2022

Estimating Weighted Panel Sizes for Primary Care Providers: An assessment of clustering and novel methods of panel size estimation on electronic medical records

Martin A. Lavallee
Virginia Commonwealth University

Estimating Weighted Panel Sizes for Primary Care Providers: An assessment of clustering and novel methods of panel size estimation on electronic medical records

Martin Lavallee

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

Virginia Commonwealth University

2022

Reading Committee:

Roy T. Sabo PhD, Chair

Leroy Thacker PhD

Ekaterina Smirnova PhD

Alison Huffstetler MD

Yalda Jabbarpour MD

Program Authorized to Offer Degree:

Biostatistics

Virginia Commonwealth University

Abstract

Estimating Weighted Panel Sizes for Primary Care Providers: An assessment of clustering and novel methods of panel size estimation on electronic medical records

Martin Lavallee

Chair of the Supervisory Committee:

Dr. Roy T. Sabo PhD

Biostatistics

"Primary Care is on the frontlines of healthcare, thus they see the most diverse set of patients. In order to achieve high functioning primary care, a practice must establish empanelment, the pairing of patients to providers. Enumeration of empanelment, or estimating panel sizes, helps ensure that the demands of the patients demand the supply of providers and optimize the balance of primary care resources to improve quality of care. Further we can adjust panel sizes by using patient-level data on healthcare utilization and complexity extracted from the electronic medial record to determine the amount of care or burden of work that a patient poses to a provider. With this adjustment we can have a more informed estimation of panel sizes that can differentiate the amount of care provided to individuals instead of assuming work for each patient is the same. This dissertation attempts to evaluate different methods of estimating adjusted panel sizes and understand the best practices for extracting data from the EMR to build decision making tools. In our analysis we compare the current best panel size estimation method introduced by Rajkomar et al 2016 with processes that change the clustering method (k-means vs gaussian mixture model), provide a direct estimation of a burden score, incorporate demographics and complexity data into the clustering (*KAMILA*

and gaussian multinomial mixture model) and assess how to conduct panel size estimation with a larger set of features beyond utilization counts.”

TABLE OF CONTENTS

LIST OF FIGURES

Figure Number Page

LIST OF TABLES

Table Number Page

ACKNOWLEDGMENTS

“I would like to thank Dr. Roy T. Sabo for introducing me to this panel size estimation project. I am very grateful to have had him as an advisor through this process. I would also like to thank my dissertation committee for taking the time to review and provide critical feedback for this research. The VCU data for this project was provided by Dr. Steve Crossman. A component of this project required conversion and utilization of the OMOP Common Data Model. Thanks to the OHDSI community for all of the open-source software and documentation needed to understand the complex OMOP structure. The OHDSI community has been a vital part of my professional life, having encouraged me to become a passionate researcher of longitudinal observational databases way beyond this disseration.”

DEDICATION

“This dissertation is dedicated to my most coveted loved ones who have helped me cope through the highs and lows of graduate school and help remind me what matters in life (love and family). I have been doubted by myself and others in this journey but through your persist support I was able to tackle the obstacles and reach the top! To my partner Kara Dennis and my family Gracia, Chirs, Patrick, and Olivia...I love you and thank you.”

INTRODUCTION

In our healthcare system, the primary care provider serves as the gate-keeper. They are often the first provider a patient consults for a medical situation and help patient's navigate the healthcare sector either by providing medical advise, prescribing medication or referring them to a specialist. Given this gatekeeper role, primary care providers serve the most diverse patient populace and it is important that practices have a good understanding of patient-provider pairings in order to provide high-quality service. The process that practices can understand balancing the diverse patient population and the limits of their provider workforce is through empanelment. Empanelment is the process of linking a patient to a primary care provider or care team. For the patient, this is important because it ensures that patients can establish continuity and trust with someone who understands their unique medical history. From the provider and practice perspective, empanelment is important because it helps enumerate the patients they serve to ensure they are not being overworked and can provide the personalized care their patients demand. Enumeration of empanelment is known as panel size estimation. Quantification of panel sizes helps address a supply and demand problem in healthcare; matching the demand of medical services from the patients with the supply of providers to address medical requests. A primary care practices' ability to understand panel size can help them be more efficient in assigning new patients to providers and properly address the medical needs of each patient.

A simplistic way of estimating a panel size is to count the number of patients matched to a provider. The issue with this simple approach is that it does not consider the burden of work posed by the patient to the provider. Each patient is different. An elder patient who smokes and has heart issues is going to be more burdensome to a provider than a young patient who goes in for an annual check up. If we were to aggregate all these patients, a provider who treats several older and sicker patients will take on much more work and follow-up care than a patient who only does annual check-ups. One thought for panel size estimation is to adjust them based on the work done for the patient, instead of treating each patient the same. Returning to our example, the elder patient with more health problems would have a higher weight than the young patient. When aggregated we can better assess which providers have more burdensome panels and properly balance the provider resources so that they are not overburdened. A few attempts have been made for coming up with adjusted panel sizes but few use the full power of data available from electronic medical records. The electronic medical record is a treasure trove of data that can be used improve healthcare decision making and quality of care. In the context of panel size estimation, we can extract patient-level data from the EMR to more informatively derive how burdensome (in terms of care effort) a patient poses to a provider. Features such as number of medications, number of primary care visits or number of specialty visits can help gauge how often a patient utilizes healthcare resources. Understanding patient complexity such as indication of certain drugs and chronic disease can also improve our ability to estimate the relative burden of a patient. Panel size estimation can benefit from the multitude of information from the EMR in order to better define certain types of patients who would be more burdensome or less burdensome. One such example was presented by Rajkomar et al in 2016 where they used k-means clustering to identify phenotypes of patient utilization and project a weight of burden to adjust panel sizes. This is the most sophisticated approach to panel size estimation at the moment, however there are opportunities to investigate improvements to this process and practically understand ways that we can squeeze the most information from the EMR to improve healthcare decision making.

The goal of this dissertation is to analyze methods of estimating adjusted panel sizes, improve our practical understanding of using unsupervised methods for data analysis and exploration and leveraging the EMR as a ample source of information to refine our analysis using a large set of medical features. This dissertation is split into two parts. Part 1 looks at methods of estimating panel sizes using only utilization data from the EMR to estimate adjusted panel sizes, where utilization data are counts of services rendered such as medications, visits, and specialty visits. The second part looks at estimating panel sizes using a mixture of utilization and complexity data from the EMR. Complexity data ranges from baseline demographics to indication of commorbidities and drugs. We combine both types of data to estimate adjusted panel sizes. The main goal of this dissertation is evaluation and providing practical guidance for panel size estimation using the electronic medical record. In addition to evalutating the various methods of panel size estimation we also consider other important factors for analyzing EMRs such as using variance stabilizing transformations for count data, performing low-dimensional projections to help differentiate clusters composed of a high number of features. We also explore important data processing concerns to using the EMR such as provider attribution and mapping to the OMOP common data model. Results of this research are made available throughout this document and further in a web application (https://lavalleema-webaps.shinyapps.io/pseDashboardv4/). If the web application is down please email me at mdlavallee92@gmail.com.

Part 1 of the dissertation looks to compare the current best method of panel size estimation using the EMR introduced by Rajkomar et al in 2016 to alternatives that try to address their pitfalls. Our first alternative is substituting the k-means method with the gaussian mixture model to construct clusters of medical phenotype. The advantage of the gaussian mixture model for clustering is that it has a probabilistic basis to the model, as opposed to k-means. The second alternative introduces the novel patient burden scoring method which directly derives a score to a patient instead of projecting a latent classification to a weight. In theory this direct approach avoids a potential issue in mis-identifying the mapping feature for creating weights and provides a more balanced contribution of all features in creating an adjusted panel. We look to evaluate the three panel size methods (k-means, gaussian miture model and patient burden scoring) first using real world data provided from VCU Health systems and second through simulated data. We want to see if we can improve upon the Rajkomar method and provide practical recommendations on how to conduct panel size estimation through EMRs.

In Part 2, we expand upon the Rajkomar approach by incorporating patient complexity data into the analysis. Complexity data is categorical so traditional clustering methods like k-means no longer work since they only analyze continuous data. Therefore we explore approaches of mixed data clustering for handling both data types. One such method is the gaussian multinomial mixture model, a type of finite mixture model. This model naturally handles categorical data using a multinomial distribution and continuous utilization data under the gaussian distribution. Another candidate method is *KAMILA*, an approach proposed by Foss et al in 2018 that uses a semi-parametric approach to clustering mixed data. *KAMILA* relaxes the parametric assumption of the continuous data by using a kernal density estimate in a model that emulates a finite mixture model. Relaxing this parametric assumption can be important since count data follows a poisson distribution, not a gaussian distribution, so being more robust to distributional assumptions may improve estimation. Finally, we introduce an extension the patient burden scoring called mixed patient burden scoring to handle the mixed data. In this approach we adjust the utilization data with predictions from the categorical data. The idea being a set of complexity would indicate greater or less utilization which is used to derive the burden score. Part 2 of the analysis has three approaches to evaluating the mixed data methods. Similar to part 1 we use the VCU health systems data and simmulated data to evaluate the three mixed data approaches to panel size estimation. Additionally we conduct an analysis using synthetic EMR data from Synthea. Full access to EMR data is difficult to obtain and the VCU data was based on a specific extraction that omitted patient complexity information. Using synthea, which we mapped to the OMOP CDM, we were able to experiment with mixed data panel size estimation in the situation that we have access to a very large number of medical features.

Chapter 1

PANEL SIZE ESTIMATION FOR UTILIZATION DATA

1.1 Introduction

Primary Care physicians are on the front line of the healthcare workforce, typically the first step for patients' medical consultations. Primary care physicians are crucial because they help patients navigate the complexities of the health care system and are key promoters of public health initiatives. Well functioning primary care relies on ease of access and continuity (Bodenheimer, Ghorob, Willard-Grace, & Grumbach, 2014). Patients demand that primary care providers are readily available to address their questions and are able to establish trust and familiarity with their clinician.

Meeting these demands can be challenging for providers. Some patients are annual visitors who are otherwise healthy, others with concerning comorbidities may not frequent a practice currently but could increase their number of visits in the future, and others who frequent a practice and demand a lot of attention due to concurrent health concerns. The diversity of the patient populace and the need to individualize care to each of these patients can overburden providers and hamper their ability to achieve the aims of continuity and ease of access. Further, a growing insured population (through the Affordable Care Act and Medicaid expansions in certain states), public health crises, an aging population and a shortage of physicians has made primary care even more important for the US healthcare system. Given this context, establishing a high-functioning primary care arm of the workforce is vital.

Key to striking the balance between supply and demand for primary care services is empanelment: the process of linking a patient to a primary care physician or care team (Bodenheimer, Ghorob, Willard-Grace, & Grumbach, 2014; Grumbach & Olayiwola, 2015).Understanding empanelment is crucial for several reasons: it establishes the patient-provider relationship, it improves access and continuity of primary care, it helps monitor care performance, promotes and assesses population health, improves management of chronic diseases, improves provider performance, and improves patient satisfaction (Bodenheimer, Ghorob, Willard-Grace, & Grumbach, 2014; Grumbach & Olayiwola, 2015; Murray, Davies, & Boushon, 2007). While a foundational building block for high-performing primary care practices, empanelment and panel size (enumeration of provider empanelment) are often overlooked. Peterson et al showed that only about a third of primary care physicians who responded to a survey could estimate their panel size. However, about 70% of practices in the same survey had access to an electronic medical record (EMR) meaning the data and technology required to calculated panel size is available to most practices (Peterson, Cochrane, Bazemore, Baxley, & Phillips, 2015).

Despite the importance of empanelment in the literature, there is not a lot of available information on how to properly estimate panel size; let alone methods that leverage the quantity of information available from electronic medical records. The simplest panel size procedure would be to count the number of unique patients attributed to a provider. However, we know that patient demands are complex and unique. A simple count does not reflect the amount of time and effort a provider dedicates to their particular paients. Thus, panel size estimation requires methods that adjust for variations in patient complexity and utilization. Kamentz et al propose a panel size adjustment approach that creates groupings based on age, sex and insurance type and then develops weights based on the average number of visits and telephone calls for patients in the group compared to the entire population. A weighted sum then yields the adjusted panel size (Kamnetz, Trowbridge, Lochner, Koslov, & Pandhi, 2018). Other approaches to panel size estimation use disease burden (Potts, Adams, & Spadin, 2011) or work resource value units (wRVUs) (Ajorlou, Shams, & Yang, 2015; Arndt, Tuan, White, & Schumacher, 2014; Chung, Eaton, & Luft, 2012). Many of these panel size estimation methods have limited predictive power since they use basic patient demographic data. Many of these methods do not properly account for direct and asynchronous care that is necessary for empanelment methods.

Rajkomar provides the most sophisticated approach to adjusted panel size estimation by using k-means clustering to help identify healthcare utilization phenotypes by considering counts from all patient encounters found through the EMR (Rajkomar, Yim, Grumbach, & Parekh, 2016). In this method healthcare utilization data were extracted from electronic medical record over two one-year intervals covering information on primary care visits, no shows, specialty visits, emergency department and hospitalization visits, telephone calls, portal messages, and number of medications prescribed, which were capped and weighted. The Rajkomar method begins by defining healthcare utilization phenotypes through a combination of clustering and up-front assignment. The method constructs phenotypes on the first year of patient data training set. Patients considered high outliers (greater than 6 standard deviations from mean annual primary care visits), inactive (no utilization in any category) or minimal activity group (a set of patients who met a minimal threshold defined by a physician panel) were set as an initial set of healthcare utilization phenotypes. The remaining patients were clustered into 4 groups using k-means clustering. The group with the smallest median primary care visit count was further clustered into another 2 groups. The final eight healthcare utilization phenotypes were then condensed into 4 categories, with the inactive group a separate cluster. Next, weights are calculated by taking the median primary care visit amount of the low, medium and high groups and dividing it by median visit count of the low group. The inactive group receives a weight of 0.05. A weighted sum of each provider panel then gives the panel size. Rajkomar validates the utilization phenotypes by evaluating how well the clustering of training set can predict the year 2 primary care using log linear models compounded with patient complexity covariates like gender, age and insurance payer type. Rajkomar showed that the models incorporating the utilization phenotypes constructed through k-means clustering had better predictive performance compared to raw primary care visit counts. While this approach is the sophisticated approach to panel size estimation using the EMR, other clustering methods exist and should also be evaluated. This study seeks to review and evaluate possible improvements to panel size estimation outlined by Rajkomar.

One way of estimating weights of patient burden is to first perform a cluster analysis to latently assign a level of burden to patients who exhibit similar medical features and then map the group to weight based on how often they visit their primary care doctor. This is the framework established by Rajkomar who developed the most sophisticated approach to panel size optimization using k-means clustering and the wealth of healthcare information available from the electronic medical record. The criticism to k-means clustering is that it deterministically assigns clusters. The gaussian mixture model is a natural extension to the current best practice as a probabilitic approach to clustering utilization data.

Instead of using latent assignments as both the k-means and gaussian mixture model both do, we also created a novel process for panel size estimation problem which evaluates dissimilarity between each patient to a reference value (i.e. a “typical” patient). This process develops a score from 0 to 2, where values approaching zero indicate less-than-typical burden and values approaching two indicate higher-than-typical burden (by literally doubling the patient’s count). A sum of these burden scores gives an estimation of panel size that theoretically balances all patient features and does not rely on mapping latent groups to a value. We introduce the approach of patient burden score to handle numeric healthcare utilization data and evaluate it against the clustering methods of panel size estimation.

For this paper we assess ways of leveraging utilization data from the electronic medical records towards estimating panel sizes of primary care physicians. In this analysis we compare the three methods for estimating panel sizes using utilization data: kmeans clustering, the gaussian mixture model and the novel patient burden scoring. There are two parts to this analysis: first, we apply the three panel size estimation methods to real world data from VCU Health Systems and second, we evaluate and assess the operating characteristics of the panel size estimation methods using simulated data. From this analysis we plan to provide suggestions on best practices for panel size estimation for primary care using utilization data from electronic medical records.

1.2 Methods

1.2.1 Panel Size Estimation Methods

Clustering Methods

Our first option for estimating provider panel sizes using healthcare utilization data is to use clustering. According to Hastie et al the goal of clustering is to minimize pairwise dissimilarities between observations assigned to the same cluster (Hastie, Tibshirani, & Friedman, 2009). There are several types of clustering methods that may be used, but in this context we focus on the finite mixture modeling class of clustering methods as these are the methods previously used in the field. In this analysis we also focus on squared Euclidean dissimilarity as our default measure in clustering. Choice of dissimilarity measure is integral to cluster performance (Hastie, Tibshirani, & Friedman, 2009). However, we do not consider other dissimilarity measures for healthcare utilization clustering because first, in this aim we are not using categorical or ordinal variables, second this simplifies our parametric assumption for finite mixture model to multivariate normal and third we believe it is the most suitable assumption for potential high-dimensional settings.

When using clustering for panel size estimation, the first step is to identify “like” groups of patients. Our expectation is that our chosen clustering algorithm will assign high utilizing patients together and low utilizing patients together so that when these groups are counted a reasonable weight can be assigned to balance the burden that comes from these clusterings of patients. In the following we describe our different clustering techniques, however there is a key step that follows clustering: mapping the latent assignment to a value. We reflect what was done by Rajkomar and use the median visit count of each cluster as the basis of our weighting scheme for panel size estimation (Rajkomar, Yim, Grumbach, & Parekh, 2016).

The median visit count for each cluster is scaled by the median visit count of the cluster with the lowest visit count, thus the order range has a lower bound of one. Next an overall weight is calculated by dividing the total number of patients by the sum of the scale times the cluster count. This overall weight is then multiplied by the scale of the cluster group to get the cluster weight. The sum of these weights stratified by the physician produces the panel size estimation. It is assumed that this group has the lowest overall utilization so it will have the smallest attributed weight when counting the panel. We consider this to be the best and most reasonable assumption to map each cluster to a weight value for panel size estimation.

Panel size estimation can be thought of as a clustering problem, where the goal is to find groupings of similar patients and assign a weight to each patient for counting a panel size. This is how panel sizes were estimated by Rajkomar, using a k-means clustering algorithm. K-means assigns each observation to the cluster with the nearest mean that minimizes variance. Next the cluster means are recalculated. These steps are iterated until the cluster assignments no longer change. The most popular k-means clustering algorithm is Hartigan and Wong (Hartigan & Wong, 1979), which is used by Rajkomar. Some key assumptions for using k-means clustering are that the data is quantitative (no binary values), dissimilarity is measured by Euclidean distance and there is a deterministic "hard" assignment of each data point to a cluster. The k-means method is a special case of a gaussian mixture model (presented in greater detail in the next section), where the covariance matrix of the multivariate gaussian distribution is $\Sigma_k = \sigma_k^2 \mathbf{I}_{n_k}$ for each cluster. Given this assumption, we do not account for the probability of the *ith* observation having group assignment k; it greedily chooses the closest current cluster center as group assignment for each observation. The advantage of this deterministic assumption is that our computational cost is linear. The disadvantage of k-means is that we are making a assumption that the covariance of the multivariate normal modelis zero. In this analysis we use k-means clustering as a point of comparison for panel size estimation of healthcare utilization. The k-means algorithm for panel size estimation is described in Algorithm A.1. We do k-means clustering using the *kmeans* function in the *stats* package in R (R Core Team, 2021).

