Spring 2022

Using Machine Learning Techniques on Real-World Data to Understand the Characteristics of the Manchester, NH Health Care for the Homeless Patient Population for Risk Factor Identification and Intervention Improvement

Kyle Partridge Rasku
University of New Hampshire, Durham

Using Machine Learning Techniques on Real-World Data to Understand the Characteristics of the Manchester, NH Health Care for the Homeless Patient Population for Risk Factor Identification and Intervention Improvement

By

Kyle Partridge Rasku

BSN, Granite State College, 2017

THESIS

Submitted to the University of New Hampshire

in Partial Fulfillment of

the Requirements for the Degree of

Master of Science

in

Health Data Science

May 2022

This thesis/practicum was examined and approved in partial fulfillment of the requirements for the degree of Master of Science in Health Data Science by:

Robert McGrath II, M.S. M.A. Ph.D. Everett B. Sackett Professor and Chair of the Dept. of Health Management and Policy

Semra Aytur, M.P.H. Ph.D. Associate Professor, Health Management & Policy

John McInally R.N. M.B.A., Adjunct Professor, UNH Online

Brian Paciotti Ph.D. M.S., Instructor, University of California, Davis

On May 9, 2022

Approval signatures are on file with the University of New Hampshire Graduate School.

ACKNOWLEDGEMENTS

Firstly, thanks are due to the supportive team of professors in the Health Data Science program at the University of New Hampshire, who have been there for me since the beginning of this project, providing their unfailing expertise, guidance, and support, including Dr. Semra Aytur (primary advisor), John McInally (committee member), Dr. Joanna Gyory, Dr. Esmail Bahalkeh, Lyn Ferrara, and Dr. Mac Bonafede. Gratitude is also due to Dr. Brian Paciotti (committee member) of the University of California, Davis, who encouraged my interest in data analysis when I was his student in the UC Davis CDE Healthcare Analytics certification program and offered helpful tips and guidance throughout.

Without the Health Care for the Homeless team at Catholic Medical Center, Manchester, NH, none of this work would have been possible. Particular thanks are due to Matthew Augeri, HCHM's Health Information Systems Analyst, who advocated for this work from our first conversation. He always made himself available to answer my questions and review my progress. Additional thanks go out to Bonnie Frisard, Catholic Medical Center's IRB Administrator, and Melissa McGee, Assistant Director of Research Integrity Services at UNH. They both helped the project find its way through the IRB approval process.

I would also like to thank my friend Dr. Erin Griffin of the Elson S. Floyd College of Medicine at Washington State University, who encouraged my academic and analytic work, and my mother, Lauren O'Donnell, RHIA, who loves to use data to help people as much as I do and who always made herself available to step through my latest presentation or analysis. Last but not least, tremendous thanks go out to my husband, Tom, my daughter, Lexi, and the rest of my family, who supported me in taking on this program and project.

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

This thesis aims to use health care domain knowledge, statistical techniques, and machine learning methods to conduct an exploratory real-world evidence study of the characteristics of the Health Care for the Homeless of Manchester, NH (HCHM) clinics' patients in collaboration with academic and clinic partners and the public and community health stakeholders supporting their work. By constructing and analyzing a multivariate feature set created from a sample of anonymized patient data from January 1, 2018, through December 31, 2019, I hope to use machine learning methods to accurately represent 2,265 HCHM clinic patients experiencing homelessness or housing insecurity during the period. By regularly collaborating with analytics and clinical experts at HCHM, I hope to accurately describe the clinics' service populations and aid staff in identifying care gaps, enabling the enrichment of future interventions for homeless people in the primary care setting. By engaging in strategic science (Bunnell, Ryan & Kent, 2021), I hope to reduce bias around the study of this vulnerable population. The study period pre-dates the COVID-19 pandemic and is designed to provide a baseline analysis that will allow for future comparisons of HCH patients' sub-population characteristics and health care needs before, during, and after the pandemic.

The introduction outlines the public health crisis of homelessness in our country, connects the goal of providing care for people experiencing homelessness with the ongoing work of ensuring health equity, introduces the National Health Care for the Homeless Council and its care paradigm, and describes care provided by the Manchester, NH clinics within the city context. The chapter on Data describes the data sources used to create the aggregated data set and the data safeguards put in place to protect the privacy and dignity of people whose medical records were

used in the study. The Feature Development section details the dataset cleaning process and the development of the multivariate features, including local weather-based features and the creation of ICD-10 code-based condition categories specific to the challenges of persons experiencing homelessness. The Description chapter provides descriptive statistics related to the patient sample and outlines the health risks of clinic patients. The modeling goal was to utilize the full feature set, without removing outliers, to describe the variation in characteristics of clinic patients and group them into meaningful sub-populations by their utilization patterns. The Modeling section provides a detailed discussion of model evolution, and details about the dimension reduction and clustering algorithms applied to partition the data into service groups with specific characteristics, and how those characteristics were discoverable. The Service Groups chapter outlines the relationships between discovered clusters and patient service groups validated by HCH partners. The Discussion and Limitations chapter expands on and summarizes how the insights gleaned from this study may be helpful to the clinics, the community, the clients, and the health care system in providing future care to people experiencing homelessness and advancing health equity. It then discusses the limitations of the data, features, approach, and algorithms used in the study. It touches on study generalizability and ethics and bias considerations in research and algorithmic use and how these considerations were applied here. The thesis concludes with an endorsement of directions for building upon this work in the future.

TABLE OF CONTENTS

CHAPTER 1: INTRODUCTION

1.1 Addressing Homelessness with Data-Driven Insights

After steady reductions between 2010 and 2016, homelessness in the United States increased in the four consecutive years following and climbed even more acutely after the advent of the COVID-19 pandemic (US Department of Housing and Urban Development, 2022). In many ways, the struggles of communities and health care services organizations to provide compassionate, evidence-based, and patient-centered care to people experiencing homelessness are symbolic of the broader struggle our society faces to dismantle systemic and internalized bias toward vulnerable and marginalized people and deliver on the promise of health equity. It is a complex one-step forward, two-steps back sort of process, confounded by social, political, financial, individual, and institutional barriers and misunderstandings about who is homeless and why, and what can or should be done to assist homeless people.

Homelessness can strike anyone, but some are more vulnerable than others. Risk factors well-understood to increase the likelihood of homelessness include childhood trauma, poverty, job loss, divorce, economic downturn, foreclosure, lack of health insurance, mental illness, diseases of addiction, domestic violence, disability, being discriminated against (for one's race, ethnicity, sexual orientation, disability status or neurodivergence) and refugee status (Shelton et al., 2009). Homelessness is a transient state for most, with an estimated 1% of the U.S. population, or between 2.3 and 3.5 million people, experiencing homelessness during a given year (Urban Institute, 2000). It does not always mean sleeping in the street; but may mean taking up residence at a campsite, living in a vehicle, moving in with other families, with friends or relatives, or temporarily staying with a series of different people ('couch-surfing'). The

typical mental picture many have of a homeless person as an older white male alcoholic does not reflect the diverse ages, races, backgrounds, and circumstances of people experiencing homelessness today (Homeless Hub, 2022), nor the regularity with which the systemic socioeconomic hardships impacting marginalized communities draw Black people, indigenous people, and the LGBTQ community into the web of housing insecurity (Hwang & Henderson, 2010).

Although Jay Forrester introduced the concept of system dynamics in the 1950s (System Dynamics Society, 2022), it was not until recently that scientists, policymakers, and public health advocates began examining the problem of homelessness from a data-driven or systems theory perspective (Edwards, 2019; Seelos, 2021). Why is this shift in the discussion around methods of ending homelessness occurring only now – decades after Forrester provided astute insights into urban areas' complex dynamics (Forrester, 1969)? There are many potential answers. A critical change in thinking has occurred since the evidence-based successes of Housing First initiatives (Larimer et al., 2009). Advocates listened to those experiencing homelessness and prioritized permanent housing and supportive services for people who asked for and needed them. These programs showed long-lasting repeated success in reducing the number of chronically unhoused. They have become a best-practice recommendation in the United States and eight other countries (United States Interagency Council on Homelessness, 2018). For many, the efficacy of Housing First was both indisputable and unintuitive. Shouldn't housing people with issues such as alcoholism, drug addiction, and untreated mental illness always incentivize their recovery efforts by tying them to the possibility of a roof? We now know that this common and long-standing way of thinking about solutions to the problems of one of the highest-risk and most resource-intensive sub-populations of people experiencing

homelessness (Trick et al., 2021) is not always right (National Alliance to End Homelessness, 2019).

Other significant changes in the landscape around policy analyses of homelessness and homeless care include changes in technology and skillsets that assist with the timely analysis of large and complex data. However, even as computing power and technological sophistication have increased during the past fifty years, few have talked – until recently – about developing solutions to the problems of homelessness or poverty by modeling the cumulative interplay of social and economic forces on individuals or collecting data specifically to model such complexities (Seelos, 2021). What may, more likely, motivate the recent increase in the application of these long-important approaches to policy setting is the growing activism of public and community health leaders in response to decades of discussion around the need for equity as an antidote to well-researched, long-standing systemic bias in every aspect of our society that disproportionately disadvantages marginalized groups. The gentle but insistent pressure on leaders, policymakers, and public officials to begin to act in response to intelligence collected so many times over is finally starting to shift how we view both the problem of homelessness and its solutions. 'Common sense' solutions with a low basis in evidence are finally being challenged and jettisoned by those closest to homeless people and listening to what they need (Seelos, 2021). Competition and disagreement among service groups once trapped communities in a cycle that focused on creating and sustaining temporary shelters (Stroh, 2013); now, community leaders and mayors, as well as health, psychiatric, and social service workers, are coming together to collaborate on strategies that, once implemented, are making significant improvements in homelessness in communities all over the nation (Community Solutions, 2022). Data-driven insights into these implementations are, at last, being stressed at every stage of the

process. Organizations like Built for Zero, founded by Rosanne Haggerty – a long-standing advocate for the needs of homeless people – focus on addressing complex systems problems to end homelessness, insisting on data collection and feedback – not to overstudy already well-understood realities – but to fuel the ongoing planning and adaptations necessary to navigate the way to a zero-homelessness future (Harvard T.H. Chan School of Public Health, 2021).

These recent changes in methodology along with policy incentives brought on by the 2008 recession, have led to further validation of housing first, coupled with increasing emphasis on the prevention of homelessness through programs including rapid rehousing initiatives (Colburn, 2014). A recent systems dynamics model created by Fowler et al. (2019) at Washington University in St. Louis validates the growing understanding that homelessness prevention and housing first have the greatest potential to dramatically decrease homelessness in the United States. The authors stress the complexity and multifactorial nature of the problem of homelessness, and the ongoing need for widespread cross-programs collaboration to improve the consistency of efforts to reduce and eliminate homelessness in our nation (Fowler et al., 2019).

This work hopes to be a small part of the growing focus on using data-driven approaches and patient-centered multidisciplinary partnerships dedicated to continuing the essential work of establishing meaningful, contextually appropriate best practices in health care tailored to homeless and housing insecure people. Among current leaders in the fight against homelessness, there is an ongoing discussion about and research into the best ways to assist homeless people. However, there is growing consensus that due to the multifactorial nature of homelessness and the frequent voicelessness of people experiencing it, both centering the needs and opinions of homeless people and continuing to improve data-driven approaches to understanding the

complex interplay between factors that exacerbate or improve both homelessness and its health sequela are essential (Harvard T.H. Chan School of Public Health, 2021).

1.2 A Few Words on Syndemics

Another important way public health officials and researchers frame the multifactorial nature of homelessness and the complexities associated with decreasing it, is through the study and discussion of homelessness as a syndemic. The term "syndemic" originated when medical anthropologists coined the term to label the synergistic interaction of two or more coexisting diseases whose combined impact was potentially more significant than the sum of its parts (Singer & Clair, 2008). The originators take a critical view of the modern concept of "disease," asking whether it is an accurate description of a discrete thing, or more of an "explanatory model" (Good, 1994). Supposing "disease" is just a practical way of discretizing health imbalances for treatment and billing purposes, what we think of as individual diseases may be related syndromes whose impacts are difficult to tease apart. To encourage thinking about disease in a broader sense – including concomitant illnesses and the social, political, environmental and economic contexts in which they spring up and thrive – the term "syndemic" was coined and explained (Baer et al.,1997; Singer, 1996).

While expert observers agree that interactions and synergism between co-occurring diseases and environmental and social hazards have a causal connection, data-driven support for causal claims continues to improve. Alexander Tsai and Atheendar Venkataramani of Massachusetts General Hospital Global Health (2016) point out that syndemics researchers could take better statistical approaches that might lend more modeling support to their synergistic and causal claims. While researchers should consider their excellent advice, detecting and correcting for the complex confounding and sometimes circular or unidentifiable causal pathways inherent in the study of

complex social, environmental, and health problems, like homelessness, will continue to pose challenges to public health researchers, statisticians and data scientists.

1.3 Health Care for the Homeless

1.3.1 The National Council

The National Health Care for the Homeless Council (NHCHC) is one of the most visible national organizations in the United States, uniting health care professionals and people experiencing homelessness. NHCHC advocates for homeless people in the work of improving health care for both the homeless and housing insecure (National Health Care for the Homeless Council, 2022a). The council engages in advocacy and supports research into best practices to support health care providers in overcoming barriers to providing the best possible care to homeless people. The council receives almost two million dollars in grants from the Health Resources and Services Administration (HRSA) of the U.S. Department of Health and Human Services (DHHS), 20% of which comes from private sources. The council views both housing and health care as fundamental human rights.

The Council began in 1985 as a demonstration program funded by the Robert Wood Johnson Foundation (RWJF) and the Pew Memorial Trust. It expanded after the passage of the McKinney-Vento Homeless Assistance Act in 1987. Each year, it serves over 800,000 homeless people via 295 affiliated health centers (National Health Care for the Homeless Council, 2022b). Many of NHCHC's founders participated in the Community Health Center (CHC) movement. The first CHCs, called "Neighborhood Clinics," were opened in Massachusetts and in Mississippi in 1965 to improve the health and lives of Americans living in deep poverty (Health Center Partners, 2022). The movement spread across the nation, and today's CHCs provide health care to more than 27 million Americans. For thirty years, the National Health Care for the

Homeless Council has worked for an end to homelessness and advanced health care justice for the most vulnerable people in our society. In 2021, NHCHC's policy priorities included: 1) advocating for a single-payer system and Medicaid expansion, 2) mitigating the impact of COVID-19 through testing, treatment, vaccines, and housing, 3) increasing access to substance use treatment and harm reduction programs, and 4) the advancement of medical respite care – an evidence-based program (National Institute for Medical Respite Care, 2021) that provides temporary housing for homeless people recovering from illness and hospitalization (National Health Care for the Homeless Council, 2021).

1.3.2 The Care Paradigm

NHCHC and its member clinics promote dignity, respect, and patient-centered care for their primary clients – people experiencing homelessness and housing insecurity around the nation. One of the critical ways NHCHC advocates for people experiencing homelessness is by educating clinicians on the evidence for and implementation of practices related to trauma-informed care. Best practices for trauma-informed care ask clinicians to help identify trauma exposure in patients and develop policies that prevent re-traumatization by promoting healing care environments with safe, respectful, collaborative, and trustworthy communication. The evidence base for this type of care arises from research in both health and behavioral health disciplines, going back twenty years (U.S. Department of Health and Human Services, 2015). This research demonstrates that traumatic experiences are shared by Americans of all races, ethnicities and backgrounds and may have long-lasting impacts on social, psychological, emotional and physical health. As Bernie Siegel, MD famously stated, "The number one public health problem is our childhood." The now-famous ACEs study confirmed that Adverse Childhood Events (ACEs) disrupt neurodevelopment, leading to emotional and psychological

impairments, health-risk behaviors, immune-system damage, and eventually, disease, disability, and early death (Felitti et al., 1998). However, clinicians can offer hope to suffering people through advances in psychotherapy and mental health care and by listening to individuals who are still suffering from the aftereffects of a traumatic past. In one study, when ACE scores were measured and discussed by providers, this alone resulted in a 35% reduction in medical visits and an 11% reduction in ED visits (Nakazawa, 2016).

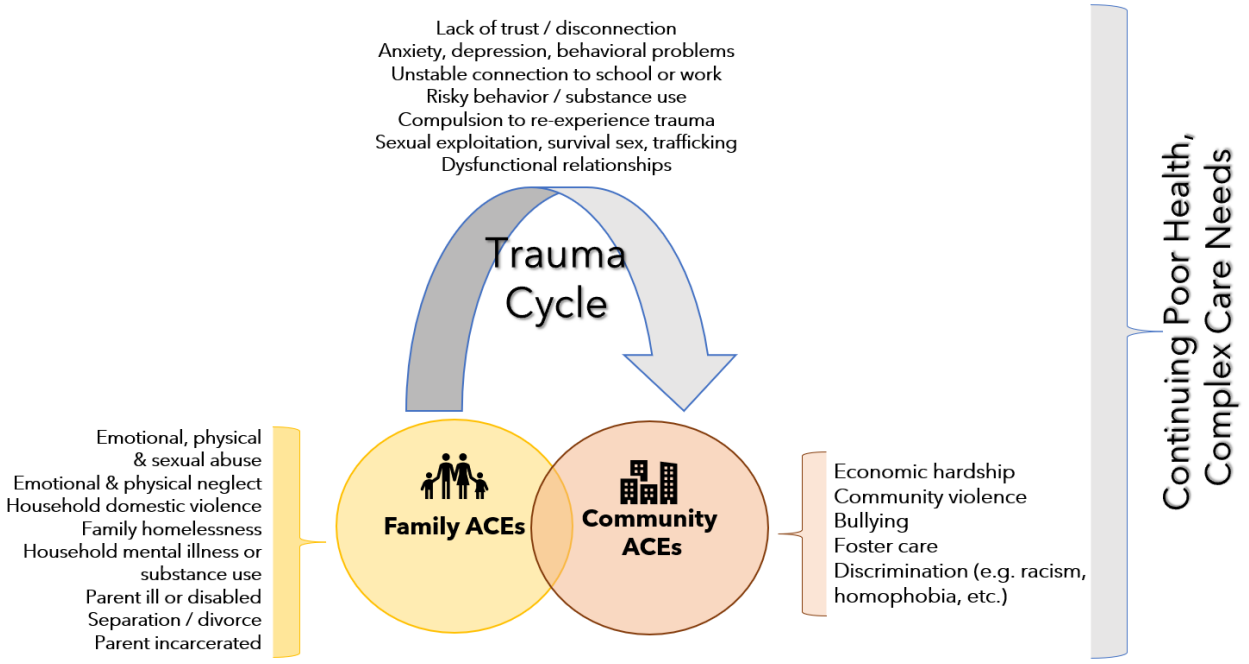


Figure 1: The Impact of The Trauma Cycle on Individual Health

(National Health Care for the Homeless Council (NHCHC) and National Network to End Family Homelessness (NNEFH), 2019)

While trauma-informed care is important for the general population, it is even more critical for people experiencing homelessness. People who become homeless are more likely to have experienced what this literature refers to as "complex trauma" – trauma that repeatedly occurs over time and consists of multiple types from both domestic (abuse, neglect) and community (discrimination, economic hardship) sources, resulting in dysregulation of the person's coping systems (Brien et al., 2019; National Child Traumatic Stress Network, 2003). Once coping

systems are overwhelmed, the person is likely to experience further trauma in the future – either through vulnerability to victimization or by seeking out familiar abusive or self-abusive situations or stimulation. In addition, both brain function and DNA expression adapt, leaving the person more susceptible to physical and mental illness (Bennett, 2016b). Those who face trauma from community or societal origins, such as systemic bias, are even more vulnerable to hardships such as job loss, eviction, foreclosure, or incarceration any of which may also result in homelessness. For decades, homeless people seeking care from the health care system have been asked questions like "What did you do?" or "What is wrong with you?" People commonly assumed the individual had somehow chosen homelessness as a lifestyle. Now, advocates for trauma-informed care teach clinicians to ask, "What happened to you, and how can we facilitate healing?" understanding that even people who seem to seek out trauma-reinforcing experiences may not do so freely, but because of damage done over time to their coping systems (Bennett, 2016a).

Changes in the neurobiology of people exposed to trauma were adaptive in their original environments but become maladaptive when the environment changes. Post-traumatic positive change and growth can occur when individuals receive needed services, including physical and mental health treatment, leading to more trusting relationships, reestablishing connection with family, and improved community support. Healing involves the development of resilience, and the ability to give others who are still suffering the byproducts of a traumatic past the hard-won gifts of understanding and acceptance (Bennett, 2016a).

1.3.3 The Queen City's Clinics

In 1846, Manchester became New Hampshire's first city. Today, it is New Hampshire's largest, with a population of 119,644 as of April 1, 2020 (U.S. Census Bureau, 2020). According to the

2020 American Community Survey (ACS), 18.4% of the city's population is under eighteen, and 13.7% is over 65 years old. The median age in 2019 was 38.7 years (City Data, 2019). The general population's racial and ethnic makeup is 82% non-Hispanic white, 5.9% Black or African American, 0.1% American Indian or Alaska Native, 5.3% Asian, 0.1% Native Hawaiian or Pacific Islander, 5.2% two or more races, and 10.7% Hispanic or Latinx. Veterans make up 6% of Manchester's population. Foreign-born persons make up 14.6%, and those over five speaking a language other than English at home account for 21.1% (U.S. Census Bureau, 2020).

Housing affordability is an issue in Manchester, where the median gross rent between 2016 and 2020 was approximately $1,160 per month. The median selected monthly housing ownership costs of people with a mortgage during the same period were estimated to be $1,788. For a person to afford a rent of $1,160 per month and have the rent be 50% of their monthly expenditures would be a precarious financial position, but they would still need to earn $580 per week. If they worked 40 hours every week, they would need to earn a minimum wage of at least $14.50 an hour. The minimum wage in New Hampshire is currently $7.25 an hour (New Hampshire Department of Labor, 2022). The estimated median household income in the city was $64,162 per year, and the estimated per capita income was $34,630 per year in 2019 (City Data, 2019).

The City of Manchester has thirty census tracts (City Builder, 2022) and stands at the northern end of Hillsborough county between Auburn and South Hooksett to the east and Goffstown and Bedford to the west. Health Care for the Homeless of Manchester, NH (HCHM), in cooperation with Catholic Medical Center (CMC) of Manchester, staffs and operates three clinics in the city – the Wilson Street Integrated Health Clinic (appointments and walk-ins), the Families in Transition Clinic (Mondays and Thursdays by appointment only), and the Families in Transition

Adult Emergency Shelter Clinic (appointments and walk-ins). The clinics provide primary medical care, mental health care and addiction counseling, health education and nurse case management, and social services connection assistance. They also offer dental and eye exams on a limited basis (Catholic Medical Center, 2022). The clinics have physicians, nurse practitioners, nurses, psychiatric nurse practitioners, and social workers on staff. The team also conducts street outreach at the Homeless Services Center, and other places around town where homeless people encamp and congregate, to assist them with care needs and encourage them to visit the clinics.

All three of HCHM's clinics are located in the Kalivas Union and Center City neighborhoods, consisting of census tracts 33011.1400.1 and 2 (also known as Census Tract 14) and tracts 33011.1500.1-3 (also known as Census Tract 15). This area is located between the Downtown and Hallsville neighborhoods and is bordered by Manchester Street to the north, Wilson and Maple Streets to the east, Cilley Road to the south, and Willow, Chestnut, and Pine Streets to the west. It is a densely populated residential section with a history of housing immigrant families over the years, including people of Irish, French and Greek heritage (City of Manchester, 2022). These two tracts both score high on the CDC's Social Vulnerability Index – a measure calculated using American Community Survey (ACS) data that evaluates areas on socioeconomic status, minority status, disability prevalence, and housing and transportation quality and availability. On a scale from 0 (least vulnerable) to 1 (most vulnerable), the Kalivas Union area (Tract 14) scores 0.99, and Center City (Tract 15) scores 0.947 (City Builder, 2022). Approximately 63.26% of households spend > 30% of their income on housing in these neighborhoods, while the metro-area average is 47.96% (Agency for Toxic Substances and Disease Registry, 2018). While the racial makeup of Manchester as a whole is approximately 82% non-Hispanic white (U.S. Census Bureau, 2020), the Kalivas Union and Center City census tracts are between 35 and

58% non-Hispanic white (City Data, 2019). Between 10% and 52% of residents in these areas have income below the federal poverty level, and median household incomes range from $22,635 to $43,185 (City Data, 2019). The table below summarizes PLACES and City Data information from 2018-2019 for the two tracts and compares them to the North End, an area of the city with low social vulnerability.

	Health Outcomes; 18+ % prevalence, 95% CI		
	Kalivas Union (SVI 0.99)	Center City (SVI 0.947)	North End (SVI 0.114)
Hypertension	**36.2% (26.6-27.9)**	29.8% (28.9-30.7)	25% (23.5-26.4)
Cancer	5.9% (5.8-6.1)	4.5% (4.3-4.6)	**7.6% (7.3-8.0)**
COPD	**11.4% (10.5-12.3)**	8.9% (8.0-9.9)	5.8% (4.5-7.1)
Heart Disease	**8.4% (8.0-8.9)**	5.8% (5.3-6.2)	5.0% (4.4-5.7)
Diabetes	**13.7% (13.2-14.3)**	10.8% (10.3-11.3)	7.9% (7.0-8.8)
Depression	**22.7% (21.8-23.6)**	**23.6% (22.4-24.8)**	20% (18.6-21.3)
Obesity	**42.1% (41.2-42.9)**	40.8% (39.7-41.8)	30.2% (28.6-31.8)
Stroke	**4.5% (4.2-4.8)**	3.3% (3.0-3.6)	2.6% (2.2-3.0)
Teeth Lost	**32.1% (26.3-37.8)**	**27.2% (20.0-35.1)**	9.9% (5.2-16.2)
	Prevention		
	Kalivas Union (SVI 0.99)	Center City (SVI 0.947)	North End (SVI 0.114)
Uninsured	**25.2% (22.8-27.6)**	**27.2% (24.0-30.4)**	9.2% (7.1-11.7)
Dental Visit	**44.5% (41.7-47.2)**	48.0% (44.5-51.5)	73.8% (69.2-77.9)
Colon CA Screening	**55.2% (52.4-57.9)**	58.4% (54.7-62.3)	73.4% (68.6-76.9)
	Health Risk Behaviors		
	Kalivas Union (SVI 0.99)	Center City (SVI 0.947)	North End (SVI 0.114)
Binge Drinking	16.0% (15.5-16.6)	17.2% (16.5-17.8)	**18.7% (17.9-19.4)**
Smoking	**28.4% (26.1-30.7)**	**27.5% (24.7-30.3)**	14.1% (11.0-17.5)
Physical Inactivity	**38.1% (36.0-40.1)**	**36.7% (34.0-39.3)**	19.5% (16.5-22.6)
Sleep < 7hr/ night	**43.3% (42.6-44.1)**	**43.2% (42.1-44.6)**	34.9% (32.3-37.3)
	Health Status		
	Kalivas Union (SVI 0.99)	Center City (SVI 0.947)	North End (SVI 0.114)
Fair/Poor General Health	**30.5% (28.2-32.9)**	**28.3% (25.6-31.2)**	12.9% (10.4-15.9)
Mental health not good last 14 days or more	**20.8% (19.6-21.9)**	**21.8% (20.2-23.3)**	12.4% (10.9-14.0)
Physical health not good last 14 days or more	**21.6% (20.3-22.8)**	19.1% (17.7-20.5)	11.6% (9.9-13.4)
*PLACES data; uses BRFSS 2018 and/or 2019; 2010 population counts and ACS 2015-2019 data			

Table 1: PLACES data (CDC, 2019) comparison of Kalivas Union, Center City, and North End

Bold text indicates worse health, higher risk behaviors, and less prevention.
Where confidence intervals overlap, more than one cell in a row is bold.

The table illustrates the impact of social determinants of health (SDoH) on community and individual health. Where communities lack adequate safety, job prospects, walkability,

affordable quality housing and transportation, recreational facilities or parks, and food availability, the physical and mental health of the population suffers. People must live in crowded and aging buildings and work harder to find and keep jobs, often sleeping less and having little time for family or recreation, and sometimes engaging in coping strategies that put their mental and physical health at risk. It's tempting to say these multifactorial impacts on the health of the Kalivas Union (Census Tract 14) and Center City (Census Tract 15) neighborhoods and the people living in them originated nearly 90 years ago when Manchester neighborhoods – including much of these tracts – were redlined by the federally sponsored Home Owners' Loan Corporation (HOLC), making it impossible for residents of these areas to get affordable mortgages, and leading to worsening living conditions in this and other similar areas of the city (Bassett, 2022b). While redlining in many parts of the United States focused on discrimination against racial minorities, in Manchester, there were few non-white residents. Instead, the HOLC's notes reveal widespread discrimination against the poor, the working class, and immigrant families (Nelson et al., 1937). Unfortunately, Manchester's history of discrimination against immigrants and the poor goes back even farther – about 185 years – to the late 1830s, when the Amoskeag Manufacturing Company owned 26,000 acres of land in the Manchester area before the city was even incorporated. Amoskeag built housing for its skilled workforce and the supervisors in its textile factories but would not even put its unskilled workforce on the waiting list. In the early 1840s, the area we now call Center City (Census Tract 15) was already becoming one of the only neighborhoods accessible to the working poor, most of whom were Irish immigrants at the time (Bassett, J., 2022a).

CHAPTER 2: DATA

2.1 The Centricity Electronic Medical Record Data

After completion of the Institutional Review Board (IRB) process at Catholic Medical Center and execution of the reliance agreement for the University of New Hampshire's IRB, Health Care for the Homeless of Manchester, NH (HCHM), and Catholic Medical Center (CMC) provided a series of Excel files containing de-identified data from HCHM's Electronic Health Record system, Centricity, for analysis of the visits of their clinic patients during the period January 1, 2018, through December 31, 2019. This included a demographic file on all clinic patients registered during the period, providing information on patients' housing status, insurance, veteran status, corrections history, highest completed education, age, self-identified race and ethnicity, sex, and preferred language. An Excel file containing raw clinical notes from all visits was also provided in chronological order. Patients were identified only by a randomly assigned identification number and not by name. No personally-identifying information (PII) such as addresses, zip codes, telephone numbers, email addresses, or social security numbers was provided.

2.2 The Outpatient Claims

Catholic Medical Center's Revenue Cycle office also provided outpatient claims data for the Health Care for the Homeless (HCHM) clinics. This Excel file contained service dates, claim IDs, clinic locations, CPT codes, and ICD-10 codes for all clinic visits. Patient data were linked to this using the same randomly assigned patient identification number used in the demographic and clinical notes data.

2.3 The Collective Medical Portal Data

Collective Medical's portal data is used by Health Care for the Homeless of Manchester (HCHM) and other federally qualified community health centers (FQCHCs) to help track their patients' visits to emergency departments all over the nation. This tracking is essential for patients experiencing homelessness because they may move about the country as transportation and temporary shelter arrangements become available to them. The Collective Medical portal data provided for the clinic population for the sample period contains the service dates, times, and locations for all emergency department visits experienced by patients between January 1, 2018, and December 31, 2019. It also provided all of the ICD-10 codes for each visit, and flags indicating whether or not the patient was admitted inpatient, had a non-emergent primary diagnosis, or died during the visit. Lastly, it provided the discharge date and time for each patient. Patients were, again, identified only by their randomly assigned patient identification numbers. Raw ED visit data for the period included 1,919 visits occurring in 2018 and 2,982 visits occurring in 2019, for a total of 4,901 ED visits representing 76 distinct locations around the nation, some as far away as Colorado, Oregon, Montana, California and Florida. The vast majority of all ED visits (75%) occurred in Manchester, at either Catholic Medical Center (1,966) or Elliot Hospital (1,692). Thirty hospitals had five or more visits by HCHM patients during the period. Of these, the most distant location was Kings Daughters Medical Center in Ashland, Kentucky.

Figure 2: Locations of 4,901 ED Visits Provided by Collective Medical Portal Data (BatchGeo, 2022)

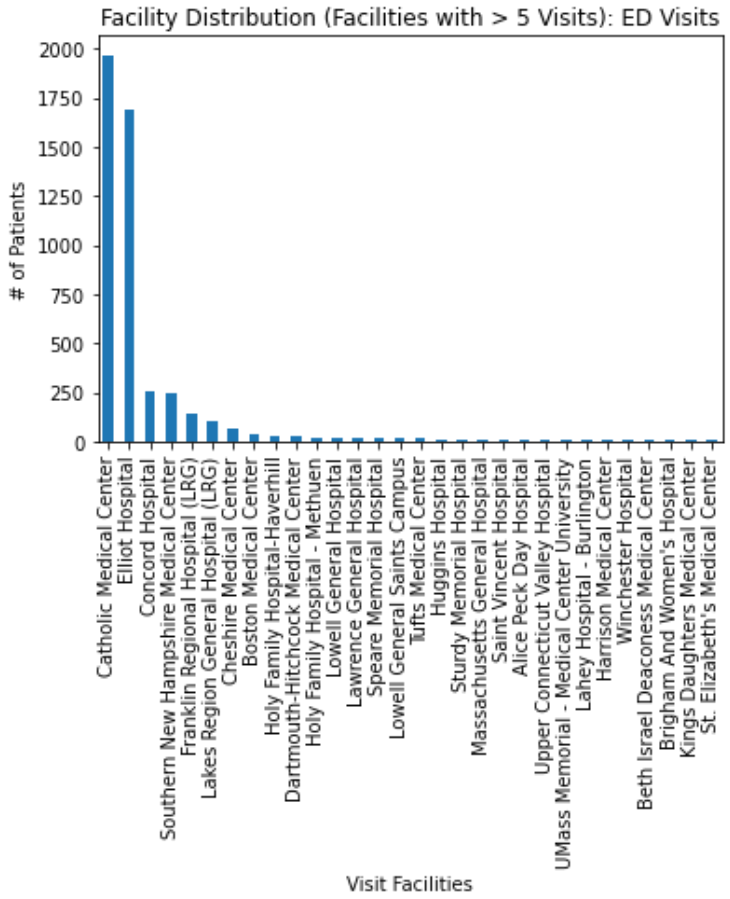


Figure 3: Facility Distribution – Locations of ED Visits with > 5 Visits (2018-2019)

2.4 Data Security and Governance

During the IRB process, data security and governance procedures to protect the privacy and identities of the HCHM patients whose de-identified records were sampled for the study period were agreed upon. During the analysis and reporting period, all patient data has been treated in a HIPAA-compliant manner at all times, and no personally identifiable information (PII) of any kind has been transmitted or displayed insecurely. Only the persons identified to the IRB as part of the analysis team have been able to view or analyze the data. No published reports or presentations of the data, analysis or research contain the personally identifying information (PII) of any patient, provider, or patient family member. Clinical notes data that accidentally included PII have been manually cleaned in four cases, and all such information has been permanently removed. Provider data contained in clinical notes was hashed during the feature set creation process so that individual providers could be identified for modeling without identification by name. Only the hashed information was retained in the final feature set, and some provider values were also aggregated together due to some providers' involvement with only a few patients. Thus, neither the final feature set used to produce the analysis, nor the final analysis contains any PII. In addition, no report displayed to the general public will present potentially identifying disaggregated data.

All data have been stored in a secure location with strictly defined view, edit, and deletion security protocols. Following completion of the analysis and reporting period, access to the original data will be revoked in accordance with HIPAA and the IRB policies of both Catholic Medical Center (CMC) and the University of New Hampshire (UNH). Only a wholly de-identified copy of the final feature set and related code and reports is retained for presentation purposes, so faculty may recreate any part of the analysis desired.

2.5 Data Cleaning

As noted above, clinical notes data were cleansed during feature development to remove all references to personally identifying data such as patient names, patient telephone numbers, the names of patients' significant others or family members, patient's locations or residences other than Families in Transition (formerly known as "New Horizons") shelter, and the names of all clinical staff caring for patients, including nurses, social services workers, physicians, nurse practitioners, and mental health professionals. Initial notes data had 442 entries that had no note, only a visit date and time. There were many exact duplicates, which were also removed. Seven rows were not visits but were entries indicating that correspondence was sent to a patient.

Additionally, thirty seven patients were removed from the analysis. Thirty six were removed by request of the sponsor, Health Care for the Homeless of Manchester, NH (HCHM), because they were no longer under the care of the clinic during the period (1/1/2018-12/31/2019), because they were EMR "shell accounts" that did not represent actual patients, or because they were non-clinic patients who received one-off care during the period, and thus had notes data but no EMR profile (demographic) data. After these patients were removed from the analysis, one patient remained who had emergency department visits but no clinic visits. Since this patient was the only such patient, this patient was also removed from the analysis. Out of 4,901 emergency department visits in the Collective Medical portal files, 3,636 visits by 720 patients were retained (74.2%), and out of 16,730 outpatient/clinic visits represented in notes or claims data, 12,061 visits by 2,265 patients were included (72.1%). Out of the 4,669 outpatient visits eliminated, twenty-six were removed because a visit was initiated for a patient who a) left the clinic without seeing the provider, b) had a medical emergency, and an ambulance was called, or c) required an

interpreter who was not able to make it to the appointment, and the appointment was rescheduled.

2.6 Missing Data

Notes data were available for 10,491 out of 12,061 visits (87%). However, all variables mined from notes data were aggregated across the individual patient, and most patients had notes data for at least one outpatient/clinic visit. For example, out of 2,265 patients, an average blood pressure reading was available for 2,206 patients (97.4%). A "calculated BMI" value was produced for 2,155 patients where average weight and height were available for at least one clinic visit (95.1%). Smoking status was assessed in clinic for 2,196 patients (97%) at least once and was available either from clinic notes or from an emergency department diagnosis code (or both) for 97.6% of patients (2,210). About 74% of patients (1,678) had at least one PHQ-2 (depression screening) score (Kronke, Spitzer & Williams, 2003), while 82% had at least one NIDA TAPS (substance use) assessment (McNeely et al., 2016). At least one HARK (Sohal, Eldridge & Feder, 2007) screening for intimate partner violence was completed for 465 patients (20.5%). A Hemoglobin A1C was available from the notes data for 233 patients, and a random office blood glucose level was mined for 232 out of the 306 patients (75.8%) who had at least one ICD-10 code recorded (either during a clinic or an emergency visit) for Diabetes Mellitus (any type).

Some demographic information was also unavailable in the electronic medical record data provided. For 1,688 patients (74.5%), no data on highest education level was recorded, and for 1,667 (73.6%), there was no current housing status information. This data stopped being collected during the clinic intake process at some time during the period. Hence, patients who were older or taken into the system earlier were more likely to have this information in their

profiles. Data on corrections history may also have been incomplete, as it followed a similar pattern.

CHAPTER 3: FEATURE DEVELOPMENT

3.1 The Demographic Data

The following demographic features were developed using electronic medical record (Centricity) demographic information provided for the 2,265 patients evaluated for the sample period:

Variable Name	Variable Value (Demographics)	Type	Original # of Values	Final # of Values	# of NULLs
housing	Housing Status	Nominal	9	5	1,667
housing 2	Housing Status	Nominal – coded	5	6	0
education	Highest Level of Education	Ordinal	32	4	1,688
education 2	Highest Level of Education	Ordinal – coded	4	5	0
insured	Primary Insurance	Binary	34	2	302
incarcerated	Corrections History	Binary	4	2	1,938
veteran	Veteran Status	Binary	2	2	0
si_race	Self-Identified Race	Nominal	8	7	50
si_race 2	Self-Identified Race	Nominal – coded	7	7	0
si_ethnicity	Self-Identified Ethnicity	Nominal	4	3	264
si_ethnicity 2	Self-Identified Ethnicity	Nominal – coded	3	3	0
age_group	Age	Ordinal	8	8	0
age_group 2	Age Group	Ordinal – coded	8	8	0
sex	Sex / Gender	Nominal	3	3	0
sex 2	Sex / Gender	Nominal – coded	3	3	0
language	Primary Language	Binary	11	2	10

Table 2: Demographic Feature Development

Demographic data for the patients analyzed was not aggregated; it was derived from a snapshot taken from the electronic medical record's registration system at the end of the sample period, and only one record per patient was provided. Some of the information may or may not have been up to date; however, all data was treated as authoritative for the period. Although housing status, highest level of education, and corrections history had many NULL values, these variables were all retained because they contained valuable information for the sample. Since there was no value for "none" under corrections history, NULL values were assumed to be none

or negative for corrections history, and the "incarcerated" variable was coded binary, with one (1) for any corrections history, and zero (0) for a NULL value. Housing status values that were not NULL were grouped into the following nominal categories: "STREET," "SHELTER," "TRANS/TRTMNT" (transitional or treatment), "DOUBLE-UP" (living temporarily with another person or family), and "OTHER/SRO" (another form of housing such as an apartment, single room, rooming house, or hotel) to create the "housing" variable. The "housing 2" variable is a coded version of this, coded as follows: -1 for NULL, 0 for "STREET" or "SHELTER", 1 for "TRANS/TRTMT", 2 for "DOUBLE-UP", and 3 for "OTHER/SRO". The "education" variable was created by grouping highest completed education values that were not NULL into the following ordinal categories: "0-8G", "9-12G", "HS/GED", and "SC/CG" for any amount of college. The "education 2" variable is the coded version with -1 for NULL, 0 for 0-8G, 1 for "9-12G", 2 for "HS/GED", and 3 for "SC/CG".

The Primary Insurance variable contained 34 distinct values for individual types of insurance or sliding scales used by patients to pay for care. "Self-pay" was also an option, and there were 302 NULL or "unknown" values. The binary "insured" variable uses one (1) for any insurance plan and zero (0) for *any other value*, including NULL, self-pay, and any cash payment arrangement.

The "veteran" binary variable is a direct translation of the existing variable, which had no NULL values, with one (1) indicating the individual is a veteran, and zero (0) indicating non-veteran. Likewise, the "sex" variable is also a direct translation of the existing variable, where there was one person who identified as neither of the two most common genders. Zero (0) was used to indicate male (because there were more males than females in the sample), one (1) to indicate female, and two (2) for "another gender". The "age_group" variable was directly translated from the continuous age variable. There were no NULL values, and ages were grouped into

seven ranges: "<10" (less than ten years), "10-18", "19-25", "26-35", "46-55", "56-65", and "65+" (sixty-five or older). The coded version, "age_group 2", codes the ordinal age ranges from 0 (under ten) to 6 (sixty-five plus). There were few clinic patients over sixty-five. This could be because the life expectancy of people experiencing homelessness is lower than the general population (the average estimated life expectancy for people experiencing homelessness is between 42 and 52 years (Health Care for the Homeless Clinicians Network, 2017)), or because many people who qualify for Medicare have more options as to where they obtain care and no longer seek primary care at the HCHM clinic.

The "language" variable was also coded as binary. This was a difficult choice, because a diverse immigrant and refugee population was well-represented by their language designations. In the end, only 368 people (16.2%) chose a primary language other than English. While 166 of these were Nepali speakers, I did not want to single out one group of people at the expense of others. For the descriptive analysis (see Chapter 4), I provide a breakdown of the languages spoken by these 368 people, but I used the binary "language" variable for the aggregate machine learning analysis.

The self-identified race data obtained from the demographic file originally contained eight categories and fifty unknown or NULL values. These were translated into seven categories by combining unknown values and (1) declined value into the category "OTHER/UNK". The final categories are "WHITE", "BLACK" (Black or African American), "ASIAN" (Asian), "MULTIPLE" (more than one), "OTHER/UNK" (another / unknown race), "AI/AK" (American Indian or Alaska Native), and "NH/PI" (Native Hawaiian or Pacific Islander) and the nominal coded version "si_race 2" uses 0 for "WHITE" (most common), and numbers the remaining categories in order of the number of patients identifying with that designation: 2 for "BLACK", 3

for “ASIAN”, 4 for “MULTIPLE”, 5 for “OTHER/UNK”, 6 for “AI/AK” and 7 for “NH/PI”. Likewise, the self-identified ethnicity variable (“si_ethnicity”) was created by combining the 264 “unknowns” with two “other” designations in the original ethnicity information to create a variable with three categories: “NON-HISPANIC”, “HISPANIC” and “OTHER/UNK”. As with race, for the coded variable “si_ethnicity 2”, the category coded zero (0) (“NON-HISPANIC”) was the most common among the patients and the second most common category (“OTHER/UNK”) was coded as one (1). “HISPANIC” was coded as two (2); this designation was chosen by 7.2% of patients. Race and ethnicity data are presented in a disaggregated fashion wherever possible. However, some categories are necessarily combined in descriptive presentations because the small numbers of patients in some categories pose a risk to their anonymity. The same thinking is applied when presenting data by gender.

3.2 Visit Counts and Intervals

To create the visit count variables “op_visit_count” and “ed_visit_count” a longitudinal data set was first created from all visits in chronological order. Then, this longitudinal data was pivoted by the de-identified patient ID and summed across the count of rows by type of visit – outpatient or emergency. The “admitted” variable was created using counts of rows by patient ID where the patient had an emergency visit that converted to inpatient, and the “non_emergent_dx” (non-emergent diagnosis) variable was created using counts of rows by patient ID where the patient had a non-emergent diagnosis associated with an emergency department visit.

To create the interval variables, “avg_ed_interval” and “avg_op_interval,” intervals were calculated between visits for all patients having more than one outpatient/clinic visit and more than one emergency department visit. Where there was more than one interval between visits, intervals were averaged for each patient. Patients with a single clinic or emergency visit had

NULL values for this variable, later filled in with the censorship value of 730 days, signifying the end of the two-year period. While these values are subject to decreased variability due to averaging, patients with greater utilization generally had shorter average intervals.

3.3 ICD-10 Codes to Homeless-Specific Condition Counts

ICD-10 stands for International Classification of Diseases, version 10, and refers to an international disease classification and coding system used for cataloging and describing diseases, disease sequelae, and surrounding information such as symptoms, family history, and situational circumstances for medical record-keeping and billing related to each health system encounter such as an office visit, emergency visit or inpatient stay (American Association of Professional Coders (AAPC), 2021). The ICD-10 system was adopted by the Centers for Medicare and Medicaid Services (CMS) in October of 2015. Codes are added to and updated every year and as needed in an emergency, such as the COVID-19 pandemic. As of the 2020 update, there were 72,184 distinct ICD-10 codes.

While codes are the accepted way to capture and categorize disease in the U.S. health care system, there are too many of them to allow for the encoding of a patient-level data set for the presence or absence of all of the specific codes associated with a given patient's visits. To capture both disease categories and the intensity of each comorbidity as thoroughly as possible, I chose to produce a series of diagnosis categories and introduce a homeless-specific comorbidity measure, the homeless-specific condition score (HSCS). I based the creation of this score on other comorbidity scoring systems, such as the Elixhauser score (Quan et al., 2005; van Walraven, 2009), where patients with particular conditions have condition flags added to their data profile for a given hospital or physician visit based on the presence of ICD-10 codes. Those flags are summed together to achieve a total visit or conditions score.

