granted access to the reasons for super-utilization prediction by the model, which allows feedback from the beneficiary to be incorporated into iterative design of the model. Also, it allows stakeholders (e.g., insurers, providers, etc.) to implement interventions that best mitigate reasons for super-utilizer prediction.

As future work, the design of a pilot program is needed to evaluate the likelihood of individual beneficiaries to respond to intervention, including the willingness of a recommended beneficiary to accept care management. Also, a follow-up study that investigates the causal mechanisms that drive differences in super-utilization between different populations (e.g., geographical, socioeconomic, etc.) is needed. To further develop the super-utilization prediction models, Object2Vec algorithm experimentation should continue with more funding. Additional features should be examined (especially features that capture social determinants of health at the individual level), and the appropriate encoder networks should be tested based on the underlying nature of the data (e.g., when the data is sequential, a recurrent neural network should be tested). Clustering should be performed on the object embeddings to discover interesting groups of clinical features and social determinants of health where similar super-utilization behaviors are shared. Similarly, reasons for super-utilization prediction should be analyzed to design successful interventions for different clusters (i.e., those with similar needs) of beneficiaries.

As the models are based on clinical information and utilization patterns, the effects of Covid-19 on model performance should be tested. For example, the Object2Vec implementation, which is more reliant on demographic variables such as social determinants of health, is likely to be more resilient to changes in utilization distributions than are the tree-based classification algorithms. Finally, it is important to test the system on the full Medicare dataset to ensure that the solution generalizes beyond the 5% subsample.

My contributions result in the foundation of a system that recommends beneficiaries who would benefit from interventions that target certain reasons for prediction—such as connecting beneficiaries to

existing ambulatory care services, or improved access—designed to prevent super-utilization of emergency departments and inpatient facilities in the future. Use cases of my system will be extended to the population level, as groups of individual beneficiaries should be identified to discover geographical "hot spots" where access to ambulatory care is inadequate but need for it is great. The immediate next steps for my system include abstraction of the models to other lines of business (e.g., Medicaid and Commercial beneficiaries), and ongoing evaluation by domain experts (e.g., decision makers in the payer and provider sectors). However, I conclude that the system is now operational for Medicare beneficiaries.

Table A1. 6-Month Feature Space XGBoost Emergency Department Event Clinical Classifications Software (CCS) for Primary Diagnosis

Clinical Classification Software (CCS) Description Principal Diagnosis	# ER Events	% ER Events	Cumulative Feature Importance
Nonspecific chest pain	2,074	6.69%	6.69%
Other lower respiratory disease	1,751	5.65%	12.34%
Other injuries and conditions due to external causes	1,548	4.99%	17.33%
Abdominal pain	1,466	4.73%	22.06%
Superficial injury; contusion	1,161	3.74%	25.80%
Other connective tissue disease	1,123	3.62%	29.42%
Spondylosis; intervertebral disc disorders; other back problems	1,120	3.61%	33.03%
Urinary tract infections	1,076	3.47%	36.50%
Other non-traumatic joint disorders	1,059	3.41%	39.91%
Headache; including migraine	795	2.56%	42.47%
Other gastrointestinal disorders	714	2.30%	44.77%
Residual codes; unclassified	698	2.25%	47.02%
Cardiac dysrhythmias	657	2.12%	49.14%
Malaise and fatigue	581	1.87%	51.01%
Conditions associated with dizziness or vertigo	565	1.82%	52.83%
Essential hypertension	535	1.73%	54.56%
Other nervous system disorders	525	1.69%	56.25%
Sprains and strains	454	1.46%	57.71%
Syncope	429	1.38%	59.09%
Fluid and electrolyte disorders	427	1.38%	60.47%
Skin and subcutaneous tissue infections	404	1.30%	61.77%
Open wounds of head; neck; and trunk	376	1.21%	62.98%
Other screening for suspected conditions (not mental disorders or infectious disease)	360	1.16%	64.14%
Nausea and vomiting	355	1.14%	65.28%
Complications of surgical procedures or medical care	344	1.11%	66.39%
Epilepsy; convulsions	338	1.09%	67.48%
Open wounds of extremities	329	1.06%	68.54%
Diabetes mellitus with complications	324	1.04%	69.58%
Genitourinary symptoms and ill-defined conditions	302	0.97%	70.55%

Chronic obstructive pulmonary disease and bronchiectasis	289	0.93%	71.48%
Complication of device; implant or graft	279	0.90%	72.38%
Other upper respiratory infections	263	0.85%	73.23%
Pleurisy; pneumothorax; pulmonary collapse	258	0.83%	74.06%
Other fractures	250	0.81%	74.87%
Hypertension with complications and secondary hypertension	248	0.80%	75.67%
Osteoarthritis	242	0.78%	76.45%
Congestive heart failure; nonhypertensive	241	0.78%	77.23%
Other skin disorders	212	0.68%	77.91%
Other circulatory disease	210	0.68%	78.59%
Fracture of upper limb	209	0.67%	79.26%
Diverticulosis and diverticulitis	192	0.62%	79.88%
Asthma	183	0.59%	80.47%
Pneumonia (except that caused by tuberculosis or sexually transmitted disease)	180	0.58%	81.05%
Coronary atherosclerosis and other heart disease	177	0.57%	81.62%
Other upper respiratory disease	171	0.55%	82.17%
Acute bronchitis	164	0.53%	82.70%
Calculus of urinary tract	147	0.47%	83.17%
Gastrointestinal hemorrhage	142	0.46%	83.63%
Conduction disorders	136	0.44%	84.07%
Other and ill-defined heart disease	136	0.44%	84.51%
Other aftercare	130	0.42%	84.93%
Noninfectious gastroenteritis	130	0.42%	85.35%
Delirium dementia and amnestic and other cognitive disorders	129	0.42%	85.77%
Allergic reactions	124	0.40%	86.17%
Anxiety disorders	124	0.40%	86.57%
Transient cerebral ischemia	123	0.40%	86.97%
Acute cerebrovascular disease	116	0.37%	87.34%
Other diseases of kidney and ureters	115	0.37%	87.71%
Abdominal hernia	115	0.37%	88.08%
Fracture of lower limb	110	0.35%	88.43%
Fever of unknown origin	105	0.34%	88.77%
Intracranial injury	102	0.33%	89.10%
Disorders of teeth and jaw	94	0.30%	89.40%

Esophageal disorders	91	0.29%	89.69%
Gout and other crystal arthropathies	91	0.29%	89.98%
Other liver diseases	90	0.29%	90.27%
Biliary tract disease	90	0.29%	90.56%
Deficiency and other anemia	82	0.26%	90.82%
Rehabilitation care; fitting of prostheses; and adjustment of devices	75	0.24%	91.06%
Intestinal obstruction without hernia	72	0.23%	91.29%
Gastritis and duodenitis	71	0.23%	91.52%
Administrative/social admission	71	0.23%	91.75%
Other ear and sense organ disorders	70	0.23%	91.98%
Viral infection	66	0.21%	92.19%
Acute and unspecified renal failure	65	0.21%	92.40%
Phlebitis; thrombophlebitis and thromboembolism	63	0.20%	92.60%
Septicemia (except in labor)	62	0.20%	92.80%
Diabetes mellitus without complication	61	0.20%	93.00%
Acute myocardial infarction	59	0.19%	93.19%
Chronic ulcer of skin	55	0.18%	93.37%
Joint disorders and dislocations; trauma-related	54	0.17%	93.54%
Other eye disorders	53	0.17%	93.71%
Other bone disease and musculoskeletal deformities	53	0.17%	93.88%
Medical examination/evaluation	52	0.17%	94.05%
Other and ill-defined cerebrovascular disease	51	0.16%	94.21%
Heart valve disorders	50	0.16%	94.37%
Other nutritional; endocrine; and metabolic disorders	48	0.15%	94.52%
Skull and face fractures	47	0.15%	94.67%
Sickle cell anemia	47	0.15%	94.82%
Other endocrine disorders	46	0.15%	94.97%
Mycoses	45	0.15%	95.12%
Inflammation; infection of eye (except that caused by tuberculosis or sexually transmitteddisease)	45	0.15%	95.27%
Intestinal infection	44	0.14%	95.41%
Other hereditary and degenerative nervous system conditions	44	0.14%	95.55%
Other inflammatory condition of skin	42	0.14%	95.69%
Fracture of neck of femur (hip)	38	0.12%	95.81%
Peripheral and visceral atherosclerosis	37	0.12%	95.93%

Occlusion or stenosis of precerebral arteries	36	0.12%	96.05%
Other female genital disorders	36	0.12%	96.17%
Chronic kidney disease	34	0.11%	96.28%
Alcohol-related disorders	33	0.11%	96.39%
Mood disorders	33	0.11%	96.50%
Poisoning by nonmedicinal substances	33	0.11%	96.61%
Respiratory failure; insufficiency; arrest (adult)	33	0.11%	96.72%
Blindness and vision defects	32	0.10%	96.82%
Otitis media and related conditions	31	0.10%	96.92%
Other acquired deformities	31	0.10%	97.02%
Other diseases of veins and lymphatics	30	0.10%	97.12%
Pancreatic disorders (not diabetes)	29	0.09%	97.21%
Anal and rectal conditions	28	0.09%	97.30%
Other disorders of stomach and duodenum	27	0.09%	97.39%
Hemorrhoids	27	0.09%	97.48%
Poisoning by other medications and drugs	27	0.09%	97.57%
Inflammatory diseases of female pelvic organs	25	0.08%	97.65%
Coma; stupor; and brain damage	25	0.08%	97.73%
Pulmonary heart disease	23	0.07%	97.80%
Hyperplasia of prostate	23	0.07%	97.87%
Adverse effects of medical drugs	23	0.07%	97.94%
Thyroid disorders	22	0.07%	98.01%
Crushing injury or internal injury	21	0.07%	98.08%
Ovarian cyst	20	0.06%	98.14%
Diseases of mouth; excluding dental	20	0.06%	98.20%
Influenza	19	0.06%	98.26%
Regional enteritis and ulcerative colitis	18	0.06%	98.32%
Other diseases of bladder and urethra	18	0.06%	98.38%
Miscellaneous mental health disorders	17	0.05%	98.43%
Diseases of white blood cells	17	0.05%	98.48%
Aspiration pneumonitis; food/vomitus	17	0.05%	98.53%
Burns	16	0.05%	98.58%
Inflammatory conditions of male genital organs	16	0.05%	98.63%
Late effects of cerebrovascular disease	16	0.05%	98.68%
Aortic; peripheral; and visceral artery aneurysms	16	0.05%	98.73%

Immunizations and screening for infectious disease	16	0.05%	98.78%
Benign neoplasm of uterus	16	0.05%	98.83%
Coagulation and hemorrhagic disorders	16	0.05%	98.88%
Peri-; endo-; and myocarditis; cardiomyopathy (except that caused by tuberculosis or sexually transmitted disease)	16	0.05%	98.93%
Other and unspecified benign neoplasm	16	0.05%	98.98%
Paralysis	15	0.05%	99.03%
Other male genital disorders	14	0.05%	99.08%
Nonmalignant breast conditions	14	0.05%	99.13%
Other infections; including parasitic	13	0.04%	99.17%
Rheumatoid arthritis and related disease	13	0.04%	99.21%
Parkinson`s disease	13	0.04%	99.25%
Bacterial infection; unspecified site	12	0.04%	99.29%
Pathological fracture	11	0.04%	99.33%
Infective arthritis and osteomyelitis (except that caused by tuberculosis or sexually transmitted disease)	11	0.04%	99.37%
Varicose veins of lower extremity	11	0.04%	99.41%
Disorders of lipid metabolism	10	0.03%	99.44%
Systemic lupus erythematosus and connective tissue disorders	10	0.03%	99.47%
Attention-deficit conduct and disruptive behavior disorders	10	0.03%	99.50%
Multiple sclerosis	8	0.03%	99.53%
Acute posthemorrhagic anemia	8	0.03%	99.56%
Menopausal disorders	8	0.03%	99.59%
Nutritional deficiencies	7	0.02%	99.61%
Other hematologic conditions	7	0.02%	99.63%
Lymphadenitis	7	0.02%	99.65%
Other congenital anomalies	6	0.02%	99.67%
Spinal cord injury	6	0.02%	99.69%
Disorders usually diagnosed in infancy childhood or adolescence	6	0.02%	99.71%
Menstrual disorders	5	0.02%	99.73%
Prolapse of female genital organs	5	0.02%	99.75%
Osteoporosis	5	0.02%	99.77%
Adjustment disorders	5	0.02%	99.79%
Personality disorders	4	0.01%	99.80%
Appendicitis and other appendiceal conditions	4	0.01%	99.81%
Developmental disorders	4	0.01%	99.82%

Immunity disorders	4	0.01%	99.83%
Glaucoma	4	0.01%	99.84%
Acquired foot deformities	4	0.01%	99.85%
Acute and chronic tonsillitis	3	0.01%	99.86%
Gastroduodenal ulcer (except hemorrhage)	3	0.01%	99.87%
Cardiac and circulatory congenital anomalies	3	0.01%	99.88%
Nephritis; nephrosis; renal sclerosis	3	0.01%	99.89%
Aortic and peripheral arterial embolism or thrombosis	3	0.01%	99.90%
Hepatitis	2	0.01%	99.91%
Poisoning by psychotropic agents	2	0.01%	99.92%
Gangrene	1	0.00%	99.92%
Substance-related disorders	1	0.00%	99.92%
Endometriosis	1	0.00%	99.92%
Sexually transmitted infections (not HIV or hepatitis)	1	0.00%	99.92%
Cataract	1	0.00%	99.92%
Cardiac arrest and ventricular fibrillation	1	0.00%	99.92%
Retinal detachments; defects; vascular occlusion; and retinopathy	1	0.00%	99.92%
Digestive congenital anomalies	1	0.00%	99.92%
Impulse control disorders NEC	1	0.00%	99.92%
Peritonitis and intestinal abscess	1	0.00%	99.92%

Table A2. 12-Month Feature Space XGBoost Emergency Department Event Clinical Classifications Software (CCS) for Primary Diagnosis

Clinical Classification Software (CCS) Description Principal Diagnosis	# ER Events	% ER Events	Cumulative Feature Importance
Nonspecific chest pain	115	6.44%	6.44%
Other injuries and conditions due to external causes	113	6.33%	12.77%
Superficial injury; contusion	104	5.82%	18.59%
Other lower respiratory disease	82	4.59%	23.18%
Spondylosis; intervertebral disc disorders; other back problems	72	4.03%	27.21%
Other connective tissue disease	64	3.58%	30.79%
Other non-traumatic joint disorders	61	3.42%	34.21%
Abdominal pain	60	3.36%	37.57%
Urinary tract infections	57	3.19%	40.76%
Headache; including migraine	44	2.46%	43.22%

Conditions associated with dizziness or vertigo	41	2.30%	45.52%
Malaise and fatigue	38	2.13%	47.65%
Open wounds of head; neck; and trunk	37	2.07%	49.72%
Residual codes; unclassified	35	1.96%	51.68%
Other gastrointestinal disorders	34	1.90%	53.58%
Syncope	34	1.90%	55.48%
Cardiac dysrhythmias	34	1.90%	57.38%
Open wounds of extremities	33	1.85%	59.23%
Sprains and strains	29	1.62%	60.85%
Essential hypertension	27	1.51%	62.36%
Fluid and electrolyte disorders	26	1.46%	63.82%
Fracture of upper limb	24	1.34%	65.16%
Skin and subcutaneous tissue infections	23	1.29%	66.45%
Nausea and vomiting	18	1.01%	67.46%
Genitourinary symptoms and ill-defined conditions	18	1.01%	68.47%
Other screening for suspected conditions (not mental disorders or infectious disease)	18	1.01%	69.48%
Complications of surgical procedures or medical care	18	1.01%	70.49%
Other fractures	17	0.95%	71.44%
Osteoarthritis	15	0.84%	72.28%
Other nervous system disorders	15	0.84%	73.12%
Fracture of lower limb	15	0.84%	73.96%
Other aftercare	14	0.78%	74.74%
Joint disorders and dislocations; trauma-related	13	0.73%	75.47%
Other circulatory disease	13	0.73%	76.20%
Chronic obstructive pulmonary disease and bronchiectasis	13	0.73%	76.93%
Other skin disorders	13	0.73%	77.66%
Pneumonia (except that caused by tuberculosis or sexually transmitted disease)	12	0.67%	78.33%
Diabetes mellitus with complications	12	0.67%	79.00%
Calculus of urinary tract	12	0.67%	79.67%
Pleurisy; pneumothorax; pulmonary collapse	12	0.67%	80.34%
Intracranial injury	11	0.62%	80.96%
Congestive heart failure; nonhypertensive	10	0.56%	81.52%
Other diseases of kidney and ureters	10	0.56%	82.08%
Noninfectious gastroenteritis	10	0.56%	82.64%

