WORCESTER POLYTECHNIC INSTITUTE

Personas Development using Data from Electronic Health Records (EHR)

by

Gaayathri Sankar

A thesis submitted in partial fulfillment for the
degree of Master of Science
in
Innovation with UX,
Future of Robots in the Workplace –
Research & Development (FORW-RD NRT)
program

May 2021

APPROVED:

____

Soussan Djamasbi, PhD (WPI)

____

Yunus Dogan Telliel, PhD (WPI)

____

Daniel J. Amante, PhD, MPH (UMMS)

____

Adarsha S. Bajracharya, MD, MMSc (UMMS)

Abstract

Grounded in the User Experience Driven Innovation (UXDI) framework, this study aims at gaining a better understanding of patients' needs through persona development. This study is part of a larger research program that attempts to design a self-supporting technological intervention for type 2 diabetes patients. In this study, Electronic Health Record (EHR) data of type 2 diabetes patients was analyzed to develop a set of data personas each representing a unique patient group belonging to UMass Memorial Accountable Care Organization (ACO). The analysis resulted in 8 clusters based on 20 different patient characteristics. A major goal of the study was to explore patterns in the patient data that related to social determinants of health (SDOH) with an aim to discover data patterns for those patient groups who not only have high health risks but also vary based on factors such as their location, access to technology-based care and race and ethnicity. These patient groups are represented by the data personas which were generated using the electronic health records. Data personas were then mapped to proto personas, which were generated for the same population in a prior study. The mapping of proto and data personas facilitated the validation of the assumption based proto persons. In addition to validation, in general, such a mapping refines the understanding of patient needs by combining what is learned from each persona development approach thereby adding to the robustness and validity of the persona development process in user experience research.

Acknowledgements

First and foremost, I immensely thank my thesis and program advisor, Dr. Soussan Djamasbi for supporting my passion for data and analytics. Her enthusiasm, timely advice and inspiring words helped me enjoy the research process. Her commitment to the project despite many pressing obligations motivated me to give my best to the project.

I am very grateful to my committee members, Dr. Yunus Telliel, Daniel J. Amante and Adarsha S. Bajracharya for sharing their time, data, and insights on my project. My sincere thanks to teammate, Qiming Shi, for lending his time to this project.

I convey my deepest thanks to the NSF Research Traineeship program for helping me hone my skills to be a responsible researcher. The requirements of the Future of Robots in the Workplace – Research & Development (FORW-RD NRT) program has broadened my perspectives on research to make an ethical and social impact through my work.

I am thankful to my professors and administrators at the Foisie Business School at WPI for facilitating this thesis at the graduate level.

My special thanks to my data science friends, Vishaal and Vandana, for their constant and timely help and guidance during the project. My loving thanks to my immediate family for supporting me in multiple ways to make this degree possible.

Last but not the least, my loving gratitude to my dear parents for their bountiful love, innumerable sacrifices, and immense prayers. Thank you for standing by me during my hardships. I am truly indebted to you both for everything you have done for me in this lifetime.

Table of Contents

Introduction

As human needs are increasingly addressed by technological products and services, gaining a deeper understanding of user needs via persona development has become a major focus in product and service design (Jain, et al., 2019). With the rise in AI and big data, data persona development is an evolving technique in User Experience (UX) research to better understand user needs (Djamasbi & Strong, 2019). Information about users is widely used in customer analytics specially using social media data wherein automated and intelligent systems can quickly help understand customer profiles (An, et al., 2016). Segmentation and clustering using data mining and machine learning algorithms are frequently used in market research and business analytics. Using existing data to create data personas, along with other types of personas (e.g., proto and user personas) provide a more comprehensive picture of user needs that is crucial in developing effective products and services (Djamasbi & Strong, 2019).

The current project is part of a larger research project that seeks to understand the needs of Type II diabetes patients at UMass Memorial Medical Center in order to design and develop a self-support technological intervention. As with customer data in any business, data personas created from medical records can help design customized solutions for patient needs, especially with chronic diseases which require a long-term interventional strategy.

In the U.S., Type II diabetes is like an epidemic with a new diagnosis made every 17 secs (Congressional Caucus on Diabetes, n.d.). A study by Lin, et al., (2018) has projected that the number of US adults with diagnosed diabetes would almost triple between 2014-2060 and that over one in six adults would be diagnosed with it by 2060. These patients also tend to have complex comorbidities (Manzella, 2020). This can prove costly particularly for historically marginalized communities because these patients face barriers to services in terms of access, language, socio-economic status etc. and are less likely to engage with their care providers and have guideline-concordant care, including completion of diabetes self-management and training (CDC, 2010).

To understand the needs of Type II diabetes patients at UMass Memorial Medical Center in order to design and develop a self-support technological intervention, this project focuses on a preliminary exploratory approach to understanding patient needs via data persona development. Such an approach will help uncover latent patient needs from a set of patient records. A prior project focused on developing proto personas for the population under investigation in this project to understand their needs through the lens of clinicians and hospital administrators. This project focuses on 1) gaining an overall understanding of the needs of this patient population through the development of data personas using Electronic Health Record (EHR) patient data and 2) combining the knowledge gained in the current and previous project by mapping proto and data personas.

There are two main research objectives to be addressed by the current study:1) conduct an exploratory cluster analysis on EHR data to discover general trends in the patient population, 2) map these general trends and proto personas that were developed for the same population in a prior project to validate and/or refine assumptions made about this group of patients.

The analysis in this project serves as a first step to provide a direction for understanding general trends in the EHR dataset and a method for mapping proto and data persons so that the subsequent projects can refine the results of this study in order to create socially meaningful interventions.

The importance of understanding user needs via persona development is highlighted in the User Experience Driven Innovation (UXDI) framework (Djamasbi &Strong, 2019). The details of this framework are elaborated in the upcoming chapter along with an explanation of persona development. Additionally, a brief discussion of the social determinants of healthcare with respect to the type 2 diabetic health scenario in the North American continent is provided.

Background

Research framework

Designing patient-centered interventions require gaining a deep understanding of patients' needs. This study uses the UXDI framework (Djamasbi & Strong, 2019) to achieve this goal. The UXDI framework provides a conceptual base for designing and developing user centered UX innovations. Since the goal of this project is to aid the design of an innovative program intervention that facilitates patient-engagement, the UXDI framework has been selected for our current study. The UXDI framework, on a general note, comprises of two worlds – the design world and the usage world.

Design world consists of the ideation phase, which involves identifying opportunities and understanding users and the implementation phase, which includes designing, building, and evaluating experiences. The **usage world** consists of the implementation phase where the experiences are applied (Djamasbi & Strong, 2019).

The current project focuses on understanding users, which is part of the ideation phase and is carried out using various UX research methods. One such frequently used method is persona development.

Persona Development

A user-centered innovation places a premium value on understanding users' emotional and utilitarian needs, goals, challenges, etc. Developing effective user experiences must be backed by data to ensure successful adoption and continued usage of a product or service (Djamasbi &

Strong, 2019). An important UX research method for gaining a holistic understating of user needs is persona development, which not only helps to make informed design decisions, but also can help to prioritize business strategies. By creating a shared understanding of the target market among project stake holders, personas in UX research create an effective language for communicating design and business/intervention strategies. Personas are realistic yet fictional representations of the target users. Representing the target market via a set of personas makes it easier to understand design needs. It also helps to model, summarize, and communicate design ideas around a business need.

There is no standardized template for a persona. Usually a persona has a name, picture, characteristic attributes, background information, users' motivations, frustrations, goals, and some bio data (Jain, et al., 2019). Persona is an artifact that is related to a project, product or service that is to be designed or created (Persons, et al., forthcoming). Since personas reflect user needs, which change overtime, personas are typically updated to keep them current before any major design decision (Jain, et al., 2019).

There are three major types of personas in UX research– proto personas, data personas and user/research personas (Djamasbi & Strong, 2019; Jain, et al., 2019).

Proto personas are cost-effective and are largely drawn on assumptions based on what the project stake holders (including designers) think of the prospective user (Jain, et al., 2019). Proto personas are developed based on the knowledge of project stake holders (e.g., managers, administrators) who have direct or indirect contact with the users. Hence, proto personas reflect the indirect knowledge of or assumptions about a user group.

Though it is assumption-based, proto personas have a great deal of business value. Proto personas help businesses understand and prioritize their target market. They also help to get the entire management team (project stake holders) on board to champion a project and set business strategies. By doing so, proto personas help bring the decision makers “on the same page” regarding what they think about the user, facilitate the expression of multi-cultural perspectives, and enable cross-pollination of ideas (Jain, et al., 2019).

While proto persona development can capture a great deal of information about user groups (including demographics, user needs and wants, and their challenges and goals), it is assumption-based and as such benefits from validation through data personas and/or user personas.

Data personas are generated from user data (e.g., from their IoT devices and/or organizational systems). The availability of large-scale quantitative data facilitates the use of data science approaches for gaining a different view of the target market (Djamasbi & Strong, 2019; An, et al., 2016). While data personas provide a data-driven abstraction of users behaviors in a given period of time, they typically lack direct information about their affects, attitudes, and perceptions. Such information is typically gained through **user personas** (Jain, et al., 2019), which are created based on actual research conducted on-field with the users. While developing user personas tends to be more time and resource intensive than developing proto and/or data personas, it provides first-hand user information (e.g., perceptions, affects and attitudes) that is typically not available through data or proto personas.

Proto personas validated with data and user personas provide the most comprehensive set of information about users. Doing so combines pre-conceived notions, empathy, and rationality to arrive at highly informed user insights. This helps the decision makers to have the most comprehensive user information to come to a consensus on what user problems exist and how they could solve them. Because people's needs change over time, personas must be regarded as living documents. Hence, persona development process needs to be repeated regularly, preferably before any major design decision to enable continuous innovation (Jain, et al., 2019; Djamasbi & Strong 2019).

Creating data personas for this project

The goal of this project, as mentioned earlier, is to conduct a basic step toward exploring patterns in the EHR data that could be used to create useful data personas, which then can be compared to their proto persona counterparts. The result of this project serves as a basic step in a larger research project which aims at designing socially relevant technological program interventions that will provide the necessary self-care support to the Type II diabetes patients. One challenge to this end goal can be seen as attrition (Huh, et al., 2016). To keep the patients engaged, these interventions have to be tailored to their needs. To ensure sustained engagement, a deeper understanding of the context, in this case, the health environment is required (Djamasbi & Strong, 2019).

Type 2 Diabetes

According to the Centers for Disease Control and Prevention (2020), about 1 in 10 Americans have diabetes and among them, approximately 90-95% have **Type II diabetes**. It generally develops in people over 45. Diabetics tend to have multiple other health conditions such as hypertension, loss of eyesight, sleep apnea etc. Comorbidities in diabetes patients has long remained a cause of premature death and a reduced quality of life. Existence of multiple chronic conditions makes it difficult for a patient to manage their self-care (SUN, et al., 2018).

Being a chronic illness arising either due to genetics or a sedentary style of living, it is also a condition that has to be managed by the individual and with support from the health care team (Centers for Disease Control and Prevention, 2020). This kind of a support is being rendered through a focus on population health management.

Broader impacts – Social implications

According to the American Hospital Association, **population health management (PHM)** refers to *"the process of improving clinical health outcomes of a defined group of individuals through improved care coordination and patient engagement supported by appropriate financial and care models."* It has three basic components – patient health outcomes, patterns of health determinants, and policies and interventions (Kindig & Stoddart, 2003). This approach aims to improve the overall health of the entire population, thereby reducing inequities and disparities in the access to quality care. As defined by the Centers for Disease Control and Prevention, (2020), *"differences in health status or access to health care among racial, ethnic, geographic, and socioeconomic groups are referred to as health disparities."* To understand these disparities so that policies and interventions can be designed to assimilate patient needs, a primary understanding of the determinants affecting health outcomes is necessary, in particular, social determinants of health (SDOH), to ensure better PHM.

○ Social determinants of health (SDOH)

Health, in a general sense, is determined by a number of factors including access to social and economic opportunities, resources and support at family, neighborhood and community levels, education level, workplace safety, access to clean food, water, and air and social interactions and relationships. According to the Office of Disease Prevention and Health Promotion, (2020), social determinants of health are "*conditions in the environments in which people are born, live, learn, work, play, worship, and age that affect a wide range of health, functioning, and quality-of-life outcomes and risks.*"

Many variables that have been included in the current study such as race/ethnicity, zip codes, income levels, age, and gender help understand how some of these SDOH factors define type 2 diabetic patient needs with respect to designing self-care interventions for them. By considering them earlier to designing the interventional program, the study aims to take a proactive approach in addressing the problem statement (Freidman & Hendry, 2019).

These variables and how they have been analyzed have been explained in the upcoming chapter.

Methodology

This chapter provides an overview about the datasets and the variables used in the study to create data personas. It explains the processes of data extraction, data cleaning and data preparation that have been employed and concludes with data analysis, particularly cluster analysis. Deploying cluster analysis has helped in segmenting the type 2 diabetes patient population and in identifying the data personas, which will then be used in a subsequent project to design self-care program interventions based on these personas.

Datasets

The datasets employed in the study contain deidentified patient Electronic Health Records (EHRs) derived from the Epic Clarity database. EHRs are a digital version of a patient's information that go beyond the standard clinical data to include other aspects of the patient's life such as their background and disease history. These digital records enable instant and secure access to patient information to authorized users (The Office of the National Coordinator for Health Information Technology (ONC), 2019). Epic is a cloud based EHR software solution that focuses on patient engagement and facilitates remote care. On the front end, it achieves this through a patient portal called MyChart which is an app available on both Android and iOS platforms (Epic EHR, n.d.). At the backend, it uses the Clarity database which is an Oracle/Microsoft SQL server database that aids in exporting, transforming, and loading data from the MyChart environment (The Trustees of the University of Pennsylvania, n.d.). The datasets extracted from Clarity, which in turn are extracted from MyChart, are static and do not provide real-time information such as those supplied by IoT devices. This means that this data and its subsequent analysis is historical in nature.

The datasets used in the study were constructed using the method described in the Data Extraction section below.

Variables used in the study

The variables that have been analyzed in this study are categorical in nature. While the data records used in this project hold a large number of variables, only those variables that have been identified as important in defining and analyzing the problem statement have been extracted and included in the analysis. Following is a list of the 22 variables that have been used in this project. Subsequent sub-sections in this chapter explain these variables and discuss how these variables have been collected, cleaned, and prepared for data analysis and interpretation.

Geographic variables:

- Zip code
- Census tract ID

Demographic variables:

- Income
- Age
- Gender
- Race
- Ethnicity
- Marital status
- Primary language
- Insurance company
- Insurance type

Technological variables:

- Home phone
- Email address
- MyChart status

Health and health care variables:

- Care duration
- LACE+ score
- Body Mass Index (BMI)
- HbA1C levels
- Alcohol/drug/smoking (social_HX)
- Accountable Care Organization (ACO)
- Appointment distance

Explanation of important variables

Zip codes, **census tract data** as well as **inflation-adjusted median household income** have been used to reflect the socio-economic status of patients. Zip codes give a higher-level picture while census tract data helps to analyze neighborhoods within these zip codes. By matching them with the median household income, one of the goals has been to identify economically disadvantaged areas so that the design of technological self-care interventions can be targeted and personalized to these patients accordingly.

Home phone, **email address** and **MyChart status** have been used to assess access to technology and comfort with technology usage. This is important in order to design a technological self-care intervention. Particularly, whether or not a patient has an email address has been used as a proxy to indicate active technology access.

As a proxy for health risk, the **LACE+ score** is an indicator of the risk of readmission to the hospital thereby predicting mortality rates. It is calculated by considering factors such as age, sex, length of stay, comorbidities, whether the patient was admitted through the ED, number of ED visits in the six months prior to admission and number of days the patient was in an alternative level of care during admission etc. (Weiss, 2017).

A major contributing risk factor in diabetes is body weight. Obesity in itself has become a global pandemic and medical research has found that a majority of type 2 diabetes patients are obese (Eckel, et al., 2011). This risk factor is represented by BMI in the EHR data where a higher BMI reflects greater obesity.

While LACE+ is a general proxy for risk and a high BMI cannot always be tied to type 2 diabetes, a patient's **HbA1C level** has been chosen as a more specific indicator of health risk. The Hemoglobin A1C levels reflect blood glucose levels therefore indicating whether the blood sugar is under control or not. This is a more direct measure of risk for type 2 diabetes.

A major player putting together the socio-economic status and health risk of patients is the **insurance** industry. Medicare and Medicaid are both government-run insurance programs. Medicare is a federal program for people above 65 years of age or those under 65 who have a disability. Age and disability are the defining factors, irrespective of income. On the other hand, Medicaid is a state and federal program for very low-income populations. State insurance is another type of insurance that is tied to a particular state, irrespective of age, disability, or income levels (Medicare Interactive, n.d.). A person can be eligible and therefore possess more than one type of insurance. These definitions of each type of insurance have led to using insurance type as another proxy to further inform patient socio-economic status.

Under Medicare, in order to aid PHM and provide high-quality coordinated care to patients, **Accountable Care Organizations (ACOs)** have been formed. They comprise of a group of doctors, hospitals and other health care providers who voluntarily come together with the aim of avoiding unnecessary duplication of services, preventing medical errors and spending health care dollars more wisely. The ACO gets a share in the savings that it achieves for the Medicare program. These providers will make more if their patients are healthy and out of the hospital (Centers for Medicare and Medicaid Services, 2021).

The UMass Memorial ACO is focused on providing Worcester and Central Massachusetts with quality, personalized and innovative care. Some of their objectives include providing personalized care coordination services across the health continuum and providing targeted health information and wellness programs that focus on disease prevention (UMass Memorial Health, 2021).

Data extraction

Datasets extracted contained a total of 53,675 deidentified patient Electronic Health Records (EHR) filtered for patients above the age of 18. The extraction of these datasets had been done from the EPIC Clarity database. For home addresses including street address, city, state, and zip code, ArcGIS had been utilized alongside EPIC Clarity to geocode and transform home addresses into latitudes and longitudes to facilitate deidentification as well as to aid geo mapping and cluster analysis.

During the period of the study, data extraction had taken place twice. Twenty variables had been extracted in the first dataset. The second extraction that occurred on 03/09/2021 included the variables HbA1C and income in addition to the 20 variables. Hence, a total of 22 variables were prepared for analysis.

LACE+ and BMI were used as overall indicators of patient health. For a more specific indicator of type 2 diabetic health risk, information on HbA1C was used. To determine socio-economic status in addition to zip codes, information on patient incomes had been included in the second extraction.

Data cleaning and Data preparation

Zip code:

The raw data had been expressed either in 4-digit, 5-digit or 9-digit zip code formats. To ensure uniformity and to aid in further analysis, in Excel, using the formula =TEXT(x, “00000”), all the 4-digits had been text-converted to 5-digits with leading zeroes. Then, the 9-digits had been truncated to 5 using the formula =LEFT(x, 5). The format of the resulting variable had then been changed to ‘Zip Code’ to ensure that it is rightly read by visualization tools like Tableau. A zip code having raw data 0 was converted to NULL to ensure it stays distinct, else conversion to a 5-digit zip code would prove of significance when there actually has been none.

Deidentified census tract ID:

This ID had been pulled along with an estimation of inflation-adjusted median household income for 2019. It had been observed that a NULL value of income had a census tract of 0. To ensure clarity while analysis, the zeroes have been converted to ‘NULL’ since the census tract is expressed and analyzed numerically.

Inflation-adjusted median household income of 2019:

The raw data had contained numerical data expressed in dollars. This had been converted to categorical data. The categories had been based on an article by Snider, (2020) from the U.S. News & World Report on the American economic class system.

- Poor or near poor – less than or equal to $32048
- Lower-middle class – between $32048 and $53413
- Middle class – between $53413 and $106827
- Upper middle class – between $106827 and $373894
- Rich – above $373894

Age:

Patient data only included those records where patient age had been above 18. The continuous numerical raw data for patient age had been converted categorically for the purpose of analysis (upper limit included) – below 30, 30-40, 40-50, 50-60, 60-70, 70-80, 80-90, 90-100 and above 100.

Home phone and email address:

The raw data for both these variables had been binary, i.e., having any of these had been represented by 1 and not having them had been represented by 0. This binary numerical data had also been converted categorically to aid better analysis and visualization.

- 1 - Has home phone/email address
- 0 – No home phone/email address

Duration of care:

The raw data had a given date for each patient record. This date had been the date on which the patient data was first ever entered in the system. The dataset pulled from the Clarity-Epic database happened on 03/09/2021. The difference between the date of data pull and the date in the dataset has given the number of years the patient record has been in the system, used as a proxy for the duration of care. (Excel formula: =DATEDIF(record date, 03/09/2021, “y”)). Based on stakeholder requirements, the number of years arrived at had then been categorized into less than 1 year, between 1-5 years and greater than 5 years to aid in analysis.

Primary language:

The dataset had 93 different languages. Each language was considered as a separate category and all the 93 different languages were included in the analysis.

LACE+ score:

The raw dataset had numeric data of a continuous nature. For the purpose of analysis, based on the report by Weiss (2017), the LACE+ scores had been converted to categorical variables.

- LACE+ score of 0-28 – low risk
- LACE+ score of 29-58 – moderate risk
- LACE+ score of 59-90 – high risk

Body Mass Index (BMI):

The raw data for BMI had been extracted in the form of continuous numeric data. For the purpose of analysis, it has been converted to categorical data. The categories have been based on those defined by the Centers of Disease Control and Prevention, (2020) for adults 20 years of age and above.

- Below 18.5 – underweight
- 18.5-24.99 – normal
- 25.0–29.99 – overweight
- 30.0 and above - obese

Insurance company:

The dataset has a list of 85 different insurance companies. Each company was considered as a separate category and all the 85 different companies were included in the analysis.

Insurance type:

The dataset has four types/categories of insurance – Medicare, Medicaid, Commercial and State. Every patient record in the dataset has only one of the four insurance types.

Social_HX - Alcohol, drug, and smoking status:

The dataset has each of these three variables expressed categorically – Former, Current, Never and Not Asked.

ACO:

The raw dataset had this variable coded binarily. To convert it categorically, 0 had been recoded as ‘not part of an ACO’ and 1 as ‘part of ACO’.

HbA1C:

HbA1C levels in the raw dataset had been of a mixed nature. It had consisted of percentages, fractions, and text such as “*SEE COMMENT*”, *“< than ^4.5*” . The records containing these values were removed since they did not denote any useful value. The A1C value of 637 in one record was an outright outlier and hence this record was removed. The rest had been converted to decimal numbers up to two places. This continuous data, for the purpose of analysis, had been converted into three categories - less than 6.5, between 6.5-8 and greater than 8.

Appointment distance:

The raw dataset had defined appointment distance as the total number of office visits made by a patient over the years to date. It has been categorized as less than 1, between 1-5 and greater than 5, expressed in terms of the number of office visits.

Data analysis

A set of charts for each variable in the entire dataset was created using Tableau (v.2020.3.2) and Excel in order to understand the data better. Since data extraction had occurred twice, data cleaning and preparation led to variables HbA1C and income (that had been added from the second extraction) to have 53314 records while the rest have 53675 records. Further, the dataset had been filtered to only contain patients belonging to an ACO. This filtering from the 53314 records resulted in 6365 records. To understand the characteristics of the ACO population, a set of charts visualizing each variable was created. These results had been compared to the overall population (in terms of percentages) for important variables. Finally, two-step cluster analysis had been employed on the ACO population to create an initial set of data personas. This analysis resulted in eight clusters.

The ACO patients have been of primary focus in the study based on hospital needs to understand this population better. In addition, the proto personas that had been developed earlier to this study focused on the ACO patients. To facilitate better validation of the persona development process, the cluster analysis focused on the ACO patient population.

Two-step cluster analysis

Consistent with a recent medical research (Lee, 2020) that examines similar variables to this current project, two step cluster analysis was used to explore major patterns in the dataset. Cluster analysis is a multivariate statistical analysis technique. A two-step cluster analysis is an exploratory tool that identifies groupings by first running a pre-clustering followed by a hierarchical methods clustering (Norusis, 2011). The two-step cluster analysis was conducted for the ACO population using IBM SPSS Statistics 26. The two-step cluster analysis has been chosen because of the large number of records in the dataset and their categorical nature (Norusis, 2011). Twenty of the 22 variables have been included in the cluster analysis, the exceptions being zip code and deidentified census tract ID. The zip codes have been retrospectively analyzed for each cluster (refer the next chapter for details) while the reason for excluding the census tract IDs has been explained in the ‘Limitations and Future work’ chapter. Even though the two-step clustering method can automatically determine the number of clusters, the objective to ensure a good cluster quality (and because proto persona development identified 8 user groups) resulted in manually assigning the number of clusters.