The Gaussian Mixture model is the natural extension to k-means clustering. It is a “soft” clustering method because it makes a probabilistic group assignment based on the maximum posterior probability for each observation. The gaussian mixture model is the most basic iteration of a parametric finite mixture model, described in general by McLachlan and Peel 2000 (McLachlan & Peel, 2000). K-means clustering is a special subset of the gaussian mixture model making a key assumptions that $\Sigma_i = \sigma^2 I$ and the mixing proportion are equal ($\pi_i = 1/K$). While this speeds up convergence for k-means, it can be an over-generalization of the true covariance structure in the normal mixture model resulting in “hard” cluster assignment. Using a gaussian mixture model we are assuming that clusters take on an ellipsoidal shape, centered at mean μ_i with other geometric features defined by the covariance. We can use eigen-decomposition of the covariance matrix $\Sigma_i = \lambda_i D_i A_i D_i^T$, where λ_i defines the volume, D_i defines the orientation and A_i defines the shape (Scrucca, Fop, Murphy, & Raftery, 2016). This leads to a flexible depiction of an appropriate covariance structure that best fits the data for clustering. Despite this advantage over k-means clustering, the trade-offs with using a gaussian mixture model are: first, the EM algorithm can be slow or get caught at a “local maxima” that does not provide an optimal solution and second, a bad parametric assumption can lead to poor clusters. The process of calculating the panel size using a gaussian mixture model is also described in algorithm A.1 with the difference that instead of k-means clustering we use the finite mixture model using the EM Algorithm to estimate the posterior probability of cluster assignment. Also we calculate the weighted sum of panel using the posterior probabilities. The gaussian mixture modelling can be performed in R using the *mclust* package. A full description of how to use the *mclust* package can be found in the package tutorial (Scrucca, Fop, Murphy, & Raftery, 2016).

A common concern for non-hierarchical clustering methods is how to determine the number of clusters to use in the analysis. For probabilistic clustering using finite mixture models, the ideal number of clusters is typically determined using an information based approach like BIC (Schwarz, 1978). For deterministic clustering like k-means, there is no superior method for selecting the number of clusters and in many cases can be left to domain expertise, depending on the purpose of clustering (Hennig & Liao, 2013). Rajkomar selected the number of clusters based on identifying the value k that exhibits the largest reduction in within-cluster dissimilarity before leveling off for larger value of k (Rajkomar, Yim, Grumbach, & Parekh, 2016). This method is popular since it is easy and visual; it relies on plotting the within-cluster dissimilarity and finding a "kink" in the graph where the values begin to level off (Hastie, Tibshirani, & Friedman, 2009). Another method for selecting the optimal number of clusters is the Gap Statistic (Tibshirani, Walther, & Hastie, 2001). Similar to Foss, we determine the number of clusters for analysis using the the prediction strength algorithm from Tibshirani and Walther (Tibshirani & Walther, 2005). The intuition is that if k is the true number of clusters then the k training set clusters will be similar to the k test set clusters. The general process of the prediction strength criteria is to first split the dataset into a training and testing set. Next, one runs a clustering algorithm on both sets to a k number of clusters. We then generate a co-membership matrix using the testing set where the ii'th element is 1 if the pair of observations fall into the same cluster and 0 otherwise. Next, we assess how well the training set cluster centers predict co-membership in the test set, therefore for each value of 1 in the co-membership matrix we see if the same cluster is assigned by the training set. Finally, the prediction strength is calculated:

$$ps(k) = \min_{1 \leq j \leq k} \frac{1}{n_{kj}(n_{kj}-1)} \sum_{i \neq i' \in A_{kj}} D[C(X_{tr}, k), X_{te}]_{ii'} \tag{1.1}$$

where X_{tr} is the training set, X_{te} is the testing set, $C(X_{tr}, k)$ is the cluster operation on k clusters and $D[C(X_{tr}, k), X_{te}]_{ii'}$ is the co-membership matrix. This equation computes, for each cluster, the proportion of observation pairs in that cluster that are also assigned to the same cluster by the training set centroids, where the minimum is the prediction strength. One can assess the optimal number of clusters graphically by choosing the the largest number of clusters that has a prediction strength value greater than 0.8 (Tibshirani & Walther, 2005).

A limitation of the prediction strength method is that it tends to favor a smaller number of clusters (A. H. Foss & Markatou, 2018).

Patient Burden Scoring

The before mentioned clustering methods are a latent approach to panel size estimation. Using the utilization matrix we are clustering "like" patients, then mapping this latent assignment to the visit count to construct a panel weight. However, it is possible to take a more direct approach by generating a patient burden score directly from the utilization matrix. In summary, this approach contrasts each patient to a "typical" patient and produces a score from 0 to 2 that assess the relative burden they present to a clinician. For example, a typical patient in a provider's panel would count as 1, while atypical patients would approach the bounds of the scores' range. Patients who have minimal primary care visits and rarely utilize healthcare resources would count fractionally, since the provider does less work for these patients. Patients who pose heavy burdens to a provider would count multiplicatively, since a provider is doing extra work. For example the most burdensome patient will have a patient burden score of 2, posing double the work. Examples of heavy burden would be arriving for several visits, needing prescription to many medications, requiring messaging between visits and seeking referrals to specialty care. In theory this approach gives a more holistic look at patient utilization, instead of mapping the cluster assignment to a single measure. Also, this approach is distinctly formulated to be more intuitive to panel size estimation logic. Patient burden scoring is described in Algorithm A.2.

In the patient burden scoring method there are a few options in the algorithm. First, we need to decide what to use as reference values of utilization; options include: 1) medians or order statistics, 2) means, and 3) domain specified values of "typical." For this analysis we chose the simplest reference option the means of the utilization measures. Second we need to decide how we aggregate the dissimilarity vectors. Our options include: 1) a crude sum, 2) a weighted sum and 3) a weighted sum using principal components. In this study we use the weighted sum using the weights from the 1st principal component.

1.2.2 Assessment

VCU Health Systems Data

For this analysis, we received EMR extracts from four primary care clinics in the Richmond, Virginia area from VCU Health Systems: Hayes E. Willis Health Center, Nelson Clinic Family Medicine, Mayland Medical Center Family Medicine and Tanglewood Family Medicine. Data was delivered in two-year segements sent at the end of each quarter. We use data from the two-year period between January 2018 to December 2019. For temporal evaluation of clustering we also used data from October 2018 to September 2020, allowing for sufficient time in between. VCU EMR data was de-identified and all patient private health information (PHI) was removed. The names of the physicians were also removed for this analysis. The EMR extracts from VCU contained all records of primary care visits taking place in the two-year period. For each patient that saw a primary care provider in the time period, we constructed measures of utilization from the EMR extract. Utilization data included: the number of primary care visits, the cumulative sum of medications, number of specialty visits, number of emergency room visits, number of portal messages to their providers, number of phone calls to their providers and the number of no show visits. Basic patient information was also provided including the patient sex, age, race and insurance type (medicare, medicaid, private or no insurance). Table A.3 and table A.2 summarize the baseline characteristics of the VCU data.

An important step in panel size estimation is attributing patients to a primary care provider. In the US insurance system, the patient is free to go to any primary care physician they choose despite the possibility of an assigned physician from insurance. As a result the assigned primary care provider in an EMR is often wrong and must be re-attributed based on the amount of care provided by a physician (Kivlahan et al., 2017). Attribution must consider a sufficient look back period, a minimum threshold of visits to be attributed to a practice and the number of times attribution assignment is done (Kivlahan et al., 2017). The most common informatics driven method of provider attribution is the “Four-Cut Methodology” outlined by Murray et al and shown in Table A.1.

Since the VCU EMR extract provided in this analysis does not provide information on physical exams, we omit cut 3 and only use cuts 1, 2 and 4. We further modify the “4-cut” method to focus on attribution to full-time physicians and not nurse practitioners or other members of a practice care team. When aggregating panel weights to determine weighted panel sizes, we use the attributed provider making this an important pre-processing step to any panel size estimation algorithm.

Another important consideration when conducting panel size estimation is handling extreme values and outliers. These values can be detrimental to accurately estimating panel sizes, so they must be handled appropriately. Extreme values are handled in a two step process in this analysis. First, we cap all utilization data to the 95th percentile, which controls extreme counts that may be clerical errors or unlikely scenarios. The second step to handling extreme values is to use a variance stabilizing transformation to reduce the influence of high counts. The most common variance stabilizing transformation is a logarithmic transformation. However, in count data this may cause a problem when there are several 0 or small values. An idea borrowed from high-throughput literature is to use a regularized log transformation proposed by Love et al that shrinks the influence of high counts but is not overly susceptible to low counts (Love, Huber, & Anders, 2014). Our process of handling extreme values was determined following experimentation with other variance stabilizing transformations and different capping thresholds. The combination of these two yielded the most promising clustering results.

Assessing Clusters

Summary tables were used to visually inspect the clustering, ensuring that cluster groups have similar, logical and plausible characteristics. The panel size estimation algorithm is dependent on identifying different median clusters, so we will look to see if there is sufficient separation. The summaries will also be plotted using bar plots, as a visual aid to the data analysis. For the patient burden scoring methods, we categorize the score range by quarters (i.e. scores less than 0.25, scores between 0.25 and 0.5 and so forth) to generate similar summary statistics, ensuring that scoring ranges are similar, logical and plausible. In addition to looking at cluster summaries, we also considered the clustering model in an attempt to understand how the final clusters were determined. For the k-means method we look at the decomposition of the sum of squares and for the gaussian mixture model we look at the selected variance-covariance structure.

A common method for visually inspecting clustering results is to use a t-SNE plot. This allows us to project high-dimenional data on two-dimensions in order to evaluate the spearation in the clusters. t-SNE stands for a t-Distributed Stochastic Neighbor Embedding which uses the student's t distribution (instead of gaussian) to determine similarities between two points in low dimensional space and simplifies the cost function used for gradient descent (van der Maaten & Hinton, 2008). We use the Rtsne package in R to construct the t-SNE plots (Krijthe, 2015).

Provider-level Data Analysis

Next, we conduct summaries after aggregating the relative burden of each patient at the provider level, generating a panel size. For each physician we generate summaries and percentages of the patient utilization and complexity and compare them to other physicians first across VCU health systems and second within practice. As a means of controlling for varying amounts of patients seen by different physicians, we standardize the estimated panel size using the following equation:

$$Z = \left(\frac{O - E}{\sqrt{E}} \right) \times 10 \tag{1.2}$$

where O is the log of observed number of patients and E is the log of the expected number of patients calculated from the panel size estimation algorithm. We multiply the standardized value by 10 to interpret that z-score roughly between -2 and 2. The observed value naively assumes that each patient requires the same amount of care and attention to the physician. The expected value is the weighted sum of the individual burden posed by each patient in the panel, assuming that a patient that is more likely to be sick and requires more care and attention is more burdensome than a patient who does not seek as much care and is less likely to be sick. With this context, a z-score less than 0 indicates that the physician has a panel that is expected to be more burdensome than its raw total of patients, while a z-score of greater than 0 indicates that the physician has panel that is expected to be less burdensome than its raw total of patients. We can plot the z-score of each physician to highlight which physicians are taking on more burdensome panels.

Correlation Assessment

An important piece of our data analysis is understanding how well the panel size estimate is associated with the patient level measures aggregated to the physician level. This association helps us understand which measures are most influential in estimating the total panel size of a physician. The correlation assessment has two parts: first internal correlation, regressing the estimated panel counts of the physician on the aggregations of the patient level measures used to derive the panel counts and second temporal correlation, using the model parameters from a starting time frame and applying them to generate predictions of panel sizes for physicians on a second or subsequent time frame. The internal correlation assessment helps assess whether the measures used to construct the panel size estimates are correctly correlated with the estimated values. In one circumstance this aggregates the raw totals of the data used for panel size estimation and in a separate circumstance aggregates on the median or most likely categorical measure. This helps us determine what is driving prediction among the measures used for constructing the weighted panel. The temporal correlation assessment helps assess how sensitive a change to year two data is on the panel size estimation prediction. In the temporal correlation we use two time frames of two year data. By regressing the predicted panel size estimated from the parameters of the first time frame on the previously mentioned aggregations, now done on the second time frame. In the temporal assessment we will extract the partial R^2 of the used features to indicate which features are more influential in modelling. We will also consider the Pearson correlations between the predicted second time frame panel size and the estimated panel size of the first time frame. Further to our correlation assessment, we will contrast a temporal z-score to see how the physician's panel has changed over time.

Simulated Data

The next part of our analysis of panel size estimation methods is to conduct a simulation study. The purpose of the simulation study is to assess the behavior and operating characteristics of the panel size estimation methods, given that we are repeating an experiment a large number of times across a variety of different scenarios. In a simulation study we are generating data based on a ground truth and evaluating how well each method does in reflecting the ground truth. Simulation studies are used to emulate 4 kinds of count data for healthcare utilization: the number of primary care visits, prescribed medications at the time of visit, phone calls, and specialty visits. We restrict the number of variables to simplify the simulation. In this scenario we choose to randomly generate count data whose marginals follow a generalized Poisson distribution that utilizes a scale parameter (λ) to account for dispersion in the distribution (H. Demirtas & Gao, 2020). We use the R package *RNGforGPD* to accomplish this task (Li et al., 2020). The simulated data are based on real-world data from the VCU health system. We generate data in four groups based on percentiles from the real-world data: first, a minimal utilization level where the majority of patients have a single primary care visit and no further utilization; second, a low utilization level; third a medium utilization level; and fourth, a high utilization level. We also randomly select observations to have extreme observations across the utilization variables. Our selection of extreme observations is chosen from the inverse cdf of a uniform distribution where the quantile is generated from a beta distribution. The shape and scale parameters stem from a sliding scale based on true cluster membership. Observations in the high utilization level will have higher probability of extreme observations than those in the low utilization group. We generate 10,000 datasets using this RNG scheme from which we diversify to a few scenarios in order to assess operating characteristics of the panel size estimation methods. Table A.29 provides all possible scenarios we will simulate to evaluate the operating characteristics. In total we will conduct 16 sets of simulations. The parameterizations of the simulations are described in the appendix from tables A.30 - A.45.

For each simulation scenario, we calculate the adjusted panel size of each simulated doctor using the different panel size estimation methods. From the estimated panel size of each doctor, we calculate the bias (difference between the panel size and the number of patients), the mean squared error (the mean of the sum of the squared bias), the coverage (the proportion of the number of times that the method correctly ranks the doctor), and the precision (the reciprocal of the variance of the panel size estimates). These summaries allow us to understand how the respective panel size estimation method behaves in different scenarios. Next we want to compare how effective each method is able to differentiate the correct ordering of clinician panel size. Before simulation we identify who should be ranked higher among the doctors so that we can compare the "truth" to the result of the panel size estimation method.

For only the clustering methods, we want to assess the performance of the clustering separately from the remainder of the panel size algorithm. We calculate the mean adjusted rand index and the cluster purity to externally evaluate the clustering to the true group in the simulation. Originally, the rand index is measure of similarity between two data clusterings and is closely related to measuring binary accuracy. Binary accuracy is a common receiver operating characteristic that measures the frequency of agreement over all pairs. The adjusted rand index was proposed by Hubert and Arabie 1985, which offers a correction-for-chance to the rand index by assuming a generalized hypergeometric distribution as the null hypothesis such that the two sets of clustering are drawn randomly from a fixed number of clusters and a fixed number of elements in each cluster (Wagner & Wagner, 2007). The adjusted rand index is the normalized difference of the rand index and its expected value under the null hypothesis. In our analysis we calculate the adjusted rand index from the *mclust* package in R. To construct the adjusted rand index, we create a contingency table partitioned by the elements in one clustering method compared to another clustering method (or the true classes). For each cell in the contingency table we sum the number of objects that are in common between the two clusterings in the same group. The row and column sums of the contingency table are used in the calculation of the adjusted rand index. The adjusted rand index is show in the following equation, where a_i are the row sums, b_j are the column sums and n_{ij} are the cell counts:

$$ARI = \frac{\sum_{ij}\binom{n_{ij}}{2} - \left[\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2}\right] / \binom{n}{2}}{\frac{1}{2}\left[\sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2}\right] - \left[\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2}\right] / \binom{n}{2}} \tag{1.3}$$

Cluster purity is another tool for external evaluation of clustering which measures how well each cluster contains a single class. For each cluster, we count the number of data points from the most common class and sum them across all clusters; then we divide by the number of observations (Manning, Raghavan, & Schutze, 2009).

1.3 Results

1.3.1 VCU Health Systems Data Anaysis

A summary of the demographics for the VCU Health Systems analysis is provided in table A.2. For the VCU data between January 2018 to December 2019, 4 primary care practices were represented: Hayes E Willis (n = 3791; 21%), Mayland Clinic (n = 5360; 29%), Nelson Clinic (n = 6035; 33%), and Tanglewood (n = 3050; 17%). Most of the patients from the VCU data are age 50-64 (n = 5131; 28%), hold private insurance (n = 10,284; 56%), are female (n = 10159; 56%) and white (n = 8788; 48%); note there is a significant black population as well (n = 6756; 37%). The demographics of the patients are quite dependent on the practice location. Mayland and Tanglewood clinics have a greater concentration of white patients, while Hayes E Willis and Nelson Clinics have larger black populations. There were 19 distinct primary care physicians in the analysis. There were 2330 patients who could not be attributed to an physician because they were either cared for by a nurse practitioner or another member of the clinic's care team and were removed from the analysis. The physicians are summarized in table A.4 with the number of patients per physician and the aggregate counts of utilization.

Table A.3 contains a summary of the utilization measures used from the VCU Health Systems dataset. With respect to these measures, the median medication count was quite high. This is related to how the medication count variable is constructed, it is a cumulative sum of all medications taken over the course of a 2 year period. Medications are correlated with visit count since many medications are re-prescribed per visit, so it may help to enumerate medications in the future based on number of unique prescriptions in a two year period. Another utilization measure worth criticism is the number of no shows. The max number of no shows in the data was 12 and the 75th percentile was 0. This measure is quite low and could impact the panel size estimation. Other measures had less problematic distributions, although most had a median count of zero. The mean count of these measures was much higher than the median meaning that outliers were present in this data set.

Kmeans

The kmeans process of panel size estimation yielded 4 clusters: cluster 1 contained 4,307 patients with a median visit count of 1 and a panel weight of 0.32; cluster 2 contained 3,008 patients with a median visit count of 7 and a panel weight of 2.22; cluster 3 contained 5,710 patients with a median visit count of 2 and a panel weight of 0.63; and cluster 4 contained 5,211 patients with a median visit count of 4 and a panel weight of 1.27 as seen in table A.6. Based on the box-plots and t-SNE plots (Figure A.3) there is a definite degree of separation between each cluster; depicting a 4-tier stratification of minimal, low, medium and high users of healthcare services. Table A.5 provides the per cluster summaries of the kmeans derived panels. As with the boxplots we see that cluster 2 contains most of the high utilizers. It is also of note that the within-group variability (shown in table A.8) captured in cluster 1 is low, indicating a minimal utilization group, probably patients who registered a single visit and did not return for further services. However for the rest of the groups the variability is larger, with clusters 1 and 4 having the largest suggesting that there is uncertainty when categorizing the “typical” utilization population.

In terms of provider aggregation (described in Table A.7), the panel size estimations largely followed a patient ranking, despite a few deviations. The first noticeable deviation was the Mayland clinic physician with 731 patients. This physician had a large positive z-score meaning that their patient panel did not require a lot of work per patient and hence their panel size estimation was smaller than other physicians. They had the third fewest total visits and the second fewest specialty visits among their patient panel. Another deviation was the Hayes E Willis physician with 502 patients who ranked higher than some of the Nelson clinic physicians. This physician had a large negative z-score meaning that the work per patient was much higher than expected. We can see that they had more visits and patient communication (messages and phone calls) compared to the other physicians. We notice in the provider aggregation that the z-scores are dependent on practice. Mayland clinic physicians all had positive z-scores, suggesting less work per patient. This could be a result of the patient population they serve (Mayland is in the Richmond suburbs) and the style of their practice (they could require less follow-up work per patient). Since physician panel sizes are practice specific, panel sizes should maybe best considered within the context of the practice and not across the health system as depicted here.

The correlation assessment helps us first, determine if the utilization variables are properly correlated with the panel size estimation and second, identify the variables that explain the most amount of variation. Table A.9 describes the model results and table A.10 describes the ANOVA results for the internal model. Emergency room visits and medication counts are not significantly correlated with the panel size. Both variables have issues with colinearity with other variables in the model, like visit count and specialty visit. In future models it may be worth dropping these variables when constructing the weights for the panel size estimates. No show is also noticeably the largest slope in the model. The small number of no shows could be overly influential in the panel size estimation. We see that the model explains 98% of the variability in the panel size estimation, with specialty visit and visit count accounting for the most. Messages and phone calls are the next most explanatory variables. No shows, med count and er visit explain little variability, suggesting these could be dropped. For the temporal model (table A.11 conveys the model summary and table A.12 conveys the ANOVA) we see similar trends to the internal correlation meaning that this mechanism is consistent over time. In the temporal model, the variables that explain the variability change slightly, however specialty visit still explains most of the variability. Phone calls and messages have more influence while the influence of visit count goes down. It is unclear if these changes impact the effectiveness of the model or are artifacts.