Delirium dementia and amnestic and other cognitive disorders	9	0.50%	83.14%
Acute bronchitis	9	0.50%	83.64%
Anxiety disorders	9	0.50%	84.14%
Conduction disorders	9	0.50%	84.64%
Gastrointestinal hemorrhage	9	0.50%	85.14%
Hypertension with complications and secondary hypertension	9	0.50%	85.64%
Other and ill-defined heart disease	9	0.50%	86.14%
Other upper respiratory infections	9	0.50%	86.64%
Complication of device; implant or graft	8	0.45%	87.09%
Biliary tract disease	8	0.45%	87.54%
Diverticulosis and diverticulitis	8	0.45%	87.99%
Coronary atherosclerosis and other heart disease	8	0.45%	88.44%
Acute cerebrovascular disease	7	0.39%	88.83%
Other upper respiratory disease	7	0.39%	89.22%
Asthma	6	0.34%	89.56%
Esophageal disorders	6	0.34%	89.90%
Intestinal obstruction without hernia	6	0.34%	90.24%
Abdominal hernia	6	0.34%	90.58%
Medical examination/evaluation	5	0.28%	90.86%
Peripheral and visceral atherosclerosis	5	0.28%	91.14%
Skull and face fractures	5	0.28%	91.42%
Poisoning by nonmedicinal substances	5	0.28%	91.70%
Transient cerebral ischemia	5	0.28%	91.98%
Gout and other crystal arthropathies	5	0.28%	92.26%
Deficiency and other anemia	5	0.28%	92.54%
Rehabilitation care; fitting of prostheses; and adjustment of devices	4	0.22%	92.76%
Other ear and sense organ disorders	4	0.22%	92.98%
Epilepsy; convulsions	4	0.22%	93.20%
Hemorrhoids	4	0.22%	93.42%
Viral infection	4	0.22%	93.64%
Allergic reactions	4	0.22%	93.86%
Inflammation; infection of eye (except that caused by tuberculosis or sexually transmitteddisease)	4	0.22%	94.08%
Disorders of teeth and jaw	4	0.22%	94.30%
Other acquired deformities	4	0.22%	94.52%

Fracture of neck of femur (hip)	3	0.17%	94.69%
Alcohol-related disorders	3	0.17%	94.86%
Adverse effects of medical drugs	3	0.17%	95.03%
Septicemia (except in labor)	3	0.17%	95.20%
Administrative/social admission	3	0.17%	95.37%
Occlusion or stenosis of precerebral arteries	3	0.17%	95.54%
Heart valve disorders	3	0.17%	95.71%
Diabetes mellitus without complication	3	0.17%	95.88%
Intestinal infection	3	0.17%	96.05%
Coma; stupor; and brain damage	3	0.17%	96.22%
Other eye disorders	3	0.17%	96.39%
Other inflammatory condition of skin	3	0.17%	96.56%
Other liver diseases	3	0.17%	96.73%
Acute and unspecified renal failure	3	0.17%	96.90%
Chronic ulcer of skin	2	0.11%	97.01%
Varicose veins of lower extremity	2	0.11%	97.12%
Other and ill-defined cerebrovascular disease	2	0.11%	97.23%
Poisoning by other medications and drugs	2	0.11%	97.34%
Miscellaneous mental health disorders	2	0.11%	97.45%
Phlebitis; thrombophlebitis and thromboembolism	2	0.11%	97.56%
Pancreatic disorders (not diabetes)	2	0.11%	97.67%
Mycoses	2	0.11%	97.78%
Paralysis	2	0.11%	97.89%
Gastritis and duodenitis	2	0.11%	98.00%
Other nutritional; endocrine; and metabolic disorders	2	0.11%	98.11%
Acute myocardial infarction	2	0.11%	98.22%
Other disorders of stomach and duodenum	2	0.11%	98.33%
Nutritional deficiencies	2	0.11%	98.44%
Other diseases of veins and lymphatics	2	0.11%	98.55%
Influenza	2	0.11%	98.66%
Cardiac and circulatory congenital anomalies	1	0.06%	98.72%
Inflammatory conditions of male genital organs	1	0.06%	98.78%
Fever of unknown origin	1	0.06%	98.84%
Mood disorders	1	0.06%	98.90%
Diseases of white blood cells	1	0.06%	98.96%

Diseases of mouth; excluding dental	1	0.06%	99.02%
Other bone disease and musculoskeletal deformities	1	0.06%	99.08%
Other diseases of bladder and urethra	1	0.06%	99.14%
Other endocrine disorders	1	0.06%	99.20%
Other female genital disorders	1	0.06%	99.26%
Coagulation and hemorrhagic disorders	1	0.06%	99.32%
Other hereditary and degenerative nervous system conditions	1	0.06%	99.38%
Other infections; including parasitic	1	0.06%	99.44%
Chronic kidney disease	1	0.06%	99.50%
Cardiac arrest and ventricular fibrillation	1	0.06%	99.56%
Other male genital disorders	1	0.06%	99.62%
Burns	1	0.06%	99.68%
Parkinson`s disease	1	0.06%	99.74%
Prolapse of female genital organs	1	0.06%	99.80%
Pulmonary heart disease	1	0.06%	99.86%
Adjustment disorders	1	0.06%	99.92%
Acute posthemorrhagic anemia	1	0.06%	99.98%
Glaucoma	1	0.06%	100.00%

Table A3. 6-Month Feature Space XGBoost Inpatient Event Diagnosis Related Groups (DRGs)

Diagnosis Related Group (DRG)	# IP Events	% IP Events	Cumulative Feature Importance
871: SEPTICEMIA OR SEVERE SEPSIS W/O MV 96+ HOURS W MCC	401	5.51%	5.51%
291: HEART FAILURE & SHOCK W MCC	381	5.24%	10.75%
690: KIDNEY & URINARY TRACT INFECTIONS W/O MCC	261	3.59%	14.34%
872: SEPTICEMIA OR SEVERE SEPSIS W/O MV 96+ HOURS W/O MCC	180	2.48%	16.82%
392: ESOPHAGITIS, GASTROENT & MISC DIGEST DISORDERS W/O MCC	173	2.38%	19.20%
689: KIDNEY & URINARY TRACT INFECTIONS W MCC	154	2.12%	21.32%
603: CELLULITIS W/O MCC	148	2.04%	23.36%
683: RENAL FAILURE W CC	147	2.02%	25.38%
292: HEART FAILURE & SHOCK W CC	139	1.91%	27.29%
378: G.I. HEMORRHAGE W CC	137	1.88%	29.17%
641: MISC DISORDERS OF NUTRITION,METABOLISM,FLUIDS/ELECTROLYTES W/O MCC	109	1.50%	30.67%
57: DEGENERATIVE NERVOUS SYSTEM DISORDERS W/O MCC	99	1.36%	32.03%

312: SYNCOPE & COLLAPSE	95	1.31%	33.34%
309: CARDIAC ARRHYTHMIA & CONDUCTION DISORDERS W CC	93	1.28%	34.62%
177: RESPIRATORY INFECTIONS & INFLAMMATIONS W MCC	93	1.28%	35.90%
812: RED BLOOD CELL DISORDERS W/O MCC	88	1.21%	37.11%
194: SIMPLE PNEUMONIA & PLEURISY W CC	87	1.20%	38.31%
698: OTHER KIDNEY & URINARY TRACT DIAGNOSES W MCC	86	1.18%	39.49%
193: SIMPLE PNEUMONIA & PLEURISY W MCC	85	1.17%	40.66%
65: INTRACRANIAL HEMORRHAGE OR CEREBRAL INFARCTION W CC OR TPA IN 24 HRS	85	1.17%	41.83%
552: MEDICAL BACK PROBLEMS W/O MCC	79	1.09%	42.92%
101: SEIZURES W/O MCC	78	1.07%	43.99%
682: RENAL FAILURE W MCC	78	1.07%	45.06%
189: PULMONARY EDEMA & RESPIRATORY FAILURE	76	1.05%	46.11%
481: HIP & FEMUR PROCEDURES EXCEPT MAJOR JOINT W CC	75	1.03%	47.14%
202: BRONCHITIS & ASTHMA W CC/MCC	59	0.81%	47.95%
948: SIGNS & SYMPTOMS W/O MCC	57	0.78%	48.73%
308: CARDIAC ARRHYTHMIA & CONDUCTION DISORDERS W MCC	52	0.72%	49.45%
69: TRANSIENT ISCHEMIA	52	0.72%	50.17%
389: G.I. OBSTRUCTION W CC	51	0.70%	50.87%
287: CIRCULATORY DISORDERS EXCEPT AMI, W CARD CATH W/O MCC	49	0.67%	51.54%
560: AFTERCARE, MUSCULOSKELETAL SYSTEM & CONNECTIVE TISSUE W CC	48	0.66%	52.20%
853: INFECTIOUS & PARASITIC DISEASES W O.R. PROCEDURE W MCC	47	0.65%	52.85%
178: RESPIRATORY INFECTIONS & INFLAMMATIONS W CC	47	0.65%	53.50%
293: HEART FAILURE & SHOCK W/O CC/MCC	45	0.62%	54.12%
638: DIABETES W CC	45	0.62%	54.74%
310: CARDIAC ARRHYTHMIA & CONDUCTION DISORDERS W/O CC/MCC	45	0.62%	55.36%
313: CHEST PAIN	45	0.62%	55.98%
640: MISC DISORDERS OF NUTRITION,METABOLISM,FLUIDS/ELECTROLYTES W MCC	44	0.61%	56.59%
92: OTHER DISORDERS OF NERVOUS SYSTEM W CC	42	0.58%	57.17%
483: MAJOR JOINT & LIMB REATTACHMENT PROC OF UPPER EXTREMITY W CC/MCC	41	0.56%	57.73%
280: ACUTE MYOCARDIAL INFARCTION, DISCHARGED ALIVE W MCC	41	0.56%	58.29%
247: PERC CARDIOVASC PROC W DRUG-ELUTING STENT W/O MCC	41	0.56%	58.85%

305: HYPERTENSION W/O MCC	40	0.55%	59.40%
195: SIMPLE PNEUMONIA & PLEURISY W/O CC/MCC	39	0.54%	59.94%
699: OTHER KIDNEY & URINARY TRACT DIAGNOSES W CC	39	0.54%	60.48%
190: CHRONIC OBSTRUCTIVE PULMONARY DISEASE W MCC	39	0.54%	61.02%
602: CELLULITIS W MCC	37	0.51%	61.53%
377: G.I. HEMORRHAGE W MCC	37	0.51%	62.04%
391: ESOPHAGITIS, GASTROENT & MISC DIGEST DISORDERS W MCC	36	0.50%	62.54%
100: SEIZURES W MCC	36	0.50%	63.04%
390: G.I. OBSTRUCTION W/O CC/MCC	35	0.48%	63.52%
64: INTRACRANIAL HEMORRHAGE OR CEREBRAL INFARCTION W MCC	35	0.48%	64.00%
281: ACUTE MYOCARDIAL INFARCTION, DISCHARGED ALIVE W CC	34	0.47%	64.47%
394: OTHER DIGESTIVE SYSTEM DIAGNOSES W CC	33	0.45%	64.92%
191: CHRONIC OBSTRUCTIVE PULMONARY DISEASE W CC	30	0.41%	65.33%
536: FRACTURES OF HIP & PELVIS W/O MCC	30	0.41%	65.74%
300: PERIPHERAL VASCULAR DISORDERS W CC	30	0.41%	66.15%
372: MAJOR GASTROINTESTINAL DISORDERS & PERITONEAL INFECTIONS W CC	29	0.40%	66.55%
442: DISORDERS OF LIVER EXCEPT MALIG,CIRR,ALC HEPA W CC	29	0.40%	66.95%
460: SPINAL FUSION EXCEPT CERVICAL W/O MCC	29	0.40%	67.35%
884: ORGANIC DISTURBANCES & MENTAL RETARDATION	29	0.40%	67.75%
91: OTHER DISORDERS OF NERVOUS SYSTEM W MCC	28	0.39%	68.14%
286: CIRCULATORY DISORDERS EXCEPT AMI, W CARD CATH W MCC	27	0.37%	68.51%
393: OTHER DIGESTIVE SYSTEM DIAGNOSES W MCC	27	0.37%	68.88%
74: CRANIAL & PERIPHERAL NERVE DISORDERS W/O MCC	27	0.37%	69.25%
811: RED BLOOD CELL DISORDERS W MCC	27	0.37%	69.62%
605: TRAUMA TO THE SKIN, SUBCUT TISS & BREAST W/O MCC	26	0.36%	69.98%
388: G.I. OBSTRUCTION W MCC	26	0.36%	70.34%
203: BRONCHITIS & ASTHMA W/O CC/MCC	26	0.36%	70.70%
563: FX, SPRN, STRN & DISL EXCEPT FEMUR, HIP, PELVIS & THIGH W/O MCC	25	0.34%	71.04%
556: SIGNS & SYMPTOMS OF MUSCULOSKELETAL SYSTEM & CONN TISSUE W/O MCC	25	0.34%	71.38%
243: PERMANENT CARDIAC PACEMAKER IMPLANT W CC	25	0.34%	71.72%
637: DIABETES W MCC	24	0.33%	72.05%
315: OTHER CIRCULATORY SYSTEM DIAGNOSES W CC	24	0.33%	72.38%
949: AFTERCARE W CC/MCC	24	0.33%	72.71%