Cluster quality – The silhouette measure and the ratio of sizes together indicate the quality of the clusters that have been generated.

The Silhouette Index is a good indicator of cluster quality which is automatically generated by the algorithm. The silhouette measure of cohesion and separation explains how distinct the clusters are from one another and how close together the variables are within the cluster as well as between the clusters. The measure ranges from -1 to +1 indicating cluster quality range from poor to good (values above zero indicate good cluster quality). According to (Norusis, 2008), a measure above 0.0 suggests the validity of within and between-cluster distances. Therefore, in our current study, the silhouette measure standard had been set at above 0.0.

A good ratio of cluster sizes is debatable, however, ratios < 3 in two-step cluster analysis in SPSS are considered to be a suitable criteria. Hence, in this project the ratio of cluster sizes had been set at < 3.

Determining the number of clusters – To explore the data, an automatic determination of the number of clusters was performed which resulted in 4 clusters with an average silhouette score of 0.1 and a ratio of sizes at 3.43. Since the ratio had been greater than 3, a manual determination of the number of clusters had been undertaken. Again, to explore the data, manual trials included the number of clusters (n) ranging from 2-10. Every trial resulted in an average silhouette score of 0.1 with a ratio of sizes greater than 3 except for n=8 where the ratio of sizes had been 2.77. Further, prior proto-persona analysis had resulted in n=8. Therefore, the number of clusters for data persona development had been manually set at 8.

Selecting the distance measure – *"A distance measure is an objective score that summarizes the relative difference between two objects in a problem domain."* (Brownlee, 2020). Two-step cluster analysis in IBM SPSS offers two choices of distance measures – log-likelihood and Euclidean distance. For the current study, the log-likelihood measure has been selected as the distance measure as opposed to Euclidean distance measure because Euclidean distance is best suited for continuous variables (IBM, 2021). Since the prepared dataset consisted of categorical variables, the log-likelihood distance measure has been applied.

According to (Norusis, 2011), the algorithm is believed to run reasonably well even when the assumptions for best performance are not met. To the extent of these assumptions, while it is quite reasonable to assume independence, some variables, such as income and technology access, age and HbA1C etc. could be somewhat related. To this extent, the two-step cluster analysis technique is at best an approximation to reality (IBM, n.d.). Cluster analysis is an unsupervised technique that does not test any specific hypothesis. So, the solutions at best reflect the needs for which it is run, and the definition of a satisfactory solution is best defined by the authors (Norusis, 2011).

Clustering criterion – Clustering criterion is the rule based on which the clustering algorithm decides the number of clusters. For automatic determination of the number of clusters, two-step clustering in IBM SPSS offers two clustering criteria to choose from – AIC and BIC. The initial automatic clustering was run using the BIC criterion. This was chosen based on a similar study by Lee, (2020). For manual determination, since the number of clusters are specified by the investigator, SPSS does not need a criterion (it uses a default algorithm to calculate it).

As stated previously, based on the number of proto personas (and exploratory analysis), the manual cluster analysis was chosen to identify overall patterns in the data. This procedure showed that 8 clusters with an average silhouette score of 0.1 and the ratio of sizes at 2.77 were best fit for the data.

The basic charts for the overall patient population as well as the ACO population along with the results of the cluster analysis and resulting patient characteristics are discussed in detail in the upcoming chapter.

Results

Data visualization with charts

Overall population

A basic analysis of the overall type 2 diabetic patient population containing 53675 records resulted in the following observations. (Refer Appendix B.1 for details)

- Geographic distribution - Patient distribution across the North American continent shows maximum distribution in Massachusetts. This could be attributed to the location of the hospital as well as patient location and subsequent access.

- Patient demographics - The UMass EHR system has an almost equal representation between male and female patients. 72% of the patients fall in the age group 50-80. 52% of the patients are married followed by 23% single, 11% widowed and 10% divorced patients. 72% of the patients are white, non-Hispanic/Latino. 87% of the patients speak English as their primary language, followed by Spanish at 8%. There are no rich patients in the dataset. 52.97% of the patients belong to middle class while 5.5% of the patients are poor or near poor. Almost half the patient population (49%) has Medicare. 34% of the patients have Commercial insurance while 14% have State insurance and 2% have Medicaid.

- Access to technology – As mentioned previously, the study has used 'email' as a proxy for active technology access. Based on this, around 60% of the patients have an email.

Out of this 60%, 34% of them have an activated MyChart status while for 26%, its inactivated. The other 40% do not have an email among whom less than 1% have an active MyChart status.

- Health risk - Concerning the risk of readmission as indicated by LACE+ scores, 36% of the patients have a moderate risk, 25% have a high risk while 21% have a low risk and 18% are NULL values. In terms of BMI, 54% of the patients are obese, 28% are overweight, 14% are normal, less than 1% are underweight while 5% are NULL values. Between the obese, overweight, and normal categories, it has been observed that each group is twice as big as the category prior. In other words, the number of obese patients are almost twice the number of the overweight patients and overweight patients are twice as high as the normal patients. The proportion of patients with a higher risk tends to double as the risk category increases. Most of the patients have low-moderate HbA1C levels with 38% less than 6.5 and 31% between 6.5-8. Patients with high blood sugar constitute 17% of the patient population, having HbA1C levels greater than 8. The data shows that 19% of the patients have never smoked, had alcohol or drugs.

- Healthcare - Around 87% of the patients have been with the UMass EHR system for more than 5 years. 12% of the patients have been with the system for 1-5 years while 1% of them have been with the system for less than a year. Around 51% of the patients have had between 1-5 appointments while 17% have had less than 1 appointment and 8% of them have had more than 5 appointments. 12% of the patients belong to an ACO while the rest 88% are not part of any ACO.

Other observations made about the overall population include patient characteristics in the most frequently occurring zip codes and a geo-mapping of insurance types to better understand socio-economic spread. (Refer Appendix B.2 for details)

- Worcester, Leominster, and Fitchburg are the top three most frequently occurring zip codes in the dataset. In other words, more patients are concentrated in these regions.
- Leominster and Fitchburg have the highest number of patients covered by Medicaid while Marlborough has the highest proportion of patients covered by Medicare. This information indicates that not only are there more patients in Leominster and Fitchburg, but also that they belong to a lower socio-economic status.
- Compared to the other regions, in Shrewsbury, which is among the top ten most frequently occurring zip codes in the dataset, the spread of patients is concentrated to an older age group with a majorly low-moderate risk of readmission. This data pattern might be indicative of better existing patient care in this region. Very few patients are covered by Medicaid and the proportion of patients covered by commercial insurance is highest in this region. This data pattern might be indicative of better socio-economic status in this region.

ACO population

The objective of this project is to create data personas for the ACO patient population. A basic statistical analysis of the ACO patient population containing 6365 records (12% of the overall patient population) resulted in the following observations. (Refer Appendix C for details)

- General characteristics - The ACO population has older patients with higher income levels and active technology access. They are mostly enrolled in Medicare. Their risk of readmission majorly varies from moderate-high, with patients being obese/overweight with lower HbA1C levels. Some have been former smokers. The ACO patients are quite actively engaged with the system and have been receiving care for more than 5 years.

- Distinct characteristics from the overall population - ACO has an older population (60-90) as compared to the overall population (majority being in 50-80 age range). ACO population has a slightly better active technology access than the overall population (by around 10%). Compared to 61% in the overall population, 67% of the ACO population has moderate-high risk of readmission to the hospital. ACO patients seem to have better HbA1C levels when compared to the overall population (mean of 6.74 in ACO as compared to 7.01 in overall). Double the percent of ACO patients (90%) have Medicare as compared to the overall population (45%). This could be due to the ACO population inclined to an older age group. More ACO patient population (95%) has been with the system for more than 5 years as compared to 87% in the overall population.
- Other observations – The ACO population has 11% widowed women as compared to 7% in the overall population. Almost half of ACO patients (48%) have been former smokers as compared to 38% in the overall population. The ACO population has a lesser proportion of NULL values in all the variables as compared to the overall population. This could mean that data collected from the ACO patients is more complete but could still be improved.

Cluster analysis

To explore the overall patterns in the dataset, the cluster analysis included all variables in the data set (Table 1). The two-step cluster analysis resulted in eight clusters with an average silhouette measure of 0.1. In Figure 1 below, the pie-chart shows the distribution of cluster sizes with the biggest cluster containing 20% of the records and with the smallest having 7.2%. Thus, the ratio of sizes (largest cluster to smallest cluster) is 2.77 which is less than the accepted maximum ratio threshold of 3. Table 1 shows the importance assigned by the clustering algorithm to each of the variables in determining the clusters. The predictor importance varies on a scale ranging from 0-1. Every patient record has been assigned to only one of the eight clusters.

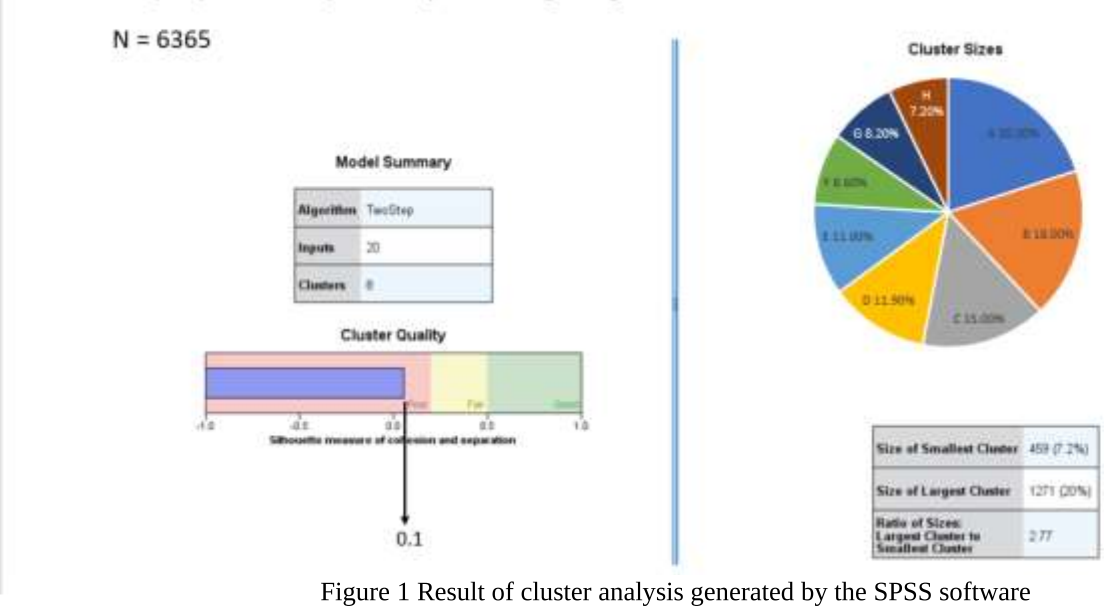
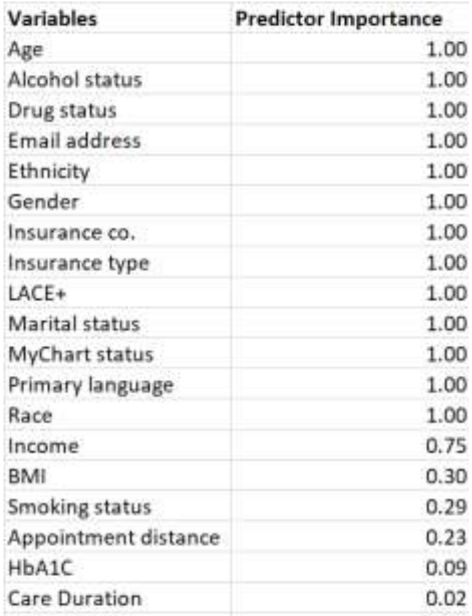


Figure 1 Result of cluster analysis generated by the SPSS software



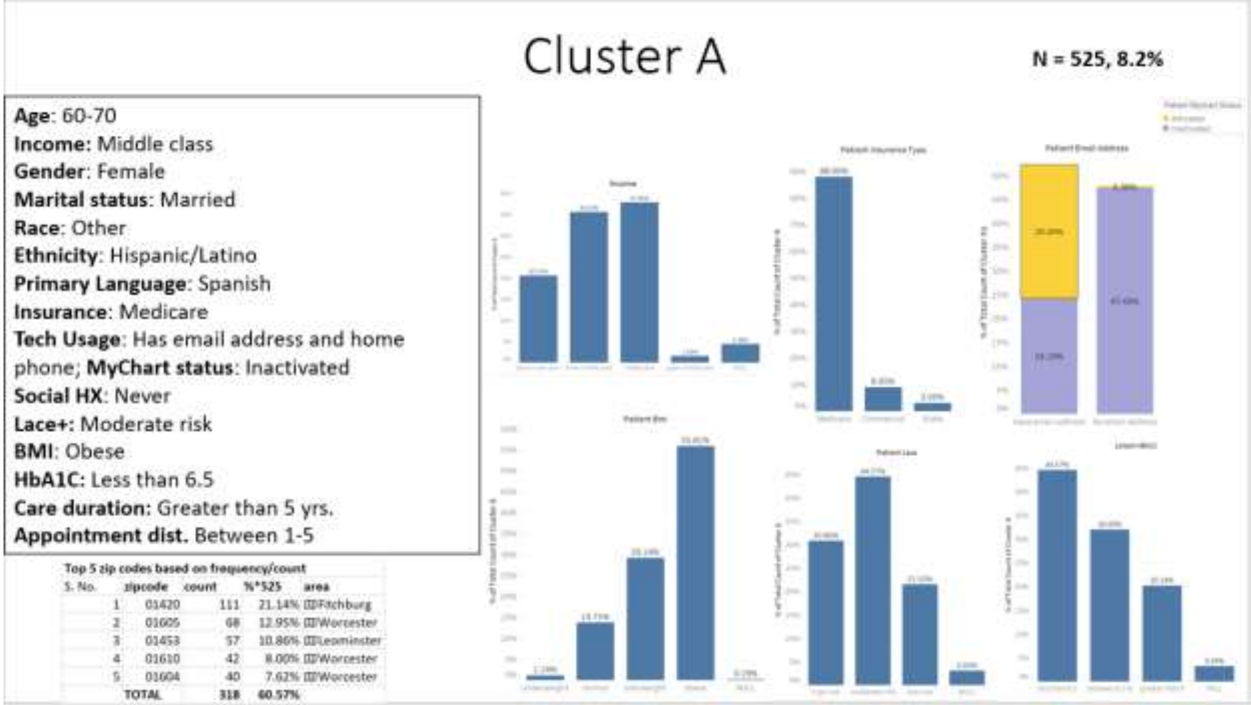
Variables	Predictor Importance
Age	1.00
Alcohol status	1.00
Drug status	1.00
Email address	1.00
Ethnicity	1.00
Gender	1.00
Insurance co.	1.00
Insurance type	1.00
LACE+	1.00
Marital status	1.00
MyChart status	1.00
Primary language	1.00
Race	1.00
Income	0.75
BMI	0.30
Smoking status	0.29
Appointment distance	0.23
HbA1C	0.09
Care Duration	0.02

Table 1 Predictor importance of each variable

In the following figures 2-9, each of the eight clusters has been analyzed separately to visualize the distribution of important variables such as income, insurance type, tech usage, BMI, LACE+ and HbA1C in each cluster. These variables were identified as important for identifying patterns in the data. An important observation about these clusters is that the distinct patient characteristics that represent each of them are based on the mode. So, this means that a patient characteristic is not absolute but is relatively defined for each cluster. For example, if in a particular cluster, the maximum number of patients belong to Medicare, then the insurance type of that cluster is defined as Medicare. But this does not mean that patients belonging to other insurance types do not exist in the cluster. It is also for this reason that the basic stats for the important variables mentioned have been drawn up. This gives a clearer view of the distribution of patient characteristics in each cluster.

While the figures portray distinct characteristics, 4 characteristics namely, BMI, home phone, duration of care and drug status have been identified as common to all the eight clusters. To elaborate, in all of the clusters, patients have a home phone, are majorly obese, have been with the system for more than 5 years and have never consumed drugs. Again, these common characteristics are based on the mode or most frequently occurring category of each variable which happens to be the same for all the clusters for these 4 variables.

Also, as mentioned in the previous chapter, the zip codes have been retrospectively analyzed. This had been done in the following manner. The clustering algorithm assigned every patient record to a cluster. This helped analyze the zip codes belonging to each cluster and thereby drawing up the top five most frequently occurring zip codes for every cluster. This has been included in a tabular format along with the cluster characteristics and charts for every cluster.



Cluster A

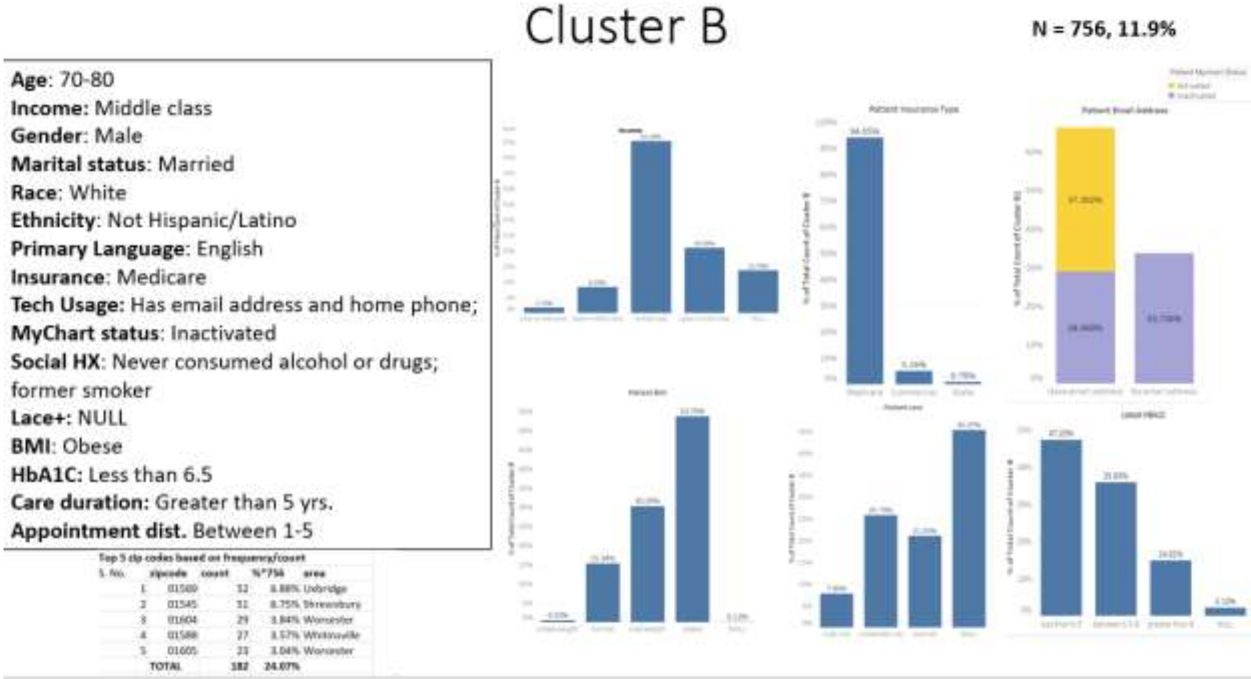
N = 525, 8.2%

Age: 60-70
Income: Middle class
Gender: Female
Marital status: Married
Race: Other
Ethnicity: Hispanic/Latino
Primary Language: Spanish
Insurance: Medicare
Tech Usage: Has email address and home phone; **MyChart status**: Inactivated
Social HX: Never
Lace+: Moderate risk
BMI: Obese
HbA1C: Less than 6.5
Care duration: Greater than 5 yrs.
Appointment dist. Between 1-5

Top 5 zip codes based on frequency/count

S. No.	zipcode	count	%*525	area
1	01420	111	21.14%	□□Fitchburg
2	01605	68	12.95%	□□Worcester
3	01453	57	10.86%	□□Leominster
4	01610	42	8.00%	□□Worcester
5	01604	40	7.62%	□□Worcester
	TOTAL	318	60.57%	

Figure 2 Cluster A characteristics and charts



Cluster B

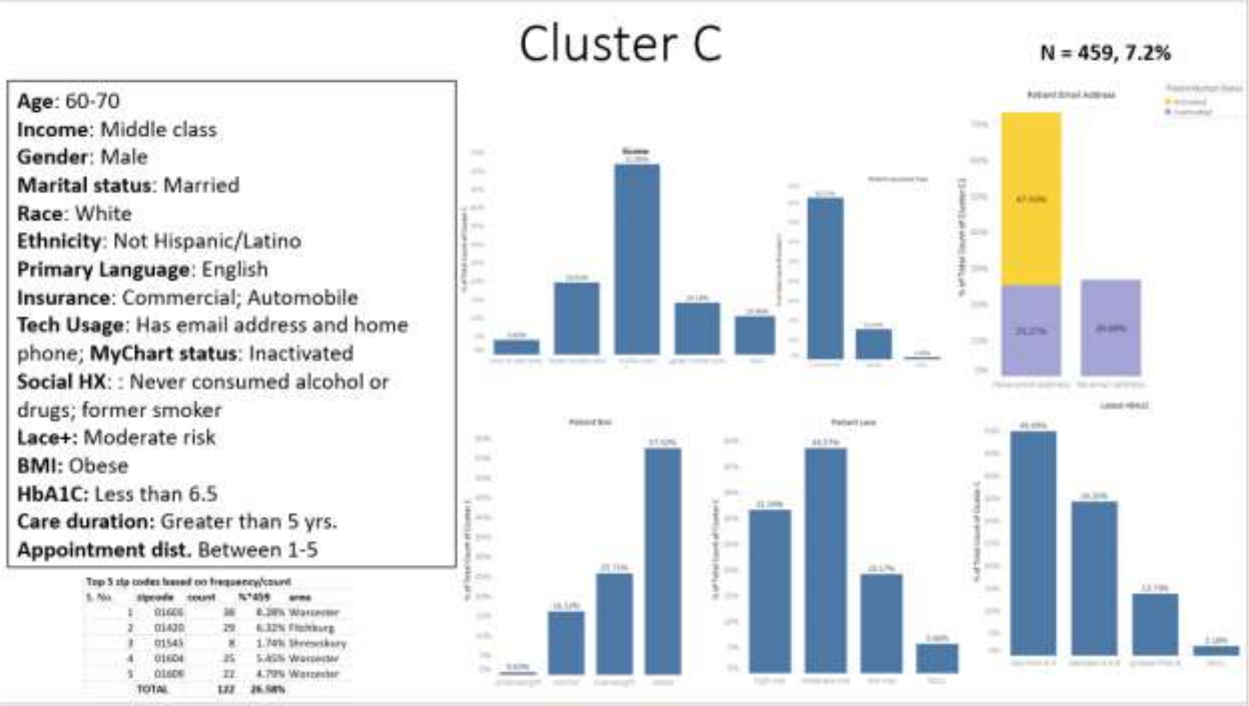
N = 756, 11.9%

Age: 70-80
Income: Middle class
Gender: Male
Marital status: Married
Race: White
Ethnicity: Not Hispanic/Latino
Primary Language: English
Insurance: Medicare
Tech Usage: Has email address and home phone;
MyChart status: Inactivated
Social HX: Never consumed alcohol or drugs; former smoker
Lace+: NULL
BMI: Obese
HbA1C: Less than 6.5
Care duration: Greater than 5 yrs.
Appointment dist. Between 1-5

Top 5 zip codes based on frequency/count

S. No.	zipcode	count	%*756	area
1	01569	52	6.88%	Uxbridge
2	01545	51	6.75%	Shrewsbury
3	01604	29	3.84%	Worcester
4	01588	27	3.57%	Whitinsville
5	01605	23	3.04%	Worcester
	TOTAL	182	24.07%	