To produce the visit reason/diagnosis categories, I conducted research using literature reviews and reviews of literature reviews, searching for information on the most common conditions impacting homeless people (Edidin et al., 2012; Lewer et al., 2014; Medlow, Klineberg & Steinbeck, 2014; Aldridge et al., 2018; Nanjo et al., 2020; Tannis & Rajupet, 2021). I reviewed the results and discussed the conditions and categories with John McInally, a thesis committee member with decades of experience in clinical informatics and emergency nursing. Then, I formulated the draft categories using our conversation as a reference, along with my own familiarity with medical conditions and ICD-10 codes. I reviewed the draft conditions with contacts at Health Care for the Homeless of Manchester (HCHM), who approved the following nineteen visit condition groups:

Condition Grp Abbr	Condition Group Name & Description	ICD-10 Codes	Description of Included Codes
CVD	Cardiovascular Disease	I05-I99, except I10-I16 (Hypertension), I50, I30-33 and I38-41; R00-R01	Includes: chronic rheumatic heart disease, ischemic heart disease, pulmonary heart diseases, PA aneurysm, valve diseases, and dysrhythmias
HF	Heart Failure	I50	All heart failure, regardless of etiology
HTN	Hypertension	I10-I16 and R03	All hypertensive conditions and the symptom “high blood pressure”
URI/PNA	Acute upper respiratory infections and pneumonias	J00-J06; J09-J18; J20-J22; J60-J70; J80-J84; J90-J91, R05-R07	Acute upper respiratory infections, influenza, pneumonia, acute bronchitis and bronchiolitis, airway disease of external exposures, ARDS, pulmonary edema, pleural effusion, pneumothorax, cough, breathing abnormalities, throat and chest pain
AS/COPD	Asthma, Chronic Obstructive Pulmonary Disease (COPD), and other chronic respiratory	J40-45 and J47; J98; J96.1, J96.2, J96.9	Chronic bronchitis, emphysema, COPD, asthma, bronchiectasis, bronchospasm, atelectasis, chronic respiratory failure

	diseases		
NEURO	Neurological diseases that do not fall into another category (Sensory, Pain-related, or Cognitive)	G06-G47; G70-G99, except G30, G31 (Dementias) and G89 (Pain NOS); R55, R56	Spinal or brain abscess, inflammation not caused by infection, Huntington's, Parkinson's, tremor, chorea, tics, restless leg syndrome, multiple sclerosis, epilepsy, migraines, TIA, stroke/CVA, sleepwalking, myasthenia gravis, muscular dystrophy, cerebral palsy, hydrocephalus, toxic encephalopathy, autonomic dysreflexia, syncope and convulsions
SUD	Substance use disorders and related symptoms	F10-F19 except F17 (nicotine dependence); T40-T43; T52; Y90	Alcohol, opioid, cannabis, sedative/hypnotic, anxiolytic, cocaine, meth, other stimulants, hallucinogens, inhalants, and other mood and thought-altering substance use/abuse and related disorders; poisoning by or adverse effects of narcotics, anesthetics, sedatives, and other psychotropic drugs and evidence of alcohol involvement by blood alcohol level
MHD	Mental health diseases and related problems	F20-F69, F90-F99, R44-R46, X71-X83, Z72.81	Schizophrenia, schizotypal, delusional, mood, anxiety, dissociative, somatoform, personality, conduct, and other mental and behavioral disorders; hallucinations, nervousness, anger, violence, worries, homicidal or suicidal ideation, intentional self-harm, and antisocial behavior
CA	All cancers and neoplasms	C00-D49; R97	All cancers and neoplasms
PREG	All pregnancy and obstetrics-related conditions (except gestational diabetes and tobacco use)	O00-O9A except O24 (Diabetes in pregnancy) and O9933 (Tobacco use in pregnancy)	All obstetric codes except where otherwise noted
DM	Diabetes Mellitus (any type) and related	E08-E13, R73, E88.81, O24	All Diabetes codes, metabolic syndrome, gestational diabetes (in

	diagnoses and symptoms		pregnancy), prediabetes, high blood sugar
INF	All infections not covered under acute or chronic respiratory diseases	A00-B99, G00-G05, I00-I02, L00-L08, M00-M02, N10-N12, N30, N39 (UTI), I30-I33, I38-I41, R50, R65	Brain, heart, bone, blood, skin, joint, urinary, and other infections by any bacteria, virus or mycoplasma (fungi); septic shock, SIRS, and fevers of unknown origin
LIV	Liver, pancreatic, and gallbladder diseases (not cancers)	K70-K87; R16-R18	Alcoholic and toxic liver disease, hepatic failure, chronic hepatitis, cirrhosis, NASH, cholelithiasis, cholecystitis, pancreatitis, hepatic or splenomegaly, jaundice, ascites
REN	Kidney diseases (not cancers)	I12 and I13 (Only these two codes counted as BOTH an HTN and a REN diagnosis in this algorithm); N00-N19, N25-N27, N28.0, N28.1, N28.81 and N29	Glomerular diseases, renal tubulointerstitial disorders, acute kidney failure and chronic kidney disease (all stages), diseases of impaired renal tubular function, congenital kidney diseases, ischemia or infarction, cyst, or hypertrophy
COG	Diseases impacting cognition, whether congenital, acute or chronic	F70-F89, F01-F09, G30-31, S06, I69.01, I69.11, I69.21, I69.31, I69.81, I69.91	Intellectual disabilities, developmental disorders, vascular dementia, dementia NOS, delirium, Alzheimer's, Lewy-body dementia, intracranial injury, sequelae of cerebrovascular accidents, cognitive deficits
SENS	Sensory deficits; diseases impacting the senses, including vision, hearing, smell, touch/sensation in limbs, and balance	H40-H42, H46-H47, H53-H54, G50-G65, H80-H94, R20, R40-R44	Glaucoma, disorders of the optic nerve, visual disturbances and blindness, nerve, nerve root, and plexus disorders, polyneuropathies, diseases of the inner ear, hearing loss, disturbances of skin sensation, somnolence, dizziness, disturbances of smell and taste, other problems of sensation and perception (not mental health-related)
PAIN	Acute or chronic pain in an area or body part or pain	G43, G54.6, G89, G90.5, H57.1, M54, M79.1, M79.2,	Migraine, phantom limb pain, pain NOS, CRPS, ocular pain, all dorsalgias (panniculitis, radiculopathy, sciatica,

	syndromes	M79.6, M79.7, M25.5, R10, R14.1, R51, R52	etc.), all myalgias, neuralgia, fibromyalgia, pain in joints, abdominal and pelvic pain, gas pain, headache
TOB	Tobacco use	F17, Z720, O9933, U070	Tobacco use, dependence, tobacco use in pregnancy, vaping-related disorder
ACC/INJ	Accidents and injuries including assaults, falls, work-related accidents, and other misfortunes	R29.6, S00-S99; T07-T34; T66-T79; V00-V99; W00-W99; X00-X58; X92-X99; Y00-Y09; Y21-Y33	Repeated falls, injuries to any body part or organ, including those involving multiple or unspecified body regions, effects of foreign bodies entering through natural orifices, burns and corrosions (any site), frostbite; all transportation accidents; all sports-related accidents; work-related accidents, drowning, tripping/stumbling, smoke/fire exposure, steam or chemical burns, heat and cold exposure, natural disaster-related injuries, assault, war, terrorism, bombing, overexertion, medical errors and device failures, adverse drug effects, accidental poisoning, injuries resulting from firearm discharges

Table 3: Visit Reason Categories by ICD-10 Codes and Illness Descriptions

After collaborative development of the visit reason categories, all ICD-10 codes related to all visits were grouped into their appropriate categories, and the visit reason category counts were summed separately across each patient's emergency and outpatient/clinic visits to produce a total of thirty-eight features: CVD_ed, HF_ed, HTN_ed, URI/PNA_ed, AS/COPD_ed, NEURO_ed, SUD_ed, MHD_ed, CA_ed, PREG_ed, DM_ed, INF_ed, LIV_ed, REN_ed, COG_ed, SENS_ed, PAIN_ed, TOB_ed, ACC/INJ_ed and CVD_op, HF_op, HTN_op, URI/PNA_op, AS/COPD_op, NEURO_op, SUD_op, MHD_op, CA_op, PREG_op, DM_op, INF_op, LIV_op, REN_op, COG_op, SENS_op, PAIN_op, TOB_op, and ACC/INJ_op. Naturally, patients with

more visits had higher numbers of visit reason category counts. However, there was still high variation in the features' values because some patients had large numbers of visits and few comorbidities, while others had large numbers of visits and many comorbidities. Patients with no emergency department visits had zero visit reason category counts for all emergency-related categories.

To produce aggregate, comparative scores for each patient, I first scored each visit in the longitudinal data set using both the Elixhauser comorbidity categories summed to produce a final score (Quan et al., 2005; Wasey, 2020) and the visit reason categories, summed to produce the homeless-specific condition score (HSCS). Then, I pivoted the data using the de-identified patient ID, divided the data between emergency and outpatient/clinic visits, and averaged the scores across each patient's visit sets. This resulted in four additional features – the patient's average Elixhauser score for their emergency visits (avg_elix_ed) and outpatient/clinic visits (avg_elix_op), and their average HSCSs for their emergency visits (HSCS_ed) and outpatient/clinic visits (HSCS_op).

3.4 Comparison of a Homeless-Specific Condition Score to the Elixhauser Score

To get an idea of how well the HSCS scores aligned with the Elixhauser scores across emergency and outpatient visits, I compared correlations and distributions and compared each measure's ability to predict the number of visits in a Poisson regression model with dispersion and Firth-Adjusted estimates. The Elixhauser score is a highly validated risk-adjustment score used in many prediction models against outcomes such as morbidity, mortality, and inpatient re-admissions (Chu, Ng & Wu, 2010; Fortin et al., 2017). I did not expect my scores to match Elixhauser's performance in visit predictions. Still, I thought it would help validate the visit

reason categories if a relationship between the HSCS and Elixhauser scores could extend to a basic visit prediction model.

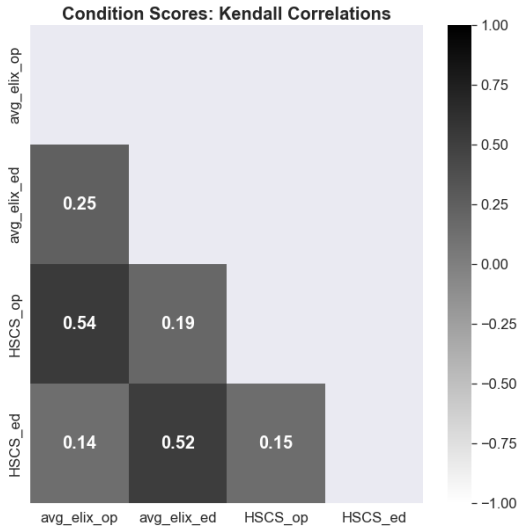


Figure 4: Kendall Correlation of the Average Elixhauser and Homeless-Specific Condition Scores by Patient

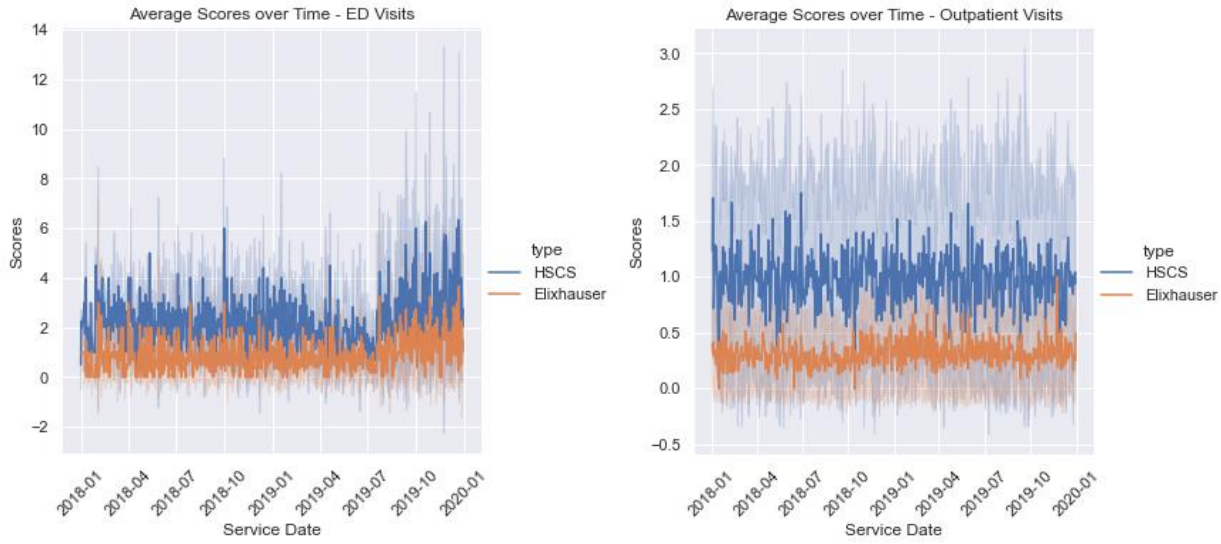


Figure 5: Comparison over time of the Average HSCS and Elixhauser for ED visits (left) and Clinic visits (right) – with standard deviations

In the Figure 4 comparison, it is apparent that the average Elixhauser and homeless-specific scores (HSCS) are coordinated for emergency visits. However, for outpatient visits, the scores diverge. It is also evident that there is more variation in all outpatient scores. This is not

surprising, since outpatients can range from healthy people seeking preventive care (their scores might be zero) to those with higher acuity or more comorbidities (having higher scores). There is also more correlation between emergency and outpatient scores for the more complex Elixhauser measure than the simple homeless-specific condition scores. These findings confirm the sensibility of the regression results.

Figure 6: Comparison of the predictive ability of the Elixhauser score (left) and HSCS (right) against the outcome of the number of emergency department visits

Figure 7: Comparison of the predictive ability of the Elixhauser score (left) and HSCS (right) against the outcome of the number of outpatient/clinic visits

While both the Elixhauser score and HSCS alone had some predictive power with respect to the number of visits – whether clinic or emergency; after adjusting for demographic variables, only the Elixhauser score retained a significant p-value against the outcome of outpatient visits. Even so, the parameter estimate was exceedingly small. Both scores had a larger effect on the outcome of emergency department visits, but unsurprisingly, Elixhauser had more robust predictive abilities. Many more factors contribute to the variation in the number of outpatient visits among the sample than the variation in emergency visits, which are more dependent on acuity and comorbidity.

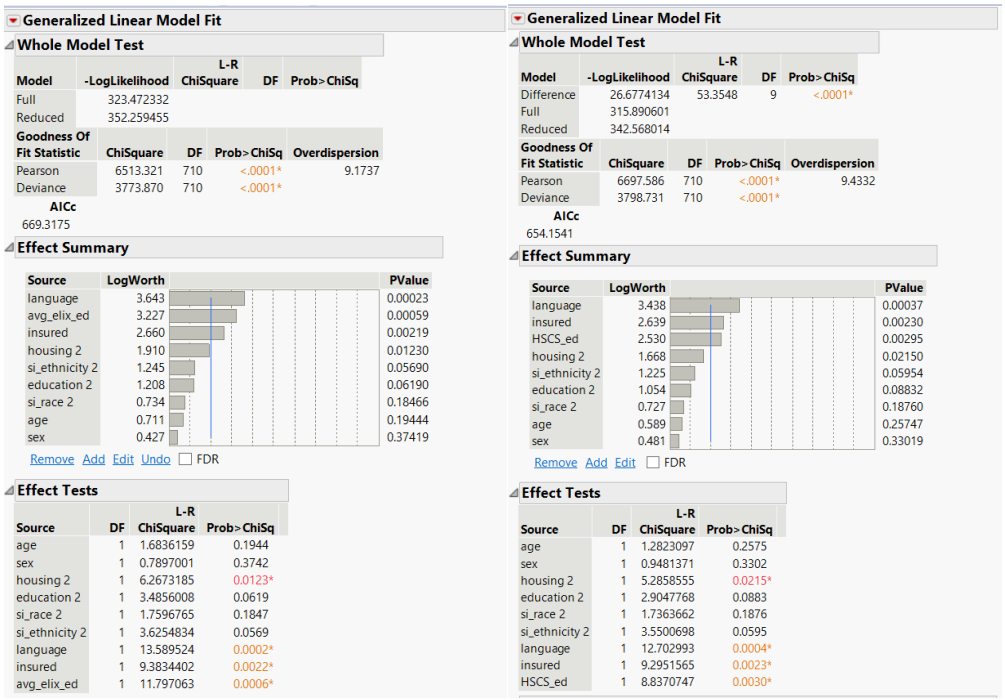


Figure 8: Comparison of the predictive ability of the Elixhauser score (left) and HSCS (right) against the outcome of the number of emergency department visits in an adjusted model with demographic information

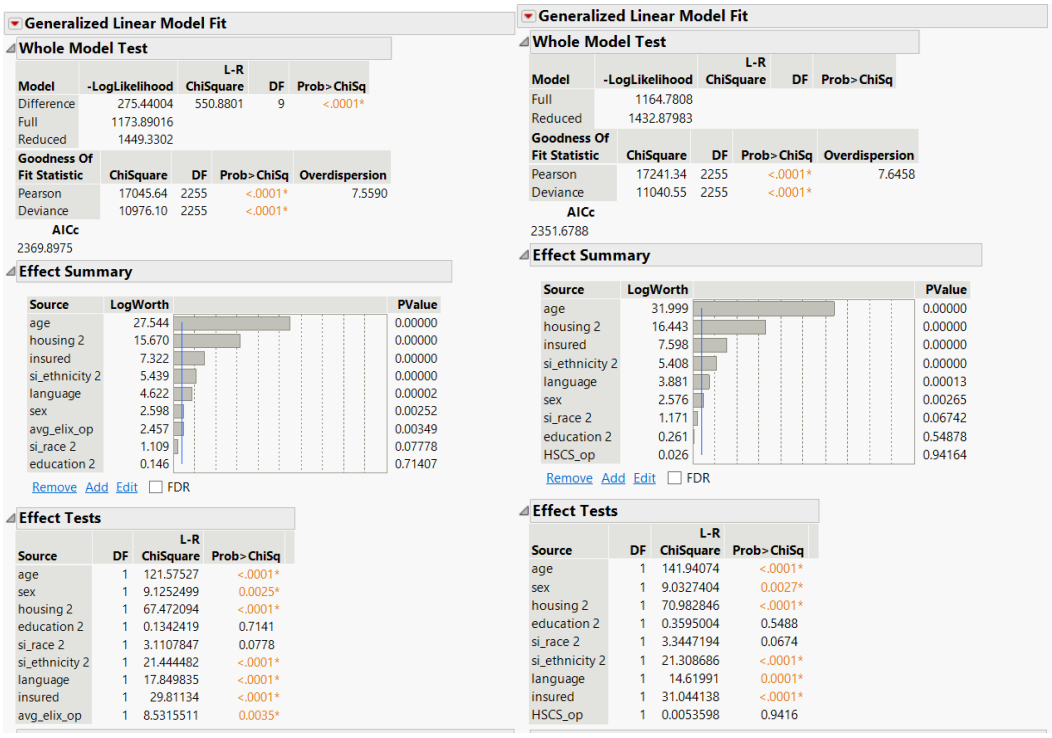


Figure 9: Comparison of the predictive ability of the Elixhauser score (left) and HSCS (right) against the outcome of the number of outpatient/clinic visits in an adjusted model with demographic information. Only Elixhauser (left) retained significance with a small effect size.

3.5 Features Created from CPT Codes

The outpatient claims data also contained CPT codes for the outpatient/clinic service dates. CPT stands for "Current Procedural Terminology." The coding scheme was first developed in 1966 by the American Medical Association (AMA) to track healthcare utilization and identify services for payment (Dotson, 2013). The codes specify the levels of visits from one to five, where a level one visit requires the least amount of clinician time and the lowest amount of complexity, and level five requires the greatest. The codes also specify any procedures performed or treatments given and may identify the type of service, such as 'psy' for a psychiatric visit. The features developed from this data include:

- "TREAT_VISIT_op" – a variable indicating the number of a patient's outpatient visits that had one or more CPT codes indicating a medical treatment was provided, such as respiratory / nebulizer treatments, antibiotic administration, surgical destruction, or joint injection.
- "PREV_VISIT_op" – a variable indicating the number of a patient's outpatient visits that had one or more CPT codes indicating preventive care was provided, such as immunization, preventive injection, screening (such as hearing or vision), or episode for testing.
- "visit_level_op" – a variable indicating the average level of all of the patient's outpatient visits with each visit's level coded between one and five as follows, based on the 2018 Office and Outpatient E/M guidelines (American College of Surgeons, 2020). (Note: In 2021, The 201 & 202 and 211 & 212 codes were combined into a single visit level reflecting "straightforward" decision-making on the part of the clinician and a lower level of visit complexity irrespective of time spent).

Visit Level	CPT Codes
5	99205, 99215, 90839, 99243, 99205psy, 99215psy, 99215MH
4	99204, 99214, 90792, 99242, 90837, 99204psy, 99214psy, 99214MH, 99214MAT
3	99203, 99213, 99213MAT, 99213MH, 90834, 99203psy, 99213psy, 99381-99387, 99391-99397, G0438, H0007, 90791
2	99202, 92012, 99212, 99212MAT, 99402CM, 99402MAT, 90832, 90832MAT, 991212psy, 98960, 99408, 99407, 99402, H0049, G0108
1	All other visits

Table 4: CPT Codes in Outpatient Claims and Corresponding Visit Levels

As many outpatient/clinic visits combined treatments and preventive care, the "TREAT_VISIT_op" and "PREV_VISIT_op" variables were 73% correlated with Kendall's tau. This correlation reflects the HCHM care team's determination to assist patients with preventive care needs whenever they are present in the clinic.

3.6 The Role of Weather Data

To get an accurate picture of the relationship between emergency and clinic utilization and diagnoses, clinical measures, and life circumstances of the Health Care for the Homeless of Manchester (HCHM) patient population, it is desirable to consider adjusting for external visit reasons. Many people speculate that homeless people gravitate toward the emergency department seeking shelter during extreme weather. Measuring the truth of this claim is complicated because no authoritative definition of extreme weather exists. This is not surprising, because the definition of extreme weather can vary in different parts of the country or world depending on what kind of weather the population in that area is accustomed to. To attempt to count extreme weather days during the two years, I downloaded weather data for Manchester, NH, from the National Weather Service (NWS) from January 1, 2018, through December 31, 2019, and arranged it into a continuous data set. The NWS data provided weather information

for the city for each of the 730 days in the sample period, including date, maximum and minimum temperatures, average temperature, and departure from typical temperatures on that date in previous years (all in degrees Fahrenheit). It also provided precipitation, new snow, and existing snow (all in inches). To produce a definition of extreme weather that would fit this data set, I looked at heat and cold advisory guidelines from the Centers for Disease Control and Prevention (CDC) and regional authorities (CDC, 2017; City of Manchester Health Department, 2021), before settling on the following criteria, at least one of which needed to be met on a given day for that day to be considered an "extreme weather" day:

- Max temperature > 90 degrees Fahrenheit
- Max temperature < 28 degrees Fahrenheit (National Weather Service, n.d.)
- Absolute daily temperature departure > 20 degrees Fahrenheit
- Precipitation (rain) > .9 inches
- New snow > 3 inches
- Snow depth > 10 inches

These criteria identified 114 out of the 730 visit days as "extreme weather" days. In a graph of average temperature data over time, the seasonal trend is apparent:

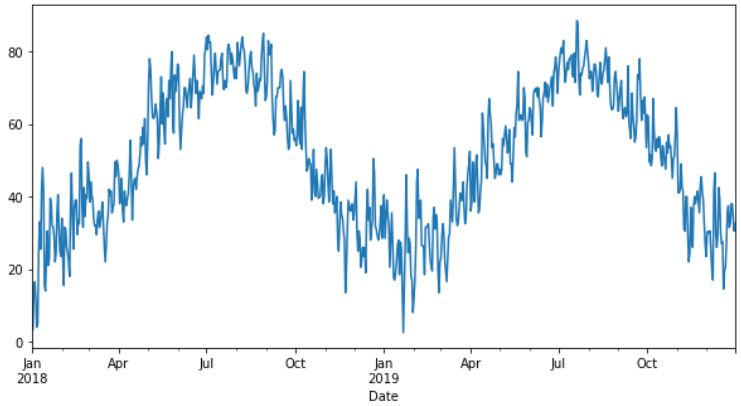


Figure 10: Trend and Seasonality of two years of average temperature data – Manchester, NH

For each of the 114 days designated extreme weather days, a flag identifying visits on these days was added to the longitudinal data set. For the patient-level aggregate data, these flags were summed, and each patient's total number of extreme weather visit days was recorded in the feature "extreme_days."

3.7 Metrics Derived from Clinical Notes

Over ten thousand clinical notes were provided for outpatient/clinic visits taking place during the two years. The notes, derived from the Centricity medical record, contained summaries of a variety of visits, ranging from nurse visits for education or immunization, to physical examinations, treatment visits, testing and screening visits, and behavioral health, psychotherapy, and social services visits. Approximately 96% of the 2,265 patients had at least one outpatient physical examination encounter recorded during the two years. Notes from these visits contained vital signs (height, weight, BMI, temperature, blood pressure, pulse, and oxygen saturation), pain assessments, physical examination details, medical histories, and smoking statuses. In addition, they contained important screening data, including NIDA TAPS (McNeely et al., 2016) scores to screen for substance use and PHQ-2 scores to screen for depression (Kronke, Spitzer & Williams, 2003). A few patients had hemoglobin A1C results (n=233), random office blood glucose readings (n=232), or results from a HARK screening (Sohal, Eldridge & Feder, 2007) for intimate partner violence (n=465). These variables are of great interest in assessing behavioral factors.

I used regular expressions to extract data points from the notes, then summed or averaged values across visits as appropriate. Instead of trying to average BMI values, height and weight values for each patient were extracted and averaged across visits. BMI values were then calculated from these averaged values wherever both values were available.

Three word-count variables were produced from the notes data. For each outpatient visit, all of the mentions of the words or lemmas "pain," "jail," or "incarcerate," and "disability" were counted. The total number of instances of the words or closely-related words with the same stem were counted for each visit, and the counts were averaged across outpatient visits for each patient. The idea is that one or two mentions of a word might be an assessment (perhaps the clinician asking the patient, "Are you in pain?" or "Do you have any incarceration history?" for example), but repeated mentions may indicate that the patient is reporting a problem with pain, has had recent corrections system involvement, or may struggle with a mental or physical disability.

Providers, including nurses, nurse practitioners, physicians, social workers, and counselors, signed their visit notes and these signatures were not anonymous. Because the provider a patient sees can be a key factor in their care outcomes, I wanted to record provider information for as many visits as possible. However, I did not want providers' names outwardly exposed in the analysis or presentation(s) to reduce bias and protect privacy. To achieve both goals, I applied a weak hashing algorithm (SHA) to the primary providers' names for each visit note and stored the hashed representations of the providers' and mental health providers' names in the longitudinal data set. For the patient-level data, I wrote a function that examined all of the hashed provider and mental health provider representations for each patient's visits and chose the provider and mental health provider most frequently associated with them. Some patients had only one outpatient visit or did not have a most commonly used provider or mental health provider. The first provider or mental health provider listed was assigned for these patients. Table 5 lists the features derived from notes data and the number of patients represented by each.

Variable Name	Description	Number of patients with at least one value	% represented (n=2265)
avg_height	Average height of patient	2,185	96.5%
avg_weight	Average weight of patient	2,183	96.4%
avg_pain_ct	Average number of times the word or lemma 'pain' appeared in the patient's visit notes	2,173	95.9%
avg_jail_ct	Average number of times the word or lemma 'jail' or 'incarcerate' appeared in the patient's visit notes	896	39.6%
avg_disability_ct	Average number of times the word or lemma 'disability' or 'SSDI' appeared in the patient's visit notes	193	8.5%
tobacco	Patient's most recent smoking status at the end of the sample period, as assessed in clinic and recorded in notes (2,196) or available through an ED diagnosis code(s)	2,210	97.5%
avg_nida	Average NIDA TAPS screening result	1,857	82.0%
avg_a1c	Average HgbA1C	233/338 with DM_ed or op dx(s)	68.9% of DM patients
avg_obg	Average office blood glucose reading	232/338	68.6% of DM patients
avg_phq2	Average depression screening result	1,678	74.1%
avg_hark	Average screening result for domestic and intimate partner violence	465	20.5%
calc_BMI	Calculated BMI using avg_height and avg_weight	2,155	95.1%
avg_systolic_bp	Average systolic blood pressure reading	2,206	97.4%
avg_diastolic_bp	Average diastolic blood pressure reading	2,206	97.4%
mf_provider	Most frequent provider	1,704	75.2%
mf_mhprovider	Most frequent mental health provider	584	25.8%

Table 5: Features Derived from Outpatient Visit Notes

3.8 The Final Feature Set

The final patient-level feature set, with ninety-two (92) variables is shown in Table 6. Not all of these variables were used in every step of the analysis process. Some were used only for descriptive purposes.

Domain	Variable	Value Description	NULLs	Used in Classification
Identifier	patient_id	Consistent, de-identified patient identifier	no	no
Visit Count and Interval Features	ed_visit_count	Total # of emergency department visits	no	yes
	op_visit_count	Total # of outpatient/clinic visits	no	yes
	avg_ed_interval	Average # of days between ED visits, censorship value = 730 days	no	yes
	avg_op_interval	Average # of days between clinic visits, censorship value = 730 days	no	yes
	avg_ip_interval	Average # of days between ED visits that became inpatient admissions, censorship value = 730 days	no	yes
	admitted	Number of times patient was admitted to the hospital from an ED visit	no	yes
	deceased	0: Patient alive at the end of the period 1: Patient died during the period	no	yes
	non_emergent_dx	Number of times patient went to the ED for a non-emergent diagnosis	no	yes
	extreme_days	Number of times patient had a visit on an extreme weather day	no	yes
	TREAT_VISIT_op	Number of outpatient treatment visits during the period	no	yes
	PREV_VISIT_op	Number of outpatient preventive visits during the period	no	yes
	ed_visit_group		no	no
	0	No ED visits		
	1	One ED visit		
	2	Two ED visits		
	3 or 4	3 or 4 ED visits		
	5 - 7	5 – 7 ED visits		
	8 - 30	8 – 30 ED visits		
	> 30	More than 30 ED visits		
	op_visit_group		no	no
	1	One clinic visit		
	2	Two clinic visits		
	3 or 4	3 or 4 clinic visits		
	5 - 7	5 – 7 clinic visits		
	8 - 30	8 – 30 clinic visits		
	> 30	More than 30 clinic visits		
Conditions and Acuity	visit_level_op	Average outpatient visit level for the patient	yes	yes
	avg_pain_ct	Average number of times the word or lemma ‘pain’ is mentioned in the patient’s outpatient notes	no	yes

	avg_disability_ct	Average number of times the word or lemma 'disable' or 'SSDI' is mentioned in the patient's outpatient notes	no	yes
	avg_elix_ed	Average Elixhauser score for patient's emergency visits; zero if no visits	no	yes
	avg_elix_op	Average Elixhauser score for patient's clinic visits; zero if preventive visits only	no	yes
	HSCS_ed	Average # of homeless-specific visit categories per emergency visit; zero if no visits	no	yes
	HSCS_op	Average # of homeless-specific visit categories per clinic visit; zero if preventive visits only	no	yes
	Emergency Visit Reason Categories: **See Table 3 for ICD-10 code details**		no	yes
	CVD_ed	Total number of times a cardiovascular disease-related code was applied to a patient's emergency visits		
	HF_ed	Total number of times I50 (heart failure) was applied to a patient's emergency visits		
	HTN_ed	Total number of times a hypertension-related code was applied to a patient's emergency visits		
	URI/PNA_ed	Total number of times an acute respiratory disease-related code was applied to a patient's emergency visits		
	AS/COPD_ed	Total number of times a chronic respiratory disease-related code was applied to a patient's emergency visits		
	NEURO_ed	Total number of times a neurological disease-related code was applied to a patient's emergency visits		
	SUD_ed	Total number of times a substance use-related code was applied to a patient's emergency visits		
	MHD_ed	Total number of times a mental health-related code was applied to a patient's emergency visits		
	CA_ed	Total number of times a cancer-related code was applied to a patient's emergency visits		
	PREG_ed	Total number of times an obstetric code was applied to patient's emergency visits		
	DM_ed	Total number of times a diabetes-related code was applied to a patient's emergency visits		
	INF_ed	Total number of times an infection-related code was applied to a patient's emergency visits		

	LIV_ed	Total number of times a liver, gallbladder, or pancreatic disease-related code was applied to a patient's emergency visits		
	REN_ed	Total number of times a renal/kidney disease-related code was applied to a patient's emergency visits		
	COG_ed	Total number of times a cognitive disease or deficit-related code was applied to a patient's emergency visits		
	SENS_ed	Total number of times a sensory disease or deficit-related code was applied to a patient's emergency visits		
	PAIN_ed	Total number of times a pain or pain syndrome-related code was applied to a patient's emergency visits		
	TOB_ed	Total number of times tobacco use was recorded during a patient's emergency visits		
	ACC/INJ_ed	Total number of times an accident or injury-related code was applied to a patient's emergency visits		
	Outpatient/clinic Visit Reason Categories: **See Table 3 for ICD-10 code details**		no	yes
	CVD_op	Total number of times a cardiovascular disease-related code was applied to a patient's clinic visits		
	HF_op	Total number of times I50 (heart failure) was applied to a patient's emergency visits		
	HTN_op	Total number of times a hypertension-related code was applied to a patient's clinic visits		
	URI/PNA_op	Total number of times an acute respiratory disease-related code was applied to a patient's clinic visits		
	AS/COPD_op	Total number of times a chronic respiratory disease-related code was applied to a patient's clinic visits		
	NEURO_op	Total number of times a neurological disease-related code was applied to a patient's clinic visits		
	SUD_op	Total number of times a substance use-related code was applied to a patient's clinic visits		
	MHD_op	Total number of times a mental health-related code was applied to a patient's clinic visits		
	CA_op	Total number of times a cancer-related code was applied to a patient's clinic visits		
	PREG_op	Total number of times an obstetric code		

		was applied to patient's clinic visits		
	DM_op	Total number of times a diabetes-related code was applied to a patient's clinic visits		
	INF_op	Total number of times an infection-related code was applied to a patient's clinic visits		
	LIV_op	Total number of times a liver, gallbladder or pancreatic disease-related code was applied to a patient's clinic visits		
	REN_op	Total number of times a renal/kidney disease-related code was applied to a patient's clinic visits		
	COG_ op	Total number of times a cognitive disease or deficit-related code was applied to a patient's clinic visits		
	SENS_op	Total number of times a sensory disease or deficit-related code was applied to a patient's clinic visits		
	PAIN_op	Total number of times a pain or pain syndrome-related code was applied to a patient's clinic visits		
	TOB_op	Total number of times tobacco use was recorded during a patient's clinic visits		
	ACC/INJ_op	Total number of times an accident or injury-related code was applied to a patient's clinic visits		
Clinic Measures (derived from notes)	avg_height	Average height of the patient across their outpatient/clinic visits	yes	yes
	avg_weight	Average weight of the patient across their outpatient/clinic visits	yes	yes
	tobacco	Most recent tobacco use status for patient, derived either from clinical notes or from an emergency visit diagnosis code	yes	yes
	avg_nida	Average NIDA / TAPS substance use screening score	yes	yes
	avg_a1c	Average hemoglobin A1C value (%)	yes	yes
	avg_obg	Average office / random blood glucose reading (mg/dL)	yes	yes
	avg_phq2	Average PHQ-2 depression screening score	yes	yes
	avg_hark	Average HARK screening score for intimate partner violence	yes	yes
	calc_BMI	BMI calculated from average height and weight, where both values are available (kg/M^2)	yes	yes
	avg_systolic_bp	Average systolic blood pressure reading	yes	yes

		(mmHg)		
	avg_diastolic_bp	Average diastolic blood pressure reading (mmHg)	yes	yes
	mf_provider	Most frequent primary provider (hashed value)	yes	yes
	mf_provider 2	Coded version of most frequent primary provider; providers with >1 patient visits coded 0-7 with provider with the most patients coded 0, and remaining providers coded 1-6. Other less common providers grouped as code 7.		
	mf_mhprovider	Most frequent mental health provider (hashed value)	yes	yes
	mf_mhprovider 2	Coded version of most frequent mental health provider; providers with five or more patient visits coded 0-10 with provider with the most patients coded 0, and remaining providers coded 1-9. Other providers grouped as code 10.	yes	yes
Demographic Information	veteran	0: Not a veteran or unknown 1: Patient is a veteran	no	yes
	age	Patient's age at the end of the study period	no	yes
	age_group – Grouped ages		no	no
	<10	Less than 10 years old		
	10-18	10 – 18 years		
	19-25	19 – 25 years		
	26-35	26 – 35 years		
	35-45	35 – 45 years		
	46-55	46 – 55 years		
	56-65	56 – 65 years		
	>65	65 years or older		
	age_group 2 – Coded version of age_group		no	no
	<10	-3		
	10-18	-2		
	19-25	-1		
	26-35	0: Reference group (most common)		
	35-45	1		
	46-55	2		
	56-65	3		
	>65	4		
	sex	0: male (reference), 1: female, 2: another gender	no	yes
	housing – Housing status		yes	no
	UNKNOWN	NULL or unknown		
	STREET	Street or encampment		
	SHELTER	New Horizons / FIT shelter		

	TRANS/TRTMNT	Transitional or treatment housing		
	DOUBLE-UP	Living with another family in the community or couch-surfing		
	OTHER/SRO	Living in a rooming house, hotel or apartment		
	housing 2 – Coded version of housing variable		no	no
	UNKNOWN	-1		
	STREET	0		
	SHELTER	0		
	TRANS/TRTMNT	1		
	DOUBLE-UP	2		
	OTHER/SRO	3		
	education – Highest completed education		yes	no
	UNKNOWN	NULL or unknown		
	0-8G	No education through 8th grade		
	9-12G	9th – 12th grade		
	HS/GED	Graduated high school or obtained GED		
	COL	Any amount of college		
	education 2 – Coded version of education variable		no	no
	UNKNOWN	-1		
	0-8G	0		
	9-12G	1		
	HS/GED	2		
	COL	3		
	si_race		no	no
	WHITE	white		
	BLACK	Black or African American		
	ASIAN	Asian or South Asian		
	MULTIPLE	More than one		
	OTHER/UNK	Another race, unknown or declined (n=1)		
	AI/AK	American Indian or Alaska Native		
	NH/PI	Native Hawaiian or Pacific Islander		
	si_race 2 – Coded version of self-identified race variable		no	yes
	WHITE	0 (most frequent)		
	BLACK	1		
	ASIAN	2		
	MULTIPLE	3		
	OTHER/UNK	4		
	AI/AK	5		
	NH/PI	6		
	si_ethnicity		no	no
	NONHISPANIC	Non-Hispanic		
	OTHER/UNK	Another ethnicity or unknown		
	HISPANIC	Hispanic		
	si_ethnicity 2 – Coded version of self-identified ethnicity variable		no	yes
	NONHISPANIC	0 (most frequent)		

	OTHER/UNK	1		
	HISPANIC	2		
	language – Primary language		yes	yes
	0	English (most frequent)		
	1	Another language		
	insured – Patient has insurance		no	yes
	0	No insurance, self-pay, sliding scale, or unknown		
	1	Any insurance		
	incarcerated – any corrections history		no	yes
	0	No corrections history indicated		
	1	Corrections history of any length		
	avg_jail_ct	Average number of times the word or lemma ‘jail’ or ‘incarcerate’ is mentioned in the patient’s outpatient notes	no	yes

Table 6: Final Features Set – Patient-Level Data

CHAPTER 4: DESCRIPTION

4.1 Demographics

In several ways, the demographics of the patient population matched those of the part of Manchester (Kalivas Union and Center City) where the Health Care for the Homeless (HCHM) clinics are located. For example, the self-identified race for the sample is 58% non-Hispanic white, and for the area between 35% and 58% non-Hispanic white (City Data, 2019), while for the city in general, 82% (U.S. Census Bureau, 2020). Likewise, the median age for the general population of Manchester is 38.7 years (City Data, 2019), and the median age for the HCHM patient sample is a very-similar 42 years. For Manchester as a whole, the Census identified 10.7% of the city's population as Hispanic (U.S. Census Bureau, 2020); for the HCHM patient sample, a slightly lower percentage, 7.2%, self-identified this way.

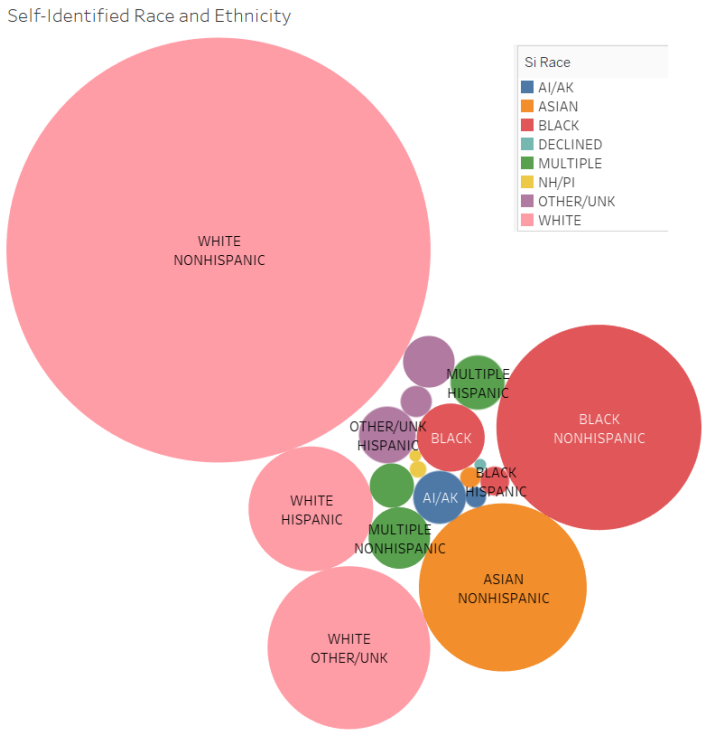


Figure 11: Self-Identified Race and Ethnicity

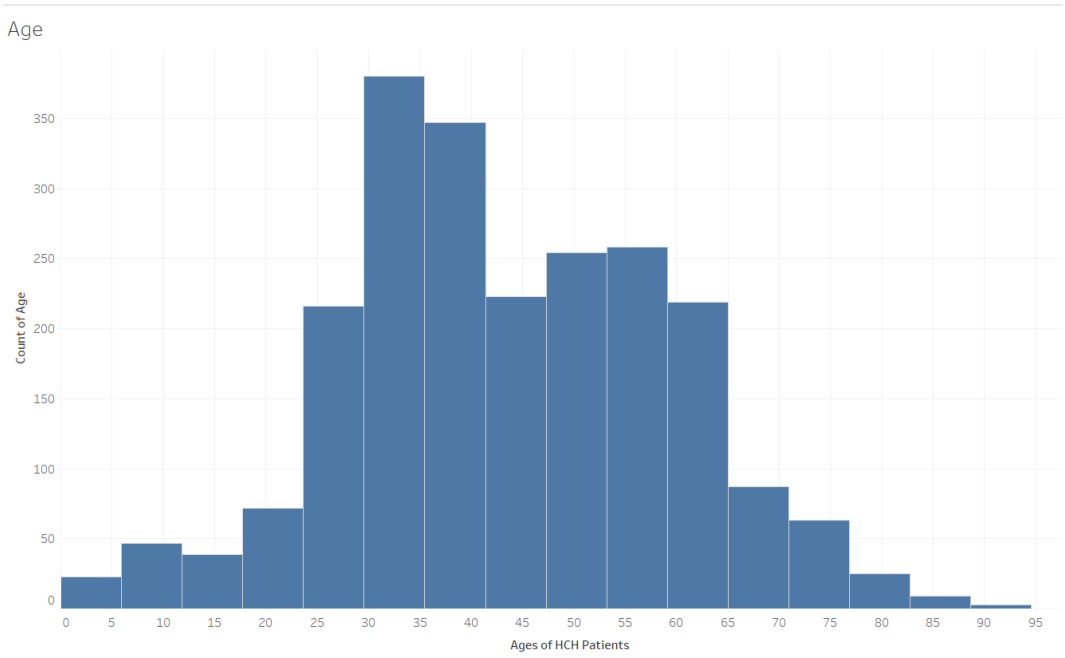


Figure 12: Distribution of Age

One point of departure is the proportion of males to females in the HCHM patient sample. Part of the reason for this skew may be that homelessness is more common in the general population among males than females. According to the 2018 Department of Housing and Urban Development Point-In-Time (PIT) count, the ratio of male to female homeless people in the general population was 2.54 to 1 (Moses & Janosko, 2018). However, another significant reason may be that pregnant women seeking care at the HCHM clinics are often referred to other primary and obstetric care sources. While the city of Manchester as a whole has equal numbers of males and females (there were 103.2 males for every 100 females with a ±3.3 margin of error, according to the 2020 U.S. Census data), the HCHM patient sample's proportion of males to females is 1.7 to 1. There are also far fewer HCHM patients sixty-five and older (9.5%) than people in the general Manchester population of the same age group (13.7%) (City Data, 2019). Reasons for the lack of older adults in the clinic sample may include the availability of primary care from other sources due to Medicare eligibility, and the lower life expectancy of homeless

people. According to a 2017 study by Romaszko et al., the average life expectancy for homeless men was about 56.27 years (SD 10.38) and 52.00 years (SD 9.85) for homeless women.

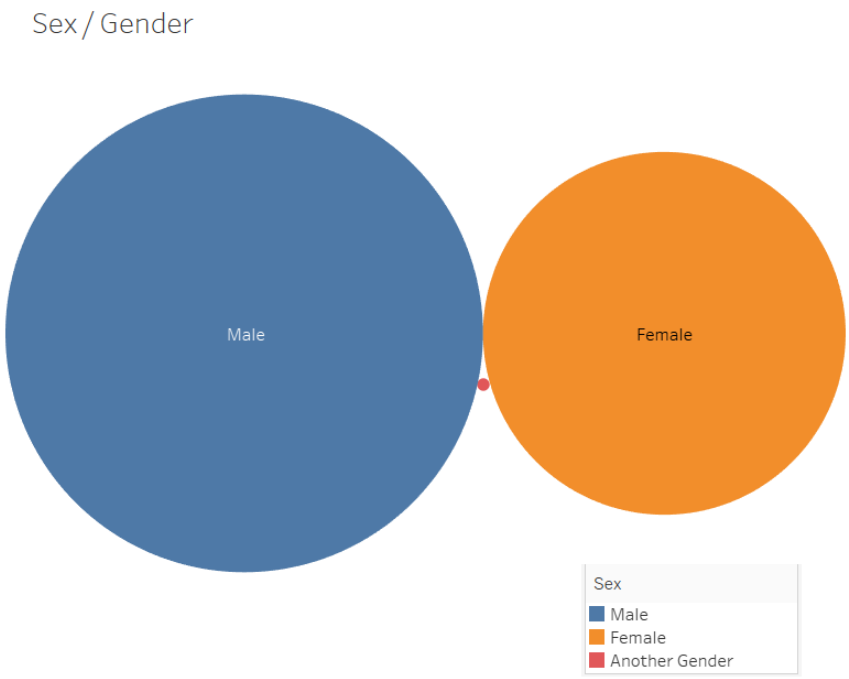


Figure 13: Sex / Gender

The housing and education variables both had many NULL values. The housing variable had 1,667 unknowns (73.6%), and highest level of education completed had 1,688 unknowns (74.5%). Older patients were more likely to have these values recorded, since the clinic stopped recording the values at some time during the two-year period. It was not possible to determine when this occurred, because demographic file entries were not dated. It is possible that because the housing variable value would change often for people with no fixed address, values of this variable were deemed less valuable or accurate.

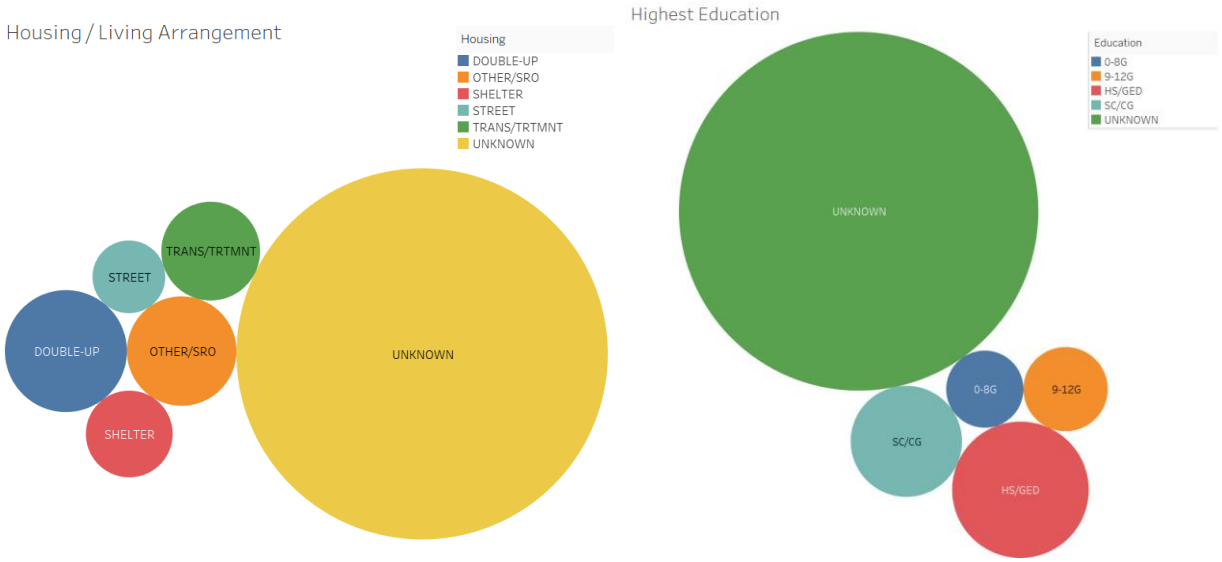


Figure 14: Housing Status and Highest Level of Education

Three hundred fifty-six people (15.7%) reported their primary language as a language other than English. There were nine distinct languages other than English reported as primary by HCHM patients, including Nepali (n=166), Swahili (n=93), Kinyarwanda (n=30), Kirundi (n=18), Spanish (n=17), Yoruba (n=15), and French, Somali or Arabic (n=18).

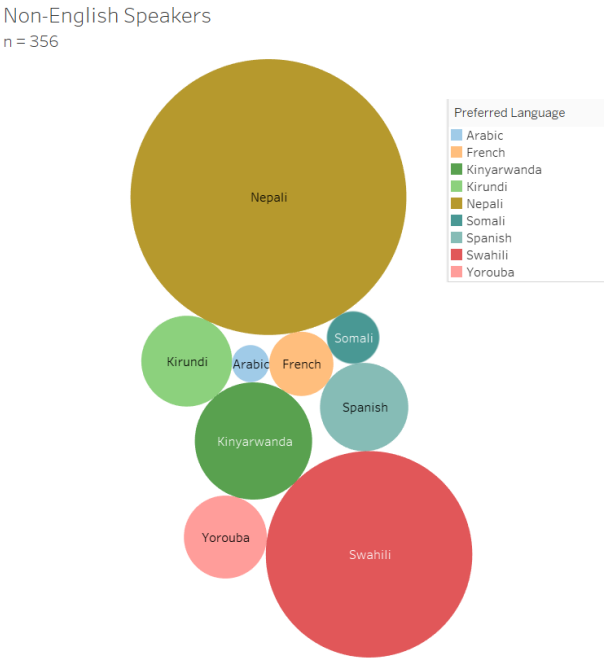


Figure 15: Preferred Languages other than English

Three hundred ninety-six people (17.5% of the sample) had either some corrections history or had the word or lemma “jail” or “incarcerate” mentioned an average of two or more times in their clinic visit notes. Only seventy-seven people in the patient sample (3.4%) were veterans. Many veterans receive free health care from the U.S. Department of Veterans Affairs health system and have no reason to utilize the HCHM clinics.

4.2 Average Clinical Measures

As previously discussed in the Feature Development chapter, many patient-level clinical measures are average measures for each patient, obtained by averaging across visit values. Not all measures had values obtained at every visit. For example, patients who attend social services or behavioral health visits did not have vital signs obtained at these visits. Additionally, average values such as hemoglobin A1C and office blood glucose would only be obtained for patients with diabetes, and screening scores for NIDA TAPS, PHQ-2, and HARK would be obtained where clinician judgment dictated these screenings were indicated. Many patients in the sample (n=790) had only one outpatient office visit. For them, “average” measures were their only measures.