86: TRAUMATIC STUPOR & COMA, COMA <1 HR W CC	23	0.32%	73.03%
242: PERMANENT CARDIAC PACEMAKER IMPLANT W MCC	23	0.32%	73.35%
870: SEPTICEMIA OR SEVERE SEPSIS W MV 96+ HOURS	22	0.30%	73.65%
330: MAJOR SMALL & LARGE BOWEL PROCEDURES W CC	22	0.30%	73.95%
71: NONSPECIFIC CEREBROVASCULAR DISORDERS W CC	22	0.30%	74.25%
149: DYSEQUILIBRIUM	21	0.29%	74.54%
554: BONE DISEASES & ARTHROPATHIES W/O MCC	21	0.29%	74.83%
66: INTRACRANIAL HEMORRHAGE OR CEREBRAL INFARCTION W/O CC/MCC	20	0.28%	75.11%
207: RESPIRATORY SYSTEM DIAGNOSIS W VENTILATOR SUPPORT 96+ HOURS	20	0.28%	75.39%
439: DISORDERS OF PANCREAS EXCEPT MALIGNANCY W CC	20	0.28%	75.67%
493: LOWER EXTREM & HUMER PROC EXCEPT HIP,FOOT,FEMUR W CC	20	0.28%	75.95%
314: OTHER CIRCULATORY SYSTEM DIAGNOSES W MCC	19	0.26%	76.21%
329: MAJOR SMALL & LARGE BOWEL PROCEDURES W MCC	19	0.26%	76.47%
480: HIP & FEMUR PROCEDURES EXCEPT MAJOR JOINT W MCC	19	0.26%	76.73%
208: RESPIRATORY SYSTEM DIAGNOSIS W VENTILATOR SUPPORT <96 HOURS	19	0.26%	76.99%
56: DEGENERATIVE NERVOUS SYSTEM DISORDERS W MCC	19	0.26%	77.25%
175: PULMONARY EMBOLISM W MCC	19	0.26%	77.51%
981: EXTENSIVE O.R. PROCEDURE UNRELATED TO PRINCIPAL DIAGNOSIS W MCC	19	0.26%	77.77%
379: G.I. HEMORRHAGE W/O CC/MCC	18	0.25%	78.02%
684: RENAL FAILURE W/O CC/MCC	18	0.25%	78.27%
854: INFECTIOUS & PARASITIC DISEASES W O.R. PROCEDURE W CC	17	0.23%	78.50%
885: PSYCHOSES	17	0.23%	78.73%
558: TENDONITIS, MYOSITIS & BURSITIS W/O MCC	17	0.23%	78.96%
561: AFTERCARE, MUSCULOSKELETAL SYSTEM & CONNECTIVE TISSUE W/O CC/MCC	17	0.23%	79.19%
467: REVISION OF HIP OR KNEE REPLACEMENT W CC	17	0.23%	79.42%
303: ATHEROSCLEROSIS W/O MCC	17	0.23%	79.65%
176: PULMONARY EMBOLISM W/O MCC	16	0.22%	79.87%
445: DISORDERS OF THE BILIARY TRACT W CC	16	0.22%	80.09%
813: COAGULATION DISORDERS	16	0.22%	80.31%
482: HIP & FEMUR PROCEDURES EXCEPT MAJOR JOINT W/O CC/MCC	16	0.22%	80.53%
864: FEVER	15	0.21%	80.74%
516: OTHER MUSCULOSKELET SYS & CONN TISS O.R. PROC W CC	15	0.21%	80.95%

253: OTHER VASCULAR PROCEDURES W CC	15	0.21%	81.16%
246: PERC CARDIOVASC PROC W DRUG-ELUTING STENT W MCC OR 4+ VESSELS/STENTS	14	0.19%	81.35%
644: ENDOCRINE DISORDERS W CC	14	0.19%	81.54%
617: AMPUTAT OF LOWER LIMB FOR ENDOCRINE,NUTRIT,& METABOL DIS W CC	14	0.19%	81.73%
70: NONSPECIFIC CEREBROVASCULAR DISORDERS W MCC	13	0.18%	81.91%
862: POSTOPERATIVE & POST-TRAUMATIC INFECTIONS W MCC	13	0.18%	82.09%
920: COMPLICATIONS OF TREATMENT W CC	13	0.18%	82.27%
639: DIABETES W/O CC/MCC	13	0.18%	82.45%
153: OTITIS MEDIA & URI W/O MCC	13	0.18%	82.63%
417: LAPAROSCOPIC CHOLECYSTECTOMY W/O C.D.E. W MCC	13	0.18%	82.81%
395: OTHER DIGESTIVE SYSTEM DIAGNOSES W/O CC/MCC	13	0.18%	82.99%
93: OTHER DISORDERS OF NERVOUS SYSTEM W/O CC/MCC	13	0.18%	83.17%
947: SIGNS & SYMPTOMS W MCC	13	0.18%	83.35%
371: MAJOR GASTROINTESTINAL DISORDERS & PERITONEAL INFECTIONS W MCC	12	0.17%	83.52%
621: O.R. PROCEDURES FOR OBESITY W/O CC/MCC	12	0.17%	83.69%
204: RESPIRATORY SIGNS & SYMPTOMS	12	0.17%	83.86%
473: CERVICAL SPINAL FUSION W/O CC/MCC	12	0.17%	84.03%
694: URINARY STONES W/O ESW LITHOTRIPSY W/O MCC	12	0.17%	84.20%
908: OTHER O.R. PROCEDURES FOR INJURIES W CC	12	0.17%	84.37%
982: EXTENSIVE O.R. PROCEDURE UNRELATED TO PRINCIPAL DIAGNOSIS W CC	12	0.17%	84.54%
381: COMPLICATED PEPTIC ULCER W CC	12	0.17%	84.71%
299: PERIPHERAL VASCULAR DISORDERS W MCC	11	0.15%	84.86%
179: RESPIRATORY INFECTIONS & INFLAMMATIONS W/O CC/MCC	11	0.15%	85.01%
282: ACUTE MYOCARDIAL INFARCTION, DISCHARGED ALIVE W/O CC/MCC	11	0.15%	85.16%
206: OTHER RESPIRATORY SYSTEM DIAGNOSES W/O MCC	11	0.15%	85.31%
419: LAPAROSCOPIC CHOLECYSTECTOMY W/O C.D.E. W/O CC/MCC	11	0.15%	85.46%
592: SKIN ULCERS W MCC	11	0.15%	85.61%
863: POSTOPERATIVE & POST-TRAUMATIC INFECTIONS W/O MCC	10	0.14%	85.75%
184: MAJOR CHEST TRAUMA W CC	10	0.14%	85.89%
373: MAJOR GASTROINTESTINAL DISORDERS & PERITONEAL INFECTIONS W/O CC/MCC	10	0.14%	86.03%
103: HEADACHES W/O MCC	10	0.14%	86.17%
418: LAPAROSCOPIC CHOLECYSTECTOMY W/O C.D.E. W CC	10	0.14%	86.31%

87: TRAUMATIC STUPOR & COMA, COMA <1 HR W/O CC/MCC	10	0.14%	86.45%
472: CERVICAL SPINAL FUSION W CC	10	0.14%	86.59%
551: MEDICAL BACK PROBLEMS W MCC	10	0.14%	86.73%
386: INFLAMMATORY BOWEL DISEASE W CC	10	0.14%	86.87%
192: CHRONIC OBSTRUCTIVE PULMONARY DISEASE W/O CC/MCC	10	0.14%	87.01%
205: OTHER RESPIRATORY SYSTEM DIAGNOSES W MCC	10	0.14%	87.15%
571: SKIN DEBRIDEMENT W CC	9	0.12%	87.27%
917: POISONING & TOXIC EFFECTS OF DRUGS W MCC	9	0.12%	87.39%
857: POSTOPERATIVE OR POST-TRAUMATIC INFECTIONS W O.R. PROC W CC	9	0.12%	87.51%
987: NON-EXTENSIVE O.R. PROC UNRELATED TO PRINCIPAL DIAGNOSIS W MCC	9	0.12%	87.63%
444: DISORDERS OF THE BILIARY TRACT W MCC	9	0.12%	87.75%
535: FRACTURES OF HIP & PELVIS W MCC	9	0.12%	87.87%
543: PATHOLOGICAL FRACTURES & MUSCULOSKELET & CONN TISS MALIG W CC	8	0.11%	87.98%
440: DISORDERS OF PANCREAS EXCEPT MALIGNANCY W/O CC/MCC	8	0.11%	88.09%
166: OTHER RESP SYSTEM O.R. PROCEDURES W MCC	8	0.11%	88.20%
593: SKIN ULCERS W CC	8	0.11%	88.31%
446: DISORDERS OF THE BILIARY TRACT W/O CC/MCC	8	0.11%	88.42%
155: OTHER EAR, NOSE, MOUTH & THROAT DIAGNOSES W CC	8	0.11%	88.53%
454: COMBINED ANTERIOR/POSTERIOR SPINAL FUSION W CC	8	0.11%	88.64%
559: AFTERCARE, MUSCULOSKELETAL SYSTEM & CONNECTIVE TISSUE W MCC	8	0.11%	88.75%
384: UNCOMPLICATED PEPTIC ULCER W/O MCC	8	0.11%	88.86%
328: STOMACH, ESOPHAGEAL & DUODENAL PROC W/O CC/MCC	8	0.11%	88.97%
264: OTHER CIRCULATORY SYSTEM O.R. PROCEDURES	8	0.11%	89.08%
565: OTHER MUSCULOSKELETAL SYS & CONNECTIVE TISSUE DIAGNOSES W CC	8	0.11%	89.19%
660: KIDNEY & URETER PROCEDURES FOR NON-NEOPLASM W CC	8	0.11%	89.30%
945: REHABILITATION W CC/MCC	8	0.11%	89.41%
515: OTHER MUSCULOSKELET SYS & CONN TISS O.R. PROC W MCC	8	0.11%	89.52%
502: SOFT TISSUE PROCEDURES W/O CC/MCC	7	0.10%	89.62%
669: TRANSURETHRAL PROCEDURES W CC	7	0.10%	89.72%
880: ACUTE ADJUSTMENT REACTION & PSYCHOSOCIAL DYSFUNCTION	7	0.10%	89.82%
643: ENDOCRINE DISORDERS W MCC	7	0.10%	89.92%
441: DISORDERS OF LIVER EXCEPT MALIG,CIRR,ALC HEPA W MCC	7	0.10%	90.02%

62: ACUTE ISCHEMIC STROKE W USE OF THROMBOLYTIC AGENT W CC	7	0.10%	90.12%
229: OTHER CARDIOTHORACIC PROCEDURES W CC	7	0.10%	90.22%
40: PERIPH/CRANIAL NERVE & OTHER NERV SYST PROC W MCC	7	0.10%	90.32%
234: CORONARY BYPASS W CARDIAC CATH W/O MCC	6	0.08%	90.40%
357: OTHER DIGESTIVE SYSTEM O.R. PROCEDURES W CC	6	0.08%	90.48%
607: MINOR SKIN DISORDERS W/O MCC	6	0.08%	90.56%
348: ANAL & STOMAL PROCEDURES W CC	6	0.08%	90.64%
539: OSTEOMYELITIS W MCC	6	0.08%	90.72%
311: ANGINA PECTORIS	6	0.08%	90.80%
562: FX, SPRN, STRN & DISL EXCEPT FEMUR, HIP, PELVIS & THIGH W MCC	6	0.08%	90.88%
468: REVISION OF HIP OR KNEE REPLACEMENT W/O CC/MCC	6	0.08%	90.96%
534: FRACTURES OF FEMUR W/O MCC	6	0.08%	91.04%
307: CARDIAC CONGENITAL & VALVULAR DISORDERS W/O MCC	6	0.08%	91.12%
261: CARDIAC PACEMAKER REVISION EXCEPT DEVICE REPLACEMENT W CC	6	0.08%	91.20%
301: PERIPHERAL VASCULAR DISORDERS W/O CC/MCC	6	0.08%	91.28%
39: EXTRACRANIAL PROCEDURES W/O CC/MCC	6	0.08%	91.36%
856: POSTOPERATIVE OR POST-TRAUMATIC INFECTIONS W O.R. PROC W MCC	6	0.08%	91.44%
897: ALCOHOL/DRUG ABUSE OR DEPENDENCE W/O REHABILITATION THERAPY W/O MCC	6	0.08%	91.52%
331: MAJOR SMALL & LARGE BOWEL PROCEDURES W/O CC/MCC	6	0.08%	91.60%
41: PERIPH/CRANIAL NERVE & OTHER NERV SYST PROC W CC OR PERIPH NEUROSTIM	6	0.08%	91.68%
327: STOMACH, ESOPHAGEAL & DUODENAL PROC W CC	6	0.08%	91.76%
700: OTHER KIDNEY & URINARY TRACT DIAGNOSES W/O CC/MCC	6	0.08%	91.84%
433: CIRRHOSIS & ALCOHOLIC HEPATITIS W CC	6	0.08%	91.92%
438: DISORDERS OF PANCREAS EXCEPT MALIGNANCY W MCC	6	0.08%	92.00%
326: STOMACH, ESOPHAGEAL & DUODENAL PROC W MCC	6	0.08%	92.08%
183: MAJOR CHEST TRAUMA W MCC	6	0.08%	92.16%
964: OTHER MULTIPLE SIGNIFICANT TRAUMA W CC	6	0.08%	92.24%
988: NON-EXTENSIVE O.R. PROC UNRELATED TO PRINCIPAL DIAGNOSIS W CC	6	0.08%	92.32%
254: OTHER VASCULAR PROCEDURES W/O CC/MCC	5	0.07%	92.39%
83: TRAUMATIC STUPOR & COMA, COMA >1 HR W CC	5	0.07%	92.46%
354: HERNIA PROCEDURES EXCEPT INGUINAL & FEMORAL W CC	5	0.07%	92.53%
252: OTHER VASCULAR PROCEDURES W MCC	5	0.07%	92.60%

758: INFECTIONS, FEMALE REPRODUCTIVE SYSTEM W CC	5	0.07%	92.67%
950: AFTERCARE W/O CC/MCC	5	0.07%	92.74%
544: PATHOLOGICAL FRACTURES & MUSCULOSKELET & CONN TISS MALIG W/O CC/MCC	5	0.07%	92.81%
907: OTHER O.R. PROCEDURES FOR INJURIES W MCC	5	0.07%	92.88%
432: CIRRHOSIS & ALCOHOLIC HEPATITIS W MCC	5	0.07%	92.95%
443: DISORDERS OF LIVER EXCEPT MALIG,CIRR,ALC HEPA W/O CC/MCC	5	0.07%	93.02%
1: HEART TRANSPLANT OR IMPLANT OF HEART ASSIST SYSTEM W MCC	5	0.07%	93.09%
696: KIDNEY & URINARY TRACT SIGNS & SYMPTOMS W/O MCC	5	0.07%	93.16%
918: POISONING & TOXIC EFFECTS OF DRUGS W/O MCC	5	0.07%	93.23%
463: WND DEBRID & SKN GRFT EXC HAND, FOR MUSCULO-CONN TISS DIS W MCC	5	0.07%	93.30%
464: WND DEBRID & SKN GRFT EXC HAND, FOR MUSCULO-CONN TISS DIS W CC	5	0.07%	93.37%
244: PERMANENT CARDIAC PACEMAKER IMPLANT W/O CC/MCC	5	0.07%	93.44%
167: OTHER RESP SYSTEM O.R. PROCEDURES W CC	5	0.07%	93.51%
596: MAJOR SKIN DISORDERS W/O MCC	5	0.07%	93.58%
570: SKIN DEBRIDEMENT W MCC	5	0.07%	93.65%
494: LOWER EXTREM & HUMER PROC EXCEPT HIP,FOOT,FEMUR W/O CC/MCC	5	0.07%	93.72%
564: OTHER MUSCULOSKELETAL SYS & CONNECTIVE TISSUE DIAGNOSES W MCC	5	0.07%	93.79%
196: INTERSTITIAL LUNG DISEASE W MCC	5	0.07%	93.86%
623: SKIN GRAFTS & WOUND DEBRID FOR ENDOC, NUTRIT & METAB DIS W CC	5	0.07%	93.93%
546: CONNECTIVE TISSUE DISORDERS W CC	5	0.07%	94.00%
200: PNEUMOTHORAX W CC	5	0.07%	94.07%
540: OSTEOMYELITIS W CC	5	0.07%	94.14%
304: HYPERTENSION W MCC	4	0.06%	94.20%
475: AMPUTATION FOR MUSCULOSKELETAL SYS & CONN TISSUE DIS W CC	4	0.06%	94.26%
919: COMPLICATIONS OF TREATMENT W MCC	4	0.06%	94.32%
501: SOFT TISSUE PROCEDURES W CC	4	0.06%	94.38%
940: O.R. PROC W DIAGNOSES OF OTHER CONTACT W HEALTH SERVICES W CC	4	0.06%	94.44%
504: FOOT PROCEDURES W CC	4	0.06%	94.50%
457: SPINAL FUS EXC CERV W SPINAL CURV/MALIG/INFEC OR 9+ FUS W CC	4	0.06%	94.56%
316: OTHER CIRCULATORY SYSTEM DIAGNOSES W/O CC/MCC	4	0.06%	94.62%