Gaussian Mixture Model

The gaussian mixture model method of panel size estimation yielded 3 clusters: cluster 1 contained 6,871 patients with a median visit count of 4 and a panel weight of 1.34; cluster 2 contained 3,554 patients with a median visit count of 1 and a panel weight of 0.33; and cluster 3 contained 7,811 patients with a median visit count of 3 and a panel weight of 1, as shown in table A.14. We can see the distribution of the utilization per cluster in table A.13. The gaussian mixture model did not yield as good separation between the clusters, where clusters 1 and 3 had very similar distributions in the box plot and did not show vertical separation in the t-SNE plot as show in Figure A.4. Tables A.15 and A.16 depict the modelling results for the gaussian mixture model using the `Mclust` package in R. The gaussian mixture model chose an ellipsoidal within-group covariance with variable volume, shape and orientation. While this covariance structure offers the most flexibility it seems to have been to conservative in estimating the clusters for the panel size estimation.

Looking at the provider aggregation for the gaussian mixture model provided in table A.17, the ordering of the panel size aggregation is similar to that of the kmeans methods, with some noticeable variations. The Mayland physician with 731 patients has a higher weighted panel size under the gaussian mixture model than in the kmeans process. Further, all of the Mayland physicians had higher weighted panel sizes compared to the k-means process. The gaussian mixture model selected fewer clusters and the weights are lower, meaning when aggregating there is less differentiation between a high utilizing patient and a “typical” utilization patient. The z-scores from the gaussian mixture model also do not deviate as far from zero with a range from -0.61 to 0.52. Again since there are less clusters and less differentiation from higher utilization patients the panel size estimations are closer to the patient counts.

Looking at the correlations in the gaussian mixture model shown for the internal model and ANOVA in tables A.18 and A.19, we see that only number of ER visits is not significantly associated with the panel size estimation. We also see that, again, no show has the largest slope of any of the covariates (0.719; SE: 0.158). This suggests having more no shows will increase the panel size estimation, but we think this is overly influential in the model. Looking at the ANOVA table, the model explains 94% of the variability in the panel size estimation. Most of the variability explained in the model comes from the specialty visit and phone call variables. However, when we see the temporal correlation (shown in tables A.20 and A.21) there is a change, the phone call variable is no longer significant but the ER visit is now significant. When looking at the ANOVA results of the temporal model, the visit count variable now explains more of the variability. This suggests that the gaussian mixture model is not consistent over time. It is unclear whether this is in fact problematic. However, since the panel weights have a tighter range between 0 and 2, we can guess that there is less panel size variability in aggregation.

Patient Burden Scoring

Since the patient burden scoring method does not produce clusters, we summarize its performance in 4 cuts: scores between 0 to 0.5 have 4853 patients, scores between 0.5 to 1 have 7081 patients, scores between 1 to 1.5 have 5065 patients and scores between 1.5 to 2 have 1237 patients as shown in Table (A.22). We can see the distribution of the utilization per cluster in Table A.23. From the boxplots and t-SNE plots (Figure A.5) there seems to be clear separation between the clusters, however we should note that this is also a structural artifact of dichotomizing the burden scores. From Figure (A.6), we see there is a skew in the scores to the left of 1, with the median score of 0.8043 and mean score of 0.8133. The group with scores between 0 and 0.5 is similar to the low cluster from the clustering methods in that it is a group within minimal utilization (1 visit and 1 medication). The patient burden score shows similarities to the *k*-means in that it recognizes a split in the middle scoring patients (between 0.5 and 1.5). There are likely a low and high middle class. The gaussian mixture model on the other hand seems to not differentiate these groups and omits an outlier class. The outlier class is small but definitely present.

For the provider aggregation (Table A.24), nearly all the z-scores are positive except for three physicians with some of the smallest patient counts. This suggests that all of the physicians have less burden per patient than expected. This is unlikely to be true but the result makes sense given the heavy left skew of the patient burden scoring. Another interesting thing to note is the top two physicians have switched in ranking by panel size estimation. This could be a result of more weight being placed on patient communication (messages and phone calls). The weights yielded from the principal component analysis from the matrix of difference between the observed utilization and typical utilization are: 0.178 for medication count, 0.168 for specialty visit, 0.157 for messages, 0.139 for visit count, 0.136 for phone calls, 0.119 for ER visits and 0.103. Messages in particular may be upweighted in the burden score, if someone has to handle more messages from a patient they are doing more work. The patient burden score seems to be more sensitive to other forms of utilization instead of pure patient count and visit count.

The correlation model for aggregated panel size estimations show no association with phone calls, ER visits and medication counts as seen in the model summary Table A.25. Similar to the kmeans model, ER visits and medication counts are likely colinear with specialty visits and visit count respectively. As for phone calls, when we look at the mean and median values within the 4 dichotomized groups we see that the number of phone calls per patient is quite small. Group 4, scores between 1.5 to 2, has a median phone call count of 8 and a mean count of 6.5. Phone calls were really only prominent in this high utilization group so its effect is minimized for most of the score range. From the ANOVA shown in Table A.26, we see that the model explains 94% of the variability in the panel size estimation. Most of the variability is explained by specialty visits, messages and visit counts. As noted before messages has more importance in the panel size estimation than in the clustering models. The no show variable still has quite a large slope, probably due to the small counts in the data. From the temporal model (Table A.27) the same three variables are not significantly associated with the panel size estimation (ER visits, phone calls, and medication counts). The temporal ANOVA (Table A.28) shows that the same three variables (visit count, messages, and specialty visits) still explain the most variability in aggregated panel sizes, however the messages variable now explains a larger proportion of the variability.

1.3.2 Simulation Analysis

Kmeans

The simulation results for the kmeans method are shown in Table A.46. The overall trend in the simulation analysis for the kmeans method is that when the means in the utilization are the same, the bias (the difference between the top doctors panel size estimation and its expected patient count) and the mean squared error are minimized. This makes sense since when the means are different between the doctors than there is a clear deviation in the work. This makes sense because there is more utilization and more healthcare interaction that is causing the burden to be higher. When the means are different it also effects the clustering performance metrics. The ARI and cluster purity (extent to which clusters contain a single class) are lower than when the utilization means are the same. The classification error rate also increases. It could be that the doctor with the higher utilization means also has more patients with extreme values of utilization which causes more error in the panel size estimation.

The other noticeable trend is that the kmeans method has a difficult time differentiating between clusters when there are more doctors (3 instead of 2), particular when the number of patients, utilization means and variances are not the same (simulation number 16). The 16th simulation has a coverage, ability to identify the true top doctor, of only 42%. The panel size estimation methods have more difficulty identifying clusters when there are more doctors to aggregate and the patients have more possibilities for variation in utilization counts. When there are three doctors with different samples the tendency is to pick the doctor that sees the most patients, especially when means can not be easily distinguished. For simulations 13 and 15 the top doc is 2 because they have the most patients, but when the means of utilization are higher for docs with lower patient counts the coverage is less complete. It no longer favors doctors with more patients, it favors doctors with more utilization. Breaking the ties of who has higher utilization becomes harder as the number of doctors increases.

Gaussian Mixture Model

The simulation results for the gaussian mixture model method for panel size estimation are shown in Table A.47. The gaussian mixture model has similar behavior to the kmeans model when evaluating the simulation results. However, the deviation from the expected patient count is less severe. The gaussian mixture model does not put as much weight into the differing classes, being more uniform in weighting than the kmeans approach. We notice that convergance in the gaussian mixture model is not always 100%, as it was in the kmeans model. There is more probability of a misclassification, especially when more physicians are involved. We note simulation 14 in particular where we only get a coverage of 0.194, where the true doctor should be doctor 1, since it has the most patients with higher utilization. It seems that in the gaussian mixture model method the weights assigned are more evenly distributed so at times its there can be some missclassification. This observation of the simulations matches what we saw in the VCU analysis, in that there is more overlap in the panel weights towards the middle. The gaussian mixture model can easily identify clear separation as in the minimal utilizers and the extreme values in the simulation, but does not create separation when the utilizations are bit closer together. When looking at the clustering metrics we notice that the ARI and cluster purity are lower when the means are different compared to the kmeans method, likewise the classification error rate is higher. It could be that the gaussian mixture model is more susceptible to mistaken clusters when there are more outliers present. This could be hampering the gaussian assumption required in the mixture model. When the means are the same, the cluster metrics are better than the kmeans model possibly because the gaussian assumption within each cluster is not as vulnerable.

PBS

The simulation results for the patient burden scoring method for panel size estimation are shown in Table A.48. The clear take-away of the patient burden scoring method is that the bias (difference from the panel size estimate to the expected number of patients) and means squared error are high. The patient burden scoring method assumes greater burden in the physician panel than the other methods because it is synthesizing the results to a score. It is not quite fair to comapre the patient burden scoring method to the clustering methods because the basis of their designs are different. It is also hard to interpret the clustering performance because the dichotomization follows the stratifications within the data. In fact we would probably expect them to be better. We would need to come up with a better way to evaluate the patient burden scoring method, since there is not much to take-away from this evaluation.

1.4 Discussion

The results of the panel size estimation analysis are not totally clear cut but there is interesting insight and suggestions that can be made for practices hoping to use an informatics approach to panel size estimation. The big take-aways are: first, a variance stabilizing transformation such as rlog improves the clustering and variance decomposition; second, analysts need to account for practice specific variation in panel size aggregation; third, selection of count variables can not be totally ignored, there can be issues with multicollinearity and too few counts among selected covariates for clustering; fourth, there is no obvious advantage to using either clustering method however the gaussian mixture model tends to be more conservative in constructing panel weights; fifth, the patient burden score does not necessarily improve upon panel size estimation but it does lead to greater consideration of dimension reduction methods to synthesizing utilization vectors to a single score; and sixth, is it possible to also incorporate patient complexity in estimating panel size estimation.

1.4.1 Variance stabilizing transformation

One thing that became clear when preparing the data for panel size estimation was how to handle the heteroskedastic nature of the data. Some patients had extreme utilization counts and others had counts of zero. We knew that there were partitions in the data beyond a non-utilizer and utilizer and want to better see clusters within correlated data. Capping outliers at a percentile helped because it removed the most extreme cases, however the high variance structure still hampered our clustering models. Variance stabilizing transformations (like a log or square root transformation) is a common pre-proccessing step in data analysis. However, with count data we are also faced with the issue of neglible (zero) counts. The rlog transformation proposed by Love et al is a tool often used in genomic analysis but works really well in this case, when there are several patients with zero utilization. The rlog transformation improved the results from clustering. We suggest using the rlog transformation, as opposed to a normal logarithmic transformation because it does a better job limiting the influence from low counts which are common in primary care data.

1.4.2 Practice Specific Variation

A major observation from this analysis is that the provider aggregation should only be compared within a practice. In the VCU data analysis we noticed that there were clear differences in panel size estimation and z-scores across practices. The Mayland clinic physicians had high positive z-scores and larger panel sizes due to seeing more patients. The physicians in this clinic see far more patients than other clinics and do not have as much follow up per patient. The demographics of the patients are also different to other practices and there could be differences in the practice due to its geographic location. Comparing panel size estimates is best done between physicians in the same practice. This is probably obvious since the point of these tools are to optimize provider allocation within a clinic. However, it should be noted that comparing panels across different practices is not advised. The practices may have different styles to primary care, they could see different types of a patients more frequently. The panel size estimation methods help give a sense of who within a practice is over or under-empaneled. In fact, we could change the z-score to be standardized with respect to the average patient panel within the practice in order to provide a more equitable comparison. It should also be noted that we only considered physicians in this analysis, however there remain other important participants to a practice care team such as nurse practitioners and DOs. Panel size estimation should be done with respect to care teams and not just clinicians.

1.4.3 Selecting Utilization Measures

One should be selective when selecting utilization measures towards estimating panel sizes. An issue with this analysis was that the medication counts were cumulative, meaning they compounded upon a primary care visit. This made medication counts highly colinear with visit count. Despite all variables having colinearity with visit count, since more utilization comes from more interaction, the medication count in particular became quite inflated. One should be careful when extracting features from the electronic medical record. It is suggested to extract medications based on eras, where a medication is counted once if it is continuously prescribed and there is no washout period. The correlation models showed this issue quite easily and perhaps the clustering models could be more accurate without extreme medication counts. Another issue was the selection of no shows as a utilization metric. While it does make sense that not showing up to an appointment triggers work to a provider, it should have a minimal influence towards panel size estimation. We saw in the correlation models that no show had a large slope probably because if a patient had 1 or 2 no shows this is a large deviation from average. It is suggested to drop this kind of variable because its spurious nature overly influences the panel sizes, while also being an awkward way of capturing utilization.

One of the reasons for suggesting the patient burden scoring method was that in the clustering methods weights are mapped based on median visit counts. While this is a plausible assumption for projecting utilization, it could potentially be a mis-specification of the level of utilization towards the provider. While we did not change this process from the original Rajkomar paper, it is worth evaluating other ways of projecting clusters to a panel weights. Perhaps this is a better use of a dissimilarity score as proposed with the patient burden scoring. More research is required in this domain.

1.4.4 Selecting Panel Size Estimation Method

Between the two clustering methods, there is no stand out winner of best method. The gaussian mixture model tends to more evenly distribute the panel weights since it is more conservative towards creating new clusters, depending on the selected covariance structure. The selection of panel size estimation method is a matter of preference. One should know that the gaussian mixture model tends to be more conservative in creating partitions while the kmeans approach is more aggresive in creating partitions. A key constraint in the gaussian mixture model approach is that we limited the number of possible groups to select to 5. This was done because the model with lowest BIC tended to always be the model with the most groups, as the number of groups increased. The gaussian mixture model may also be susceptible to over-partitioning due to model selection by BIC. However, given these limitations there was no clear cut winner for panel size estimation. The gaussian mixture model tended to perform better when there was less outliers and the kmeans approach worked better when there were clear differences between the means in the patient roster. Neither proved to be better than the other when the panels of two doctors were virtually the same.

1.4.5 Dimension Reduction

The patient burden score method can not be outright dismissed because it did provide some utility. It was able to highlight which patients were exhibiting more work to providers than others. It was quite susceptible to heavy outliers and low utiliziers, which could explain why there was a left skew in the mean burden score. Dimension reduction can be quite helpful when trying to make sense of high dimensional data. The patient burden scoring method helped reduce the dimensionality of the data and show that were more partitions in the middle of the data than the two shown in the gaussian mixture model. Patient burden scoring is a novel method designed particularly for this data question, however there are several ways that an analyst can perform dimension reduction: the most traditional principal component analysis, the t-SNE method used for assessing clusters and neural network based methods like autoencoder. It is worth exploring low dimensional projections of the data prior to clustering and panel size estimation so that one better understands the data. It is also worth exploring dimension reduction methods that can fit more seamlessly into the panel size estimation pipeline.

1.4.6 Patient Complexity

One area that we did not explore in this analysis was patient complexity. Complexity is often characterized by categorical data such as patient demographics or characteristics and either presence or absence of a drug exposure or a condition occurrence. Complexity may also be a numeric measure if one uses a patient complexity score like a CHADS2Vasc score. Complexity and baseline characteristics of a patient are also important towards panel size because we want to be able to assess what types of patient characteristics will lead to more visits over time. The issue is that adding complexity into a clustering model is not completely straightforward. The kmeans method only uses numeric data as an input, since it is based on euclidean distance. It is worth exploring how to better incorporate complexity from the EMR into the panel size estimation problem.

Chapter 2

PANEL SIZE ESTIMATION FOR MIXED DATA

2.1 Introduction

Panel size estimation is an important consideration for primary care practices because it helps them understand the number of patients they are serving and determine if they can appropriately accommodate the demands of their patient population. Panel size estimation is the enumeration from empanelment, which is the process of linking patients to a provider or a care team within a practice (Bodenheimer, Ghorob, Willard-Grace, & Grumbach, 2014; Grumbach & Olayiwola, 2015). Much research highlights the importance of empanelment in primary care, how it encourages continuity in the provider-patient relationship, leads to improved understanding of the patient population and quality of care and can help practices prioritize resources to improve clinical outcomes and reduce costs (Grumbach & Olayiwola, 2015).

While panel size estimation is considered important, not many practices implement a formal process for panel size enumeration, let alone leverage electronic medical records (EMR) to guide panel size estimation. According to a survey of primary care providers conducted by Peterson et al only about a third of physcians estimated their panel sizes, even though about 70% of practices in the survey had access to an EMR (Peterson, Cochrane, Bazemore, Baxley, & Phillips, 2015). This indicates that the resources are largely available to conduct panel size estimation but are not often used. There is limited research in regards to how to conduct panel size estimation and even less research on processes that leverage the full scope of the EMR to help understand patient populations. Panel size estimation methods that do leverage EMRs only use numeric data such as healthcare utilization, coming in the form of counts of visits or medications. While this is good information to use to determine aggregated panel sizes, it is limited in that it only represents a small amount of available EMR data. The purpose of this research is to present methods of panel size estimation that leverage as much information as possible from the EMR that can lead to informative ways of adjusted panel sizes.

The simplest approach to determine a physician panel size is to count the number of patients attributed to each provider in a practice. The issue with this method is that it treats each patient as the same; for example, a young patient with no chronic conditions who only goes to their primary care provider for a yearly wellness visit would no ascribe the same level of burden to a provider as an older patient with hypertension, takes 4 different medications and goes to their primary care provider once a quarter. Instead of treating patients as the same, we want to estimate an adjusted panel size that incorporates the relative burden that each patient poses to a provider, with patients assigned weights representative of the amount of effort it takes to treat this patient over time.

More sophisticated methods are required to estimate the relative burden of each individual patient to a physician. One such method of panel size adjustment was proposed by Kamentz et al who created groupings of patients based on age, sex, and insurance type. From these groups weights were developed based on the average amount of “work” done by a physician in the form of office visits and telephone visits (Kamnetz, Trowbridge, Lochner, Koslov, & Pandhi, 2018). An issue with this method is that it does not use any statistical model and uses a limited scope of data available from the electronic medical record. The most sophisticated approach to adjusting panel sizes based on patient-level EMR information comes from Rajokar et al who propose using kmeans clustering to generate patient phenotypes based on healthcare utilization data such as the number of primary care visits, number of other healthcare visits such as a visit to the emergency department, radiology, surgery or urgent care, and the quantity of communication between the patient and the physician in the form of messages and phone calls (Rajkomar, Yim, Grumbach, & Parekh, 2016). While currently the most sophisticated approach, the limitation of the Rajkomar method is that it does not consider patient complexity in the clustering step. Like most clustering approaches, K-means clustering, as used in that study, can only handle a single data type, which limits the amount of data we used for clustering to determine patient phenotypes. Ideally we would want to incorporate not only patient utilization, representing the amount of work done by a physician, but also their complexity, such as exposure to certain drugs, presence of chronic conditions or other observations such as smoking status or obesity. This would require clustering on mixed-type data.

The most common approaches to mixed data clustering either turn all the data into a single data type (for example using a "one-hot" matrix) or using a mixed type distance measure like Gower's distance. While these methods are practical, their performance has been question (A. H. Foss & Markatou, 2018). We aim to consider clustering methods that truly embrace the mixed characteristics of the data. We consider two methods for clustering mixed type data: the gaussian-multinomial mixture model and *KAMILA*. The gaussian-multinomial mixture model is the natural extension of model-based clustering to handle mixed data. *KAMILA* is a newer approach to mixed data clustering proposed by Foss et al 2018 which uses similar ideas to model based clustering but replaces the gaussian distribution with a kernal density estimate so that it is not reliant on parametric assumptions. This can be particularly helpful when we have count data or non-normal data, as is likely the case in aggregations from electronic medical records. By substituting the gaussian-multinomial mixture model and *KAMILA* methods of mixed type clustering for the K-means method from Rajkomar we hope to evaluate how incorporating a larger amount of data into the clustering method effects panel size estimation and determine if it is a useful addition for real world use.