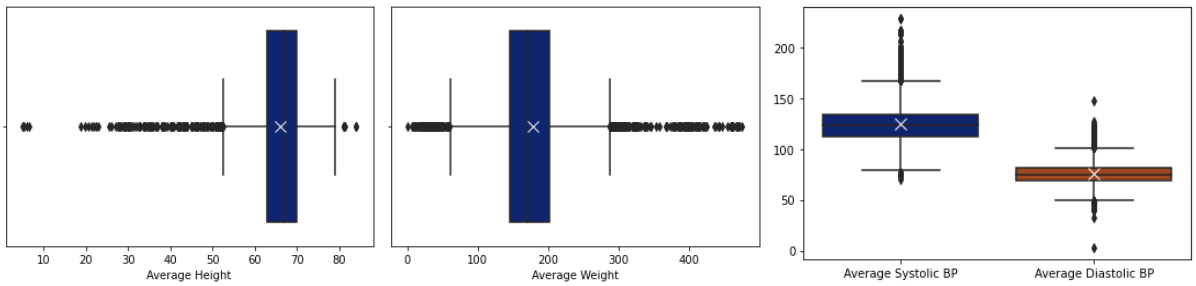


Figure 16: Average Height, Weight, and Blood Pressure Readings with Standard Deviation and Outliers

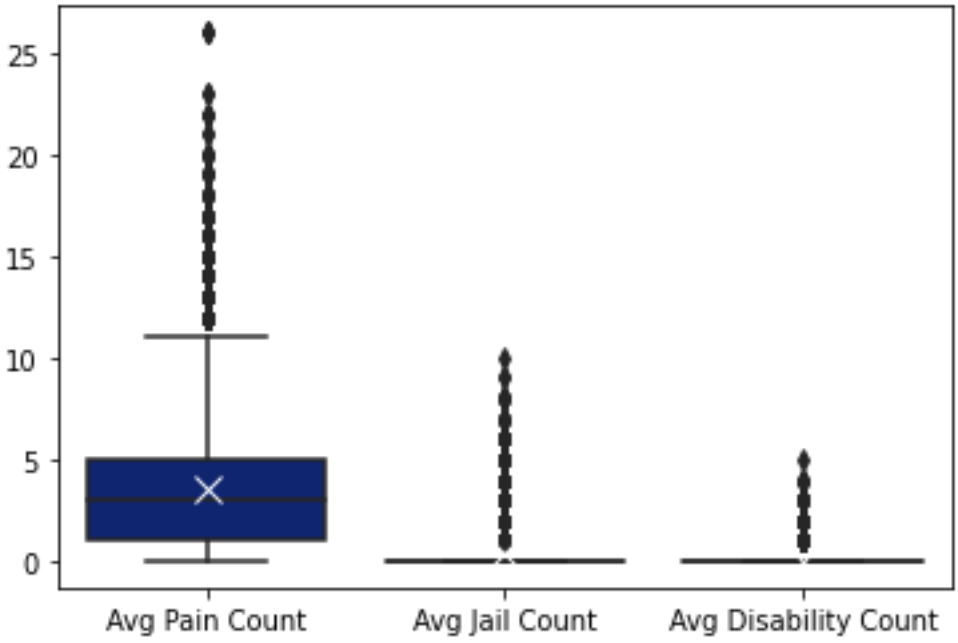


Figure 17: Average Count Variables with Standard Deviation and Outliers

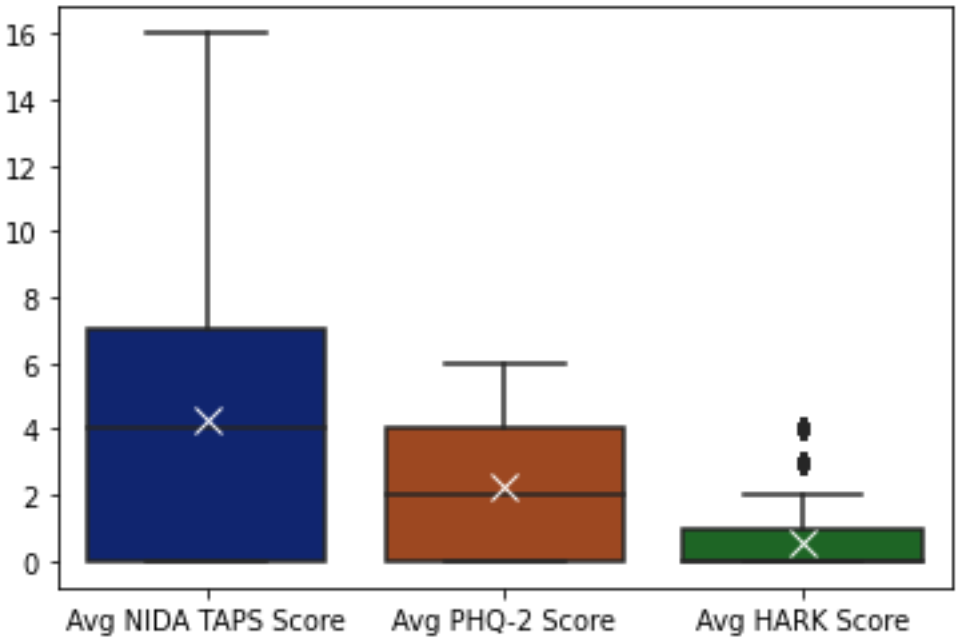


Figure 18: Average NIDA TAPS, PHQ-2, and HARK Screening Scores with Standard Deviation and Outliers

The average height and weight of clinic patients were available for 2,185 patients (96.5% of the sample) and 2,183 patients (96.4%) respectively. These values were skewed by the presence of both children and refugees. The overall mean height for female adult patients who identified English as their primary language was 63 inches (5 feet, 3 inches), and 61 inches (5 feet, 1 inch) for those who did not. However, the overall mean height for male adult patients who identified English as their primary language was 70 inches (5 feet, 10 inches), whereas it was 65 inches (5 feet, 5 inches) for those who did not. Female children with English as their primary language had an average height of 50 inches (about 4 feet, 2 inches). In comparison, those who identified another language as primary had average height of 48 inches (4 feet). For male children, those who reported English as their primary language had an average height of 48 inches, and those with another primary language were similar, at 49 inches. The mean ages for child patients differed, however. The mean ages were eleven years for English-speaking females, eight years for males, and nine years for females with another primary language, ten years for males.

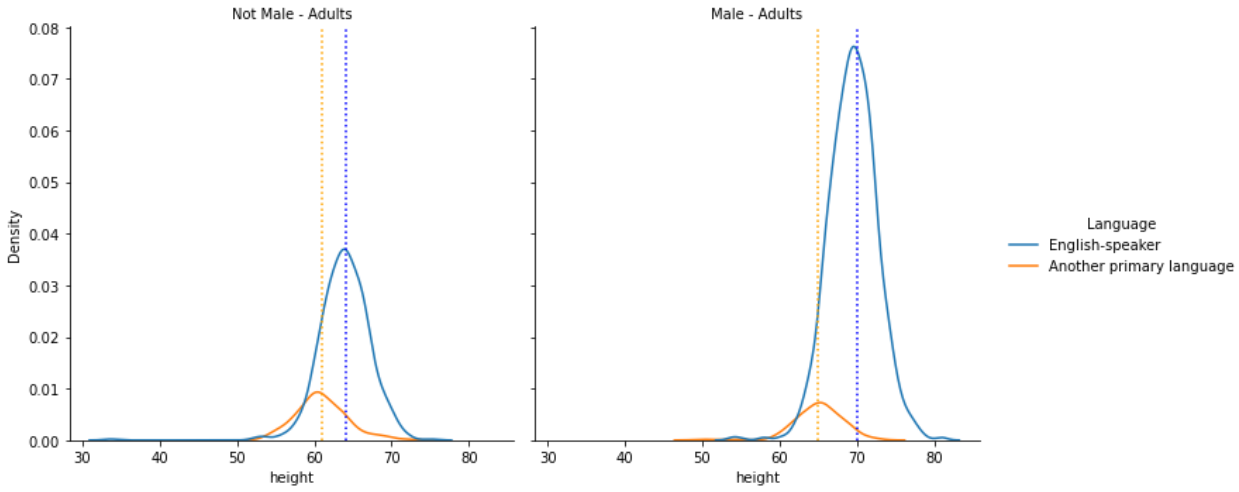


Figure 19: Average Height Comparison – Adults by Gender, Primary Language

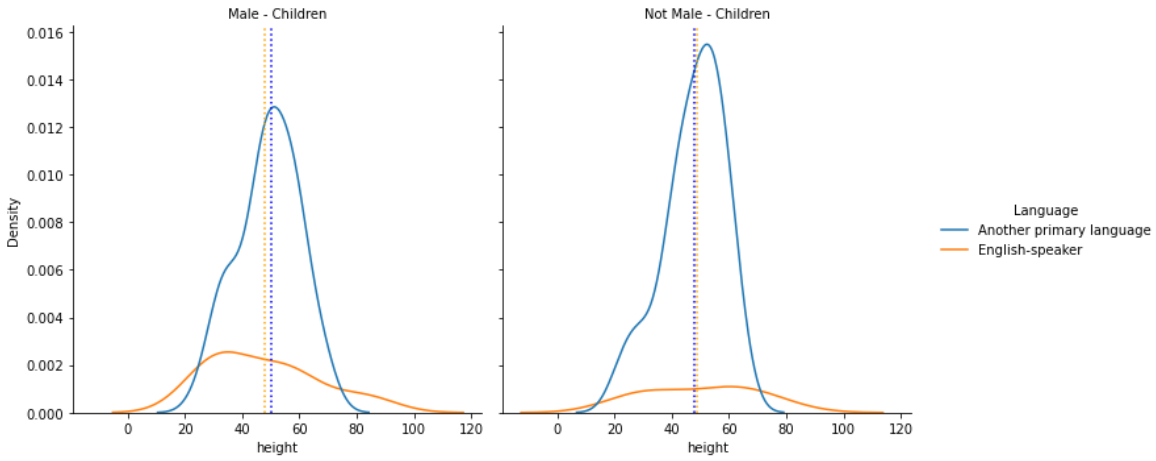


Figure 20: Average Height Comparison – Children by Gender, Primary Language

Just as heights were taller on average for adults, males, and those with English as a primary language, weights were also higher on average. Among those who identified English as their primary language, the average weight for males was 188 pounds, and for females, 167 pounds. For those who identified another language as primary, the average weight for males was 158 pounds, and for females, 153 pounds. Among children, those who identified English as their primary language had average weights of 49 pounds for males (average age eight) and 72 pounds for females (average age eleven), while for those who identified another language as primary, males had an average weight of 67 pounds (average age ten) and females, 57 pounds (average age nine). Many children of refugees – particularly females – had average weights and BMI measurements that placed them below the 75th percentile.

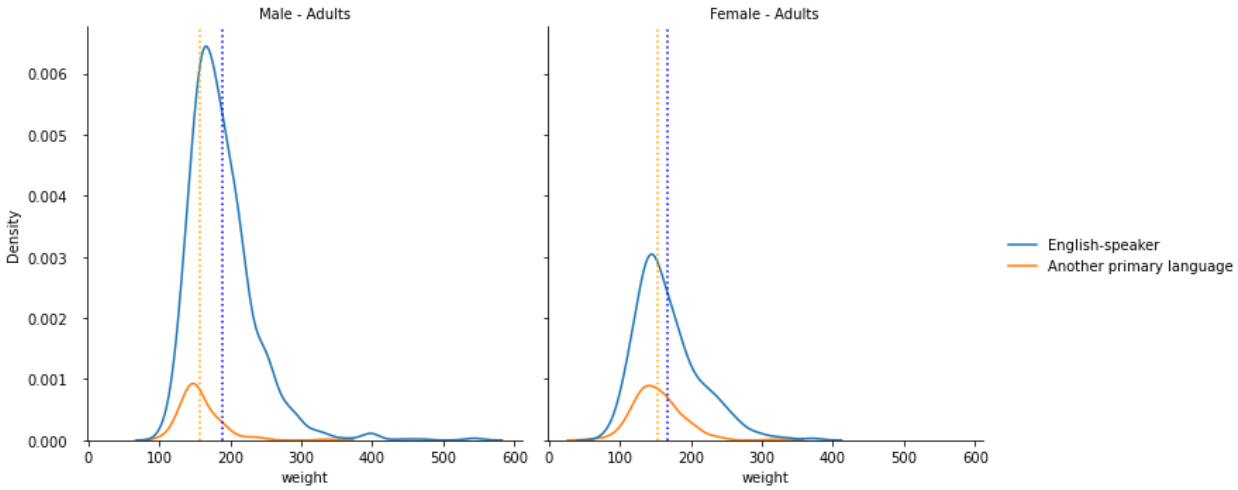


Figure 21: Average Weight Comparison – Adults by Sex, Primary Language

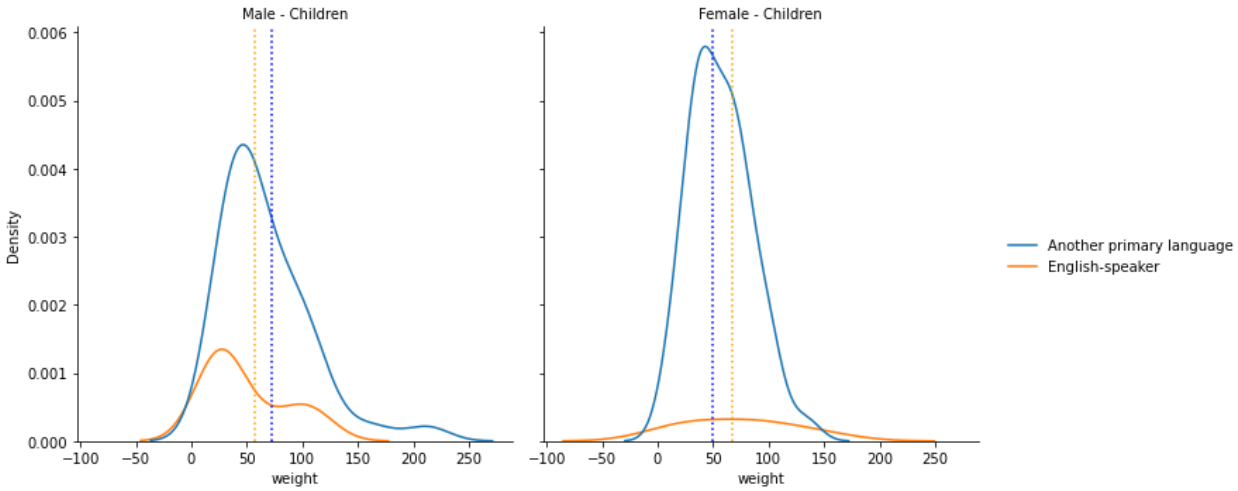


Figure 22: Average Weight Comparison – Children by Sex, Primary Language

Males – avg. ages: 10 for those with another primary language, 8 for English-speakers
Females – avg. ages: 9 for those with another primary language, 11 for English-speakers

BMI values were calculated for adult patients with both average height and weight available (95.1% of the sample). These values show the distribution of BMI for the adult clinic patients as follows: underweight (BMI < 18.5 kg/m^2), 2.2% of adults; normal range (BMI 18.5 to 24.9 kg/m^2), 36.1% of adults; overweight (BMI 25 to 29.9 kg/m^2), 32.9% of adults; and obese (BMI

30 kg/m^2 or greater), 28.8% of adults. Of the 565 patients in the obese category for BMI, 16.8% had a BMI greater than 40.

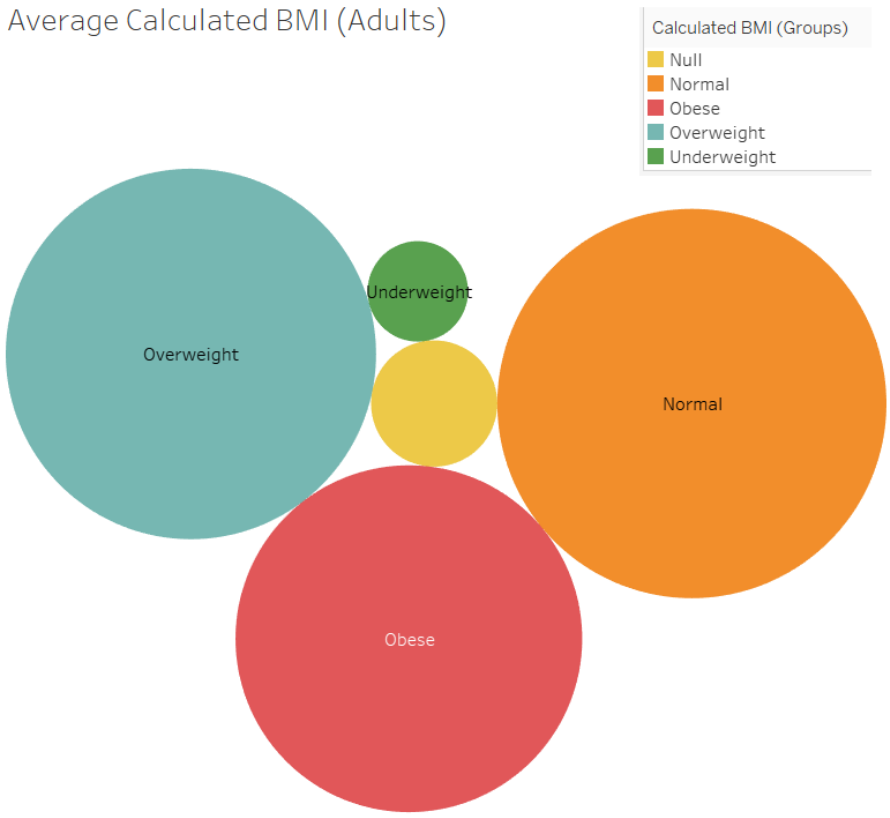


Figure 23: Adult BMI, Calculated from Average Height and Weight

At least one blood pressure reading was available for 2,206 clinic patients (97.4% of the sample). Some patients, particularly those diagnosed with hypertension, were likely to have more check-in visits and, thus, more readings. Because of these repeated readings, some patients' average pressures skewed higher, impacting the overall sample average.

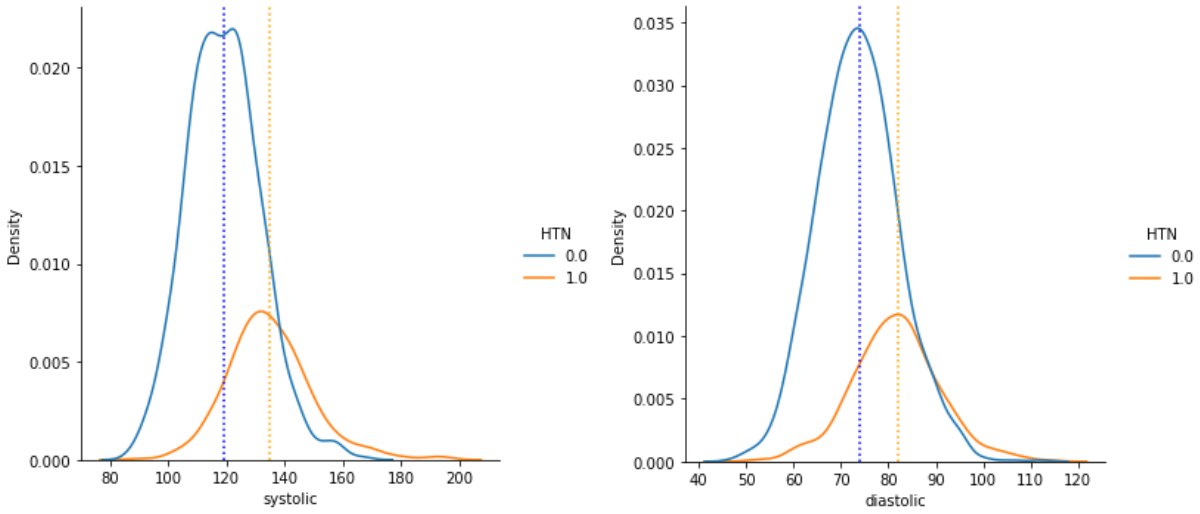


Figure 24: Adults, Average Blood Pressure Systolic (left) and Diastolic (right) by Hypertensive Diagnosis (0=No hypertension, 1=At least one hypertension diagnosis code)

According to the 2021 American Heart Association Guidelines, blood pressure is considered elevated when systolic pressure is repeatedly above 120 mmHg (millimeters of mercury) but below 130 mmHg, and diastolic pressure remains less than or equal to 80 mmHg. Stage 1 hypertension is defined as a systolic pressure greater than or equal to 130 mmHg, but less than 140 mmHg or a diastolic pressure that frequently falls between 81 and 89 mmHg. Hypertension graduates to Stage 2 when systolic pressure is regularly above 139 mmHg but still below 181 mmHg, or diastolic pressure frequently falls between 90 and 120 mmHg. A hypertensive crisis, considered a medical emergency, occurs when systolic pressure exceeds 180 mmHg or diastolic pressure exceeds 120 mmHg. Clinical guidelines for the treatment of hypertension have gradually become more aggressive because studies show that earlier control of hypertension limits the associated risks of heart attack and stroke. It is now recommended that even low-risk adults be treated with medication if they reach guidelines for a Stage 1 hypertensive diagnosis

and lifestyle modification fails to produce a change within three to six months. (Buelt, Richards & Jones, 2021; Goetsch, Tumarkin, Blumenthal & Whelton, 2021).

Among the clinic patients' sample, 543 adults (24.0%) had at least one diagnosis code for a hypertensive disease sometime during the two years. Among these patients, 308 (56.7% of hypertensive patients) had more than one outpatient visit for hypertension and had more outpatient visits than ED visits for the diagnosis. Of the adult patients with a hypertension diagnosis, 226 of them had an average blood pressure of >130 systolic and >80 diastolic (41.6%), and 154 had hypertension that was controlled, with an average reading of ≤130 systolic and ≤ 80 diastolic (28.4%).

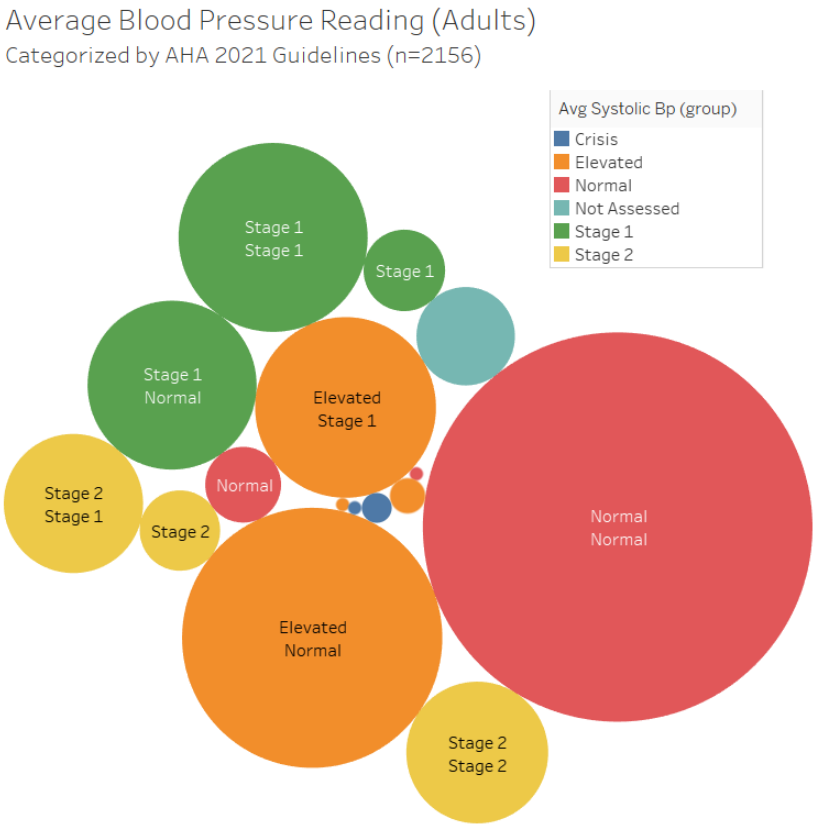


Figure 25: Average Blood Pressure Readings – Adults by AHA 2021 Guidelines
Top Classification: Systolic group, Bottom Classification: Diastolic group

While 543 patients received at least one hypertension diagnosis during outpatient or emergency treatment, many more patients had average readings that qualified as hypertensive. Overall, 375 patients had an average blood pressure that could be classified as "elevated" based on 2021 AHA guidelines, 659 had an average blood pressure that could be classified as "Stage 1" by either systolic or diastolic criteria, and 312 had an average blood pressure that could be classified as "Stage 2" in the same fashion. Of the 369 people with "elevated" average readings, 84 had the diagnosis (22.8%). For those with "Stage 1" average readings, 333 people had received the diagnoses (50.7%), and for those with average readings qualifying as "Stage 2", 232 were diagnosed (79.7%).

Average office blood glucose readings were calculated for 232 people, and average hemoglobin A1C results were available for 233 individuals out of a patient sample where 306 people had one or more emergency department or outpatient visits related to diabetes or gestational diabetes during the two-year period (13.5%). This amounts to laboratory monitoring of the condition for 76.1% of diagnosed patients, a substantial number given a vulnerable and often transient patient population. While more primarily English-speaking patients had diabetes or a diabetes-related diagnosis (229 vs. 77 who reported another language as primary), those whose primary language was not English had lower average office blood glucose values (168 vs. 200 mg/dl) and lower average A1C values (6% glycosylated hemoglobin vs. 7%). Only 14 out of the 77 people with diabetes whose primary language was not English had no HgbA1C recorded the two years (18.2%), while 82 out of the 229 diabetes patients with English as a primary language had no HgbA1C (35.8%). Ninety-one patients out of the 306 with diabetes or a diabetes-related

diagnosis (29.7%) demonstrated control with an average office blood glucose of ≤ 150 mg/dl, and 116 patients had an average HgbA1C of 6% or less (37.9%).

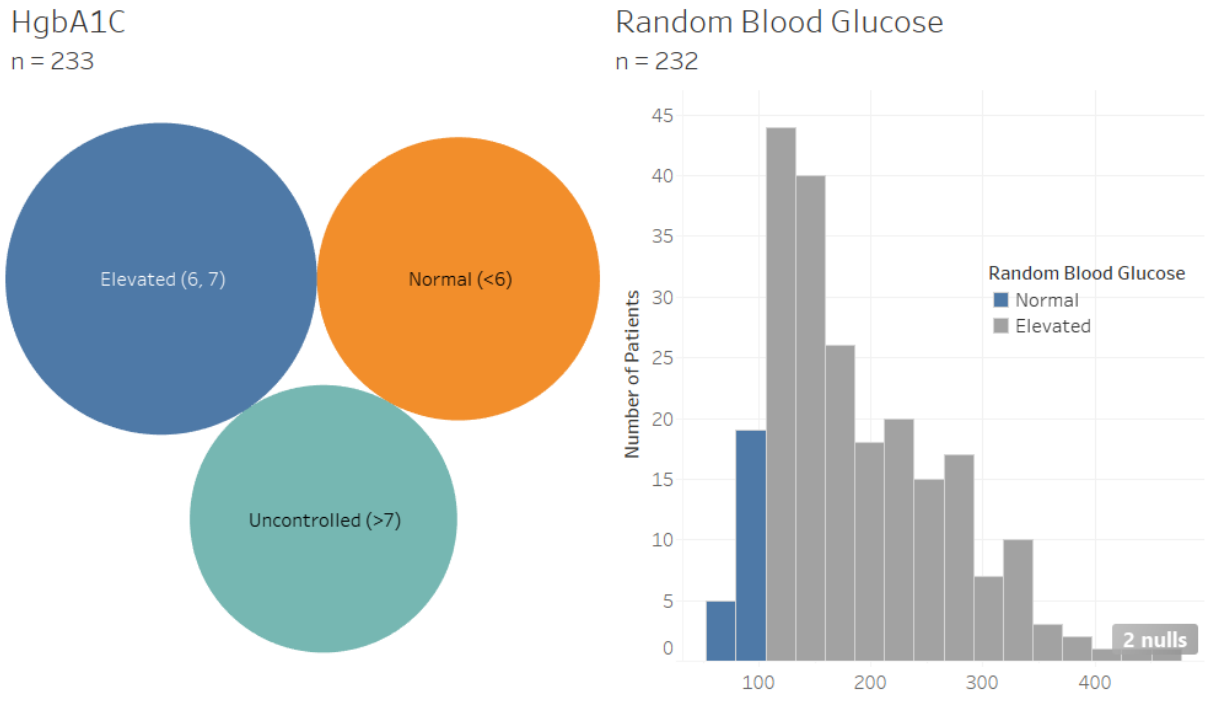


Figure 26: Overall Distribution of Average Values: HgbA1C and Office Blood Glucose

Many patients had at least one screening for substance use (n=1,857), depression (n=1,665), or intimate partner violence (n=465) during the period. These screenings were conducted for adult patients only (n=2,148). The TAPS tool (McNeely et al., 2016) demonstrated 70% sensitivity for detection of DSM-5 substance use of alcohol, tobacco, and marijuana at the 2+ cut-off ("highest-risk" rating). Of those screened in clinic with this tool (86.5% of adults), 476 patients scored 0 (no substance use, past three months) (25.6%), 52 patients scored 1 (at least one instance of problem use) (2.8%), and 1,329 patients scored 2+ (highest-risk for substance use, including tobacco and alcohol) (71.6%).

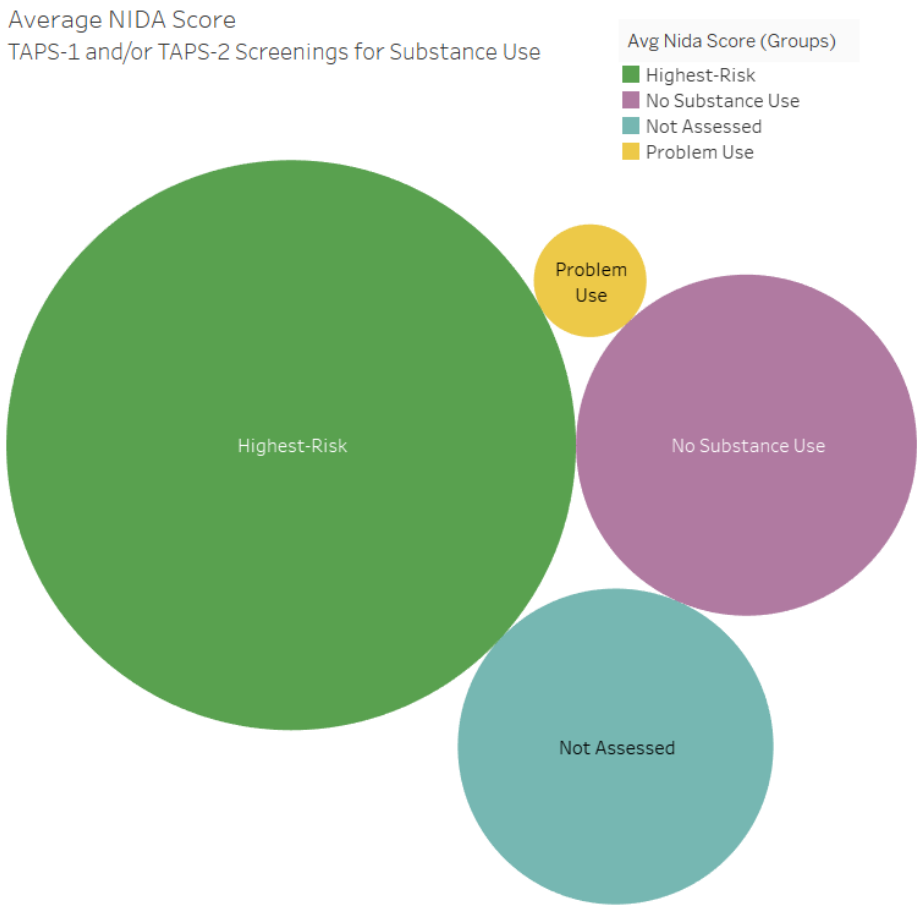


Figure 27: NIDA TAPS Scores Distribution – All Adults (n=2,148)

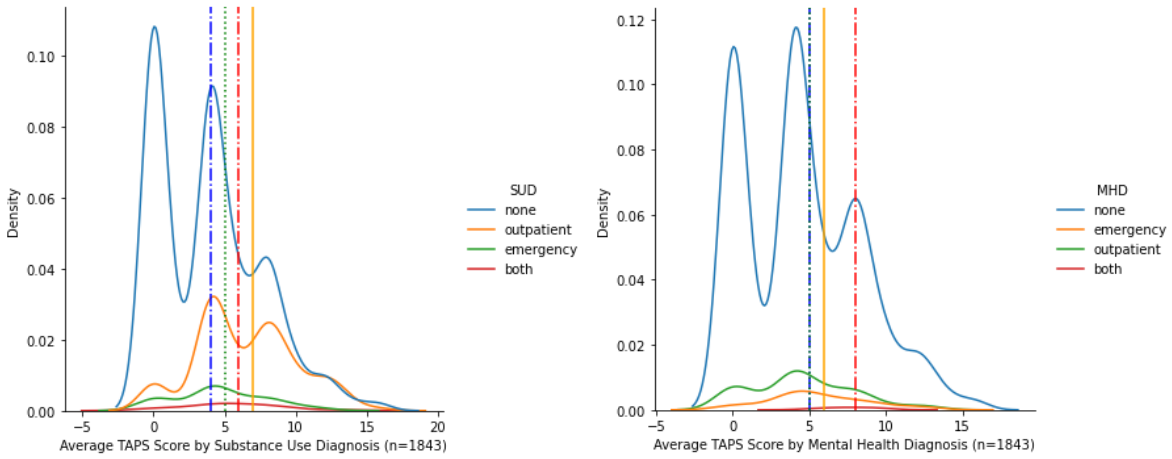


Figure 28: Average NIDA TAPS Scores by Substance Use and Mental Health Diagnoses

Although all average scores were in the "highest-risk" category, average TAPS scores were lowest for patients with either no substance use (4) or mental health diagnoses (5) or those with only outpatient visits for their mental health diagnoses (5). Average scores were highest for patients with only outpatient visits related to substance use (7) and those with outpatient and emergency department visits related to mental health diagnoses (8). The tails of the distributions show that those with the highest TAPS scores either had no substance use or mental health-related visits or had outpatient visits related to their substance use. These often consisted of medication-assisted treatment (MAT) visits or physicals conducted for patients seeking drug treatment.

Of adults screened for depression using the PHQ-2 (Patient Health Questionnaire-2) (n=1,665), a short-form screening tool for depression where a positive score (three or higher) is 38.4% predictive of major depressive disorder and 75% predictive of any depressive disorder (Kroenke, Spitzer & Williams, 2003), six hundred forty (38.4%) patients scored above the threshold. Patients with mental health or substance use diagnoses treated outpatient, or at both the clinic and the emergency department, had average scores of three (positive for depression).

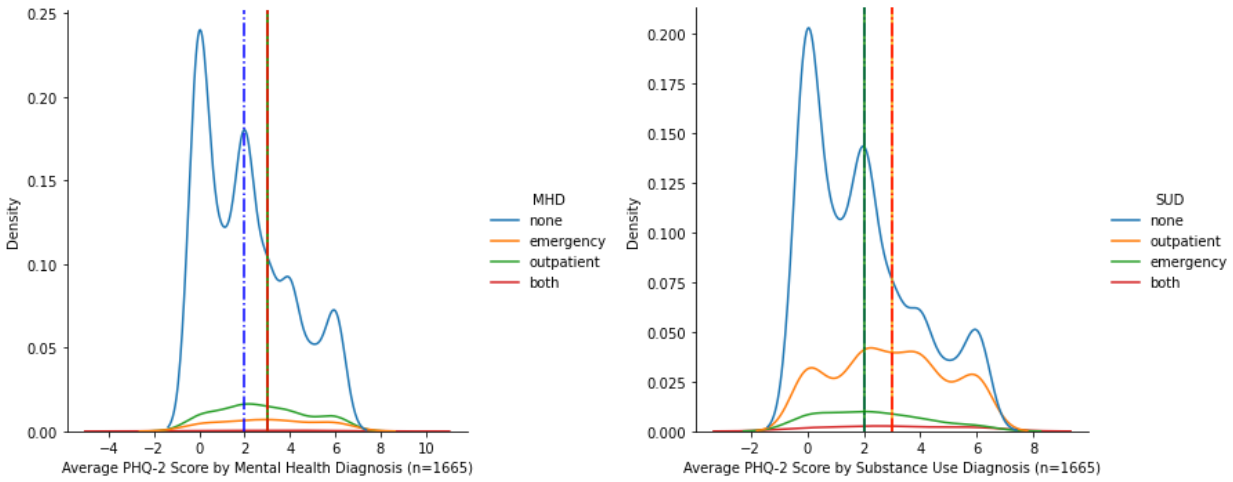
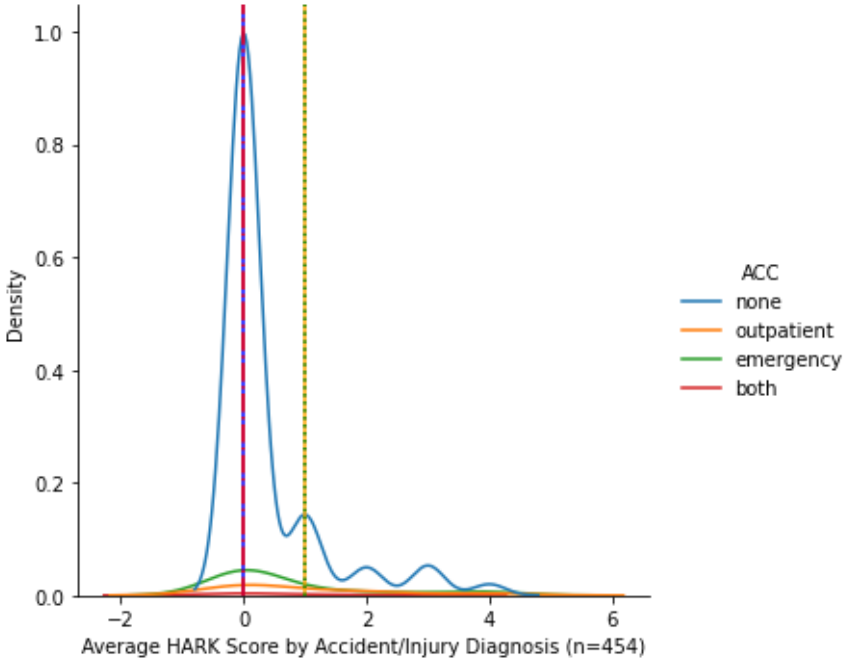


Figure 29: Average PHQ-2 Scores by Mental Health & Substance Use Diagnoses

Tails of these graphs reveal that, unlike NIDA TAPS scores, patients being treated for mental health or substance use disorders were more likely to have the highest average PHQ-2 scores.

The HARK (Humiliation, Afraid, Rape & Kick) is a four-question screening tool for intimate partner violence based on the thirty-question Composite Abuse Scale (CAS). The HARK cut-off score of ≥1 demonstrated a positive predictive value of 83% in detecting intimate partner violence and minimized false positives (Sohal, Eldridge & Feder, 2007). Average HARK scores were derived from the clinical notes for 454 adults (20.0%). Of these, 277 were men (61.0%), and 177 were women (39.0%). Three hundred ninety-two (86.3%) indicated English as their primary language, while only 62 (13.7%) did not. A total of 106 patients had a positive score (≥1) (23.3% of those screened), 43 males and 63 females. Of those with a positive score, only a few individuals indicated a language other than English was primary.



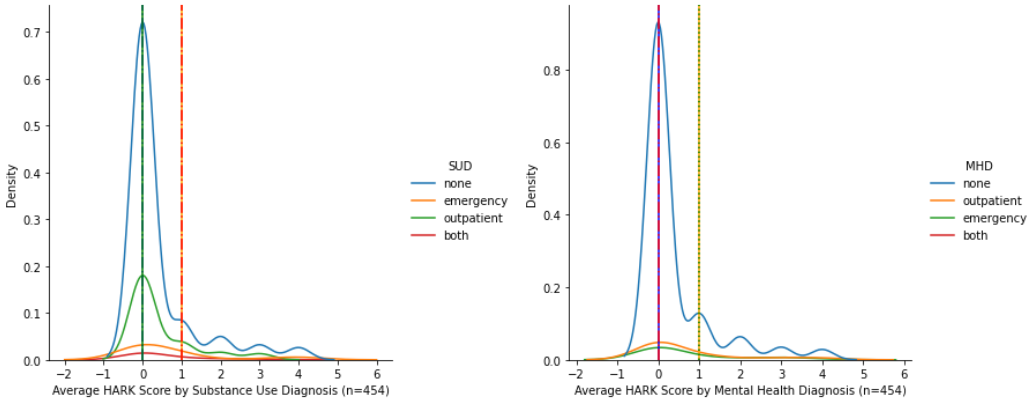


Figure 30: Average HARK Scores by Accidents/Injuries, Mental Health & Substance Use Diagnoses

Among those screened, average HARK scores of 1 (positive) were present for patients who received outpatient or emergency treatment for accidents and injuries (25/69 positive scores) or mental health disorders (26/78 positive scores), and for some patients who received treatment for substance use in the emergency department or both the clinic and the ED (16/47 positive scores).

A total of 2,173 clinic patients (95.9%) had notes from their outpatient visits scanned for particular words or lemmas – base forms of a word that represent all the other forms of the same word. The notes were searched for the word or lemma "pain," "jail" or "incarcerate," and "disable" (to cover both "disability" and "disabled") or "SSDI." The word pain is used at least once in every clinic note, with a single occurrence likely to be associated with a clinical assessment. Since one or two mentions of the word or lemma "pain" might not increase the probability that a patient is struggling with pain issues, I looked more closely at patients with an average of three or more mentions (1,562 patients, or 69.0% of the sample), and separated the distributions by the number of outpatient or emergency visits for pain that each patient had. Interestingly, those seeking treatment for pain only in the ED were less likely to have a high number of pain mentions in their clinic notes. While the mean number of pain-mentions for

those with ED diagnoses only was four, the mean number of pain-mentions for all other patients with three or more mentions (regardless of diagnosis) was five.

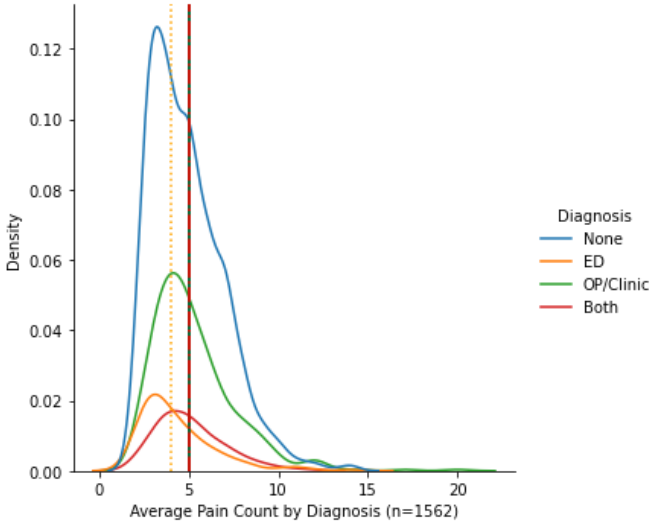


Figure 31: Number of Pain-Mentions by Pain-related Diagnosis Location for all patients with three or more mentions (n=1562)

Only 39.6% of patients (896 people) had one or more mentions of words or lemmas relating to jail or incarceration. Singling out those who had an average of two or more mentions isolated seventy-two people with a likely corrections history. Of these seventy-two people, eleven indicated corrections history in the demographic data, and the other sixty-one did not. Adding these sixty-one people to the number of people indicating a corrections history in the demographic data increased the total number of patients with a history of incarceration to 396 people, or 17.5% of the patient sample.

Since there was no demographic data indicating patients' disability status, I attempted to discover it by searching for any word related to the lemma "disable" or "SSDI," an acronym for Social Security Disability Insurance – the benefit that disabled people receive from the government when necessary. One hundred ninety-three people (8.5% of the patient sample) had

an average of one or more disability-related mentions in their clinic notes, twelve had two or more average mentions, and only two people had three or more mentions. Upon investigation, most of these mentions did not indicate patients with a confirmed disability. Instead, they often had to do with patients consulting social services to attempt to receive the SSDI benefit, with varying degrees of success. Instead of people self-declaring their disability status, most patients who discussed their functional challenges in a visit with a clinician did not describe themselves as disabled in relation to such challenges. While it is understandable that people might not wish to define themselves in terms of their functional challenges, this may be limiting their ability to access services or accommodations that could assist them.

4.3 Visit Reason Counts

All 2,265 patients in the sample had one or more outpatient/clinic visits during the two years; however, many patients had care in the clinic for different conditions than those they sought care for in the emergency department. This is something to keep in mind when interpreting the results of this research. When a table or graph shows the differences in the numbers of patients with a given diagnosis being treated in the emergency department vs. the HCHM clinic, many of the people being treated in the emergency department for a condition are different people from those being treated in the clinic for the same condition. For example, take one patient who had twelve emergency department visits that included codes for hypertension but three clinic visits – two for infection and one for preventive care, but zero related to hypertension.

Another critical point when interpreting these numbers is that they are not counts of only the primary diagnosis for a given visit; instead, they are counts – by homeless-specific diagnostic category – of *all* sample patients' diagnosis codes for all of their visits during the two years. For example, if a patient went to the emergency department for a visit and that patient had many

comorbidities, they might have received three codes falling into the substance use (SUD) category, two for the mental health (MHD) category, one in the accidents and injuries (ACC/INJ) category, and two in the infection (INF) category. That *single visit* would add three to that patient's SUD_ed value, two to their MHD_ed value, one to their ACC/INJ_ed value, and two to their INF_ed value. In this way, the visit reason counts for each patient reflect both the types of conditions they received treatment for during the two years, and the number of times treatment was indicated, providing a holistic picture of their health challenges. Many patients who received primarily preventive care in the clinics had low to exceptionally low emergency department and outpatient visit reason counts. Instead, some of these patients had high numbers for the preventive visits (PREV_VISIT_op) variable. When looking at the percentages of diagnoses counted within a given emergency department visit reason category (e.g., SUD), the percentage reflects the proportion of all emergency department visit diagnoses that fall within that category. In this case, patients are not weighted equally since those with more diagnoses (sicker patients) will contribute many more diagnoses to the total number or percentage of diagnoses that fall into each category. While this does not treat patients equally, it focuses on the amount of total utilization that falls into each visit reason category, quantifies the overall impact of each category, and describes the clinic's patient population in terms of their most significant health needs and challenges.

Visit Reason Category	Visit Reason Category Abbr.	Total #/% of Diagnoses: Emergency Visits		Total #/% of Diagnoses: Outpatient Visits		Total #/% Diagnoses: All Visits		# of Patients Affected	
Cardiovascular Disease	CVD	318	(3.91%)	370	(3.06%)	688	(5.63%)	258	(11.39%)
Heart Failure	HF	47	(0.58%)	35	(0.29%)	82	(0.67%)	24	(1.06%)
Hypertension	HTN	488	(6.00%)	**1702**	(14.09%)	**2190**	(17.94%)	543	(23.97%)
Resp. Infection / Pneumonia	URI/PNA	**683**	(8.40%)	681	(5.64%)	1364	(11.17%)	**577**	(25.47%)
Asthma /	AS/COPD	385	(4.74%)	476	(3.94%)	861	(7.05%)	314	(13.86%)

COPD, Chronic Respiratory									
Neurological Diseases	NEURO	254	(3.12%)	262	(2.17%)	516	(4.23%)	243	(10.73%)
Substance Use Disorders	SUD	**1340**	(16.48%)	**2611**	(21.61%)	**3951**	(32.36%)	**1,025**	(45.25%)
Mental Health Disorders	MHD	655	(8.06%)	**1291**	(10.69%)	**1946**	(15.94%)	**549**	(24.24%)
Cancers and Neoplasms	CA	30	(0.37%)	98	(0.81%)	128	(1.05%)	70	(3.09%)
Pregnancy-related Conditions	PREG	69	(0.85%)	7	(0.06%)	76	(0.62%)	30	(1.32%)
Diabetes and Related Conditions	DM	257	(3.16%)	**1517**	(12.56%)	**1774**	(14.53%)	307	(13.55%)
Infections	INF	**577**	(7.10%)	711	(5.89%)	1288	(10.55%)	**602**	(26.58%)
Liver, Pancreatic, and Gallbladder Diseases	LIV	81	(1.00%)	97	(0.80%)	178	(1.46%)	79	(5.49%)
Renal Diseases	REN	39	(0.48%)	69	(0.57%)	108	(0.88%)	48	(2.12%)
Cognitive Deficits	COG	57	(0.70%)	81	(0.67%)	138	(1.13%)	80	(3.53%)
Sensory Deficits	SENS	233	(2.87%)	326	(2.70%)	559	(4.58%)	315	(13.91%)
Pain and Pain Syndromes	PAIN	**818**	(10.06%)	**1219**	(10.09%)	**2037**	(16.68%)	**783**	(34.57%)
Tobacco / Nicotine Use	TOB	**1226**	(15.08%)	166	(1.37%)	1392	(11.40%)	447	(19.74%)
Accidents & Injuries	ACC/INJ	572	(7.04%)	362	(3.00%)	934	(7.65%)	433	(19.12%)
Totals		**8,129**		**12,081**		**20,210**		**2265**	**(100%)**

Table 7:Visit Reason Counts & Percentages:

Emergency Visits, Outpatient Visits, All Visits; # of Patients

Bold numbers in each column represent top 5 categories for that measure.

The number of diagnoses in each category varies between emergency and outpatient/clinic visit types. This reflects the differences in conditions that are coded for, and conditions whose treatment is emphasized in a setting. For example, accidents and injuries are often treated in the emergency department, and diabetes and hypertension in the primary care setting. The top five categories overall, in terms of the total numbers of diagnoses, are: Substance Use Disorders (SUD) with 3,951 codes (32.4% of all codes), Hypertension (HTN) with 2,190 codes (17.9%),

Pain and Pain Syndromes (PAIN) with 2,037 codes (16.7%), Mental Health Disorders with 1,946 codes (15.9%) and Diabetes (DM) with 1,774 codes (14.5%). Also of note, the top five condition groups among emergency department visits that were converted to inpatient admissions were Substance Use Disorder (SUD), Upper Respiratory Infections and Pneumonia (URI/PNA), Infections (INF), Cardiovascular Disease (CVD), and Pain and Pain Syndromes (PAIN).

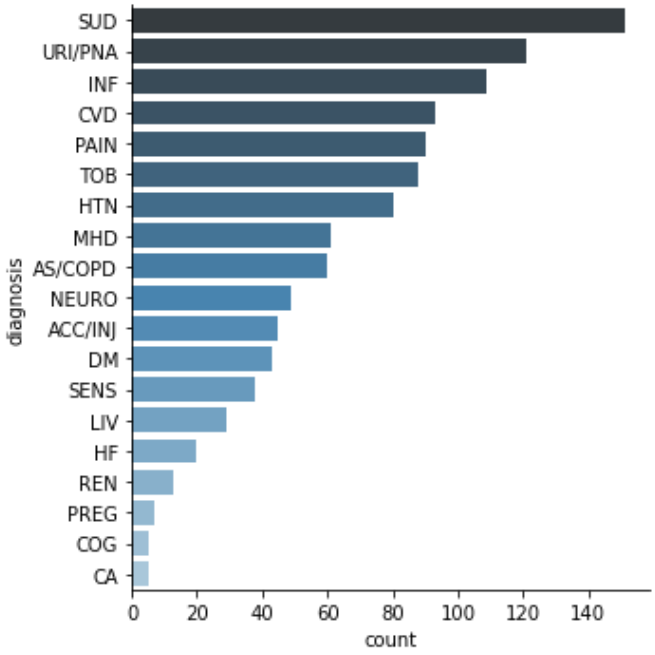


Figure 32: Numbers of Diagnosis Codes by Category for Emergency Visits that Became Inpatient Admissions (visits=349; patients=181)

While these categories represent the areas where the most service is being provided, the numbers of patients with diagnoses in each group provide insight into the top condition categories impacting the patient population. The top five condition categories in terms of the number of patients affected are: Substance Use Disorders (SUD) impacting 1,025 people (45.25% of the sample), Pain and Pain Syndromes (PAIN) impacting 783 (34.57%), Infections (INF) impacting

602 (26.58%), Upper Respiratory Infections and Pneumonia (URI/PNA) impacting 577 (25.47%), and Mental Health Disorders (MHD) impacting 549 (24.24%).

Examining these top five condition groups more closely, some relationships are worth noting. The top category, Substance Use Disorder (SUD), is a category with a lot of differentiating power. Among the 1,025 patients (45.25%) impacted by this diagnostic category, people with a corrections history, or mental health conditions (either an MHD diagnosis or positive PHQ-2 screening or both) had a median of two substance use diagnoses during their visits. In contrast, those without had a median of one (Figure 33).