914: TRAUMATIC INJURY W/O MCC	4	0.06%	94.68%
336: PERITONEAL ADHESIOLYSIS W CC	4	0.06%	94.74%
60: MULTIPLE SCLEROSIS & CEREBELLAR ATAXIA W/O CC/MCC	4	0.06%	94.80%
726: BENIGN PROSTATIC HYPERTROPHY W/O MCC	4	0.06%	94.86%
256: UPPER LIMB & TOE AMPUTATION FOR CIRC SYSTEM DISORDERS W CC	4	0.06%	94.92%
199: PNEUMOTHORAX W MCC	4	0.06%	94.98%
517: OTHER MUSCULOSKELET SYS & CONN TISS O.R. PROC W/O CC/MCC	4	0.06%	95.04%
553: BONE DISEASES & ARTHROPATHIES W MCC	4	0.06%	95.10%
757: INFECTIONS, FEMALE REPRODUCTIVE SYSTEM W MCC	4	0.06%	95.16%
866: VIRAL ILLNESS W/O MCC	4	0.06%	95.22%
227: CARDIAC DEFIBRILLATOR IMPLANT W/O CARDIAC CATH W/O MCC	4	0.06%	95.28%
151: EPISTAXIS W/O MCC	4	0.06%	95.34%
248: PERC CARDIOVASC PROC W NON-DRUG-ELUTING STENT W MCC OR 4+ VES/STENTS	4	0.06%	95.40%
604: TRAUMA TO THE SKIN, SUBCUT TISS & BREAST W MCC	4	0.06%	95.46%
620: O.R. PROCEDURES FOR OBESITY W CC	4	0.06%	95.52%
59: MULTIPLE SCLEROSIS & CEREBELLAR ATAXIA W CC	4	0.06%	95.58%
345: MINOR SMALL & LARGE BOWEL PROCEDURES W CC	4	0.06%	95.64%
351: INGUINAL & FEMORAL HERNIA PROCEDURES W CC	4	0.06%	95.70%
478: BIOPSIES OF MUSCULOSKELETAL SYSTEM & CONNECTIVE TISSUE W CC	4	0.06%	95.76%
73: CRANIAL & PERIPHERAL NERVE DISORDERS W MCC	3	0.04%	95.80%
27: CRANIOTOMY & ENDOVASCULAR INTRACRANIAL PROCEDURES W/O CC/MCC	3	0.04%	95.84%
250: PERC CARDIOVASC PROC W/O CORONARY ARTERY STENT W MCC	3	0.04%	95.88%
302: ATHEROSCLEROSIS W MCC	3	0.04%	95.92%
353: HERNIA PROCEDURES EXCEPT INGUINAL & FEMORAL W MCC	3	0.04%	95.96%
355: HERNIA PROCEDURES EXCEPT INGUINAL & FEMORAL W/O CC/MCC	3	0.04%	96.00%
356: OTHER DIGESTIVE SYSTEM O.R. PROCEDURES W MCC	3	0.04%	96.04%
228: OTHER CARDIOTHORACIC PROCEDURES W MCC	3	0.04%	96.08%
37: EXTRACRANIAL PROCEDURES W MCC	3	0.04%	96.12%
380: COMPLICATED PEPTIC ULCER W MCC	3	0.04%	96.16%
387: INFLAMMATORY BOWEL DISEASE W/O CC/MCC	3	0.04%	96.20%
226: CARDIAC DEFIBRILLATOR IMPLANT W/O CARDIAC CATH W MCC	3	0.04%	96.24%

38: EXTRACRANIAL PROCEDURES W CC	3	0.04%	96.28%
3: ECMO OR TRACH W MV 96+ HRS OR PDX EXC FACE, MOUTH & NECK W MAJ O.R.	3	0.04%	96.32%
408: BILIARY TRACT PROC EXCEPT ONLY CHOLECYST W OR W/O C.D.E. W MCC	3	0.04%	96.36%
414: CHOLECYSTECTOMY EXCEPT BY LAPAROSCOPE W/O C.D.E. W MCC	3	0.04%	96.40%
416: CHOLECYSTECTOMY EXCEPT BY LAPAROSCOPE W/O C.D.E. W/O CC/MCC	3	0.04%	96.44%
42: PERIPH/CRANIAL NERVE & OTHER NERV SYST PROC W/O CC/MCC	3	0.04%	96.48%
459: SPINAL FUSION EXCEPT CERVICAL W MCC	3	0.04%	96.52%
466: REVISION OF HIP OR KNEE REPLACEMENT W MCC	3	0.04%	96.56%
486: KNEE PROCEDURES W PDX OF INFECTION W CC	3	0.04%	96.60%
496: LOCAL EXCISION & REMOVAL INT FIX DEVICES EXC HIP & FEMUR W CC	3	0.04%	96.64%
4: TRACH W MV 96+ HRS OR PDX EXC FACE, MOUTH & NECK W/O MAJ O.R.	3	0.04%	96.68%
52: SPINAL DISORDERS & INJURIES W CC/MCC	3	0.04%	96.72%
542: PATHOLOGICAL FRACTURES & MUSCULOSKELET & CONN TISS MALIG W MCC	3	0.04%	96.76%
549: SEPTIC ARTHRITIS W CC	3	0.04%	96.80%
555: SIGNS & SYMPTOMS OF MUSCULOSKELETAL SYSTEM & CONN TISSUE W MCC	3	0.04%	96.84%
580: OTHER SKIN, SUBCUT TISS & BREAST PROC W CC	3	0.04%	96.88%
629: OTHER ENDOCRINE, NUTRIT & METAB O.R. PROC W CC	3	0.04%	96.92%
645: ENDOCRINE DISORDERS W/O CC/MCC	3	0.04%	96.96%
68: NONSPECIFIC CVA & PRECEREBRAL OCCLUSION W/O INFARCT W/O MCC	3	0.04%	97.00%
742: UTERINE & ADNEXA PROC FOR NON-MALIGNANCY W CC/MCC	3	0.04%	97.04%
78: HYPERTENSIVE ENCEPHALOPATHY W CC	3	0.04%	97.08%
84: TRAUMATIC STUPOR & COMA, COMA >1 HR W/O CC/MCC	3	0.04%	97.12%
85: TRAUMATIC STUPOR & COMA, COMA <1 HR W MCC	3	0.04%	97.16%
865: VIRAL ILLNESS W MCC	3	0.04%	97.20%
882: NEUROSES EXCEPT DEPRESSIVE	3	0.04%	97.24%
909: OTHER O.R. PROCEDURES FOR INJURIES W/O CC/MCC	3	0.04%	97.28%
915: ALLERGIC REACTIONS W MCC	3	0.04%	97.32%
916: ALLERGIC REACTIONS W/O MCC	3	0.04%	97.36%
946: REHABILITATION W/O CC/MCC	3	0.04%	97.40%
977: HIV W OR W/O OTHER RELATED CONDITION	3	0.04%	97.44%

983: EXTENSIVE O.R. PROCEDURE UNRELATED TO PRINCIPAL DIAGNOSIS W/O CC/MCC	3	0.04%	97.48%
743: UTERINE & ADNEXA PROC FOR NON-MALIGNANCY W/O CC/MCC	2	0.03%	97.51%
673: OTHER KIDNEY & URINARY TRACT PROCEDURES W MCC	2	0.03%	97.54%
547: CONNECTIVE TISSUE DISORDERS W/O CC/MCC	2	0.03%	97.57%
881: DEPRESSIVE NEUROSES	2	0.03%	97.60%
216: CARDIAC VALVE & OTH MAJ CARDIOTHORACIC PROC W CARD CATH W MCC	2	0.03%	97.63%
217: CARDIAC VALVE & OTH MAJ CARDIOTHORACIC PROC W CARD CATH W CC	2	0.03%	97.66%
479: BIOPSIES OF MUSCULOSKELETAL SYSTEM & CONNECTIVE TISSUE W/O CC/MCC	2	0.03%	97.69%
628: OTHER ENDOCRINE, NUTRIT & METAB O.R. PROC W MCC	2	0.03%	97.72%
956: LIMB REATTACHMENT, HIP & FEMUR PROC FOR MULTIPLE SIGNIFICANT TRAUMA	2	0.03%	97.75%
158: DENTAL & ORAL DISEASES W CC	2	0.03%	97.78%
497: LOCAL EXCISION & REMOVAL INT FIX DEVICES EXC HIP & FEMUR W/O CC/MCC	2	0.03%	97.81%
545: CONNECTIVE TISSUE DISORDERS W MCC	2	0.03%	97.84%
654: MAJOR BLADDER PROCEDURES W CC	2	0.03%	97.87%
465: WND DEBRID & SKN GRFT EXC HAND, FOR MUSCULO-CONN TISS DIS W/O CC/MCC	2	0.03%	97.90%
335: PERITONEAL ADHESIOLYSIS W MCC	2	0.03%	97.93%
32: VENTRICULAR SHUNT PROCEDURES W CC	2	0.03%	97.96%
262: CARDIAC PACEMAKER REVISION EXCEPT DEVICE REPLACEMENT W/O CC/MCC	2	0.03%	97.99%
511: SHOULDER,ELBOW OR FOREARM PROC,EXC MAJOR JOINT PROC W CC	2	0.03%	98.02%
455: COMBINED ANTERIOR/POSTERIOR SPINAL FUSION W/O CC/MCC	2	0.03%	98.05%
99: NON-BACTERIAL INFECT OF NERVOUS SYS EXC VIRAL MENINGITIS W/O CC/MCC	2	0.03%	98.08%
156: OTHER EAR, NOSE, MOUTH & THROAT DIAGNOSES W/O CC/MCC	2	0.03%	98.11%
453: COMBINED ANTERIOR/POSTERIOR SPINAL FUSION W MCC	2	0.03%	98.14%
233: CORONARY BYPASS W CARDIAC CATH W MCC	2	0.03%	98.17%
154: OTHER EAR, NOSE, MOUTH & THROAT DIAGNOSES W MCC	2	0.03%	98.20%
492: LOWER EXTREM & HUMER PROC EXCEPT HIP,FOOT,FEMUR W MCC	2	0.03%	98.23%
974: HIV W MAJOR RELATED CONDITION W MCC	2	0.03%	98.26%
923: OTHER INJURY, POISONING & TOXIC EFFECT DIAG W/O MCC	2	0.03%	98.29%
533: FRACTURES OF FEMUR W MCC	2	0.03%	98.32%

668: TRANSURETHRAL PROCEDURES W MCC	2	0.03%	98.35%
36: CAROTID ARTERY STENT PROCEDURE W/O CC/MCC	2	0.03%	98.38%
369: MAJOR ESOPHAGEAL DISORDERS W CC	2	0.03%	98.41%
29: SPINAL PROCEDURES W CC OR SPINAL NEUROSTIMULATORS	2	0.03%	98.44%
415: CHOLECYSTECTOMY EXCEPT BY LAPAROSCOPE W/O C.D.E. W CC	2	0.03%	98.47%
258: CARDIAC PACEMAKER DEVICE REPLACEMENT W MCC	2	0.03%	98.50%
24: CRANIO W MAJOR DEV IMPL/ACUTE COMPLEX CNS PDX W/O MCC	2	0.03%	98.53%
98: NON-BACTERIAL INFECT OF NERVOUS SYS EXC VIRAL MENINGITIS W CC	2	0.03%	98.56%
61: ACUTE ISCHEMIC STROKE W USE OF THROMBOLYTIC AGENT W MCC	2	0.03%	98.59%
485: KNEE PROCEDURES W PDX OF INFECTION W MCC	2	0.03%	98.62%
344: MINOR SMALL & LARGE BOWEL PROCEDURES W MCC	2	0.03%	98.65%
579: OTHER SKIN, SUBCUT TISS & BREAST PROC W MCC	2	0.03%	98.68%
728: INFLAMMATION OF THE MALE REPRODUCTIVE SYSTEM W/O MCC	2	0.03%	98.71%
343: APPENDECTOMY W/O COMPLICATED PRINCIPAL DIAG W/O CC/MCC	2	0.03%	98.74%
72: NONSPECIFIC CEREBROVASCULAR DISORDERS W/O CC/MCC	2	0.03%	98.77%
868: OTHER INFECTIOUS & PARASITIC DISEASES DIAGNOSES W CC	2	0.03%	98.80%
405: PANCREAS, LIVER & SHUNT PROCEDURES W MCC	2	0.03%	98.83%
566: OTHER MUSCULOSKELETAL SYS & CONNECTIVE TISSUE DIAGNOSES W/O CC/MCC	2	0.03%	98.86%
186: PLEURAL EFFUSION W MCC	2	0.03%	98.89%
197: INTERSTITIAL LUNG DISEASE W CC	2	0.03%	98.92%
557: TENDONITIS, MYOSITIS & BURSITIS W MCC	2	0.03%	98.95%
138: MOUTH PROCEDURES W/O CC/MCC	2	0.03%	98.98%
342: APPENDECTOMY W/O COMPLICATED PRINCIPAL DIAG W CC	2	0.03%	99.01%
251: PERC CARDIOVASC PROC W/O CORONARY ARTERY STENT W/O MCC	2	0.03%	99.04%
869: OTHER INFECTIOUS & PARASITIC DISEASES DIAGNOSES W/O CC/MCC	1	0.01%	99.05%
125: OTHER DISORDERS OF THE EYE W/O MCC	1	0.01%	99.06%
124: OTHER DISORDERS OF THE EYE W MCC	1	0.01%	99.07%
341: APPENDECTOMY W/O COMPLICATED PRINCIPAL DIAG W MCC	1	0.01%	99.08%
339: APPENDECTOMY W COMPLICATED PRINCIPAL DIAG W CC	1	0.01%	99.09%
26: CRANIOTOMY & ENDOVASCULAR INTRACRANIAL PROCEDURES W CC	1	0.01%	99.10%
123: NEUROLOGICAL EYE DISORDERS	1	0.01%	99.11%

338: APPENDECTOMY W COMPLICATED PRINCIPAL DIAG W MCC	1	0.01%	99.12%
88: CONCUSSION W MCC	1	0.01%	99.13%
895: ALCOHOL/DRUG ABUSE OR DEPENDENCE W REHABILITATION THERAPY	1	0.01%	99.14%
337: PERITONEAL ADHESIOLYSIS W/O CC/MCC	1	0.01%	99.15%
89: CONCUSSION W CC	1	0.01%	99.16%
901: WOUND DEBRIDEMENTS FOR INJURIES W MCC	1	0.01%	99.17%
902: WOUND DEBRIDEMENTS FOR INJURIES W CC	1	0.01%	99.18%
906: HAND PROCEDURES FOR INJURIES	1	0.01%	99.19%
90: CONCUSSION W/O CC/MCC	1	0.01%	99.20%
31: VENTRICULAR SHUNT PROCEDURES W MCC	1	0.01%	99.21%
25: CRANIOTOMY & ENDOVASCULAR INTRACRANIAL PROCEDURES W MCC	1	0.01%	99.22%
232: CORONARY BYPASS W PTCA W/O MCC	1	0.01%	99.23%
23: CRANIO W MAJOR DEV IMPL/ACUTE COMPLEX CNS PDX W MCC OR CHEMO IMPLANT	1	0.01%	99.24%
11: TRACHEOSTOMY FOR FACE,MOUTH & NECK DIAGNOSES W MCC	1	0.01%	99.25%
245: AICD GENERATOR PROCEDURES	1	0.01%	99.26%
929: FULL THICKNESS BURN W SKIN GRAFT OR INHAL INJ W/O CC/MCC	1	0.01%	99.27%
117: INTRAOCULAR PROCEDURES W/O CC/MCC	1	0.01%	99.28%
935: NON-EXTENSIVE BURNS	1	0.01%	99.29%
306: CARDIAC CONGENITAL & VALVULAR DISORDERS W MCC	1	0.01%	99.30%
249: PERC CARDIOVASC PROC W NON-DRUG-ELUTING STENT W/O MCC	1	0.01%	99.31%
28: SPINAL PROCEDURES W MCC	1	0.01%	99.32%
102: HEADACHES W MCC	1	0.01%	99.33%
584: BREAST BIOPSY, LOCAL EXCISION & OTHER BREAST PROCEDURES W CC/MCC	1	0.01%	99.34%
581: OTHER SKIN, SUBCUT TISS & BREAST PROC W/O CC/MCC	1	0.01%	99.35%
594: SKIN ULCERS W/O CC/MCC	1	0.01%	99.36%
578: SKIN GRAFT EXC FOR SKIN ULCER OR CELLULITIS W/O CC/MCC	1	0.01%	99.37%
577: SKIN GRAFT EXC FOR SKIN ULCER OR CELLULITIS W CC	1	0.01%	99.38%
188: PLEURAL EFFUSION W/O CC/MCC	1	0.01%	99.39%
187: PLEURAL EFFUSION W CC	1	0.01%	99.40%
614: ADRENAL & PITUITARY PROCEDURES W CC/MCC	1	0.01%	99.41%
616: AMPUTAT OF LOWER LIMB FOR ENDOCRINE,NUTRIT,& METABOL DIS W MCC	1	0.01%	99.42%