In this study we also introduce a novel concept for adjusting panel sizes using mixed data called mixed patient burden scoring. This is an extension to the patient burden scoring method proposed for utilization data in Electronic Medical Records (Lavallee, 2021). A challenge with estimating panel sizes using a clustering algorithm is that we are finding a latent assignment of patient types and mapping that onto a single weight, which can have an impact on panel size estimation because we must arbitrarily select a numeric feature to construct a weight. A lingering question is whether we selected an appropriate feature for that mapping. The patient burden scoring approach attempts to condense the information from the utilization vector into a single value representing the burden possed by a patient. This value is constructed by contrasting the utilization of the observed patient with a "typical" patient, which could be based either on a summary statistic of the utilization vectors or domain-specified values. Once the contrast between the actual and "typical" patient is made, the difference is condensed and transformed into a continuous score between 0 and 2, with 0 representing a patient with minimal burden posed to a provider, a score of 1 representing a patient with "typical" burden (i.e. a single "typical" patient), and a score of 2 representing a patient with double the burden posed to a provider.

For mixed data, we need to incorporate the patient complexity represented through categorical variables. In order to do this we need to predict the burden posed by a patient if they were to have a particular set of characteristics. For example, a patient who is a smoker, greater than 65 years of age, and has hypertension would likely have higher expected utilization than a younger patient without comorbidities. Ignoring this information would treat these patients as identical, regardless of any impact these features would have on their actual care utilization. Ultimately, the mixed patient burden scoring method would distinguish between these patients and is then a more direct way to estimate panel size when including all aspects of the patient level data. We boost the score rendered from the utilization data with the complexity data to in theory provide a more informative burden score.

In this analysis we expand the Rajkomar approach to incorporate a mixed-type clustering method to replace *K-means* in identifying patient phenotypes used for panel size estimation with more sophisticated methods. Incorporating mixed-type data allows us to evaluate the full range of data available to us from the EMR. We compare the gaussian-multinomial mixture model, *KAMILA* and mixed patient burden scoring approaches for estimating panel sizes with mixed-type data. We evaluate these methods in three ways: first, we consider real world data from VCU health system; second, we conduct a simulation to study operating characteristics and performance of the methods for panel size estimation; and third, we consider a synthetic EMR dataset to evaluate how mixed data methods handle a large set healthcare features.

2.2 Methods

2.2.1 Panel Size Estimation Pipeline

In order to estimate adjusted panel sizes based on patient characteristics we must construct a pipeline where we can transform data from the EMR into panel size estimations. The panel size estimation pipeline presented here largely follows that of Rajkomar et al with adjustments to accommodate mixed data types. Unlike a simple count of patients, we want to estimate a value that corresponds to the relative burden posed by the patient to the physician. We can estimate this relative burden by using several patient features available in most EMRs. Rajkomar used only numeric features such as the number of primary care visits, specialty visits and patient communications, we wish to expand the features to include categorical patient features such as demographics and presence of comorbidities or drugs. Our panel size estimation pipeline extracts all patients with a primary care visits within a 2 year range (matching the setup of Rajkomar) and extract all features corresponding to this cohort. Important steps in the panel size estimate pipeline are: first, attributing patients to a provider; second pre-processing data to handle outliers and improve clustering results; third, selection of a method to estimate panel weights; and fourth, aggregating panels by providers and interpreting results.

Provider Attribution

The first step upon receiving the EMR data is provider attribution. Provider attribution is important because we need to accurately determine who is doing the predominant work for the patient at a practice. Often in the EMR, provider assignment for primary care providers are missing or misleading. Inaccurate attribution leads to inaccurate panel sizes and misleading results about the amount of work providers are doing. To help reduce this potential inaccuracy, attribution can be informatically derived from EMR extracts, with the most common way attribution approach being the 4-cut methodology outlined by Murray et al and shown in Table B.1 (Kivlahan et al., 2017). This heuristic approximation for who could realistically be identified as a patient's primary care physician (PCP) is a common-sense approach, even though patients in the US are free to select their PCP, regardless of who they actually see. In our situation, the provider who performed the last physical was not available so we only use cuts 1,2 and 4. While we acknowledge it as a limitation, in this analysis we only attribute to providers who have an physician, even though practices often use nurse practitioners or other care providers to serve patients' primary care needs.

Data Transformation

Numeric data from the EMR is typically extracted as count data, such as the number of medications or the number of primary care visits for a patient. These counts can have outlier values as a result of clerical errors or patients with heavy utilization in comparison to central tendency in the patient populace. These extreme numeric values can be overly influential during the clustering phase, so we cap all values up to the 95th percentile, replacing observed value exceeding the 95th percentile with this new max value. This removes most extreme values without altering the distribution of the observed data, which is important because high utilizers are a likely group of interest. There are other methods of controlling for outliers such as using data up to a certain standard deviation from the mean, but we erred toward preserving as much of the original observed data as possible.

An issue with poisson count data is that the mean and variance are equivalent. This is a problem when observations with high counts have exceedingly higher variability than observations with low. This leads to poor approximation of the probability distribution by the Poisson distribution. Another issue is that measures with high counts tend to be positively correlated with other count measures. For example, patients that only go to the doctor once will have a lower variance of utilization in other aspects like medication counts and phone calls since they are engaging health services less. However someone who goes to their primary care doctor a lot will have a lot more variance with how they interact with health services. The variance is non-constant and mean-dependent, which affects our, analytical methods by giving a greater emphasis to high count groups than lower count groups.

Variance stabilizing transformations can help reduce these limitations on count data, and in turn make them more suitable for use in clustering methods. A common variance stabilizing transformation with count data is a logarithmic transformation, which is the natural link function for Poisson data in generalized linear models and is effective in stabilizing the variance in data with heavy spread. A key issue with the log transformation is that while it helps stabilize the influence of highly expressed counts, it is susceptible to overweighing low expressed counts or zero counts (Love, Huber, & Anders, 2014). To balance the control of high and low counts and their influence in analysis, use the regularized logarithm transformation (rlog) on utilization counts, like in gene expression data from which it was originally used, it follows a Poisson distribution and shrinks low count values from different observations together so as not to generate arbitrary signal, which commonly occurs when log transformations over-emphasize zero counts (Love, Huber, & Anders, 2014). The *rlog* transformation is provided from the *DSeq2* package in R from Bioconducter. We should note that in future analyses researchers should consider other transformations that limit the influence of zero counts like *rlog*.

Issues with mixed data clustering

After transforming the patient-level data in preparation for analysis, the next step is to determine a means of estimating panel weights. If we wished to estimate panel size using patient complexity and categorical risk factors, we would need to augment the clustering approach to handle a mixture of numeric and categorical data types. For clustering mixed-type data, including both quantitative and categorical features, there are three generally established approaches: first, converting categorical data to numerical data (typically by dummy coding) and then using a single data type clustering approach like K-means (Hennig & Liao, 2013); second, using a dissimilarity function designed for mixed-type data like Gower's distance (Gower, 1971) with the K-medoids PAM algorithm (Kaufman & Rousseeuw, 1990); and third, using a gaussian-multinomial mixture model (Hunt & Jorgensen, 2011). These standard approaches exhibit varying limitations furthering the difficulty of mixed-data clustering. Despite being the simplest and most common approach, the dummy coding is problematic due to dimensionality issues (i.e. creating a one-hot matrix for a very large number of features) or by making a non-trivial choice of weights to represent categorical levels (A. Foss, Markatou, Ray, & Heching, 2016). Say a categorical variable with l levels could be dummy coded by indicators $0 - c$ where 0 indicates absence and c_l indicates presence of level of the categorical variable. If c is improperly determined, it can over or under influence the categorical variables compared to the continuous variables in determining the final set of clusters (A. H. Foss & Markatou, 2018; A. H. Foss, Markatou, & Ray, 2019; A. Foss, Markatou, Ray, & Heching, 2016). Finding the ideal weighting scheme for categorical variables is difficult if not impossible. Using a mixed-type distance function like Gower's distance presents the same problem as weight selection for numerical coding and may be computationally intractable for large data sets that require large storage (A. Foss, Markatou, Ray, & Heching, 2016).

Gaussian Multinomial Mixture Model

Model-based approaches have been present for several years in mixture model literature (Hunt & Jorgensen, 1999, 2011; Krzanowski, 1993; McLachlan & Peel, 2000; McNicholas, 2016), and offer a natural means for handling mixed-type data as long as parametric assumptions are met. The gaussian-multinomial mixture model for mixed data can be applied to panel size estimation because it handles numeric data using a gaussian distribution and categorical data using a multinomial distribution, two natural choices for those data types. More specifically, we can partition the numeric information into levels comprised of groupings of the categorical variables, thus establishing “weak” local independence between the numeric information at different categorical groupings (Hunt & Jorgensen, 1999). The parameters of the gaussian-multinomial mixture model can be solved using the EM algorithm with the optimal number of clusters determined by the Bayesian Information Criterion (BIC(), selecting the model with lowest BIC. This model can be done in R using the *flexmixedruns* command in the *fpc* package (Hennig, 2020). The major drawback to model-based clustering is that parametric assumptions must be met. Some other issues with model-based approaches are consistent with any approach that uses the EM algorithm; for instance there is a possibility of getting caught a local maxima leading to slow convergence, there could be undefined objective functions such as a zero variance that make it difficult to converge, and sometimes required numerical integration slows down convergence.

KAMILA

In our data we clearly have poisson distributed counts from the EMR so we also need to test robust alternatives to the gaussian-multinomial mixture model, the assumptions for which are likely to be violated. One such method is called *KAMILA*, which is a semiparametric approach to mixed data clustering (A. Foss, Markatou, Ray, & Heching, 2016). Foss et al describe the *KAMILA* algorithm as a solution to three main issues that exist among mixed type clustering: first, variables are not transformed to a single data type resulting in loss of information; second, it accounts for a better balance between categorical and continuous measures without the use of arbitrary weighting; and third, it does not require a strong parametric assumption (A. Foss, Markatou, Ray, & Heching, 2016). The *KAMILA* approach leverages aspects of K-means clustering and the finite mixture model. *KAMILA* uses a similar formulation for the gaussian-multinomial mixture model but instead of modelling the numeric data using a gaussian distribution it uses a kernal density estimate so that it is not dependent on the underlying distribution assumption ((A. Foss, Markatou, Ray, & Heching, 2016)). Full details of the theoretical approach to *KAMILA* can be found in Foss et al and are not covered in detail in this paper. There are other robust methods to mixed type clustering such as the weighted K-means approach from Modha and Spangler, however we did not incorporate those methods in this analysis because of limitations expressed by Foss; these could be areas of future research for panel size estimation. The *KAMILA* approach is available as an R package, *kamila*, which used in this analysis (A. H. Foss & Markatou, 2018).

A common concern for clustering is determining the number of clusters, particularly for approaches like K-means and *KAMILA*. There are variety of recommendations for determining the number of clusters ranging from domain expertise, minimizing within cluster dissimilarity or using a clustering statistic (Hennig & Liao, 2013). Foss shows a preference towards the prediction strength algorithm from Tibshirani and Walther, which is used in this analysis. The prediction strength algorithm leverages the idea of spliting data into a testing and training set. The true number of clusters should match in the training and testing set. The prediction strength algorithm helps determine the proportion of observation pairs in the testing set that are also assigned to the same cluster by the centroids of the training set, per cluster (Tibshirani & Walther, 2005). Typically, the prediction strength algorithm is run for several cluster number k and selects the largest k whose prediction strength value is greater than 0.8 (Tibshirani & Walther, 2005). A drawback of this method is that is usually favors a smaller number of clusters (A. H. Foss & Markatou, 2018).

The two clustering methods (gaussian multnomial mixture model and *KAMILA*) can be used in a general pipeline for panel size estimation of mixed data. The algorithms for each process are described in the appendix B.1 which summarize the key decisions and steps made to determine adjusted panel sizes.

Mixed Patient Burden Scoring

These clustering approaches suffer limitations in their application to determine weights for panel size estimation. A cluster is a latent assignment of similar patients that must be mapped to a weight that is numerically aggregated at the provider level. This is an indirect method and assumes that the mapping feature chosen is appropriate for determining the weight. The novel patient burden scoring approach attempts to provide a more direct approach to contextualizing the relative burden posed by a patient to a provider. Originally, the patient burden scoring approach was constructed as an alternative to panel size estimation using K-means clustering for numeric data. In this analysis we extend the patient burden scoring approach to handle mixed type data.

The patient burden scoring approach is a domain-specific method of estimating the relative burden posed by a patient in an easily interpreted score. We construct a score between 0 and 2 representing burden of a patient to a provider, taking multiple values based on the patient level information. A score of zero corresponds to negligible burden, a score of one represents typical burden to a provider and a score of two presents the most burden to a provider. Thinking of it simplistically a score of 2 would be "double" a patient, in terms of effort for the provider. We calculate these scores in three steps: first, we contrast the vector of utilization with a vector of "typical" utilization; second, we synthesize the vector of differences into a single value; and third, we scale this value to an interpretable score between 0 and 2. Our first decision – to determine a "typical" value – can involve either a domain-specific value or a summary statistic for each utilization measure. For instance, a practice manager can supply a series of values that they would "typically" see from a patient at the practice in terms of visits, number of medications and any other numeric vector used to calculate weighted panels. These "typical" values can also be data driven through the use of a summary statistic such as a mean, median or other percentile. Our choice of reference vector of "typical" values causes a significant amount of sensitivity in the downstream analysis. In this analysis we use the means of the utilization vectors as a straightforward approach. Our second major decision is to determine how we want to synthesize the vector of differences into a single value. The simplest way to synthesize is to sum the differences across the vector, though it is possible that the contribution of different features could be more significant in some features than for others. Alternatively, we can derive weights for each feature and perform a weighted summation. Weights can be derived either from the proportion of contribution from the data matrix or by something more complex such as taking the square of the vector of the first principal component of the data matrix. In this analysis we conduct our weighted sum based off the first principal component.

To account for information provided by the categorical features extracted from the EMR in the patient burden scoring approach, we view patients in terms of combinations of categorical features. A provider would care for a elder patient with several comorbidites differently with how they would care for a younger patient with no comorbidities. We incorporate the information from the categorical complexity based on prediction by regressing each utilization individually with the categorical features and extracting the predicted values from the model for each patient. We then add the vector of predicted utilizations to the original utilization, adjusting the data based off the predicted risk from baseline. We then take the new utilization data and run the patient burden scoring algorithm as depicted above. If we are considering several features of complexity to predict the utilization, we could use regularization techniques to attempt basic feature selection, minimizing the contribution of negligible complexity variables. Our mixed patient burden approach is described step by step in algorithm B.2.

Provider Aggregation

The last step is to aggregate the weighted panels by provider, which provides the quantification of the panel. As previously mentioned, the naive approach to panel size estimation is to simply count the number of attributed patients. However we know that each patient is different in the amount of work and burden they will pose to the provider. Therefore, the panel size estimation pipeline as designed by Rajkomar involves a weighted summation, where the relative burden for each patient is multiplied by the “average” visit count of their cluster, and then averaged across all attributed patients. Though we want to compare providers by their panel size estimation, this comparison can be misleading because panel size estimates will be highly correlated with the raw number of patients a provider sees, which is dependent upon their clinical status as full-time or part-time. Thus, comparisons of even these adjusted panel sizes between providers may not be equitable since it is possible that a provider who is on less hours or time is still being overworked relative to a full-time provider. Therefore, we calculate a z-score in order to contrast a burden metric that is not dependent upon the number of hours worked. The z-score is calculated by taking the log of the observed number of patients and subtracting it by the log the expected number of patients calculated by the panel size estimation, with the difference divided by the square root of the log of the expected number of patients (the estimated panel size) and multiplied by 10 (as a scaling factor for interpretation). this transformation permits the direct comparison of providers accounting for differences in raw patient load. A negative z-score is interpreted as a provider with a more burdensome panel compared to the number of patients they see, while a positive z-score is interpreted as a provider with a less burdensome panel compared to the number of patients they see.

Another important consideration in provider aggregation is mapping the clusters to a weight. This only impacts the clustering methods (gaussian-multinomial mixture model and *KAMILA*). Clustering gives latent assignments to groups of similar observations, which is an arbitrary value with no quantitative meaning other than to identify “like” patients. The problem is to map this assignment to a weight or numeric value we can use to aggregate at the provider level. In this analysis we follow the same process as Rajkomar et al, who took the median primary care visit count of each cluster to be a scaling factor to quantify the cluster weight. The weight vector was calculated by taking the total number of patients and dividing it by the dot product of the median visit count and the cluster size (Rajkomar, Yim, Grumbach, & Parekh, 2016). The mapping feature is an important assumption in this analysis. We agree that the number of primary care visits is the most logical mapping feature, however this does not utilize any information from the complexity. Basing the weights on the primary care visits could cause misleading aggregations, since we do not know which feature to use to represent weights. While we stick to the Rajkomar approach, future work could look to other ways of projecting the cluster assignment onto weights used for provider aggregation.

2.2.2 Data

VCU Health Systems Data

We received two-year extracts of EMR data from VCU Health Systems for four primary care practices in Richmond, Virginia: Hayes E. Willis Health Center, Nelson Clinic Family Medicine, Mayland Medical Center Family Medicine and Tanglewood Family Medicine. We received new extracts every quarter, however in this analysis we only used the two-year extracts from January, 2018 to December, 2019 and from October, 2018 to September, 2020 to allow for temporal contrasts. We did not have EMR direct access and were provided a limited number of features for the patient-level data. Utilization features included the number of primary care visits, cumulative number of medications prescribed, number of specialty visits, number of emergency department visits, number of missed appointments or "no shows," number of portal messages, and number of phone calls with the provider. The categorical features provided were limited to demographic information such as patient sex, age, race and insurance type (private, medicare, medicaid or none). Provider information was also received including name and practice location. We masked provider names in the analysis.

We evaluated this real-world EMR data in three major sections: first, we summarized and evaluated the clusters from the various methods; second, we performed the provider aggregation; and third we performed a correlation assessment of the aggregated panel sizes at the provider level against the aggregated numeric features. Our cluster evaluation consisted of a balance of visualization, summary tables and model output information. We visualized clustering in three ways: first, we used a simple boxplot of the numeric features stratified by cluster; second, we used a stacked barplot of the baseline demographics per cluster; and third we looked a map of low-dimensional rendering of the data highlighted by cluster identification.

We received two year extracts of EMR data from VCU Health Systems for four primary care practices in Richmond, Virginia. These practices are Hayes E. Willis Health Center, Nelson Clinic Family Medicine, Mayland Medical Center Family Medicine and Tanglewood Family Medicine. We received a new two year data extract every quarter, however in this analysis we only used the two year extract from Janurary 2018 to December 2019 and October 2018 to September 2020 as a temporal contrast in the analysis. We did not have direct access to the EMR so we were provided a limited number of features for the patient level data. The utilization data provided was: the number of primary care visits, the cumulative number of medications prescribed, the number of specialty visits, the number of emergency department visits, the number of no shows, the number of portal messages and the number of phone calls with the provider. The categorical data provided was limited to demographic information such as: patient sex, age, race, and insurance type (private, medicare, medicaid and none). Provider information was also received including the practice location of the provider and their name. We masked the provider names in the analysis. Low-dimensional rendering was achieved t-distributed stochastic neighborhood embedding (t-SNE) plots (van der Maaten & Hinton, 2008)), which are a common dimensionality reduction method used to visual data, particularly for clustering performance. We attempt to distinguish unique bands or groupings of clusters and ensure that the clusters do not overlap or “bleed” into each other.

We used the R package Rtsne to perform the dimensionality reduction of the full data set and ggplot2 to plot the data (Krijthe, 2015; Wickham, 2016). Two summary tables are provided, one for the continuous data and a second for the categorical data. The continuous data summary table contained the min, max, 1st and 3rd quantile, the median and mean of the measures. For the categorical table, we provide the number of observations and the percentage of the category relative to the cluster.