Figure 3 Cluster B characteristics and basic charts



Cluster C

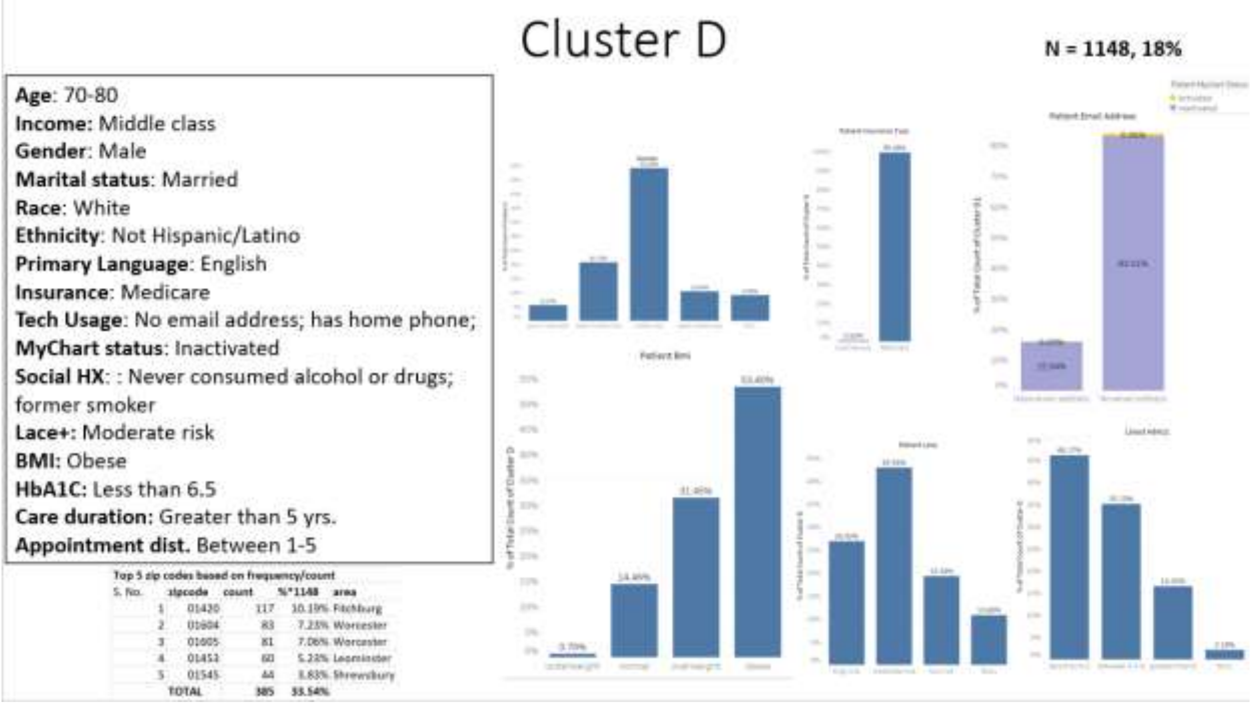
N = 459, 7.2%

Age: 60-70
Income: Middle class
Gender: Male
Marital status: Married
Race: White
Ethnicity: Not Hispanic/Latino
Primary Language: English
Insurance: Commercial; Automobile
Tech Usage: Has email address and home phone; **MyChart status**: Inactivated
Social HX: : Never consumed alcohol or drugs; former smoker
Lace+: Moderate risk
BMI: Obese
HbA1C: Less than 6.5
Care duration: Greater than 5 yrs.
Appointment dist. Between 1-5

Top 5 zip codes based on frequency/count

S. No.	zipcode	count	%*459	area
1	01605	38	8.28%	Worcester
2	01420	29	6.32%	Fitchburg
3	01545	8	1.74%	Shrewsbury
4	01604	25	5.45%	Worcester
5	01606	22	4.79%	Worcester
	TOTAL	122	26.58%	

Figure 4 Cluster C characteristics and charts



Cluster D

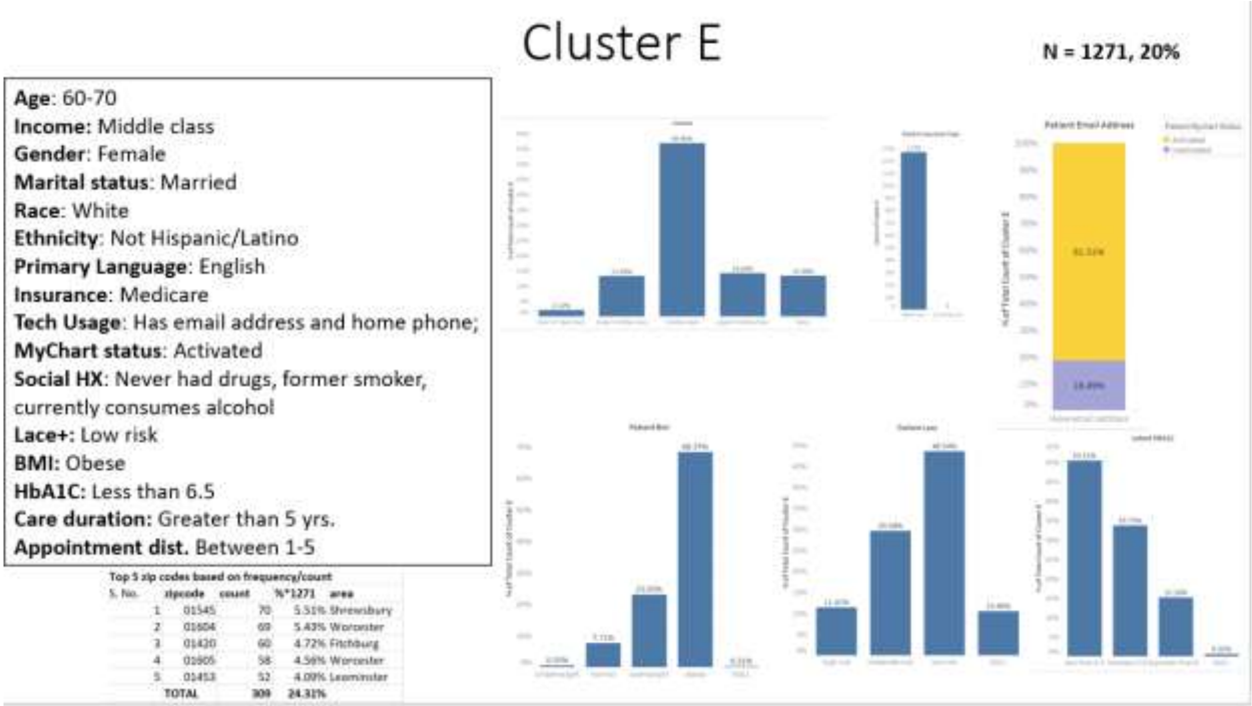
N = 1148, 18%

Age: 70-80
Income: Middle class
Gender: Male
Marital status: Married
Race: White
Ethnicity: Not Hispanic/Latino
Primary Language: English
Insurance: Medicare
Tech Usage: No email address; has home phone; **MyChart status**: Inactivated
Social HX: : Never consumed alcohol or drugs; former smoker
Lace+: Moderate risk
BMI: Obese
HbA1C: Less than 6.5
Care duration: Greater than 5 yrs.
Appointment dist. Between 1-5

Top 5 zip codes based on frequency/count

S. No.	zipcode	count	%*1148	area
1	01420	117	10.19%	Fitchburg
2	01604	83	7.23%	Worcester
3	01605	81	7.06%	Worcester
4	01453	60	5.23%	Leominster
5	01545	44	3.83%	Shrewsbury
	TOTAL	385	33.54%	

Figure 5 Cluster D characteristics and charts



Cluster E

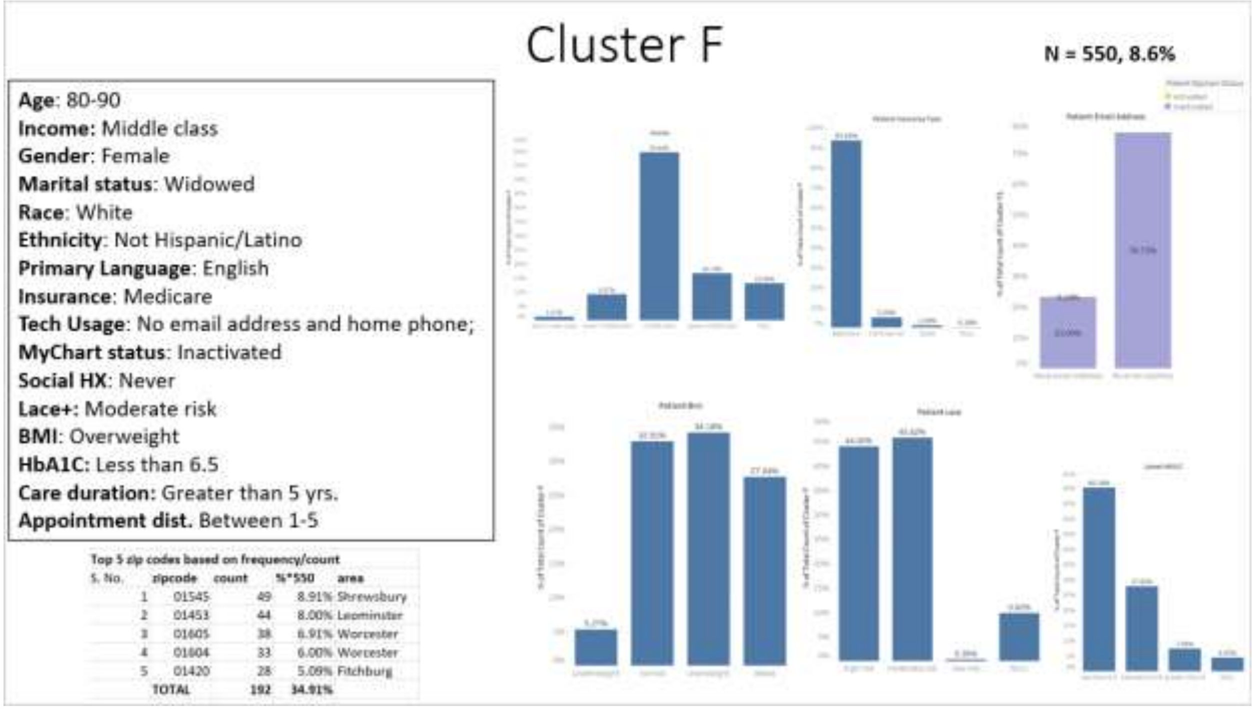
N = 1271, 20%

Age: 60-70
Income: Middle class
Gender: Female
Marital status: Married
Race: White
Ethnicity: Not Hispanic/Latino
Primary Language: English
Insurance: Medicare
Tech Usage: Has email address and home phone;
MyChart status: Activated
Social HX: Never had drugs, former smoker, currently consumes alcohol
Lace+: Low risk
BMI: Obese
HbA1C: Less than 6.5
Care duration: Greater than 5 yrs.
Appointment dist. Between 1-5

Top 5 zip codes based on frequency/count

S. No.	zipcode	count	%*1271	area
1	01545	70	5.51%	Shrewsbury
2	01604	69	5.43%	Worcester
3	01420	60	4.72%	Fitchburg
4	01605	58	4.56%	Worcester
5	01453	52	4.09%	Leominster
TOTAL		309	24.31%	

Figure 6 Cluster E characteristics and charts



Cluster F

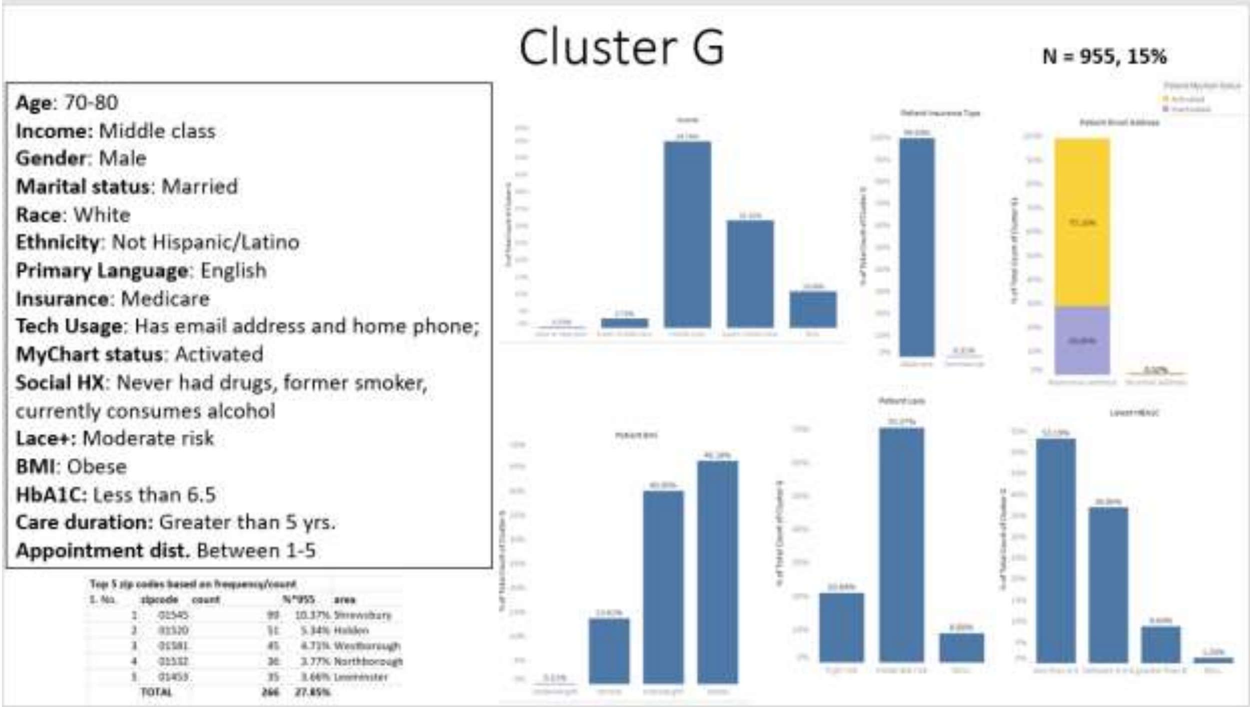
N = 550, 8.6%

Age: 80-90
Income: Middle class
Gender: Female
Marital status: Widowed
Race: White
Ethnicity: Not Hispanic/Latino
Primary Language: English
Insurance: Medicare
Tech Usage: No email address and home phone;
MyChart status: Inactivated
Social HX: Never
Lace+: Moderate risk
BMI: Overweight
HbA1C: Less than 6.5
Care duration: Greater than 5 yrs.
Appointment dist. Between 1-5

Top 5 zip codes based on frequency/count

S. No.	zipcode	count	%*550	area
1	01545	49	8.91%	Shrewsbury
2	01453	44	8.00%	Leominster
3	01605	38	6.91%	Worcester
4	01604	33	6.00%	Worcester
5	01420	28	5.09%	Fitchburg
TOTAL		192	34.91%	

Figure 7 Cluster F characteristics and basic charts



Cluster G

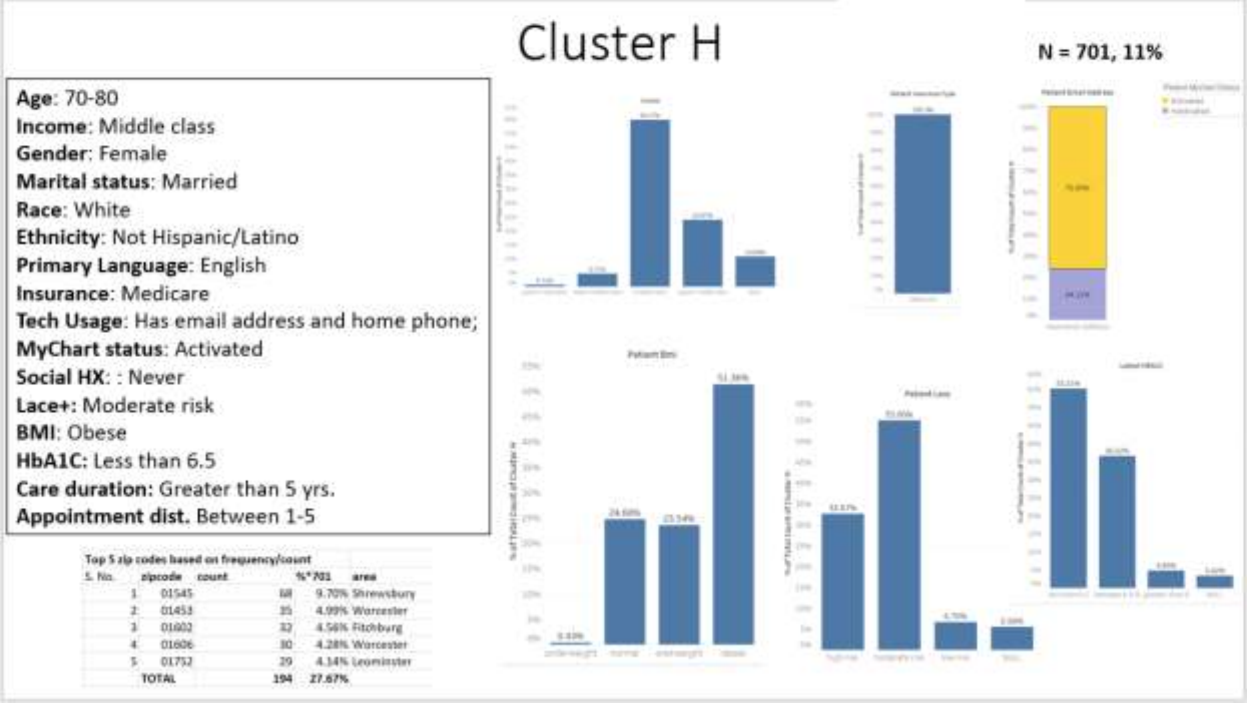
N = 955, 15%

Age: 70-80
Income: Middle class
Gender: Male
Marital status: Married
Race: White
Ethnicity: Not Hispanic/Latino
Primary Language: English
Insurance: Medicare
Tech Usage: Has email address and home phone;
MyChart status: Activated
Social HX: Never had drugs, former smoker, currently consumes alcohol
Lace+: Moderate risk
BMI: Obese
HbA1C: Less than 6.5
Care duration: Greater than 5 yrs.
Appointment dist. Between 1-5

Top 5 zip codes based on frequency/count

S. No.	zipcode	count	%*955	area
1	01545	99	10.37%	Shrewsbury
2	01520	51	5.34%	Holden
3	01581	45	4.71%	Westborough
4	01532	36	3.77%	Northborough
5	01453	35	3.66%	Leominster
	TOTAL	266	27.85%	

Figure 8 Cluster G characteristics and basic charts



Cluster H

N = 701, 11%

Age: 70-80
Income: Middle class
Gender: Female
Marital status: Married
Race: White
Ethnicity: Not Hispanic/Latino
Primary Language: English
Insurance: Medicare
Tech Usage: Has email address and home phone;
MyChart status: Activated
Social HX: : Never
Lace+: Moderate risk
BMI: Obese
HbA1C: Less than 6.5
Care duration: Greater than 5 yrs.
Appointment dist. Between 1-5

Top 5 zip codes based on frequency/count

S. No.	zipcode	count	%*701	area
1	01545	68	9.70%	Shrewsbury
2	01453	35	4.99%	Worcester
3	01602	32	4.56%	Fitchburg
4	01606	30	4.28%	Worcester
5	01752	29	4.14%	Leominster
	TOTAL	194	27.67%	

Figure 9 Cluster H characteristics and basic charts

Data and Proto personas

In a prior project, 8 proto personas were developed for the same ACO population that is considered in this current project. Proto personas, which are assumption based, are often refined/validated with user or data personas (Djamasbi &Strong; 2019; Jain, et al., 2019). A major objective of the current study as outlined in the research question was to use the data artifact generated in the usage world (patient records) so that it can be used to verify and refine the data generated in the design world (proto personas) (Djamasbi & Strong, 2019). Cluster analysis in this project provided the basis for developing a set of initial data personas. Hence, data personas and proto personas were manually inspected and analyzed to identify similarities and differences between the two sets. Figure 10 shows an initial attempt at matching the two sets of personas. The green arrows indicate direct matches while the orange ones indicate those that were made using expert interpretation. It is important to point out the assumption-based nature of proto personas and the exploratory nature of data personas in this project. Hence, a proto persona in this project can have more than one data persona matches. In corollary, a data persona can match with more than one proto persona. Note that 2 of proto personas did not have a match in the data persona as are 2 of the data personas without any match. Again, due to the nature of data and proto personas this outcome is not unusual.

This initial attempt has led to valuable insights which will pave the way for next steps which are discussed in the upcoming chapters.

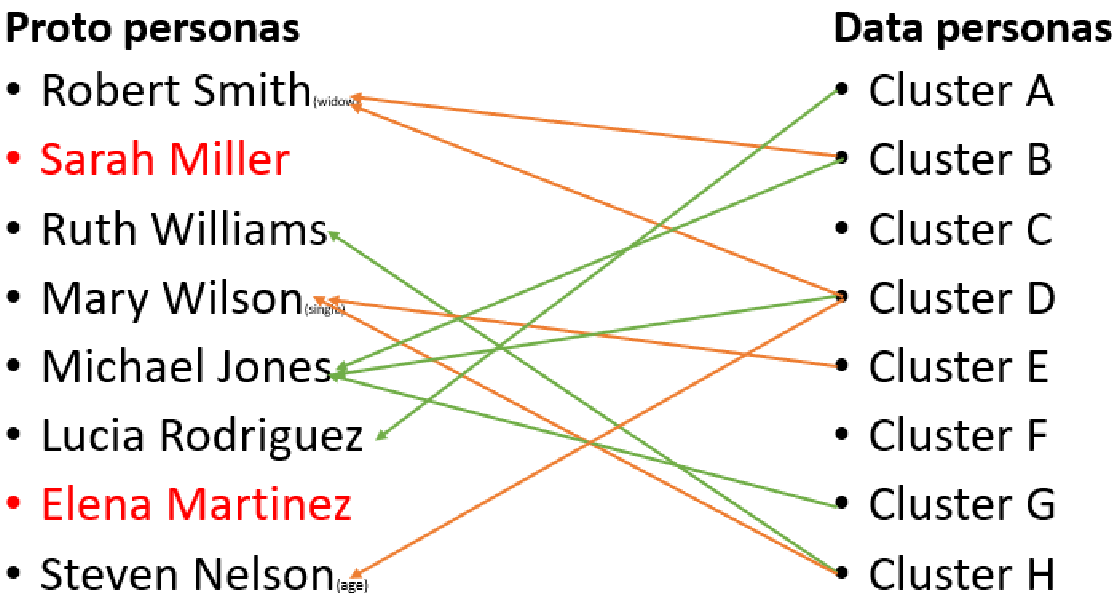


Figure 10 Initial attempt at validating proto personas and data personas

Discussion & Broader Impacts

This chapter includes insights for a broader application of the current study. As discussed in the previous chapter, initial attempts at matching the proto personas and data personas have led to important insights which will guide the next steps in persona development.

The two proto personas, Sarah Miller, and Elena Martinez, who belong to a younger age group do not have any data persona matches. This is because while the experts in the proto persona development project presented younger age groups of patients as two personas, the younger patient groups were not represented by the data personas. The two proto personas that are missing from the initial set of data personas, represent patients from a minority community (Hispanic/Latino). These observations have important implications for using patient records data for persona creation. Any data-driven decision making, whether manual or automated, must be free of gender and racial bias (Keyes, et al., 2019). Data when run by algorithms provides only limited information, which are generally skewed towards the majority of cases in the data set unless specified otherwise. Such an algorithmic approach results in underrepresenting cases that are fewer in number in the data set, hence, introducing bias in the analysis. A possible solution could be to ensure that, when possible, at the data collection stage, data is collected in the same proportion from all groups that are being represented. Another possible solution, as our analysis shows, is to use proto personas in conjunction with data personas. Because proto personas are developed based on remembering or envisioning a user population, they are likely to be more diverse in representation. In fact, generating diverse cases in design thinking methodology is often encouraged because it facilitates imagining possible boundaries of a project. Hence using proto personas in conjunction with data personas can help to identify biases in patient representation in the dataset. In corollary, data personas as well help to understand what is not being captured by the proto personas indicating that there are existing user groups that are not captured in the immediate memory of the decision makers.

In the current study, it has been observed that there is an almost equal representation of the two genders, male and female. The other sub-category in gender is ‘Unknown’. While this may seem fair on the surface, today, there exists so many gender categories. Missing out on the rest of the sub-categories, especially in the context of healthcare, may prove detrimental to fair decision making in the long run. Having the term ‘Unknown’ does not solve for the problem and this reflects on the need for further improvements in data collection.