619: O.R. PROCEDURES FOR OBESITY W MCC	1	0.01%	99.43%
548: SEPTIC ARTHRITIS W MCC	1	0.01%	99.44%
624: SKIN GRAFTS & WOUND DEBRID FOR ENDOC, NUTRIT & METAB DIS W/O CC/MCC	1	0.01%	99.45%
626: THYROID, PARATHYROID & THYROGLOSSAL PROCEDURES W CC	1	0.01%	99.46%
289: ACUTE & SUBACUTE ENDOCARDITIS W CC	1	0.01%	99.47%
514: HAND OR WRIST PROC, EXCEPT MAJOR THUMB OR JOINT PROC W/O CC/MCC	1	0.01%	99.48%
185: MAJOR CHEST TRAUMA W/O CC/MCC	1	0.01%	99.49%
513: HAND OR WRIST PROC, EXCEPT MAJOR THUMB OR JOINT PROC W CC/MCC	1	0.01%	99.50%
512: SHOULDER,ELBOW OR FOREARM PROC,EXC MAJOR JOINT PROC W/O CC/MCC	1	0.01%	99.51%
63: ACUTE ISCHEMIC STROKE W USE OF THROMBOLYTIC AGENT W/O CC/MCC	1	0.01%	99.52%
507: MAJOR SHOULDER OR ELBOW JOINT PROCEDURES W CC/MCC	1	0.01%	99.53%
642: INBORN AND OTHER DISORDERS OF METABOLISM	1	0.01%	99.54%
505: FOOT PROCEDURES W/O CC/MCC	1	0.01%	99.55%
500: SOFT TISSUE PROCEDURES W MCC	1	0.01%	99.56%
653: MAJOR BLADDER PROCEDURES W MCC	1	0.01%	99.57%
495: LOCAL EXCISION & REMOVAL INT FIX DEVICES EXC HIP & FEMUR W MCC	1	0.01%	99.58%
661: KIDNEY & URETER PROCEDURES FOR NON-NEOPLASM W/O CC/MCC	1	0.01%	99.59%
489: KNEE PROCEDURES W/O PDX OF INFECTION W/O CC/MCC	1	0.01%	99.60%
674: OTHER KIDNEY & URINARY TRACT PROCEDURES W CC	1	0.01%	99.61%
476: AMPUTATION FOR MUSCULOSKELETAL SYS & CONN TISSUE DIS W/O CC/MCC	1	0.01%	99.62%
474: AMPUTATION FOR MUSCULOSKELETAL SYS & CONN TISSUE DIS W MCC	1	0.01%	99.63%
95: BACTERIAL & TUBERCULOUS INFECTIONS OF NERVOUS SYSTEM W CC	1	0.01%	99.64%
157: DENTAL & ORAL DISEASES W MCC	1	0.01%	99.65%
693: URINARY STONES W/O ESW LITHOTRIPSY W MCC	1	0.01%	99.66%
695: KIDNEY & URINARY TRACT SIGNS & SYMPTOMS W MCC	1	0.01%	99.67%
423: OTHER HEPATOBILIARY OR PANCREAS O.R. PROCEDURES W MCC	1	0.01%	99.68%
707: MAJOR MALE PELVIC PROCEDURES W CC/MCC	1	0.01%	99.69%
420: HEPATOBILIARY DIAGNOSTIC PROCEDURES W MCC	1	0.01%	99.70%
409: BILIARY TRACT PROC EXCEPT ONLY CHOLECYST W OR W/O C.D.E. W CC	1	0.01%	99.71%

727: INFLAMMATION OF THE MALE REPRODUCTIVE SYSTEM W MCC	1	0.01%	99.72%
748: FEMALE REPRODUCTIVE SYSTEM RECONSTRUCTIVE PROCEDURES	1	0.01%	99.73%
215: OTHER HEART ASSIST SYSTEM IMPLANT	1	0.01%	99.74%
75: VIRAL MENINGITIS W CC/MCC	1	0.01%	99.75%
761: MENSTRUAL & OTHER FEMALE REPRODUCTIVE SYSTEM DISORDERS W/O CC/MCC	1	0.01%	99.76%
77: HYPERTENSIVE ENCEPHALOPATHY W MCC	1	0.01%	99.77%
965: OTHER MULTIPLE SIGNIFICANT TRAUMA W/O CC/MCC	1	0.01%	99.78%
79: HYPERTENSIVE ENCEPHALOPATHY W/O CC/MCC	1	0.01%	99.79%
800: SPLENECTOMY W CC	1	0.01%	99.80%
802: OTHER O.R. PROC OF THE BLOOD & BLOOD FORMING ORGANS W MCC	1	0.01%	99.81%
80: NONTRAUMATIC STUPOR & COMA W MCC	1	0.01%	99.82%
810: MAJOR HEMATOL/IMMUN DIAG EXC SICKLE CELL CRISIS & COAGUL W/O CC/MCC	1	0.01%	99.83%
222: CARDIAC DEFIB IMPLANT W CARDIAC CATH W AMI/HF/SHOCK W MCC	1	0.01%	99.84%
224: CARDIAC DEFIB IMPLANT W CARDIAC CATH W/O AMI/HF/SHOCK W MCC	1	0.01%	99.85%
385: INFLAMMATORY BOWEL DISEASE W MCC	1	0.01%	99.86%
814: RETICULOENDOTHELIAL & IMMUNITY DISORDERS W MCC	1	0.01%	99.87%
382: COMPLICATED PEPTIC ULCER W/O CC/MCC	1	0.01%	99.88%
139: SALIVARY GLAND PROCEDURES	1	0.01%	99.89%
370: MAJOR ESOPHAGEAL DISORDERS W/O CC/MCC	1	0.01%	99.90%
858: POSTOPERATIVE OR POST-TRAUMATIC INFECTIONS W O.R. PROC W/O CC/MCC	1	0.01%	99.91%
35: CAROTID ARTERY STENT PROCEDURE W CC	1	0.01%	99.92%
358: OTHER DIGESTIVE SYSTEM O.R. PROCEDURES W/O CC/MCC	1	0.01%	99.93%
97: NON-BACTERIAL INFECT OF NERVOUS SYS EXC VIRAL MENINGITIS W MCC	1	0.01%	99.94%
867: OTHER INFECTIOUS & PARASITIC DISEASES DIAGNOSES W MCC	1	0.01%	99.95%

Table A4. 12-Month Feature Space XGBoost Inpatient Event Diagnosis Related Groups (DRGs)

Diagnosis Related Group (DRG)	# IP Events	% IP Events	Cumulative Feature Importance
871: SEPTICEMIA OR SEVERE SEPSIS W/O MV 96+ HOURS W MCC	22	4.54%	4.54%
291: HEART FAILURE & SHOCK W MCC	20	4.12%	8.66%
683: RENAL FAILURE W CC	13	2.68%	11.34%

392: ESOPHAGITIS, GASTROENT & MISC DIGEST DISORDERS W/O MCC	12	2.47%	13.81%
481: HIP & FEMUR PROCEDURES EXCEPT MAJOR JOINT W CC	10	2.06%	15.87%
690: KIDNEY & URINARY TRACT INFECTIONS W/O MCC	10	2.06%	17.93%
65: INTRACRANIAL HEMORRHAGE OR CEREBRAL INFARCTION W CC OR TPA IN 24 HRS	9	1.86%	19.79%
483: MAJOR JOINT & LIMB REATTACHMENT PROC OF UPPER EXTREMITY W CC/MCC	9	1.86%	21.65%
309: CARDIAC ARRHYTHMIA & CONDUCTION DISORDERS W CC	8	1.65%	23.30%
194: SIMPLE PNEUMONIA & PLEURISY W CC	8	1.65%	24.95%
292: HEART FAILURE & SHOCK W CC	8	1.65%	26.60%
378: G.I. HEMORRHAGE W CC	8	1.65%	28.25%
552: MEDICAL BACK PROBLEMS W/O MCC	8	1.65%	29.90%
281: ACUTE MYOCARDIAL INFARCTION, DISCHARGED ALIVE W CC	7	1.44%	31.34%
872: SEPTICEMIA OR SEVERE SEPSIS W/O MV 96+ HOURS W/O MCC	7	1.44%	32.78%
280: ACUTE MYOCARDIAL INFARCTION, DISCHARGED ALIVE W MCC	7	1.44%	34.22%
57: DEGENERATIVE NERVOUS SYSTEM DISORDERS W/O MCC	7	1.44%	35.66%
689: KIDNEY & URINARY TRACT INFECTIONS W MCC	7	1.44%	37.10%
394: OTHER DIGESTIVE SYSTEM DIAGNOSES W CC	6	1.24%	38.34%
641: MISC DISORDERS OF NUTRITION,METABOLISM,FLUIDS/ELECTROLYTES W/O MCC	6	1.24%	39.58%
293: HEART FAILURE & SHOCK W/O CC/MCC	6	1.24%	40.82%
243: PERMANENT CARDIAC PACEMAKER IMPLANT W CC	6	1.24%	42.06%
554: BONE DISEASES & ARTHROPATHIES W/O MCC	6	1.24%	43.30%
560: AFTERCARE, MUSCULOSKELETAL SYSTEM & CONNECTIVE TISSUE W CC	5	1.03%	44.33%
682: RENAL FAILURE W MCC	5	1.03%	45.36%
312: SYNCOPE & COLLAPSE	5	1.03%	46.39%
884: ORGANIC DISTURBANCES & MENTAL RETARDATION	5	1.03%	47.42%
389: G.I. OBSTRUCTION W CC	5	1.03%	48.45%
101: SEIZURES W/O MCC	5	1.03%	49.48%
603: CELLULITIS W/O MCC	5	1.03%	50.51%
812: RED BLOOD CELL DISORDERS W/O MCC	5	1.03%	51.54%
377: G.I. HEMORRHAGE W MCC	5	1.03%	52.57%
388: G.I. OBSTRUCTION W MCC	5	1.03%	53.60%
189: PULMONARY EDEMA & RESPIRATORY FAILURE	4	0.82%	54.42%
202: BRONCHITIS & ASTHMA W CC/MCC	4	0.82%	55.24%

308: CARDIAC ARRHYTHMIA & CONDUCTION DISORDERS W MCC	4	0.82%	56.06%
313: CHEST PAIN	4	0.82%	56.88%
372: MAJOR GASTROINTESTINAL DISORDERS & PERITONEAL INFECTIONS W CC	4	0.82%	57.70%
66: INTRACRANIAL HEMORRHAGE OR CEREBRAL INFARCTION W/O CC/MCC	4	0.82%	58.52%
482: HIP & FEMUR PROCEDURES EXCEPT MAJOR JOINT W/O CC/MCC	3	0.62%	59.14%
390: G.I. OBSTRUCTION W/O CC/MCC	3	0.62%	59.76%
62: ACUTE ISCHEMIC STROKE W USE OF THROMBOLYTIC AGENT W CC	3	0.62%	60.38%
247: PERC CARDIOVASC PROC W DRUG-ELUTING STENT W/O MCC	3	0.62%	61.00%
69: TRANSIENT ISCHEMIA	3	0.62%	61.62%
193: SIMPLE PNEUMONIA & PLEURISY W MCC	3	0.62%	62.24%
253: OTHER VASCULAR PROCEDURES W CC	3	0.62%	62.86%
195: SIMPLE PNEUMONIA & PLEURISY W/O CC/MCC	3	0.62%	63.48%
853: INFECTIOUS & PARASITIC DISEASES W O.R. PROCEDURE W MCC	3	0.62%	64.10%
242: PERMANENT CARDIAC PACEMAKER IMPLANT W MCC	3	0.62%	64.72%
282: ACUTE MYOCARDIAL INFARCTION, DISCHARGED ALIVE W/O CC/MCC	3	0.62%	65.34%
561: AFTERCARE, MUSCULOSKELETAL SYSTEM & CONNECTIVE TISSUE W/O CC/MCC	3	0.62%	65.96%
287: CIRCULATORY DISORDERS EXCEPT AMI, W CARD CATH W/O MCC	3	0.62%	66.58%
86: TRAUMATIC STUPOR & COMA, COMA <1 HR W CC	3	0.62%	67.20%
149: DYSEQUILIBRIUM	3	0.62%	67.82%
310: CARDIAC ARRHYTHMIA & CONDUCTION DISORDERS W/O CC/MCC	3	0.62%	68.44%
299: PERIPHERAL VASCULAR DISORDERS W MCC	3	0.62%	69.06%
177: RESPIRATORY INFECTIONS & INFLAMMATIONS W MCC	3	0.62%	69.68%
391: ESOPHAGITIS, GASTROENT & MISC DIGEST DISORDERS W MCC	3	0.62%	70.30%
91: OTHER DISORDERS OF NERVOUS SYSTEM W MCC	2	0.41%	70.71%
920: COMPLICATIONS OF TREATMENT W CC	2	0.41%	71.12%
439: DISORDERS OF PANCREAS EXCEPT MALIGNANCY W CC	2	0.41%	71.53%
92: OTHER DISORDERS OF NERVOUS SYSTEM W CC	2	0.41%	71.94%
460: SPINAL FUSION EXCEPT CERVICAL W/O MCC	2	0.41%	72.35%
300: PERIPHERAL VASCULAR DISORDERS W CC	2	0.41%	72.76%
305: HYPERTENSION W/O MCC	2	0.41%	73.17%
445: DISORDERS OF THE BILIARY TRACT W CC	2	0.41%	73.58%