Next, we report the provider aggregations. In these summaries we provide the number of patients seen by the provider, aggregated panel size estimation, z-score, aggregate counts of the utilization per provider, and number of demographic variables observed per provider. For the correlation assessment we regress the aggregate panel size estimations for each provider on the aggregate utilization counts. From here we produce model summaries using the `broom` package (Robinson, Hayes, & Couch, 2021). We estimate two correlations: first an internal correlation by regressing the panel size estimations on the utilization counts of the same year, and second a temporal correlation where we regress a new set of aggregated panel size estimates on the utilization counts from a different time-frame (October 2018 to September 2020). The new set of panel size estimates, aggregated by provider, were created by taking the clustering information derived from the first time-frame (January 2018 to December 2019) as inputs to calculate the panel weights for the second time-frame, which helps us evaluate whether the panel size estimations are consistent over time. If the same regression coefficients are prioritized in the temporal correlation we know that variability is consistently captured by the same features across time. In the correlation assessments we look at the partial R-squared values from the regression to help understand which features explain a greater proportion of the variability in the panel size estimates. We can look at these correlations to help consider variable selection for future model, understanding if there is multicollinearity affecting the model or if there are unnecessary features that can be dropped.

Simulated Data

In addition to the real world data, we want to assess the efficiency, accuracy and reproducibility of these panel size estimation algorithms behave under 10,000 repeated simulations. To randomly generate a mixture of correlated poisson data (to capture healthcare utilization) as well as correlated multinomial data representing the patient demographics, we adapt the approach by Demirtas et al 2012 used to generate data from correlated mixtures of distributions representative of public health data (Hakan Demirtas, Hedeker, & Mermelstein, 2012). We augment to approach from the R package `PoisBinOrd` to include the generalized poisson case provided in the the R package `RNGforGPD` (Inan, Demirtas, & Gao, 2021; Li et al., 2020). We vary the randomly generated data to accomodate for 16 situations defined in Table B.28. These varying situations give us some context on how panel size estimation performs in varying scenarios adjusting for the number of providers, the number of patients seen, the mean in utilization and the proportion of demographics. More details of the simulation are provided in the appendix.

We evaluate the simulations by considering the clustering performance and basic operating characteristics. For our operating characteristics we define our "true value" as the number of patients for the doctor expected to have the largest panel. From this perspective we determine the bias, mean squared error, and coverage of the panel size estimation algorithms. We also consider coverage, which is simply the proportion of simulations where the top doctor is correctly identified by the panel size estimation algorithm. For the clustering performance, we consider three common performance measures: the adjusted rand index (ARI), cluster purity, and classification error rate.

Synthea synthetic EMR Data

A problem we faced analyzing the VCU data is the limited scope. The VCU data was a fixed extract from the VCU Health Systems EMR delivered for analysis, so we could consider additional healthcare features for determining panel sizes, being limited to patient demographics like sex, race, age and insurance type. While these are helpful, there are other features of complexity such as indicators of comorbidities, taking specific drugs and other kinds of features available to us from the full EMR. In order to test the panel size estimation process using a larger set of features, we created a synthetic EMR dataset, using the open source tool Synthea. The EMR dataset was rendered in a way that it would reflect the patient demographics of Richmond, Virginia based on census data. Synthea is a program that first generates a defined number of patients at a specified location and then it simulates a series of events and encounters. These events and encounters are tallied based on healthcare modules that probabilistically issue health related events and determine a care plan based on the disease status (Walonoski et al., 2017). As a result of this process, Synthea creates a realistic synthetic EMR database for evaluating methods and constructing software solutions for analyzing EMR data without needing or having to risk using the real data. Access to real EMR has several logistical and privacy hurdles, so Synthea offers a flexible solution to construct proofs of concept, although the downside is that it is not real data.

Another important step made when analyzing the Synthea data, was mapping the EMR data to the OMOP common data model (CDM) (Sciennces & Informatics, 2021), which is an informatics model designed for longitudinal observational databases (such as EMRs, claims and registries) and sets a standardized representation of the healthcare data. If we construct an analytic model and method on healthcare data targeting the OMOP structure, then we can use this same analytic approach on any other data source mapped to the OMOP CDM. This avoids creating customized queries and analysis to source data, which can not be reproduced in a different database. Mapping the Synthea data to the OMOP CDM also allows us to use open source software from the OHDSI community. OHDSI is an open science initiative centered around building tools and generating evidence on longitudinal observational data mapped to the OMOP CDM. OHDSI has created a series of tools that researchers can use to extract and analyze data in the OMOP CDM. For this analysis we leverage two tools in particular *Capr*, an R interface to constructing cohort definitions on the OMOP CDM; and *FeatureExtraction*, a tool to create and extract features from the OMOP CDM (Lavallee, Evans, & Black, 2021; Schuemie, Suchard, Ryan, Reps, & Sena, 2021). We use *Capr* to define our cohort of primary care visits in Synthea and *FeatureExtraction* to define a large set of healthcare features that we wish to use to estimate panel size. Conducting the analysis using OMOP and OHDSI allows us to conduct tests on panel size estimation that can be reproduced by other interested sites who have their data mapped to the OMOP CDM.

Analyzing the synthetic EMR data generated from synthea gives us an opportunity to evaluate the panel size estimation methods if we had access to the entirety of the EMR data. First, we created a cohort of patients over the age of 18 who had an outpatient visit between January 1st 2018 and December 31st 2019. We stopped observing patients at December 31st 2019 so as to only include this 2 year time period. Following cohort creation, we extracted utilization and complexity variables from the patients in the cohort. The utilization variables extracted from the cohort include: the number of condition eras, drug eras, procedures, observations, inpatient visits, outpatient visits, ER visits and measurements within the past 730 days (equivalent to 2 years). In terms of patient complexity variables we extracted numeric comorbidity scores like the Charlson Index (Romano adaptation), Diabetes Comborbidity Severity Index and the CHADS2VASc. We also extracted several categorical variables from Synthea ranging from patient demographics, chronic comorbidities, and drug prescriptions. All of the complexity features were based on an indicator of occurrence within a two-year period from baseline. A full list of these features is included in the appendix.

Our analysis of the synthea data is similar to how we analyze the VCU data. We analyzed the clustering results and summarized the continuous and categorical data to inspect how the results from clustering impacted aggregation at the provider level. Since the Synthea data are not real and are only an emulation of EMR data, the interpretations from this simulation study should not be extrapolated upon.

2.3 Results

2.3.1 VCU Health Systems Data Anaysis

Tables B.2 and B.3 provide summaries of the continuous and categorical variables used in this analysis. From January, 2018 to December, 2019, there are 18,236 unique patients seen and 19 unique physicians. Additionally 2,330 patients were attributed to a nurse practitioner or other provider because there was no more information available. The patients were seen in four primary care practices: Hayes E Willis (n = 3791; 21%), Mayland Clinic (n = 5360; 29%), Nelson Clinic (n = 6035; 33%), and Tanglewood (n = 3050; 17%). Patients had four different types of insurance coverage: private (10,284; 56%), medicare (3,559; 20%), medicaid (3,295; 18%), and other/none (1,098; 6%). The two demographic variables provided in the VCU data were race [black: 6,756 (37%); white: 8,788 (48%), and other/unknown: 2,692 (15%)] and sex [male: 8,077 (44%) and female: 10,159 (56%)]. The majority of the population in the VCU primary care data had private insurance and were white females. Age was dichotomized into five categories: age 0 to 17 with 1,651 patients (9.1%); age 18 to 34 with 3,848 patients (21%); age 35 to 49 with 4,081 patients (22%); age 50 to 64 with 5,131 patients (28%); and patients age 65 and greater with 3,525 (19%). Most patients in the VCU primary care dataset were in the age 50-64 bracket, though the representation is age was not overly dominated by any one category. As we are limited by the patient complexity variables available to us (insurance, sex and race), we suggest that other analyses use more complexity measures from the EMR in mixed type panel size estimation. From the seven continuous utilization variables considered in the analysis, only three had a non-zero median (visit count: 3, medication count 8, and specialty visit 2). The continuous variables had some extreme values but they were capped at the 95%. Still there remain several extreme values in the data set which could affect the clustering. While this suggests that the 95th percentile was itself an outlier for some measures, we did not alter our decision to cap at the 95th percentile. The medication count measure in particular was constructed based on an accumulation per visit, and so exhibited high counts.

KAMILA

The results of VCU data analysis from the *KAMILA* approach can be found in the appendix under tables B.5 - B.12 and figure B.3 of results. Using the *KAMILA* approach the VCU data was partitioned into five clusters: cluster 1 contained 2,853 observations and a weight of 2.268; cluster 2 contained 2,641 observations and a weight of 0.972; cluster 3 contained 3,977 observations and a weight of 0.324; cluster 4 contained 3,438 observations and a weight of 1.296; and cluster 5 contained 5,327 observations and a weight of 0.648. We can see that cluster 1 holds the most utilization and some patient characteristics that we would likely lead to more required care such as medicare insurance and patients older than 50. Cluster 3 had the smallest weight and its utilization and complexity seemed to correspond. Cluster 3 had a median visit count of 1 and a median medication count of 1, while the other utilization metrics were zero. There were more patients in the age 0-17 category in cluster 3 and several patients with no insurance. The boxplots show definite separation between clusters 1, 3 and 5 as these would be the high, minimal and low utilizing categories respectively. Interestingly clusters 2 and 4 have quite a bit of overlap, where the major deviation is in the messages utilization variable. Cluster 2 has a higher median message count than cluster 4, where as the other metrics are largely similar.

The t-SNE plots are a projection of all features (both utilization and complexity), which causes a tilt in the data map which can be interpreted as series of U-bands facing “northeast.” Cluster 3, the group with the lowest panel weight, is depicted in the outermost green U-band. Cluster 5 is the second lowest panel weight and is depicted by the purple U-band. Clusters 2 and 4 (yellow and blue respectively) make up a mixed band that difficult to distinguish, suggesting that these two categories might both represent different subsets of patients with medium utilization. If we look at the differences between clusters 2 and 4 in terms of demographics, cluster 2 contains more younger (18 to 49) white patients on private insurance, whereas cluster 4 contained more older (greater than 50) black patients with medicare and medicaid insurance. The *KAMILA* method seems to do well incorporating a balance between continuous and categorical covariates in determining clusters. Another interesting difference between clusters 2 and 4 is that the median visit count (the metric used to map the cluster to a panel weight) is higher in cluster 4, though it is unclear if cluster 4 is more burdensome than cluster 2. This brings doubt as to whether mapping the panel weight by median visit count is the most prudent choice, especially when we add patient complexity into the equation. The *KAMILA* method seems to have more difficulty distinguishing the middle category of patients in terms of burden to the physician. Finally, cluster 1 is the top band depicted in red. This cluster is clearly on the top, however there are some interesting blends with clusters 2 and 4. It is interesting to consider how much the complexity distorts the clustering. Clearly they are adding a bit more explanation of the variability, but it is unclear if the complexity variables have a significant contribution to determining panel sizes.

Looking at the patient aggregation, the panel size estimation is not an exact ranking of the number of patients the physicians see. For example, the physician from Hayes E Willis with 831 patients ranks higher than a physician from Nelson Clinic with 896 patients. Looking through the patient panels, the Hayes E Willis physician has more patient visits and medication prescriptions while handling a panel with more government insurance. Recall that the weights for clusters 2 (0.972) – which had more messages and less visits – and 4 (1.296) – which had more visits, the mapping by visits may have labeled the panel for the Hayes E Willis physician as more burdensome panel. Further, we see that only three physicians had positive z-scores (meaning the burden per patient is less than the number of patients they saw) all of which came from the Mayland clinic. This would suggest that practice adds a degree of variation when comparing panel size estimation in the VCU health system, the implication being that a doctor from Mayland clinic is not necessarily comparable to a doctor at Nelson clinic. This could be due to style of care or that the patient populations are systematically different. If we compare the Mayland physicians against each other we see that the physician with a panel size of 532.07 has more room to accept more patients over time. If we were to rank the panel size estimates within each practice, we see that this basically follows a patient number ranking, reinforcing the idea that practice characteristics have a large impact on panel size estimation and that panel size estimation should be conducted within the context of the practice.

Finally, looking at the correlation assessment the medication count and ER visit variables are not significant in the model. These two variables may have issues with multicollinearity with the visit count and specialty visit variables respectively. It is also worth noting that the no show variable has quite a high slope value. The no show metric has small counts and so may be overly influential in the model if there are patients with particularly large values. When looking at the ANOVA model, the specialty visit and visit count variables explain the most amount of variability in the model. With all the utilization variables, the model explains about 96% of the variability in aggregated weighted panel sizes. While the complexity variables were not included in this model, it shows that the complexity variables explain little of the variability in the model. From the temporal model, there is consistency over time. The ER visit and medication count variables remain not significant. From the ANOVA model, specialty visit and visit count still explain most of the variability in the model. However, messages and phone calls no longer explain a larger amount of the variability in the temporal model.

Gaussian Multinomial Mixture Model

We summarize the results from the gaussian multinomial mixture model in the appendix in tables B.13 - B.20 and figure B.4. The gaussian multinomial mixture model approach to the VCU data also partitioned into five clusters: cluster 1 contained 3,207 observations and a weight of 2.097; cluster 2 contained 1,670 observations and a weight of 0.300; cluster 3 contained 2,807 observations and a weight of 0.449; cluster 4 contained 4,825 observations and a weight of 0.599; and cluster 5 contained 5,727 observations and a weight of 1.98. The gaussian multinomial mixture model shows a clear gradual increase in the weights across the clusters from 2 to 5, where cluster 2 is the lowest and cluster 5 is the second highest after cluster 1. The boxplots of the utilization show that most of the extreme values fit in cluster 1, while most of the patients with negligible utilization were in cluster 2. In terms of complexity, cluster 1 contains mostly older patients covered by medicare. Cluster 2 contains mostly younger patients with no insurance or medicaid. Interestingly clusters 2 and 3 have the same visit count, but cluster 3 contains patients that had more median medication prescriptions. From a complexity stand point, clusters 2 and 3 are very similar meaning the utilization portion of the model explains more of the final cluster than patient characteristics. Cluster 4 also has mostly low utilization and its complexity covers mostly young white males on private insurance. Cluster 5 seems to cover the “middle” kind of patient, those that go to the doctor frequently but not to an extreme degree. This population had 4 visits, 14 medications, and 4 specialty visits. Their demographics were mostly older female patients covered by medicare. Cluster 1 is again the group with the largest panel weight, however compared to the KAMILA approach, the weight of this group is less. While the vectors are similar, the process of determining the weight depends on the median and sizes of the other groups. Since the gaussian multinomial mixture model splits the minimal utilization group into two clusters, it lowers the weights of the other clusters.

Looking at the t-SNE plots we see a similar story to the *KAMILA* method, with the t-SNE projection interpreted in U-bands facing northeast. Clusters 2 and 3 (yellow and green, respectively) are on the outer rim. There are few points in the t-SNE plot with a designation of cluster 2 because there are likely several patients with identical utilization. In terms of their demographics, most of these patients are young with no specified insurance, characteristics known with limited healthcare interaction. Cluster 4 (in blue) is the next rim, followed by cluster 5 (purple) and cluster 1 (red). Cluster 1 is clearly the group with largest demand for healthcare resources. This group contains persons with high utilization and mostly older medicare patients who are thought to have a lot more need for healthcare interaction. The gaussian multinomial mixture model may be over-splitting the low utilization groups (ex. groups 2 and 3), but does a better job at distinguishing the patients in the middle.

Aggregating the panel size estimation created from the gaussian multinomial mixture model per physician gives identical physician ranking as the *KAMILA* approach. The difference is the aggregated weighted panels are different between the two methods. For example, the Mayland physicians have larger combined weighted panels under the gaussian multinomial mixture model approach, but the Nelson clinic and Hayes E. Willis physicians have smaller combined weighted panels compared to the *KAMILA* approach. Again, this raises the concern of comparing panel size estimation between practices. The Mayland and Tanglewood practices see different types of patients compared to the Hayes E Willis and Nelson clinic. When adding the patient demographics into the model it is further emphasizing the differences in the patient populace that each practice serves. There are less patients with high demands at Mayland and Tanglewood. Under the gaussian mixture model this increases their panel sizes because the weights from cluster 1 (the group with projecting the most healthcare utilization) are smaller and the weights from the middle clusters are increased. Mapping the weight based on a single utilization, as done by Rajkomar, may have a negative impact on the overall process. While visit count is the logical best measure to map this weight, it could also not be the best choice depending on the circumstance.

Looking at the correlation tables we see a similar story to the *KAMILA* method. The ER visit and medication count variables are no longer significant in the aggregated panel size model. The no show variable has a large slope, meaning that a single no show has a greater effect on panel size estimation than it probably should. From the ANOVA, we see that specialty visit and the visit count explain most of the variability in the aggregated panel size model. The model expalins about 97% of the variability. This would suggest that the patient demographics have far less influence on the panel size estimation than the utilization vectors. From the temporal correlation we can see that the model is consistent over time since the same two variables are not significant (er visit and medication count). Similar to the temporal *KAMILA* model, the specialty visit and visit count variables explain most of the variability in the temporal model while the messages and phone call variables have increased in the amount of variability they explain.

Mixed Patient Burden Scoring Method

We summarize the results from the mixed patient burden scoring approach to the VCU data analysis in the appendix in tables B.22 - B.27 and figure B.5. The mixed patient burden scoring method created scores ranging from 0 to 2, which were further dichotomized into four categories: group 1 contained scores between 0 and 0.5 (n = 2,829), group 2 contained scores between 0.5 and 1 (n = 7,498), group 3 contained scores between 1 and 1.5 (n = 6,391) and group 4 contained scores between 1.5 and 2 (n = 1,518). The median score was 0.926 and the mean score was 0.9331, however the 75th percentile was a score of 1.22 meaning that most of the data is gathered in the middle. Scores closer to 2 represented patients who accessed healthcare way more frequently than typical. From each cut in the scores, we see a noticeable increase in the utilization vectors. The first three cuts largely reflect what is seen in both the *KAMILA* and gaussian multinomial mixture model. The first cluster is the minimal utilization group, the second cluster is the low-medium and the third is the high-medium group. The last group in the mixed patient burden scoring method has much higher utilization than the other methods. This would suggest it is way more susceptible to outliers. This may not be completely a bad thing because we know there are cases of patients who require much greater attention regardless of their demographic characteristics. As for the patient demographics, they also followed how we would think most patients in those categories would behave. Young patients with no insurance went to the doctor the least, while old patients on Medicare went to the doctor the most. The biggest group, cluster 4, has much less percentages of the complexity because it is a smaller group compared to the other methods. This group consists of mostly major outliers. The boxplots reinforce this point showing increasing utilization for each group. In the mixed patient burden score approach, there seems to be more within group variance for each cluster than there was in the *KAMILA* or gaussian multinomial mixture model methods. The t-SNE plots also have a similar depiction to the other two methods. There are four u-shapped rings facing northeast where the bottom group is the set with the least utilization and the subsequent bands stack up until group 4 which has the set with the highest scores. By in large there is clear separation between each cluster but this is an artifact of the dichotomization of the scores.

The mixed patient burden scoring approach changes the order of the physician rankings compared to the *KAMILA* and gaussian multinomial mixture model approaches. For example the order of the top two Mayland clinic physicians is flipped based on the panel size estimation. We see that the new top physician only has more messages and no shows than the physician with the most patients. It could be that this method could be more susceptible to increases in vectors with sparse values. Any non-zero value could trigger a larger increase than anticipated. The mixed patient burden scoring method also increased the ranking of the third Mayland clinic physician, who is much lower in the other two methods, and does not flip the order of a few other physicians. The mixed patient burden scoring method largely follows the patient ranking. While the hope was to weight things based more on utilization and patient complexity it ended up not differentiating the patient rank as well. And when it broke the rank, it typically did it in a bad way. The mixed patient burden scoring method does not seem to do a great job ordering the physicians.