The upcoming chapter delves into the limitations of the current study and the opportunities it opens up for future work.

Limitations and Future Work

Deidentified census tract IDs had been extracted for neighborhoods along with median household income. However, these IDs require GIS software for further analysis of socio-economic status within a particular zip code. Future work can undertake analysis of these details to aid health officials and the local government in ensuring better access to care for socio-economically disadvantaged neighborhoods. Furthermore, these IDs can also be used to analyze access to care and subsequent problems in transportation (an issue that was discussed in proto personas). Transportation has not been touched in the current study for two main reasons. One, the covid-19 pandemic has made telehealth the only option, especially in cases such as long-term care of chronic diseases. So, the current study has focused more on access to technology. Two, with time, people's addresses change, and this may or may not be reflected in the database. Therefore, assessing their distance from the hospital may or may not produce fruitful results. However, challenges to access has been frequently represented by the proto personas for this population.

On a similar note, qualitative aspects of proto personas such as a patient's management of medication or desiring a social life are difficult to capture in data personas. The Epic database does not have variables that monitor a patient's management of his/her medicines on a daily basis. Moreover, having a social life is a proven indicator for better overall health. However, these patient needs and challenges, though identified by proto personas, are difficult to capture quantitatively with data personas. Attitudinal data are best captured through one-on-one interviews during user personas development (Jain, et al., 2019). However, advances in data science approaches such as text mining may also help provide insight from patient comments and feedback to understand patient emotional needs and social challenges.

The limitations of data personas in providing insight for attitudinal and subjective perceptions of patients can be addressed in future studies that focus on the development of user personas which are developed through direct contact with patients. User personas, which are considered as generative UX research, are created by conducting unstructured or semi structured interviews with patients.

Another limitation observed in this project revolves around the variable ‘income’ which is an important social determinant of health. The current breakdown of income categories is applicable to a three-person household. However, the data regarding the number of members in the patients’ household is not available in the dataset, which can have a significant impact on the assessment of income category. For example, if a patient has currently been classified as belonging to upper middle class based on the 3-person household classification but the patient’s household consists of 6 members, then on average, he/she will fall in middle-middle or lower-middle socio-economic strata.

In this project, the exploratory cluster analysis used a large set of variables to identify general/macro patterns in the data set represented by clusters. Within each cluster, geo mapping and basic charts were used to provide insight for the distribution of variables that impact health and wellness. Future projects can refine this analysis by looking into specific micro patterns (e.g., high burden of disease) which can be uncovered by creating clusters using a smaller subset of variables (e.g., A1C, LACE+, and BMI).

On a concluding note, data personas are generated purely using data stored in IT systems. For commercial purposes, researchers have combined data from social media sites such as YouTube, Facebook, and Twitter to generate personas in real-time (An, et al., 2016). In this project, cluster analysis was conducted manually to create data personas. Future projects can create software that can conduct this analysis automatically in real-time to create data personas, thus enabling easy access to the most up-to-date data personas (Djamasbi & Strong, 2019). However, designers and developers must be mindful of the concept of autonomy in data-driven decision making (IEEE, 2018) especially in the context of healthcare where the interaction is more complex owing to the fact that these decisions have a direct impact on human life.

References

The Trustees of the University of Pennsylvania. (n.d.). *Data Warehousing: Epic Clarity*. Retrieved from Penn Medicine: Data Analytics Center: https://www.med.upenn.edu/dac/epic-clarity-data-warehousing.html

American Hospital Association. (n.d.). *Population Health Management*. Retrieved from AHA Center for Health Innovation: https://www.aha.org/center/population-health-management

An, J., Cho, H., Kwak, H., Hassen, M. Z., & Jansen, B. J. (2016). Towards Automatic Persona Generation Using Social Media. *2016 IEEE 4th International Conference on Future Internet of Things and Cloud Workshops (FiCloudW)* (pp. 206-211). IEEE.

Brownlee, J. (2020, August 19). *Blog: Python Machine Learning.* Retrieved from Machine Learning Mastery: https://machinelearningmastery.com/distance-measures-for-machine-learning/

CDC. (2010, October 22). *Media: Page Release*. Retrieved from cdc.gov: https://www.cdc.gov/media/pressrel/2010/r101022.html

Centers for Disease Control and Prevention. (2020, September 17). *Body Mass Index (BMI) | Healthy Weight, Nutrition and Physical Activity*. Retrieved from cdc.gov: https://www.cdc.gov/healthyweight/assessing/bmi/index.html

Centers for Medicare and Medicaid Services. (2021, January 14). *ACOs*. Retrieved from CMS.gov: https://innovation.cms.gov/innovation-models/aco

Congressional Caucus on Diabetes. (n.d.). *About Diabetes: Facts and Figures*. Retrieved from diabetescaucus-degette.house.gov: https://diabetescaucus-degette.house.gov/facts-and-figures#:~:text=Every%2017%20seconds%2C%20another%20individual,will%20be%20diagnosed%20this%20year.

Djamasbi, S., & Strong, D. (2019). User Experience-driven Innovation in Smart and Connected Worlds. *AIS Transactions on Human-Computer Interaction*, 215-231.

Eckel, R. H., Kahn, S. E., Ferrannini, E., Goldfine, A. B., Nathan, D. M., Schwartz, M. W., . . . Smith, S. R. (2011). Obesity and Type 2 Diabetes: What Can Be Unified and What Needs to Be Individualized? *Diabetes Care*, 1424-1430.

Epic EHR. (n.d.). Retrieved from EHR in Practice: https://www.ehrinpractice.com/epic-ehr-software-profile-119.html

Freidman, B., & Hendry, D. G. (2019). *Value Sensitive Design: Shaping Technology with Moral Imagination.* MIT Press.

Huh, J., Kwon, B. C., Kim, S.-H., Lee, S., Choo, J., Kim, J., . . . Yi, J. S. (2016). Personas in online health communities. *Journal of Biomedical Informatics*, 212-225.

IBM. (2021). *SPSS Statistics: SaaS-TwoStep Cluster Analysis*. Retrieved from IBM: https://www.ibm.com/docs/en/spss-statistics/SaaS?topic=features-twostep-cluster-analysis

IBM. (n.d.). *Support: IBM*. Retrieved from IBM: https://www.ibm.com/support/pages/how-log-likelihood-distance-method-applied-twostep-cluster-analysis

IEEE. (2018). *Ethically Aligned Design.* IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems.

Jain, P., Djamasbi, S., & Wyatt, J. (2019). Creating Value with Proto-Research Persona Development. In *HCI in Business, Government and Organizations. Information Systems and Analytics* (pp. 72-82). Springer International Publishing.

Keyes, O., Hutson, J., & Durbin, M. (2019). A Mulching Proposal: Analysing and Improving an Algorithmic System for Turning the Elderly into High-Nutrient Slurry. *Proceedings of ACM CHI Conference on Human Factors in Computing Systems (CHI 2019).* New York: ACM.

Kindig, D., & Stoddart, G. (2003). What Is Population Health? *American Journal of Public Health*, 380-383.

Lee, C.-Y. (2020). A Two-step Clustering Approach for Measuring Socioeconomic Factors Associated with Cardiovascular Health among Older Adults in South Korea. *Korean Journal of Adult Nursing*, 551-559.

Lin, J., Thompson, T. J., Cheng, Y. J., Zhuo, X., Zhang, P., Gregg, E., & Rolka, D. B. (2018). Projection of the future diabetes burden in the United States through 2060. *Population Health Metrics*.

Manzella, D. (2020, September 27). *Type 2 Diabetes: Comorbid Conditions and Diabetes*. Retrieved from verywell health: https://www.verywellhealth.com/comorbidity-disease-diabetes-1087365

McGinn, J. S., & Kotamraju, N. P. (2008). Data-driven persona development. *Proceeding of the twenty-sixth annual CHI conference.*

Medicare Interactive. (n.d.). *Differences between Medicare and Medicaid*. Retrieved from MedicareInteractive.org: https://www.medicareinteractive.org/get-answers/medicare-basics/medicare-coverage-overview/differences-between-medicare-and-medicaid

Norusis. (2008). *SPSS 16.0 Guide to Data Analysis, 2nd Edition.* Pearson.

Norusis, M. J. (2011). *IBM SPSS Statistics 19 Guide to Data Analysis.*

Office of Disease Prevention and Health Promotion. (2020). *Social Determinants of Health*. Retrieved from HealthyPeople.gov: https://www.healthypeople.gov/2020/topics-objectives/topic/social-determinants-of-health

Persons, B., Jain, P., Chagnon, C. J., & Djamasbi, S. (forthcoming). Designing the Empathetic Research IOT Network (ERIN) Chatbot for Mental Health Resources. *Conference on HCI in Business, Government and Organizations.*

Snider, S. (2020, December 8). Where Do I Fall in the American Economic Class System? *U.S. News & World Report*.

SUN, D., ZHOU, T., LI, X., HEIANZA, Y., SHANG, X., FONSECA, V., & QI, L. (2018). Twenty-Year Secular Trend of Diabetes Comorbidities in the United States, 1997-2016. *Diabetes*.

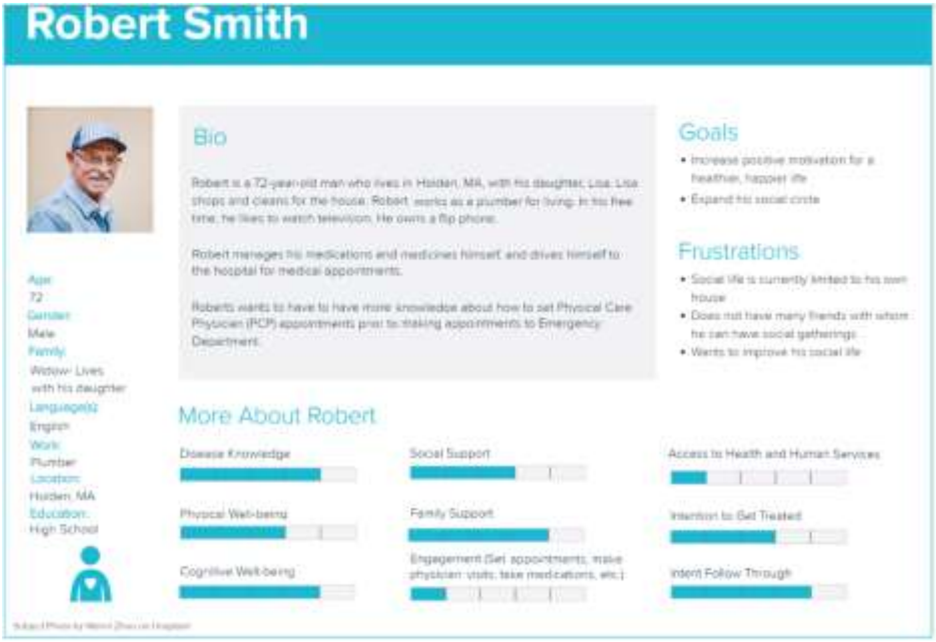
The Office of the National Coordinator for Health Information Technology (ONC). (2019, September 10). *What is an electronic health record (EHR)?* Retrieved from HealthIT.gov: https://www.healthit.gov/faq/what-electronic-health-record-ehr

UMass Memorial Health. (2021). *UMASS MEMORIAL MEDICARE ACCOUNTABLE CARE ORGANIZATION (ACO)*. Retrieved from UMass Memorial Health: https://www.ummhealth.org/umass-memorial-medicare-accountable-care-organization-aco

Weiss, A. P. (2017). *Morning CMO Report.* Upstate University Hospital.

Appendix A

Proto personas for the ACO population



Robert Smith

Age:
72
Gender:
Male
Family:
Widow- Lives with his daughter
Language(s):
English
Work:
Plumber
Location:
Holden, MA
Education:
High School

Bio

Robert is a 72-year-old man who lives in Holden, MA, with his daughter, Lisa. Lisa shops and cleans for the house. Robert works as a plumber for living. In his free time, he likes to watch television. He owns a flip phone.

Robert manages his medications and medicines himself and drives himself to the hospital for medical appointments.

Roberts wants to have to have more knowledge about how to set Physical Care Physician (PCP) appointments prior to making appointments to Emergency Department.

Goals

- Increase positive motivation for a healthier, happier life
- Expand his social circle

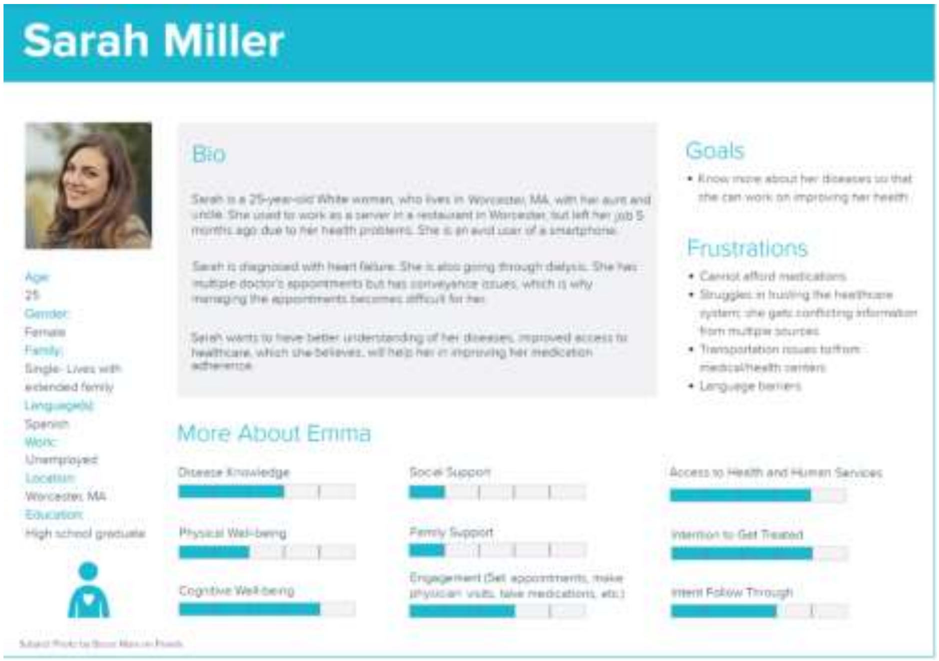
Frustrations

- Social life is currently limited to his own house
- Does not have many friends with whom he can have social gatherings
- Wants to improve his social life

More About Robert

Subject Photo by [illegible] on Unsplash

Figure A.1 Proto persona Robert Smith



Sarah Miller

Age:
25
Gender:
Female
Family:
Single- Lives with extended family
Language(s):
Spanish
Work:
Unemployed
Location:
Worcester, MA
Education:
High school graduate

Bio

Sarah is a 25-year-old White woman, who lives in Worcester, MA, with her aunt and uncle. She used to work as a server in a restaurant in Worcester, but left her job 5 months ago due to her health problems. She is an avid user of a smartphone.

Sarah is diagnosed with heart failure. She is also going through dialysis. She has multiple doctor's appointments but has conveyance issues, which is why managing the appointments becomes difficult for her.

Sarah wants to have better understanding of her diseases, improved access to healthcare, which she believes, will help her in improving her medication adherence.

Goals

- Know more about her diseases so that she can work on improving her health

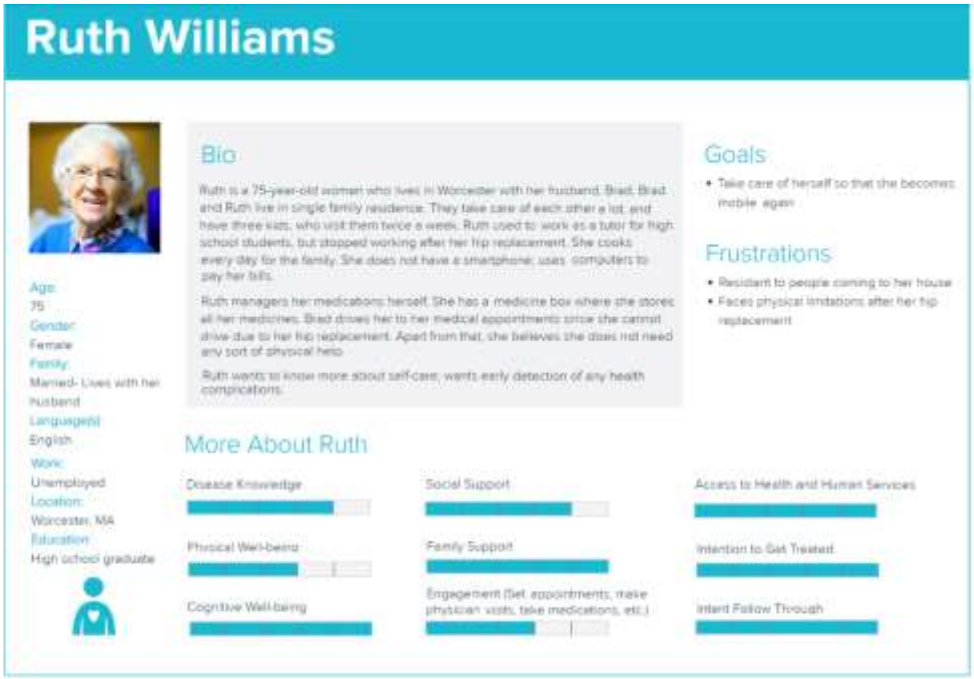
Frustrations

- Cannot afford medications
- Struggles in trusting the healthcare system; she gets conflicting information from multiple sources
- Transportation issues to/from medical/health centers
- Language barriers

More About Emma

Subject Photo by [illegible] on Pexels

Figure A.2 Proto persona Sarah Miller



Ruth Williams

Age:
75
Gender:
Female
Family:
Married- Lives with her husband
Language(s)
English
Work:
Unemployed
Location:
Worcester, MA
Education:
High school graduate

Bio

Ruth is a 75-year-old woman who lives in Worcester with her husband, Brad. Brad and Ruth live in single family residence. They take care of each other a lot, and have three kids, who visit them twice a week. Ruth used to work as a tutor for high school students, but stopped working after her hip replacement. She cooks every day for the family. She does not have a smartphone; uses computers to pay her bills.

Ruth managers her medications herself. She has a medicine box where she stores all her medicines. Brad drives her to her medical appointments since she cannot drive due to her hip replacement. Apart from that, she believes she does not need any sort of physical help.

Ruth wants to know more about self-care, wants early detection of any health complications.

Goals

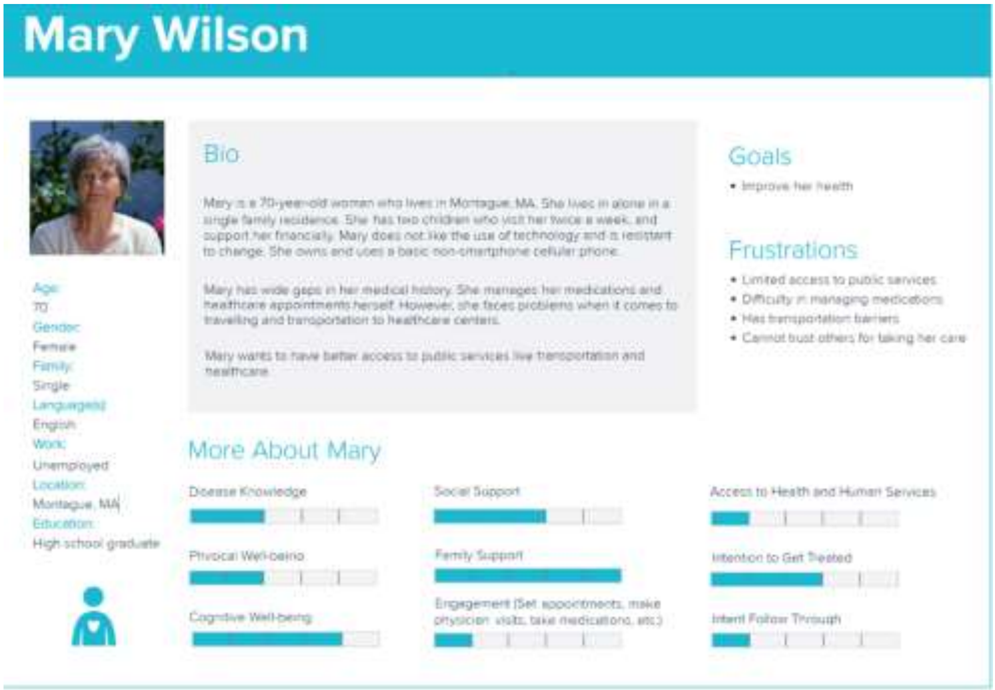
- Take care of herself so that she becomes mobile again

Frustrations

- Resident to people coming to her house
- Faces physical limitations after her hip replacement

More About Ruth

Figure A.3 Proto persona Ruth Williams



Mary Wilson

Age:
70
Gender:
Female
Family:
Single
Language(s)
English
Work:
Unemployed
Location:
Montague, MA
Education:
High school graduate

Bio

Mary is a 70-year-old woman who lives in Montague, MA. She lives in alone in a single family residence. She has two children who visit her twice a week, and support her financially. Mary does not like the use of technology and is resistant to change. She owns and uses a basic non-smartphone cellular phone.

Mary has wide gaps in her medical history. She manages her medications and healthcare appointments herself. However, she faces problems when it comes to travelling and transportation to healthcare centers.

Mary wants to have better access to public services live transportation and healthcare.

Goals

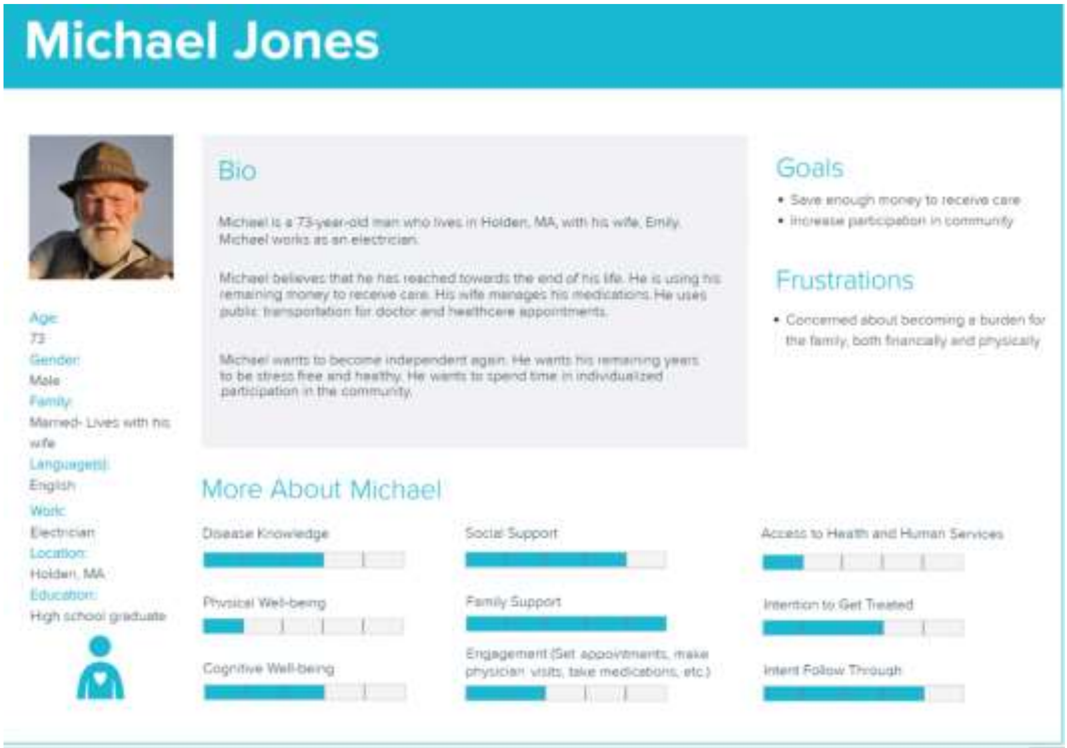
- Improve her health

Frustrations

- Limited access to public services
- Difficulty in managing medications
- Has transportation barriers
- Cannot bust others for taking her care

More About Mary

Figure A.4 Proto persona Mary Wilson



Michael Jones

Age:
73
Gender:
Male
Family:
Married- Lives with his wife
Language(s):
English
Work:
Electrician
Location:
Holden, MA
Education:
High school graduate

Bio

Michael is a 73-year-old man who lives in Holden, MA, with his wife, Emily. Michael works as an electrician.