558: TENDONITIS, MYOSITIS & BURSITIS W/O MCC	2	0.41%	73.99%
640: MISC DISORDERS OF NUTRITION,METABOLISM,FLUIDS/ELECTROLYTES W MCC	2	0.41%	74.40%
246: PERC CARDIOVASC PROC W DRUG-ELUTING STENT W MCC OR 4+ VESSELS/STENTS	2	0.41%	74.81%
190: CHRONIC OBSTRUCTIVE PULMONARY DISEASE W MCC	2	0.41%	75.22%
696: KIDNEY & URINARY TRACT SIGNS & SYMPTOMS W/O MCC	2	0.41%	75.63%
87: TRAUMATIC STUPOR & COMA, COMA <1 HR W/O CC/MCC	2	0.41%	76.04%
698: OTHER KIDNEY & URINARY TRACT DIAGNOSES W MCC	2	0.41%	76.45%
699: OTHER KIDNEY & URINARY TRACT DIAGNOSES W CC	2	0.41%	76.86%
336: PERITONEAL ADHESIOLYSIS W CC	2	0.41%	77.27%
39: EXTRACRANIAL PROCEDURES W/O CC/MCC	2	0.41%	77.68%
166: OTHER RESP SYSTEM O.R. PROCEDURES W MCC	2	0.41%	78.09%
74: CRANIAL & PERIPHERAL NERVE DISORDERS W/O MCC	2	0.41%	78.50%
948: SIGNS & SYMPTOMS W/O MCC	2	0.41%	78.91%
809: MAJOR HEMATOL/IMMUN DIAG EXC SICKLE CELL CRISIS & COAGUL W CC	2	0.41%	79.32%
175: PULMONARY EMBOLISM W MCC	2	0.41%	79.73%
605: TRAUMA TO THE SKIN, SUBCUT TISS & BREAST W/O MCC	2	0.41%	80.14%
949: AFTERCARE W CC/MCC	2	0.41%	80.55%
371: MAJOR GASTROINTESTINAL DISORDERS & PERITONEAL INFECTIONS W MCC	2	0.41%	80.96%
908: OTHER O.R. PROCEDURES FOR INJURIES W CC	1	0.21%	81.17%
914: TRAUMATIC INJURY W/O MCC	1	0.21%	81.38%
93: OTHER DISORDERS OF NERVOUS SYSTEM W/O CC/MCC	1	0.21%	81.59%
945: REHABILITATION W CC/MCC	1	0.21%	81.80%
947: SIGNS & SYMPTOMS W MCC	1	0.21%	82.01%
963: OTHER MULTIPLE SIGNIFICANT TRAUMA W MCC	1	0.21%	82.22%
982: EXTENSIVE O.R. PROCEDURE UNRELATED TO PRINCIPAL DIAGNOSIS W CC	1	0.21%	82.43%
100: SEIZURES W MCC	1	0.21%	82.64%
98: NON-BACTERIAL INFECT OF NERVOUS SYS EXC VIRAL MENINGITIS W CC	1	0.21%	82.85%
103: HEADACHES W/O MCC	1	0.21%	83.06%
137: MOUTH PROCEDURES W CC/MCC	1	0.21%	83.27%
176: PULMONARY EMBOLISM W/O MCC	1	0.21%	83.48%
184: MAJOR CHEST TRAUMA W CC	1	0.21%	83.69%
185: MAJOR CHEST TRAUMA W/O CC/MCC	1	0.21%	83.90%

192: CHRONIC OBSTRUCTIVE PULMONARY DISEASE W/O CC/MCC	1	0.21%	84.11%
199: PNEUMOTHORAX W MCC	1	0.21%	84.32%
200: PNEUMOTHORAX W CC	1	0.21%	84.53%
203: BRONCHITIS & ASTHMA W/O CC/MCC	1	0.21%	84.74%
206: OTHER RESPIRATORY SYSTEM DIAGNOSES W/O MCC	1	0.21%	84.95%
229: OTHER CARDIOTHORACIC PROCEDURES W CC	1	0.21%	85.16%
234: CORONARY BYPASS W CARDIAC CATH W/O MCC	1	0.21%	85.37%
244: PERMANENT CARDIAC PACEMAKER IMPLANT W/O CC/MCC	1	0.21%	85.58%
248: PERC CARDIOVASC PROC W NON-DRUG-ELUTING STENT W MCC OR 4+ VES/STENTS	1	0.21%	85.79%
259: CARDIAC PACEMAKER DEVICE REPLACEMENT W/O MCC	1	0.21%	86.00%
261: CARDIAC PACEMAKER REVISION EXCEPT DEVICE REPLACEMENT W CC	1	0.21%	86.21%
286: CIRCULATORY DISORDERS EXCEPT AMI, W CARD CATH W MCC	1	0.21%	86.42%
29: SPINAL PROCEDURES W CC OR SPINAL NEUROSTIMULATORS	1	0.21%	86.63%
303: ATHEROSCLEROSIS W/O MCC	1	0.21%	86.84%
315: OTHER CIRCULATORY SYSTEM DIAGNOSES W CC	1	0.21%	87.05%
328: STOMACH, ESOPHAGEAL & DUODENAL PROC W/O CC/MCC	1	0.21%	87.26%
330: MAJOR SMALL & LARGE BOWEL PROCEDURES W CC	1	0.21%	87.47%
348: ANAL & STOMAL PROCEDURES W CC	1	0.21%	87.68%
349: ANAL & STOMAL PROCEDURES W/O CC/MCC	1	0.21%	87.89%
36: CAROTID ARTERY STENT PROCEDURE W/O CC/MCC	1	0.21%	88.10%
373: MAJOR GASTROINTESTINAL DISORDERS & PERITONEAL INFECTIONS W/O CC/MCC	1	0.21%	88.31%
384: UNCOMPLICATED PEPTIC ULCER W/O MCC	1	0.21%	88.52%
395: OTHER DIGESTIVE SYSTEM DIAGNOSES W/O CC/MCC	1	0.21%	88.73%
414: CHOLECYSTECTOMY EXCEPT BY LAPAROSCOPE W/O C.D.E. W MCC	1	0.21%	88.94%
418: LAPAROSCOPIC CHOLECYSTECTOMY W/O C.D.E. W CC	1	0.21%	89.15%
41: PERIPH/CRANIAL NERVE & OTHER NERV SYST PROC W CC OR PERIPH NEUROSTIM	1	0.21%	89.36%
42: PERIPH/CRANIAL NERVE & OTHER NERV SYST PROC W/O CC/MCC	1	0.21%	89.57%
433: CIRRHOSIS & ALCOHOLIC HEPATITIS W CC	1	0.21%	89.78%
438: DISORDERS OF PANCREAS EXCEPT MALIGNANCY W MCC	1	0.21%	89.99%
440: DISORDERS OF PANCREAS EXCEPT MALIGNANCY W/O CC/MCC	1	0.21%	90.20%
443: DISORDERS OF LIVER EXCEPT MALIG,CIRR,ALC HEPA W/O CC/MCC	1	0.21%	90.41%

444: DISORDERS OF THE BILIARY TRACT W MCC	1	0.21%	90.62%
454: COMBINED ANTERIOR/POSTERIOR SPINAL FUSION W CC	1	0.21%	90.83%
455: COMBINED ANTERIOR/POSTERIOR SPINAL FUSION W/O CC/MCC	1	0.21%	91.04%
458: SPINAL FUS EXC CERV W SPINAL CURV/MALIG/INFEC OR 9+ FUS W/O CC/MCC	1	0.21%	91.25%
468: REVISION OF HIP OR KNEE REPLACEMENT W/O CC/MCC	1	0.21%	91.46%
479: BIOPSIES OF MUSCULOSKELETAL SYSTEM & CONNECTIVE TISSUE W/O CC/MCC	1	0.21%	91.67%
480: HIP & FEMUR PROCEDURES EXCEPT MAJOR JOINT W MCC	1	0.21%	91.88%
492: LOWER EXTREM & HUMER PROC EXCEPT HIP,FOOT,FEMUR W MCC	1	0.21%	92.09%
493: LOWER EXTREM & HUMER PROC EXCEPT HIP,FOOT,FEMUR W CC	1	0.21%	92.30%
500: SOFT TISSUE PROCEDURES W MCC	1	0.21%	92.51%
502: SOFT TISSUE PROCEDURES W/O CC/MCC	1	0.21%	92.72%
511: SHOULDER,ELBOW OR FOREARM PROC,EXC MAJOR JOINT PROC W CC	1	0.21%	92.93%
514: HAND OR WRIST PROC, EXCEPT MAJOR THUMB OR JOINT PROC W/O CC/MCC	1	0.21%	93.14%
516: OTHER MUSCULOSKELET SYS & CONN TISS O.R. PROC W CC	1	0.21%	93.35%
536: FRACTURES OF HIP & PELVIS W/O MCC	1	0.21%	93.56%
542: PATHOLOGICAL FRACTURES & MUSCULOSKELET & CONN TISS MALIG W MCC	1	0.21%	93.77%
543: PATHOLOGICAL FRACTURES & MUSCULOSKELET & CONN TISS MALIG W CC	1	0.21%	93.98%
556: SIGNS & SYMPTOMS OF MUSCULOSKELETAL SYSTEM & CONN TISSUE W/O MCC	1	0.21%	94.19%
557: TENDONITIS, MYOSITIS & BURSITIS W MCC	1	0.21%	94.40%
563: FX, SPRN, STRN & DISL EXCEPT FEMUR, HIP, PELVIS & THIGH W/O MCC	1	0.21%	94.61%
565: OTHER MUSCULOSKELETAL SYS & CONNECTIVE TISSUE DIAGNOSES W CC	1	0.21%	94.82%
570: SKIN DEBRIDEMENT W MCC	1	0.21%	95.03%
593: SKIN ULCERS W CC	1	0.21%	95.24%
623: SKIN GRAFTS & WOUND DEBRID FOR ENDOC, NUTRIT & METAB DIS W CC	1	0.21%	95.45%
628: OTHER ENDOCRINE, NUTRIT & METAB O.R. PROC W MCC	1	0.21%	95.66%
637: DIABETES W MCC	1	0.21%	95.87%
638: DIABETES W CC	1	0.21%	96.08%
639: DIABETES W/O CC/MCC	1	0.21%	96.29%
645: ENDOCRINE DISORDERS W/O CC/MCC	1	0.21%	96.50%

64: INTRACRANIAL HEMORRHAGE OR CEREBRAL INFARCTION W MCC	1	0.21%	96.71%
669: TRANSURETHRAL PROCEDURES W CC	1	0.21%	96.92%
68: NONSPECIFIC CVA & PRECEREBRAL OCCLUSION W/O INFARCT W/O MCC	1	0.21%	97.13%
694: URINARY STONES W/O ESW LITHOTRIPSY W/O MCC	1	0.21%	97.34%
697: URETHRAL STRICTURE	1	0.21%	97.55%
700: OTHER KIDNEY & URINARY TRACT DIAGNOSES W/O CC/MCC	1	0.21%	97.76%
70: NONSPECIFIC CEREBROVASCULAR DISORDERS W MCC	1	0.21%	97.97%
75: VIRAL MENINGITIS W CC/MCC	1	0.21%	98.18%
813: COAGULATION DISORDERS	1	0.21%	98.39%
815: RETICULOENDOTHELIAL & IMMUNITY DISORDERS W CC	1	0.21%	98.60%
84: TRAUMATIC STUPOR & COMA, COMA >1 HR W/O CC/MCC	1	0.21%	98.81%
854: INFECTIOUS & PARASITIC DISEASES W O.R. PROCEDURE W CC	1	0.21%	99.02%
857: POSTOPERATIVE OR POST-TRAUMATIC INFECTIONS W O.R. PROC W CC	1	0.21%	99.23%
85: TRAUMATIC STUPOR & COMA, COMA <1 HR W MCC	1	0.21%	99.44%
862: POSTOPERATIVE & POST-TRAUMATIC INFECTIONS W MCC	1	0.21%	99.65%
868: OTHER INFECTIOUS & PARASITIC DISEASES DIAGNOSES W CC	1	0.21%	99.86%
870: SEPTICEMIA OR SEVERE SEPSIS W MV 96+ HOURS	1	0.21%	100.00%
882: NEUROSES EXCEPT DEPRESSIVE	1	0.21%	100.00%

REFERENCES

100-04. (2020). In *cms. gov*. https://www.cms.gov/Regulations-and-Guidance/Guidance/Manuals/Internet-Only-Manuals-IOMs-Items/CMS018912

2019 National Healthcare Quality and Disparities Report Executive Summary (20(21)-0045-EF). (2020). Agency for Healthcare Research and Quality. https://www.ahrq.gov/sites/default/files/wysiwyg/research/findings/nhqrdr/2019qdr-final-es.pdf

About the Challenge | Agency for Healthcare Research & Quality. (2019). In *Ahrq. gov*. https://www.ahrq.gov/predictive-analytics-challenge/about.html

Abu-Rmileh, A. (2019, February 8). *The Multiple faces of "Feature importance" in XGBoost*. Towards Data Science. https://towardsdatascience.com/be-careful-when-interpreting-your-features-importance-in-xgboost-6e16132588e7

Acharya, A. S., Prakash, A., Saxena, P., & Nigam, A. (2013). Sampling: Why and how of it. *Indian Journal of Medical Specialties*, *4*(2), 330–333. https://www.researchgate.net/profile/Anita_Acharya/publication/256446902_Sampling_Why_and_How_of_it_Anita_S_Acharya_Anupam_Prakash_Pikee_Saxena_Aruna_Nigam/links/0c960527c82d449788000000.pdf

Admes, J., & Garets, D. (2014). The healthcare analytics evolution: moving from descriptive to predictive to prescriptive. *Analytics in Healthcare: An Introduction*. https://books.google.com/books?hl=en&lr=&id=15guBQAAQBAJ&oi=fnd&pg=PA13&dq=healthcare+analytics+evolution+moving+descriptive+predictive+prescriptive+Admes+Garets&ots=Gs-koWtz9p&sig=k61h-VCoLQvu2hHKEZTg9K4q1U8

Albarracin, D. (2020, November 9). *Predicting & understanding patterns of disease using big data*. University at Albany School of Public Health Lecture Series, Virtual (Zoom).

Alonso, S. G., de la Torre-Díez, I., Hamrioui, S., López-Coronado, M., Barreno, D. C., Nozaleda, L. M., & Franco, M. (2018). Data Mining Algorithms and Techniques in Mental Health: A Systematic Review. In *Journal of Medical Systems* (Vol. 42, Issue 9). https://doi.org/10.1007/s10916-018-1018-2

Amazon Sagemaker. (2020). In *Amazon Web Services*. https://aws.amazon.com/sagemaker/

Amazon SageMaker Clarify. (n.d.). Amazon. Retrieved March 3, 2021, from https://aws.amazon.com/sagemaker/clarify/

Ancker, J. S., Kim, M.-H., Zhang, Y., Zhang, Y., & Pathak, J. (2018). The potential value of social determinants of health in predicting health outcomes [Review of *The potential value of social determinants of health in predicting health outcomes*]. *Journal of the American Medical Informatics Association: JAMIA*, *25*(8), 1109–1110. https://doi.org/10.1093/jamia/ocy061

Ancona, M., Öztireli, C., & Gross, M. (2019). Explaining Deep Neural Networks with a Polynomial Time Algorithm for Shapley Values Approximation. In *arXiv [cs.LG]*. arXiv. http://arxiv.org/abs/1903.10992

Aoyama, K., D'Souza, R., Pinto, R., Ray, J. G., Hill, A., Scales, D. C., Lapinsky, S. E., Seaward, G. R., Hladunewich, M., Shah, P. S., & Fowler, R. A. (2018). Risk prediction models for maternal mortality: A systematic review and meta-analysis. *PloS One*, *13*(12), e0208563. https://doi.org/10.1371/journal.pone.0208563

Artetxe, A., Beristain, A., & Graña, M. (2018). Predictive models for hospital readmission risk: A systematic review of methods. *Computer Methods and Programs in Biomedicine*, *164*, 49–64. https://doi.org/10.1016/j.cmpb.2018.06.006

Becker, D., van Breda, W., Funk, B., Hoogendoorn, M., Ruwaard, J., & Riper, H. (2018). Predictive modeling in e-mental health: A common language framework. In *Internet Interventions* (Vol. 12, pp. 57–67). https://doi.org/10.1016/j.invent.2018.03.002

Berkowitz, S. A., Parashuram, S., Rowan, K., Andon, L., Bass, E. B., Bellantoni, M., Brotman, D. J., Deutschendorf, A., Dunbar, L., Durso, S. C., Everett, A., Giuriceo, K. D., Hebert, L., Hickman, D., Hough, D. E., Howell, E. E., Huang, X., Lepley, D., Leung, C., … Johns Hopkins Community Health Partnership (J-CHiP) Team. (2018). Association of a Care Coordination Model With Health Care Costs and Utilization: The Johns Hopkins Community Health Partnership (J-CHiP). *JAMA Network Open*, *1*(7), e184273. https://doi.org/10.1001/jamanetworkopen.2018.4273

Berwick, D. M., Nolan, T. W., & Whittington, J. (2008). The triple aim: care, health, and cost. *Health Affairs* , *27*(3), 759–769. https://doi.org/10.1377/hlthaff.27.3.759

Brabec, J., & Machlica, L. (2018). Bad practices in evaluation methodology relevant to class-imbalanced

problems. In *arXiv [cs.LG]*. arXiv. http://arxiv.org/abs/1812.01388

Bradley, A. P. (1997). The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition*, *30*(7), 1145–1159. https://doi.org/10.1016/S0031-3203(96)00142-2

Breiman, L., Friedman, J., Stone, C. J., & Olshen, R. A. (1984). *Classification and regression trees*. CRC press.