While people with mental health conditions and positive PHQ-2 scores were more likely to have positive HARK scores, there was no difference in the median number of substance use diagnoses between those with positive HARK scores and those without (Figure 34). Fewer patients with a primary language other than English had any substance use diagnoses; however, those who did have this diagnosis had a higher median number of total diagnoses than English-speaking users. There was an insignificant difference between males and females within groups.

The relationships between NIDA scores or nicotine use, and the median number of substance use diagnoses appear counterintuitive (Figure 35). Never and former nicotine users had a median of two substance use diagnoses, while current users had a median of one. Likewise, patients who scored three or more on the NIDA TAPS screening ("highest-risk") had a median of one substance use diagnosis, while those who scored less than three had a median of two. Women were more likely than men to have a substance use diagnosis, but a low NIDA TAPS score. Both findings may show alignment between people's willingness to admit to having a problem and receiving treatment for it. Some patients in the sample who were older and had given up

smoking had many outpatient visits for mental health and substance use. In this case, their larger total numbers of diagnosis codes may reflect the treatment they are seeking and receiving.

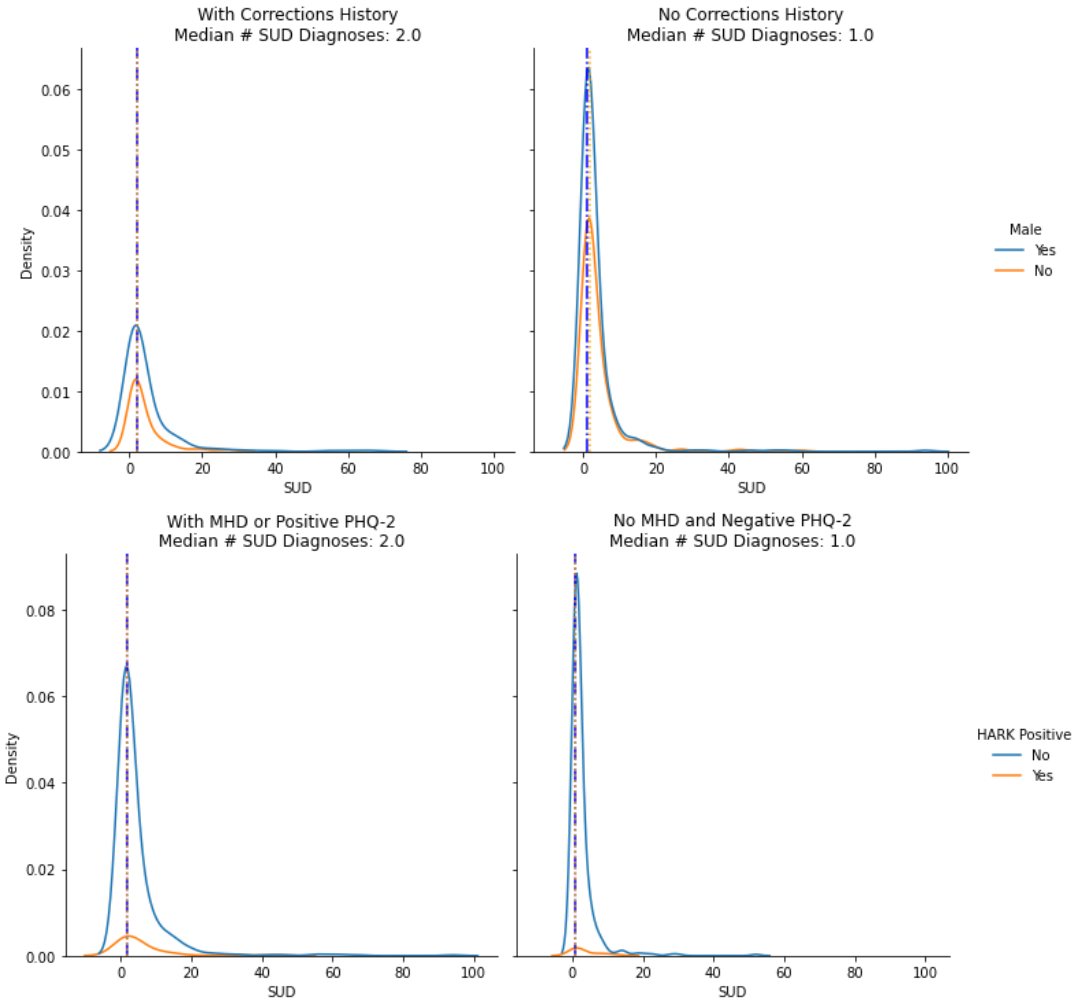


Figure 33: Median # Substance Use Diagnoses (n=1,025), Corrections History and Mental Health Diagnosis / Positive PHQ-2 Score

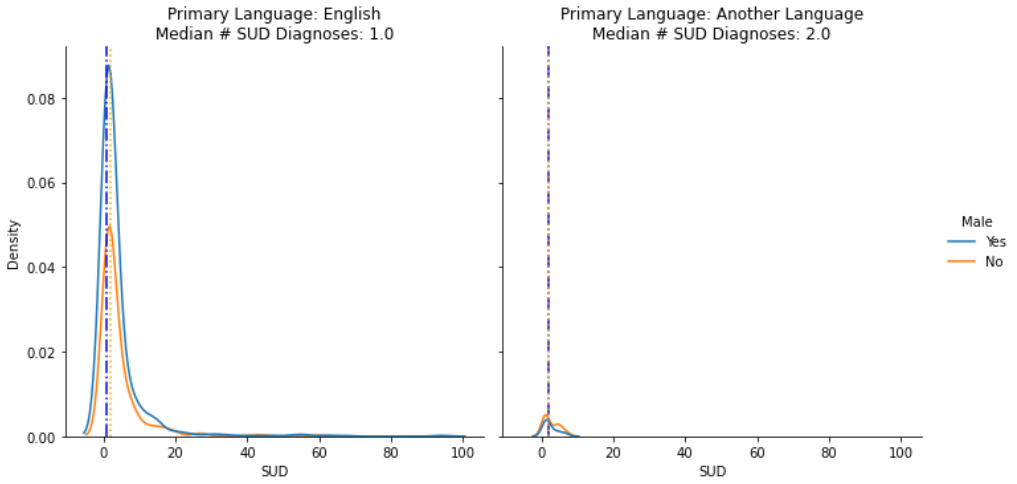


Figure 34: Median # Substance Use Diagnoses (n=1,025), Primary Language

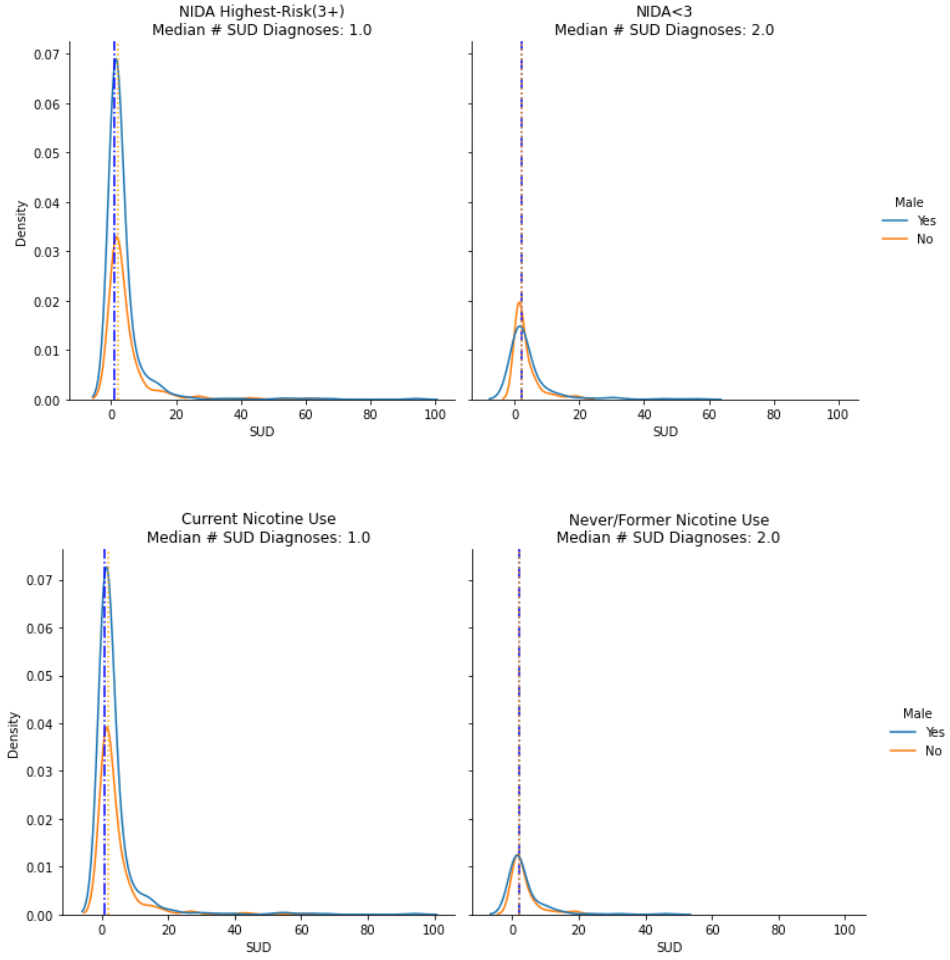


Figure 35: Median # Substance Use Diagnoses (n=1,025), Nicotine Use & NIDA TAPS Scores

PAIN, the diagnosis group impacting the second-largest number of people in the patient sample (783, 34.57%) had a median number of diagnoses of two, regardless of strata. Having or not having a mental health condition or positive PHQ-2 score, or a substance use diagnosis, did not impact the overall median number of PAIN diagnosis codes. However, there were diagnosis categories that were correlated with more pain diagnoses and higher pain counts, including acute upper respiratory conditions (URI/PNA), hypertension (HTN), diabetes (DM), other infections (INF), and some sensory conditions (SENS). The density plots in Figure 36 show the relationship between more pain diagnoses, and more of each of these types of diagnoses, broken out between patients with average pain count from clinical notes of less than 5 vs. 5 or more.

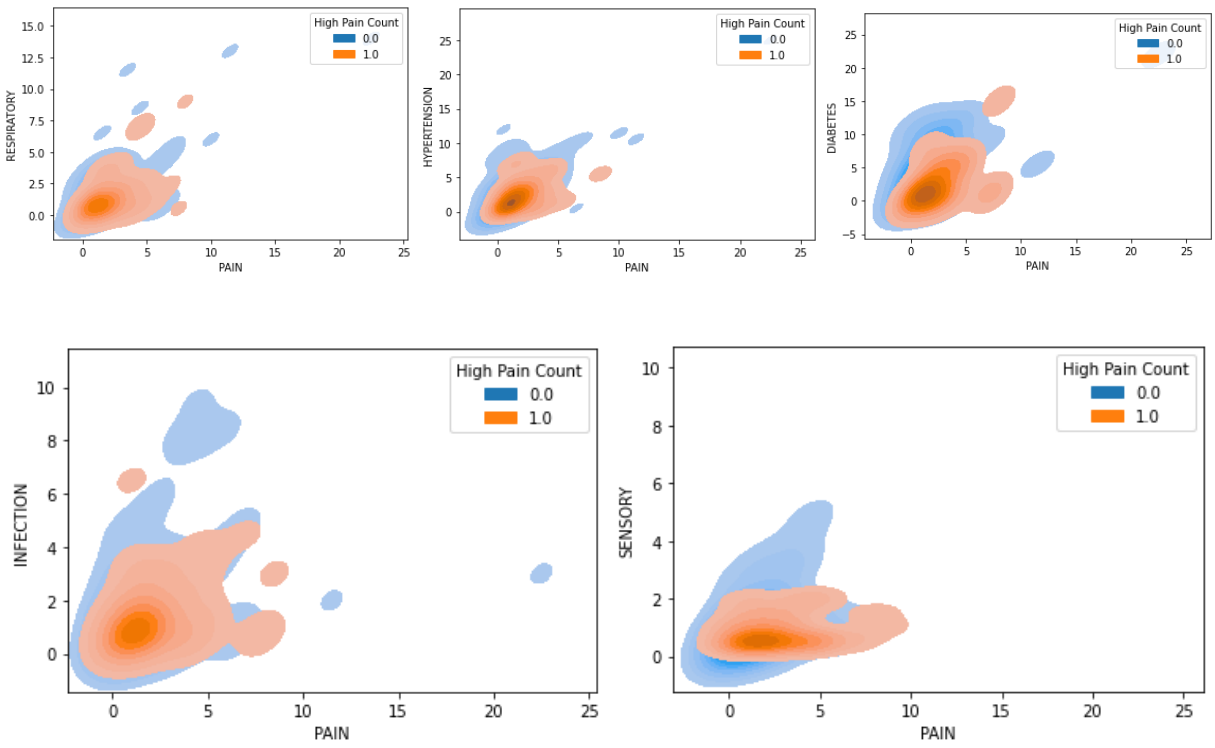


Figure 36: Relationship between total average Pain diagnoses and total average Acute Respiratory, Hypertension, Diabetes, Infection, and Sensory diagnoses by pain count 5 or more (High=1), vs. less than 5 (High=0)

Diagnoses for Infections (INF) impacted 602 (26.58%) patients during the two years. There were a larger median number of diagnoses for infections among male substance users, and among people without mental health diagnoses or positive PHQ-2 screenings and negative HARK scores. These groups' median infection diagnosis count was two, while all other strata had a median of one. It is possible that male substance users' increased likelihood of infection can be attributed to males' more frequent use of IV drugs (Powis et al., 1996).

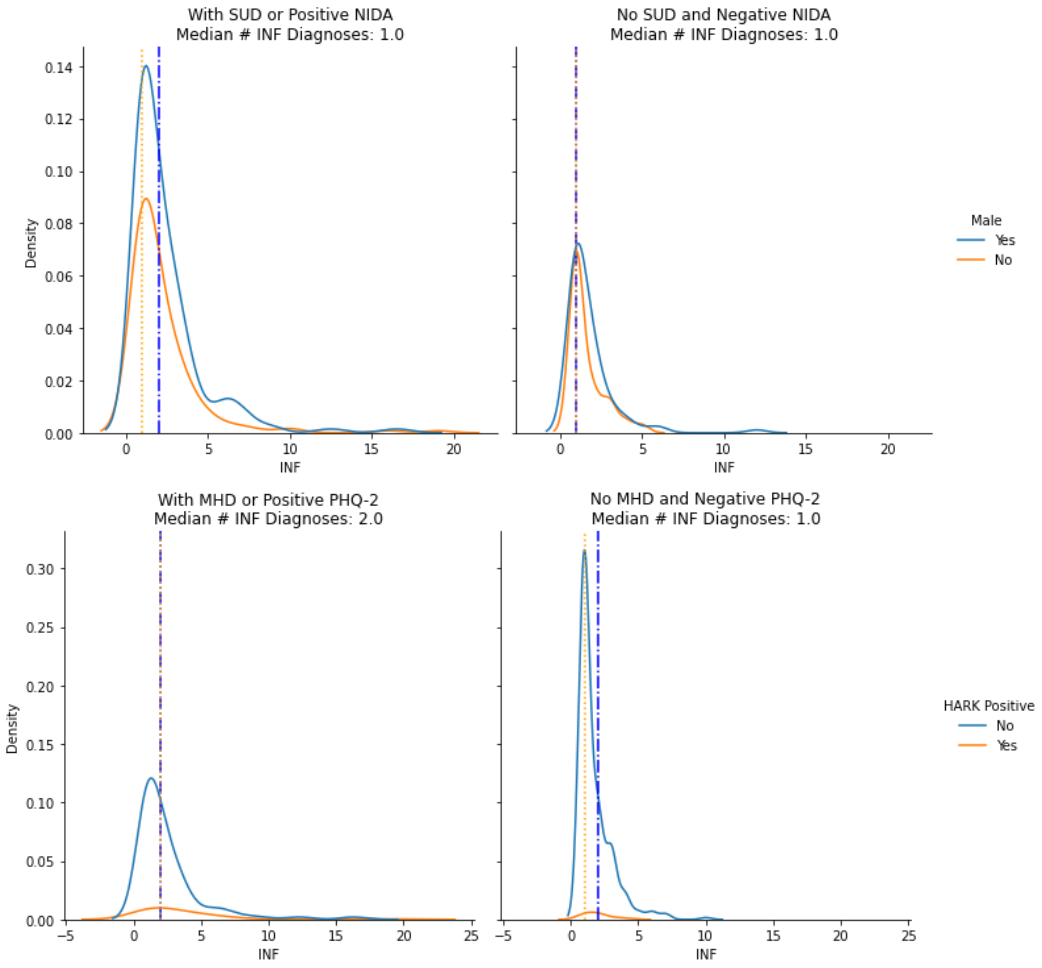


Figure 37: Median # of Infection Diagnoses (n=602), Substance Use / Positive NIDA and Mental Health Diagnosis / Positive PHQ-2 Score by HARK Score (Positive: ≥1)

Upper Respiratory Infections and Pneumonia (URI/PNA) impacted 577 (25.47%) patients and were the second-most-common diagnoses associated with inpatient admissions. While the most common median number of diagnoses in this group was two across all strata, the distribution of diagnoses for nicotine/tobacco users (and for those with one or more substance use diagnoses) had a wider tail, indicating at least some smokers and other substance users had an increase in acute respiratory diagnoses. Those with mental health disorders or positive PHQ-2 scores had a median of two URI/PNA diagnoses for the two years. However, those without had a median of one.

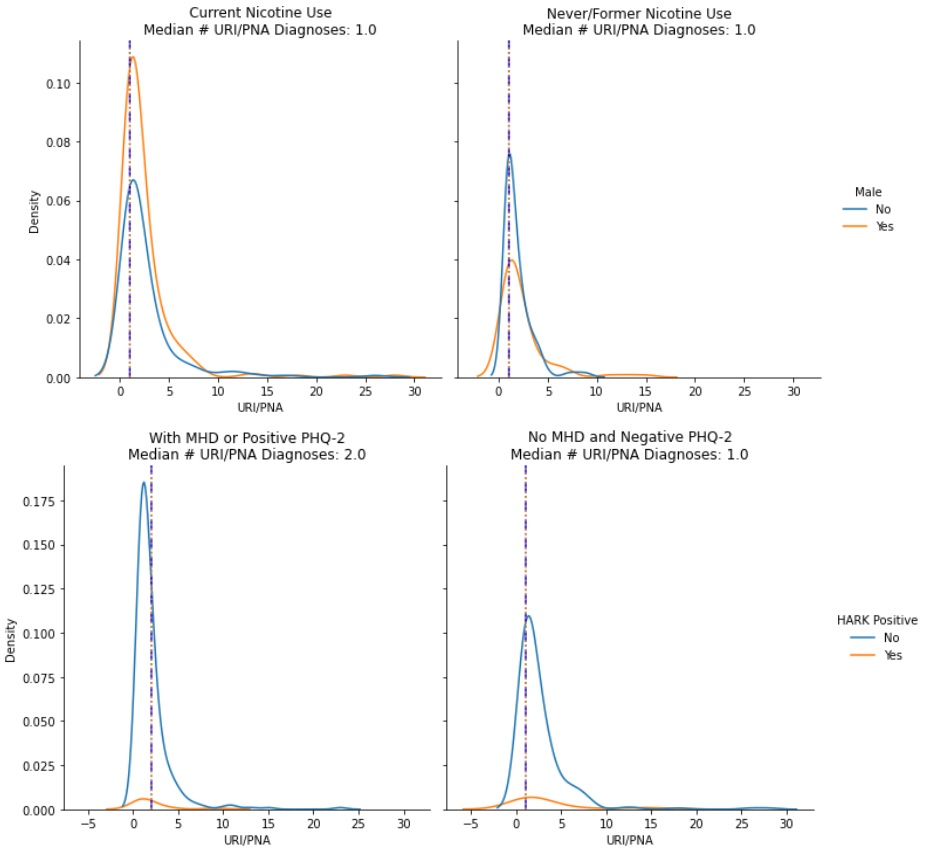


Figure 38: Median # of Upper Respiratory / Pneumonia Diagnoses (n=577), Nicotine Use and Mental Health Diagnosis / Positive PHQ-2 Score by HARK Score (Positive: ≥1)

Mental Health Disorders (MHD) impacted 549 (24.24%) patients during the sample period. This is approximately the same percentage of people (25%) estimated to have mental health conditions among the national homeless population (SAMHSA, 2011). We've already seen how mental health was a modifier for the median number of several other high-impact conditions. The median number of mental health conditions, not surprisingly, was influenced by a diagnosis of substance use and/or a "highest-risk" NIDA score (3+). The median for most groups was two mental health diagnoses across all visit types.

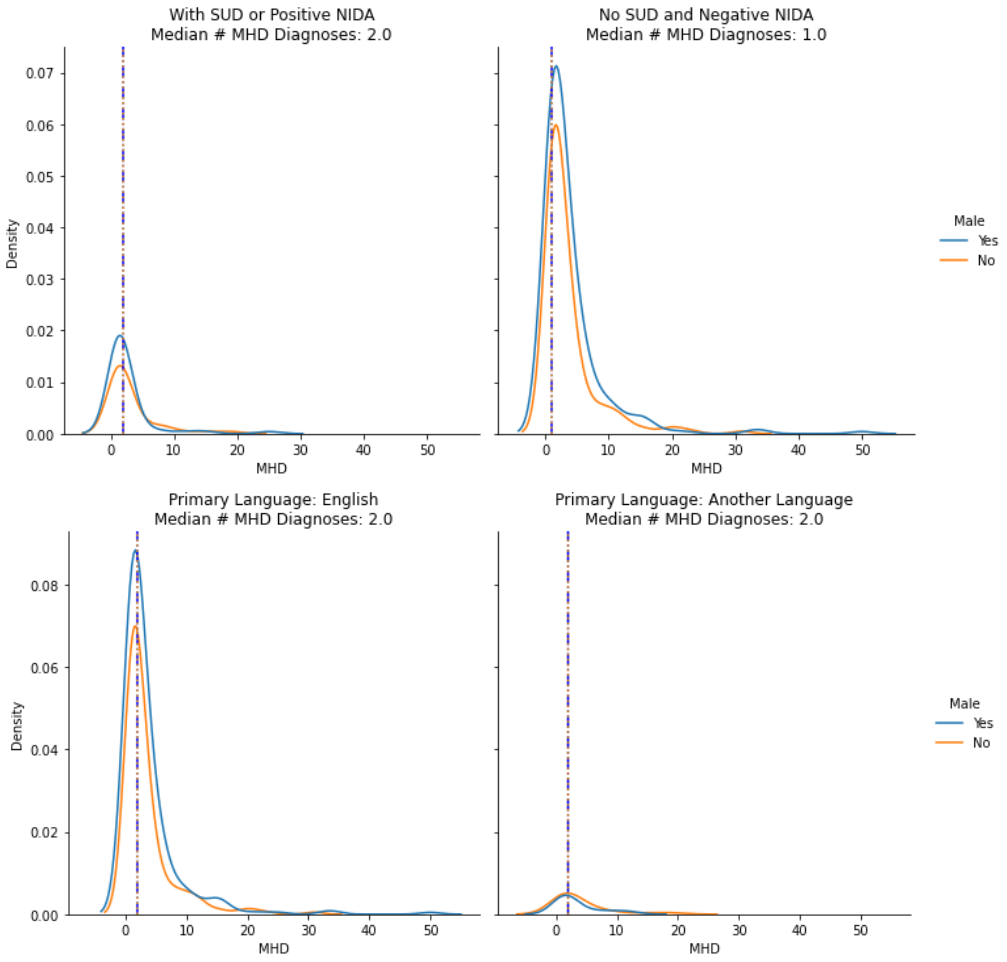


Figure 39: Median # of Mental Health Diagnoses (n=549), Substance Use Diagnosis / Positive NIDA TAPS (3+), Primary Language (wide tails)

It is easy to see a pattern in many of these relationships – for example, substance use diagnoses and mental health diagnoses seem to co-occur and co-influence each other (National Institute on Drug Abuse (NIDA), 2021). Examining correlations is a way to better understand the co-occurrence of visit reasons among patients; however, these relationships should never be interpreted causally. The fact that diagnoses for diabetes and pain (for example) occur in many of the same patients does not tell us why, which came first, or whether or not the relationship represents a trend among homeless persons or human beings in general. Correlations can, however, help characterize variation in the visit reason counts. To see correlation plots for emergency and outpatient visit reason groups, see Appendix B, Section 1.

4.4 Visits and Intervals

4.4.1 Visit Counts

As with the distributions of diagnosis codes among visits, the distributions of visit counts cluster around zero (emergency visits) or one (outpatient/clinic visits) and incrementally progress towards a long and heavy tail, where a few patients have remarkably high counts. As mentioned previously, it is not the same patients who have many emergency visits and many outpatient/clinic visits. To protect the privacy of patient outliers and decrease bias in interpretation, visit counts have been grouped for the descriptive analysis in the following way:

Group Number	Group Name	#/% of Patients, Outpatient Groups	#/% of Patients, Emergency Groups
0	Zero Visits	n/a*	1,545 (68.2%)
1	One Visit	790 (35.0%)	210 (9.3%)
2	Two Visits	446 (19.7%)	149 (6.6%)
3	Three or Four Visits	388 (17.1%)	135 (6.0%)
4	Five to Seven Visits	343 (15.1%)	124 (5.5%)
5	Eight to Thirty Visits	259 (11.4%)	92 (4.1%)

6	More than Thirty Visits	39 (1.7%)	10 (0.4%)

Table 8: Outpatient and Emergency Department Visit Groups

**All patients had at least one outpatient/clinic visit during the two years.*

While these visit groups' ranges may seem arbitrary, they were chosen to allow the data to cluster together into groups of visit counts that are meaningful and properly distributed within the patient data *as it is*. This allows retention of an accurate picture of patients' visit behavior and removes the need to delete important outliers from the analysis.

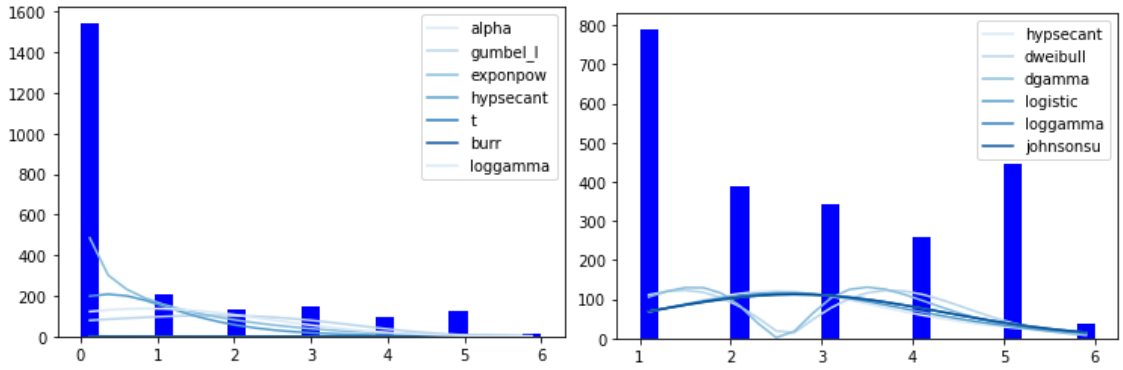


Figure 40: Distribution Fits: ED Visit Groups (left) and Outpatient/Clinic Visit Groups

Visit reason counts vary widely between emergency department and outpatient/clinic visits, as is illustrated in Table 7. Reason counts also vary between patients belonging to each visit group, with patients who are seeking more clinic or more emergency department visits having distinct characteristics and diagnoses from patients seeking fewer.

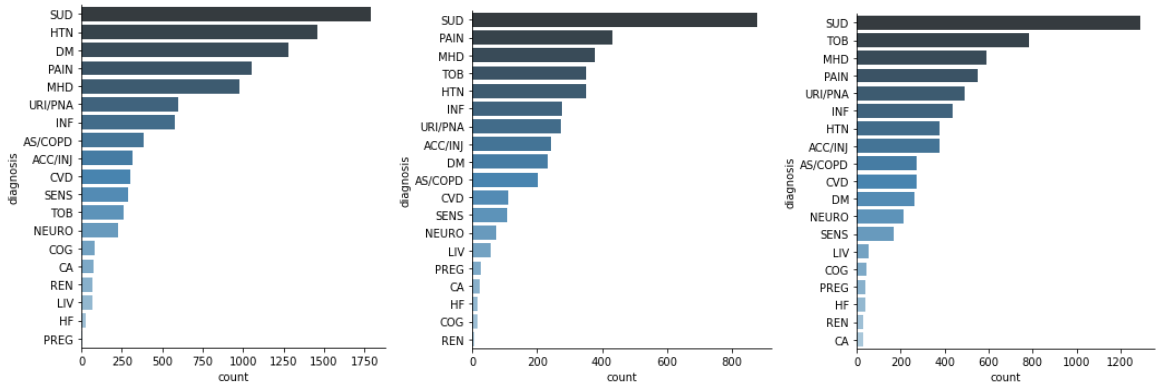


Figure 41: Overall Diagnosis Counts for Patients in ED Visit Groups 0 to 2 (left, n=1,890), 3 and 4 (center, n=241), or 5 and 6 (right, n=134)

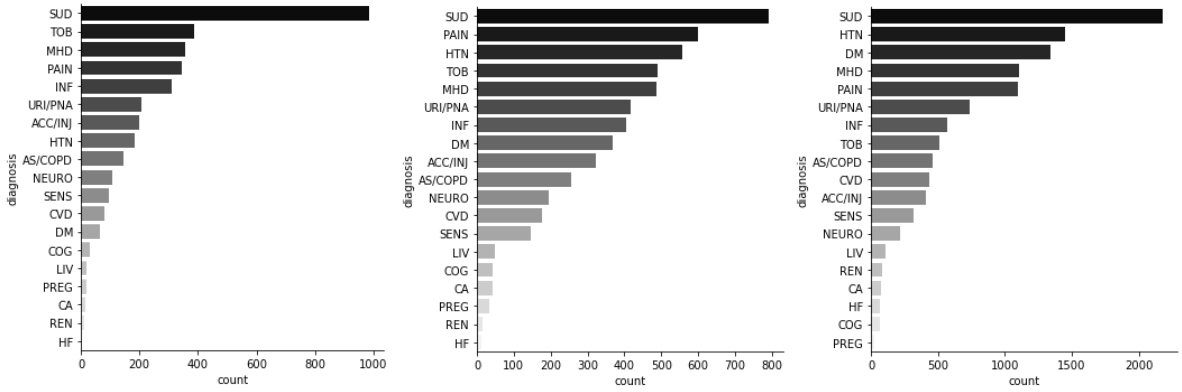


Figure 42: Overall Diagnosis Counts for Patients in OP Visit Groups 1 and 2 (left, n=1,178), 3 and 4 (center, n=602), or 5 and 6 (right, n=485)

While the diagnosis of Substance Use Disorder (SUD) continued to be a dominant one, it is evident that patients with fewer ED visits and more clinic visits had more diagnoses typically addressed in the primary care setting, including Hypertension (HTN), Diabetes (DM), and Mental Health Disorders (MHD). Additionally, the shorter the bars in these graphs, the lower the patients' overall utilization related to that diagnosis. Therefore, patients in outpatient visit groups 1 and 2 predominantly sought preventive care (for example, immunizations and physicals) at the clinic, otherwise having few visits unless there was an emergency. This could account for the considerable number of Substance Use Disorder (SUD) diagnoses among these

patients, and the smaller number of diagnoses in every other category. The visit reason graphs for the highest emergency utilizers (ED groups 5 and 6) and the lowest clinic utilizers (OP groups 1 and 2) look similar, except for the addition or increase for ED groups 5 and 6 of some diagnostic categories indicating worsening health and increasing chronic illness, such as Cardiovascular Disease (CVD), Cognitive (COG) and Sensory (SENS) deficits, Liver (LIV) and Renal (REN) diseases, and Heart Failure (HF).

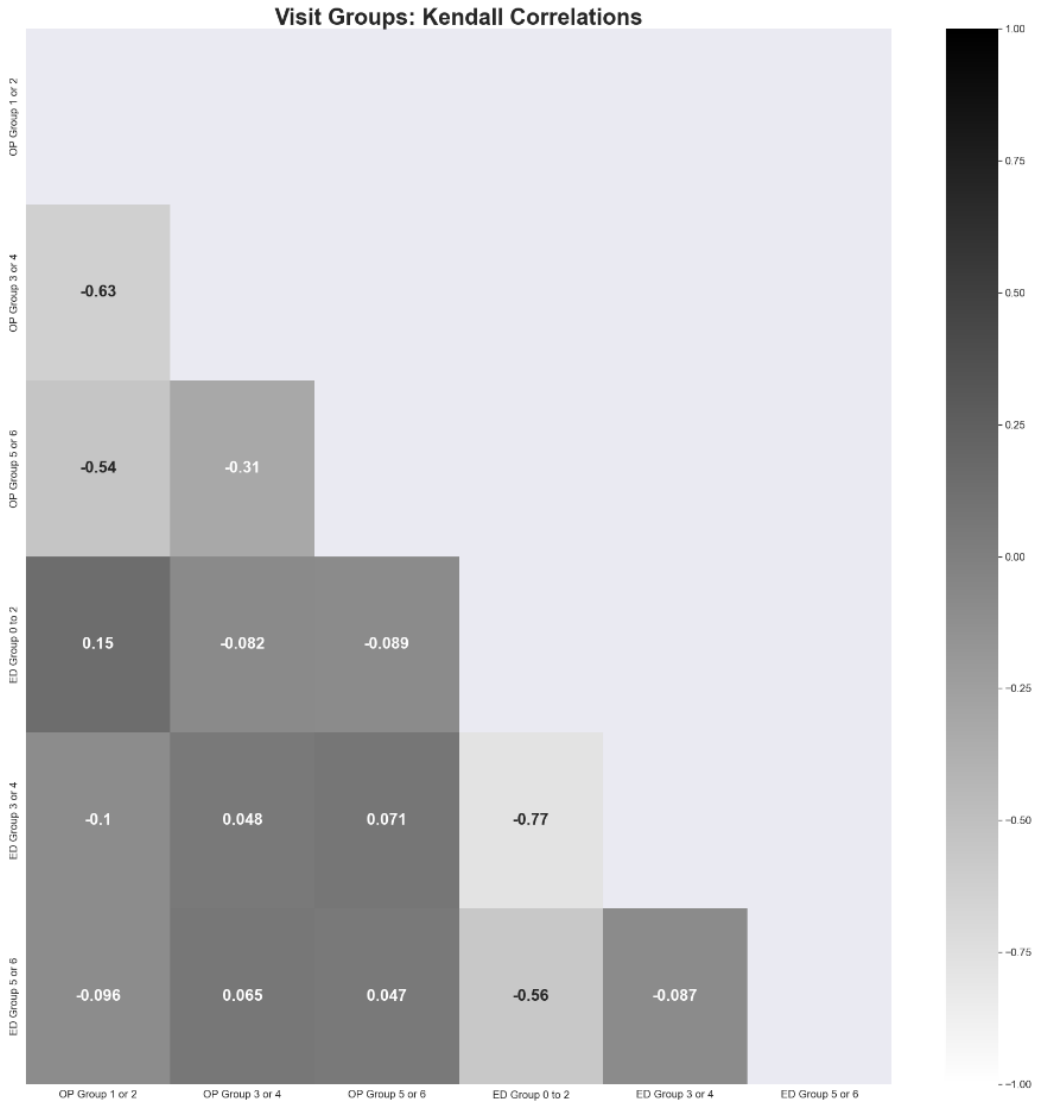


Figure 43: Kendall Correlations Between Outpatient and Emergency Visit Groups

A correlation map showing the relationships between patient group memberships reveals that it is less likely that the same patient would be in both OP group one or two and ED group five or six. Not only is there a negative correlation between being in both OP group one or two and in ED group five or six (-0.096), but there is also a stronger positive correlation between OP group one or two membership and ED group zero to two membership (0.15). This demonstrates that many HCHM patients are low utilizers of all services. Only forty-four patients (1.9% of the total sample) were members of both OP group one or two and ED group five or six. Therefore, the most likely interpretation of the similarities between the visit reason categories (Figures 41 and 42) of those in the lowest outpatient utilization groups and those in the highest emergency department utilization groups is that patients who start out as low utilizers of all services eventually develop chronic conditions and worsening health due to lack of health maintenance. Those in the highest emergency department utilization groups have similarities with patients in the lowest outpatient utilization groups because they used to be those patients when they were younger. Indeed, the median age for the lowest outpatient utilizers is thirty-six years, and that of the highest emergency department utilizers is forty-three years.

Additional interesting correlations between visit group memberships include the negative correlation between being a member of OP groups one or two and being a member of ED groups three or four (-0.1). So those in OP groups one or two are rarely in ED groups three, four, five *or* six. For those in OP groups three, four, five or six (the highest clinic utilizers), there is a slight negative correlation with being in ED groups zero to two, but a slight positive correlation with being in ED groups three, four, five, or six. It may be that a few additional outpatient visits reduce emergency department utilization, but only for those who are less sick. As patients reach

the point where they need many outpatient visits, they also may be sick enough to require more trips to the emergency department.

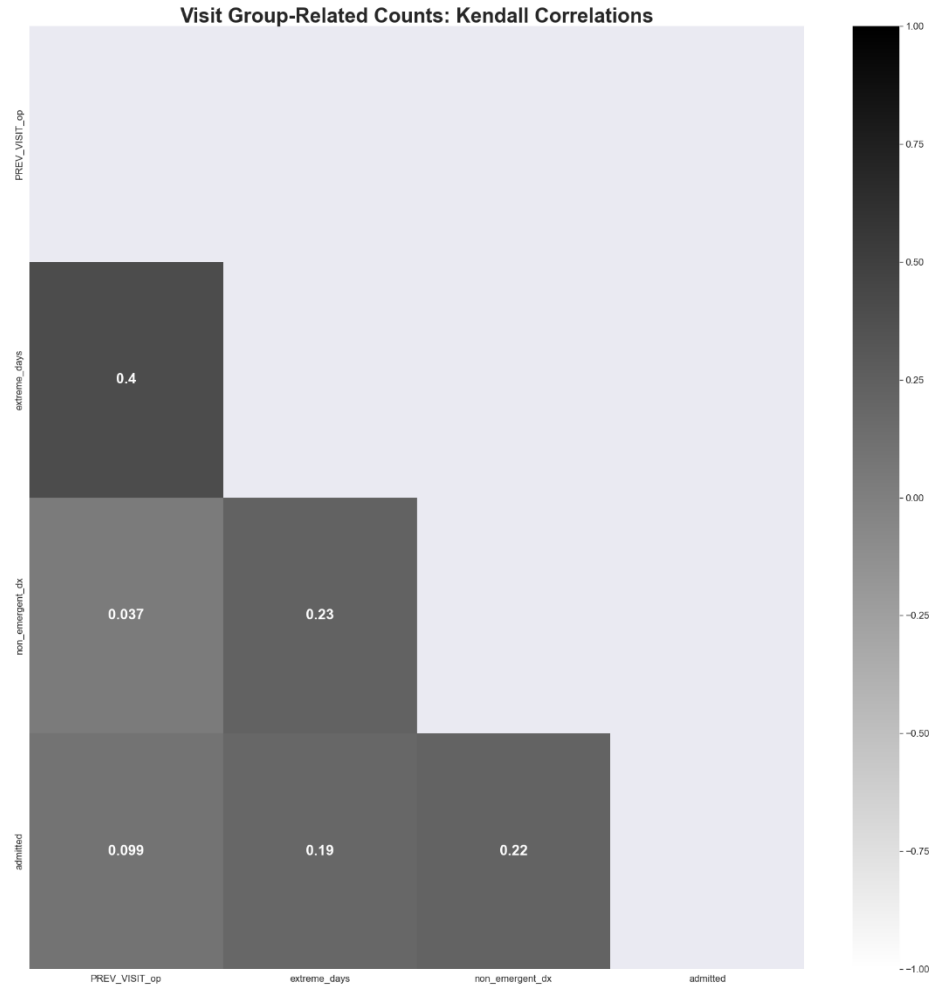


Figure 44: Kendall Correlations Between Visit-Related Features

4.4.1 Visits on Extreme Weather Days

Many patients had one or two visits on an extreme weather day (see Chapter 3, Section 3.6 for more information about the definition of an extreme weather day). However, a few patients had many such visits. It is not readily apparent why a few people would have many visits on inclement days – they may have had frequent visits to begin with, leading to an increased

probability that at least some of them will fall on a day with severe weather, or they may have taken ill because of the weather, or otherwise sought out the clinic or emergency department during a storm. Assuming that one could need to go to an emergency visit on any of the 730 days in the sample period, for example, and defining 114/730 as extreme weather days, the probability of three emergency visits in a row being extreme weather visits would be 114/730 * 113/729 * 112/728 or .37%, an improbable occurrence on its face. However, this doesn't factor in situations that would increase the probability of more visits, such as a period of greater illness, a large numbers of comorbidities, or a recent new diagnosis. There were 100 patients (13.9% of those with any ED visits (n=720)) who had more than two ED visits where extreme day visits made up at least half of all of their ED visits, and there were 105 patients (4.6% of those with any OP/clinic visits (n=2265)) who had more than two OP/clinic visits where extreme day visits made up at least half of all of their clinic visits.

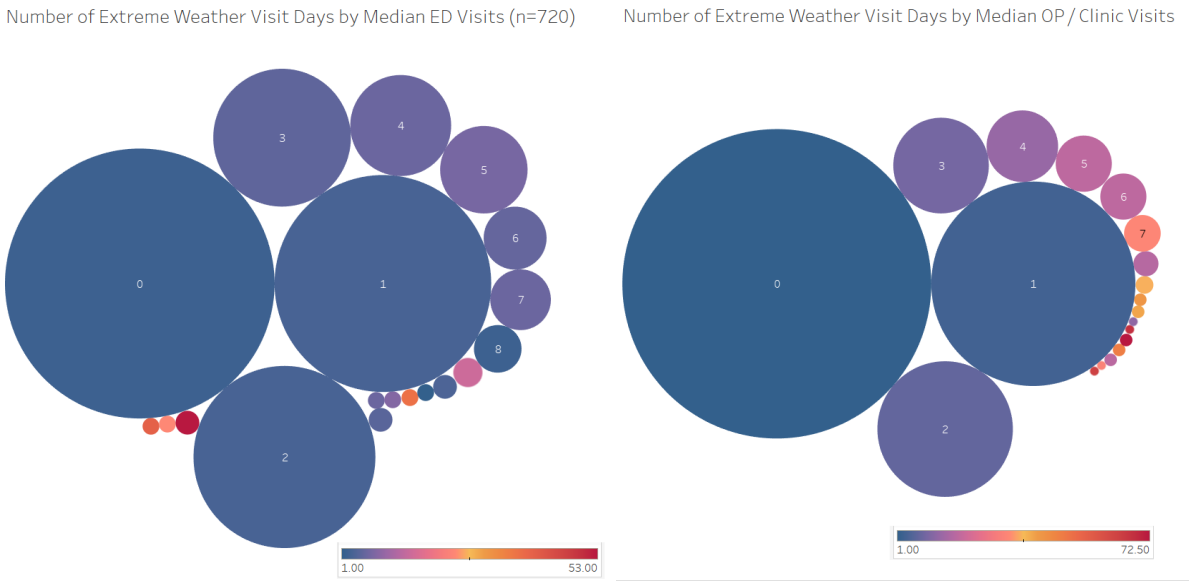


Figure 45: Numbers of Extreme Weather Visits by Patients' Median #s of ED (left) and Outpatient (right) Visits

No conclusions can be drawn based on this information, however. There are many unmeasured factors influencing conditional transitions from one visit to another that are not considered. In addition, the definition of an 'extreme weather day' used to identify the days when patients might seek shelter could be erroneous. For example, the exclusion of important measures, such as heat index (Wellenius et al., 2017), could reduce the ability to detect a relationship between seeking shelter in the clinic or emergency department and weather events. An accurate analysis of pertinent transitions would require both more detailed weather information and analysis using a longitudinal data set.

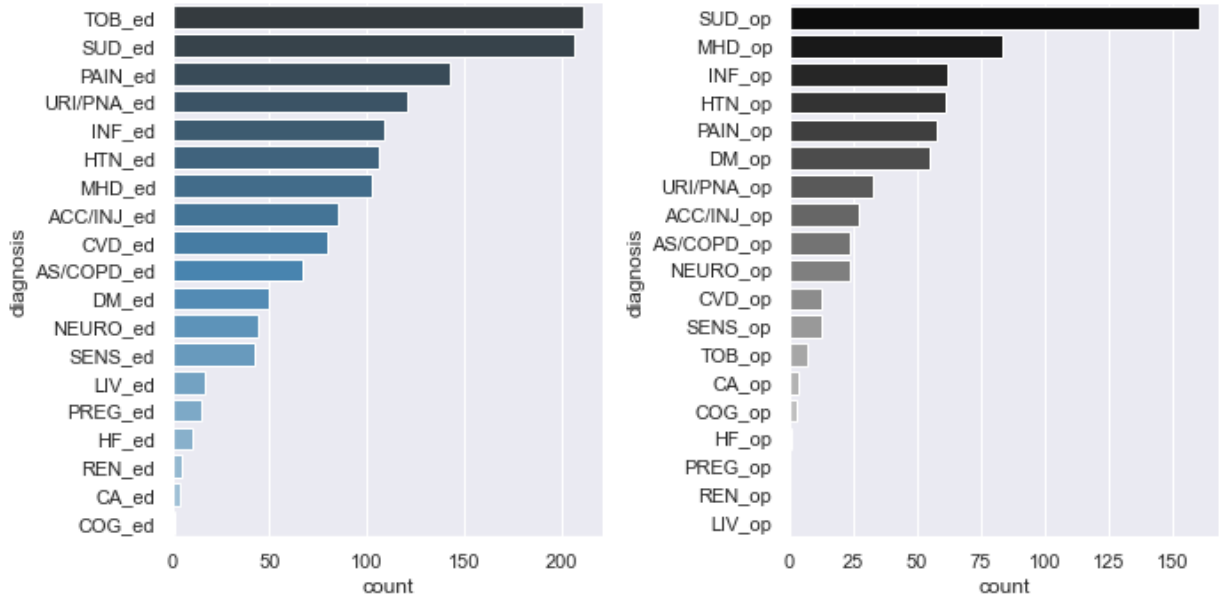


Figure 46: Top ED & OP Visit Reason Categories

Among patients (n=205) with at least one extreme weather appointment day for every two appointments; note the lack of difference between these diagnosis groups and top diagnosis groups overall.

4.4.2 Visit Intervals

Another characteristic of high utilizers is the frequency of their visits. People with chronic conditions that are poorly controlled due to lack of health maintenance are more likely to need

frequent emergency interventions. People with chronic conditions that can be difficult to manage, such as Mental Health Disorders (MHD), Substance Use (SUD), Hypertension (HTN) and Diabetes (DM), may need frequent outpatient visits to keep these concerns under the best possible control. Average visit intervals were calculated wherever patients had more than one emergency or outpatient visit, as discussed in Chapter 3, Section 3.2. Where average intervals could not be calculated, values were simply set to 730 days, indicating interval censorship. The same value was used for all patients whose intervals were censored, regardless of the true number of days between their first/only visit and the end of the sample period. Average ED intervals that were not censored were calculated for 510 patients (70.8% of those with any ED visits (n=720)), and average OP/clinic intervals were calculated for 1,524 patients (67.3% of the sample). The median calculated average ED interval was 68.5 days, and the median calculated average OP/clinic interval was 32.5 days. The most common overall diagnosis categories for patients whose average intervals were below the median were chronic conditions that can be heavily relapsing and hard to manage, such as substance use disorders (SUD), mental health disorders (MHD), and frequently associated conditions such as pain (PAIN), upper respiratory infections (URI/PNA), other infections (INF), and accidents and injuries (ACC/INJ).

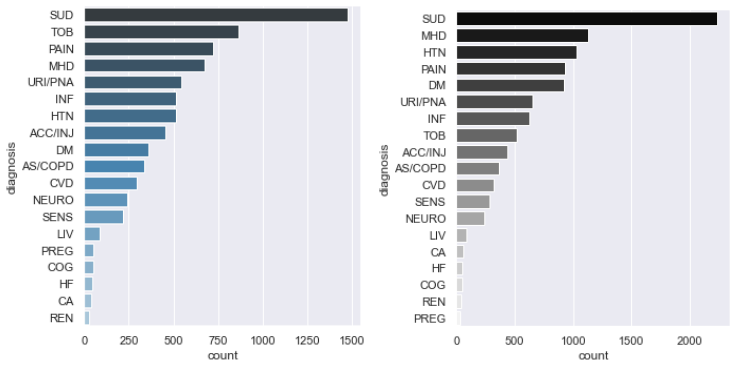


Figure 47: Top ED (left) & OP (right) Visit Reason Categories

Among patients with average visit intervals less than the median (ns=255 (ed), 762 (op))

CHAPTER 5: MODELING

5.1 Objectives and Initial Approaches

The completed feature set consists of many variables that could potentially do a good job describing the patient population based on a two-year visit sample, including demographic information, meaningful visit reason categories, Elixhauser scores, and measures derived from clinical notes that contained not only health measures (height, weight and blood pressure readings), but some indicators of patient behaviors (substance use, depression and intimate partner violence screenings, nicotine use and visit intervals). Descriptive analysis alone cannot divide a complex data set with features having a variety of distributions in a holistic way; I needed a modeling technique that could reduce complexity while considering many features. The goal was to utilize the full feature set, remove no outliers, and describe the characteristics of clinic patients by grouping them into meaningful sub-populations by their utilization patterns. The idea was to both determine which patients were high utilizers and describe all patients, revealing how they might be better served.

Initially, the HCHM clinic sponsors had requested a model to look at predictors related to the outcome of three or more emergency department visits per year. An early study proposal centered around using a traditional regression approach to model the effects of many features descriptive of the patient sample against this logistic outcome. In addition to this logistic regression, many commonly used, traditional statistical modeling techniques (GLM) were tried – including binomial, gamma, exponential and Poisson models with dispersion – in an attempt to predict the outcomes of a) the numbers of emergency department visits, and b) the numbers of outpatient/clinic visits for the patient sample. Both zero-inflated models (for the outcome of

emergency department visit counts) and models without zero-inflation (for the outcome of outpatient/clinic visit counts) were attempted. Regression models were produced both with and without interaction terms. Then, additional models were produced by grouping visit categories together based on their correlations and performing regression with and without interactions among the groupings. None of the GLM models was a particularly good fit for high-dimensional, non-linear data with important outliers that I did not want to remove. Individual model log-likelihoods were low, and combination models had low variation in AICc scores. Residual plots had high heteroskedasticity, and the models did not make accurate predictions of the numbers of visits. There was also questionable accuracy in the direction and strength of effect sizes in all models, particularly the logistic model. Upon sensitivity testing, there was trivial difference in effect sizes or significance levels of covariates in a logistic model against the outcome of three or more emergency visits per year vs. two or more or four or more visits per year. Moreover, these approaches did not help characterize the high utilization population.