Michael believes that he has reached towards the end of his life. He is using his remaining money to receive care. His wife manages his medications. He uses public transportation for doctor and healthcare appointments.

Michael wants to become independent again. He wants his remaining years to be stress free and healthy. He wants to spend time in individualized participation in the community.

Goals

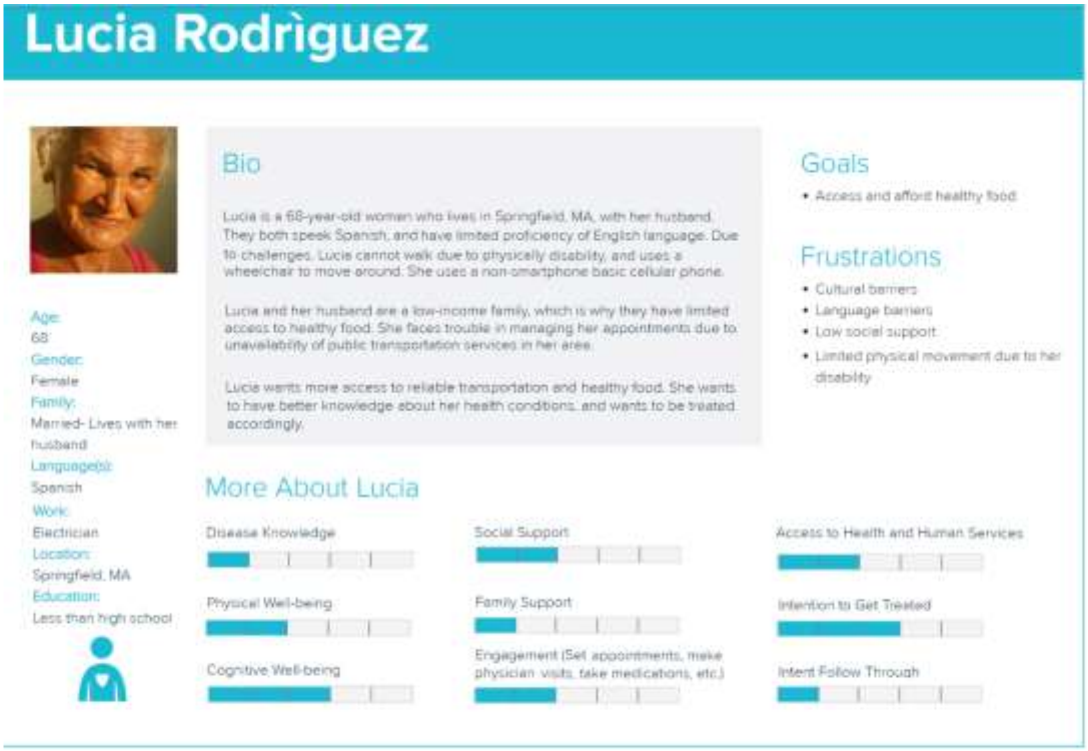
- Save enough money to receive care
- Increase participation in community

Frustrations

- Concerned about becoming a burden for the family, both financially and physically

More About Michael

Figure A.5 Proto persona Michael Jones



Lucia Rodrìguez

Age:
68
Gender:
Female
Family:
Married- Lives with her husband
Language(s):
Spanish
Work:
Electrician
Location:
Springfield, MA
Education:
Less than high school

Bio

Lucia is a 68-year-old woman who lives in Springfield, MA, with her husband. They both speak Spanish, and have limited proficiency of English language. Due to challenges, Lucia cannot walk due to physically disability, and uses a wheelchair to move around. She uses a non-smartphone basic cellular phone.

Lucia and her husband are a low-income family, which is why they have limited access to healthy food. She faces trouble in managing her appointments due to unavailability of public transportation services in her area.

Lucia wants more access to reliable transportation and healthy food. She wants to have better knowledge about her health conditions, and wants to be treated accordingly.

Goals

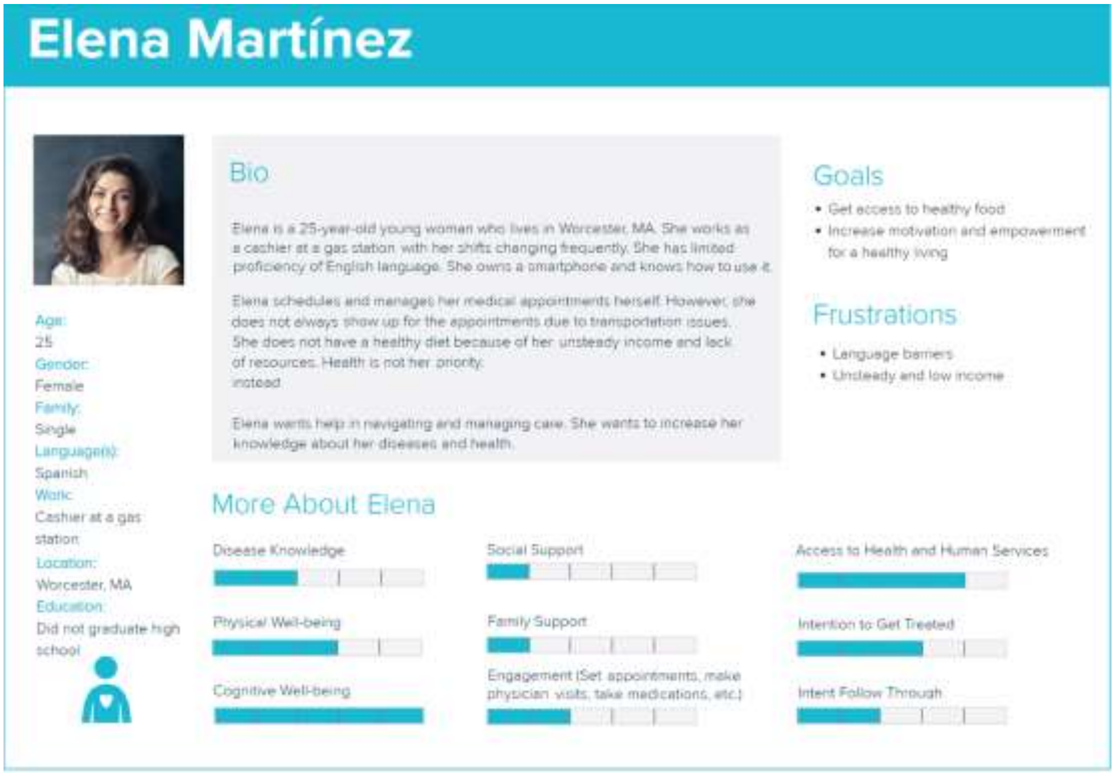
- Access and afford healthy food

Frustrations

- Cultural barriers
- Language barriers
- Low social support
- Limited physical movement due to her disability

More About Lucia

Figure A.6 Proto persona Lucia Rodriguez



Elena Martínez

Age:
25
Gender:
Female
Family:
Single
Language(s):
Spanish
Work:
Cashier at a gas station
Location:
Worcester, MA
Education:
Did not graduate high school

Bio

Elena is a 25-year-old young woman who lives in Worcester, MA. She works as a cashier at a gas station with her shifts changing frequently. She has limited proficiency of English language. She owns a smartphone and knows how to use it.

Elena schedules and manages her medical appointments herself. However, she does not always show up for the appointments due to transportation issues. She does not have a healthy diet because of her unsteady income and lack of resources. Health is not her priority.
instead

Elena wants help in navigating and managing care. She wants to increase her knowledge about her diseases and health.

Goals

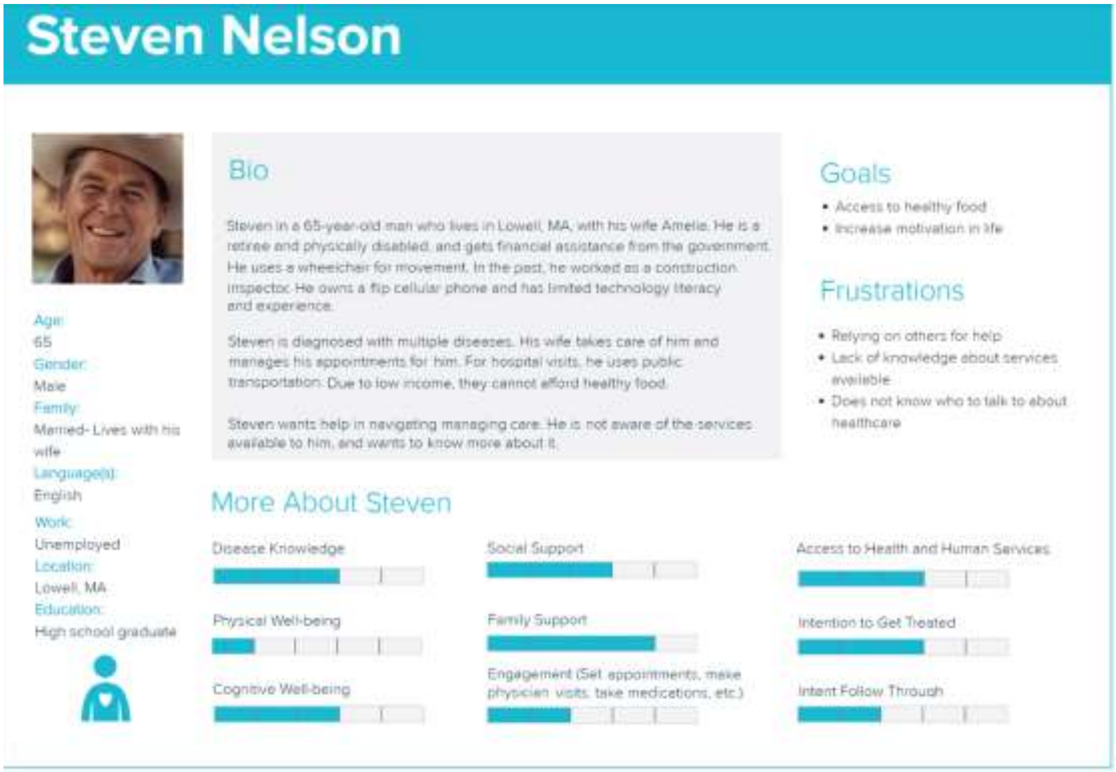
- Get access to healthy food
- Increase motivation and empowerment for a healthy living

Frustrations

- Language barriers
- Unsteady and low income

More About Elena

Figure A.7 Proto persona Elena Martinez



Steven Nelson

Age:
65
Gender:
Male
Family:
Married- Lives with his wife
Language(s):
English
Work:
Unemployed
Location:
Lowell, MA
Education:
High school graduate

Bio

Steven in a 65-year-old man who lives in Lowell, MA, with his wife Amelie. He is a retiree and physically disabled, and gets financial assistance from the government. He uses a wheelchair for movement. In the past, he worked as a construction inspector. He owns a flip cellular phone and has limited technology literacy and experience.

Steven is diagnosed with multiple diseases. His wife takes care of him and manages his appointments for him. For hospital visits, he uses public transportation. Due to low income, they cannot afford healthy food.

Steven wants help in navigating managing care. He is not aware of the services available to him, and wants to know more about it.

Goals

- Access to healthy food
- Increase motivation in life

Frustrations

- Relying on others for help
- Lack of knowledge about services available
- Does not know who to talk to about healthcare

More About Steven

Figure A.8 Proto persona Steven Nelson

Appendix B.1

Geo mapping and basic data analysis for all the records

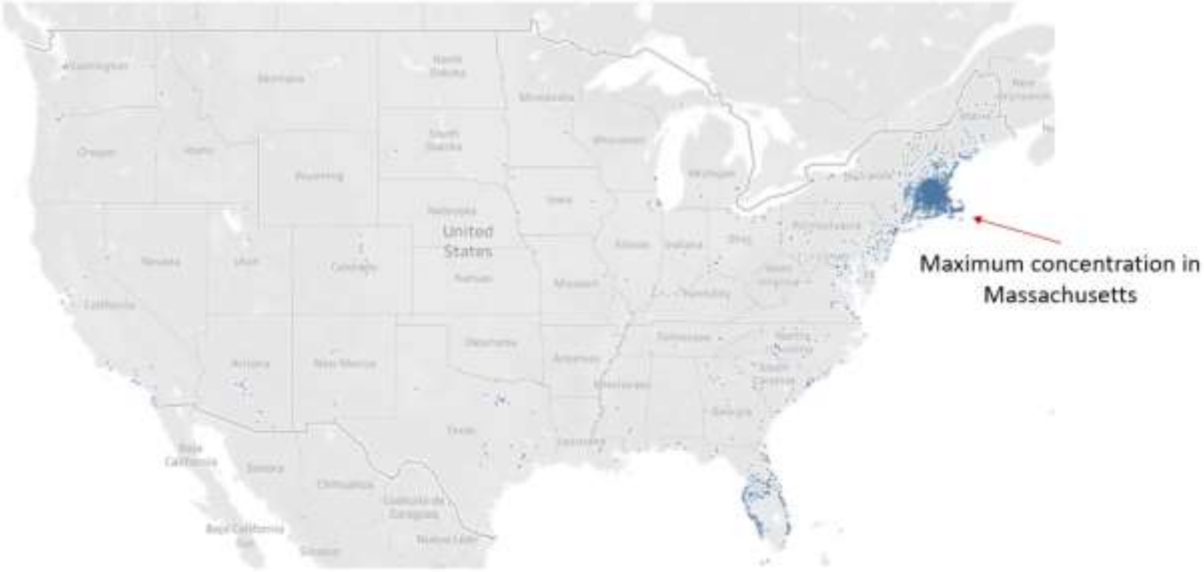


Figure B.1.1 Patient distribution across the North American continent

N = 53675

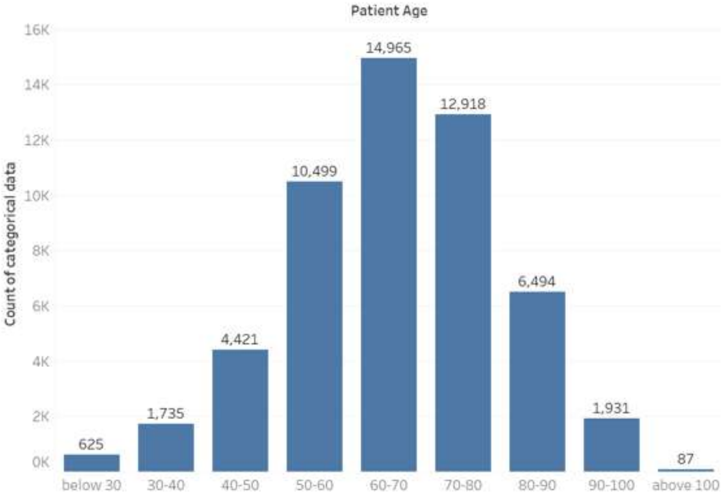


Figure B.1.2 Distribution of Patient Age

N = 53675

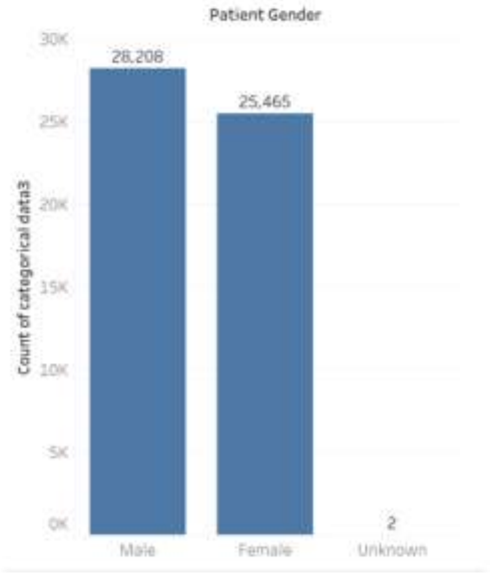


Figure B.1.3 Distribution of Patient Gender

N = 53675

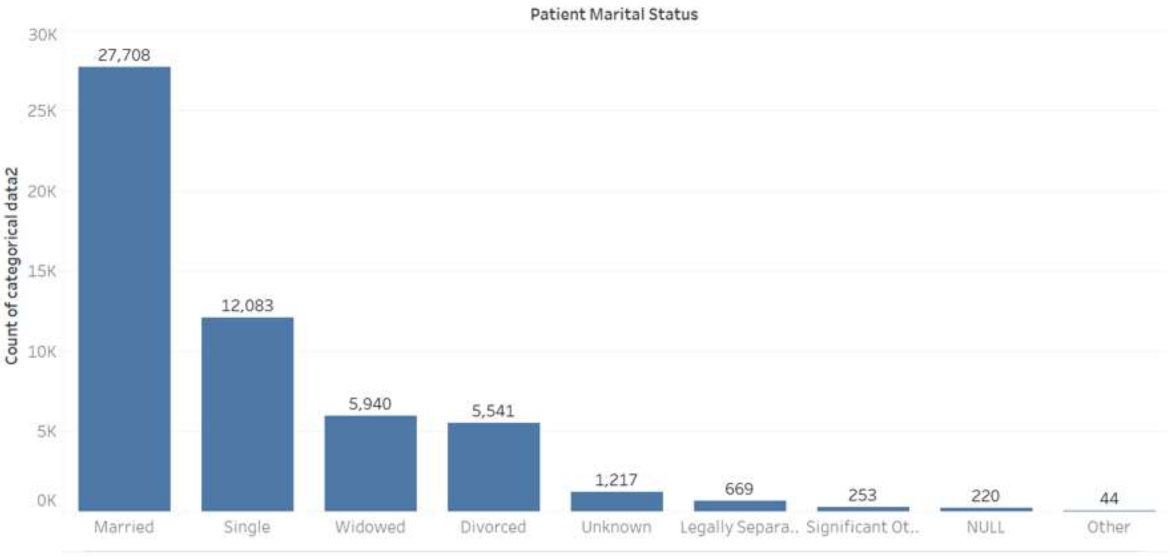
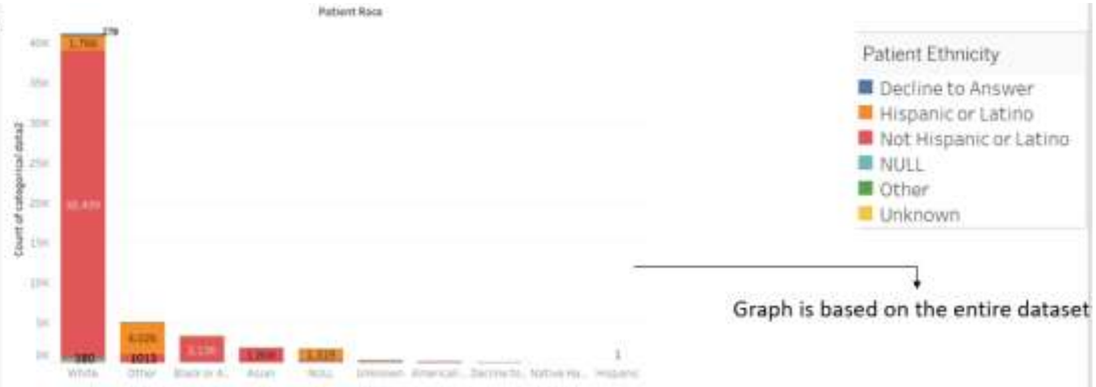


Figure B.1.4 Distribution of Patient Marital Status

N = 53675



Graph is based on the entire dataset

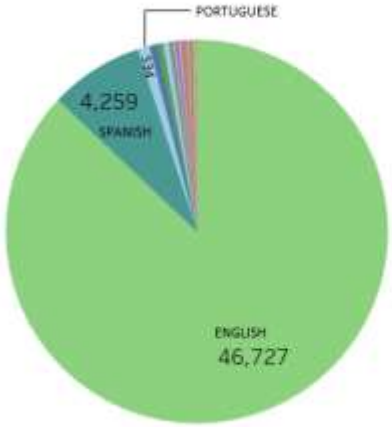
	Patient_Race										
Patient_Ethnicity	American Indian/AN	Asian	African American	Decline to answer	Hispanic	Native Hawaiian/OPI	NULL	Other	Unknown	White	Total
Decline to answer	3	18	26	13			29	37	3	279	408
Hispanic or Latino	10	18	138	34	1	7	1319	4026	130	1766	7449
Not Hispanic or Latino	131	1808	3136	19		5	218	1013	97	38499	44926
NULL		15	27				155	17	37	380	631
Other	1	8	6					1		59	75
Unknown	1	7	9				1	8	49	111	186
Total	**146**	**1874**	**3342**	**66**	**1**	**12**	**1722**	**5102**	**316**	**41094**	**53675**

Highlighted cells show the top 10 frequently occurring categories

Figure B.1.5 Distribution of Patient Race and Ethnicity

N = 53675

Pie chart depicts the entire dataset



Top 10 frequently occurring languages

English	46,727
Spanish	4,259
Portuguese	534
Vietnamese	349
Arabic	269
Albanian	261
NULL	125
Russian	85
Creole, Hatian	84
Chinese, Mandar..	83

87% of the patients speak English as their primary language, followed by Spanish at 8%