Looking at the correlation assessment for the mixed patient burden scoring method, ER visit, no show, phone calls, and medication count are not significantly associated with aggregated panel size estimation. That left only the specialty visits, messages and visit count as the variables significantly associated with aggregated panel size estimation. This suggests that most of the utilization variables do not explain the variability in the model. When looking at the ANOVA tables, 12% of the variability is explained by residuals. Some of this could be comprised by the patient demographics, but we can not know for sure what part of this is explained by random error. Specialty visits (0.47), visit count (0.17) and messages (0.15) explain most of the variability in the aggregated panel. There is temporal consistency with the panel size aggregation models, the same variables (visit count, messages and specialty visits) are significant in the model. From the temporal ANOVA, the influence of the messages variable increases and the residuals also increases.

2.3.2 Simulation Analysis

The results of the simulation are found in tables B.29, B.3.2, B.30. When we evaluated the *KAMILA* approach through simulation, the trend we noticed was that when the utilization means are the same the bias (the mean difference between the panel size estimate of the top doctor and the expected number of patients) is small, especially when there are just two providers. When the means are different the bias is larger, which makes sense since there are clear differences in the utilization tiers. As we saw with the real world data, the *KAMILA* method can highlight clear differences in utilization and will prioritize this over differences in the patient demographics. Interestingly the coverage in simulation 3 is about 59.1% (not that different from the coverage in simulation 1 of 50.7%) which would indicate that the demographics offer little in terms of distinguishing weighted panels. It is unclear if this has more to do with the mapping to a utilization measure or if the patient demographics are simply unimportant to the overall clustering. The trend is the same when the patient counts of the two doctors are different and when there are 3 providers with the same or different number of patients, the bias and MSE are smaller. In the 3 provider scenarios the bias and MSE are larger when the simulations have different proportions (as in simulations 11 and 15). The variability seems to increase as more providers are added into the analysis, further complicating the panel size estimation process. Simulation 11 interestingly had 81.3% coverage, suggesting that differences in proportion of the patient demographics do not fully distinguish weighted panels, however this lessens as more providers are added.

In terms of the clustering metrics, cluster purity is consistently in the high 80s, suggesting that adding patient demographics helps the clustering identify a single class well. There are a few occurrences when the cluster purity is around 79% and 80%, all of which come in situations when there are three providers, which has shown to add more variance in the clustering. For the adjusted rand index, most of the simulations have a score around 0.5, meaning they are “toss up” matching the benchmark clusters. This score decreases to around 0.37-0.38 in simulations when there are 3 providers and the means are different. More providers and a greater range in utilization seems to further prohibit finding the true clusters. When the mean utilizations are more tiered, clinicians with lower overall utilization could be categorized into smaller categories since their higher values would be thought as medium values compared to another provider. In terms of the classification error rate, it is fairly low (around 20%), again this increases to about 30% when there are differences in means and there are three doctors. This further emphasizes the point that adding my providers increases the variability in the clustering leading to panel size estimation.

Similar to the *KAMILA* approach, the gaussian multinomial mixture model has a general trend that shows when the means of the utilization measures are different there is little bias and a smaller MSE. There is also a similar pattern to *KAMILA* with simulation 3 (when there are different proportions in patient demographics) in that the coverage is less than 1. This could indicate that differences in demographics highlight some differences between the providers that is slightly better than a toss up, which is helpful. However, the simulations show that utilization measures mostly dominate the determination of the panel size estimation. In the simulations with three providers, the gaussian multinomial mixture model has similar results to the *KAMILA* approach. One difference is that the bias, the mean difference between the estimated panel size and the known number of patients, is quite small. The gaussian multinomial mixture model comes closer to returning the patient count when there is no obvious difference in utilization means. As for the clustering metrics, it is also a similar story to the *KAMILA* approach. Only when the utilization means are different is there a noticeable drop in clustering performance.

The simulation results of the mixed patient burden scoring approach are not very promising. The bias and mean square error are very high, meaning the panel size estimation largely deviates from the number of patients. This suggest that the mixed patient burden scoring approach is over quantifying the burden of the provider panel. It is prone to outliers who have high scores and increases the panel size overall based on high utilization patients and minimal utilization patients. The bias is less when it is just a difference in proportions than when the means are different. This generally stays the same across all simulations. The clustering metrics are a bit better with higher ARI and lower classification error but this may be more a result of the dichotomization of the scores. Having said this, the mixed PBS method is different to the clustering methods which makes it difficult to assess in head to head comparisons. The mixed pbs approach does seem to do a good job at projecting the patient burden for data analysis but it is unclear if this score is justifiable as a way of weighting panels.

2.3.3 Synthea Analysis

Synthea Summary

The synthea data is summarized in tables B.31 - B.32. We see that most of the utilization counts are low while the number of measurements is high. The median number of measurements is 21,including taking a patient's blood pressure, checking their height and weight and blood tests. For an outpatient visit there could be several basic tests that a patient could take during their visit, therefore we are not surprised by the large counts for this variable. The median number of outpatient visits is 4 over a two year period, so patients are going about 2 times a year. The commorbidity scores are all pretty low in this analysis. The majority of patients are young (age 18 to 34) black females. Some of the more common drugs taken include: sex hormones and contraception medications (n = 1996; 25%), calcium channel blockers (n = 1923; 24%), ace inhibitors (n = 1631; 20%), lipid modifying agents such as statins (n = 1101; 14%) and opioids (n = 1078; 13%). The most common conditions include: normal pregnancy (n = 1195; 15%), anemia (n = 297; 3.7%), hypertension (n = 189; 2.5%), concussions (n = 173; 2.1%) and heart disease (n = 127, 1.6%). Some other common occurrences in the cohort where colonoscopy (n = 999; 12%) and observation of obesity (n = 164; 2%). It was more common for patients to be on different drugs rather than having conditions or other medical encounters. There was a total of 15 physicians caring for 8077 unique patients. The mean number of patients seen by the physicians was 539 with a minimum of 385 and a maximum of 700.

Panel Size Estimation in Synthea

Results for the panel size estimation for the synthea data are summarized in the appendix in tables B.33 - B.41 and figures B.7, B.8, and B.9. Starting with the *KAMILA* approach, we see that there is a more variation in the clustering than there was in the VCU analysis. This makes sense because are using more covariates for clustering, capturing variability in more ways. The order of the cluster weights is specified in increasing order as follows: 3, 1, 4, 5 and 2. Cluster 2 has by far the highest amount of utilization. There is a median of 13 outpatient visits and 11 procedures in this cluster. Interestingly in terms of medications and measurements, cluster 5 reflects more utilization than cluster 2, calling to question whether mapping to number of visits is the best choice for panel size estimation. When we look at the complexity measures, this becomes more apparent. We can confirm what we are seeing with the panel weights through the t-SNE plot. Cluster 2, in yellow, is positioned in the bottom of the t-SNE plot. Next cluster 5, in pink, is positioned in the bottom left part of the graph. In some parts, this cluster is even with cluster 2 in the t-SNE plot. Clusters 4, 1, and 3 make up the next three sections going up. There is pretty solid vertical separation in the clusters, suggesting that the clustering is doing a good job identifying different groups across all the data. There seems to be a but of overlap in the center of the t-SNE plot when cluster 1 superseeds cluster 4. This could be a result of more variability incorporated into the clustering problem.

Assessing the complexity variables for panel size estimation, we see that cluster 2 contains several patients that were pregnant or had complications during pregnancy; 83% of all pregnancies and over 90% of all pregnancy complications were found in group 2. Accordingly, cluster 2 has several more visits due to this pregnancy factor, where there are probably more structured follow up visits. Mapping panel weights based on the number of visits may not be the best choice here because it is not reflective of a more burdensome panel. While pregnant patients require a lot of attention, it is anticipated that this burden will subside over time. Cluster 5 on the other hand has a larger percentage of conditions and drugs overall; 97% of all insulins, 93% of all chronic kidney disease, 86% of heart failure and 86% of all beta blockers. Cluster 5 seems to contain patients with high comorbidities. Combining the information from the complexity measures and the utilization measures, perhaps basing the panel weights on the number of outpatient visits is not the best way to reflect burden because logically we see a sicker set of patients in panels.

From the provider aggregation for the *KAMILA* approach we see a spread of normalized z-scores from -0.78 to 2.01, where more positive scores suggest physicians with less burden per patient and more negative scores suggest more burden per patient. The panel size estimation ranking does not completely follow the patient count ranking meaning that not every physician is seeing equal panels. The created physician Marilou Conn has the largest z-score, suggesting a less burdensome panel than 4 physicians with higher patient counts. It is likely that they had seen more patients in cluster 2 (possibly more pregnant patients who are predominant in this cluster), where the panel weight is larger. In the synthea simulation there is no variation due to practices. This makes the z-scores more comparable across physicians because there are no practice-specific variation.

With the gaussian multinomial mixture model, we see similar results to the *KAMILA* approach to panel size estimation on the synthea data. From the t-SNE plot we see a similar distribution of data points in the plots, despite the labels having changed. From Table B.36, we see that the ranking of the clusters by panel weight from lowest to highest is: cluster

3, 1, 5, 4, and 2. Clusters 4 and 5 have switched their order in the gaussian multinomial mixture model. Between the *KAMILA* aapproach and the gaussian multinomial mixture model approach there is about 91% agreement in terms of placing observations in clusters with the same rank. Aside for label switch, these two models are yielding nearly identical cluster rankings, however the rankings are different between the two methods. From the same table, we see that the weight of cluster 2 is higher in the gaussian multinomial mixture model but the second largest weight (this time for cluster 4 instead of 5) is smaller. The gaussian multinomial mixture model found the median outpatient visit count to be 6 for cluster 4 compared to *KAMILA* had a median of 7 for the analogous cluster 5. In terms of the utilization measure some other minor differences was the median number of measurements in cluster 4 was 30 instead of 30 and the median number of inpatient visits was 1 instead of 2.

From the categorical variables, we further see the result of the label switching between the gaussian multinomial mixture model and the *KAMILA* approaches. Cluster 4 now contains patients that clearly have more commorbidites, such as different cancers (most above 97%), heart disease (81.1%), dementia (80.6%) and diabetes (70%), and more drug prescriptions, such as blood glucose medications (89%) and antithrombotic agents (59.4%). Patients in cluster 4 also tend to be over the age of 65. We notice that in the gaussian multinomial mixture model approach there is a greater concentration of patients with cancer in the cluster containing sicker patients. Like with the *KAMILA* approach, cluster 2 contained mostly patients who were pregnant or had pregnancy complications. We reach a similar conclusion to the *KAMILA* approach in that it is no longer clear if mapping weights to the number of visits is the best choice when conducting mixed clustering. Other approaches for mapping these clusters to a weight should be tried, but as of right now mapping to the number of visits is the most rational choice.

Looking at the provider aggregation for the gaussian multinomial mixture model, we see that the z-scores range from -0.75 to 1.86, which is close to the *KAMILA* approach. The physician with the highest positive z-score was the same between two approaches, though the negative z-scores were different. Now the created physician Sharyl Hilpert has the highest z-score and Marilou Conn is second, where as in the *KAMILA* approach they were tied. If we ranked the physicians based on panel size estimation, again we see that the created physician Marilou Conn jumps ahead of 4 physicians. The only change in the panel size estimation rankings between the *KAMILA* and gaussian multinomial mixture model approach is the created physicians Harlan Beer and Paulene Schultz switch in rankings. There are very minor changes in panel size estimation between the gaussian multinomial mixture model and the *KAMILA* approach. Besides a few minor variations in clustering, the solutions presented are very similar.

Finally, we looked at the mixed patient burden scoring method. With this method there are only 4 clusters because we cut the score into: cluster 1 with score between 0 and 0.5, cluster 2 with score between 0.5 and 1, cluster 3 with score between 1 and 1.5, and cluster 4 with score between 1.5 and 2 as was done before. The median score was 0.7812 and the mean score is 0.8165. Also the 75th percentile of the score is 1.0492, which means that the score is skewed heavily to the left of 1. This point is further emphasized when we see that cluster 4 only contains 222 points. It could be that the data is too influenced by outliers when converting to a score. The few values with extreme utilization are minimizing the scores of the other values. From the synthea analysis, the mixed patient burden scoring method is not very promising when adding multiple comorbidites into the analysis.

When we consider the complexity, most of the pregnancy patients (n = 1029; 86.1%) and patients who saw a complication due to pregnancy (n = 155; 77.9%) are in cluster 3. Since cluster 4 is so small it does not dominate any specific category in terms of complexity. Heart failure (n = 17; 60.7%) was the only condition that was predominant in cluster 4. Cluster 4 also contained a decent amount of patients on beta blockers (n = 19; 68%). Cluster 4 doesn't seem to have a clear group of sicker patients like the *KAMILA* and gaussian multinomial mixture model approaches. This confirms what we saw in the summary of the burden score. It is so influenced by outlier utilization that the burden score is shifted to the left. It undervalues the burden of most patients because it overvalues the burden from the most extreme patients. When we look at the physician aggregation, this point is further emphasized. All of the z-scores are positive, ranging from a low of 0.21 to a high 2.22. This suggest that all of the physicians have a less burdensome panel per patient. From this assessment, we are not sold on the benefit of the mixed patient burden scoring method for panel size estimation.

2.4 Discussion

Following this analysis, the results of mixed type panel size estimation is not totally clear. As is the case with a lot of unsupervised learning methods it is hard to determine if the methods are doing a good job separating items into good groups. It is also unclear how helpful adding patient complexity to the equation improves panel size estimation. While the results lead to a lot more questions than answers, there are a number of important take-aways that progress the way we think about panel size estimation. These discussion points are: first, a variance stabilizing transformation and data pre-processing steps are vital for effective panel size estimation; second, selecting the correct features to use in an analysis is important yet quite challenging; third, mapping the panel weights is a harsh but justified assumption; fourth, accurately phenotyping the population of interest is important and helpful for physician aggregation; and fifth, using unsupervised methods is a good way to do preliminary analysis for decision making on panel sizes. Before going into greater depth on the key-takeaways we also summarize the findings of the analysis for each method.

2.4.1 Summary of Analysis Results

From the VCU analysis, the *KAMILA* approach curiously added a partition to the “middle-class” of patients. One of those groups had patients with a lot of messages and median of 3 visits, while the other group had patients with no messages and a median of 4 visits.

Since the visit count was used as the mapping variable, the group with more primary care visits received a greater amount of weight. Looking at the patient demographics, these two groups showed a split in socio-economic status. Those in the lower visit high message group tended to be white, middle age and have private insurance. Those in the higher visit, low message group tended to be black, older and on public insurance. The *KAMILA* approach was able to distinguish patient populations that belonged to different practices which seems encouraging. However when we looked at the correlations, the amount of variability explained by the model was mostly captured by the utilization variables. When it came to the simulations, the *KAMILA* approached was not effected much by changes in the categorical percentages. It was more impacted by changes in the utilization and the number of physicians used in panel size estimation. Finally, when it came to looking at the synthea results, the *KAMILA* approach split the middle class of patients. This time the split was more a result in distinguishing very sick patients to not as sick patients. The most interesting cluster generated from *KAMILA* in the synthea data analysis was the cluster that highlighted pregnant patients. These patients go to the doctor a lot due to the nature of care during pregnancy. The process may have over-emphasized these patients compared to sicker patients since they visited more.

In comaprison, the gaussian multinomial mixture model produced very similar results to the *KAMILA* approach across the three sections of the study. One noticeable difference was that the gaussian multinomial mixture model did not get as hung up on the middle class of patients but created more partitions towards the low utilization class of patients in the VCU data analysis. The method tried to partition between non-utilizers and minimal utilizers where as in the *KAMILA* method this was seen as a single group. Perhaps with the VCU data, the mixture model was getting caught in regions with undefined variance such as these minimial utilization categories. When it came to the simulations, the gaussian multinomial mixture model performed nearly on par with the *KAMILA* approach perhaps with some slight increases to the bias and MSE. From the simulations we see minor advantages towards using

KAMILA, but the evidence is not overly convincing in any direction. From the synthea analysis we that the gaussian multinomial mixture model performed very similar to the *KAMILA* approach. It identified a similar population of pregnant patients who require a lot of attention. The difference between the methods was that the gaussian multinomial mixture model down-weighted the middle-class of patients. By in large the results were pretty similar between the two methods.

The mixed patient burden scoring method performed quite poorly in comparison to the clustering methods in the synthea analysis, however it performed ok in the VCU analysis. The biggest problem for the synthea analysis was that the mean and median score were skewed to the left at around 0.7. This indicates that the mixed patient burden scoring method is highly susceptible to outliers and could not properly account for the multitude of complexity variables. Despite controlling for extreme values in data pre-processing, the few extreme values remaining in the data set forced the other patients to be heavily down-weighted in terms of burden. Meanwhile in the VCU data analysis, the mixed patient burden scoring approach did a decent job at creating a normal score distribution centered around 1 (recall the mean and median scores for the VCU analysis were 0.9331 and 0.926 respectively). It could be that when only including demographics the adjustment of the burden scores performs fairly well because we are able to account for the variation in practice characteristics. This is a promising result if we wanted to better isolate practice specific variation to panel size estimation. The contrasting results from the VCU and synthea analysis perhaps render lessons on how to use dimension reduction in preliminary analysis to better understand patients and how to empanel them rather than using the mixed patient burden scoring method directly. We will discuss this further in the “key-takeaways” section.

By in large, there is no single method we would recommend for mixed data panel size estimation. Neither the *KAMILA* or gaussian multinomial mixture model showed dramatic differences in results. It is also unclear if incorporating categorical features vastly improves panel size estimation. From the VCU data analysis we saw that the categorical features did not explain a lot of variability in the model. Also in the simulations we saw that changes to the categorical proportions did not improve the accuracy of aggregations. However, in the synthea analysis we saw that adding the categorical variable definitely helped define groups. The pregnancy patients had a clear pattern that was different to a typical patient and the methods could identify very sick patients. There is likely an association with healthcare interaction and patient comorbidities, meaning the sicker a patient the more they utilize healthcare resources. So when adding complex patient features this helped clustering whereas when we only add basic demographics it does not help the clusters as much. We would need to conduct more analysis on real data to determine if a mixed-type approach does improve panel size estimation. Regardless of that results, what we do know is understanding the complexity to the primary care patient population can only improve healthcare resource management. We can use a higher dimensional healthcare data to perform preliminary analysis that can be used to find patterns and tendencies among patients. There is a lot more research that should be done on the utility of unsupervised learning methods for longitudinal healthcare data analysis.

2.4.2 Key Take-aways

Data Preprocessing

Processing the data before analysis is essential when performing panel size estimation. In this analysis we conducted to major preprocessing steps: first, we algorithmically attributed patients to physicians and second, we used a variance stabilizing transformation to normalize the count data before using it in a unsupervised learning tool (clustering or dimension reduction). Starting with patient attribution, this step is important because it is too difficult to tell in US data who is a patient's primary care physician without looking at the amount of work that they do. There are a few considerations to think about with attribution such as how much time is taken into account (i.e. a year, 2 years or 6 months), how to segment outpatient visits and whether to use only physicians or include members of the care team, like NPs, to the attribution. Physician attribution is sensitive to the time allocated for accrument of visits. In this analysis we used 2 years since this refelcted the Rajkomar paper, however we would need to further understand if this is the best time frame. Including NPs to the panel size estimation process seems warranted as they do a substantial amount of care. What remains unclear from informatic attribution is whether to count short visits to the main physician or to whoever saw them. In the instance that the primary care provider is the physician but an NP sees the patient several times for minor assessments who is attributed to the patient. This step is vital to properly attribute the work to the correct person so that panel size estimation is equitable.

The other key pre-processing step is performing a variance stabilizing transformation on the continuous data. As mentioned a variance stabilizing transformation is used to curtail the effect of non-constant variance in the data. Utilization counts can have very high values or zero values which can lead to confusing results when analyzing the raw data. High counts have wider variability that impacts downstream analysis, whereas zero and minimal counts can overly influence associations following a basic log transformation. In this analysis we borrowed a principal from genomic count data in performing appropriate variance stabilizing transformations so that the count data can follow a gaussian distribution. In particular, we borrowed the rlog transformation from Love et al to normalize our utilization count data. This method stabilizes the variance in the counts across each patient sample. Using the rlog data on unsupervised methods (clustering and ordination methods) reduces the influence from high counts and lessens the influence from zero counts. We found after some experimentation that the clusters produced after the rlog transformation were much more consistent and stable. For example when we used the mixture model on the raw data the results were poor because we likely violated the gaussian assumption. Once the rlog was applied the clustering was much improved. There are more methods becoming available for transforming count data before using unsupervised methods to find clusters and associations within the underlying data. One such method known as glmpca **(citation)**, similarly uses a glm model to limit the influence of low counts but then goes a step further and conducts pca prior to clustering. This is an interesting idea, using a lower dimension representation of the data prior to clustering. It also presents a possibility dimension reduction on mixed data prior to using a more conventional clustering method like kmeans. We did not perform this assessment in this analysis but the idea generates an interesting scenario for next steps on constructed weighted panels from EMR data.