A causal modeling approach using non-parametric g-methods to attempt to quantify the relationship between outpatient/clinic visits and the number of emergency department visits through weight adjustment was heavily considered. However, it became clear that there were too many unmeasured confounders at play among these complex, real-world interactions. Therefore the variance in these variables is almost certainly explained by far more than the labels placed upon the features and the lines drawn between them. Causal modeling of the relationships at play in high-complexity, systemic interactions – like those that lead to homelessness and its sequelae – would be a fascinating and revealing undertaking but requires a high level of expertise in the systems dynamics modeling domain (Fowler et al., 2019). Since identification of patient subgroups was the desired outcome of this analysis, I settled on the unsupervised

machine learning technique of clustering as a viable way to divide the aggregated patient-level data.

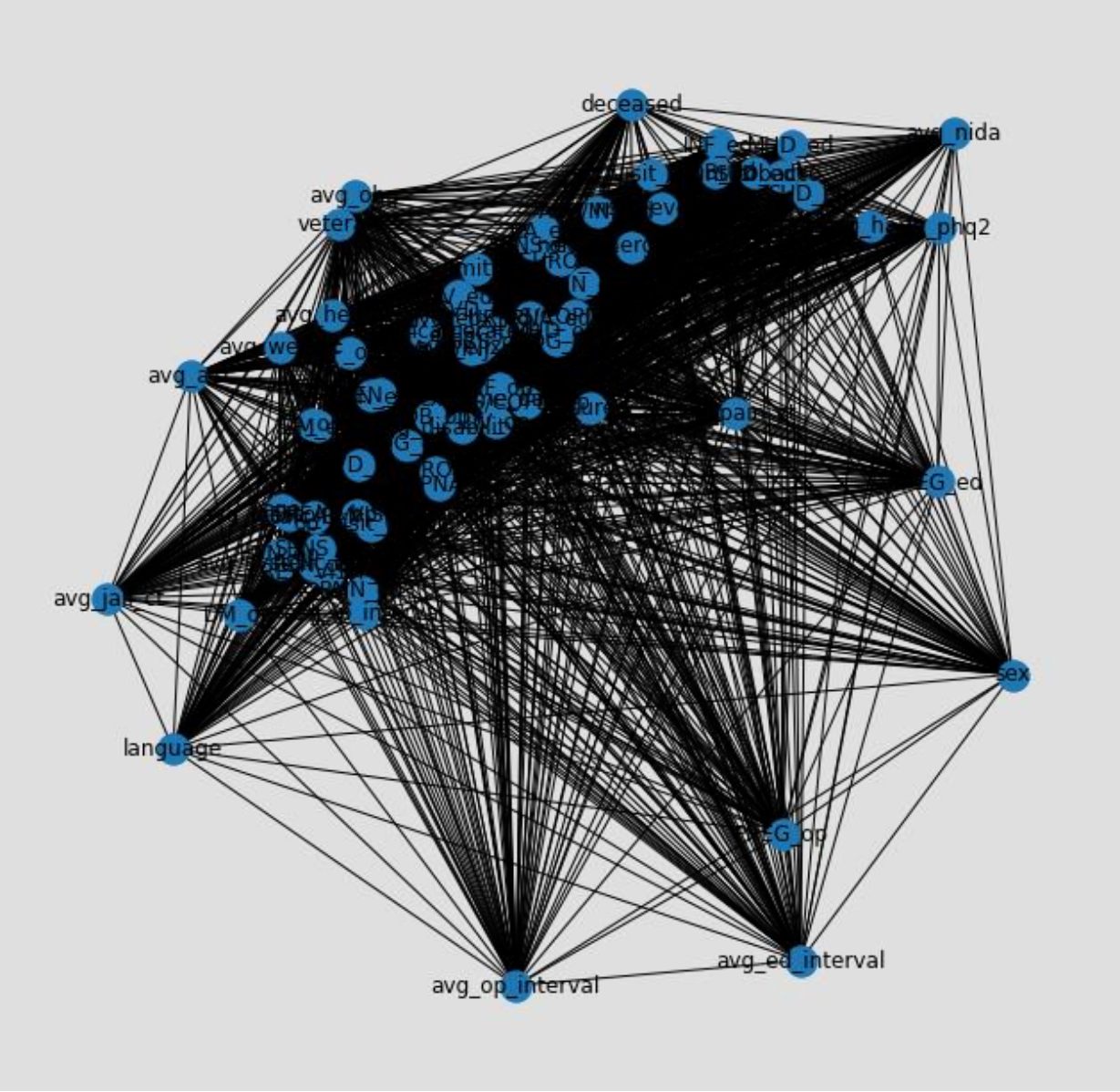


Figure 48: NetworkX Visualization of the Dataset's Interaction Graph
Produced from a Kendall Correlation Matrix of the Features (Hagberg et al., 2008)

It was necessary to try many different clustering and dimension reduction techniques on the data before settling on the successful approach I will describe in the remainder of this chapter.

Strategies that work well on data with many normally-distributed features and many linearly related data points, such as principal components analysis (PCA) or k-Means clustering using Euclidean distance measures, did not produce results of interest with this feature set. In many ways, the key to a successful outcome was discovery of a) a suitable dimension reduction algorithm, b) use of the right projection metric to allow the algorithm to create the dimension-reduced data set, and c) the use of 'double reduction' by application of both the right, carefully tuned, dimension reduction algorithm and an additional dimension-reducing clustering algorithm. This combination of approaches led to the discovery of a manageable number of clusters that were useful in describing the characteristics of the clinic's patient sub-populations, allowing for a deeper analysis of each one using both descriptive statistics and classification algorithms. The iterative process used to select the algorithms and choose from among several cluster sets was highly collaborative. The Health Care for the Homeless of Manchester clinic team carefully reviewed many sets of clusters to help determine which approaches produced those most descriptive of their clinic patients.

5.2 Dimension Reduction, Imputation & Dissimilarity Calculation

Manifold learning assumes that datasets can be represented as lying on smooth, non-linear manifolds of low dimension by finding a distance mapping function that will preserve the properties of the higher dimensional data in a lower dimension (Ihler, 2003). Distance-preserving methods may maintain spatial or graph distances and assume linear relationships or not. In general, dimension reduction algorithms using eigenvalue decomposition tend to be more effective on normal and linearly related features, while methods that attempt to keep the order or rank of dissimilarity metrics intact do better on non-linear data. The choice of algorithm depends on an understanding of the dataset's features and their meanings. In Figure 54, various

techniques are applied to reduce the three-dimensional, spherical dataset shown on the left. The "correct" algorithm depends entirely on what original proximities from the high-dimensional data need to be preserved in the low-dimensional representation. For example, in the original three-dimensional spherical dataset, the red points are near the purple points, separated by a gap. Whether it is better to "unroll" the sphere and place the red and purple points far from one another in the low-dimensional representation while preserving the other data points' relationships more accurately (as with the majority of the below methods), or better to preserve the distance between the red and purple points while obscuring the relationship between some of the other points by "squishing" the sphere on its side (as with Modified LLE) depends entirely on the meaning and importance of the proximity between the features in the original dataset.

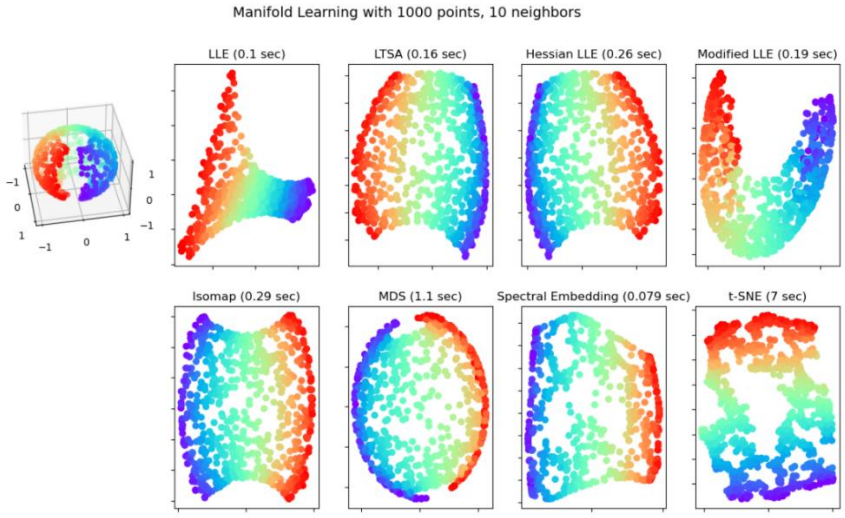


Figure 49: Examples of manifold learning dimension reduction algorithms applied to a three-dimensional data set (Credit: Grobler, J., 2022)

Several dimension reduction algorithms were tried to reduce the dimensionality of the patient-level data set while retaining the most informative relationships in its lower-dimensional representation. Two dimension reduction techniques stood out as preserving key relationships

present in the original dataset: Iso-linear Mapping (Isomap) and Uniform Manifold Approximation and Projection (UMAP).

The Isomap algorithm allows for a piecewise linear approximation of the geodesic distance for non-neighboring points. It can handle non-linear data because while it utilizes Euclidean distance to approximate the geodesic distance for neighboring or nearby points, it uses a series of incremental distance approximations (estimated using tangential vectors from every moment along the shortest path) to piecewise-estimate the distances between non-neighboring points (Das, 2020). Once it has calculated all the distances it needs, Isomap produces a weighted graph, recording distance values as the edge weights. It then takes the pairwise square distances between all the points and extracts low-dimensional coordinates for each point, producing a new, lower-dimensional data set (Tenenbaum, de Silva & Langford, 2000).

A typical example of how Isomap differs from prior dimension reduction techniques uses a "jelly roll"-shaped data set for illustration (Figure 56). With typical projection-style dimension reduction, the roll-shaped data might be flattened across the top of the roll, intermingling the different-colored points in the example manifold. This produces a lower-dimensional representation that does not retain the local relationships between the data points. Since the geodesic distance approximation technique used by the Isomap algorithm allows it to estimate point-to-point distances – regardless of locality – as long as more points can be found, the resulting distance graph allows for the projection of a lower-dimensional data set that preserves the local relationships of the data points. A common misconception about Isomap is that it only works well on convex data. However, a recent paper by Trosset & Buyukbas of Indiana University (2021) provides mathematical evidence that what Isomap really does is "produce a Euclidean representation of a non-Euclidean geometry," even if the low-dimensional mapping it

produces distorts some of the distances in the original geometry and the complexity of the surface estimated does not allow for parameterization recovery (p. 17-24) as it does when a convex manifold – such as the Swiss roll – is estimated.

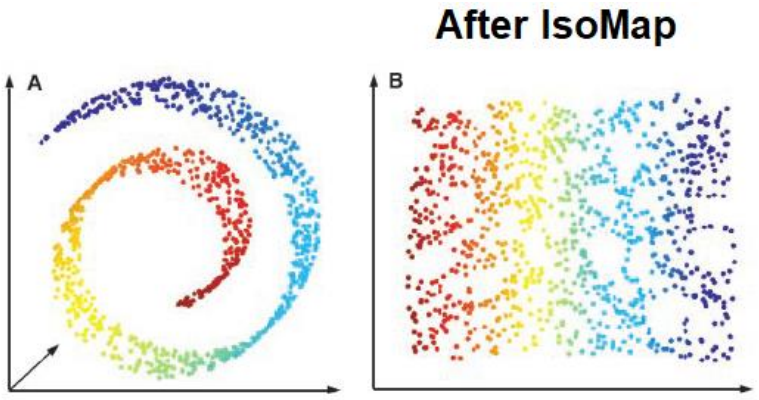


Figure 50: Isomap: Jelly Roll Example
Photo Reference: https://i.stack.imgur.com/pa1FR.png

A new and extremely useful dimension reduction technique, Uniform Manifold Approximation and Projection (UMAP) was developed by Leland McInnes, John Healy, Nathaniel Saul, and Lukas Großberger in 2018. UMAP's popularity is justified by its speed and mathematical sophistication, allowing it to produce a low-dimensional graph from complex, high-dimensional data maintaining the proportional distances and relationships between original features. UMAP creates a raw distance matrix using similarity scores based on the number of high-dimensional neighbors each data point has. Fuzzy union operation makes the similarity scores symmetrical. UMAP then projects the similarity graph into a lower-dimensional space using Spectral embedding, a non-linear calculation process that preserves the local distances between data points – much like Isomap but using a different methodology (Laplacian eigenmaps) (McInnes, Healy & Melville, 2020). Details of Spectral embedding will be discussed further in the Clustering section (Section 5.3). This initial projection is then iteratively updated using lower-dimensional similarity calculations (Starmer, 2022).

Both UMAP and Isomap allow for flexibility in distance metrics or dissimilarity calculations. With either approach, the user can choose the method the algorithm will use to calculate the similarity scores between high-dimensional neighbors from a series of commonly-used metrics and dissimilarity measures. UMAP even allows users to implement their own measures (McInnes et al., 2018). Another essential advantage of UMAP is its ability to preserve the global structure of data due to its choice of cost function (Oskolkov, 2019). While another popular dimension reduction algorithm, t-distributed Stochastic Neighbor Embedding (t-SNE)'s (van der Maaten & Hinton, 2008) performance often degrades as perplexity increases, UMAP's use of nearest neighbor calculation reduces its sensitivity to increases in n-neighbors once the n-neighbor parameter value reaches a threshold (Figure 56) (Oskolkov, 2020).

Before utilizing any dimension reduction technique, it is necessary to have a complete and fully-scaled feature set. Elimination of NULL values is always a challenge for the researcher, because imputation typically either complicates or over-simplifies feature ranges. Fortunately, because the patient-level data set's most key features primarily consist of count data, it was possible to fill most NULL count values with zeros without changing the meaning of these variables. Interval values were likewise filled with a standard "censorship" value of 730 days.

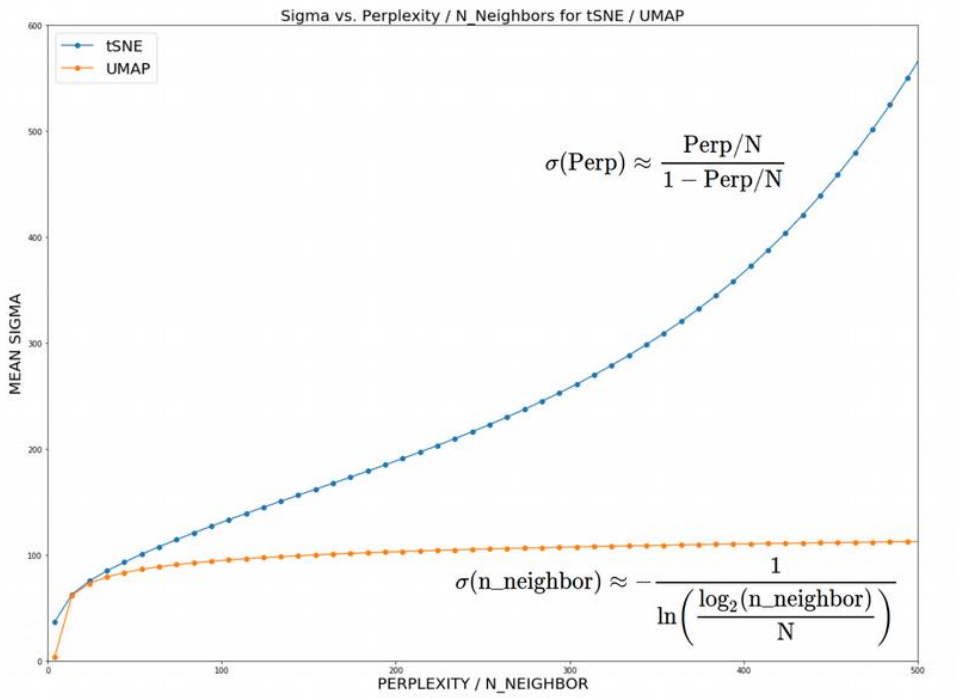


Figure 51: Behavior of mean sigma as a function of perplexity / n_neighbors for tSNE / UMAP (Credit: Oskolkov, 2020)

The remaining missing data were imputed using Sci-kit Learn's (Pedregosa et al., 2011) experimental implementation of Multivariate Imputation by Chained Equations (MICE) (van Buuren & Groothuis-Oudshoorn, 2011). This iterative algorithm estimates values for each missing data point using all the other data points in the data set. As MICE adds data points, it uses these new data points to continue the process of estimating additional missing values. It continues the process until all of the data points are filled in. Following imputation, the data set was scaled using Sci-kit Learn's StandardScaler (Pedregosa et al., 2011).

Each dimension reduction technique requires careful parameter tuning to produce the best results. For both algorithms, a number of nearest neighbors (k) needed to be selected to establish the amount of global vs. local structure that would be preserved in the reduced feature sets. This parameter was backed into based on the final choice of clustering algorithm after a great many

methods were tried. A high number of neighbors (k=200) and a moderate number of features (Isomap=13, UMAP=15) were selected for both embeddings, a) to preserve relationships between data points in an area of high density with data points in areas of lower density, and b) to produce a "soft reduction" in the data that would preserve its overall structure while still reducing dimensionality. In the same spirit, a higher minimum distance (0.45) was chosen for the UMAP algorithm to help spread apart densely packed data points from the original data in the low-dimensional representation.

One of the most critical choices was the distance or dissimilarity measure to pass to both algorithms to produce a low-dimensional feature set that would preserve the essential information in the original data. The most commonly used metrics, including Euclidean distance, Manhattan distance, and Mahalanobis or covariance distances, typically work well on evenly-spaced, primarily linearly-related feature sets. Attempts to produce meaningful low-dimensional representations of the patient-level dataset using these metrics were mainly failures, as the below images show (Figures 52 & 53). While correlation distance used against the Isomap features produced a better result than the others, apart from the lowest utilizers, utilization levels were not well separated in the low dimensional features.

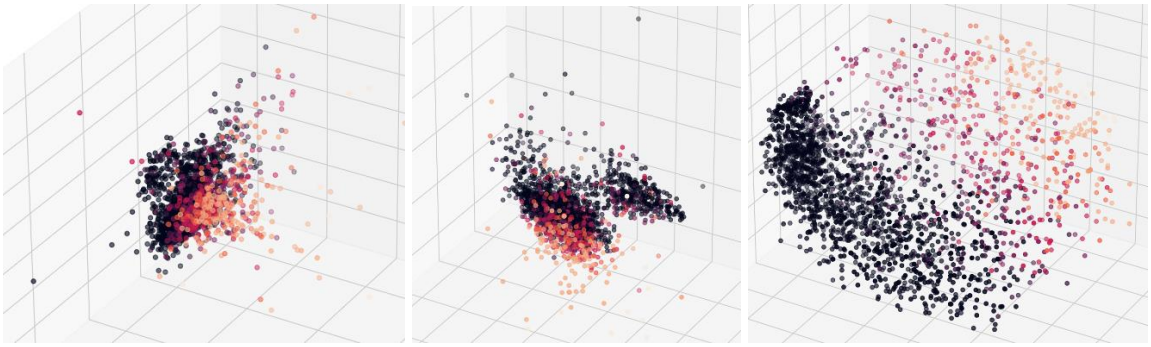


Figure 52: Isomap Reductions (3/5 dimensions, k=200) using Euclidean (left), Mahalanobis (center) and Correlation (right) distances

Plotted against y of "ed_visit_group" (black=group 0, light yellow= group 6)

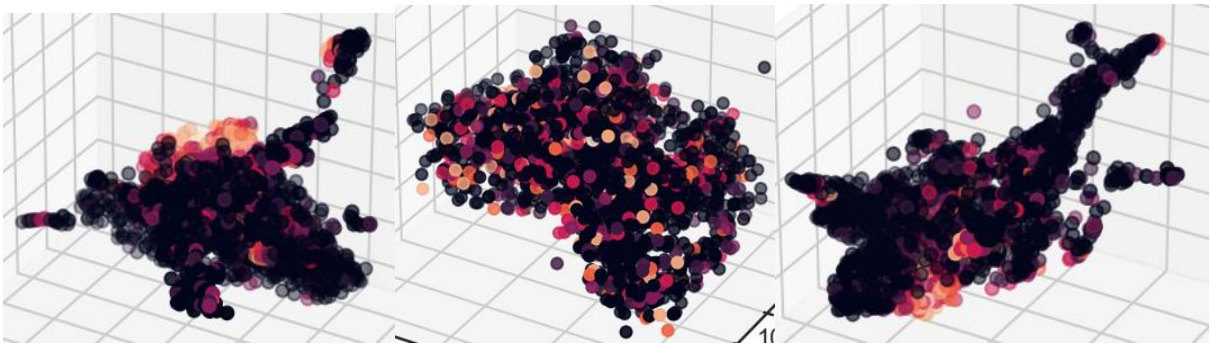


Figure 58: UMAP Reductions (3/5 dimensions, k=15, minimum distance=0.25) using Euclidean (left), Mahalanobis (center) and Correlation (right) distances

Plotted against y of "ed_visit_group" (black=group 0, light yellow= group 6)

The Bray-Curtis dissimilarity is a non-Euclidean method of obtaining the absolute proportion of dissimilarity between rows of raw count data (Bray & Curtis, 1957). The measure is typically used in environmental biology, where counts of species across various sites are commonly compared to one another to pinpoint areas where changes in the ecology may be impacting species counts. The Bray-Curtis dissimilarity assumes count origins are equal in every way; for example, the measure does not typically scale by computing differences on relative counts. The usual objection to the use of the Bray-Curtis dissimilarity is that it is not a "true metric" because it violates the triangle inequality. This Euclidean axiom demands that, in every case, the distance between two points a and b must be less than the distance from a to b via another point, c. While many measures violate this property, they are referred to as "dissimilarities" and not "metrics" because this property is part of the mathematical definition of a metric (Greenacre, 2008). While environmental biologists would need to be concerned with the question of whether or not species counts at sites that vary by size or importance should be considered equal, no such problem exists when using the Bray-Curtis dissimilarity to compare the number of visits or visit reasons by category across individuals in a patient sample. Considering every patient equal to every

other while focusing on the absolute level of difference in each patient's set of visit counts and visit reasons was precisely the behavior I was hoping the chosen measure would achieve. Another helpful behavior of the Bray-Curtis dissimilarity in the dataset context is that zero values across pair-wise comparisons result in a NULL value for the dissimilarity score. Since most algorithms ignore NULL values, this helps increase the salience of information that differentiates patients from one other.

Implementation of the Bray-Curtis dissimilarity in the context of the patient-level data works as follows. The Bray-Curtis dissimilarity – $BC_{ij} = 1-(2*C_{ij})/(S_i+S_j)$ – is computed pair-wise, where C_{ij} is the sum of the lesser values for each of two patients i & j, S_i is the sum of counts for the first patient, and S_j is the sum of counts for the second patient (Bobbitt, 2021). For instance, suppose the following data represented diagnosis and visit counts for two patients in the aggregate dataset:

	SUD	MHD	CVD	INF	DM	ED VISITS	OP VISITS
Patient 1	12	3	0	5	0	3	1
Patient 2	0	1	2	2	6	1	5

Table 9: Toy Example Used to Illustrate Calculation of the Bray-Curtis Dissimilarity Between Patients

To calculate C_{ij}, we must first sum the blue numbers in the table. These represent the lesser values for each of the two patients:

$$C_{ij} = 0 + 1 + 0 + 2 + 0 + 1 + 1 = \mathbf{5}$$

Then, we calculate the sum of each row for each patient, which is 24 for patient 1 and 17 for patient 2. We plug our numbers into the formula, $BC_{ij} = 1-(2*5)/(24+17)$, and get $BC_{ij} = 0.756$, indicating that these two patients are 75.6% different in terms of diagnosis and visit counts. Passing this parameter to the dimension reduction algorithms, Isomap and UMAP, causes each

algorithm to utilize this calculation to create a matrix of – in this case – dissimilarity scores, computed pair-wise, for all of the patients in the dataset. The algorithms then carry out their additional respective calculations to produce a lower-dimension dataset projection.

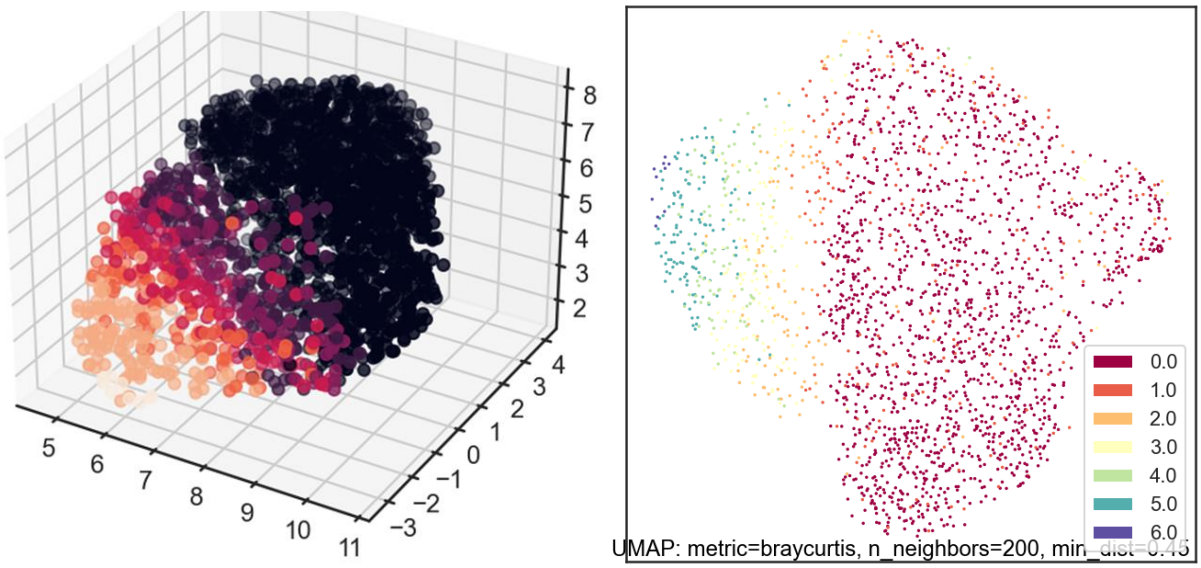


Figure 54: Isomap (left) and UMAP (right) reductions (2-3/13-15 dimensions shown) produced using the Bray-Curtis dissimilarity (k=200, minimum distance (UMAP)=0.45) Plotted against a y of "ed_visit_group."

Compared to previous attempts to produce a low-dimension representation of the data that preserved essential information about utilization and diagnosis patterns from the original data, these feature sets showed much more promise. The repurposing of the Bray-Curtis dissimilarity to calculate differences between patients in the sample was critical to the success of this project.

5.3 Clustering

Although the reduced feature sets exhibited more normality and linear-relatedness than the original data, I first attempted k-Means clustering on the two reduced data sets with skepticism.

k-Means clustering was used to separate the data set into three utilization groups using either the Isomap or UMAP features. In both instances, the number of clusters was selected using Silhouette scores, although these were low for both feature sets. These three groups consisted of a) high outpatient/clinic utilizers, b) high emergency department utilizers, and c) low utilizers. The cluster assignments were then isolated, and ensemble classification methods were applied to the original feature set (without the uncoded categorical variables or other redundant variables) to predict the cluster assignments. Both the sci-kit learn Random Forest classifier (Pedregosa et al., 2011) and XGBoost classification (Chen & Guestrin, 2016) were successful in predicting the clusters with high accuracy. Permutation feature importances were also calculated for each classifier.

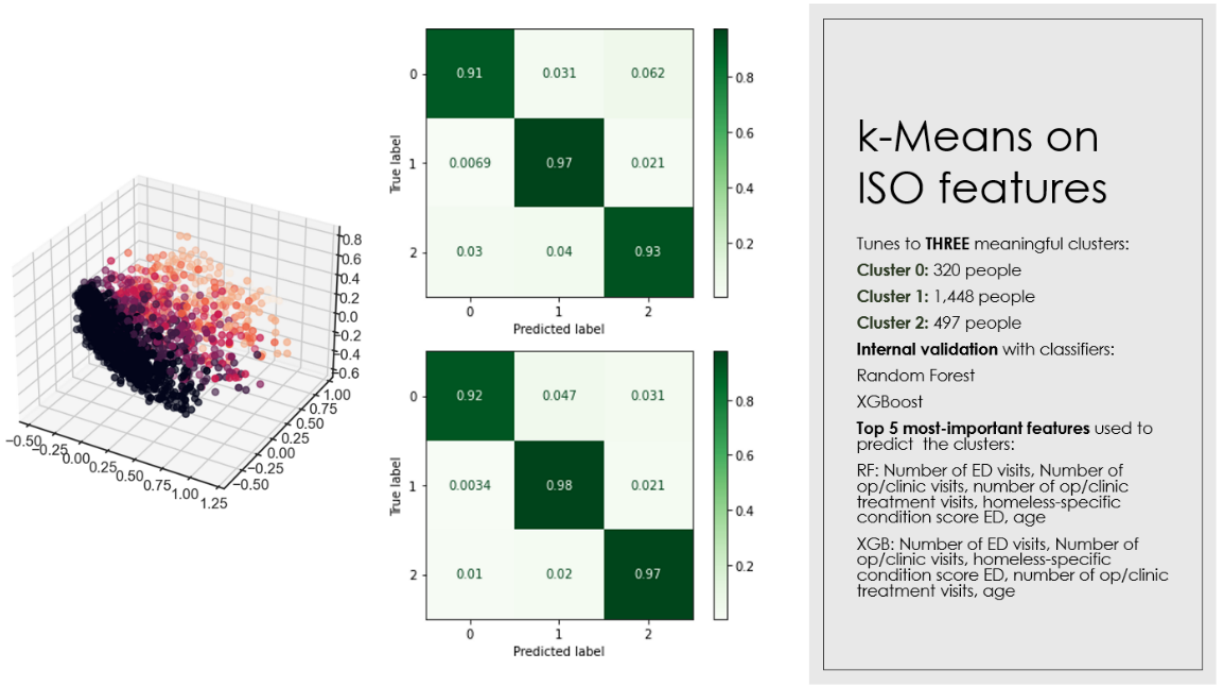


Figure 55: k-Means Clustering Results: ISO Features

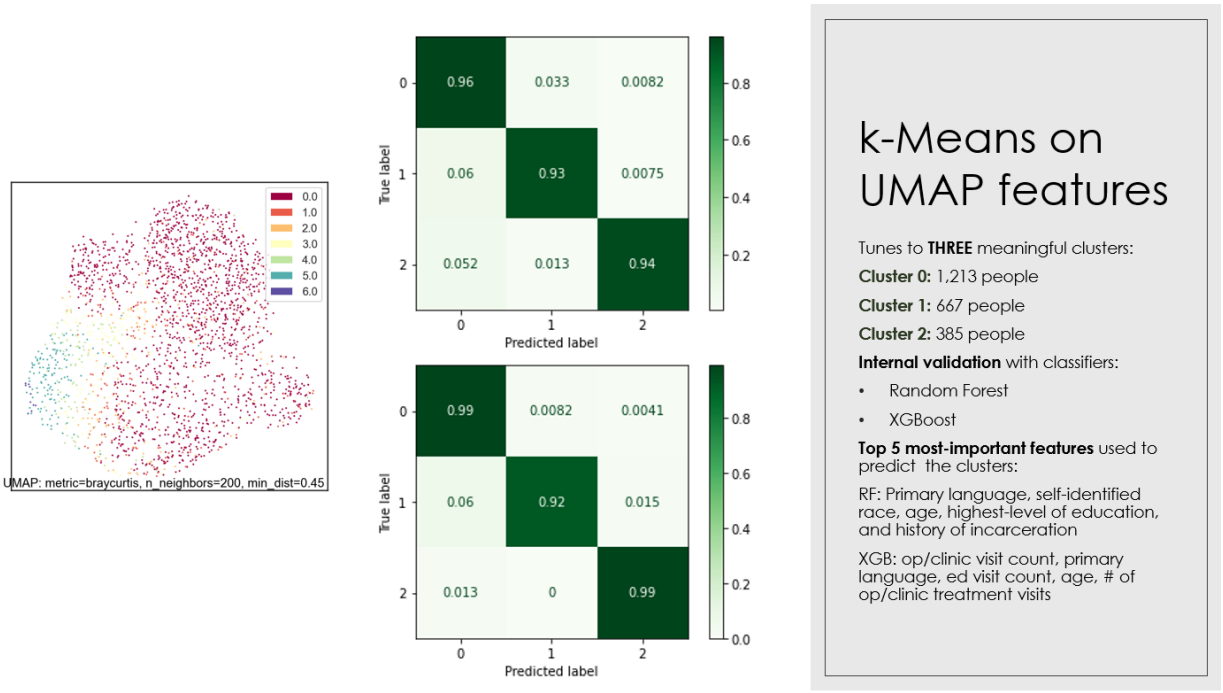


Figure 56: k-Means Clustering Results: UMAP Features

While k-Means successfully identified the three most important general clusters, I was hoping for a final result that would provide more detail into the clinic's patient sub-populations. An application of Spectral clustering (von Luxburg, 2007) using both a radial basis function (referred to as 'Spectral A') and k-nearest neighbors' approach (referred to as 'Spectral B') yielded interesting results when applied to both the Isomap and UMAP feature sets.

The name "Spectral" in Spectral clustering comes from the mathematical definition of a matrix's "spectrum" as its eigenvalues (*German*: intrinsic values) (Strang, 2019). Eigenvalues determine the magnitude of corresponding vectors (eigenvectors) together summarizing the variance of a multi-dimensional dataset. Spectral clustering starts with calculating a graph matrix (G) representing the relationships in the dataset. To compute the graph matrix, there must be a way of quantifying these relationships. The proper way of doing so depends upon the data. Are the data colored pixels? If so, perhaps the right measurement of difference tells how far away each

color is from other colors based on RGB values, for example. Many graph representations are computed using k-nearest neighbors, the Gaussian kernel/radial basis function, or Euclidean distance. Once the graph representation (G) is calculated, three matrixes can be obtained from it:

1. An incidence matrix (A): An *n* x *m* matrix where *n* is the number of nodes and *m* is the number of edges or connections. The incidence matrix summarizes the relationships between the graph's nodes (points) and edges (links).
2. A degree matrix (D): An *n* x *n* diagonal matrix summarizing the number of connections between every node and all other nodes.
3. An adjacency matrix (B): A binary *n* x *n* matrix with a diagonal of zeros that establishes whether or not every node is connected to every other node.

The second and third matrixes (the degree and adjacency matrixes) can be used to compute a symmetric, positive, semi-definite matrix called the Laplacian matrix. It is called "Laplacian" because when Laplace's finite difference equation is applied to a discrete graph, the resulting matrix represents an undirected graph's state of equilibrium (Strang, 2019). A typical Laplacian matrix is obtained by subtracting the adjacency matrix from the degree matrix, in this case: L = D - B. A normalized version of the Laplacian can also be calculated as $D^{-1}(D - B) = I\text{-}D^{-1}B$, according to the idea that a graph separation can be optimized by computing the probability of a transition from one area of the graph to another via random walk (Meila & Shi, 2001). This fits in nicely with the perturbation theory, necessary for computing clusters via eigenvalues of the normalized Laplacian when the separation between areas of the graph is imperfect and varies in density (Stewart & Sun, 1990 in von Luxburg, 2007) – a situation that applies to the separation of this dataset.

Many methodologies are proposed for making an appropriate separation in the connected graph representation of a dataset's connections symbolized by its Laplacian. I used the eigengap heuristic proposed by Chung (1997). If the eigenvalues of a graph Laplacian are represented by $\lambda_1 \ldots \lambda_k$ and the perturbation theory applies, an optimal number of clusters *n* can be determined by the differences in λ_k and λ_{k+1} where *n* is equal to the first *k* where the difference shows a significant gap between itself and the prior eigenvalue (von Luxburg, 2007; Ciortan, 2019).

The reduced Isomap and UMAP feature sets, created using a large number of nearest neighbors (k=200) and the Bray-Curtis dissimilarity, consisted of normally-distributed, linearly-related and highly-connected features. For the "Spectral A" clustering approach, the Euclidean distance metric was employed, and a value for gamma was tuned for the radial basis function for each feature set separately to create a graph representation of the reduced features' relationships from which the normalized Laplacian could be obtained and its eigengaps analyzed. The optimal value for gamma ended up being exceedingly small for the UMAP features (0.00001) and larger for the Isomap features (0.2). For the "Spectral B" approach, a value for k was tuned (k=40 was used for both feature sets) to create a graph representation of the features using the k-nearest neighbors' algorithm before following the remaining steps to compute and analyze the Laplacian eigengaps. Once the numbers of clusters were tuned, Sci-kit Learn's implementation of SpectralClustering (Pedregosa et al., 2011) was used to produce the clusters, and the original feature set was used to predict them using ensemble methods. Permutation "feature importances" allowed a first insight into the key features used to divide the data into the target clusters; however, these features and the order of their importance varied between models, leaving the true importance of each feature in classifying patients unclear.

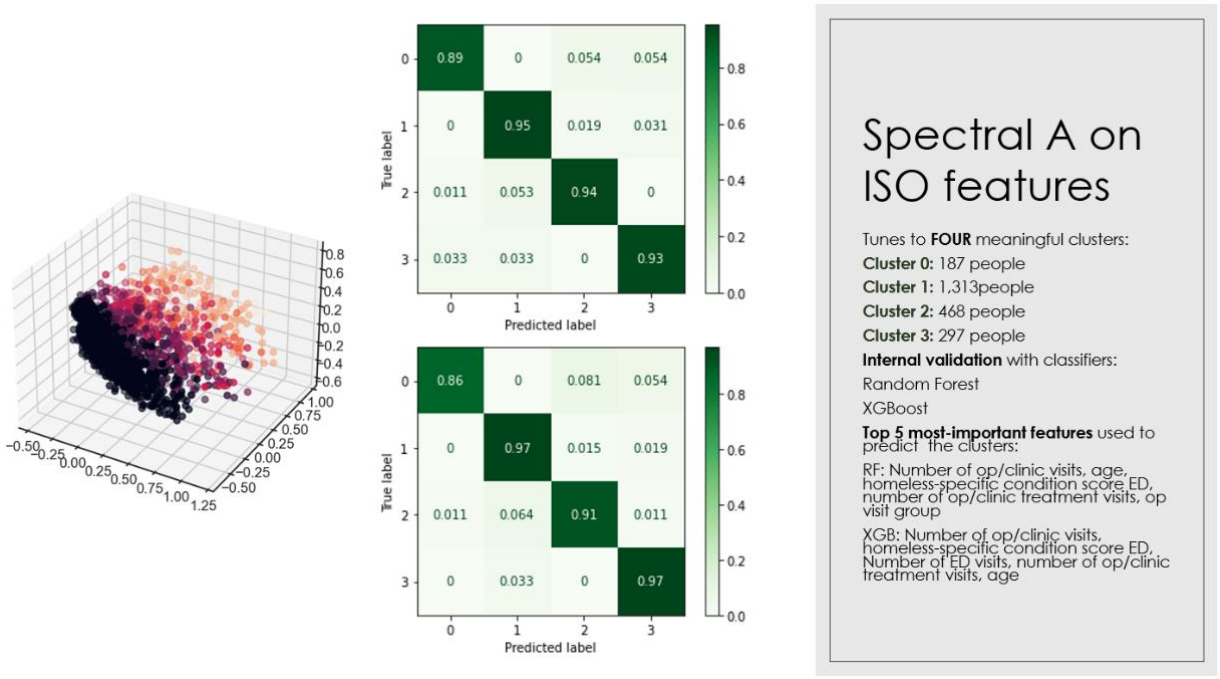


Figure 57: Spectral Clustering Results using the Isomap features and Radial Basis Function

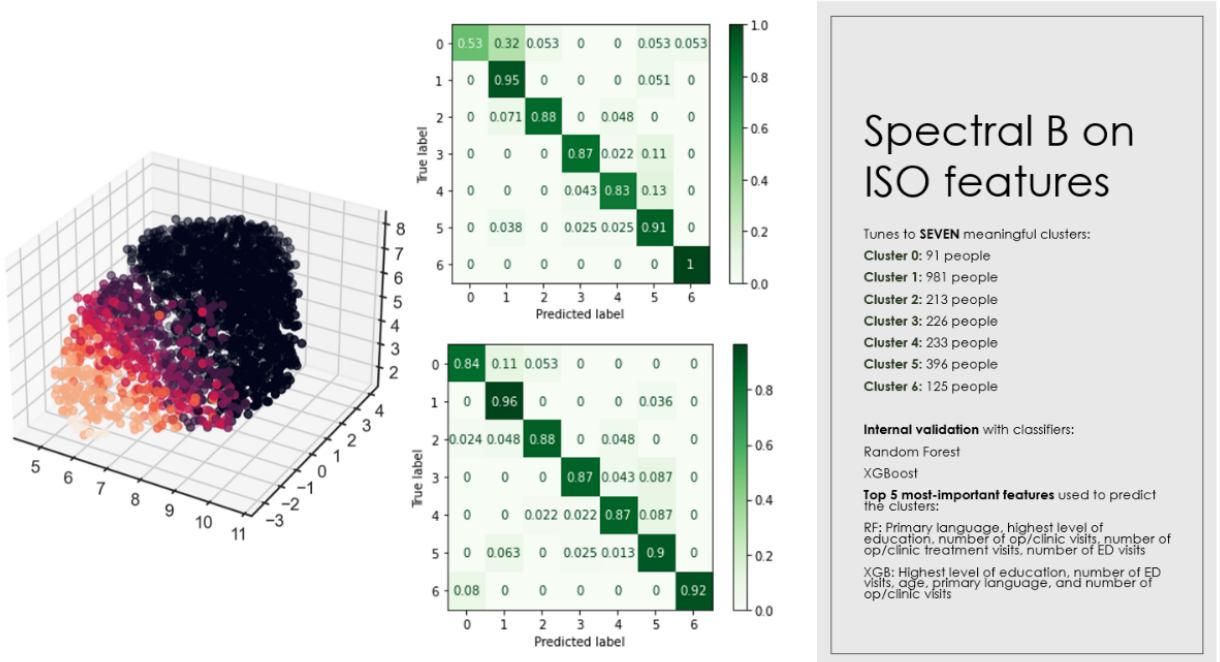


Figure 58: Spectral Clustering Results using the Isomap features and k-Nearest Neighbors

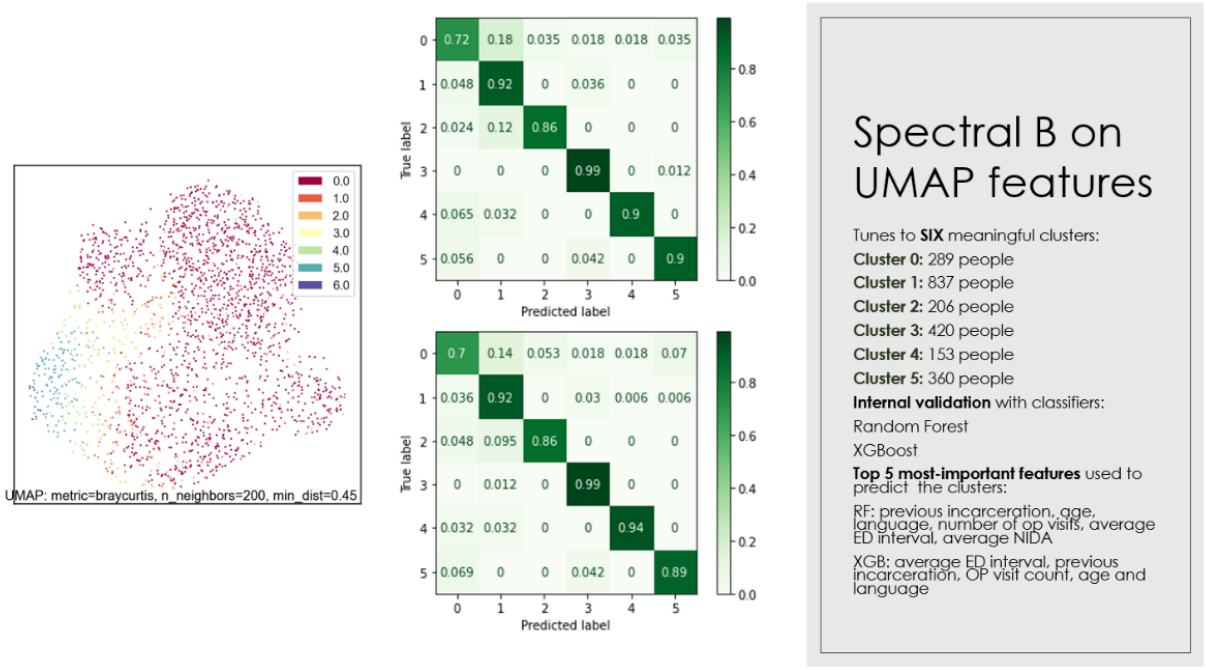


Figure 59: Spectral Clustering Results using the UMAP features and k-Nearest Neighbors

The results of "Spectral A" clustering on the UMAP features are not presented here because they did not obtain satisfactory results when classification methods were applied to this cluster set using the original data. Results using the "Spectral B" method on the Isomap features are presented, even though the Random Forest classifier results for cluster 0 show the classifier's confusion, and a sub-standard result predicting cluster membership.

5.5 Clinical Feedback

A crucial step when utilizing unsupervised learning to describe patient sub-populations is the validation and feedback of the clinicians who work with the patients daily. Whether or not these algorithms could divide the clinic's patients into meaningful groups worthy of further analysis was something only the HCHM team could advise on. The clustering results were presented twice. First, we presented to my primary contact, Matthew Augeri, HCHM's Health Information Systems Analyst. Then, we expanded the presentation to a wider audience, including HCHM's

Director, Practice Manager, and other stakeholders, including Timothy Soucy, Catholic Medical Center's Senior Executive Director of Community Health & Mission. I presented basic statistics describing demographic and visit data for each cluster set, and the ensemble classification results and feature importances. The team then reviewed and discussed the clustering results independently and eventually agreed that the six "Spectral B" clusters produced using the UMAP feature set should be subject to a deeper analysis.

5.6 The "Spectral B" UMAP Clusters

The "Spectral B" UMAP clusters were imperfect due to the difficulty of dividing the patient data set across many complex features. However, in large part, the clusters were able to describe recognizable sub-populations of clinic users only hinted at in the descriptive analysis.