Figure B.1.6 Distribution of Patient Primary Language

N = 53314

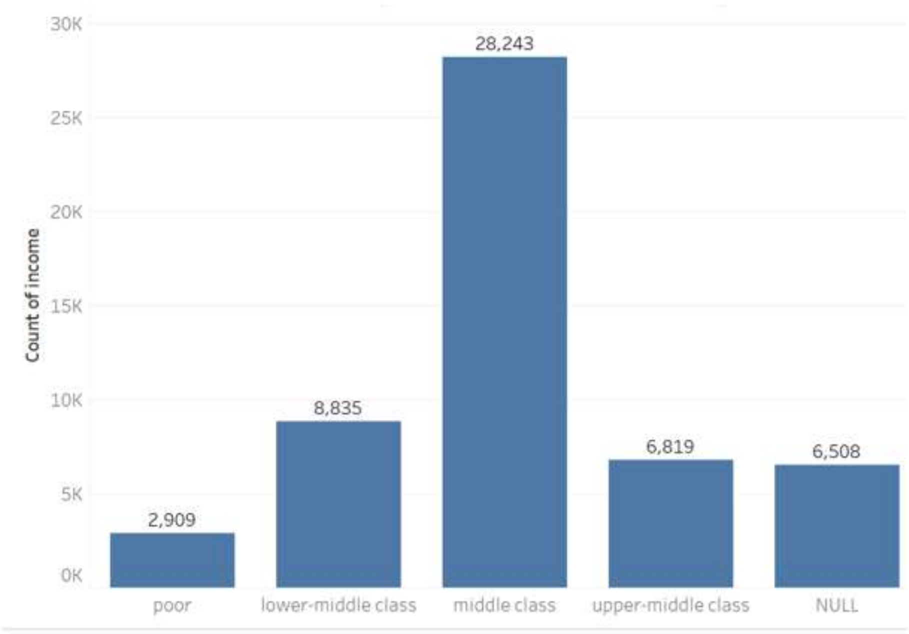


Figure B.1.7 Distribution of patient income levels

N = 53675

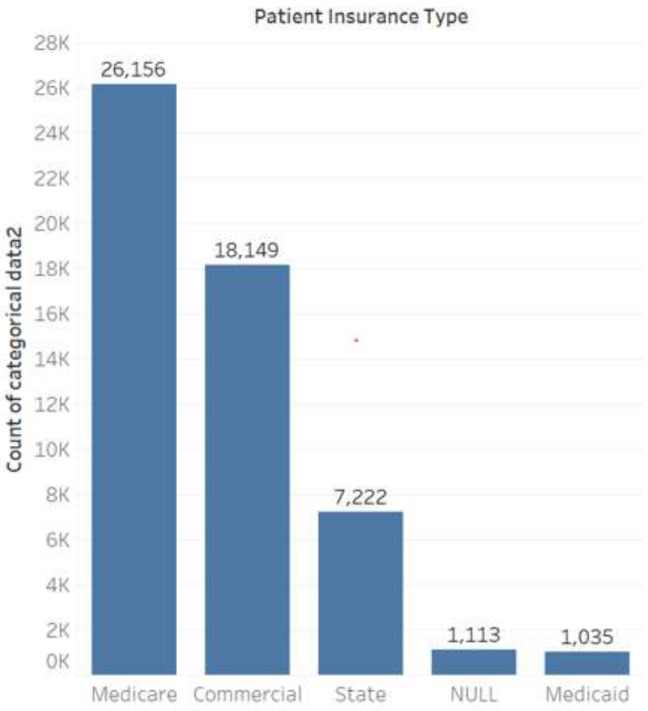


Figure B.1.8 Distribution of patient insurance types

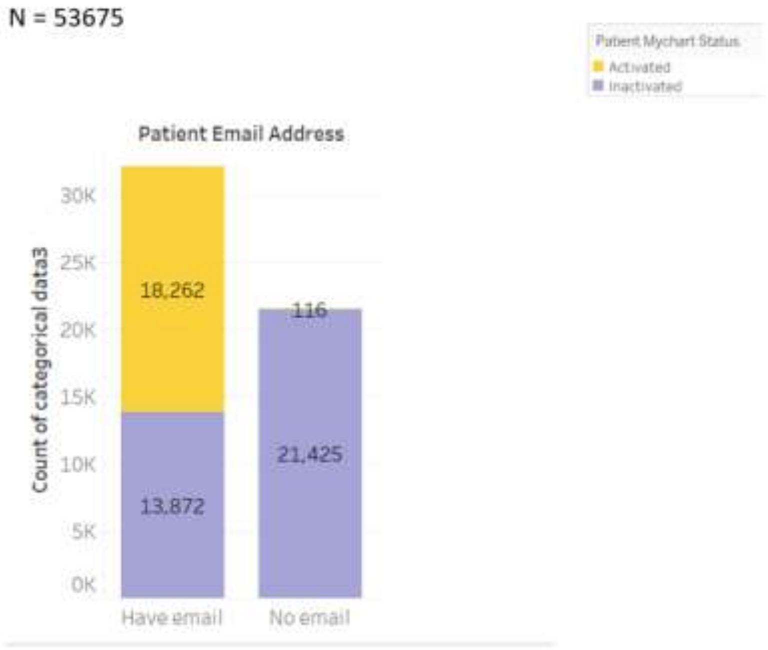


Figure B.1.9 Patient access to tech

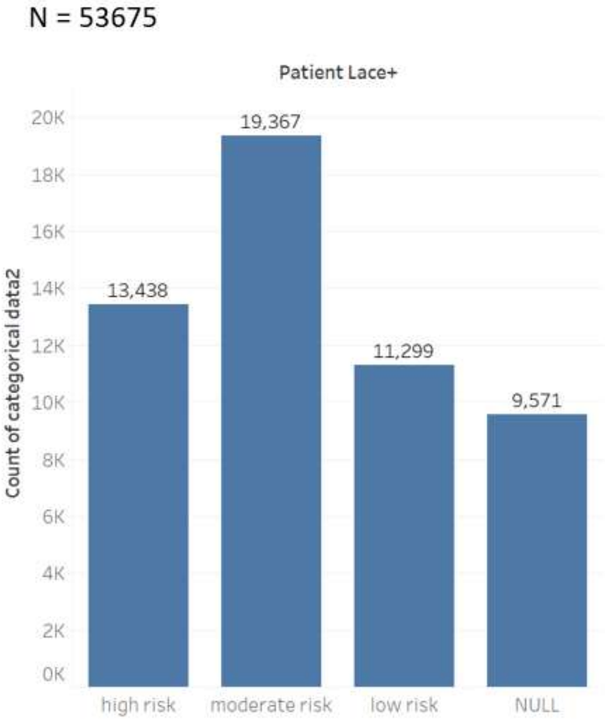


Figure B.1.10 Distribution of patient LACE+ scores

N = 53675

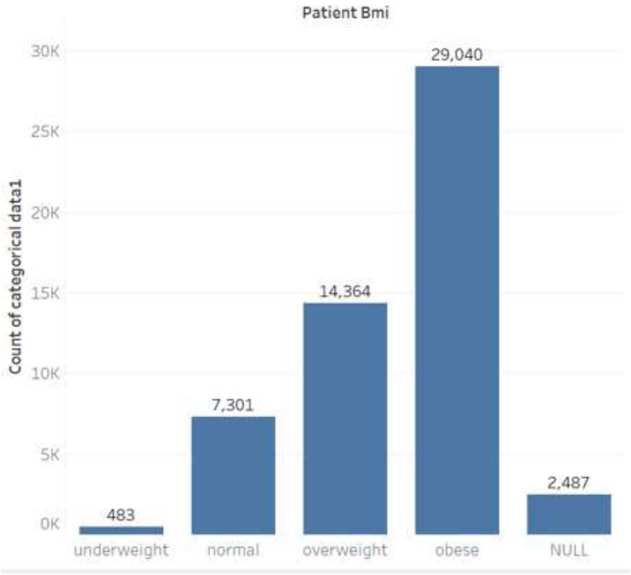
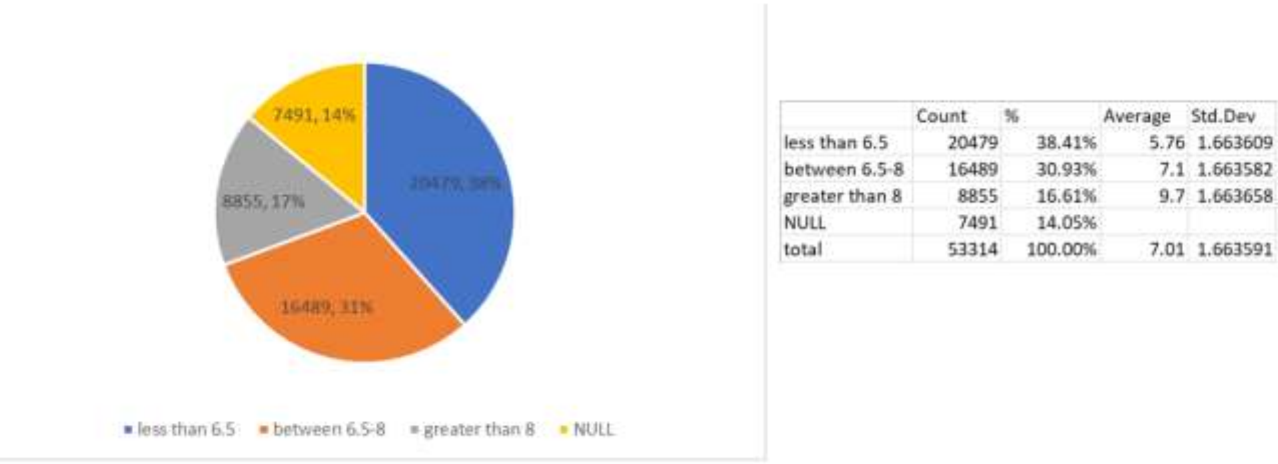


Figure B.1.11 Distribution of patient BMI

N = 53314



	Count	%	Average	Std.Dev
less than 6.5	20479	38.41%	5.76	1.663609
between 6.5-8	16489	30.93%	7.1	1.663582
greater than 8	8855	16.61%	9.7	1.663658
NULL	7491	14.05%		
total	53314	100.00%	7.01	1.663591

Figure B.1.12 Distribution of patient HbA1C levels

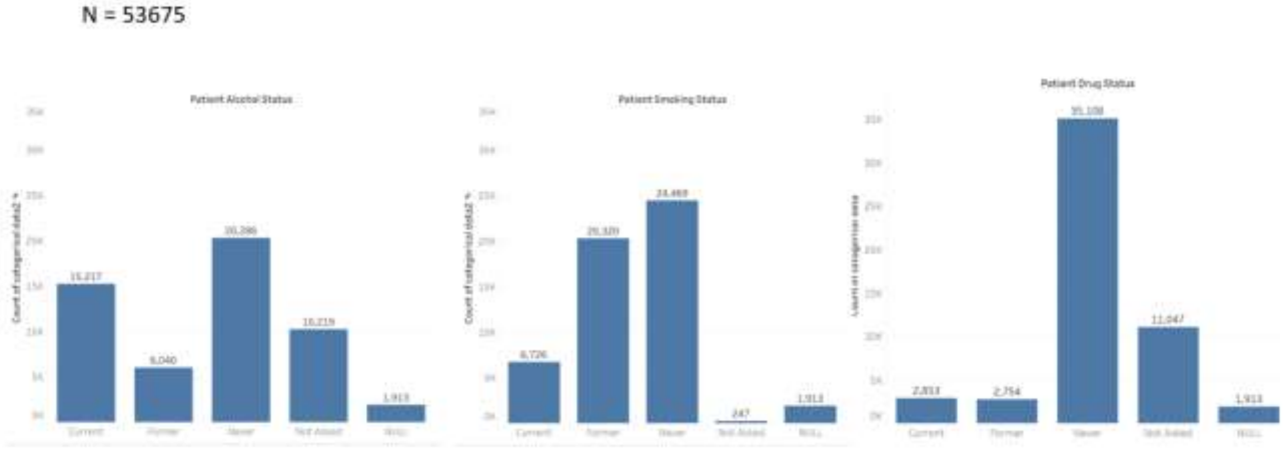


Figure B.1.13 Distribution of patient social_HX

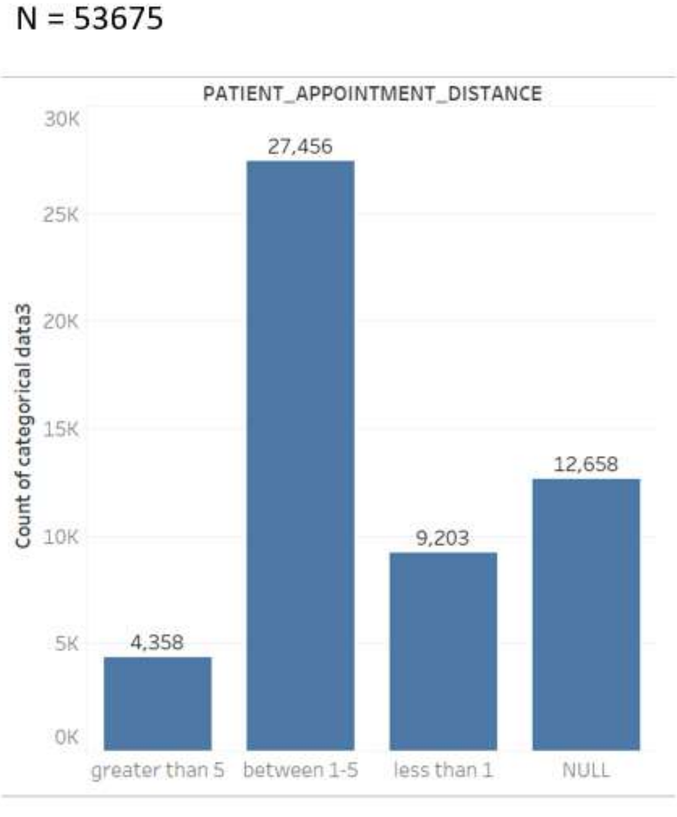


Figure B.1.14 Distribution of patient appointment distance

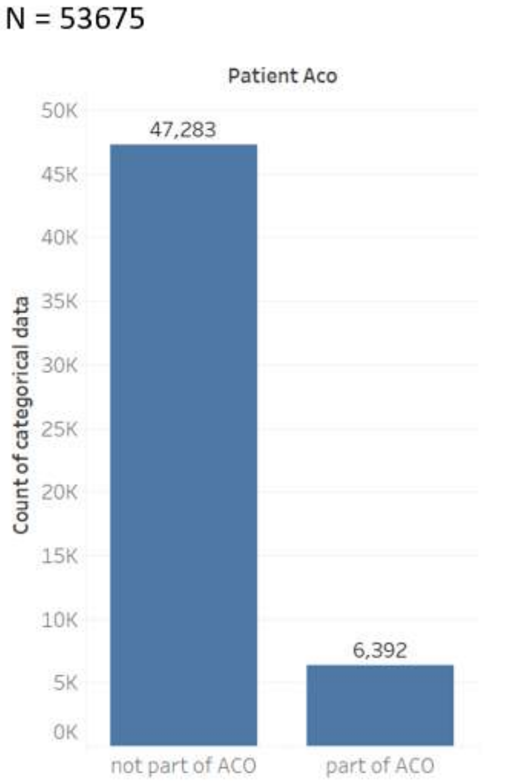
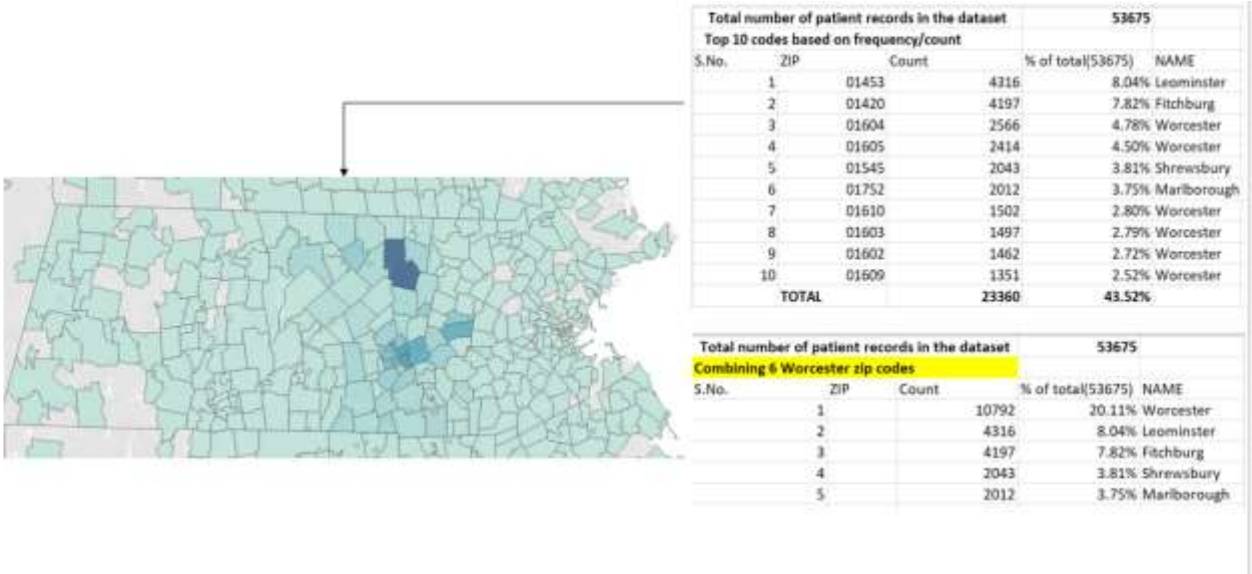


Figure B.1.15 Distribution of patients by ACO

Appendix B.2

Geo mapping and socio-economic analysis of all patient records



Total number of patient records in the dataset			53675	
Top 10 codes based on frequency/count				
S.No.	ZIP	Count	% of total(53675)	NAME
1	01453	4316	8.04%	Leominster
2	01420	4197	7.82%	Fitchburg
3	01604	2566	4.78%	Worcester
4	01605	2414	4.50%	Worcester
5	01545	2043	3.81%	Shrewsbury
6	01752	2012	3.75%	Marlborough
7	01610	1502	2.80%	Worcester
8	01603	1497	2.79%	Worcester
9	01602	1462	2.72%	Worcester
10	01609	1351	2.52%	Worcester
	TOTAL	23360	43.52%	

Total number of patient records in the dataset			53675	
Combining 6 Worcester zip codes				
S.No.	ZIP	Count	% of total(53675)	NAME
	1	10792	20.11%	Worcester
	2	4316	8.04%	Leominster
	3	4197	7.82%	Fitchburg
	4	2043	3.81%	Shrewsbury
	5	2012	3.75%	Marlborough

Figure B.2.1 Top 10 frequently occurring zip codes in Massachusetts

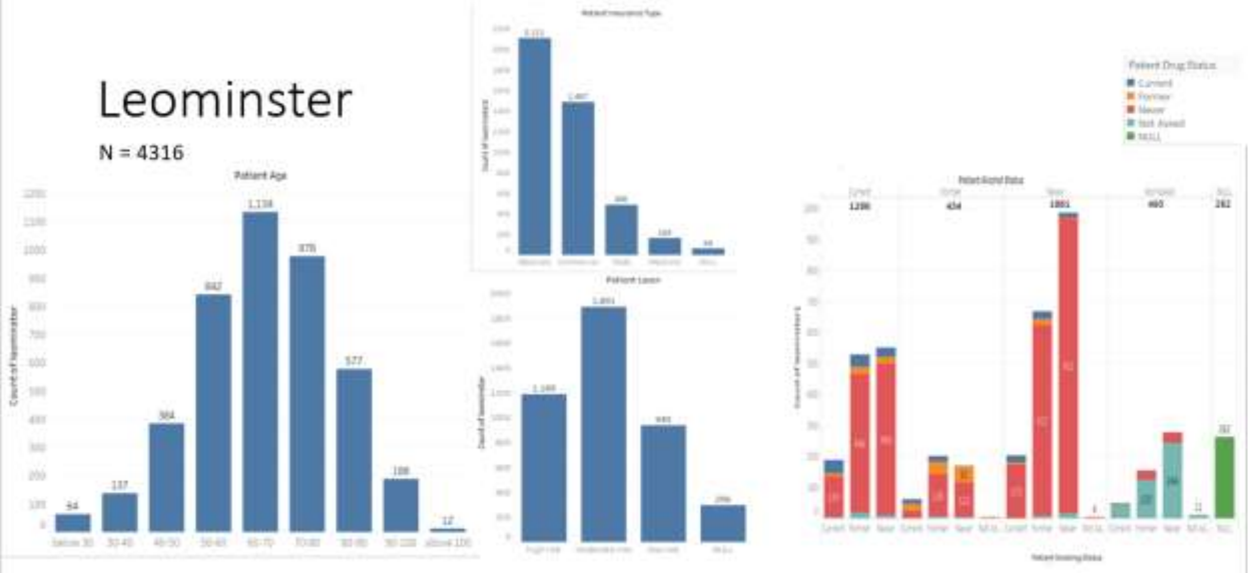


Figure B.2.2 Distribution of patient age, insurance type, LACE+ and social_HX in Leominster

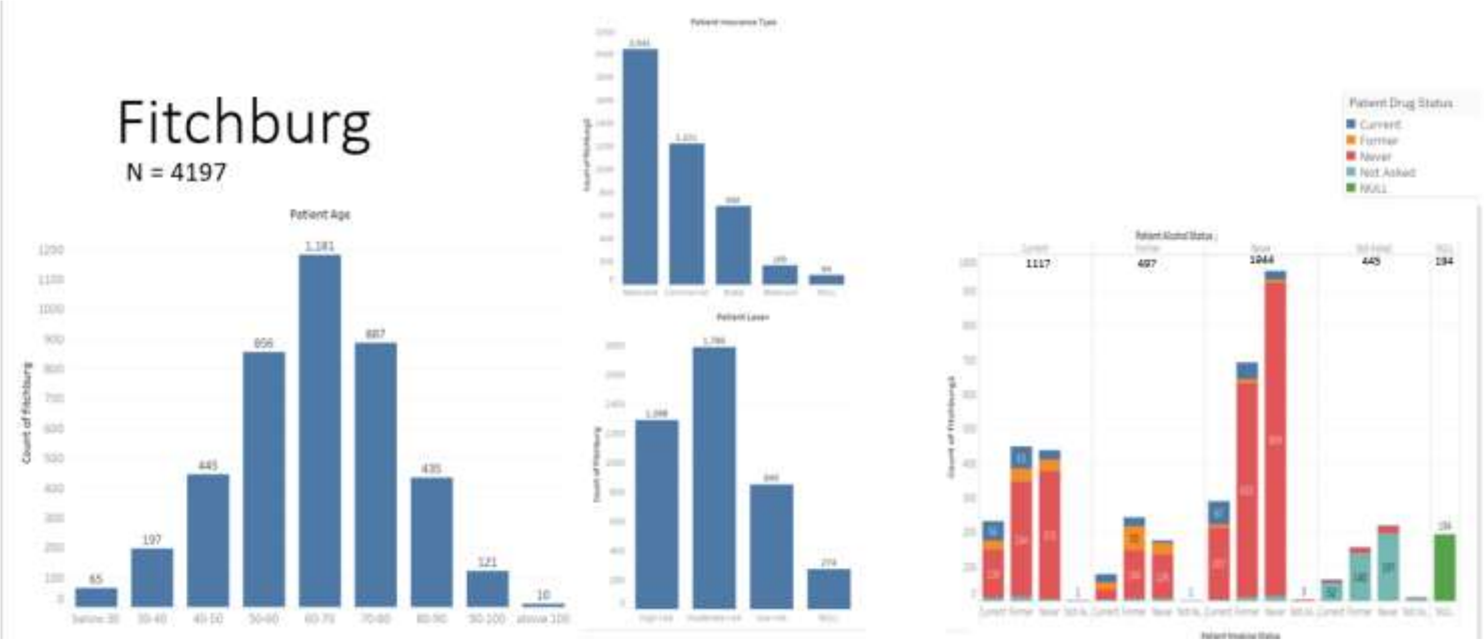


Figure B.2.3 Distribution of patient age, insurance type, LACE+ and social_HX in Fitchburg

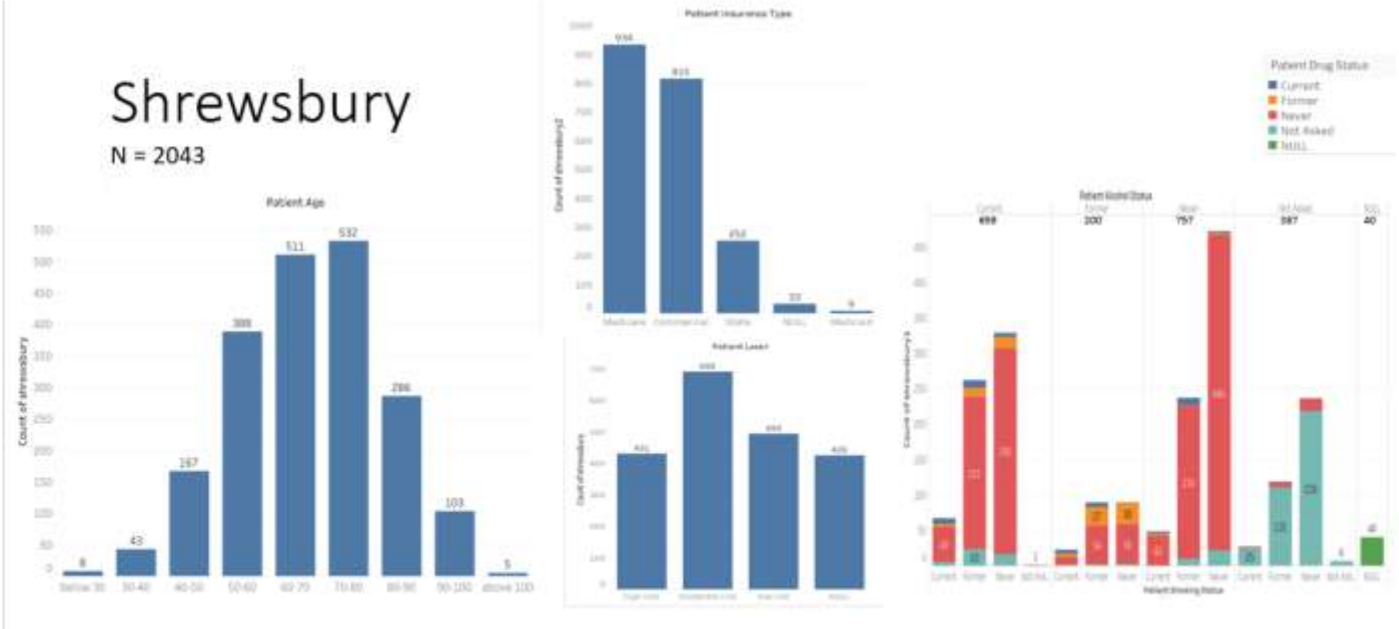


Figure B.2.4 Distribution of patient age, insurance type, LACE+ and social_HX in Shrewsbury