Selecting Patient Features

In the VCU analysis we were limited to a small number of patient features we could use to derive adjusted panel sizes. A feature like medication count was not the best feature because it was based on a cummulative sum of all medications over two years. This means that medication count is highly correlated with the number of patient visits. When we looked at the correlations, we saw that medication count was largely excluded because it suffered from multicolinearity. It is important that the constructed patient features are not highly correlated with each other and are the best depiction of the clinical context of the patient. A more appropriate way of working with medications is based on eras. An era is compounded timeline that is extended for a subsequent indication over a period of time. For example if a patient was prescribed an ace inhibitor every 60 days, then we would not count each refill as a unique drug exposure but rather count it as a single instance if the refill takes place within the 60 day time frame. Constructing accurate patient features should improve the accuracy of panel size estimation.

Another variable that caused problems in the analysis was the no show variable. In the original Rajkomar paper, no show represents work done by the physician in the event that a patient does not show up to an appointment. The arguement is that they are still performing work. However, since the number of no shows is so small, any patient who had at least one no show proved to be overwhelmingly burdensome compared to another patient. From the correlation assessments, we saw that the no show variable had the largest beta estimate meaning it lead to the largest increase in panel size estimation compared to all the other variables. This should not be the case and we should not let a largely zero valued feature have so much influence on our estimation. It is important to conduct data analysis and understand the features used in the panel size estimation before coming up with an adjusted count. While there is a plausible justification for using no shows in estimating panel weights through clustering, it does not tangibly add value to the process and should be eliminated.

Feature extraction is often a difficult exercise with longitudinal observational databases such as EMRs. There are potentially millions of arbitrary features one can construct from the EMR based on indicators for occurrences across several clinical domains such as drugs, conditions, procedures and more. Building features requires a concept, a time frame and a starting point and may need a roll-up (for example if a condition is chronic). The concept can be a clinical definition such as atrial fibrillation. The time frame is the period where the concept is observed for a patient relative to a starting point. The starting point can be at baseline, from the start of a hospitalization or from the end of an encounter. A time frame can be any length of time usually done in relation to days (i.e. 30 days, 180 days or 365 days). A time frame should be appropriately constructed based on the concept, for example a ischemic stroke may be best suited as seen within the past 3 days as opposed to over the course of several years. All of these considerations make the creation of patient features very difficult. Extracting the features can also be very tedious depending on the orientation of the database. In our example we created a set of 70 features of patient complexity, which may or may not have been the best set of choices. Our selection was arbitrary aimed at getting a sense of how the panel size estimation methods would handle this quantity of data. Future panel size estimation algorithms should spend more time determining features that should be extracted that balance healthcare utilization and complexity. The features also need to have some clinical significance, for example a dental procedure is not applicable to a primary care panel.

Mapping to Weights

A key assumption made for the panel size estimation process, taken from Rajkomar, was to generate weights by scaling the median visit count of each cluster by the cluster size. This is a very reasonable assumption since the number of visits is the best term to measure the amount of interaction with a primary care physician. We also agree that this is still the best practice for a panel size estimation pipeline, however it is a harsh assumption that will have downstream ramifications on aggregating panels by physician. We see this effect particularly with the *KAMILA* approach on the VCU data. Looking at the t-SNE plots, clusters 2 and 4 seem to cover similar sets of patients. The median visit count for cluster 2 is 3 while the median visit count for cluster 4 was 4. However, when looking at the messages cluster 2 has a median count of 13 compared to cluster 4 which has a median count of 0. So the clusters show more visits in one cluster than the other but one can argue that handling that amoung of messages is just as much work. That brings the question if projecting the cluster on to a single utilization feature is the best idea since it could have a narrow scope on what ultimate contributes to the panel weight. Likewise, projecting based on the visit count ignores the contribution of the categorical variables. Is it possible that the categorical variable distinguish some of the variability not seen by the visit count? These lingering questions bring doubt towards whether mapping on a utilization feature is the best decision. The issue is that there are not many suitable alternatives. The other utilization and complexity features could be just as problematic, if not more, towards biasing the weights in unintended ways. Also as evidenced by the failed patient burden scoring there is not a ideal way to present a score of burden. So that leaves us with mapping based on visit count; while this could cause downstream effects we need to understand its limitations before using it within a panel size estimation algorithm.

Accurate Phenotype of Population

One note from the Rajkomar paper is they excluded pediatric patients from panel size estimation, only considering patients ages 18 and older. In the VCU analysis we included all patients regardless of age. However, upon reflection we should have eliminated pediatric patients because the way they engage with a primary care physician is fundamentally different to an adult. As a result there were several newborn patients with a single primary care visit in the 2 year period, with no other forms of healthcare utilization, that formed a very particular group in the analysis. These patients should have been excluded from the analysis because they are not comparable to adult patients without insurance who go to a primary care provider. This lead to an important lesson in this analysis, correctly phenotyping the population of interest. An issue with panel size estimation is that the patient populace is extremely diverse and by our process we want to highlight this variability among patients to project burden. However, if this population were to diverse then it would effect the analysis since we are no longer comparing similar kinds of patients. For example, in the synthea analysis we looked at all patients who had an outpatient visit of any kind. This is too general of a visit type because we incorporate too many patients that go for a non-primary care outpatient visit. It could have been why there is clearly a different pregnancy group. Since the synthea data is synthetic and has a bare representation of visit type, we did not try to narrow the search to primary care. But, the experience leads us to improve the way we phenotype the primary care patient population in order to estimate an accurate panel size.

Another consideration for phenotyping the patient population is understanding which practices are involved with the set of patients. Each practice sees different sets of patients. For example with the VCU data, Mayland clinic encompasses patients from a suburban area where as the Nelson Clinic encompasses patients from an urban area. The demographics of these patients are different and the style of care across these practices are also different. Isolating practices may be important for panel size estimation as there may too many practice specific differences to estimate panel weights all together. In theory by adding the categorical variables, such as patient demographics, this should help capture differences between practices. But it is unclear how well the methods do in separating panel weights by practice. In phenotyping the patient population it is suggested that the primary care patients should be separated out by practice so that panel weights are estimated among a similarly treated population.

Unsupervised Methods for Data Analysis

The last lesson we garner from the mixed panel size estimation process is using unsupervised methods such as clustering and dimension reduction for exploratory data analysis of physician panels. In addition to estimating panel sizes, it is also important for physicians to understand the types of patients in their panel. Clustering and dimension reduction methods are helpful for understanding data traits of diverse populations. While in panel size estimation those traits are projected towards a weight, these visualizations can be helpful in general. More exploration is needed in understand how to best leverage methods such as the t-SNE plots, clustering and principal component analysis to understand patient populations.

CONCLUSION

Our analysis of panel size estimation has shown that there is no clear cut answer on what serves as the best method. However, we have found solid suggestions that help improve the panel size estimation process and ideas for next steps in

2.5 Part 1 Conclusions

In part 1 of the dissertation, we found that the k-means clustering method was slightly better than the gaussian mixture model for finding patient utilization phenotypes. The k-means approach was able to find more clusters of patients and differentiate patients in the "middle," meaning it could split utilization phenotypes beyond the obvious low utilization and high outliers. The gaussian mixture model only found 3 groups in the VCU data analysis and could not separate the middle group. However, the choice of the two methods is a matter of preference. The gaussian mixture model is more conservative in where it creates partitions in the cluster space. According to the simulations, the k-means method had better clustering metrics while the gaussian mixture model had a smaller bias from the true count of patients. This suggests that the gaussian mixture model is more conservative in creating partitions and could be more susceptible to very small utilization and big outliers while not being able to distinguish the middle group. Our suggestion for utilization panel size estimation is to use the k-means method however there is not a overwhelming amount of evidence to say that the gaussian mixture model should not be used.

We did find that novel patient burden scoring method did not perform well for panel size estimation. The issue with this method is that it was highly susceptible to outliers and extreme values of utilization. There was a skew to the left of 1 in the distribution of the burden scores. The scores of most patients ended up being too small. Another issue is that there are too many parameterizations of this method. The analyst needs to choose how to contrast the patient from “typical” and they also need to determine how to weigh the features. These choices could greatly alter how the results were presented making it an impossible task to find the right patient burden setting for the study. We would advise against using the patient burden scoring method and rely on identifying clustering phenotypes. However, the patient burden scoring method hinted at the possibility of using dimension reduction methods for understanding types of patients and clusters. By projecting the patients into a low dimensional space we can help assess how well the clusters are doing. Looking at the t-SNE plots proved helpful in better understanding the clusters. Further using a more informative dimension reduction could help understand why patients were placed into particular clusters. Similarly, it is worth exploring further if principal component analysis prior to k-means could help with the panel size estimation. Moving the scores into a orthogonal space could help find good clusters and help remove the issue of mapping the weight to a single feature. These areas need to be explored further and were outside the original scope of the dissertation.

Our biggest take-aways for panel size estimation came from aspects outside the statistical models and part of the data extraction and pre analysis phase. First, a variance stabilizing transformation like r-log was very helpful prior to clustering. It removed the influence of extreme values and of trivial values (such as patients with 1 visit and no further utilization). The r-log transformation was borrowed from genomic research showing that we can translate ideas from other domains and apply them to EMR data analysis. Another important consideration was how to construct features. Feature extraction is important in this analysis because we are using this data to orient cluster since there is no outcome to drive our model, hence unsupervised learning. When bad features are used this can tremendously impact our analysis; for example features that are colinear could be masking each others influence or features with too trivial of information can be too influential when a single event is observed. It is vital that we extract good, independent features from the EMR, which is not a simple task. In the VCU analysis the medication count was based on a cummulative addition of medications. Thus it was highly correlated with visits; if the person went to the doctor more they were prescribed more medicaitons. This was seen as an issue in our correlation assessments. The no show feature was also contentious; while showing some "utilization" the counts were very rare. So when a no show was registered by a patient their burden shot up, which we should not expect. Finally, another important consideration was constructing a strong phenotype of primary care patients. In the VCU data, we included pediatric patients who have very different visit tendencies than adults. This was problematic since new-born patients would have single visits or young patients were not comparable to adult patients in terms of needs. This most likely impacted the clustering downstream since it added more "non-utilizers." Another aspect concerning phenotypes was how there was no distinction for practice location. Providers in urban practices see a different patient populace than those in a suburban setting so their panel sizes are not quite comparable. We introduced a z-score to attempt to standardize the estimation of burden across practice, however the underlying data still contained practice variation. It would be better to isolate these providers by practice and understand panel sizes within a practice rather than a system.

2.6 Part 2 Conclusions

In part 2 of the dissertation we found that there was even less of a clear-cut answer to which clustering method was superior for mixed data. The KAMILA and gaussian multinomial mixture model performed similarly in all three sections of the analysis: the real-world application to VCU data, the simulation and the synthetic EMR data analysis. In the VCU data, the KAMILA approach partitioned two clusters of patients that had similar utilizations, the difference coming in the number of messages and demographics. The gaussian multinomial mixture model had more conservative cuts along the utilization of the patients. We would recommend using the KAMILA approach because it is more robust to violations in parametric assumptions, but we did not identify overwhelming evidence to disuade users from using a finite mixture model. The KAMILA also had smaller bias and better clustering metrics in the simulation. The biggest source of optimism was the ability to use a larger array of features for clustering. Using either the gaussian multinomial mixture model or the KAMILA approach can yield panel size estimations that consider a much larger scope of sources of variation from demorgraphic differences to condition differenes. The the synthea analysis when we explored these methods for panel size estimation with a larger number of features we were able to see how utilization changed based on someone's characteristics. Patients who were pregnant were different from those who were old and sick. This was very promising in terms of using more advanced clustering methods than k-means to handle different data types.

The mixed patient burden scoring method was not successful. Just like the patient burden scoring method, the mixed extension was very susceptible to extreme values and trivial utilization causing a left skew in the mean score away from 1. The patient burden scoring approach should not be used for panel size estimation because it requires to many parameter choices and is too susceptible to high variance data. Similar to our conclusion in part 1, we do see the value in low-dimensional projections for data analysis and data processing but to not use the approach for panel size estimation.

Similar to part 1, we arrived at suggestions to help improve panel size estimation beyond the statistical model. Using variance stabilizing transformations on the continuous data is a very important piece in data pre-processing. Also finding good features and correctly phenotyping patients is important towards panel size estimation. The synthea experiment showed the potential of using a large number of features to build adjusted panel sizes. We do not need to use every single feature in the EMR but we can make a set of useful features to help build clusters. Rare conditions and drugs are not helpful, but using common conditions and drugs can help inform the clustering process a bit more than the simple k-means process. We found that inclusion of features like drug exposures to insulin and chronic kidney disease helped highlight a high utilization group in the synthea data. The group with the most visits was one that was full of pregnancy patients, but the patients with chronic conditions was highlighted as a high utilizer group. Including these extra features help compliment information from the utilization. When using so many features from the EMR as we did with the synthea data requires good data infrastructure to facilitate the analysis. We converted synthea to the OMOP CDM to use standardized open-source tools to aid in feature extraction. This made it easier to find chronic conditions and add them into our clustering pipeline. We recommend converting any longitudinal observational database to the OMOP CDM because it standardizes the analysis, targeting a consistent structure as opposed to creating custom analyses for each project.

Appendix A

APPENDIX FOR CHAPTER 1

A.1 Algorithms

A.2 VCU Data Analysis

A.2.1 Overall

Table A.1: Four Cut Methodology

Cut	Assignment	Description
1	Sole provider	Patient has only seen one provider in the past year
2	Majority provider	Patient has seen multiple providers but has seen one a majority of the time
3	Provider who performed last physical	No majority provider seen by a patient in past year
4	Last seen provider	First 3 cuts do not settle pcp assignment so it goes to the last seen provider

Table A.2: VCU Health Systems Demographic Summary

covariate	factor	num	pct
age	0-17	1651	9.1%
age	18-34	3848	21%

covariate	factor	num	pct
age	35-49	4081	22%
age	50-64	5131	28%
age	65+	3525	19%
insurance	None/Other	1098	6%
insurance	Medicaid	3295	18%
insurance	Medicare	3559	20%
insurance	Private	10284	56%
practice	HEWHC FamMed	3791	21%
practice	Mayland FamMed	5360	29%
practice	Nelson FamMed	6035	33%
practice	Tanglewood FamMed	3050	17%
race	Other/Unknown	2692	15%
race	Black	6756	37%
race	White	8788	48%
sex	F	10159	56%
sex	M	8077	44%

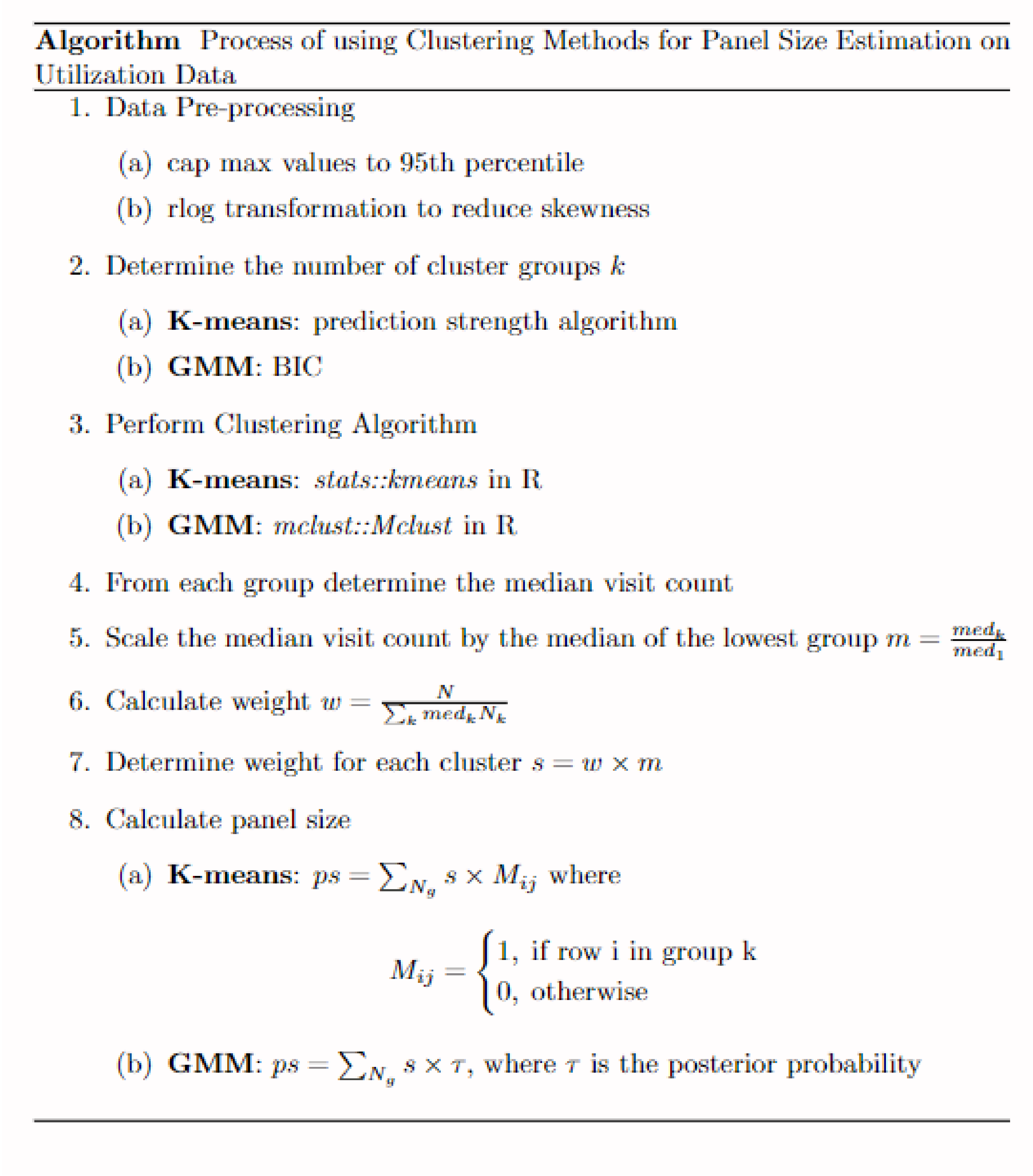
Table A.3: VCU Health Systems Utilization Summary

utilization	min	p25	median	mean	p75	max
visit_count	0	1	3	3.80	5	66
specialty_visit	0	0	2	5.94	7	176
pt_call	0	0	0	1.95	2	114
no_show	0	0	0	0.26	0	12
messages	0	0	0	5.55	5	532

utilization	min	p25	median	mean	p75	max
med_count	0	2	8	22.95	24	917
er_visit	0	0	0	0.68	1	97