Bruce Sherman, K. W. (2020, December 8). *Ignoring inequitable benefit design is not an option*. Benefits Pro. https://www-benefitspro-com.cdn.ampproject.org/c/s/www.benefitspro.com/2020/12/08/ignoring-inequitable-benefit-design-is-not-an-option/?amp=1

Buchan, K., Filannino, M., & Uzuner, Ö. (2017). Automatic prediction of coronary artery disease from clinical narratives. *Journal of Biomedical Informatics*, *72*, 23–32. https://doi.org/10.1016/j.jbi.2017.06.019

CDC Social Vulnerability Index. (2014). [Data set]. https://doi.org/10.4211/hs.c2df2a80b9d6490788704a24854f4879

CDC SVI Documentation. (2018). https://www.atsdr.cdc.gov/placeandhealth/svi/documentation/SVI_documentation_2018.html

Chang, S.-F., & Lin, P.-L. (2015). Frail phenotype and mortality prediction: a systematic review and meta-analysis of prospective cohort studies. *International Journal of Nursing Studies*, *52*(8), 1362–1374. https://doi.org/10.1016/j.ijnurstu.2015.04.005

Chen, J. H., & Asch, S. M. (2017). Machine Learning and Prediction in Medicine - Beyond the Peak of Inflated Expectations. *The New England Journal of Medicine*, *376*(26), 2507–2509. https://doi.org/10.1056/NEJMp1702071

Chen, M., Mao, S., & Liu, Y. (2014). Big Data: A Survey. *Mobile Networks and Applications*, *19*(2), 171–209. https://doi.org/10.1007/s11036-013-0489-0

Choi, E., Xiao, C., Stewart, W. F., & Sun, J. (2018). MiME: Multilevel Medical Embedding of Electronic Health Records for Predictive Healthcare. In *arXiv [cs.LG]*. arXiv. http://arxiv.org/abs/1810.09593

CMS creates new chief data officer post | CMS. (2014). In *Cms. gov*. https://www.cms.gov/newsroom/press-releases/cms-creates-new-chief-data-officer-post

CMS is releasing Medicare Advantage encounter data for researchers to analyze. (2019). In *Healthcare Finance News*. https://www.healthcarefinancenews.com/news/cms-releasing-medicare-advantage-encounter-data-

researchers-analyze

Consortium, T. G. O., & T. G. O. Consortium. (2001). Creating the Gene Ontology Resource: Design and Implementation. In *Genome Research* (Vol. 11, Issue 8, pp. 1425–1433). https://doi.org/10.1101/gr.180801

Coordinating your care | Medicare. (2019). In *Medicare. gov*. https://www.medicare.gov/manage-your-health/coordinating-your-care

Cost, H., Project, U., & Others. (2016). *Clinical Classifications Software (CCS) for ICD-10-PCS (beta version)*. https://www.hcup-us.ahrq.gov/toolssoftware/ccs10/ccs10.jsp

Createspace Independent Pub, & Office of the Legislative Counsel. (2010). *Compilation of Patient Protection and Affordable Care Act*. CreateSpace Independent Publishing Platform. https://play.google.com/store/books/details?id=zndIMQAACAAJ

DeCaprio, D., Gartner, J., Burgess, T., Garcia, K., Kothari, S., Sayed, S., & McCall, C. J. (2020). Building a COVID-19 Vulnerability Index. In *arXiv [stat.AP]*. arXiv. http://arxiv.org/abs/2003.07347

Distribution of Medicare beneficiaries by race/ethnicity. (2020, October 23). https://www.kff.org/medicare/state-indicator/medicare-beneficiaries-by-raceethnicity/?currentTimeframe=0&selectedDistributions=black&selectedRows=%7B%22wrapups%22:%7B%22united-states%22:%7B%7D%7D%7D&sortModel=%7B%22colId%22:%22Location%22,%22sort%22:%22asc%22%7D

Distribution of Medicare beneficiaries by sex. (2020, October 23). https://www.kff.org/medicare/state-indicator/medicare-beneficiaries-by-sex/?currentTimeframe=0&sortModel=%7B%22colId%22:%22Location%22,%22sort%22:%22asc%22%7D

Duhigg, C. (2012, February 16). How Companies Learn Your Secrets. *The New York Times*. https://www.nytimes.com/2012/02/19/magazine/shopping-habits.html

Dworkis, D. A., Taylor, L. A., & Peak, D. A. (2016). 145 Multi-Patient High-Utilizer Addresses: Defining a Unique Population of High Utilizers of Emergency Care. *Annals of Emergency Medicine*, *4*(68), S58–S59. https://www.infona.pl/resource/bwmeta1.element.elsevier-b377ea08-e9b2-3fa2-a833-a82fa47bd7f0

Elixhauser, A. (1996). *Clinical Classifications for Health Policy Research, Version 2: Hospital Inpatient*

Statistics. U.S. Department of Health and Human Services, Public Health Service, Agency for Health Care Policy and Research. https://play.google.com/store/books/details?id=B_QhAQAAMAAJ

Evans, E. A., Delorme, E., Cyr, K., & Goldstein, D. M. (2020). A qualitative study of big data and the opioid epidemic: recommendations for data governance. *BMC Medical Ethics*, *21*(1), 101. https://doi.org/10.1186/s12910-020-00544-9

Fahey, M., Crayton, E., Wolfe, C., & Douiri, A. (2018). Clinical prediction models for mortality and functional outcome following ischemic stroke: A systematic review and meta-analysis. *PloS One*, *13*(1), e0185402. https://doi.org/10.1371/journal.pone.0185402

Federhen, S. (2012). The NCBI Taxonomy database. *Nucleic Acids Research*, *40*(Database issue), D136–D143. https://doi.org/10.1093/nar/gkr1178

Fedus, W., Zoph, B., & Shazeer, N. (2021). Switch Transformers: Scaling to Trillion Parameter Models with Simple and Efficient Sparsity. In *arXiv [cs.LG]*. arXiv. http://arxiv.org/abs/2101.03961

Feinleib, D. (2014). The Big Data Landscape. In D. Feinleib (Ed.), *Big Data Bootcamp: What Managers Need to Know to Profit from the Big Data Revolution* (pp. 15–34). Apress. https://doi.org/10.1007/978-1-4842-0040-7_2

Feurer, M., & Hutter, F. (2019). Hyperparameter optimization. In *Automated Machine Learning* (pp. 3–33). Springer, Cham. https://library.oapen.org/bitstream/handle/20.500.12657/23012/1007149.pdf?sequence=1#page=15

Fragoso, G., de Coronado, S., Haber, M., Hartel, F., & Wright, L. (2004). Overview and utilization of the NCI thesaurus. *Comparative and Functional Genomics*, *5*(8), 648–654. https://doi.org/10.1002/cfg.445

Friedman, J., Hastie, T., & Tibshirani, R. (2001). *The elements of statistical learning* (Vol. 1). Springer series in statistics New York. http://statweb.stanford.edu/~tibs/book/preface.ps

Goffman, R. M., Harris, S. L., May, J. H., Milicevic, A. S., Monte, R. J., Myaskovsky, L., Rodriguez, K. L., Tjader, Y. C., & Vargas, D. L. (2017). Modeling Patient No-Show History and Predicting Future Outpatient Appointment Behavior in the Veterans Health Administration. In *Military Medicine* (Vol. 182, Issue 5, pp. e1708–e1714). https://doi.org/10.7205/milmed-d-16-00345

Goldberg, Y., & Levy, O. (2014). word2vec Explained: deriving Mikolov et al.'s negative-sampling word-

embedding method. In *arXiv [cs.CL]*. arXiv. http://arxiv.org/abs/1402.3722

Grover, A., & Leskovec, J. (2016). node2vec: Scalable Feature Learning for Networks. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 855–864. https://doi.org/10.1145/2939672.2939754

Groves, P., Kayyali, B., Knott, D., & Van Kuiken, S. (2016). *The "big data" revolution in healthcare: Accelerating value and innovation*.

Hao, K. (2019). *An AI app that "undressed" women shows how deepfakes harm the most vulnerable*. MIT Technology Review.

Harris, L. J., Jeff Harris, L., Graetz, I., Podila, P. S. B., Wan, J., Waters, T. M., & Bailey, J. E. (2016). Characteristics of Hospital and Emergency Care Super-utilizers with Multiple Chronic Conditions. In *The Journal of Emergency Medicine* (Vol. 50, Issue 4, pp. e203–e214). https://doi.org/10.1016/j.jemermed.2015.09.002

Harvey, H. B., Benjamin Harvey, H., Liu, C., Ai, J., Jaworsky, C., Guerrier, C. E., Flores, E., & Pianykh, O. (2017). Predicting No-Shows in Radiology Using Regression Modeling of Data Available in the Electronic Medical Record. In *Journal of the American College of Radiology* (Vol. 14, Issue 10, pp. 1303–1309). https://doi.org/10.1016/j.jacr.2017.05.007

Hasselman, D. (2013). Super-Utilizer Summit: common themes from innovative complex care management programs. *Center for Health Care Strategies*.

Health And Economic Costs Of Chronic Diseases | CDC. (2020). In *cdc. gov*. https://www.cdc.gov/chronicdisease/about/costs/index.htm

Hill, K. (2012, February 16). How Target Figured Out A Teen Girl Was Pregnant Before Her Father Did. *Forbes Magazine*. https://www.forbes.com/sites/kashmirhill/2012/02/16/how-target-figured-out-a-teen-girl-was-pregnant-before-her-father-did/

Hospital Readmissions Reduction Program (HRRP). (2020). In *cms. gov*. https://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/Value-Based-Programs/HRRP/Hospital-Readmission-Reduction-Program

How Target gets the most out of its guest data to improve marketing ROI « machine learning times. (2010).

http://www.predictiveanalyticsworld.com/machinelearningtimes/how-target-gets-the-most-out-of-its-guest-data-to-improve-marketing-roi/6815/

Hughes, J. S., Averill, R. F., Eisenhandler, J., Goldfield, N. I., Muldoon, J., Neff, J. M., & Gay, J. C. (2004). Clinical Risk Groups (CRGs): a classification system for risk-adjusted capitation-based payment and health care management. *Medical Care*, *42*(1), 81–90. https://doi.org/10.1097/01.mlr.0000102367.93252.70

Improving Language Understanding With Unsupervised Learning. (2020). In *OpenAI*. https://openai.com/blog/language-unsupervised/

Indicators, A. Q. (2001). Prevention Quality Indicators Technical Specifications. *Department of Health and Human Services. Agency for Healthcare Research and Quality [accessed on February 10, 2018]. Available at Https://sharepoint. Fdrhpo. org/public/nchipPublic/Public% 20Documents/Preventative% 20Quality% 20Indicator% 20Documents/PQI_Appendices. Pdf*.

Islam, M. S., Hasan, M. M., Wang, X., Germack, H. D., & Noor-E-Alam, M. (2018). A Systematic Review on Healthcare Analytics: Application and Theoretical Perspective of Data Mining. *Healthcare (Basel, Switzerland)*, *6*(2). https://doi.org/10.3390/healthcare6020054

Jencks, S. F., Williams, M. V., & Coleman, E. A. (2009). Rehospitalizations among patients in the Medicare fee-for-service program. *The New England Journal of Medicine*, *360*(14), 1418–1428. https://doi.org/10.1056/NEJMsa0803563

Johnson, A. E. W., Pollard, T. J., Shen, L., Lehman, L.-W. H., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Celi, L. A., & Mark, R. G. (2016). MIMIC-III, a freely accessible critical care database. *Scientific Data*, *3*, 160035. https://doi.org/10.1038/sdata.2016.35

Jupp, S., Burdett, T., Leroy, C., & Parkinson, H. E. (2015). A new Ontology Lookup Service at EMBL-EBI. *SWAT4LS*, 118–119. http://ceur-ws.org/Vol-1546/paper_29.pdf

Kalyan, K. S., & Sangeetha, S. (2020). SECNLP: A survey of embeddings in clinical natural language processing. In *Journal of Biomedical Informatics* (Vol. 101, p. 103323). https://doi.org/10.1016/j.jbi.2019.103323

Kandhasamy, J. P., & Balamurali, S. (2015). Performance Analysis of Classifier Models to Predict Diabetes Mellitus. *Procedia Computer Science*, *47*, 45–51. https://doi.org/10.1016/j.procs.2015.03.182

Kansagara, D., Englander, H., Salanitro, A., Kagen, D., Theobald, C., Freeman, M., & Kripalani, S. (2011). Risk

Prediction Models for Hospital Readmission. In *JAMA* (Vol. 306, Issue 15, p. 1688). https://doi.org/10.1001/jama.2011.1515

Kiela, D., & Bottou, L. (2014). Learning Image Embeddings using Convolutional Neural Networks for Improved Multi-Modal Semantics. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 36–45. https://doi.org/10.3115/v1/D14-1005

Kojima, G., Iliffe, S., & Walters, K. (2018). Frailty index as a predictor of mortality: a systematic review and meta-analysis. *Age and Ageing*, *47*(2), 193–200. https://doi.org/10.1093/ageing/afx162

Kozyrkov, C. (2021a, February 9). *MFML 016 - Why trust AI?* Youtube. https://www.youtube.com/watch?v=pBYG5_kGMGY

Kozyrkov, C. (2021b, February 19). *MFML 017 - Explainability and AI*. Youtube. https://www.youtube.com/watch?v=J-cst3PBK4E

Krishnan, M. (2020). Against Interpretability: a Critical Examination of the Interpretability Problem in Machine Learning. *Philosophy & Technology*, *33*(3), 487–502. https://doi.org/10.1007/s13347-019-00372-9

Kronick, R., Bella, M., Gilmer, T. P., & Somers, S. (2007). The faces of Medicaid II: recognizing the care needs of people with multiple chronic conditions. *Center for Health Care Strategies, Inc*. https://www.policyarchive.org/handle/10207/11701

Kwon, O., Lee, N., & Shin, B. (2014). Data quality management, data usage experience and acquisition intention of big data analytics. *International Journal of Information Management*, *34*(3), 387–394. https://doi.org/10.1016/j.ijinfomgt.2014.02.002

Lewis, S. L., Feighner, B. H., Loschen, W. A., Wojcik, R. A., Skora, J. F., Coberly, J. S., & Blazes, D. L. (2011). SAGES: a suite of freely-available software tools for electronic disease surveillance in resource-limited settings. *PloS One*, *6*(5), e19750. https://doi.org/10.1371/journal.pone.0019750

Liberty, E., Karnin, Z., Xiang, B., Rouesnel, L., Coskun, B., Nallapati, R., Delgado, J., Sadoughi, A., Astashonok, Y., Das, P., Balioglu, C., Chakravarty, S., Jha, M., Gautier, P., Arpin, D., Januschowski, T., Flunkert, V., Wang, Y., Gasthaus, J., … Smola, A. (2020). Elastic Machine Learning Algorithms in Amazon SageMaker. *Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data*, 731–737. https://doi.org/10.1145/3318464.3386126

Li, C. (2020). OpenAI's GPT-3 Language Model: A Technical Overview. In *lambdalabs. com*.