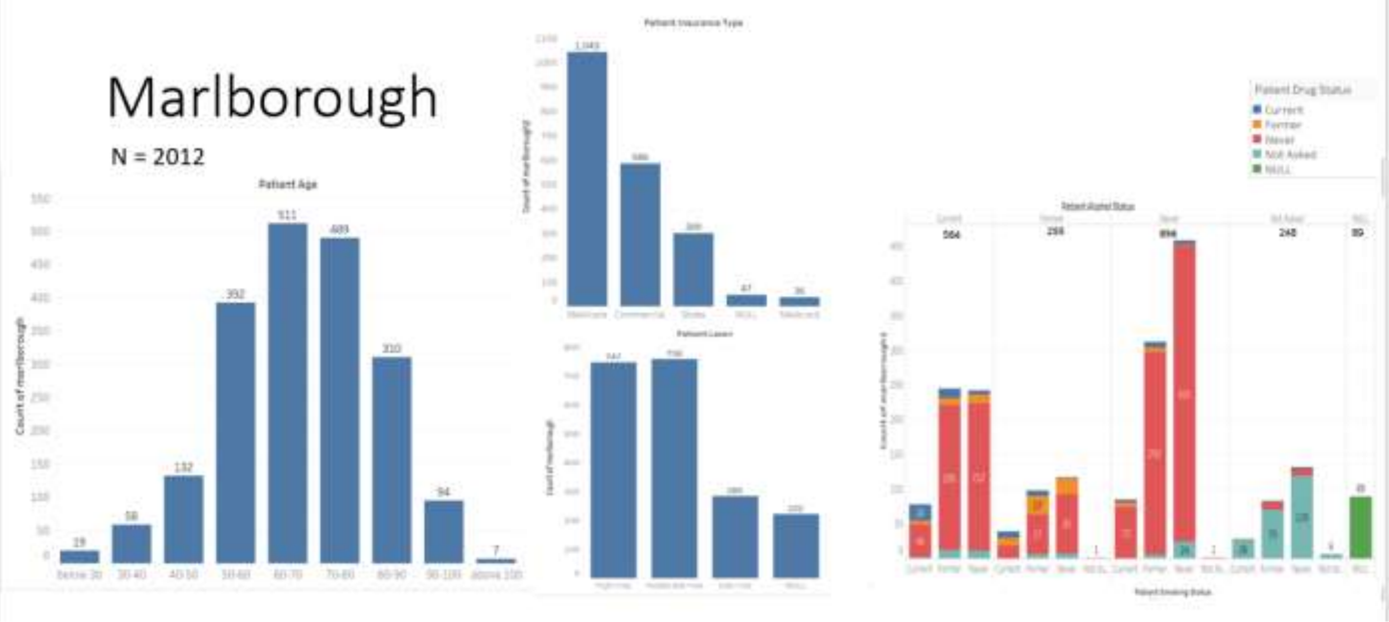


Figure B.2.5 Distribution of patient age, insurance type, LACE+ and social_HX in Marlborough

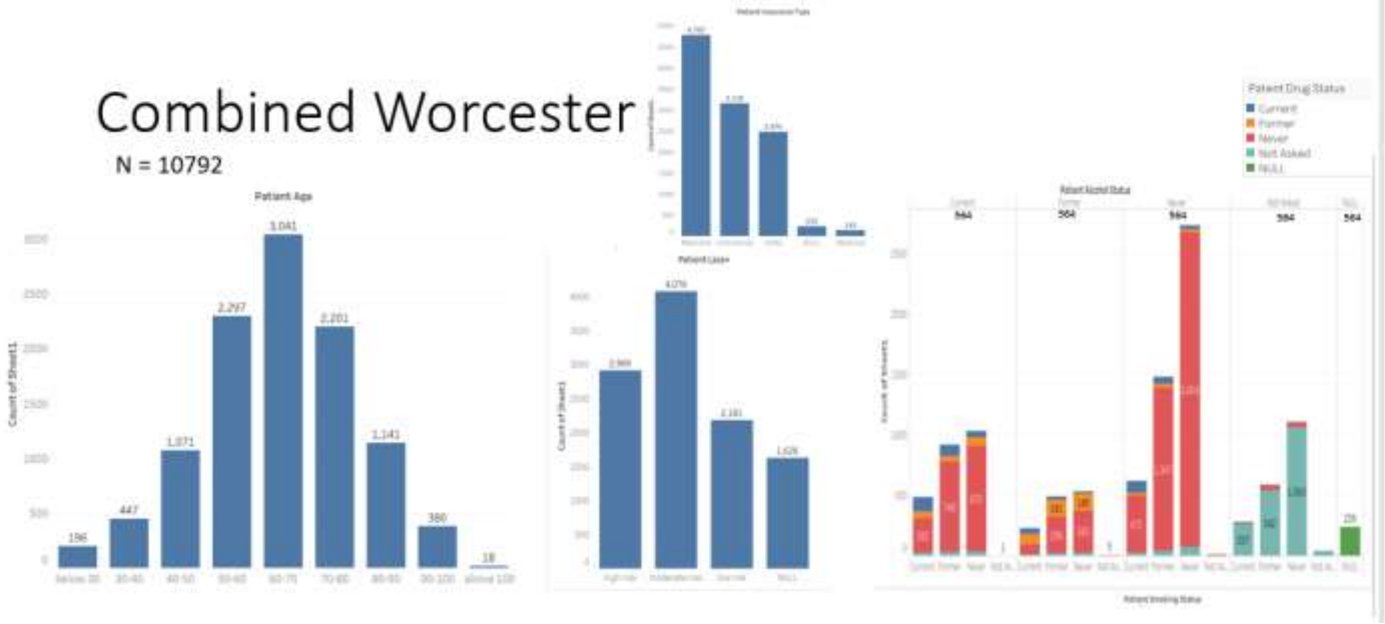
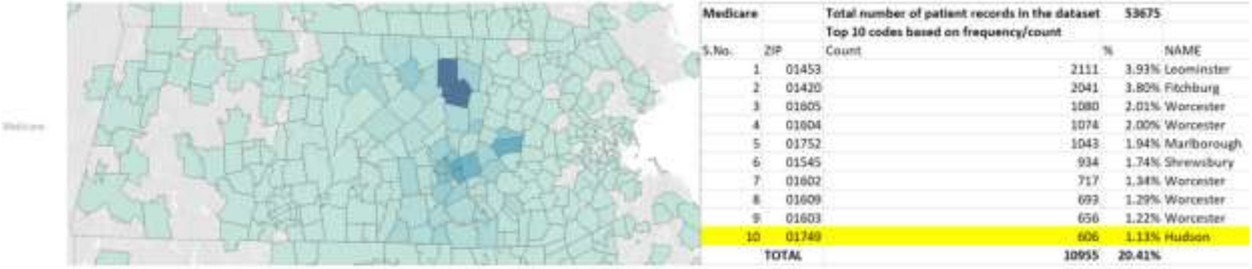


Figure B.2.6 Distribution of patient age, insurance type, LACE+ and social_HX in Worcester*

*(6 of the 10 Worcester zip codes have been combined)



Medicare		Total number of patient records in the dataset		53675
		Top 10 codes based on frequency/count		
S.No.	ZIP	Count	%	NAME
1	01453	2111	3.93%	Leominster
2	01420	2041	3.80%	Fitchburg
3	01605	1080	2.01%	Worcester
4	01604	1074	2.00%	Worcester
5	01752	1043	1.94%	Marlborough
6	01545	934	1.74%	Shrewsbury
7	01602	717	1.34%	Worcester
8	01609	693	1.29%	Worcester
9	01603	656	1.22%	Worcester
10	01749	606	1.13%	Hudson
	TOTAL	10955	20.41%	

Figure B.2.7 Top 10 zip codes covered by Medicare**

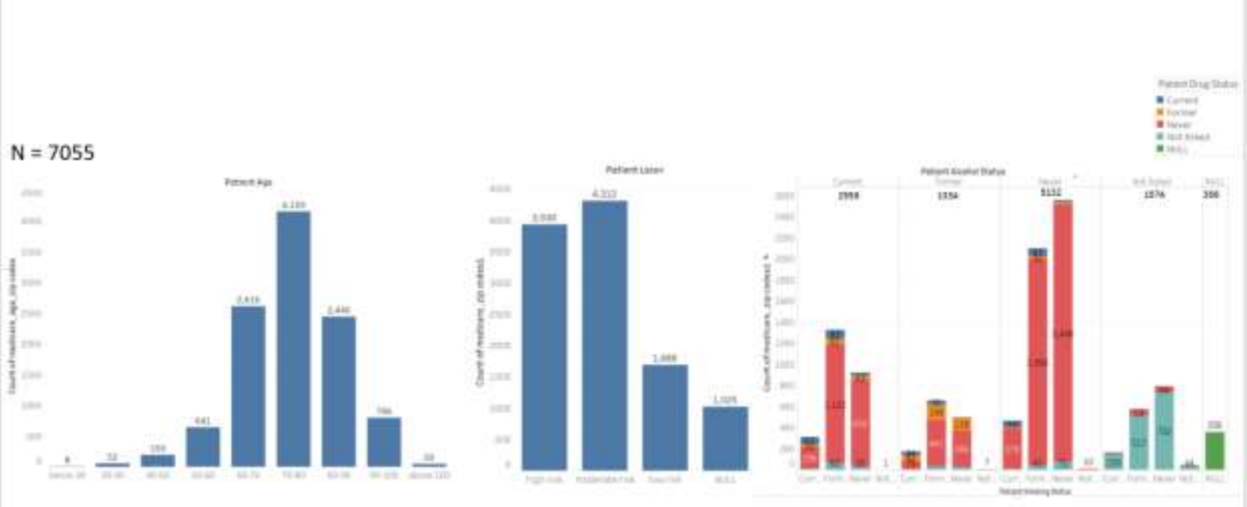
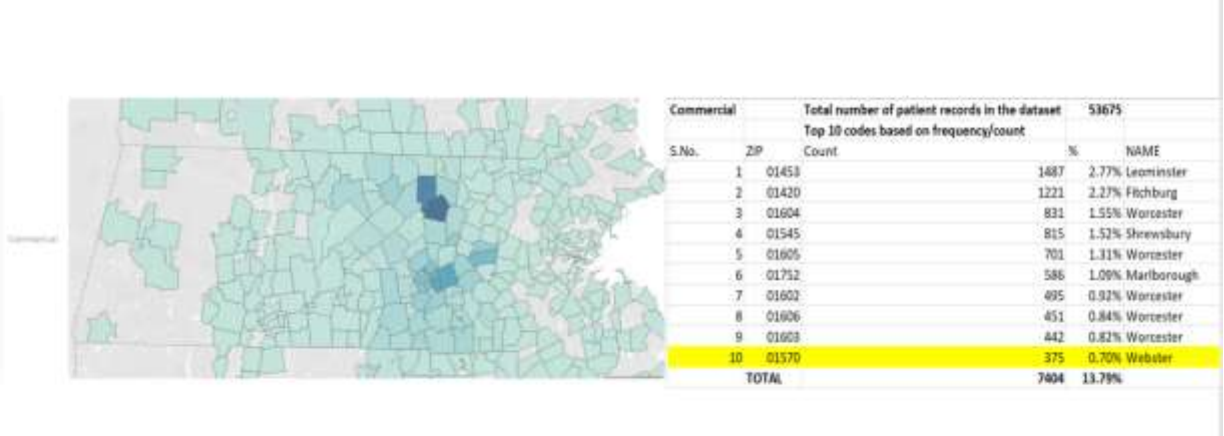


Figure B.2.8 Distribution of patient age, LACE+ and social_HX in the top 10 zip codes covered by Medicare



Commercial		Total number of patient records in the dataset		53675
		Top 10 codes based on frequency/count		
S.No.	ZIP	Count	%	NAME
1	01453	1487	2.77%	Leominster
2	01420	1221	2.27%	Fitchburg
3	01604	831	1.55%	Worcester
4	01545	815	1.52%	Shrewsbury
5	01605	701	1.31%	Worcester
6	01752	586	1.09%	Marlborough
7	01602	495	0.92%	Worcester
8	01606	451	0.84%	Worcester
9	01603	442	0.82%	Worcester
10	01570	375	0.70%	Webster
	TOTAL	7404	13.79%	

Figure B.2.9 Top 10 zip codes covered by Commercial insurance**

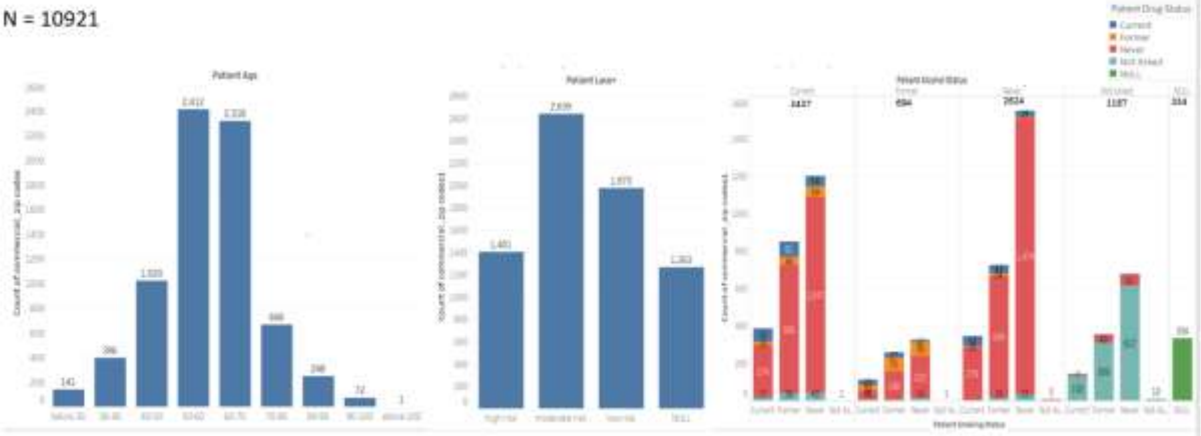
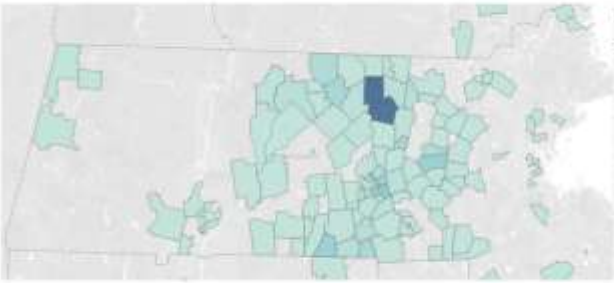


Figure B.2.10 Distribution of patient age, LACE+ and social_HX in the top 10 zip codes covered by Commercial insurance



Medicaid		Total number of patient records in the dataset		53675
		Top 10 codes based on frequency/count		
S.No.	ZIP	Count	%	NAME
1	01420	169	0.31%	Fitchburg
2	01453	164	0.31%	Leominster
3	01604	43	0.08%	Worcester
4	01752	36	0.07%	Marlborough
5	01605	33	0.06%	Worcester
6	01550	33	0.06%	Southbridge
7	01510	28	0.05%	Clinton
8	01440	24	0.04%	Gardner
9	01570	23	0.04%	Webster
10	01603	21	0.04%	Worcester
	TOTAL	574	1.07%	

Figure B.2.11 Top 10 zip codes covered by Medicaid**

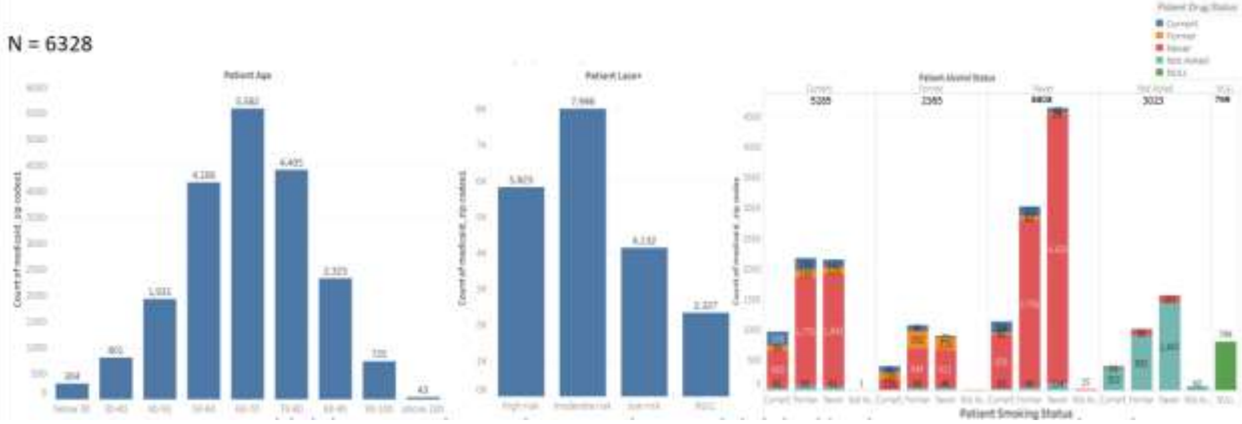
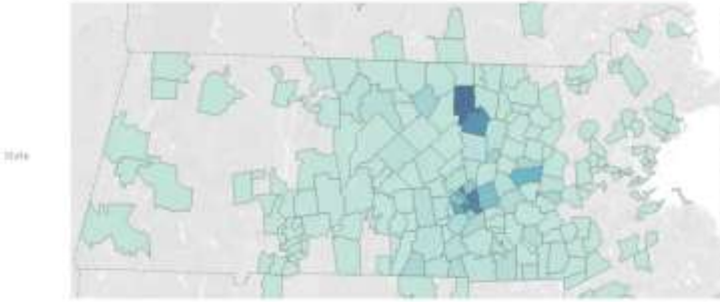


Figure B.2.12 Distribution of patient age, LACE+ and social_HX in the top 10 zip codes covered by Medicaid



State		Total number of patient records in the dataset		53675
		Top 10 codes based on frequency/count		
S.No.	ZIP	Count	%	NAME
1	01420	682	1.27%	Fitchburg
2	01604	555	1.03%	Worcester
3	01605	553	1.03%	Worcester
4	01610	520	0.97%	Worcester
5	01453	490	0.91%	Leominster
6	01603	353	0.66%	Worcester
7	01752	300	0.56%	Marlborough
8	01609	278	0.52%	Worcester
9	01545	252	0.47%	Shrewsbury
10	01602	217	0.40%	Worcester
	TOTAL	4200	7.82%	

Figure B.2.13 Top 10 zip codes covered by State insurance**

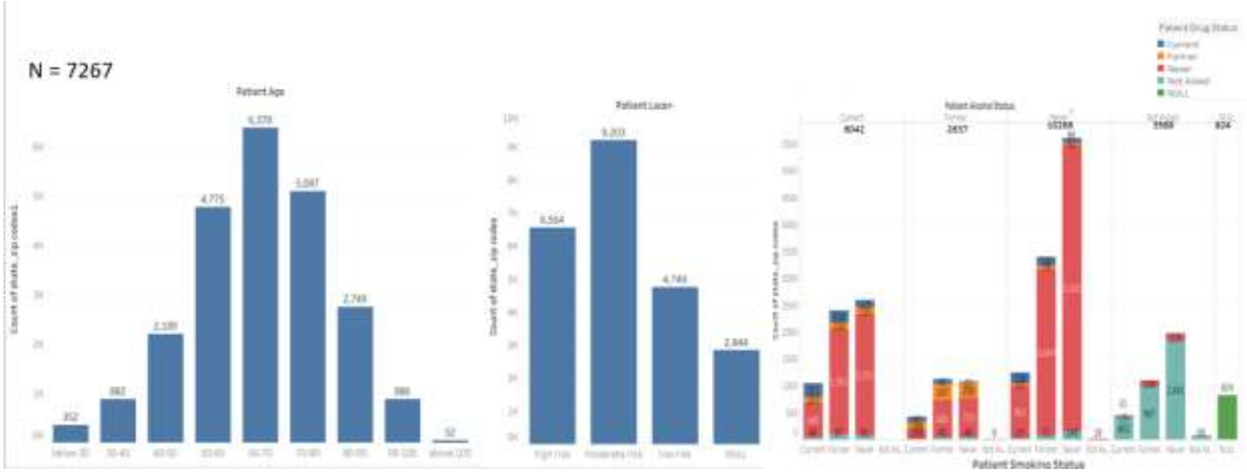


Figure B.2.12 Distribution of patient age, LACE+ and social_HX in the top 10 zip codes covered by State insurance

(** Tables are highlighted to show emerging patterns in the areas across the insurance types. There are some new areas which do not occur in the top 10 category while some areas that hold constant positions or interchange based on the insurance type)

Appendix C

ACO population

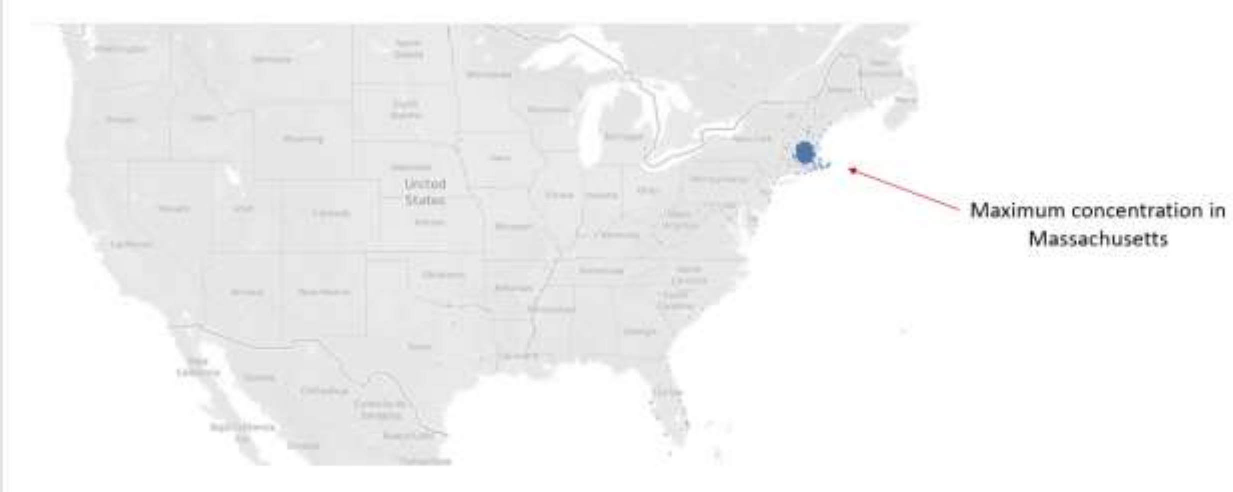
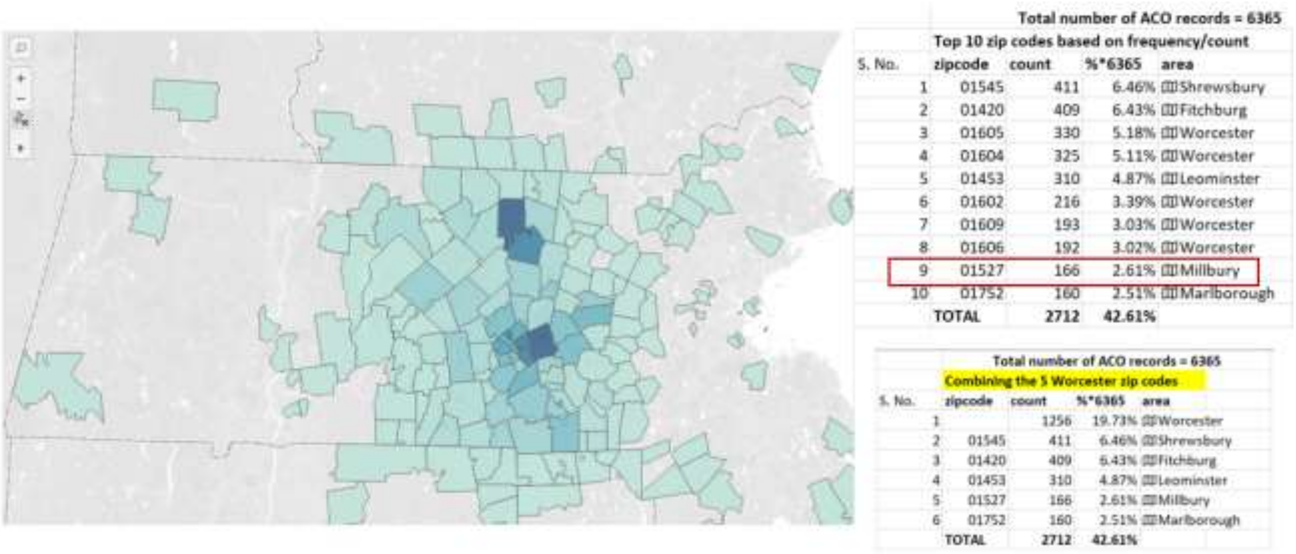