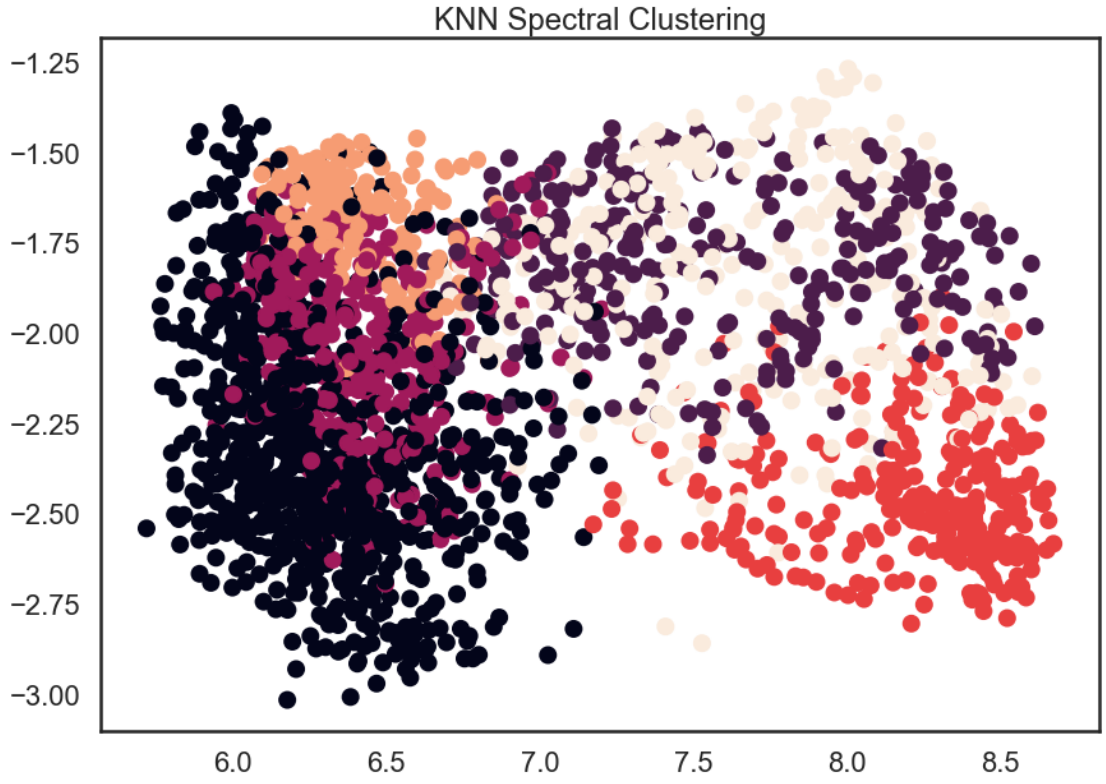


Figure 60: A Two-dimensional Visualization of the Six "Spectral B" UMAP Clusters

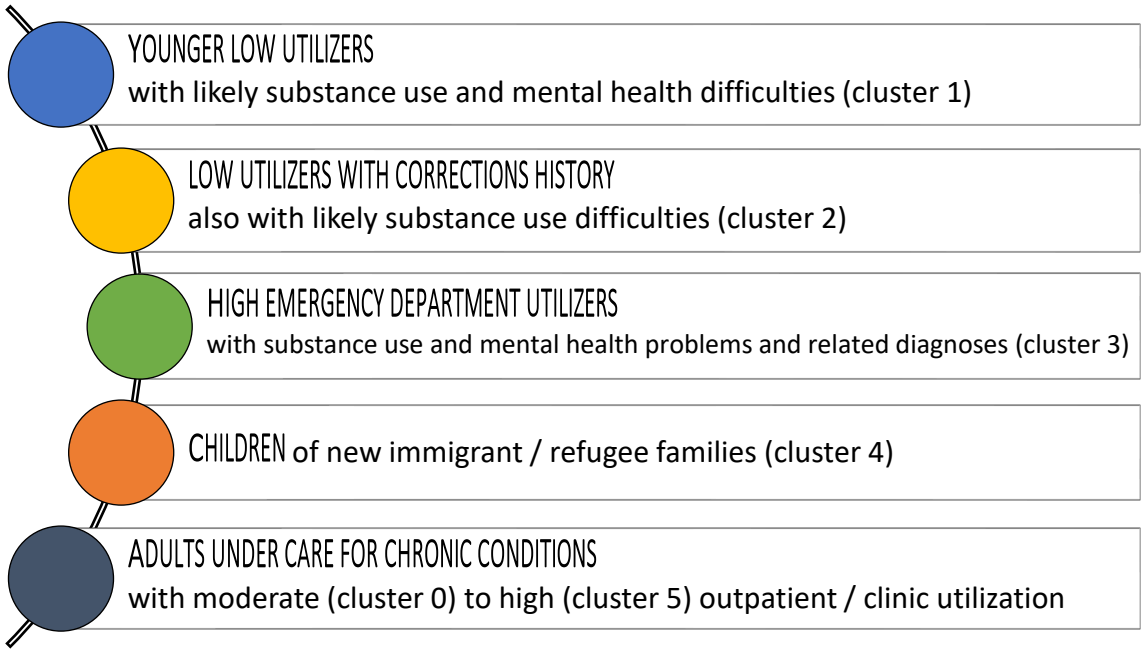
Results of the cluster set description are detailed in the next chapter, including:

- Basic statistics and correlations
- Shapley values produced from the XGBoost classification (Chen & Guestrin, 2016) using the SHAP package (Lundberg & Lee, 2017)
- Visualization and analysis of the cluster classification of using Sci-kit Learn's DecisionTreeClassifier (Pedregosa, et al., 2011)
- Summary descriptions of each cluster and how the clusters relate to clinic patient sub-populations/service groups

CHAPTER 6: FROM CLUSTERS TO SERVICE GROUPS

6.1 Introduction: Clusters and Service Groups

The six "Spectral B" UMAP clusters roughly correspond to five service groups:



Data-driven cluster descriptions and their relationships to the service groups will be explored in detail in the following sections. People experiencing homelessness may fall into one or more sub-populations, including: homeless families with children, unaccompanied youth, parenting youth, chronically or long-term homeless (often including disabled persons), homeless veterans, persons suffering from severe mental illness, people with chronic substance use problems, and victims of domestic violence (United States Department of Housing and Urban Development (HUD), 2021). Because veterans and pregnant patients are referred to other sources of care, these sub-populations – likely served by clinics in other metropolitan areas – are not all a part of the service groups described here. The HCHM service groups, instead, likely correspond to:

- Previously incarcerated, who may be chronically homeless (cluster 2)
- People with severe mental illness and substance use problems; some may be chronically homeless (cluster 3, and likely also some people in clusters 1 and 2)
- Victims of domestic violence (majority in clusters 1 and 3), and
- Housing insecure and refugee families (adults in cluster 5, children in cluster 4)

6.2 Statistics and Correlations

Domain	Variable	Metrics	Clusters					
			0 n=289	1 n=837	2 n=206	3 n=420	4 n=153	5 n=360
Visit Count and Interval Features	ED Visit Group 0: No visits 1: 1 visit 2: 2 visits 3: 3-4 visits 4: 5-7 visits 5: 8-30 visits 6: > 30 visits	0	249 (86.2%)	736 (87.9%)	185 (89.8%)	0	126 (82.4%)	249 (69.2%)
		1	34	72	19	4	14	67
		2	6	23	2	70	8	26
		3-4	0	6	0	126	4	13
		5-7	0	0	0	88	1	3
		8-30	0	0	0	122	0	2
		> 30	0	0	0	10	0	0
	OP Visit Group 1: 1 visit 2: 2 visits 3: 3-4 visits 4: 5-7 visits 5: 8-30 visits 6: > 30 visits	1	18	542 (64.8%)	86	82	60	2
		2	54	182	35	77	32	8
		3-4	73	85	43	82	41	19
		5-7	61	26	24	61	14	73
		8-30	83	2	8	106	6	231
		> 30	0	0	0	12	0	27
	Average ED Visit Interval (Days) – excludes censored	*#/n*	*6*	*29*	***2***	*416*	*13*	*44*
		min	1	0	21	0	0	1
		median	106.50	190.00	x	66.00	118.00	114.50
		mean	150.00	192.60	x	93.29	152.91	161.48
		max	453.00	466.00	178.00	549.00	607.00	566.00
		IQR	172.00	240.00	x	86.75	175.00	183.00
	Average OP Visit Interval (Days) – excludes censored	*#/n*	*272*	*328*	*122*	*350*	*93*	*359*
		min	8.00	0.00	74.00	0.00	7.00	7.00
		median	55.50	26.00	80.50	34.00	63.00	29.50
		mean	55.00	66.00	80.50	58.01	109.09	43.34
		max	95.00	399.00	87.00	587.00	602.00	319.00
		IQR	57.25	49.00	69.75	62.25	105.00	35.00
	Admissions	count	4	5	0	78	0	17
	Deceased	count	0	0	0	3	0	0
	ED Visits with Non-Emergent Primary Diagnoses	0-1	288 (99.7%)	837 (100%)	206 (100%)	237 (56.4%)	152 (99.3%)	353 (98.1%)
		2-3	1	0	0	121	1	7
		4-5	0	0	0	40	0	0
		> 5	0	0	0	22	0	0
	Number of Any Visit Days During Extreme Weather	0-1	236 (81.7%)	804 (96.1%)	188 (91.3%)	222 (52.9%)	143 (93.5%)	148 (41.1%)
		2-3	39	28	2	125	9	133
		4-5	12	4	2	43	1	43
		> 5	2	1	1	30	0	36

	Variable	Metrics	Clusters					
			0 n=289	1 n=837	2 n=206	3 n=420	4 n=153	5 n=360
	Number of Outpatient Preventive Visits	0-1	111	734	138	202	97	20
		2-3	97	97	49	116	52	81
		4-5	42	6	10	47	3	70
		> 5	39	0	9	55	1	189
Conditions and Acuity	Average Outpatient Visit Level	min	1	1	1	1	1	1.56
		median	2.67	3	3	3	2.86	2.67
		mean	2.60	2.73	2.89	2.81	2.69	2.68
		std	0.60	0.64	0.72	0.60	0.46	0.40
		max	4	5	5	5	4	4.67
		IQR	0.79	1	0.80	0.61	0.50	0.57
	Average # of Times 'Pain' in Visit Notes	0-1	61	182	30	105	89	78
		2-3	124	218	62	185	56	207
		4-5	109	239	73	154	24	159
		> 5	72	255	73	83	8	69
	Average # of Times 'Disability' in Notes	> 1	8	1	1	1	1	1
	Average Elixhauser Score based on ED diagnoses	min	0	0	0	0	0	0
		median	0	0	0	0.67	0	0
		mean	0.08	0.04	0.03	0.88	0.02	0.32
		std	0.32	0.22	0.21	0.81	0.10	0.83
		max	3.00	2.00	2.00	5.17	1.00	7.00
		IQR	0	0	0	0.82	0	0
	Average Elixhauser Score based on OP / clinic diagnoses	min	0	0	0	0	0	0
		median	0.40	0.67	0.29	0.71	0	0.99
		mean	0.53	0.63	0.52	0.73	0.02	0.96
		std	0.59	0.67	0.61	0.64	0.10	0.59
		max	3.00	3.00	3.00	3.22	0.75	2.75
		IQR	0.94	1.00	1.00	0.96	0	0.80
	Average Homeless-Specific Condition Score, ED Visits	min	0	0	0	0	0	0
		median	0	0	0	2.00	0	0
		mean	0.28	0.17	0.18	2.21	0.22	0.69
		std	0.85	0.57	0.68	1.19	0.51	1.32
		max	6.00	5.00	4.00	8.00	2.50	8.50
		IQR	0	0	0	1.30	0	1.00
	Average Homeless-Specific Condition Score, OP Visits	min	0	0	0	0	0	0
		median	0.83	0	0	0	0	1.17
		mean	0.86	0.95	0.86	1.05	0.23	1.19
		std	0.67	0.77	0.69	0.60	0.31	0.50
		max	4.00	4.00	4.00	3.00	1.00	3.58
		IQR	0.77	0.67	0.67	0.71	0.50	0.65
	Emergency Visit Reason Categories:							
	CVD_ed	count	5	0	0	94	2	14
	HF_ed	count	0	0	0	12	0	2
	HTN_ed	count	7	4	2	104	0	46
	URI/PNA_ed	count	6	16	4	209	11	42
	AS/COPD_ed	count	5	4	2	119	1	19
	NEURO_ed	count	3	2	2	88	4	6

	Variable	Metric	Clusters					
			0 n=289	1 n=837	2 n=206	3 n=420	4 n=153	5 n=360
	SUD_ed	count	6	28	4	301	0	14
	MHD_ed	count	4	14	1	183	0	7
	CA_ed	count	0	0	0	13	0	3
	PREG_ed	count	0	3	0	21	3	0
	DM_ed	count	1	2	0	36	0	30
	INF_ed	count	8	14	5	214	5	16
	LIV_ed	count	2	2	0	36	0	3
	REN_ed	count	0	0	0	16	1	3
	COG_ed	count	3	0	1	21	1	5
	SENS_ed	count	2	4	3	107	2	15
	PAIN_ed	count	14	12	2	275	9	41
	TOB_ed	count	13	32	9	291	1	23
	ACC/INJ_ed	count	10	24	4	216	8	25
	Outpatient/clinic Visit Reason Categories:							
	CVD_op	count	20	20	10	40	1	88
	HF_op	count	2	1	0	3	0	7
	HTN_op	count	73	65	30	88	0	233
	URI/PNA_op	count	60	52	25	94	26	130
	AS/COPD_op	count	36	33	22	74	0	68
	NEURO_op	count	25	36	7	42	3	48
	SUD_op	count	81	398	75	218	3	92
	MHD_op	count	67	113	23	117	6	94
	CA_op	count	5	6	2	11	1	31
	PREG_op	count	1	0	1	1	3	0
	DM_op	count	36	22	12	41	2	182
	INF_op	count	61	86	43	121	18	110
	LIV_op	count	7	8	2	20	0	14
	REN_op	count	3	2	1	3	0	23
	COG_ op	count	6	5	1	10	4	23
	SENS_op	count	32	29	14	38	1	89
	PAIN_op	count	102	96	37	120	16	194
	TOB_op	count	22	14	7	31	0	44
	ACC/INJ_op	count	29	36	17	67	8	63
Clinic Measures (derived from notes)	Average Height (Inches)	min	52.50	58.00	55.00	33.65	21.66	50.10
		median	66.00	68.75	68.50	67.00	54.00	64.12
		mean	66.23	68.43	68.20	67.20	52.06	64.52
		std	4.29	3.68	3.68	4.70	12.78	4.79
		max	82.32	81.00	78.00	84.00	82.50	83.48
		IQR	6.00	5.00	5.00	6.00	18.18	7.08
	Average Weight (Pounds)	min	97.70	86.37	98.50	88.30	9.22	63.96
		median	175.40	172.05	164.35	176.30	75.66	163.12
		mean	182.27	178.92	170.61	183.42	84.78	173.84
		std	45.18	43.05	36.98	50.44	49.93	47.08
		max	373.60	394.50	322.95	467.30	236.60	441.48
		IQR	58.20	52.30	46.49	54.02	81.44	53.30

Variable	Metrics	Clusters					
		0 n=289	1 n=837	2 n=206	3 n=420	4 n=153	5 n=360
Tobacco / Nicotine Use	current	134	662	173	348	3	123
	former	21	46	2	18	0	39
	never	106	89	14	34	137	161
	quitter	4	2	3	9	0	30
Average NIDA TAPS Score	0	98	80	15	46	43	194
	1	12	14	3	8	1	14
	2	8	7	5	9	0	17
	3-5	68	256	74	146	0	70
	> 5	34	364	76	168	0	27
Average HgbA1C	*#/n*	*16*	*8*	*8*	*34*	*0*	*167*
	< 6	3	4	3	18	0	47
	6-7	2	1	1	4	0	20
	> 7	8	2	2	6	0	49
Average Random Office Blood Glucose	*#/n*	*24*	*18*	*10*	*33*	*0*	*147*
	< 120	8	5	3	10	0	28
	121-150	2	3	1	5	0	27
	151-200	3	3	2	3	0	38
	> 200	11	7	4	15	0	53
Average PHQ-2 Score	< 2	30	59	17	38	5	45
	2-3	44	258	58	120	4	67
	4-6	25	246	29	124	0	23
Average HARK Score	> 1	8	42	5	40	1	11
Calculated BMI (kg/m^2)	min	16.00	14.48	14.69	13.37	10.72	15.36
	median	28.17	25.64	25.12	27.03	17.96	27.97
	mean	29.31	26.79	25.74	28.74	19.20	29.21
	max	60.91	57.11	47.69	78.27	38.67	63.34
	IQR	9.32	6.48	5.86	7.81	6.76	7.56
Average Systolic Blood Pressure (mmHg)	min	94.00	86.00	90.00	86.00	84.00	99.00
	median	123.00	121.00	122.00	121.00	102.00	129.50
	mean	123.37	121.05	123.46	122.43	103.99	130.08
	std	13.92	13.90	16.85	15.25	11.47	14.95
	max	170.00	190.00	195.00	194.00	148.00	229.00
	IQR	19.00	19.00	21.50	19.00	15.50	18.50
Average Diastolic Blood Pressure (mmHg)	min	54.00	47.00	53.00	50.00	41.00	57.00
	median	76.00	74.00	75.00	75.00	63.00	78.00
	mean	76.49	74.69	76.27	75.66	63.63	77.66
	std	9.10	9.54	10.64	9.41	7.57	7.54
	max	109.00	114.00	109.00	104.00	86.00	109.00
	IQR	12.00	12.00	14.00	13.00	11.00	10.00
Veteran	count	12	38	5	16	0	6

Variable	Metric	Clusters					
		0 n=289	**1 n=837**	**2 n=206**	**3 n=420**	**4 n=153**	**5 n=360**
Age	min	9	20	28	13	2	26
	median	51	36	48	43	12	61
	mean	48.40	38.35	47.37	44.55	14.78	60.43
	std	14.37	10.83	11.18	12.49	9.47	11.79
	max	85	79	70	86	52	94
	IQR	22	13	19	20	11	14.25
Sex / Gender	male	136	631	155	255	74	185
	female or another	153	206	51	165	79	175
Housing Status							
UNKNOWN		228	813	0	255	151	220
STREET		5	4	23	25	0	7
SHELTER		13	3	31	31	0	13
TRANSITIONAL/TREATMENT		6	7	57	40	0	8
DOUBLE-UP		19	8	52	38	2	61
OTHER/SRO		18	2	43	31	0	51
Highest Completed Education							
UNKNOWN		226	823	1	271	147	220
0-8 GRADE		10	2	8	11	2	45
9-12 GRADE		5	6	33	24	3	22
HIGH SCHOOL/GED		26	3	103	75	0	37
ANY COLLEGE		22	3	61	39	1	36
Self-identified Race							
WHITE		183	697	183	372	14	135
BLACK or AFRICAN AMERICAN		57	63	13	20	104	82
ASIAN		30	10	3	1	31	129
MORE THAN ONE		8	27	3	17	0	7
OTHER/UNKNOWN		8	28	1	6	4	2
AI/AN or NH/PI		3	11	3	4	0	5
Self-identified Ethnicity							
NON-HISPANIC		238	587	190	341	145	338
OTHER/UNKNOWN		24	167	2	54	4	11
HISPANIC or LATINX		27	83	14	25	4	11
Primary Language							
ENGLISH (0)		247	830	204	417	28	183
ANOTHER LANGUAGE (1)		42	7	2	3	125	177
Insurance							
NONE / SELF-PAY or UNKNOWN (0)		79	291	50	52	12	64
ANY INSURANCE (1)		210	546	156	368	141	296
Corrections History							
NONE or UNKNOWN (0)		268	837	51	319	153	303
ANY CORRECTIONS HISTORY (1)		21	0	155	101	0	57
'Jail' Mentions in Notes > 2		4	48	9	9	1	1

Table 10: Description / Frequency of "Spectral B" UMAP Clusters

NOTE: *Correlation heatmaps showing relationships between each cluster and all variables are located in Appendix B, Section 2.*

6.3 Describing the Clusters

Just by looking at frequency counts, ranges of values (Table 10), and correlation plots (Appendix B, Section 2), it was possible to begin to get a picture of each cluster, and thus, each service group. Starting with the distribution of age and sex within each cluster, a wide range of ages were present in all of the clusters. Owing to the fuzziness of this clustering problem, and the choice to use the Bray-Curtis dissimilarity to focus on utilization counts, the clusters divided more by diagnosis categories and visit counts than by demographic distinctions. However, there were still some significant descriptions in the demographic data for each cluster. For example, clusters one, two, and five had no members younger than 20, 28, and 26 years of age respectively, while their median ages ranged from 36 years for cluster one (standard deviation (sd) 10.83) to 48 years for cluster two (sd 11.18), and 61 years for cluster five (sd 11.79). While cluster four contained patients ranging in age from 2 to 52 years, the median age was 12 years (sd 9.47). Cluster zero had a few very young and a few very old patients, but most patients in that cluster were between 40 and 60 years old (median 51, sd 14.37).

While the ratio of males to other genders in clusters one and two was about three males for every non-male, in clusters zero and four there were more non-males/females, and in cluster three the ratio was about 1.5 males for every non-male. In cluster five, there were approximately equal numbers of males and non-males.

With regard to housing status and educational completion, the majority of patients in clusters zero, one and four had unknown data because these data points were sparsely populated. Only clusters two and three contained a substantial number of individuals identified in the data as

either street or shelter dwellers or as living in transitional or treatment housing. In contrast, cluster five seemed to have many more individuals living doubled-up or in apartments or rooming houses. Similarly, all clusters except clusters two and three had significant numbers of people with unknown educational completion. A majority of those in clusters two and three had completed either high school, or some amount of college. A substantial number of people (n=45) in cluster five indicated their highest level of education was less than 8th grade (0-8G).

White people made up the majority in clusters one, two, and three, were more than half (63%) of cluster zero, a minority (9.2%) in cluster four, and a significant minority (37.5%) in cluster five. While white people made up 83.3% of cluster one, this cluster also had the most heterogeneity with respect to race and had more people who identified as more than one race or "another race". Cluster one also had the majority of people who identified their ethnicity as Hispanic or Latinx. English was the primary language for the majority of all patients; however, 81.7% of patients in cluster four and 49.2% of patients in cluster five identified another language as primary.

The largest numbers of people who did not identify any insurance plan (either they were uninsured, were providing self-payment, were on a sliding scale, or did not provide any information about their insurance) were in clusters zero, one, and five, with cluster one having the majority (291 out of 548 patients, or 53.1%). Although patients in cluster one had almost no identified corrections history, they also had the majority of instances where "jail" or "incarceration" was mentioned in their clinical notes more than two times (48 out of 72, or 66.7%). Patients with identified corrections history (n=334) clustered into groups two and three, with 155 (46.4%) in cluster two, and 101 (30.2%) in cluster three. Due to similarities in patterns of null data between the corrections history and housing and education variables, it is possible

that null values in the original corrections history data did not truly indicate a lack of corrections history, but rather the presence of missing data.

Moving on to assessments and readings, a few things stand out. In clusters one, two, and three the number of current smokers/nicotine users far outnumbered those who were former or never users, while in clusters zero and five there were equal numbers of users and non-users. Cluster four, which consists predominantly of children, had only three identified nicotine users.

Although there were outliers in all groups, cluster five had the highest median and mean blood pressure readings, with a median average pressure of 130/78 mmHg (sd 14.95/7.54). The highest BMIs were in clusters zero and five, with median BMI in cluster five of 27.97 kg/m^2 (sd 6.52), and in cluster zero, 28.17 kg/m^2 (sd 6.95). The lowest BMIs were in clusters two and one, with cluster two median of 25.12 kg/m^2 (sd 5.05) and cluster one median of 25.64 kg/m^2 (sd 5.75).

Cluster five had the majority of patients with diabetes (167 out of 306 patients, 54.6%) and also the highest number of patients with elevated HgbA1C (49 patients with readings > 7%) and random blood glucose readings (53 patients with readings > 200 mg/dl). The highest NIDA TAPS, PHQ-2, and HARK scores were all most common in clusters one and three. Cluster one had 364 patients with TAPS scores >5, 246 patients with PHQ-2 scores between 4 and 6, and 42 patients who were HARK positive. Cluster three had 168 patients with TAPS scores >5, 124 with PHQ-2 of 4-6, and 40 HARK positive patients. People in cluster five had many moderate scores, with 70 patients with TAPS scores of 3-5, and 67 with PHQ-2 of 2-3. Seventy-one percent of patients in cluster two, and 63% of patients in clusters zero and five had an average of four or more pain mentions in clinical notes. Based on the many commonalities between clusters one and two and cluster three, it is easy to imagine that patients in these clusters may have similar characteristics except with respect to age, as the interquartile age range for cluster one

was 29 ½ to 42 ½ years, whereas for cluster three it was 33 to 53 years and for cluster two it was 38 ½ to 57 ½ years.

With regard to visit-related features, there are few surprises. Cluster three had all the high emergency department utilizers, and the majority of those admitted inpatient after an ED visit, while other clusters consisted of people who went to the emergency department a handful of times in the two years. Patients in cluster three had the most non-emergent ED diagnoses, the most visits on severe weather days, and the shortest median average interval between ED visits at 64.5 days. Patients in cluster three also had a substantial number of outpatient/clinic visits, while patients in cluster five had the most outpatient visits and patients in cluster zero had moderate clinic utilization. Patients in clusters one and two had few visits overall, with cluster two having the fewest visits. In keeping with this, patients in cluster five also had the most preventive visits of all the patients, while those in clusters zero and four had more low-acuity outpatient visits with a maximum visit level of four.

Since almost all of the high emergency department utilizers were clustered together in group three, this cluster had the most emergency department diagnosis categories across the board. Cluster five was second and had more categorical diversity than cluster three – with more visits for cardiovascular disease, hypertension, chronic obstructive pulmonary disease (COPD), upper respiratory infections (URIs), diabetes, pain, sensory disorders and accidents and injuries than for substance use and mental health disorders. Cluster one had few emergency visits; however, where they did have visits, they had higher numbers of diagnoses in the same categories as those in cluster three, including substance use disorders, mental health diseases, infections, accidents and injuries, and URIs. Similarly, the distribution of diagnosis categories for cluster zero was

akin to that of cluster five, having visits for hypertension, injuries and abdominal pain among top emergency visit reasons.

The outpatient picture in terms of diagnosis categories is more complex. Many groups had substance use disorders at the top of their list, including clusters one, two, and three. Once again, clusters one and two and cluster three were similar, with the same top visit reason categories, including substance use, mental health disease, pain, and infections. Cluster zero patients had pain as their primary outpatient visit reason, followed by substance use disorder, hypertension, mental health disorders and infections. Group five had hypertension as its number one category, followed by pain, diabetes, URIs, and infections. Group five also had the most diagnoses related to neurological diseases, cancer, renal, and liver diseases. Sensory conditions were an issue across many clusters, including clusters two, three, and five. Cluster four had few diagnoses; however, unsurprisingly, the majority were for URIs, infections, pain, and accidents and injuries – common outpatient concerns of children and adolescents.

Elixhauser scores for emergency and outpatient/clinic visits were highest for clusters three and five, with cluster three dominating in emergency Elixhauser scores (mean score 0.88, sd 0.81), and cluster five in the outpatient area (mean score 0.99, sd 0.59). Interestingly, cluster five had a higher maximum ED Elixhauser score, at 7.00; cluster three's maximum score was 5.17. Although cluster five had higher mean and median outpatient Elixhauser scores, cluster three had the highest maximum scores (3.22). The lowest scores were in cluster four, with cluster two a close second. Cluster two's median outpatient Elixhauser score was 0.29, and their maximum was 3.00. While groups three and five again dominated in terms of homeless-specific condition scores (HSCSs), cluster zero had the second-highest outpatient HSCS, with a median score of 0.83, and a maximum score of 4.00. While groups zero, one, and two typically had low

emergency HSCS scores; they also had some higher outliers, with maximum scores of 4.00 across all three groups. The sickest patients (in terms of diagnoses and comorbidities) were in groups three and five, followed by group zero. Patients with the fewest visits and diagnoses were in groups one, two, and four, although a few patients in groups one and two had high-acuity visits.

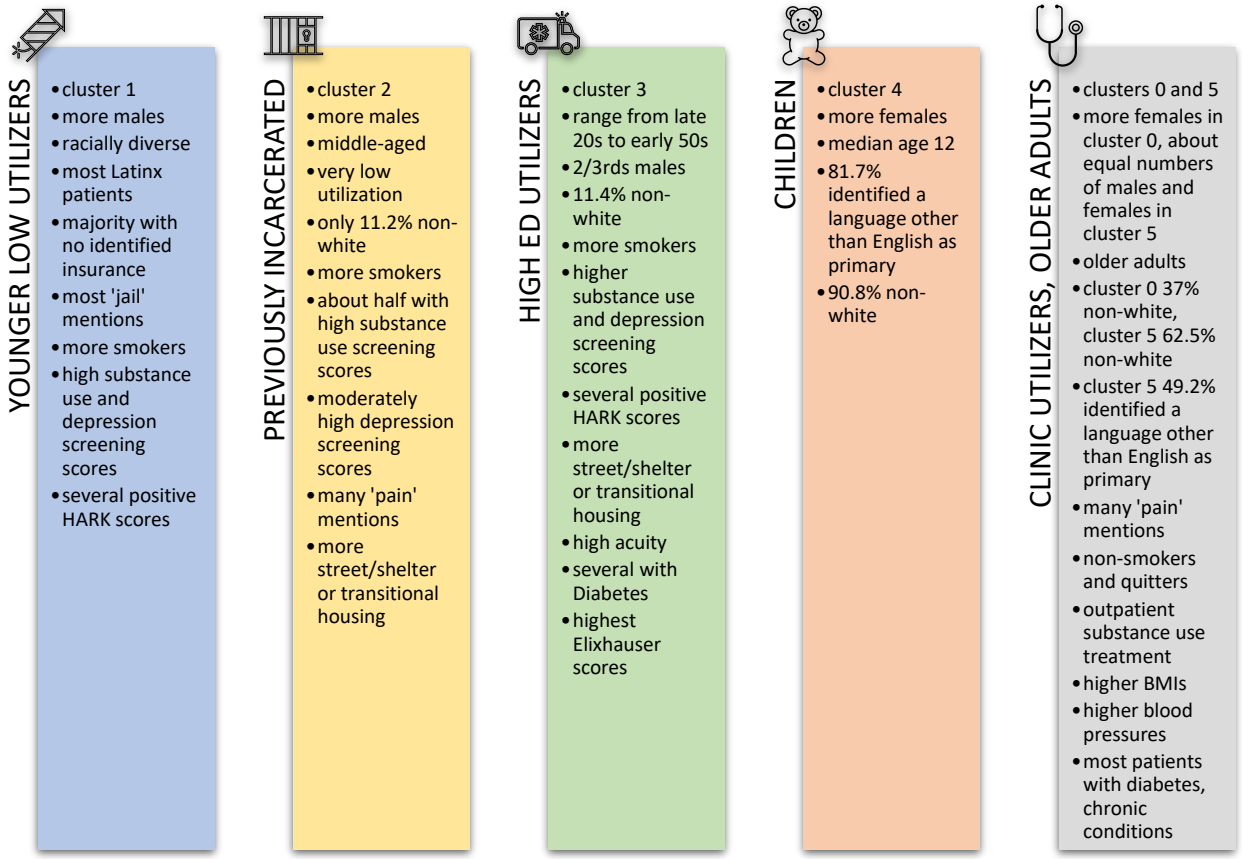


Figure 61 - Descriptive Summary of Service Groups based on Cluster Data

6.3 Shapley Values & Decision Tree Analysis

Ensemble decision tree-based methods, including gradient-boosted trees such as XGBoost (Chen & Guestrin, 2016), are powerful ways to use machine learning about features in a training set to classify withheld (test) data into categories. Classification methods are often used to make predictions, but they can also be used to explain categorizations such as unsupervised clustering. As discussed in chapter six, the Random Forrest and XGBoost classifiers were used to predict the cluster memberships of patients in the original data set, using two-thirds of the dataset to train the models, and one-third to test them. Parameters for each model were selected using Ski-kit Learn's (Pedregosa et al., 2011) GridSearchCV on data divided using StratifiedShuffleSplit with five partitions. For the tuned model, a fresh copy of the data was partitioned using StratifiedKFold, also with five partitions, on separately scaled training and testing data. Both models were evaluated using a confusion matrix of their performance on the test/holdout dataset, with a goal of 75% or better accuracy in classifying each patient into their cluster. Results for both classifiers hit the mark in every cluster except cluster zero, where both classifiers had a challenging time distinguishing some patients in cluster zero from patients in other clusters, particularly cluster one. In spite of these limitations, the Random Forrest classification model had 72% accuracy in predicting cluster zero membership, and the XGBoost model had 70% accuracy.

The concept of Shapley values was introduced to the field of game theory in 1953 by Lloyd Shapley, who was trying to produce a way to quantify the distribution of labor within cooperative enterprises or games. For every outcome that involves input by *n* entities, there is a breakdown of how much each contributor is responsible for the product. This concept can easily be applied to a linear regression model, for example, where the contribution of each covariate

could be represented by the covariate's weight multiplied by its value (Molnar, 2022, Section 9.5.1), however calculating the contribution of non-linear features to a tree-based model is not as straightforward. Shapley's idea was to estimate each entity's marginal contribution to an overall outcome by systematically excluding each contributing entity one at a time. This allows calculation of the weighted average of all marginal contributions – aka the Shapley value. The SHAP package (Lundberg & Lee, 2017) utilizes a special Shapley explaining algorithm specifically for use with tree ensemble models. This explainer is more consistent than feature importances and improves the feature attribution methods of previous Shapley explainers in the ensemble tree context by directly measuring local feature interaction effects, and better explaining global model structure using combinations of many local explanations of each classification (Lundberg, Erion & Lee, 2019).

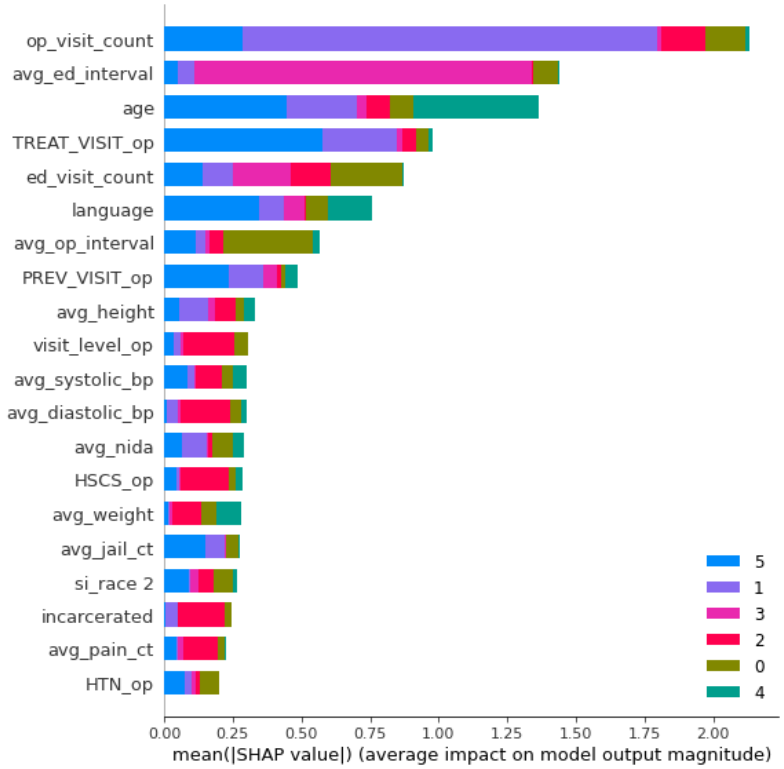


Figure 62: Feature Contributions by Shapley Value, Broken Down by Cluster Contribution

The SHAP summary chart (Figure 62) explains what features contributed most to the prediction of the clusters using the test (held out) data. Each feature's bar is broken down by its importance in classifying patients into each cluster. The most important feature XGBoost used in breaking up the patients into their clusters was the number of outpatient visits. Since the number of outpatient visits was spread across all the clusters (even though some clusters had patients with a larger proportion of them than others), this makes a lot of sense. The number of outpatient visits was particularly important in determining who belonged in cluster one. This group had very few outpatient visits, but this was important in assigning them to that cluster. The number of outpatient visits was also important in determining who belonged in clusters five, two, and zero. It had a minor impact on cluster four membership and no impact at all on cluster three membership.

The second most important feature was the average emergency department visit interval (avg_ed_interval), which was especially important in assigning patients to cluster three. Based on this information, we can see that while patients in cluster three had outpatient visits as well as emergency department visits, they were placed in cluster three based on the frequency of their emergency department visits. The fact that they also had outpatient visits appears to be incidental. That's an important detail when reflecting on feasible options to assist these patients in reducing their emergency utilization; if patients with high emergency utilization go to the clinic, it is possible to consider ways to engage them further in the outpatient setting.

Another important consideration is whether any of these characteristics or risk factors are modifiable. Interventions might be able to help patients control chronic conditions or stop smoking (secondary prevention) or avoid diagnoses like pneumonia altogether through regular vaccination (primary prevention). Variables such as high NIDA TAPS scores or PHQ-2 scores,

higher numbers of visits, or frequent visit intervals might be modifiable with the right tailored series of interventions or with better care coordination, but other factors – such as age – would not be.

Age played a significant role in placing people in clusters four (child patients) and five (older adult patients). It mattered little in placing patients in cluster three. The number of emergency department visits and average outpatient interval were the most significant features in assigning patients to cluster zero. For cluster four, it was age, language, and average weight. For cluster two, previous incarceration was an important classifying factor, along with pain count, outpatient visit level (visit_level_op), and the number of outpatient diagnoses (HSCS_op).

While ensemble methods make more accurate predictions by averaging and cross-validating across a large number of individual models, and Shapley values can shed some light on the true importance of features used by these models to divide the data, these models are limited in their explainability. The use of a single sci-kit learn DecisionTreeClassifier (Pedregosa et al., 2011) allows for a deeper analysis not only of the features used to predict each cluster but the specific values of the features used to separate the data set into categories. The same steps used to tune and train the ensemble classifiers were used to obtain the cross-validated results of a DecisionTreeClassifier on the test/withheld data. It was understood that the results would not be as accurate as those obtained using ensemble methods, however, the classifier did quite well predicting cluster memberships in the test data (Figure 76) considering the multi-class nature of the classification problem, and the difficulty ensemble methods already encountered separating cluster zero from the other clusters. The tuned classifier used the gini criterion, and the max-depth of the tree was set to nine, although the tree structure produced by the classifier had a maximum depth of 6 to 7 nodes.

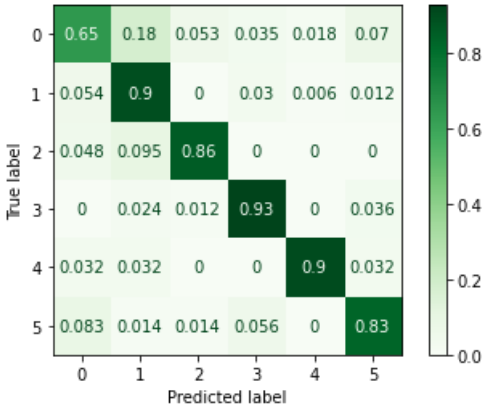


Figure 63: Confusion Matrix results of DecisionTreeClassifier, 'Spectral B' UMAP Clusters

The decision tree classifier did the best predicting the membership of clusters one, three, and four, an above-average job of predicting cluster two, an adequate job of predicting cluster five, and a fair job predicting cluster zero (as expected). Sci-kit Learn's tree library (Pedregosa et al., 2011) was used to produce a complete plot of the tree used by the classifier, and the node decision values were analyzed to help round out the picture of each cluster's characteristics.

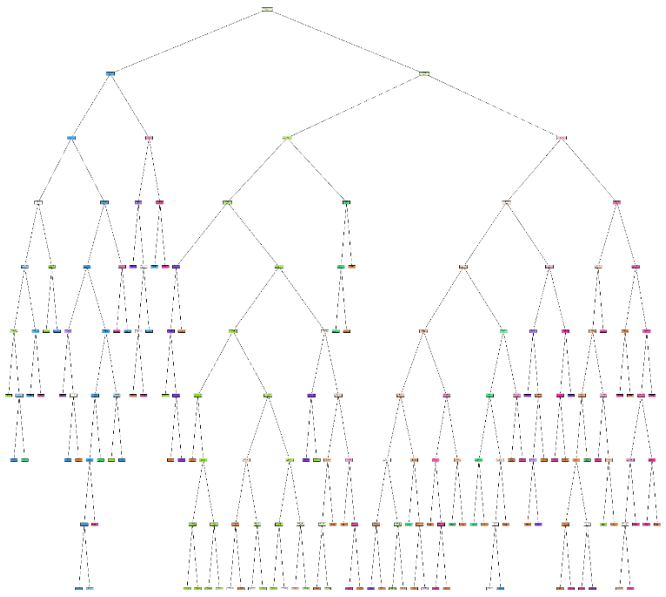


Figure 64: The Decision Tree Classifier's Output, Wide View

The overall tree can be divided into three regions – a left region, a central region, and a right region. The tree algorithm initially split the data based on the length of the average emergency department visit interval ("avg_ed_interval"). Those with an average interval value of fewer than 452 days were routed to the left side of the tree, and those with longer intervals toward the right. This right branch was then divided into two sub-trees (the central tree, and the right tree) based on the number of outpatient visits, with patients with 3.5 visits or fewer forming the central tree, and those with more than 3.5 visits forming the right-most tree.

The left tree then split patients based on their primary language, with those speaking primarily English going to the left sub-tree, and those speaking primarily another language to the right side. Those with higher acuity were regularly placed into clusters three and five, with those with more extreme weather day visits classified into group three, and those with more chronic condition diagnoses (AS/COPD, HTN) classified into group five. Younger people were divided based on age and height into groups one and four, since both clusters had patients of low acuity and few diagnosis codes. The classifier placed patients under age 43 with a moderate number of outpatient visits into group zero.

The central tree (consisting of patients with 3.5 outpatient visits or fewer, and an average ED interval of more than 452 days), was then split on previous incarceration history, with those with a history indicated by demographic data divided between clusters zero (higher BMI and average HARK scores) and two (majority of those previously incarcerated). Those without a history of incarceration indicated by the demographic data were split up into clusters zero or one, depending on acuity and visit counts, with those who were less acute and had fewer visits classified into cluster one. Those without incarceration history, but with higher 'jail' counts in their clinic notes, were placed in cluster one.

The right tree (consisting of patients with 3.5 outpatient visits or more, and an average ED interval of more than 452 days), was then divided into patients with 5.5 outpatient treatment visits (TREAT_VISIT_op) or fewer versus those with more than 5.5 visits. Patients with fewer outpatient visits were subsequently divided by primary language, and by the number of outpatient preventive visits (PREV_VISIT_op), with those with 2.5 preventive visits or fewer divided into clusters zero (moderate acuity, more outpatient visits), one (higher NIDA scores, low acuity), and two (previously incarcerated but less utilization). Those with more than 2.5 preventive visits were divided into clusters zero and five, with cluster five the higher acuity of the two. Across the board, the classifier placed patients with higher "disability" counts into cluster zero.

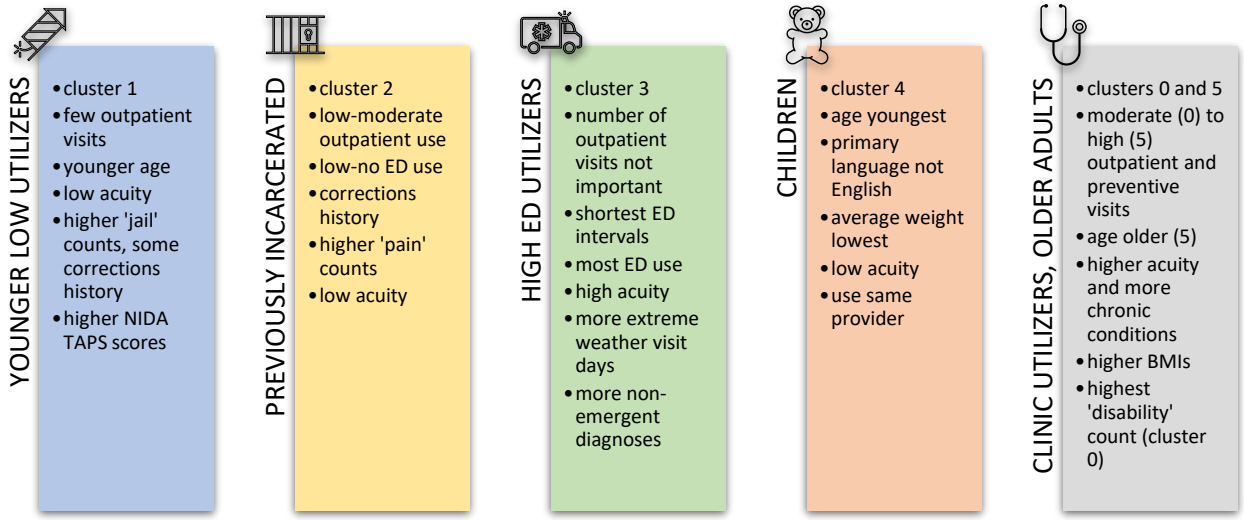


Figure 65 - Descriptive Summary of Service Groups based on Shapley Values and Results of A Decision Tree Classifier

6.4 Further Cluster Comparisons

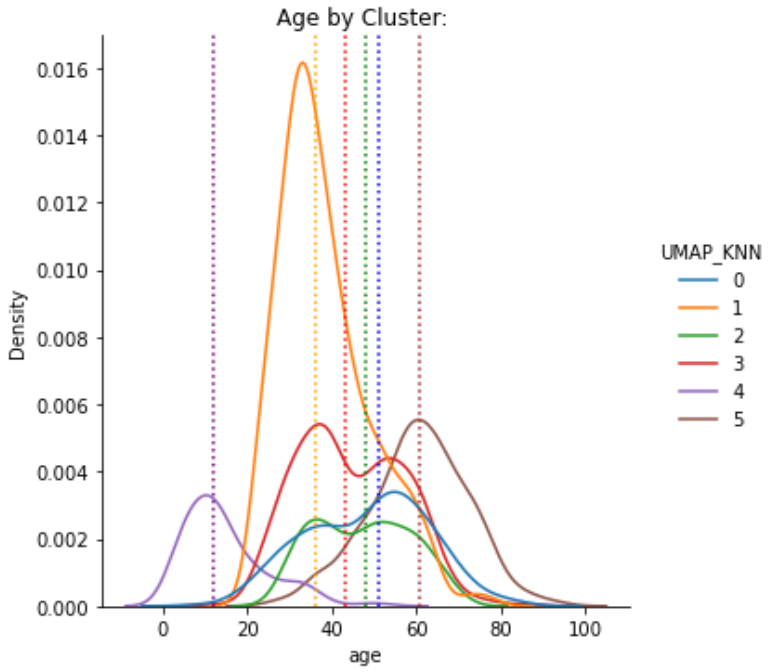


Figure 66: Cluster Comparison: Age Distributions

Following the clustering, it was possible to reexamine the data more efficiently to find differences between the patient groups. For example, age distributions and medians, shown in Figure 66, confirms that children are predominantly in cluster four, that younger adults (a larger population) are in cluster one, and that age distributions are similar in clusters two and three. Cluster zero patients were middle-aged, and cluster five had the majority of older adults.

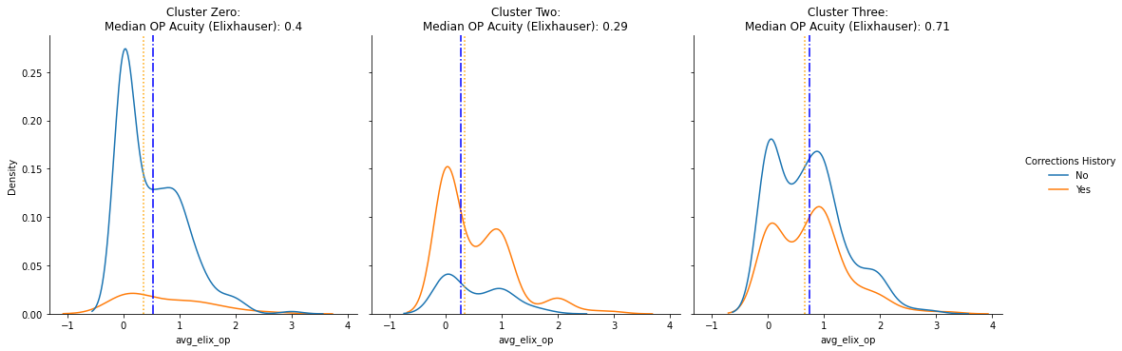


Figure 67: Clusters 0, 2, and 3: Outpatient Acuity by Corrections History

Comparing outpatient acuity using the average Elixhauser score, it was possible to break down the three clusters with the most previously-incarcerated patients and see that a longer tail of acuity exists in cluster two for patients with a corrections history (Figure 67). Outpatient acuity was used in this comparison because all three of these groups had many more outpatient than emergency visits. Clusters zero, one, and five had the majority of individuals who identified a language other than English as primary. Comparing their overall acuity within each cluster to the acuity of patients who were primarily English speakers shows that among the HCHM patient population, those who are primarily English speaking had higher overall acuity. This difference held in cluster five, where there were many older patients with chronic health conditions (Figure 68).

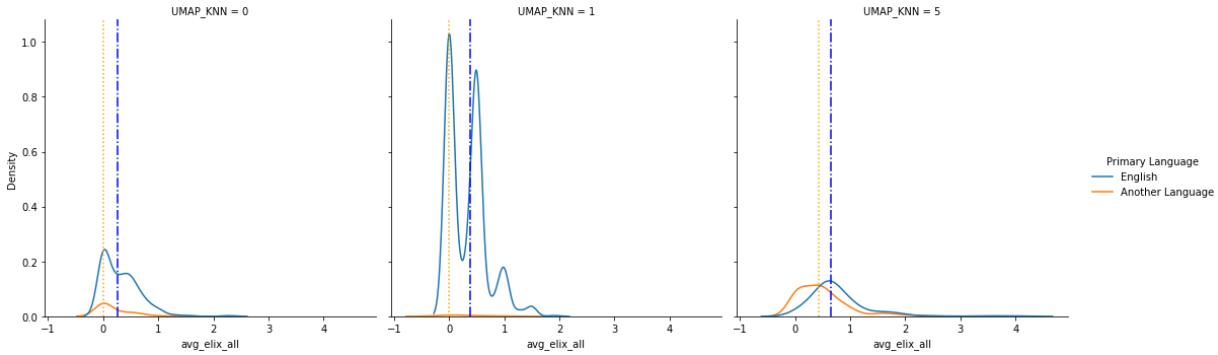


Figure 68: Clusters 0, 1, and 5: Total Acuity by Primary Language

Clusters one, two, and three had the largest numbers of patients with diagnosis codes for substance use and mental health problems. Comparing these three clusters' total acuity across the presence or absence of elevated NIDA TAPS scores (Figure 69) and the presence or absence of mental health struggles (Figure 70), a longer tail of acuity was visible for patients with substance use and mental health concerns in clusters one and three, while patients in cluster two had fewer mental health diagnoses and lower to moderate PHQ-2 scores. This difference was

most pronounced in cluster three, where the health impact of these conditions was highest. One thing to note, when comparing the age distributions of clusters one and two with acuity distributions – both cluster one and cluster two have a bi-modal distribution with regard to both age and acuity. This could indicate that the younger patients in both clusters have the least utilization and thus, lowest acuity, pulling median and mean acuity in these clusters lower.

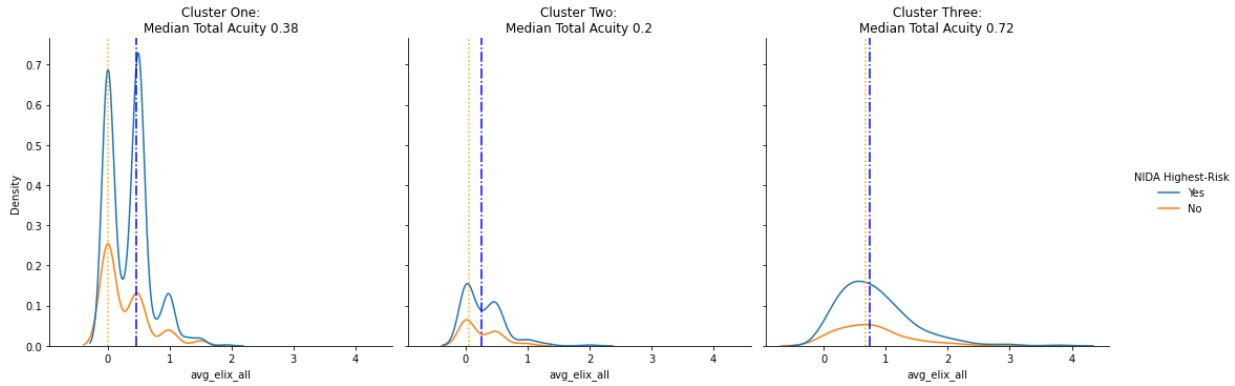


Figure 69: Clusters 1, 2, and 3: Total Acuity by Highest-Risk NIDA Score

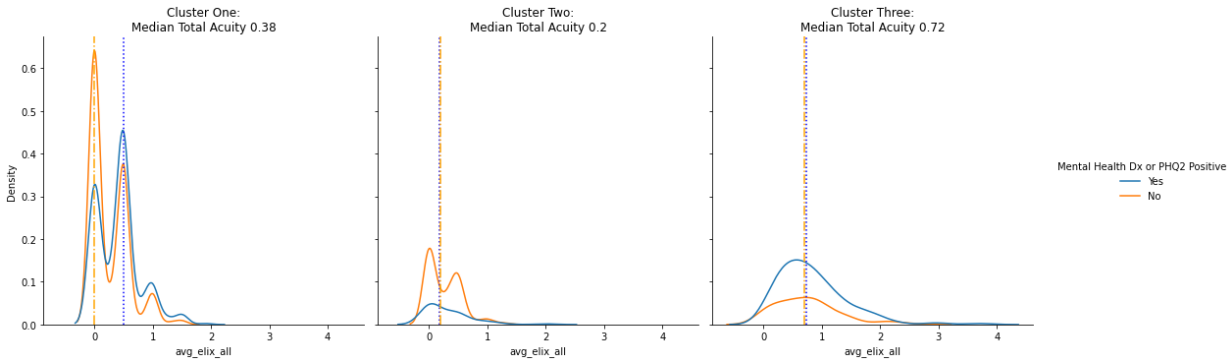


Figure 70: Clusters 1, 2, and 3: Total Acuity by Mental Health Diagnosis or Elevated PHQ-2

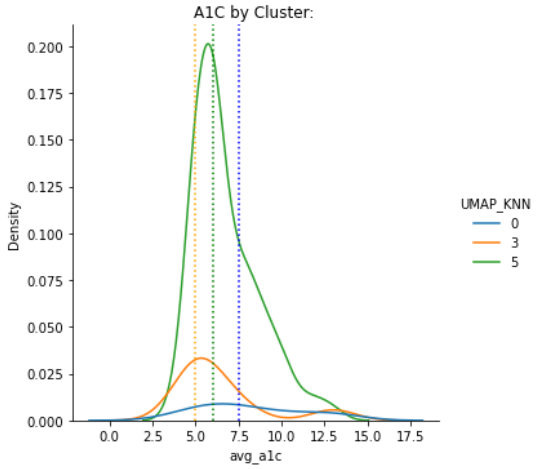


Figure 71: Clusters 0, 3 and 5: Average HgbA1C

Figure 72: Clusters 0, 3, and 5: Average Random Blood Glucose

Clusters zero, three, and five contained all the patients with diabetes. Looking at HgbA1C percentages and random office blood glucose measurements across patients in the three clusters, we see that the median values were highest for those in cluster zero. In spite of the much higher number of patients with diabetes in cluster five, those in cluster five had slightly better overall glucose control than those in the, on average, much younger cluster three. Cluster three patients with diabetes also appeared to have a bimodal distribution with regard to both their average

HgbA1C and random office glucose readings, with one crest appearing in the normal range, and the other in the uncontrolled range. By breaking down median average random glucose and HgbA1C measures for patients in cluster three who had versus did not have highest-risk NIDA scores (Figure 73) or mental health issues (Figure 74), the bimodality of the diabetes control distribution in cluster three patients was emphasized. Patients in cluster three with substance use and mental health challenges had blood sugar control levels falling into one of these two areas. While the median average HgbA1C for cluster three as a whole was only 5%, for those with highest-risk NIDA scores, the median average HgbA1C was 8%. Likewise, the median average random glucose for cluster three as a whole was 171 mg/dL, but for those with highest-risk NIDA scores, it was 209 mg/dL, and for those *without* highest-risk NIDA scores, the median average glucose reading dropped to 142 mg/dL. Here, data can back up what clinicians well know – for patients with severe mental health and substance use issues, control over both mental and physical health, particularly chronic conditions that require steady upkeep, is difficult.