Liu, H., & Cocea, M. (2017). Semi-random partitioning of data into training and test sets in granular computing context. *Granular Computing*, *2*(4), 357–386. https://doi.org/10.1007/s41066-017-0049-2

Lundberg, S. (n.d.-a). *Census income classification with XGBoost*. Slundberg.github.io. Retrieved March 15, 2021, from https://slundberg.github.io/shap/notebooks/Census%20income%20classification%20with%20XGBoost.html

Lundberg, S. (n.d.-b). *shap*. Github. Retrieved March 3, 2021, from https://github.com/slundberg/shap

Lundberg, S., & Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions. In *arXiv [cs.AI]*. arXiv. http://arxiv.org/abs/1705.07874

Lundberg, S. M., Erion, G. G., & Lee, S.-I. (2018). Consistent Individualized Feature Attribution for Tree Ensembles. In *arXiv [cs.LG]*. arXiv. http://arxiv.org/abs/1802.03888

Lund, S. R. (n.d.). *Slund - Overview*. Github. Retrieved March 16, 2021, from https://github.com/Slund

Metcalf, J., & Crawford, K. (2016). Where are human subjects in Big Data research? The emerging ethics divide. *Big Data & Society*, *3*(1), 2053951716650211. https://doi.org/10.1177/2053951716650211

Meticulous Market Research Pvt. Ltd. (2020, August 25). *Healthcare analytics market worth $84.2 billion by 2027, growing at a CAGR of 26% from 2020- pre and post COVID-19 market opportunity analysis and industry forecasts by meticulous research*®. PR Newswire. https://www.prnewswire.com/news-releases/healthcare-analytics-market-worth-84-2-billion-by-2027—growing-at-a-cagr-of-26-from-2020—pre-and-post-covid-19-market-opportunity-analysis-and-industry-forecasts-by-meticulous-research-301117822.html

Miotto, R., Wang, F., Wang, S., Jiang, X., & Dudley, J. T. (2018). Deep learning for healthcare: review, opportunities and challenges. *Briefings in Bioinformatics*, *19*(6), 1236–1246. https://doi.org/10.1093/bib/bbx044

Mitchell, E. M. (2017). *Concentration of Health Expenditures in the US Civilian Noninstitutionalized Population, 2014. Rockville, MD: Agency for Healthcare Research & Quality; 2016*.

Mobley, L. R. (2013). Spatial Sufficiency of 5% Medicare Standard Analytic Files. *Spatial Demography*, *1*(2), 202–218. https://doi.org/10.1007/BF03354899

Molnar, C. (2021, March 1). *5.10 SHAP (SHapley Additive exPlanations)*. https://christophm.github.io/interpretable-ml-book/shap.html

Mtsoulis, P. (2020). Sagify: Training And Deploying ML/DL Models On AWS Sagemaker Made Simple. In *medium. com*. https://medium.com/kenza-ai/training-and-deploying-ml-dl-models-on-aws-sagemaker-made-simple-e1132719838d

Ng, S. H. X., Rahman, N., Ang, I. Y. H., Sridharan, S., Ramachandran, S., Wang, D. D., Khoo, A., Tan, C. S., Feng, M., Toh, S.-A. E. S., & Tan, X. Q. (2020). Characterising and predicting persistent high-cost utilisers in healthcare: a retrospective cohort study in Singapore. *BMJ Open*, *10*(1), e031622. https://doi.org/10.1136/bmjopen-2019-031622

Nimmala, S., Ramadevi, Y., Nenavath, S. N., & Cheruku, R. (2018). Predicting High Blood Pressure Using Decision Tree-Based Algorithm. *Advances in Machine Learning and Data Science*, 53–60. https://doi.org/10.1007/978-981-10-8569-7_6

Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, *366*(6464), 447–453. https://doi.org/10.1126/science.aax2342

of Emergency Physicians, A. C., & Others. (2015). *2015 ACEP poll Affordable Care Act research results*. Marketing General Incorporated Alexandria.

Perera, S. (2018). *LITERATURE REVIEW FOR PREDICTING 30-DAY HOSPITAL READMISSION*. https://digitalcommons.mtu.edu/etdr/676/

Persad, G., Peek, M. E., & Emanuel, E. J. (2020). Fairly Prioritizing Groups for Access to COVID-19 Vaccines. *JAMA: The Journal of the American Medical Association*, *324*(16), 1601–1602. https://doi.org/10.1001/jama.2020.18513

Piatetsky, G. (n.d.). *Did Target really predict a teen's pregnancy? The inside story*. Retrieved March 3, 2021, from http://www.predictiveanalyticsworld.com/machinelearningtimes/target-really-predict-teens-pregnancy-inside-story/

Poole, S., Grannis, S., & Shah, N. H. (2016). Predicting Emergency Department Visits. *AMIA Joint Summits on Translational Science Proceedings. AMIA Joint Summits on Translational Science*, *2016*, 438–445. https://www.ncbi.nlm.nih.gov/pubmed/27570684

Pope, G. C., Kautter, J., Ellis, R. P., Ash, A. S., Ayanian, J. Z., Lezzoni, L. I., Ingber, M. J., Levy, J. M., & Robst, J. (2004). Risk adjustment of Medicare capitation payments using the CMS-HCC model. *Health Care Financing Review*, *25*(4), 119–141. https://www.ncbi.nlm.nih.gov/pubmed/15493448

Prevention Quality Indicators Overview. (2020). In *ahrq. gov*. https://www.qualityindicators.ahrq.gov/Modules/pqi_resources.aspx

Purushotham, S., Meng, C., Che, Z., & Liu, Y. (2017). Benchmark of Deep Learning Models on Large Healthcare MIMIC Datasets. In *arXiv [cs.LG]*. arXiv. http://arxiv.org/abs/1710.08531

Radford, A., Wu, J., Amodei, D., Amodei, D., Clark, J., Brundage, M., & Sutskever, I. (2019). Better language models and their implications. *OpenAI Blog Https://openai. Com/blog/better-Language-Models*. https://openai.com/blog/better-language-models/

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI Blog*, *1*(8), 9.

Raghupathi, W., & Raghupathi, V. (2013). An overview of health analytics. *Medical Record and Health Care Information Journal*, *4*(132), 2.

Rana, J. S., Tabada, G. H., Solomon, M. D., Lo, J. C., Jaffe, M. G., Sung, S. H., Ballantyne, C. M., & Go, A. S. (2016). Accuracy of the Atherosclerotic Cardiovascular Risk Equation in a Large Contemporary, Multiethnic Population. *Journal of the American College of Cardiology*, *67*(18), 2118–2130. https://doi.org/10.1016/j.jacc.2016.02.055

Reece, A. G., Reagan, A. J., Lix, K. L. M., Dodds, P. S., Danforth, C. M., & Langer, E. J. (2017). Forecasting the onset and course of mental illness with Twitter data. *Scientific Reports*, *7*(1), 13006. https://doi.org/10.1038/s41598-017-12961-9

Rinehart, D. J., Oronce, C., Durfee, M. J., Ranby, K. W., Batal, H. A., Hanratty, R., Vogel, J., & Johnson, T. L. (2018). Identifying Subgroups of Adult Superutilizers in an Urban Safety-Net System Using Latent Class Analysis: Implications for Clinical Practice. *Medical Care*, *56*(1), e1–e9. https://doi.org/10.1097/MLR.0000000000000628

Robinson, D., Yu, H., & Rieke, A. (2014). Civil rights, big data, and our algorithmic future. *Leadership Conference on Civil and Human Rights, Available at: Https://bigdata. Fairness. Io/wp-*

content/uploads/2014/11/Civil_Rights_big_Data_and_Our_Algorithmic-Future_v1, 1.

Robinson, R. L., Grabner, M., Palli, S. R., Faries, D., & Stephenson, J. J. (2016). Covariates of depression and high utilizers of healthcare: Impact on resource use and costs. *Journal of Psychosomatic Research*, *85*, 35–43. https://doi.org/10.1016/j.jpsychores.2016.04.002

Rogers, L. H., Gakhar, B., Singampalli, S., & Ali, S. A. (2012). Physician recommendation system (USPTO Patent No. 8103524). In *US Patent* (No. 8103524). https://patentimages.storage.googleapis.com/5b/8a/9e/ab48c7e40b72ca/US8103524.pdf

Ross, K. M., & Wing, R. R. (2016). Impact of newer self-monitoring technology and brief phone-based intervention on weight loss: a randomized pilot study. *Obesity* , *24*(8), 1653–1659. https://onlinelibrary.wiley.com/doi/abs/10.1002/oby.21536

sagify. (n.d.). Github. Retrieved November 6, 2020, from https://github.com/Kenza-AI/sagify

Samantha Artiga, J. K. (2020, December 3). *Addressing Racial Equity in Vaccine Distribution*. Kaiser Family Foundation. https://www.kff.org/racial-equity-and-health-policy/issue-brief/addressing-racial-equity-vaccine-distribution/

Santos Rutschman, A. (2020). *Mapping Misinformation in the Coronavirus Outbreak*. https://doi.org/10.2139/ssrn.3631555

Saravia, E. (2018). *Deep learning for NLP: An overview of recent trends - KDnuggets*. KDnuggets. https://www.kdnuggets.com/2018/09/deep-learning-nlp-overview-recent-trends.html

Sevo, R., & Chubin, D. E. (2010). Bias literacy: A review of concepts in research on gender discrimination and the US context. *Women in Engineering, Science and Technology*. https://www.igi-global.com/chapter/women-engineering-science-technology/43201

Sheets, L., Petroski, G., Zhuang, Y., Phinney, M., Ge, B., Parker, J., & Shyu, C.-R. (2017). Combining Contrast Mining with Logistic Regression To Predict Healthcare Utilization in a Managed Care Population. In *Applied Clinical Informatics* (Vol. 08, Issue 02, pp. 430–446). https://doi.org/10.4338/aci-2016-05-ra-0078

Siegel, E. (2020, October 23). When Does Predictive Technology Become Unethical? *Harvard Business Review*. https://hbr.org/2020/10/when-does-predictive-technology-become-unethical

Silberg, J., & Manyika, J. (2019). Tackling bias in artificial intelligence (and in humans). *McKinsey Global*

Institute.

Simpson, L., & Jain, S. H. (2021, February 10). To Make Progress, Focus On Building Trust. *Health Affairs*. https://www.healthaffairs.org/do/10.1377/hblog20210208.91982/full/

Solar, O., & Irwin, A. (2010). *A conceptual framework for action on the social determinants of health*. WHO Document Production Services. https://drum.lib.umd.edu/handle/1903/23135

Solove, D. J. (2005). A taxonomy of privacy. *University of Pennsylvania Law Review*, *154*, 477. https://heinonline.org/hol-cgi-bin/get_pdf.cgi?handle=hein.journals/pnlr154&section=20&casa_token=i3AtLH3FtFEAAAAA:8Q-pDsprIiyfPDUxroND9DSmNdH3B_mvdepnZK88CM_9C5GYVoqWI5BGHWzeoMJfWcaCuGuZ1w

STATISTICAL BRIEF #497: Concentration of Health Expenditures in the U.S. Civilian Noninstitutionalized Population, 2014. (n.d.). Retrieved September 9, 2020, from https://meps.ahrq.gov/data_files/publications/st497/stat497.shtml

Sterling, S., Chi, F., Weisner, C., Grant, R., Pruzansky, A., Bui, S., Madvig, P., & Pearl, R. (2018). Association of behavioral health factors and social determinants of health with high and persistently high healthcare costs. *Preventive Medicine Reports*, *11*, 154–159. https://doi.org/10.1016/j.pmedr.2018.06.017

Strengthening public health and improving equity through policy and data innovation. (2021, February 16). American Public Health Association. https://apha.org/Events-and-Meetings/APHA-Calendar/2021/Strengthening-Public-Health

Tam-Tham, H., Ravani, P., Zhang, J., Weaver, R. G., Quinn, R. R., James, M. T., Liu, P., Manns, B. J., Tonelli, M., Ronksley, P. E., Harrison, T. G., Thomas, C., Davison, S., & Hemmelgarn, B. R. (2020). Association of Initiation of Dialysis With Hospital Length of Stay and Intensity of Care in Older Adults With Kidney Failure. *JAMA Network Open*, *3*(2), e200222. https://doi.org/10.1001/jamanetworkopen.2020.0222

Taxonomy - Centers for Medicare & Medicaid Services. (2019). Cms.gov. https://www.cms.gov/medicare/provider-enrollment-and-certification/medicareprovidersupenroll/taxonomy.html

Tene, O., & Polonetsky, J. (2012). Big data for all: Privacy and user control in the age of analytics. *Nw. J. Tech. & Intell. Prop.*, *11*, xxvii. https://heinonline.org/hol-cgi-

bin/get_pdf.cgi?handle=hein.journals/nwteintp11&section=20&casa_token=7VrJwoU3JTAAAAAA:le8_jfNhroLow5T78Hi9FAO8sTWmo8MU0kwMu8h4HaiVVili7cQhBFfAsDkoDgCwpo7pjfHzwg

Tomasev, N., McKee, K. R., Kay, J., & Mohamed, S. (2021). Fairness for Unobserved Characteristics: Insights from Technological Impacts on Queer Communities. In *arXiv [cs.CY]*. arXiv. http://arxiv.org/abs/2102.04257

Topuz, K., Uner, H., Oztekin, A., & Yildirim, M. B. (2018). Predicting pediatric clinic no-shows: a decision analytic framework using elastic net and Bayesian belief network. In *Annals of Operations Research* (Vol. 263, Issues 1-2, pp. 479–499). https://doi.org/10.1007/s10479-017-2489-0

Treeprasertsuk, S., Leverage, S., Adams, L. A., Lindor, K. D., Sauver, J., & Angulo, P. (2012). The Framingham risk score and heart disease in nonalcoholic fatty liver disease. In *Liver International* (Vol. 32, Issue 6, pp. 945–950). https://doi.org/10.1111/j.1478-3231.2011.02753.x

Turbow, S., Fakunle, O., & Okosun, I. S. (2018). Multi- and Single-Year High-Utilizers of Inpatient Services Share Many Clinical and Behavioral Characteristics. *Journal of General Internal Medicine*, *33*(10), 1614–1615. https://doi.org/10.1007/s11606-018-4517-4

Tuttle, M. S., Blois, M. S., Erlbaum, M. S., Nelson, S. J., & Sherertz, D. D. (1988). Toward a Bio-Medical Thesaurus: Building the Foundation of the UMLS. *Proceedings of the Annual Symposium on Computer Application in Medical Care*, 191. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2245213/

University of Wisconsin School of Medicine Public Health. (n.d.). *Area Deprivation Index*. About the Neighborhood Atlas®. https://www.neighborhoodatlas.medicine.wisc.edu/

Vest, J. R., & Ben-Assuli, O. (2019). Prediction of emergency department revisits using area-level social determinants of health measures and health information exchange information. *International Journal of Medical Informatics*, *129*, 205–210. https://doi.org/10.1016/j.ijmedinf.2019.06.013

Wesson, D. E., Lucey, C. R., & Cooper, L. A. (2019). Building Trust in Health Systems to Eliminate Health Disparities. *JAMA: The Journal of the American Medical Association*, *322*(2), 111–112. https://doi.org/10.1001/jama.2019.1924

Yang, C., Delcher, C., Shenkman, E., & Ranka, S. (2018). Machine learning approaches for predicting high cost high need patient expenditures in health care. In *BioMedical Engineering OnLine* (Vol. 17, Issue S1).

https://doi.org/10.1186/s12938-018-0568-3

Zheng, H., Yuan, J., & Chen, L. (2017). Short-Term Load Forecasting Using EMD-LSTM Neural Networks with a Xgboost Algorithm for Feature Importance Evaluation. *Energies*, *10*(8), 1168. https://doi.org/10.3390/en10081168