Figure C.1 Distribution of ACO patients across the North American continent



Total number of ACO records = 6365

Top 10 zip codes based on frequency/count

S. No.	zipcode	count	%*6365	area
1	01545	411	6.46%	□□Shrewsbury
2	01420	409	6.43%	□□Fitchburg
3	01605	330	5.18%	□□Worcester
4	01604	325	5.11%	□□Worcester
5	01453	310	4.87%	□□Leominster
6	01602	216	3.39%	□□Worcester
7	01609	193	3.03%	□□Worcester
8	01606	192	3.02%	□□Worcester
9	01527	166	2.61%	□□Millbury
10	01752	160	2.51%	□□Marlborough
	TOTAL	2712	42.61%	

Total number of ACO records = 6365

Combining the 5 Worcester zip codes

S. No.	zipcode	count	%*6365	area
1		1256	19.73%	□□Worcester
2	01545	411	6.46%	□□Shrewsbury
3	01420	409	6.43%	□□Fitchburg
4	01453	310	4.87%	□□Leominster
5	01527	166	2.61%	□□Millbury
6	01752	160	2.51%	□□Marlborough
	TOTAL	2712	42.61%	

Figure C.2 Geo mapping of ACO patients in Massachusetts and the top 10 most frequently occurring zip codes***

(*** Millbury is the only area in the top 10 which does not show up in the overall patient population)

N = 6365

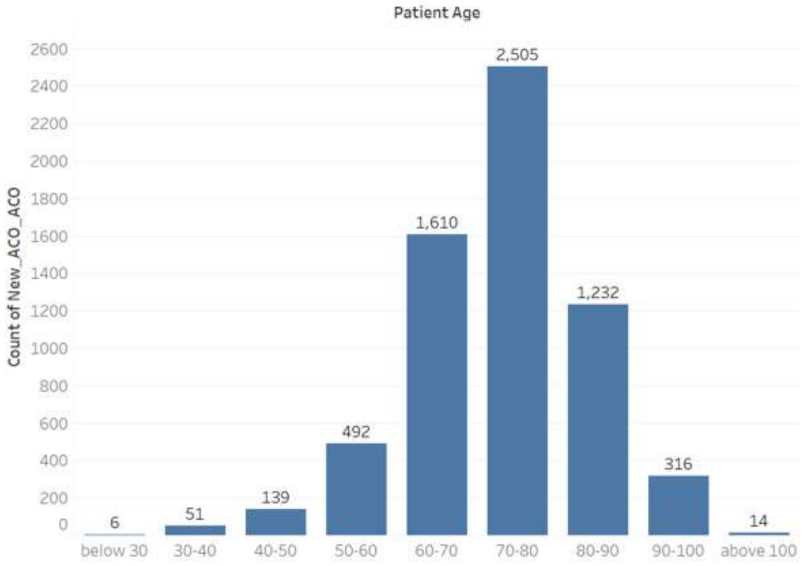


Figure C.3 Age distribution of ACO patients

N = 6365

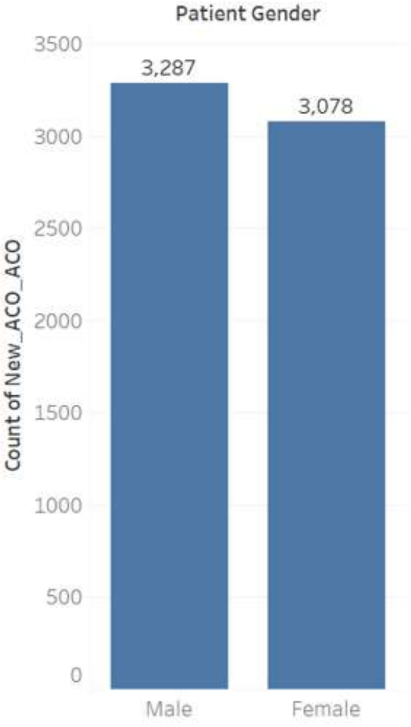


Figure C.4 Gender distribution of ACO patients

N = 6365

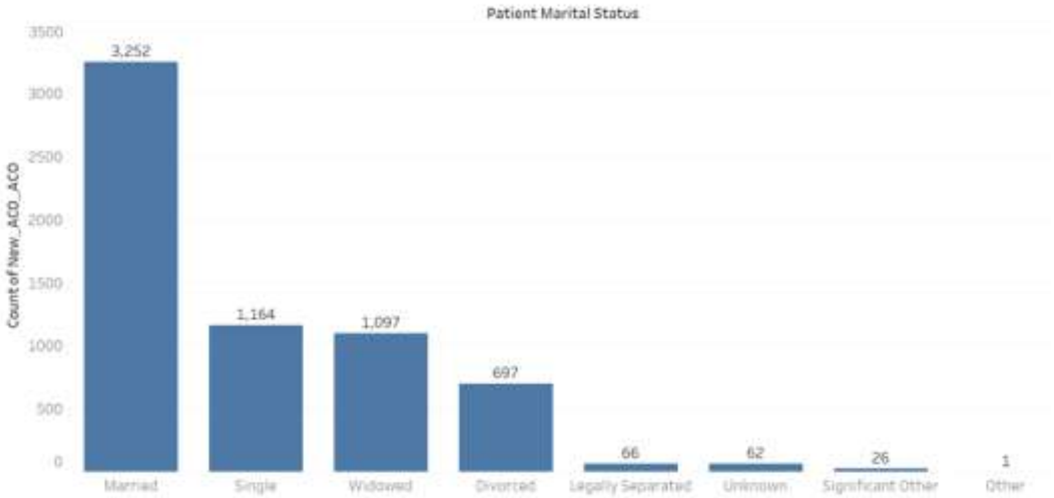
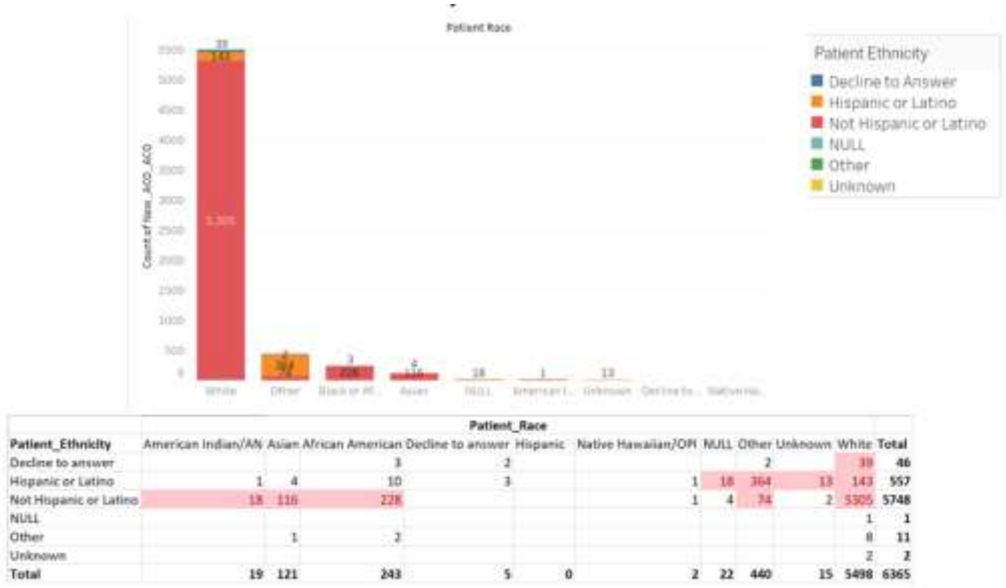


Figure C.5 Distribution of ACO patients' marital status

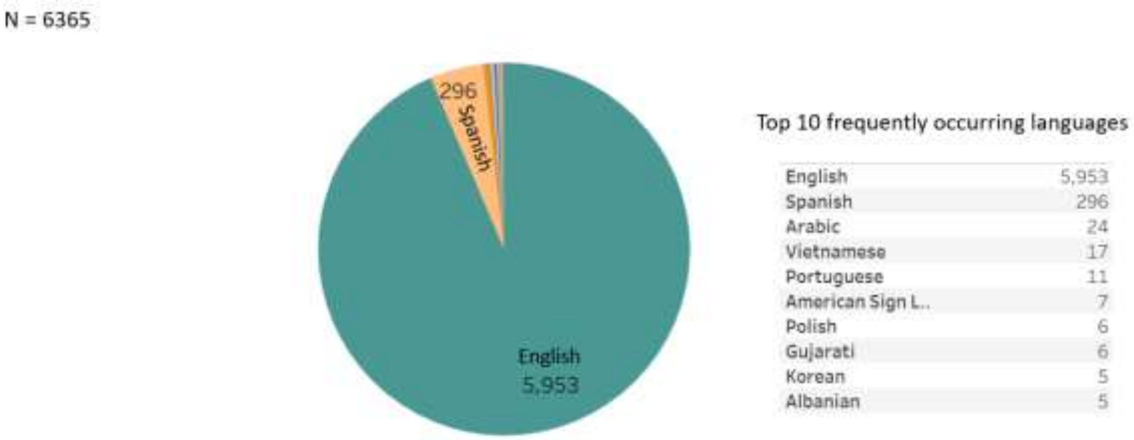
N = 6365



	Patient_Race										
Patient_Ethnicity	American Indian/AN	Asian	African American	Decline to answer	Hispanic	Native Hawaiian/OPI	NULL	Other	Unknown	White	Total
Decline to answer			3	2				2		39	46
Hispanic or Latino	1	4	10	3		1	18	364	13	143	557
Not Hispanic or Latino	18	116	228			1	4	74	2	5305	5748
NULL										1	1
Other		1	2							8	11
Unknown										2	2
Total	19	121	243	5	0	2	22	440	15	5498	6365

Highlighted cells show the top 10 frequently occurring categories

Figure C.6 Distribution of ACO patients' race and ethnicity



English	5,953
Spanish	296
Arabic	24
Vietnamese	17
Portuguese	11
American Sign L..	7
Polish	6
Gujarati	6
Korean	5
Albanian	5

Figure C.7 Distribution of ACO patients' primary language

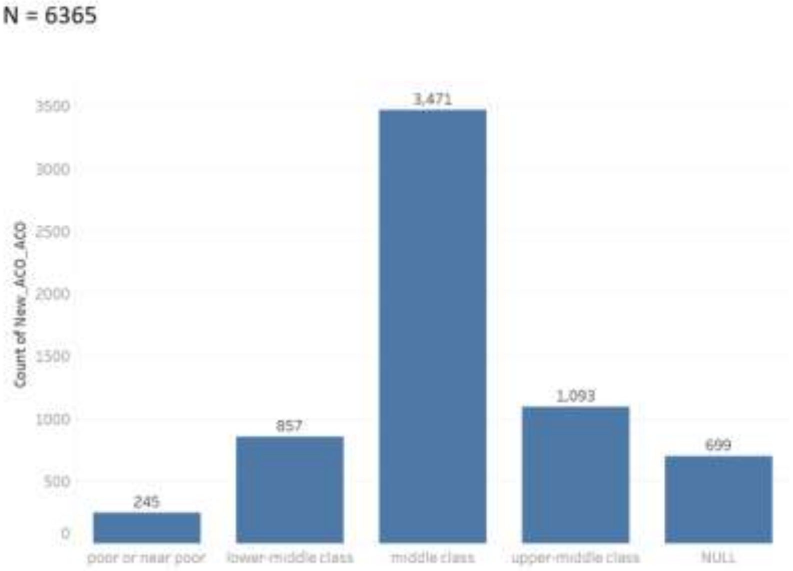


Figure C.8 Distribution of ACO patients' income

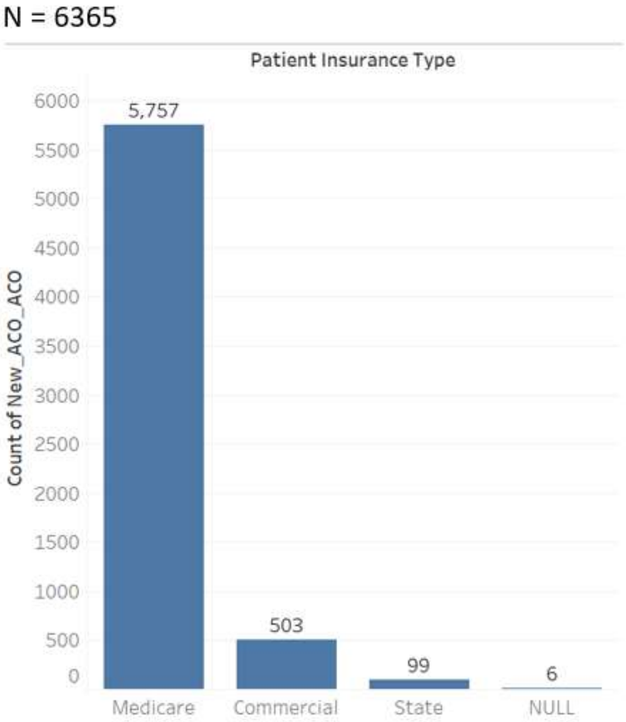


Figure C.9 Distribution of ACO patients’ insurance type

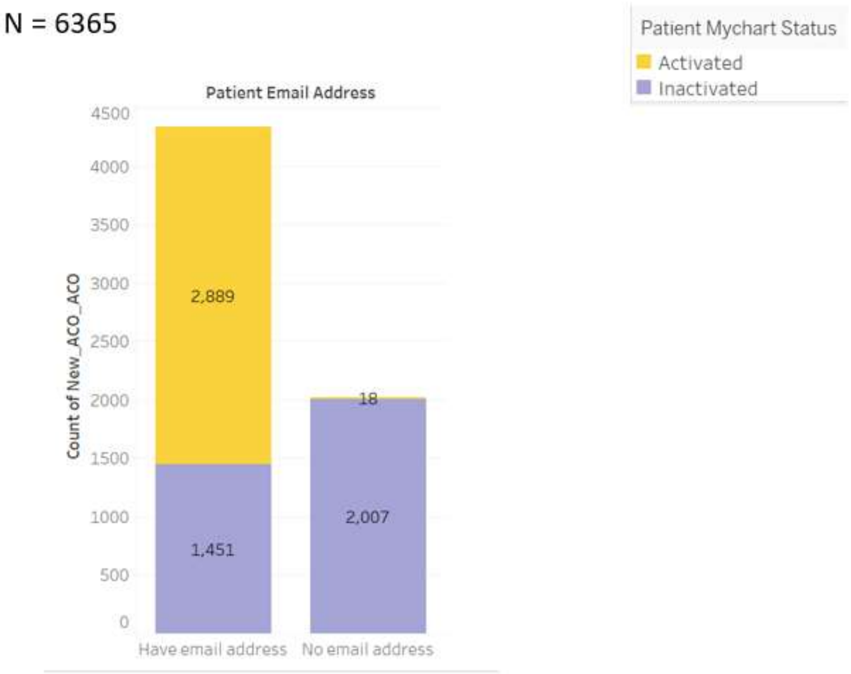


Figure C.10 Distribution of ACO patients’ access to tech

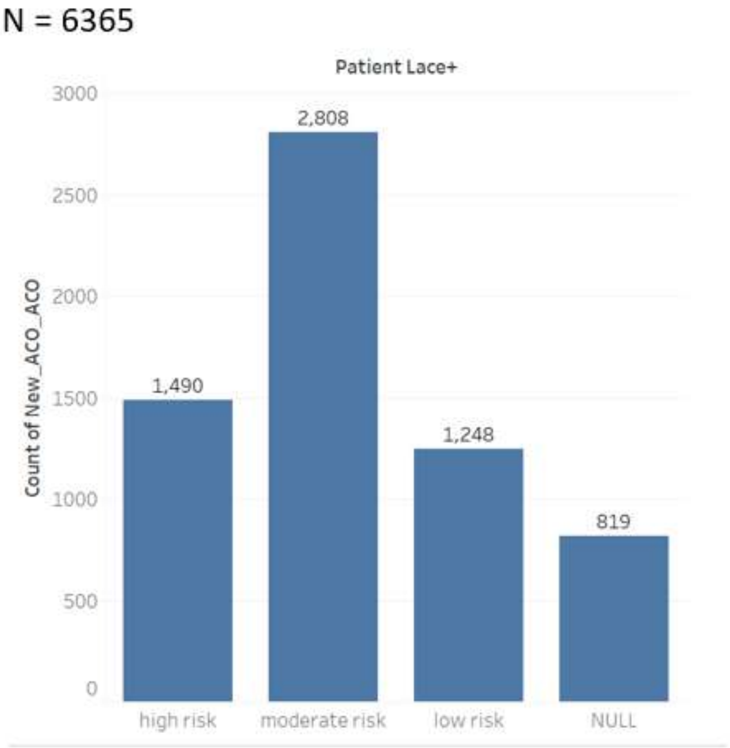


Figure C.11 Distribution of ACO patients’ LACE+ scores

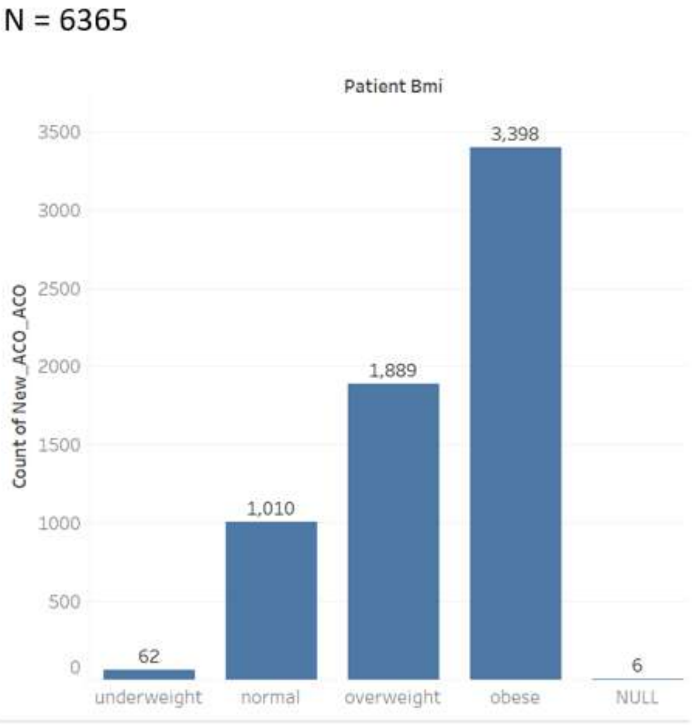


Figure C.12 Distribution of ACO patients’ BMI

N = 6365

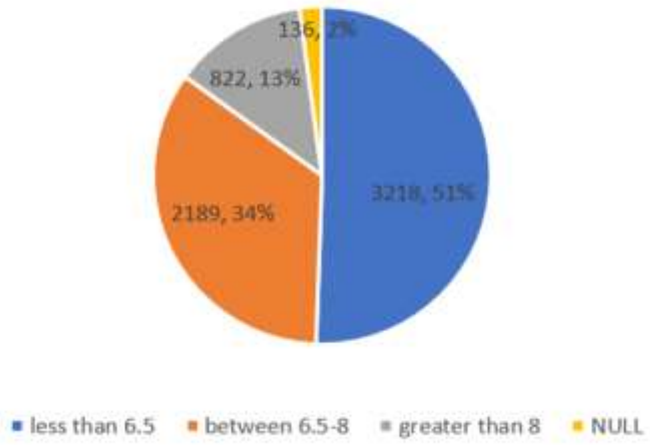


Figure C.13 Distribution of ACO patients' HbA1C

N = 6365

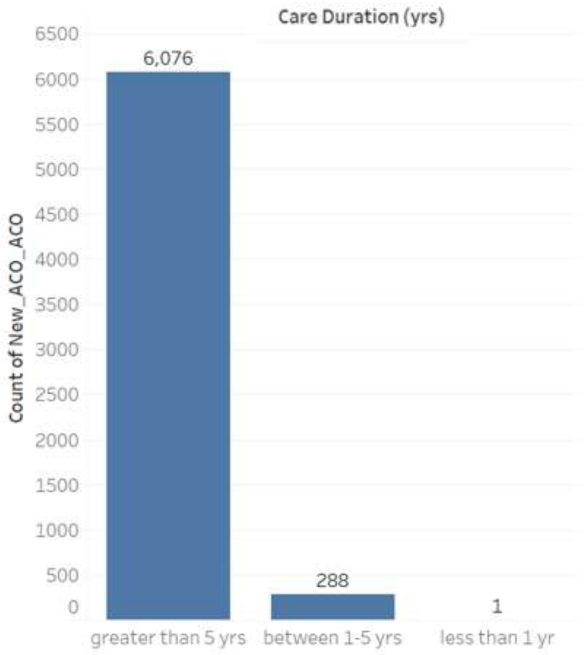


Figure C.14 Distribution of ACO patients' duration of care

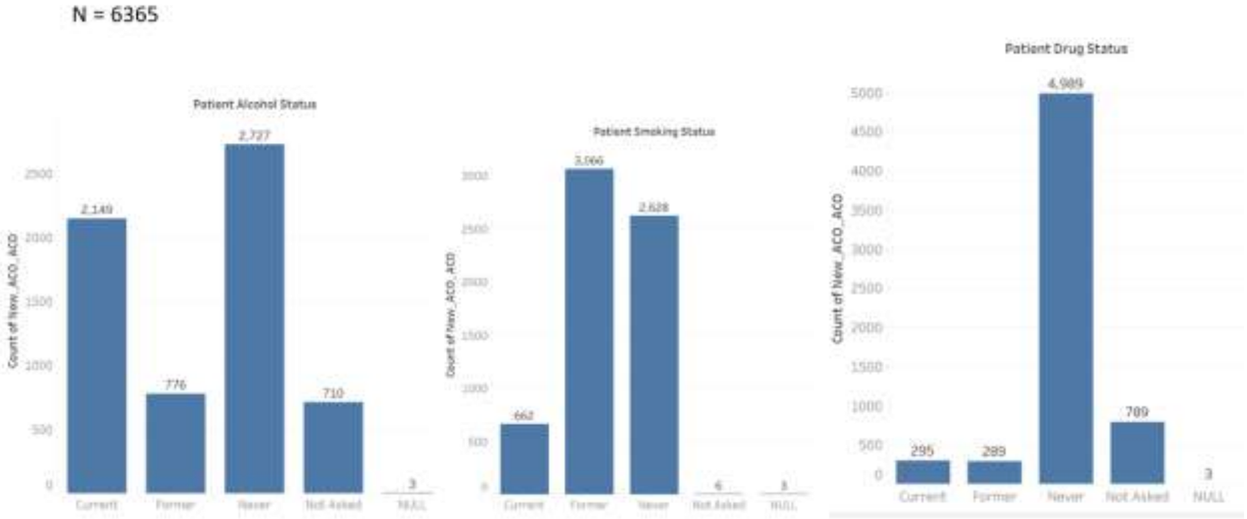


Figure C.15 Distribution of ACO patients' social_HX

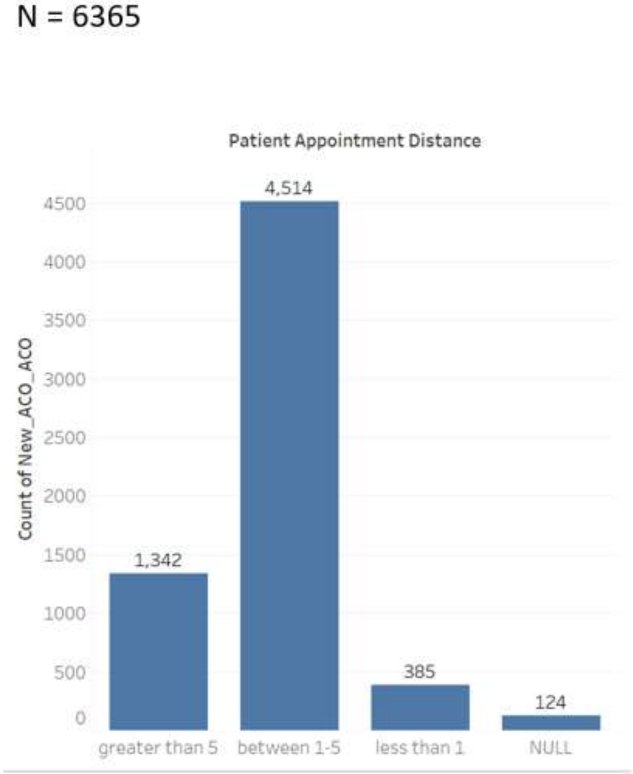


Figure C.16 Distribution of ACO patients' appointment distance