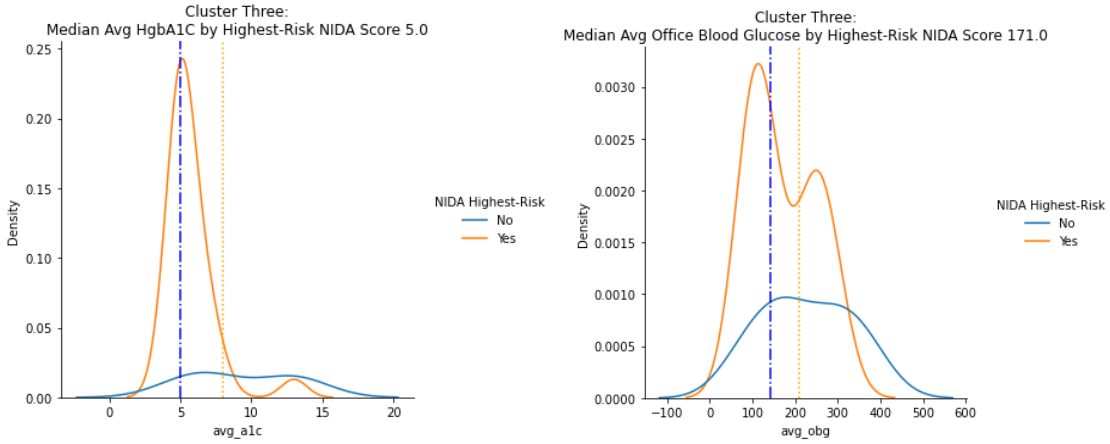


Figure 73: Cluster 3 Patients with Diabetes:

HgbA1C & Random Office Glucose by Highest-Risk NIDA Scores

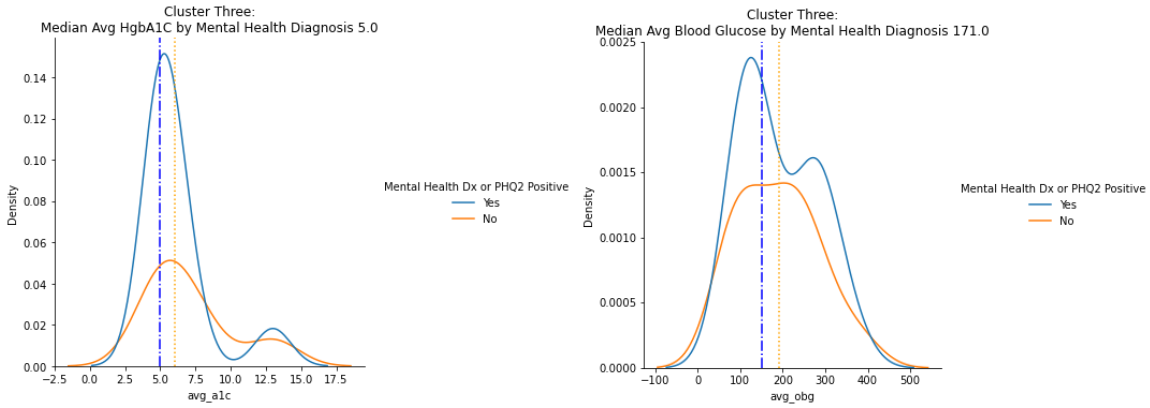


Figure 74: Cluster 3 Patients with Diabetes:

HgbA1C & Random Office Glucose by Mental Health Diagnosis or Elevated PHQ-2

Blood pressure control was also an issue across HCHM patient sub-populations, however just as clusters zero, three, and five had the majority of patients with diabetes, they also had the majority of patients with hypertension and hypertensive blood pressure readings (Figure 75). While cluster five patients had the highest median average pressures, clusters zero and three were not far behind. All three clusters' average systolic distributions had long, heavy tails; these were more pronounced in clusters three and five. The distributions of diastolic pressure were more clinically concerning, with a substantial proportion of patients from all three clusters showing average readings falling between 90 and 100 mm Hg.

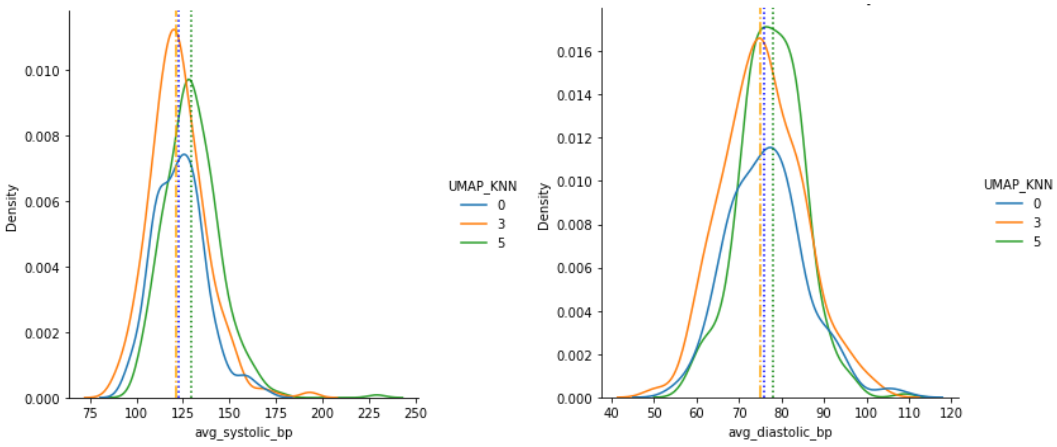


Figure 75: Clusters 0, 3, and 5: Median Average Systolic (left) and Diastolic (right) Blood Pressures

Patients in clusters three and five had the highest acuity levels, the highest numbers of visits, and the highest numbers of diagnosis codes across their visits. As discussed previously, the utilization patterns and primary condition categories of these two clusters differed dramatically.

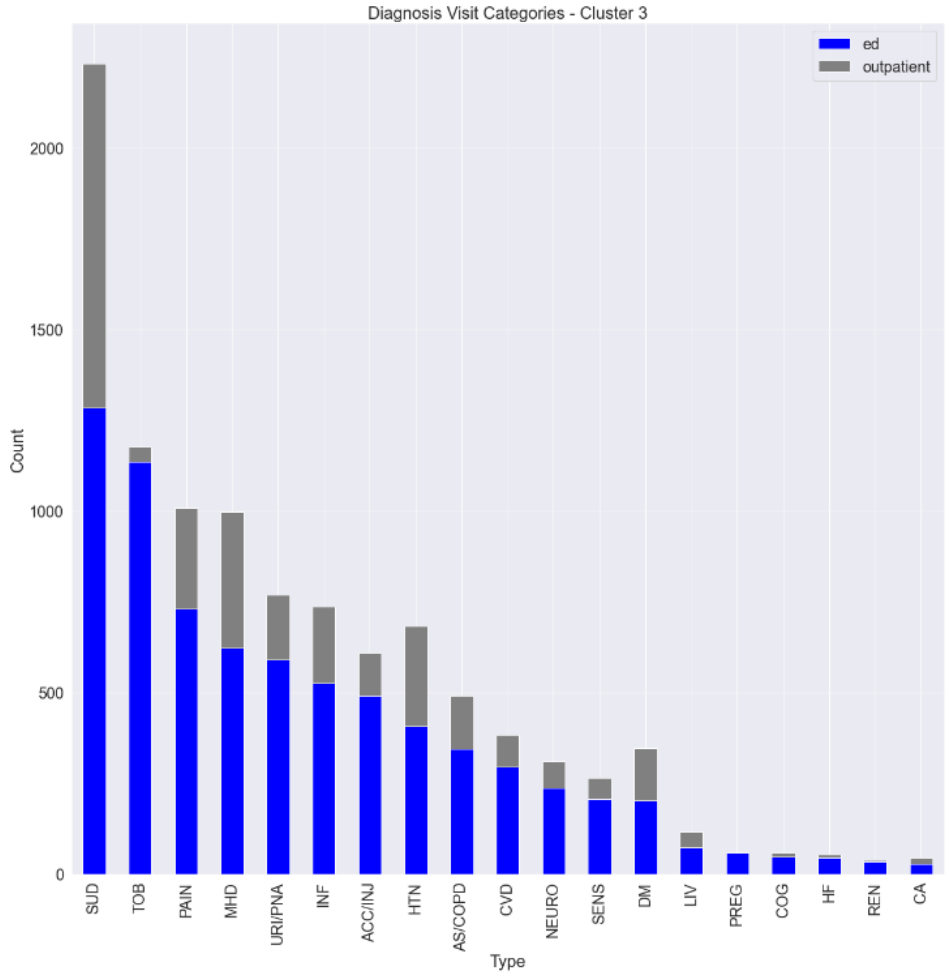


Figure 76: Cluster Three: Diagnoses by Category and Visit Type

Cluster three, the group with the highest emergency department utilizers, had more emergency department codes across all categories than outpatient codes, but their code counts for substance use visits were also elevated in the outpatient setting. Cluster three's top emergency department codes (excluding tobacco use-related codes) included alcohol dependence and intoxication, hypertension, depression and suicidal ideation, anxiety, and opioid and other psychoactive

substance abuse. Asthma, COPD exacerbation, urinary tract infections, and other infections - particularly respiratory and skin infections, including abscesses and cellulitis, were also common reasons for emergency department visits within this cluster (Table 11).

Code Example	Code Description (Similar codes have been grouped)	Total # ED Codes
F17210	Nicotine dependence	924
F1010	Alcohol-related	520
I10	Essential (primary) hypertension	357
F329	Major depressive disorder	296
F1910	Other psychoactive substance abuse	286
F419	Anxiety disorders	280
R079	Chest pain	277
R45851	Suicidal ideations	229
F1110	Opioid-related	181
R109	Abdominal pain	157
E119	Diabetes (all)	129

Table 11: Frequency of Top Emergency Codes: Highest ED Utilizers (Cluster 3)

In cluster five, outpatient visits dominated. While many cluster five patients suffered from mental health and substance use conditions like cluster three patients, they had high numbers of outpatient visits for these conditions and very few emergency visits. Cluster five patients had more chronic conditions, such as diabetes, hypertension, chronic obstructive pulmonary disease, and cardiovascular conditions. They also did a better job managing their conditions, with many patients in this cluster coming to the clinic frequently for blood pressure checks, diabetes visits, and mental health treatment. More patients in cluster five were former smokers and recent quitters, and a smaller proportion of these patients had elevated PHQ-2 and "highest-risk" NIDA scores. In spite of having many complaints of pain during outpatient, and even some emergency department visits, patients in cluster five seemed to be striving to improve their outcomes and were using the HCHM clinic as a support system for their health maintenance and well-being.

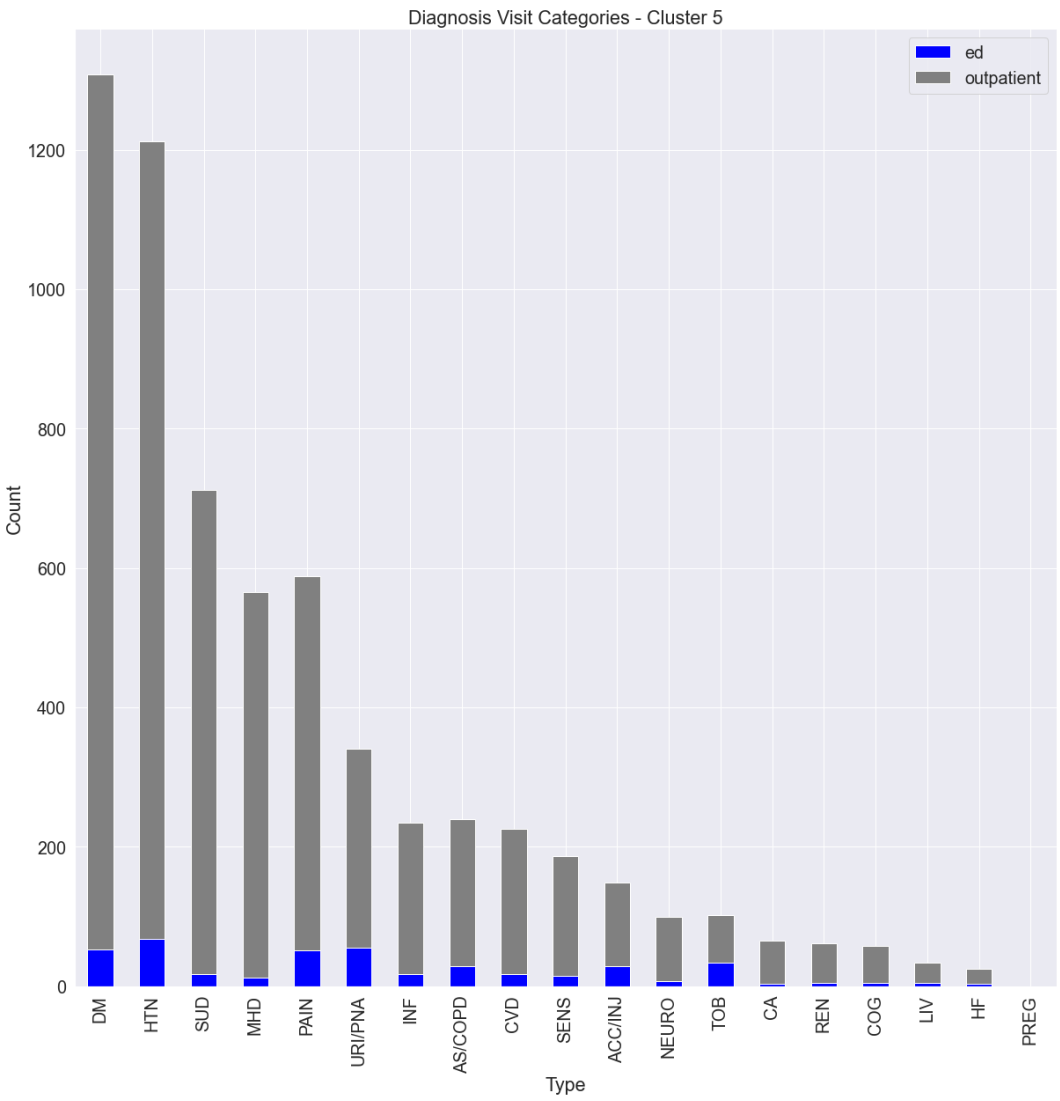


Figure 77: Cluster Five: Diagnoses by Category and Visit Type

Cluster Five		
Code Example	Code Description (Similar codes have been grouped)	Total # ED Codes
I10	Essential (primary) hypertension	65
S0591XA	Strains, sprains, cuts, and injuries	36
E119	Diabetes-related	34
F17210	Nicotine dependence	26
R05	Cough	10
R109	Abdominal pain	14
J449	Chronic obstructive pulmonary disease	13
R079	Chest pain	16
R0602	Shortness of breath	7
J45909	Unspecified asthma	7

Cluster Zero		
Code Example	Code Description (Similar codes have been grouped)	Total # ED Codes
F17210	Nicotine dependence	14
S0003XA	Strains, sprains, cuts, and injuries	14
R109	Abdominal pain (all)	9
Y09	Assaults and accidents	9
I10	Essential (primary) hypertension	8
M79604	Pain (various limbs)	6
Z79899	Other long term (current) drug therapy	5
R0789	Other chest pain	5
F10129	Alcohol-related	5
S0990XA	Unspecified injury of head, initial encounter	4

Table 12: Frequency of Top Emergency Codes: Moderate to High Clinic Utilizers (Clusters 5 and 0)

While clusters zero, one, two, and four did not have the elevated levels of acuity found in clusters three and five, there were good reasons to think that low or under-utilizers in some of these clusters – particularly clusters one and two – might become high utilizers in the future, if ways of better engaging them in health maintenance are not discovered. Each of these clusters had particular conditions that were challenging for them, some uniquely. For example, contusions, strains, sprains, bites, fractures, and other injuries were an issue for people in these clusters. Dental problems and mental health issues such as suicidal ideation were high on the list of emergency visit reasons for those in cluster one. Concussions and other head injuries were also common, as well as alcohol and substance-related diagnoses.

Cluster One		
Code Example	Code Description (Similar codes have been grouped)	Total # ED Codes
S6991XA	Contusions, strains, sprains, bites, injuries	38
F17210	Nicotine dependence	33
K0889	Dental problems	10
F1110	Opioid-related	9
K047	Cutaneous abscesses	9
Z79899	Other long term (current) drug therapy	8
F1010	Alcohol-related	8
J029	Acute pharyngitis or sinusitis	7
R112	Nausea and/or vomiting	7
Y042XXA	Assault or accidental injuries	7
R030	Hypertension or Elevated blood-pressure	6
F419	Anxiety disorders	5
R45851	Suicidal ideations	5

Cluster Two		
Code Example	Code Description (Similar codes have been grouped)	Total # ED Codes
F17210	Nicotine dependence	9
S602	Contusions, strains, sprains, bites, injuries	5
F10129	Alcohol-related	4
L03115	Cellulitis (all)	3
L02512	Cutaneous abscesses	3
S060X0A	Concussion, head injury	2
R55	Syncope and collapse	2

Table 13: Frequency of Top Emergency Codes: The Lowest Utilizers (Clusters 1 and 2)

Code Example	Code Description (Similar codes have been combined)	Total # ED Codes
S0083XA	Contusions, sprains, strains, injuries	11
R509	Fever, unspecified	7
R109	Abdominal pain	6
R05	Cough	5
O200	Threatened abortion	5
R51	Headache or migraine	5
Z7722	Contact with and (suspected) exposure to environmental tobacco smoke	4
R0789	Other chest pain	4
Z3801	Single liveborn infant, delivered by cesarean	3
O039	Spontaneous abortion	3

Table 14: Frequency of Top Emergency Codes: Children (Cluster 4)

CHAPTER 7: DISSCUSSION AND LIMITATIONS

7.1 Opportunities for Service Groups

There are several opportunities for care initiatives that might assist patients in these clinic sub-populations more comprehensively. The most important – and sometimes the most difficult – task providers face is creating durable, safe, and trusting relationships with patients so health care can be planned and executed collaboratively. In no setting is it more likely to be difficult to establish such bonds than in the provision of primary care to homeless people, many of whom have experienced trauma throughout their lives. My goal in this section is to summarize how some findings might be useful in tailoring future care for service groups. In the end, the best judge of what is right for patients is the patients themselves, and the providers who show up to care for them, and stand with them, every day.

7.1.1 High Emergency Department Utilizers

Dr. Gabor Maté, a physician specializing in the relationship between childhood trauma and both mental and physical health, defines addiction as "any behavior a person finds relief in, and therefore craves in the short-term, but suffers negative consequences from in the long-term, and does not give up despite the negative consequences" (Lee, 2022). He goes on to point out that, by this definition, addiction is neither a choice nor a disease, but something universal to human beings – a method of coping with emotional pain and life situations that seem impossible or undesirable to overcome another way. Most people know what it is like to cope with life events in a less-than-healthy way. However, for most people, their addiction is either less severe or more socially acceptable (video games, television watching, occasional overeating, mildly toxic

relationships etc.) than the addictive behaviors engaged in by people experiencing severe life consequences related to an alcohol or substance use disorder.

Important research into the long-term impacts of childhood trauma on neurological pathways in the brain reveals that childhood abuse, betrayal, and humiliation can set the stage for severe addiction and mental health problems later in life (Strathearn et al., 2019). Children naturally repress memories that they do not have words or safety to express, while also accepting self-definitions involving shame, self-blame, and the normalization of repetitive trauma and abuse (Miller, 2008). Growing up without parents or guardians who are able to handle their own emotions, life-circumstances, or addictions, children may – out of necessity – jump directly over the process of forming a positive self-image, empathy, and adaptive social behaviors, focusing instead on survival skills such as avoiding abuse while engaging in the care-taking of their abusers. These children survive to adulthood by virtue of their adaptations but habituate themselves to self-destructive self-definitions and relationship roles. Instead of being offered protection or therapy by witnesses to their difficulties, they often fall into at least some of the same maladaptive coping mechanisms many of their parents and caregivers engaged in, including the substitution of alcohol, drugs, sex, or food for the slow process of self-redefinition and healing necessary to find the motivation to engage in adaptive coping strategies. The achievement of a new self-definition as someone worthy of love and success may seem impossible and even inarticulable, especially for people whose trauma began before the formation of their self-concept and earliest memories.

Alcohol, drugs, and smoking can be attractive to people with a traumatic past for social reasons as well. Because the development of self-concepts and coping strategies necessary to the survival of a traumatic childhood lead to maladaptations in social behavior, people with such

pasts often immediately add loneliness and isolation to their list of problems. In adolescence, these people may quickly discover that asking for a light for a cigarette, drinking or engaging in risky physical or sexual behaviors, or becoming the sought-after source of illicit substances can all seem like straightforward ways to not only escape from negative emotions, but also create the desperately-desired illusion of being loved and cared for. While the expectations formed in childhood by those who grew up in abusive families often lead to disappointment or tragedy in friendships and intimate relationships, most people in our society lack the meta-language necessary to provide these people with the feedback they need to learn more adaptive social interaction strategies during the natural course of life. While they become repetitively isolated through no fault of their own, without professional intervention they may go on to more fully embrace self-destructive roles and self-definitions.

Since the time of Sigmund Freud, people interested in human psychology have been noting that individuals with a traumatic past tend to repetitively seek out trauma in their lives. Various theories exist as to why this occurs, but the phenomenon of "repetition compulsion" – the compulsive recreation of traumatic social and emotional situations by trauma survivors – remains scientifically understudied (van der Kolk, 1989). Most clinicians hypothesize that the compulsive, subconscious re-creation of these situations stems from a paradoxical combination of masochism (the ingrained belief that one's proper self-concept is negative and shameful) and the desire to re-create past negative social experiences – such as rejection by a parent – in the hope of resolving them differently. For those who do not seek help to heal from a traumatic past and re-define their self-concept, this behavior results in the reinforcement of negative self-beliefs and maladaptive social interactions, further entrenching their engagement in harmful coping strategies. Repetition compulsion may also serve as a way people create self-protection from the

deep reservoir of grief they harbor within themselves as a consequence of childhood mistreatment and neglect. By repetitively recreating situations that affirm these traumatic experiences as inevitable and "normal," at least for them, they reassert their roles and expectations as impossible to challenge or change. While this is self-destructive, it may also seem preferable to facing the profound overwhelm often associated with confronting both the need to heal and the bottomless courage required to begin and sustain the process (Miller, 2008).

The deep-seated synergistic impact of childhood trauma and its sequelae create a self-reinforcing prison for people that often leads to profoundly negative alterations in both mental and physical health. The widespread negative social and economic impacts of this individual tragedy writ large are illustrated by the data presented here; cluster three patients are living out the consequences of both an individual nightmare and a public health emergency. As overwhelmed as they likely are by the daily imperatives of their survival, those who wish to assist them may feel equally overwhelmed by the seeming complexity and nuance required to care for them. Both the Substance Abuse and Mental Health Services Administration (SAMHSA) and the National Health Care for the Homeless Council provide clinicians with excellent guidance that reflects the reality that assisting people recovering from a traumatic past, whether they suffer from overt addiction or mental health conditions or not, is often primarily a matter of encouraging their authenticity and making truthful identification of their desires and emotions less anxiety-provoking. Learning about and engaging in trauma-informed care and motivational interviewing provides a place to start, and return to, when engaging with people who may be unpredictable or struggle with self-efficacy (Bennett, 2016a).

SAMHSA also makes several important points about the process of addiction recovery (Substance Abuse and Mental Health Services Administration (SAMHSA), 2022a), and only one

of them is specifically about health. The others are about **home** – having a safe and stable dwelling place, **purpose** – having meaningful daily activities, and **community** – having relationships that provide support and hopefulness to the individual. All this might sound like a tall order, and in our current societal and political context, it is. However, the reality reflected in SAMHSA's framework is echoed by behavioral health clinicians everywhere – for recovery to be more appealing than addiction, it must a) seem possible, and b) be more supportive of a functional life than the addiction itself. To many people struggling with substance use disorders, recovery threatens their deeply-ingrained negative self-image and their existing social relationships, appearing isolating, purposeless, and less adaptive than the addiction itself. This perception is not only inaccurate but dangerous in a country where drug overdoses killed more than 100,000 people between May 2020 and April 2021 (Centers for Disease Control and Prevention, 2022). However, these realities reinforce the necessity of advocating for a supportive housing first approach to provide both shelter and social support to people struggling with addiction and homelessness, and the importance of harm reduction strategies such as the distribution of fentanyl test-strips and naloxone (SAMHSA, 2022b).

The data presented here on cluster three patients affirms previous findings that people with high-acuity mental health and substance use problems have difficulty managing their health issues, are at elevated risk for many comorbidities, and are unpredictable in their resource utilization patterns. The costs associated with providing people suffering from these conditions with the support necessary to allow them to choose recovery might seem high, but they are likely to be at least equal to the cost to families, society, and our institutions of allowing them to fend for themselves.

Actions assistive to the highest emergency utilization group may include:

- Continuing to focus on therapeutic relationship building, promoting self-acceptance and authenticity in all patient-provider interactions
- Augmenting the therapeutic relationship and advancing patient-centered care by tailoring care to both the primary care guidelines and patients' priorities
- Increasing care and case management services; considering the addition of dedicated care and case management for the highest-risk individuals
- Advocating for an increase in harm-reduction strategies and the wider availability of both temporary and permanent supportive housing
- Increasing self-service resource availability, such as recovery meeting lists and recovery literature
- Increasing access to self-service hygiene care
- Using standing order protocols to empower RNs to assist clients with common needs such as first-aid and infection prevention, obtaining commonly-needed medications, and obtaining and using blood glucose testing supplies
- Engaging in frequent community and/or street outreach efforts
- Establishing a telehealth connection with patients who have access to the internet via a mobile device
- Increasing the frequency of outpatient communication and follow-up appointments

7.1.2 Low Utilizers

Low utilizers (cluster one and two patients) may have many of the same problems as cluster three patients, while their levels of acuity and numbers of comorbidities are lower overall. Younger cluster one and two patients may be especially difficult to reach because they are still able to function and may be finding their existing coping strategies workable for now. Cluster two patients may also be avoiding institutional interactions due to deep discouragement brought on by struggles common to the previously incarcerated. In addition to health challenges, these patients commonly face widespread employment discrimination, and are ten times more likely to be homeless than the general public according to a recent Prison Policy Initiative report produced using Bureau of Justice Statistics' survey results (2018). These rates increase for those who have been incarcerated more than once, and for former inmates of color. While those who were

incarcerated face a higher probability of homelessness, those who are homeless also face a higher probability of incarceration, due to the frequent criminalization of behaviors that may be engaged in by homeless people such as sleeping in public places, public urination, and panhandling (Dupuy, Allen & Hernández, 2017). Those who offer housing, such as regional housing authorities or individual land-owners, often implement tenant screening criteria that increase housing insecurity in the previously incarcerated population. Being less likely to pass a credit or background check, people with a corrections history frequently end up living in rooming houses or hotels and motels, or doubled-up with family or friends, if they can find a housing situation at all (Couloute, 2018).

In a large study of care avoidance in the general population (Taber, Leyva & Persoskie, 2015), researchers found that people typically avoid interactions with the health care system due to combinations of factors including perceived cost, other priorities, lack of insurance, and the assumption that their symptoms will get better over time. Care avoidance among people experiencing homelessness may be explained by similar rationales but exacerbated by frequent negative encounters with the health care system and institutions in general. This is likely related to the social stigma attached to homelessness. There are also many reasons people experiencing homelessness might not go to a clinic for care that wouldn't necessarily occur to those who frequently care for them. Clinicians caring for homeless people do not expect such people to be well-groomed, however in a qualitative study of street sleepers in London, UK, researchers found that many homeless people reported being ashamed when they were not able to put themselves together, stating they might not go to a clinic even if they needed care because they were tired or unkempt, or because they would have to carry all their belongings with them when they went (Ungpakorn & Rae, 2019). In an effort to alleviate the isolation homeless people may

feel and the lack of access to care this may result in, many communities – including Manchester, NH – have put together mobile outreach programs, enabling clinicians to meet the homeless where they are. Ungpakorn & Rae (2019) also asked street sleepers what would make them more or less likely to engage with mobile clinical teams, and across the board, their respondents stated that "a relaxed and casual" approach with open body language was key to the possibility of establishing trust with health care workers. Other interviewees suggested avoiding partnerships with law enforcement, not waking people who are sleeping, and the importance of "seeing the same faces over and over again" to establishing relationships through frequent-enough contact.

Another attraction to a clinic for homeless people who tend to be low utilizers might be access to self-care services. In a new analysis in the BMJ, Hopkins & Narasimhan (2022) point out that having nowhere to call home decreases people's ability to engage in regular hygiene care or safely store self-care items. A place where homeless people could obtain access to items needed for self-care including running water, hygiene products, clean socks, frequently needed medications such as over the counter pain relievers, inhalers, and self-injectable contraception, as well as self-testing kits for pregnancy or illnesses such as HIV and HPV, would be helpful in developing relationships between health care workers and homeless people and allow homeless people more agency in caring for their health outside of a clinic or the emergency room.

In summary, actions assistive to low-utilization patients might include:

- Focusing on causal relationship-building more than health care administration
- Increasing the frequency and regularity of contact both in the clinics and with street outreach
- Advocating for harm-reduction
- Increasing opportunities for free dental assessments and care
- Increasing self-service hygiene and resource availability
- Increasing access to first-aid and commonly-needed medications

- Expanding telehealth for patients with access
- Increasing supportive programs for the previously incarcerated, connecting them with community, housing, and employment resources after release

7.1.3 Moderate and Higher-Acuity Clinic Utilizers

For patients in clusters zero and five, regular clinic visits were already a common occurrence. Many were engaging in regular appointments to help with the control of chronic conditions and participating in substance recovery and mental and behavioral health treatment programs. Integrated, trauma-informed care already happens at the clinics; however, variations in provider patterns show that some providers make more frequent follow-up appointments with patients, engaging them more frequently for health maintenance. These patients tend to participate in more follow-ups overall, and their glucose and blood pressure control is superior to that of other patients.

An area of concern for the clinic in general, but for higher-acuity patients in particular, is the high number of patients with poor control of their hypertension. Hypertension is a significant risk factor for heart attack and stroke and is poorly controlled across the U.S. population. According to the Centers for Disease Control and Prevention (2020), only about 1 in 4 adults (24%) with hypertension have their condition under control. Blood pressure control among homeless people is likely to be worse. After a retrospective chart review study of homeless New Yorkers, Asgary et al. (2016) concluded that approximately 40.1% of homeless patients had uncontrolled blood pressure ($p = .29$) and that 15.8% had stage 2 hypertension ($p = .27$). Blood pressure treatment in homeless adults is complicated by more than decreased access or reasons for care avoidance. Patients experiencing homelessness may be further limited in their ability to monitor treatment through regular self-checking of blood pressures, and in their ability to avail themselves safely of more aggressive medication regimens requiring shorter-acting medications

with frequent dosing schedules. Clinicians specializing in primary care for homeless people recommend limiting the use of diuretics due to concerns of dehydration, pursuing once-daily dosing wherever possible, and establishing more frequent follow-up appointments (Strehlow et al., 2009).

Other frequent drivers of emergency visits among higher-acuity patients were asthma and chronic obstructive pulmonary disease. Care for these common chronic respiratory diseases is extraordinarily difficult when patients are experiencing homelessness. Many environmental triggers may be present for these patients, including animal and insect droppings, pollen, mold, and smoke. These patients may also experience frequent exposure to contagious upper respiratory infections such as pneumonia, influenza, and even tuberculosis. Clinicians recommend frequent outreach to provide regular peak flow testing and immunizations, and to assess patients' access to and use of prescribed medications or inhalers. They also recommend frequent reassessment for symptoms of worsening allergies or burgeoning respiratory infections requiring further treatment or intervention (Gracy et al., 2018).

Another issue common across adult patient cluster groups was the presence of emergency department visits following prior hospitalizations or procedures. Top emergency department code lists for several of the patient clusters indicated that patients may be having difficulty caring for themselves following medical procedures, leading to readmissions or complications such as infections. A recent retrospective study of 232,373 general, vascular, and orthopedic surgeries occurring on homeless veterans in the Veterans Health Administration between 2008 to 2014 (Titan et al., 2018) found that hospital readmissions were higher in people experiencing homelessness who were discharged to the community following surgery. In addition to better care coordination and increased communication between primary care and inpatient clinical

teams, the National Health Care for the Homeless Council advocates for the establishment of Medical Respite Care for all homeless people (NHCHC, 2022c). Medical respite care provides temporary, longer-term shelter for people recovering from hospitalizations, surgeries and other acute illnesses, preventing discharge back to the street or a homeless shelter following acute care. Currently, thirty-eight out of fifty states have at least one medical respite care facility available for homeless people (NIMRC, 2021b). California has forty-one such facilities, while New Hampshire and Maine are among the twelve states with no such facilities. Neighboring Vermont has one such facility, and Massachusetts has three, including one of the earliest such facilities. Two recent studies conducted in Denmark demonstrated that Medical respite care for homeless people was cost-effective and prevented increased utilization (Bring et al., 2020), as well as providing homeless individuals with an environment where they could rest, reflect on their lives, and make plans for a better future (Pedersen et al., 2018).

A little more than half of the patients in cluster five identified a language other than English as their primary mode of communication. The majority of these patients are refugees from Asia and Africa, including Bhutan and the Democratic Republic of the Congo. Refugees in the community experience many barriers to health services access, including racism, xenophobia, reduced access to transportation, and a lack of available interpreters. Many refugees also experienced persecution or torture in their countries of origin, but few had mental health diagnoses or scored high on depression screenings. It is possible that some of these patients are reluctant to discuss difficulties with their mental health or seek out treatment for post-traumatic stress disorder related to their traumas. With face-to-face communication support and the development of a stable and trusting relationship with the HCHM care team, it is possible these patients may become more willing to receive treatment related to their experiences.

While I was not able to locate many patients struggling with disabilities through data mining, the most patients with 'disability' mentions in their clinic visit notes were in cluster zero. People facing functional challenges are a commonly-identified sub-population among homeless people (HUD, 2022) who may benefit from collaborative assessments and referrals that can improve their independence and assist them in accessing transportation, employment, health care, and other necessities.

Assistive actions for moderate and higher-acuity engaged outpatients might include:

- Increasing pre and post-hospitalization follow-up and care planning
- Advocacy for medical respite care in the State of New Hampshire
- Increasing the frequency of follow-up contacts and visits either in the community or at the clinics, using telehealth where available
- Increasing the availability of paid or volunteer translators for patients for whom English is not a primary language
- Continuing to assist refugee families with connections to additional refugee services
- Assessing patients with functional challenges for disability, and assisting them to access services that increase their health, independence, and quality of life

7.1.4 Children of Refugees

Cluster four consisted primarily of children of refugee families living in the Manchester area. Many of these children face similar difficulties to their parents in integrating into the larger English-speaking society and adapting to a new culture in the United States. Primary causes of emergency visits for these patients were similar to those of other children and adolescents around the nation. Emphasis on prevention and safety, as well as the development of good habits and the avoidance of unhealthy habits (sugary drinks, prolonged or early exposure to screens and social media, early pregnancy, smoking, and drug use), are just as appropriate for the primary care of these children as all others. The availability of face-to-face interpreters may be key to

establishing pathways of trust and communication between these children, their parents, and the care team. Better quality communication may enable the care team to better assess patients' needs and provide them, and their families, with supportive interventions and education.

Assistive actions for children of refugees might include:

- Increasing the availability of translators
- Establishing ongoing relationships and trust with children, parents, and the refugee community
- Providing education about safety, prevention, and risks associated with adolescence
- Collaborating with schools and after school programs to help build trauma-informed care teams and community programs promoting literacy and social support for these children

7.2 Limitations

7.2.1 Data and Features

The presence of missing data, as well as the use of the MICE imputer on some data in preparation for machine learning, were limitations. Some data also contained far-outlying values that skewed some of the statistical ranges of some of the patient clusters, particularly when a given cluster had a small number of patients meeting given criteria (e.g., patients within the cluster with HARK assessments, or HgbA1C values). In spite of these limitations, the data were able to produce an interpretable clustering result, identifying clinic sub-populations, albeit imperfectly due to the complexity and non-linearity of the developed feature set. As with all retrospective studies, the analysis necessarily represents a snapshot of the clinic's patients, frozen in time. It does not tell us what happened to those patients before or after the sample period.

7.2.2 Bias and Ethics

Selection bias was a factor here, impacting the generalizability of findings from this sample of the HCHM clinic's population to a) other HCH clinic populations in the nation, and b) homeless

people in general. Individuals were removed from this study who had no outpatient visits (n=37). These patients, who all had more emergency visits, were likely to have been sicker than those who were included in the study who all had at least one interaction with the primary care team. Their removal creates a situation where the sample as a whole might appear healthier or have more outpatient and less emergency utilization than other people experiencing homelessness in the nation. Another factor limiting generalizability is the differences between the patient sub-populations served by the HCHM clinic and those served by other HCH clinics around the nation. In other primary care clinics in the nation serving patients who are homeless or housing insecure, common additional service groups include parenting youth, unaccompanied minors, veterans, and people with HIV/AIDS (United States Department of Housing and Urban Development (HUD), 2021).

Unfortunately, all research contains bias because all researchers have bias. To every extent possible, I worked to identify my biases towards homeless people, those who might struggle with mental health or addiction, and other socially-reinforced sources of discriminatory thoughts and attitudes. I spent a lot of time reading and studying current thinking on the needs and challenges of homeless people, both within and outside of the health care system, and listening to interviews with and reading memoirs of individuals who struggled with homelessness, childhood trauma and addiction to gain further insight into their experiences. In conducting this research and writing about it, I was also forced to consider the impact of early childhood experiences on my life and my family of origin. Throughout this project, I made every effort to keep the dignity and privacy of the individuals about whom this research was conducted in the forefront of my mind. I presented aggregated data to reduce the identifiability of individuals within the data set, although I tried to balance this concern with advocacy for disaggregation practices intended to

improve the visibility of minorities within research presentations (Schwabish & Feng, 2021). While I completed a thorough descriptive analysis early in the research project and engaged clinic partners to collaborate on decisions around modeling choices and cluster set optimization, I purposely did not ask staff to teach me how to identify patient sub-groups prior to creating and tuning the unsupervised clustering algorithms. I wanted the unsupervised algorithm to find the patients in as unbiased a manner as possible, and then subject both the model and the cluster groups to exposition using the algorithmic tools at my disposal.

7.2.3 Methods

Limitations of the dimension reduction and clustering methods used here are about the complexity of the data set, and the "fuzzy" algorithmic processes used to imperfectly divide it into clusters. Both the UMAP and Spectral clustering algorithms involve some processes designed to optimize the preservation of the relationships between complex and non-linear features, but while they work well, they are not always able to divide the data set perfectly. Repurposing of the Bray-Curtis dissimilarity provided the best possible separation between the clusters, but there were always a few patients in each cluster who should have been placed in another cluster with respect to age, comorbidities, or other characteristics. Classification methods used to predict the clusters operated on coded features but were unable to utilize features such as housing status and highest completed education, due to the presence of many NULL values.

CONCLUSION

In this project, I used my health care domain knowledge, descriptive statistics, and machine learning in cooperation with academic and clinical partners to conduct an exploratory real-world evidence study capable of describing the characteristics of distinct Health Care for the Homeless of Manchester, NH (HCHM) patient sub-populations, identifying data-driven and evidence-based opportunities for care improvement. I placed this work in the larger context of the City of Manchester and the many complex challenges of providing effective, patient-centered primary care to people experiencing homelessness. I necessarily looked upon this work as an analyst, applying scrutiny to feature creation and algorithmic selection and tuning, but also as a nurse, acquainted with and deeply concerned about the compassionate care of every patient. It was my privilege to gain a deeper understanding of these 2,265 individuals and to become better acquainted with the thoughtful, high-quality, collaborative care provided to them by the Health Care for the Homeless team.

I hope that the collaborative partnership established by this work between the University of New Hampshire Health Data Science program and the Health Care for the Homeless and Catholic Medical Center community health partnership will continue into the future, and undertake a re-evaluation of clinic patient groups, their risk factors, and care experiences in the challenging context of the COVID-19 pandemic, comparing those results to this baseline. There are also exciting opportunities to evaluate longitudinal data to more accurately identify specific risk factors and characteristics associated with changes in health status in the population over time, and to introduce possible “interventions” via a multistate model to examine predicted changes in patient outcomes.

REFERENCES

Agency for Toxic Substances and Disease Registry (ATSDR), (2018). Social Vulnerability Index, CDC SVI Documentation 2018. https://www.atsdr.cdc.gov/placeandhealth/svi/documentation/SVI_documentation_2018.html

Aldridge, R.W., Story, A., Hwang, S. W., Nordentoft, M., Luchenski, S. A., Hartwell, G., Tweed, E. J., Lewer, D., Vittal Katikireddi, S., & Hayward, A. C. (2018). Morbidity and mortality in homeless individuals, prisoners, sex workers, and individuals with substance use disorders in high-income countries: a systematic review and meta-analysis. *The Lancet (British Edition)*, 391(10117), 241–250. https://doi.org/10.1016/S0140-6736(17)31869-X

American Association of Professional Coders (AAPC), (2021). What is ICD-10? https://www.aapc.com/icd-10/

American College of Surgeons (2020). Office E/M Coding Changes: Be prepared for 2021…learn how to navigate new CPT guidelines. https://www.facs.org/-/media/files/advocacy/practice-management/2020_emcoding.ashx#page=4

Asgary, R., Sckell, B., Alcabes, A., Naderi, R., Schoenthaler, A., & Ogedegbe, G. (2016). Rates and Predictors of Uncontrolled Hypertension Among Hypertensive Homeless Adults Using New York City Shelter-Based Clinics. *Annals of Family Medicine*, 14(1), 41–46. https://doi-org.unh.idm.oclc.org/10.1370/afm.1882

Baer, H., Singer, M. & Susser, I. (2003). *Medical Anthropology and the World System; 2nd edition*. Westport, CT: Greenwood.

Bassett, J. (2022, February 2). Invisible Walls: The Amoskeag Company legacy. *Manchester Ink Link*; https://manchesterinklink.com/invisible-walls-the-amoskeag-company-legacy/

Bassett, J. (2022, February 3). Invisible Walls: The Role of Redlining. *Manchester Ink Link*; https://manchesterinklink.com/invisible-walls-the-role-of-redlining/

BatchGeo LLC. (2022). BatchGeo Mapping Tool. https://batchgeo.com/

Bennett, M. (2016, October 5). Webinar: Being Trauma Informed and Its Role in Ending Homelessness. National Health Care for the Homeless Council. https://nhchc.org/online-courses/trauma-informed-care-webinar-series/being-trauma-informed-and-its-role-in-ending-homelessness/

Bennett, M. (2016, October 26). Webinar: The Abyss: Addiction, Homelessness and Trauma. National Health Care for the Homeless Council. https://nhchc.org/online-courses/trauma-informed-care-webinar-series/the-abyss-addiction-homelessness-and-trauma/

Bobbitt, Z. (2021, March 13). Bray-Curtis Dissimilarity: Definition & Examples. *Statology*. https://www.statology.org/bray-curtis-dissimilarity/

Brien, A., So, M., Ma, C. & Berner, L. (2019). *Homelessness & Adverse Childhood Experiences: The health and behavioral health consequences of childhood trauma*. National Health Care for the Homeless Council and National Network to End Family Homelessness, http://www.nhchc.org/aces

Bring, C., Kruse, M., Ankarfeldt, M. Z., Brünés, N., Pedersen, M., Petersen, J., & Andersen, O. (2020). Post-hospital medical respite care for homeless people in Denmark: a randomized controlled trial

and cost-utility analysis. *BMC Health Services Research*, 20(1), 1–11. https://doi-org.unh.idm.oclc.org/10.1186/s12913-020-05358-4

Buelt, A., Richards, A., & Jones, A. L. (2021). Hypertension: New guidelines from the international society of hypertension. *American Family Physician*, 103(12), 763-765.

Bunnell, R., Ryan, J. & Kent, C. (2021). Toward a New Strategic Public Health Science for Policy, Practice, Impact, and Health Equity. *American Journal of Public Health*. 111. e1-e8. 10.2105/AJPH.2021.306355.

Catholic Medical Center, (2022). Community Health: Health Care for the Homeless. https://www.catholicmedicalcenter.org/care-and-treatment/community-health/health-care-for-the-homeless

Centers for Disease Control and Prevention (CDC), (2017). Extreme Heat; National Disasters and Extreme Weather. https://www.cdc.gov/disasters/extremeheat/heat_guide.html

Centers for Disease Control and Prevention (CDC), (2019). PLACES: Local Data for Better Health. https://www.cdc.gov/places/

Centers for Disease Control and Prevention (CDC) (2020, February 25) Facts About Hypertension. Retrieved from https://www.cdc.gov/bloodpressure/facts.htm

Centers for Disease Control and Prevention (CDC), (2022). Provisional Drug Overdose Death Counts. Vital Statistics Rapid Release, National Center for Health Statistics. https://www.cdc.gov/nchs/nvss/vsrr/drug-overdose-data.htm

Chu, Y.-T., Ng, Y.-Y., & Wu, S.-C. (2010). Comparison of different comorbidity measures for use with administrative data in predicting short- and long-term mortality. *BMC Health Services Research*, 10(1), p.140–140. https://doi.org/10.1186/1472-6963-10-140

City Builder, (2022). Manchester, NH; Census Tracts. https://www.citivelocity.com/citybuilder/eppublic/cb/us/cities/2718

City of Manchester, (2022). My Manchester: Kalivas Union. https://www.mymanchesternh.com/Neighborhoods/Kalivas-Union

City of Manchester Health Department, (2021, August 10). Press Release: The Manchester Emergency Operations Center Reminds Residents to Prepare for a Period of Extended Heat and Humidity. https://www.manchesternh.gov/Portals/2/Departments/health/2021-08-10_HEAT_ALERT.PDF?ver=2021-08-10-124839-000

Colburn, G., (2014). The federal commitment to homelessness prevention: a silver lining of the economic crisis. *Poverty Public Policy*, 6(1), 33–45.

Community Solutions (2022, January 16). Mayors Play a Part in Ending Homelessness [video]. YouTube; *Community Solutions channel*. https://youtu.be/-Tozc10hmCg

Couloute, L. (2018). Nowhere to Go: Homelessness among formerly incarcerated people. *Prison Policy Institute*. https://www.prisonpolicy.org/reports/housing.html

Das, S., (2020, August 31). ISOMap. *Sanjoy Das, YouTube Channel*. https://youtu.be/bi-xB9OywlM

Dotson P. (2013). CPT® Codes: What Are They, Why Are They Necessary, and How Are They Developed? *Advances in Wound Care*, 2(10), 583–587. https://doi.org/10.1089/wound.2013.0483

Dupuy D., Allen, T., & Hernández, K.L. (2017). Policing the Houseless 2.0, Arrests by the LAPD 2012-2017. [Infographic]. *Million Dollar Hoods Research Team*. http://milliondollarhoods.org/wp-content/uploads/2017/10/Policing-the-House-2.0.FINAL_.pdf

Edidin, J. P., Ganim, Z., Hunter, S. J., & Karnik, N. S. (2012). The mental and physical health of homeless youth: A literature review. *Child Psychiatry and Human Development*, 43(3), 354-375. doi:http://dx.doi.org/10.1007/s10578-011-0270-1

Edwards, A. (2019). Computing Solutions for Homelessness. *Quest Magazine*, California State University at Long Beach. https://sites.csulb.edu/sites/quest/2019/computing-solutions-for-homelessness/

Felitti, V.J., Anda, R.F., Nordenberg, D., Williamson, D. F., Spitz, A.M., et al. (1998). Relationship of Childhood Abuse and Household Dysfunction to Many of the Leading Causes of Death in Adults; The Adverse Childhood Experiences (ACE) Study. *American Journal of Preventative Medicine*; 1998;14(4), p. 245-258. https://doi.org/10.1016/S0749-3797(98)00017-8

Forrester, J. (1969). *Urban Dynamics*. Waltham, MA: Pegasus Communications. ISBN: 978-1883823399.

Forrester, J. (1971). Counterintuitive Behavior of Social Systems, *Technology Review*. The Massachusetts Institute of Technology.

Fortin, Y., Crispo, J. A. G., Cohen, D., McNair, D. S., Mattison, D. R., & Krewski, D. (2017). External validation and comparison of two variants of the Elixhauser comorbidity measures for all-cause mortality. *PloS One*, 12(3), e0174379–e0174379. https://doi.org/10.1371/journal.pone.0174379

Fowler, P. J., Hovmand, P. S., Marcal, K. E., & Das, S. (2019). Solving Homelessness from a Complex Systems Perspective: Insights for Prevention Responses. *Annual Review of Public Health*, 40(1), 465–486. https://doi.org/10.1146/annurev-publhealth-040617-013553

Goetsch, M.R., Tumarkin, E., Blumenthal, R.S. & Whelton, S.P. (2021). New Guidance on Blood Pressure Management in Low-Risk Adults with Stage 1 Hypertension. *American College of Cardiology, Education and Meetings*. https://www.acc.org/latest-in-cardiology/articles/2021/06/21/13/05/new-guidance-on-bp-management-in-low-risk-adults-with-stage-1-htn

Good, B. (1994). *Medicine, Rationality, and Experience: An Anthropological Perspective*. Cambridge, MA: Cambridge University Press.

Gracy, D., Carlson, J., King, C., Oberg, C., Newport, S., Sherman, P., & Strehlow, A., (2018). Adapting your practice: Assessment and treatment of people with asthma who are experiencing homelessness. Adlparvar, F. (Ed.). Nashville: Health Care for the Homeless Clinicians' Network, National Health Care for the Homeless Council, Inc. https://nhchc.org/wp-content/uploads/2019/08/2018-asthma-guidelines.pdf

Greenacre, M. (2008). Chapter 5: Measures of distance between samples: non-Euclidean. STAT254: Correspondence Analysis and Related Methods, *Stanford University*, Fall 2008. http://84.89.132.1/~michael/stanford/maeb5.pdf

Harvard T. H. Chan School of Public Health, (2021, February 24). The Forum: Homelessness in America: The Search for Solutions During COVID-19. [Discussion] https://theforum.sph.harvard.edu/events/homelessness-in-america/

Health Care for the Homeless (HCH) Clinicians' Network, (2017). Preventive Care for People Experiencing Homelessness. Healing Hands; 21(2). https://nhchc.org/wp-content/uploads/2019/08/healing-hands-preventative-care-finalized.pdf

Health Center Partners (2022). The Community Health Center Movement. https://hcpsocal.org/the-community-health-center-movement/

Homeless Hub, (2022). Who are homeless people? *Canadian Observatory on Homelessness*. https://www.homelesshub.ca/resource/who-are-homeless-people

Hopkins, J., & Narasimhan, M. (2022). Access to self-care interventions can improve health outcomes for people experiencing homelessness. *BMJ (Clinical research ed.)*, 376, e068700. https://doi.org/10.1136/bmj-2021-068700

Hwang, S.W. & Henderson, M.J., (2010). Health Care Utilization in Homeless People: Translating Research into Policy and Practice. *Agency for Healthcare Research and Quality, Working Paper No. 10002*, October 2010, http://gold.ahrq.gov

Ihler, A. (2003). Nonlinear Manifold Learning; 6.454 Summary. *Laboratory for Information and Decision Systems (LIDS)*, Massachusetts Institute of Technology. http://www.mit.edu/~6.454/www_fall_2003/ihler/summary.pdf

Kertesz, S. & Roncarati, J. (2021, November 10). IHH Research Seminar - Research to measure what matters in primary care for persons experiencing homelessness: what's the point, really? *Harvard T.H. Chan School of Public Health*. https://harvard.zoom.us/rec/share/7iTmZbd6XnCOtFvKRwKFqUyJ08hTqE6qTClso6nbJ-gPPx74Bbpk-MX3muNmsO96.XlogAtrhHRzYUBbP

Kroenke, K., Spitzer, R.L., Williams, J.B. (2003). The Patient Health Questionnaire-2: validity of a two-item depression screener. *Medical Care*, 41:1284–92.

Larimer, M.E., Malone, D.K., Garner, M.D., et al. (2013). Health Care and Public Service Use and Costs Before and After Provision of Housing for Chronically Homeless Persons With Severe Alcohol Problems. *JAMA.* 2009;301(13):1349–1357. doi:10.1001/jama.2009.414

Lewer, D., Aldridge, R. W., Menezes, D., Sawyer, C., Zaninotto, P., Dedicoat, M., Ahmed, I., Luchenski, S., Hayward, A., & Story, A. (2019). Health-related quality of life and prevalence of six chronic diseases in homeless and housed people: a cross-sectional study in London and Birmingham, England. *BMJ Open*, 9(4), e025192–e025192. https://doi.org/10.1136/bmjopen-2018-025192

Lee, S. (2022, April 5). Beyond Drugs: The Universal Experience of Addiction. *Gabor Maté*. https://drgabormate.com/opioids-universal-experience-addiction/

McInnes et al., (2018). UMAP: Uniform Manifold Approximation and Projection. *Journal of Open Source Software*, 3(29), 861, https://doi.org/10.21105/joss.00861

McNeely J., Wu, L., Subramaniam, G., Sharma, G., Cathers, L.A., Svikis, D., et al. (2016). Performance of the Tobacco, Alcohol, Prescription Medication, and Other Substance Use (TAPS) Tool for Substance Use Screening in Primary Care Patients. *Annals of Internal Medicine*, 165, p. 690-699. doi: 10.7326/M16-0317

Medlow, S., Klineberg, E. & Steinbeck, K. (2014). The health diagnoses of homeless adolescents: A systematic review of the literature. *Journal of Adolescence*; 37(5), p. 531-542. https://doi.org/10.1016/j.adolescence.2014.04.003

Miller, A. (2008). *The drama of the gifted child: the search for the true self*. 30th anniversary ed., rev. and updated, Hardcover ed. New York: Basic Books/Perseus Books Group.

Molnar, C. (2022). *Interpretable Machine Learning, A Guide for Making Black Box Models Explainable; Second Edition*. Victoria, BC: LeanPub. https://christophm.github.io/interpretable-ml-book/

Moses, J. & Janosko, J. (2018). Demographic Data Project: Part II, Gender and Individual Homelessness. *Homelessness Research Institute*. https://endhomelessness.org/wp-content/uploads/2019/09/DDP-Gender-brief-09272019-byline-single-pages.pdf

Nakazawa, D. J. (2016). *Childhood Disrupted: How Your Biography Becomes Your Biology, and How You Can Heal*. New York: Atria Books.

Nanjo, A., Evans, H., Direk, K., Hayward, A. C., Story, A., & Banerjee, A. (2020). Prevalence, incidence, and outcomes across cardiovascular diseases in homeless individuals using national linked electronic health records. *European Heart Journal*, 41(41), 4011–4020. https://doi.org/10.1093/eurheartj/ehaa795

National Alliance to End Homelessness (2019, March 18). What Housing First Really Means. [Blog]. https://endhomelessness.org/blog/what-housing-first-really-means/

National Health Care for the Homeless Council (NHCHC), (2022). Health Care for the Homeless. Who We Are. https://nhchc.org/who-we-are/

National Health Care for the Homeless Council (NHCHC), (2022). Executive Summary: Health Care for the Homeless, A Vision of Health for All. https://nhchc.org/understanding-homelessness/health-care-for-the-homeless-a-vision-of-health-for-all/executive-summary/

National Health Care for the Homeless Council (NHCHC), (2022). Medical Respite Care. https://nhchc.org/clinical-practice/medical-respite-care/

National Health Care for the Homeless Council (NHCHC), (2021). 2021 Federal Policy Priorities for the HCH Community. https://nhchc.org/wp-content/uploads/2021/02/2021-Policy-Priorities.pdf

National Health Care for the Homeless Council (NHCHC) and National Network to End Family Homelessness (NNEFH), (2019). Homelessness & Adverse Childhood Experiences: The health and behavioral health consequences of childhood trauma (Authors: Avery Brien, Program Manager NNEFH; Marvin So, Co-Chair, NNEFH; Christine Ma, Pediatrician, NNEFH; Lauryn Berner, Project Manager, NHCHC) Available at: http://www.nhchc.org/aces

National Institute for Medical Respite Care (NIMRC), (2021). *Medical Respite Literature Review: An Update on the Evidence for Medical Respite Care*. National Health Care for the Homeless Council (NHCHC). https://nimrc.org/wp-content/uploads/2021/08/NIMRC_Medical-Respite-Literature-Review.pdf

National Institute for Medical Respite Care (NIMRC), (2021). Medical Respite Care Directory. https://nimrc.org/medical-respite-directory/

National Institute on Drug Abuse (NIDA), (2021, April 13). Part 1: The Connection Between Substance Use Disorders and Mental Illness. Retrieved from https://nida.nih.gov/publications/research-reports/common-comorbidities-substance-use-disorders/part-1-connection-between-substance-use-disorders-mental-illness

National Weather Service (NWS), (n.d.). Frost and Freeze Information. https://www.weather.gov/iwx/fallfrostinfo

Nelson, R.K., Winling, L., Marciano, R., Connolly, N., et al. (1937). Manchester, NH – Section D3. *Mapping Inequality; American Panorama*. https://dsl.richmond.edu/panorama/redlining/#loc=13/42.983/-71.492&city=manchester-nh&area=D3&adview=full

New Hampshire Department of Labor (2022). Minimum Wage, *RSA 279*. https://www.nh.gov/labor/inspection/wage-hour/minimum-wage.htm

Oskolkov, N. (2019, December 31). Why UMAP is Superior over tSNE, Does initialization really matter? *Towards Data Science*, Medium. https://towardsdatascience.com/why-umap-is-superior-over-tsne-faa039c28e99

Oskolkov, N. (2020, March 3). tSNE vs. UMAP: Global Structure; Why preservation of global structure is important. *Towards Data Science*, Medium. https://towardsdatascience.com/tsne-vs-umap-global-structure-4d8045acba17

Pedersen, M., Bring, C., Brünés, N., Andersen, O., Petersen, J., & Jarden, M. (2018). Homeless people's experiences of medical respite care following acute hospitalization in Denmark. *Health & social care in the community*, 26(4), 538–546. https://doi.org/10.1111/hsc.12550

Powis, B., Griffiths, P., Gossop, M. & Strang, J. (1996). The differences between male and female drug users: community samples of heroin and cocaine users compared. *Substance Use & Misuse*, 31(5), p.529-543.

Quan, H., Sundararajan, V., Halfon, P., Fong, A., Burnand, B., Luthi, J. C., Saunders, L. D., Beck, C. A., Feasby, T. E., & Ghali, W. A. (2005). Coding algorithms for defining comorbidities in ICD-9-CM and ICD-10 administrative data. *Medical Care*, 43(11), 1130–1139. https://doi.org/10.1097/01.mlr.0000182534.19832.83

Romaszko, J., Cymes, I., Dragańska, E., Kuchta, R. & Glińska-Lewczuk, K. (2017). Mortality among the homeless: Causes and meteorological relationships. *PLOS ONE* 12(12): e0189938. https://doi.org/10.1371/journal.pone.0189938

Schwabish, J. & Feng, A. (2021). *Do No Harm Guide: Applying Equity Awareness in Data Visualization*. Washington DC: The Urban Institute. https://www.urban.org/sites/default/files/publication/104296/do-no-harm-guide.pdf

Seelos, C. (2021) Homelessness, a System Perspective. The Journey of Community Solutions. Stanford Center on Philanthropy and Civil Society, *PACS working paper* GIIL002/2021; Version: 23AUG21 https://pacscenter.stanford.edu/wp-content/uploads/2021/04/Homelessness_A-System-Perspective_Seelos_GIIL_002_2021.pdf

Shelton, K.H., Taylor, P.J., Bonner, A. & van den Bree, M. (2009). Risk factors for homelessness: evidence from a population-based study. *Psychiatric Services*, 60(4), p.465-72.

Singer, M. & Clair, S. (2003). Syndemics and Public Health: Reconceptualizing Disease in Bio-Social Context. *Medical Anthropology Quarterly*, 17(4); p. 423-441. https://doi-org.unh.idm.oclc.org/10.1525/maq.2003.17.4.423

Singer, M. (1996). A Dose of Drugs, a Touch of Violence, a Case of AIDS: Conceptualizing the SAVA Syndemic. *Free Inquiry in Creative Sociology*, 24(2), p. 99-110.

Sohal, H., Eldridge, S. & Feder, G. (2007). The sensitivity and specificity of four questions (HARK) to identify intimate partner violence: a diagnostic accuracy study in general practice. *BMC Family Practice*, 8:49. doi: 10.1186/1471-2296-8-49. PMID: 17727730; PMCID: PMC2034562.

Starmer, J. (2022). UMAP: Mathematical Details (clearly explained!!!). YouTube, *StatQuest with Josh Starmer*. https://youtu.be/jth4kEvJ3P8

Strang, G. (2019, May 16). 35. Finding Clusters in Graphs [video]. Matrix Methods in Data Analysis, Signal Processing, and Machine Learning; Spring 2018. YouTube, *MIT OpenCourseWare*. https://youtu.be/cxTmmasBiC8

Strathearn, L., Mertens, C. E., Mayes, L., Rutherford, H., Rajhans, P., Xu, G., Potenza, M. N., & Kim, S. (2019). Pathways Relating the Neurobiology of Attachment to Drug Addiction. *Frontiers in Psychiatry*, 10, 737. https://doi.org/10.3389/fpsyt.2019.00737

Strehlow A, Robertshaw D, Louison A, Lopez M, Colangelo B, Silver K, Post P. *Adapting Your Practice: Treatment and Recommendations for Homeless Patients with Hypertension, Hyperlipidemia & Heart Failure*, 53 pages. Nashville: Health Care for the Homeless Clinicians' Network, National Health Care for the Homeless Council, Inc., 2009 https://nhchc.org/wp-content/uploads/2019/08/CardioDiseases.pdf

Stroh, D. P. (2013, July 12). A Systems Approach to Ending Homelessness [video]. YouTube, *Bridgeway Partners*. https://youtu.be/khFnWMZGcuM

Substance Abuse and Mental Health Services Administration (SAMHSA), (2022, April 4). Recovery and Recovery Support. https://www.samhsa.gov/find-help/recovery

Substance Abuse and Mental Health Services Administration (SAMHSA), (2022, April 6). Harm Reduction. https://www.samhsa.gov/find-help/harm-reduction

Substance Abuse and Mental Health Services Administration (SAMHSA), (2011). Current Statistics on the Prevalence and Characteristics of People Experiencing Homelessness in the United States. https://www.samhsa.gov/sites/default/files/programs_campaigns/homelessness_programs_resources/hrc-factsheet-current-statistics-prevalence-characteristics-homelessness.pdf

System Dynamics Society (2022). *Origin of System Dynamics*. https://systemdynamics.org/origin-of-system-dynamics/

Taber, J. M., Leyva, B., & Persoskie, A. (2015). Why do people avoid medical care? A qualitative study using national data. Journal of general internal medicine, 30(3), 290–297. https://doi.org/10.1007/s11606-014-3089-1

Tannis, C. & Rajupet, S. (2021). Differences in disease prevalence among homeless and non-homeless veterans at an urban VA hospital. *Chronic Illness*, 0(0), p.1-10. https://doi.org/10.1177/17423953211023959

Tenenbaum, J. B., de Silva, V., & Langford, J. C. (2000). A global geometric framework for nonlinear dimensionality reduction. *Science* (New York, N.Y.), 290(5500), 2319–2323. https://doi.org/10.1126/science.290.5500.2319

Titan, A., Graham, L., Rosen, A., Itani, K., Copeland, L. A., Mull, H. J., Burns, E., Richman, J., Kertesz, S., Wahl, T., Morris, M., Whittle, J., Telford, G., Wilson, M., & Hawn, M. (2018). Homeless Status, Post discharge Health Care Utilization, and Readmission After Surgery. *Medical care*, 56(6), 460–469. https://doi.org/10.1097/MLR.0000000000000915

Trick, W. E., Rachman, F., Hinami, K., Hill, J. C., Conover, C., Diep, L., Gordon, H. S., Kho, A., Meltzer, D. O., Shah, R. C., Stellon, E., Thangaraj, P., & Toepfer, P. S. (2021). Variability in comorbidites and health services use across homeless typologies: multicenter data linkage between healthcare and homeless systems. *BMC Public Health*, 21(1), 1–9. https://doi-org.unh.idm.oclc.org/10.1186/s12889-021-10958-8

Trosset, M.W. & Buyukbas, G. (2021). Rehabilitating Isomap: Euclidean Representation of Geodesic Structure. *arXiv*, stat, machine learning, Cornell University. https://doi.org/10.48550/arXiv.2006.10858

Tsai, A. C., & Venkataramani, A. S. (2016). Syndemics and Health Disparities: A Methodological Note. *AIDS and Behavior*, 20(2), 423–430. https://doi.org/10.1007/s10461-015-1260-2

Ungpakorn, R., & Rae, B. (2020). Health-related street outreach: Exploring the perceptions of homeless people with experience of sleeping rough. *Journal of advanced nursing*, 76(1), 253–263. https://doi.org/10.1111/jan.14225

United States Interagency Council on Homelessness (USICH), (2018, August 15). *Deploy Housing First Systemwide*. https://www.usich.gov/solutions/housing/housing-first/

United States Census Bureau (2020). *Quick Facts: Manchester, NH*. https://www.census.gov/quickfacts/fact/table/manchestercitynewhampshire/PST045221

United States Department of Health and Human Services (DHHS), (2015). Substance Abuse and Mental Health Services Administration (SAMHSA) Treatment Improvement Protocol (TIP) 57, *Trauma-Informed Care in Behavioral Health Services*; April 2015. https://store.samhsa.gov/sites/default/files/d7/priv/sma14-4816_litreview.pdf

United States Department of Housing and Urban Development (HUD), (2021). HUD 2021 Continuum of Care Homeless Assistance Programs Homeless Populations and Subpopulations. https://files.hudexchange.info/reports/published/CoC_PopSub_NatlTerrDC_2021.pdf

United States Department of Housing and Urban Development (HUD), (2022). AHAR Reports. https://www.hudexchange.info/homelessness-assistance/ahar/#2021-reports

Urban Institute, (2000). *A New Look at Homelessness in America*; February 2000. Available at: http://www.urban.org

van der Kolk, B.A. (1989). The Compulsion to Repeat the Trauma: Re-enactment, Revictimization and Masochism. *Psychiatric Clinics of North America*, 12(2), 389-411. http://www.cirp.org/library/psych/vanderkolk/

van Walraven, C., Austin, P. C., Jennings, A., Quan, H., & Forster, A. J. (2009). A modification of the Elixhauser comorbidity measures into a point system for hospital death using administrative data. *Medical Care*, 47(6), 626–633. https://doi.org/10.1097/MLR.0b013e31819432e5

Wellenius, G.A., Eliot, M.N., Bush, K.F., Holt, D., Lincoln, R.A., Smith, A.E. & Gold, J. (2017). Heat-related morbidity and mortality in New England: Evidence for local policy. *Environmental Research*, 156, 845-853. doi: 10.1016/j.envres.2017.02.005. Epub 2017 May 9. PMID: 28499499. https://doi.org/10.1016/j.envres.2017.02.005

APPENDIX A: CODE

Algorithm, Code, Language and Library References

Bray, J.R. & Curtis, J.T. (1957). An Ordination of the Upland Forest Communities of Southern Wisconsin. *Ecological Monographs*, 27(4), p. 325-349. https://doi.org/10.2307/1942268

Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. *In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794). New York, NY, USA: ACM. https://doi.org/10.1145/2939672.2939785

Chung, F.R.K. (1995). Eigenvalues of Graphs. In: Chatterji, S.D. (eds) *Proceedings of the International Congress of Mathematicians*. Birkhäuser, Basel. https://doi.org/10.1007/978-3-0348-9078-6_128 Retrieved from https://mathweb.ucsd.edu/~fan/wp/eigenval.pdf

Ciortan, M. (2019. January 1). Spectral graph clustering and optimal number of clusters estimation – an overview of spectral graph clustering and a python implementation of the eigengap heuristic. *Towards Data Science*, Medium. https://towardsdatascience.com/spectral-graph-clustering-and-optimal-number-of-clusters-estimation-32704189afbe

Grobler, J. (2022). Manifold Learning Methods on a Severed Sphere. https://notebooks.gesis.org/binder/jupyter/user/scikit-learn-scikit-learn-giyvdmm0/lab/tree/notebooks/auto_examples/manifold/plot_manifold_sphere.ipynb

Hagberg, A., Swart, P., & S Chult, D. (2008). Exploring network structure, dynamics, and function using NetworkX.

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., … Oliphant, T. E. (2020). Array programming with NumPy. *Nature*, 585, 357–362. https://doi.org/10.1038/s41586-020-2649-2

Hunter, J. D., (2007). Matplotlib: A 2D Graphics Environment, *Computing in Science & Engineering*, 9(3), p. 90-95.

Lundberg, S.M., Erion, G.G. & Lee, S-.I. (2019). Consistent Individualized Feature Attribution for Tree Ensembles. *arXiv*:1802.03888v3 [cs.LG]. https://arxiv.org/abs/1802.03888

Lundberg, S.M. & Lee, S-.I. (2017). A Unified Approach to Interpreting Model Predictions. *31st Conference on Neural Information Processing Systems* (NIPS 2017), Long Beach, CA, USA. https://proceedings.neurips.cc/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf

McInnes, L., Healy, J. & Melville, J. (2020). UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *ArXiv e-prints*, 1802.03426 (stat.ML). https://arxiv.org/pdf/1802.03426.pdf

McInnes, L., Healy, J., Saul, N. & Grossberger, L. (2018). UMAP: Uniform Manifold Approximation and Projection. *Journal of Open Source Software*, 3(29), p. 861.

Meila, M. & Shi, J. (2001). A Random Walks View of Spectral Segmentation. In Thomas S. Richardson 0001, Tommi Jaakkola eds, Proceedings of the Eighth International Workshop on Artificial Intelligence and Statistics, AISTATS 2001, Key West, Florida, US, January 4-7, 2001. *Society for Artificial Intelligence and Statistics*. Retrieved from http://proceedings.mlr.press/r3/meila01a/meila01a.pdf

Pedregosa et al., (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12, pp. 2825-2830.

R Core Team, (2022). *R: A Language and Environment for Statistical Computing, Reference Index*. Vienna, Austria: R Foundation for Statistical Computing.

Reback, J., et al. (The pandas Development Team), (2022). pandas-dev/pandas: Pandas 1.4.2 (v1.4.2). *Zenodo*. https://doi.org/10.5281/zenodo.6408044

Shi. J. & Malik, J. (2000). "Normalized Cuts and Image Segmentation." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):888–905.

Stewart, G.W. & Sun, J.-G. (1990). *Matrix Perturbation Theory*, Academic Press, New York.

van Buuren, S. & Groothuis-Oudshoorn, K. (2011). "mice: Multivariate Imputation by Chained Equations in R." *Journal of Statistical Software*, 45: 1-67.

van der Maaten, L. & Hinton, G., (2008). Visualizing Data Using t-SNE. *Journal of Machine Learning Research*, 9(86), p. 2579−2605. https://jmlr.org/papers/v9/vandermaaten08a.html

Van Rossum, G., & Drake, F. L. (2009). *Python 3 Reference Manual*. Scotts Valley, CA: CreateSpace.

Virtanen, P. et al. (The SciPy Development Team). (2020) SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17(3), 261-272.

Von Luxburg, U. (2007). A Tutorial on Spectral Clustering. *Statistics and Computing*, 17(4), p. 395–416.

Waskom, M. L., (2021). seaborn: statistical data visualization. *Journal of Open Source Software*, 6(60), 3021, https://doi.org/10.21105/joss.03021

Wasey, J. O. (2020). *icd: Fast comorbidities from ICD-9 and ICD-10 codes, decoding, manipulation and validation*. R package version 4.0.9 https://rdocumentation.org/packages/icd/versions/4.0.9

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T., Miller, E., Bache, S., Müller, K., Ooms, J., Robinson, D., Seidel, D., Spinu, V., … Yutani, H. (2019). Welcome to the tidyverse. *Journal of Open Source Software*, 4(43), 1686.

Code Access

To protect patient privacy and confidentiality, Python and R code for this project resides in a private repository on my GitHub:

https://github.com/krashr-ds/thesis

(If you try to use this link and do not have access to view the private repository, you will get a 404 (page not found) message.) Stakeholders who would like to review the code are encouraged to do so. Please email me your GitHub user information so I can grant you access to the repository. If you don't have a GitHub user account and need to sign up, please visit https://github.com/signup

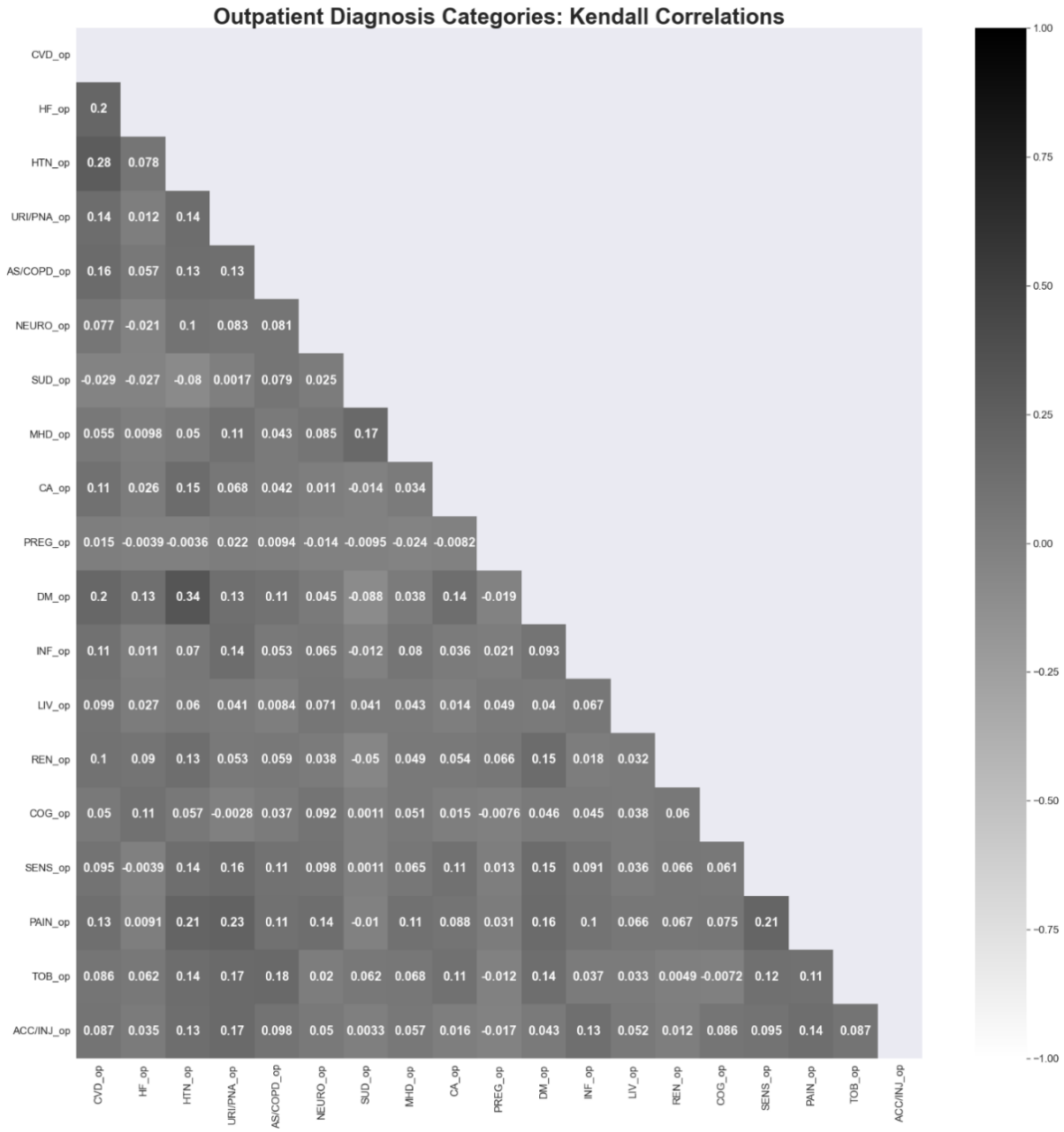
Contact

Please feel free to contact me with any questions.

Kyle Rasku – email: kylie.rasku@unh.edu
GitHub: https://github.com/krashr-ds
LinkedIn: https://www.linkedin.com/in/krashr-ds/

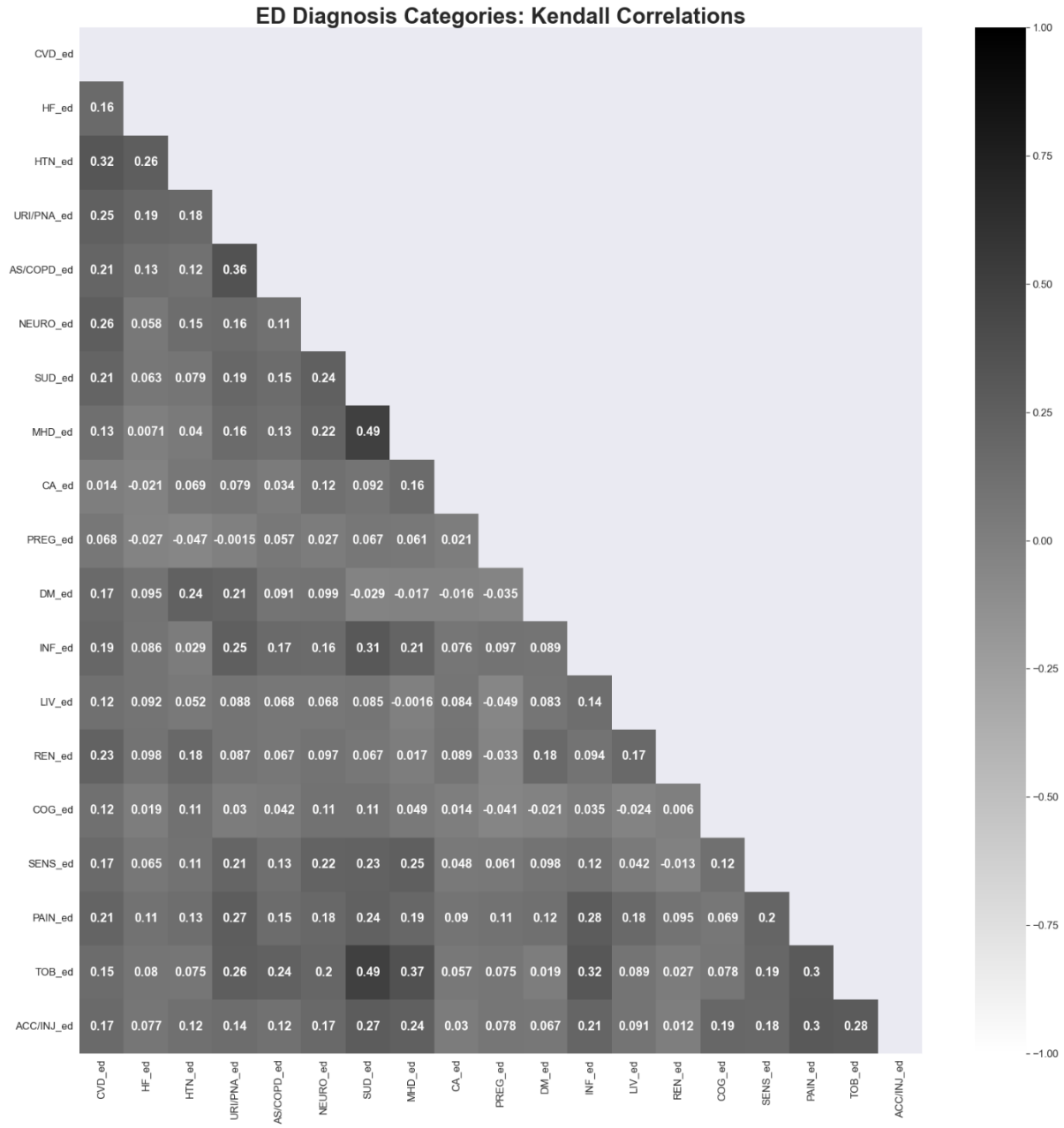
APPENDIX B: CORRELATION MATRICES

Section 1: Visit Reason Categories

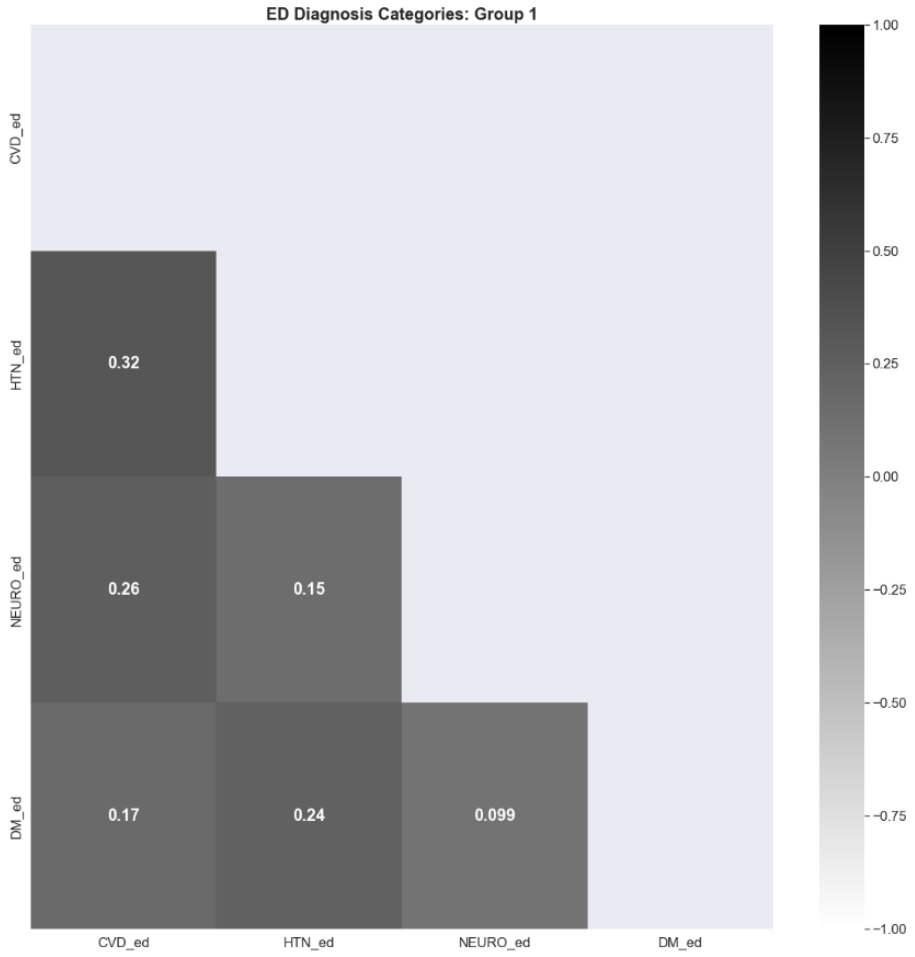


S1, Fig 1: Kendall Correlation: Outpatient Visit Diagnosis Groups

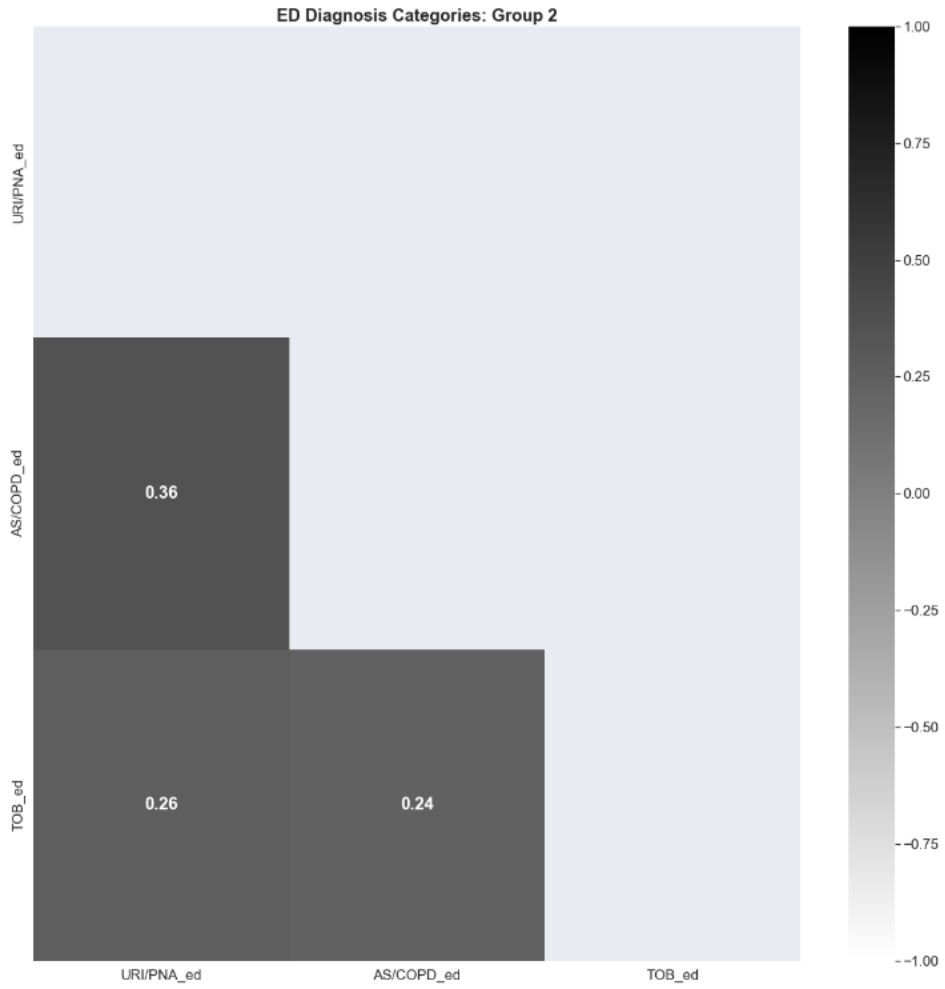
High correlations are seen here between Diabetes and Hypertension (0.34); Hypertension and Cardiovascular Disease (0.28); Hypertension, Sensory Deficits and Pain (0.21); Diabetes, Heart Failure, and Cardiovascular Disease (0.2), and Mental Health Disorders and Substance Use (0.17)



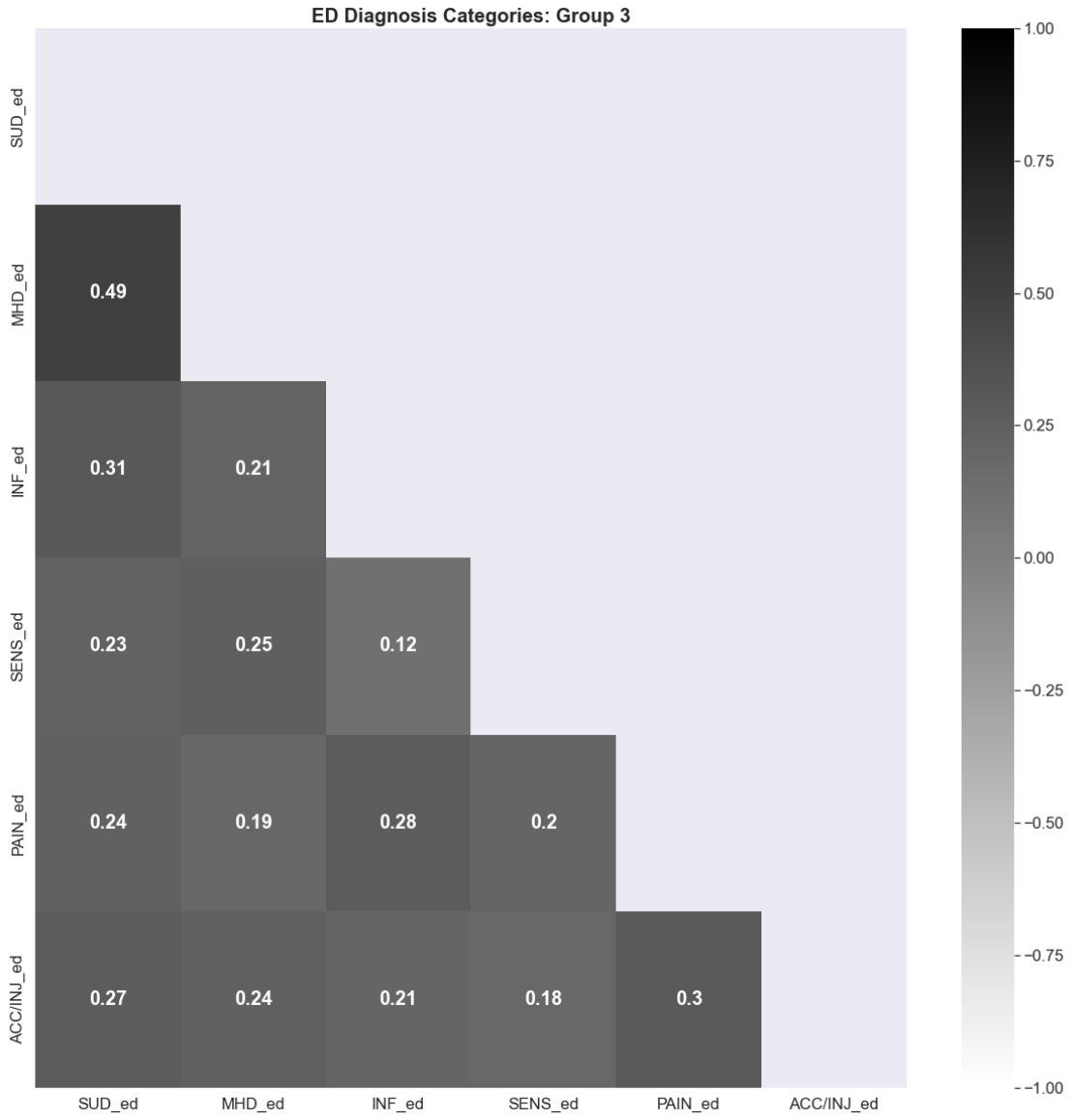
S1, Fig 2: Kendall Correlation: Emergency Department Visit Diagnosis Groups.
Some clinically related high correlations are grouped in the following sub-figures.



S1, Fig 3: Kendall Correlation: Four Highly Correlated and Clinically Related Emergency Department Visit Diagnosis Categories: Hypertension, Neurological Diseases, Diabetes & Related, and Cardiovascular Disease

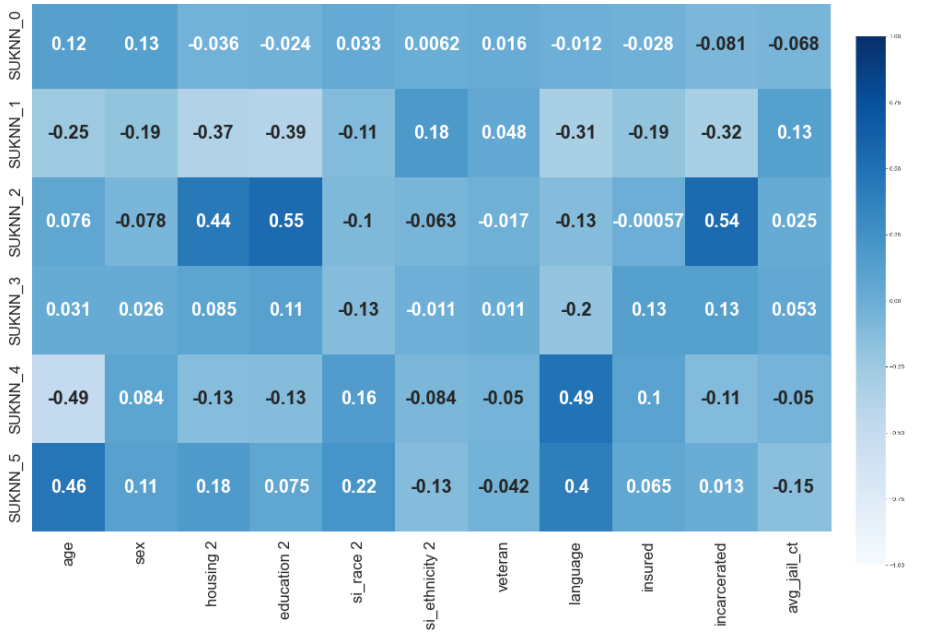


S1, Fig 4: Kendall Correlation: Emergency Department Visit Diagnosis Categories: Respiratory Diseases and Tobacco Use

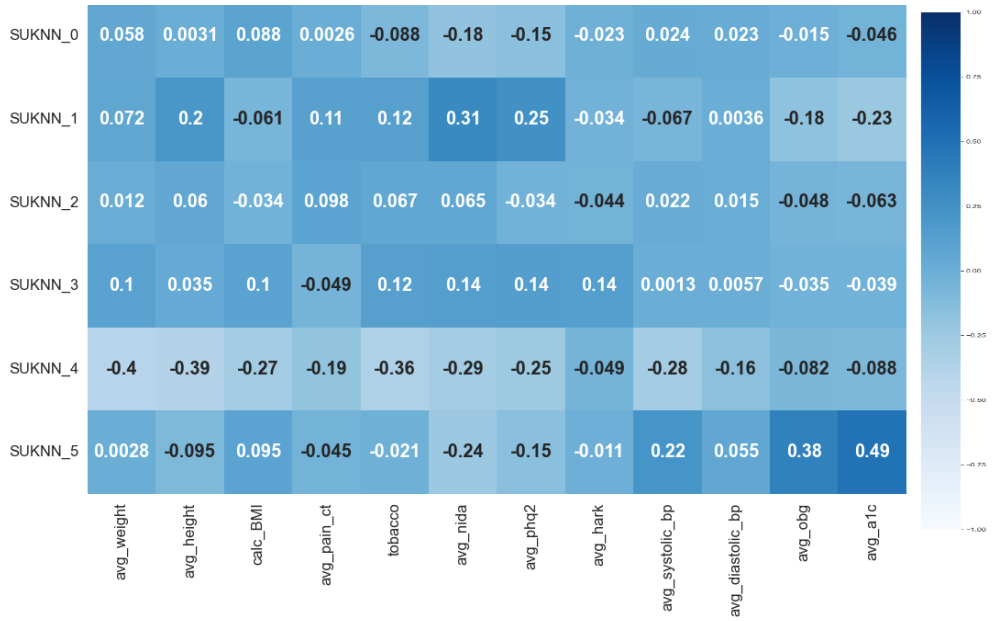


S1, Fig 5: Kendall Correlation: Emergency Department Visit Diagnosis Categories: Accidents & Injuries, Pain, Sensory Deficits, Infections, Substance Use, and Mental Health Diseases

Section 2: Cluster Correlation Heatmaps



S2, Fig 1: Correlation: "Spectral B" UMAP Clusters & Demographic Data

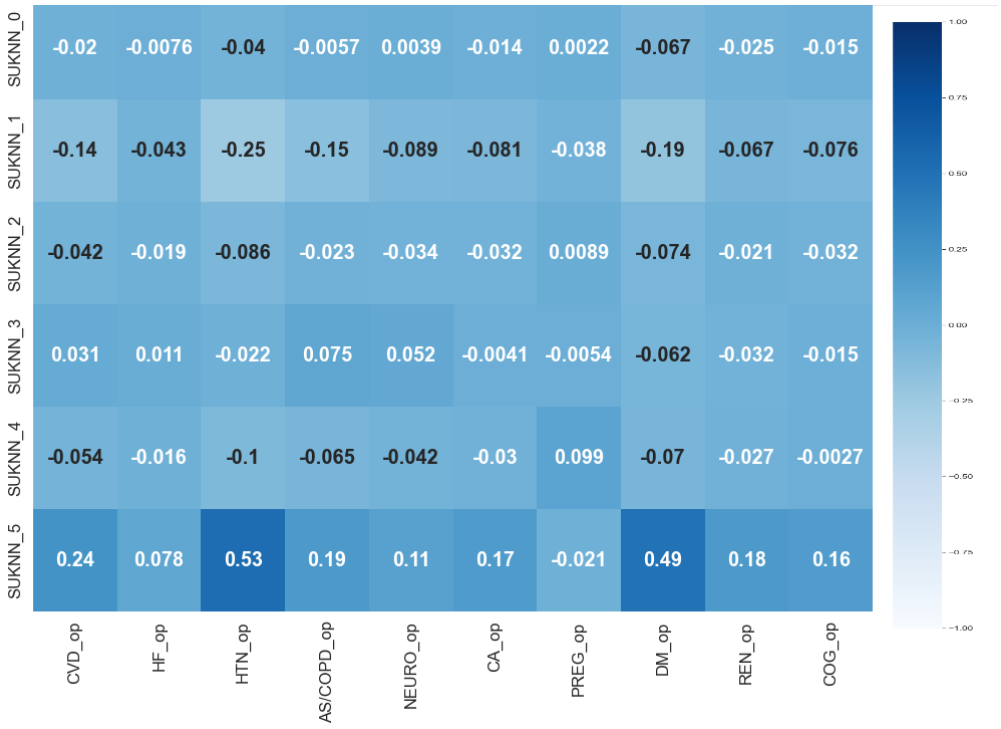


S2, Fig 2: Correlation: "Spectral B" UMAP Clusters & Assessments & Readings

S2, Fig 3: Correlation: "Spectral B" UMAP Clusters & Visit Features

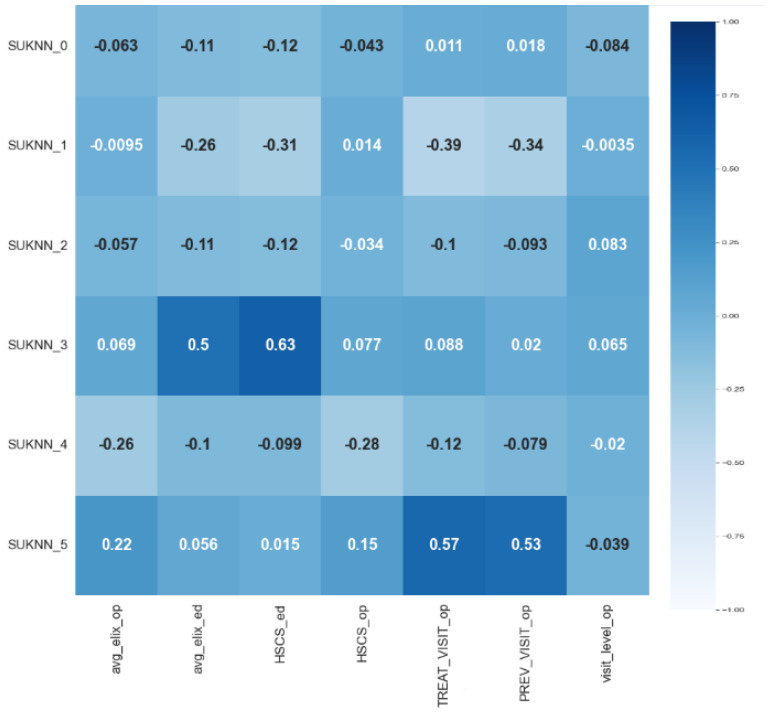
S2, Fig 4: Correlation: "Spectral B" UMAP Clusters, ED Diagnosis Groups – Part 1

S2, Fig 5: Correlation: "Spectral B" UMAP Clusters, ED Diagnosis Groups – Part 2



S2, Fig 6: Correlation: "Spectral B" UMAP Clusters, OP/Clinic Diagnosis Groups – Part 1

S2, Fig 7: Correlation: "Spectral B" UMAP Clusters, OP/Clinic Diagnosis Groups – Part 2



S2, Fig 8: Correlation: "Spectral B" UMAP Clusters, Overall Health