Table A.4: VCU Health Systems Provider Aggregation Summary

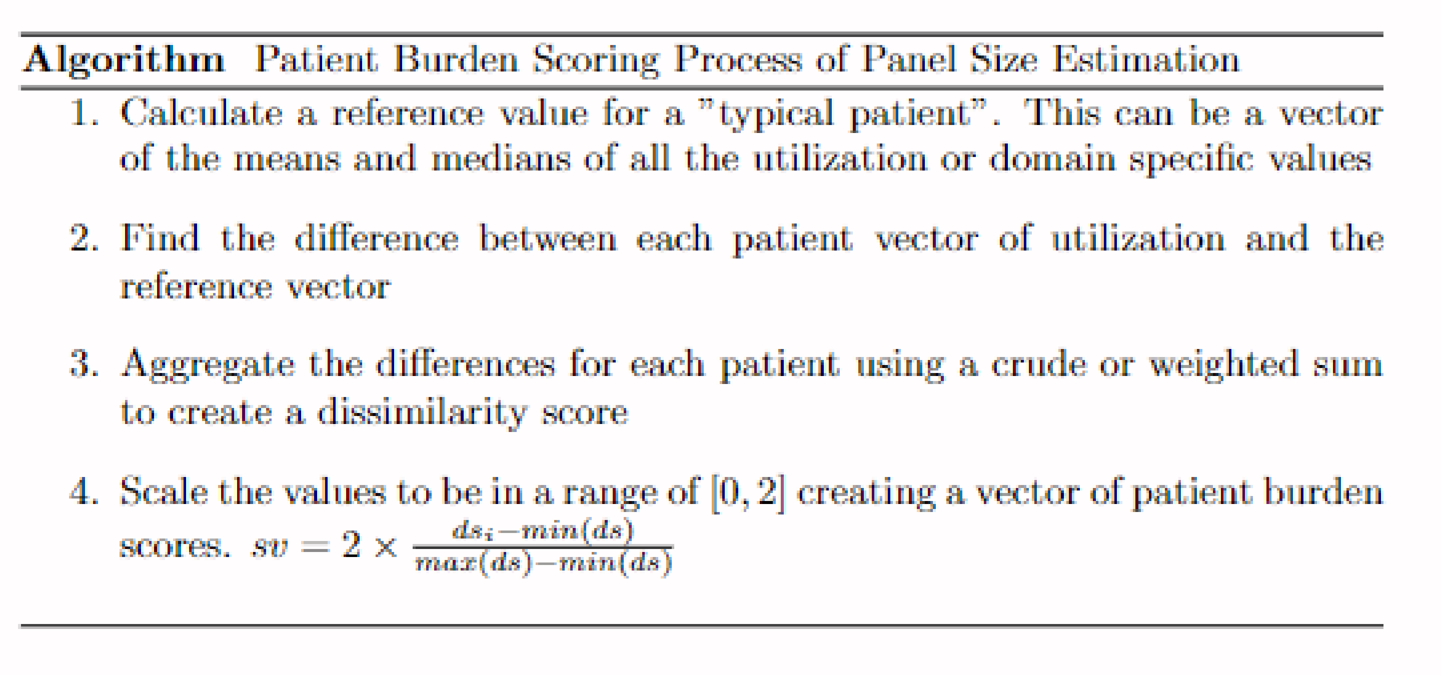
Provider	Practice	pat	visit_count	no_show	med_count	messages	er_visit	specialty_visit	pt_call
M	Mayland	2410	7065	320	35848	14681	382	9833	90
L	Mayland	2227	6523	322	34614	17197	377	9177	104
H	Nelson	1475	3783	280	20205	5919	1187	11098	2845
B	Tanglewood	1337	8058	299	53248	306	1017	6096	25
K	Nelson	1006	3392	203	19131	7494	514	6761	2530
F	Nelson	896	3218	156	16254	6164	503	5726	2028
A	HEWHC	831	4012	335	24368	1021	618	5544	3274
J	HEWHC	829	3687	386	16269	902	592	4233	2612
N	Mayland	731	1599	128	7285	2585	168	3253	31
G	HEWHC	675	2931	250	15608	1793	423	3820	2459
C	Nelson	620	2060	103	12010	3504	270	4113	1568
D	Nelson	584	1745	112	7708	2931	375	3824	1235
O	Nelson	503	1818	76	9362	3110	250	3455	1141
I	HEWHC	502	2424	312	16083	1024	406	3527	2282
P	HEWHC	372	1673	126	6893	1191	167	1401	1175
E	Nelson	362	1518	72	9910	3519	214	2561	1307
Q	Nelson	259	1493	61	10404	1750	144	2212	1235



Algorithm Process of using Clustering Methods for Panel Size Estimation on Utilization Data

```
1. Data Pre-processing
    (a) cap max values to 95th percentile
    (b) rlog transformation to reduce skewness
2. Determine the number of cluster groups $k$
    (a) K-means: prediction strength algorithm
    (b) GMM: BIC
3. Perform Clustering Algorithm
    (a) K-means: stats::kmeans in R
    (b) GMM: mclust::Mclust in R
4. From each group determine the median visit count
5. Scale the median visit count by the median of the lowest group $m = \frac{med_k}{med_1}$
6. Calculate weight $w = \frac{N}{\sum_k med_k N_k}$
7. Determine weight for each cluster $s = w \times m$
8. Calculate panel size
    (a) K-means: $ps = \sum_{N_g} s \times M_{ij}$ where
            $M_{ij} = \begin{cases} 1, & \text{if row i in group k} \\ 0, & \text{otherwise} \end{cases}$
    (b) GMM: $ps = \sum_{N_g} s \times \tau$, where $\tau$ is the posterior probability
```

Figure A.1: PSE Algorithm for Utilization data in EMR



Algorithm Patient Burden Scoring Process of Panel Size Estimation

1. Calculate a reference value for a "typical patient". This can be a vector of the means and medians of all the utilization or domain specific values

2. Find the difference between each patient vector of utilization and the reference vector

3. Aggregate the differences for each patient using a crude or weighted sum to create a dissimilarity score

4. Scale the values to be in a range of $[0, 2]$ creating a vector of patient burden scores. $sv = 2 \times \frac{ds_i - min(ds)}{max(ds) - min(ds)}$

Figure A.2: Patient Burden Scoring Algorithm

A.2.2 Kmeans

Table A.5: Kmeans Cluster Summary Table

cluster	utilization	min	q1	mean	median	q3	max
1	visit_count	0	1	1.47	1	2	11
2	visit_count	1	5	7.25	7	10	11
3	visit_count	1	1	2.71	2	3	11
4	visit_count	1	2	4.25	4	5	11
1	specialty_visit	0	0	0.44	0	0	8
2	specialty_visit	0	6	14.12	13	23	26
3	specialty_visit	0	0	2.73	1	4	26
4	specialty_visit	0	1	6.86	4	10	26
1	pt_call	0	0	0.14	0	0	5
2	pt_call	0	0	4.93	5	10	10
3	pt_call	0	0	0.63	0	1	10
4	pt_call	0	0	1.56	0	2	10
1	no_show	0	0	0.06	0	0	2
2	no_show	0	0	0.54	0	1	2
3	no_show	0	0	0.15	0	0	2
4	no_show	0	0	0.25	0	0	2
1	messages	0	0	0.23	0	0	6
2	messages	0	0	10.25	4	22	28
3	messages	0	0	2.09	0	3	28
4	messages	0	0	6.44	2	11	28
1	med_count	0	0	2.14	1	3	26
2	med_count	0	31	58.14	56	96	97
3	med_count	0	2	8.24	5	11	92

cluster	utilization	min	q1	mean	median	q3	max
4	med_count	0	8	23.23	16	30	97
1	er_visit	0	0	0.07	0	0	3
2	er_visit	0	0	1.31	1	3	3
3	er_visit	0	0	0.29	0	0	3
4	er_visit	0	0	0.59	0	1	3

Table A.6: Kmeans Panel Weight Mapping

clusterId	medianVisit	clusterSize	scale	weight
1	1	4307	1	0.32
2	7	3008	7	2.22
3	2	5710	2	0.63
4	4	5211	4	1.27

Table A.7: Kmeans PSE Provider Aggregation

Provider	Practice	pat	pse	z
M	Mayland	2410	2020.84	0.64
L	Mayland	2227	1980.65	0.43
H	Nelson	1475	1536.36	-0.15
B	Tanglewood	1337	1387.63	-0.14
K	Nelson	1006	1199.97	-0.66
F	Nelson	896	1050.29	-0.60
A	HEWHC	831	1032.26	-0.82
J	HEWHC	829	909.47	-0.35

Provider	Practice	pat	pse	z
G	HEWHC	675	789.22	-0.61
C	Nelson	620	720.87	-0.59
I	HEWHC	502	669.61	-1.13
D	Nelson	584	601.25	-0.12
O	Nelson	503	594.92	-0.66
N	Mayland	731	542.08	1.19
E	Nelson	362	493.66	-1.25
Q	Nelson	259	409.80	-1.87
P	HEWHC	372	373.73	-0.02

Table A.8: Kmeans Model Summary

visit_count	no_show	med_count	messages	er_visit	specialty_visit	pt_call	size	withinss	cluster
-0.96	-1.12	-1.36	-1.40	-1.19	-1.40	-1.37	4307	5464.31	1
2.37	1.75	2.42	2.05	1.90	2.26	1.98	3008	10795.12	2
0.33	-0.05	-0.15	-0.14	-0.10	-0.15	-0.32	5710	12043.32	3
1.31	0.77	1.06	1.07	0.79	1.02	0.64	5211	16879.04	4

Table A.9: Kmeans Internal Correlation Model

term	estimate	std.error	statistic	p.value
visit_count	0.13	0.013	9.81	0.0000042
no_show	0.20	0.069	2.94	0.0164099
med_count	0.00	0.002	-1.74	0.1154575
messages	0.03	0.003	8.46	0.0000141

term	estimate	std.error	statistic	p.value
er_visit	-0.07	0.045	-1.65	0.1335166
specialty_visit	0.08	0.006	13.46	0.0000003
pt_call	0.03	0.006	4.66	0.0011923

Table A.10: Kmeans Internal Correlation ANOVA

variable	Sum Sq	prsq	Df	F value	Pr(>F)
visit_count	16752.05	0.244	1	96.17	0.0000042
no_show	1508.52	0.022	1	8.66	0.0164099
med_count	528.67	0.008	1	3.04	0.1154575
messages	12471.05	0.182	1	71.60	0.0000141
er_visit	473.73	0.007	1	2.72	0.1335166
specialty_visit	31542.95	0.460	1	181.09	0.0000003
pt_call	3775.65	0.055	1	21.68	0.0011923
Residuals	1567.66	0.023	9		

Table A.11: Kmeans Internal Correlation Model

term	estimate	std.error	statistic	p.value
visit_count	0.13	0.020	6.37	0.0000813
no_show	0.43	0.112	3.82	0.0033861
med_count	0.00	0.003	-0.46	0.6556176
messages	0.03	0.003	8.18	0.0000097
er_visit	-0.13	0.071	-1.80	0.1024368
specialty_visit	0.07	0.008	8.86	0.0000048

term	estimate	std.error	statistic	p.value
pt_call	0.05	0.006	7.68	0.0000168

Table A.12: Kmeans Internal Correlation ANOVA

variable	Sum Sq	prsq	Df	F value	Pr(>F)
visit_count	14868.12	0.149	1	40.59	0.0000813
no_show	5339.27	0.053	1	14.58	0.0033861
med_count	77.39	0.001	1	0.21	0.6556176
messages	24500.40	0.245	1	66.88	0.0000097
er_visit	1183.86	0.012	1	3.23	0.1024368
specialty_visit	28728.43	0.287	1	78.42	0.0000048
pt_call	21600.72	0.216	1	58.97	0.0000168
Residuals	3663.29	0.037	10		

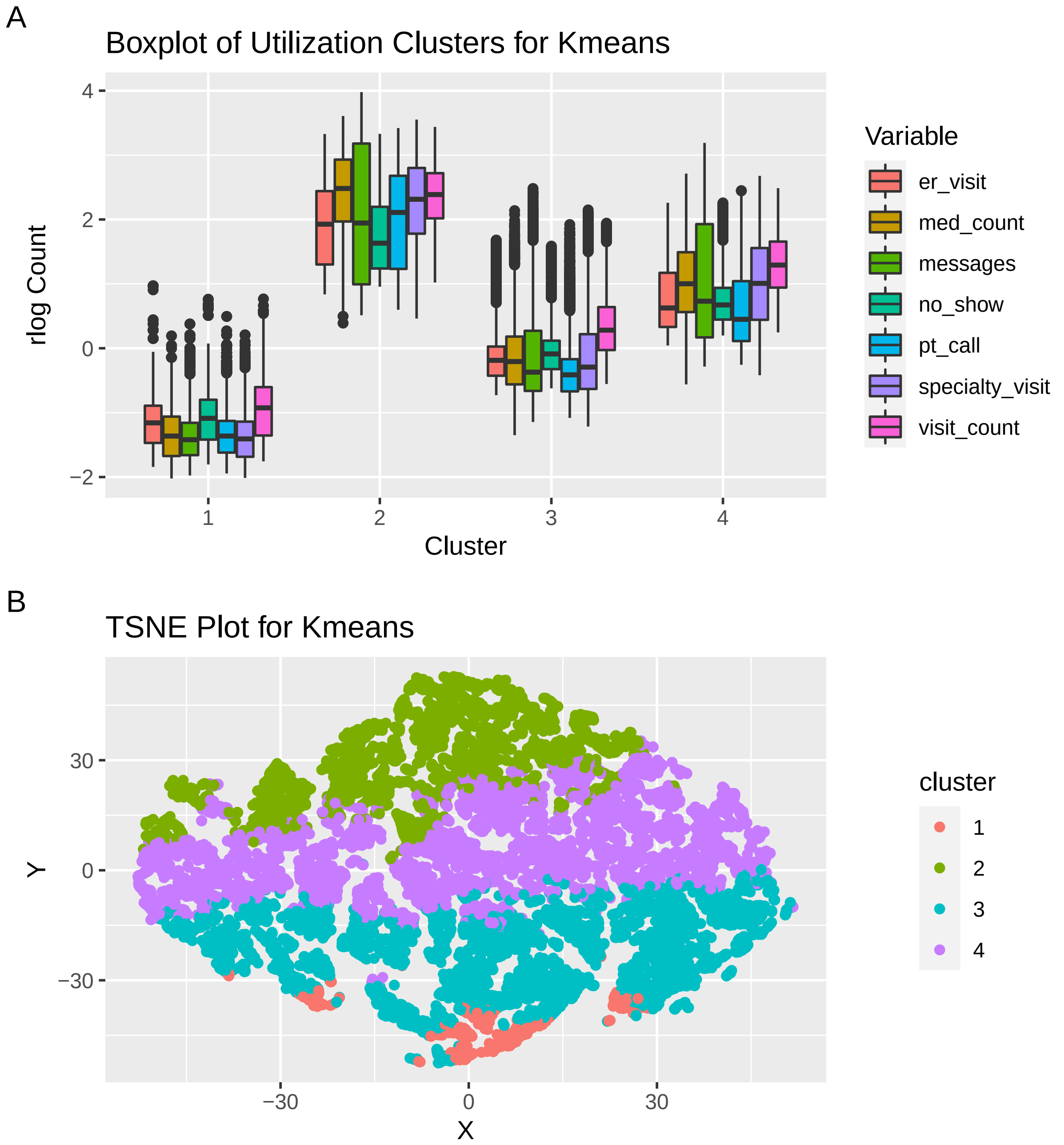


Figure A.3: Kmeans Clustering Plots

A.2.3 Gaussian Mixture Model

Table A.13: GMM Cluster Summary Table

cluster	utilization	min	q1	mean	median	q3	max
1	visit_count	0	2	4.61	4	7	11
2	visit_count	1	1	1.52	1	2	6
3	visit_count	1	2	3.67	3	5	11
1	specialty_visit	0	1	7.89	4	13	26
2	specialty_visit	0	0	0.41	0	1	4
3	specialty_visit	0	0	5.12	3	7	26
1	pt_call	0	0	2.42	1	4	10
2	pt_call	0	0	0.07	0	0	1
3	pt_call	0	0	1.32	0	2	10
1	no_show	0	0	0.59	0	1	2
2	no_show	0	0	0.00	0	0	0
3	no_show	0	0	0.00	0	0	0
1	messages	0	0	4.05	0	3	28
2	messages	0	0	0.16	0	0	2
3	messages	0	0	6.26	2	9	28
1	med_count	0	4	28.12	14	42	97
2	med_count	0	0	2.46	2	4	15
3	med_count	0	4	19.24	10	25	97
1	er_visit	0	0	1.30	1	2	3
2	er_visit	0	0	0.00	0	0	0
3	er_visit	0	0	0.00	0	0	0

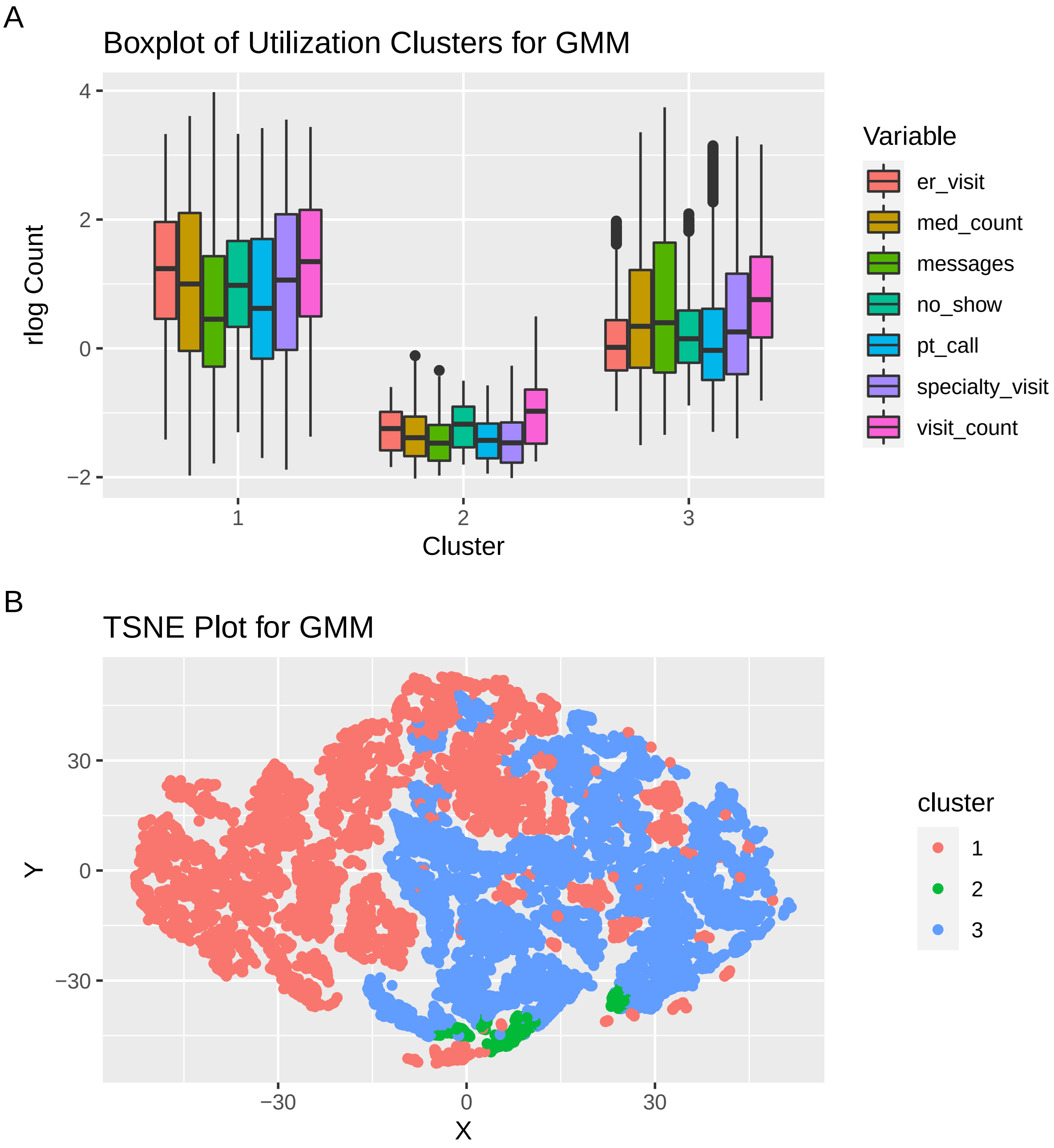


Figure A.4: Gaussian Mixture Model Clustering Plots

Table A.14: GMM Panel Weight Mapping

clusterId	medianVisit	clusterSize	scale	weight
1	4	6871	4	1.34
2	1	3554	1	0.33
3	3	7811	3	1.00

Table A.15: GMM Model Glance

model	G	BIC	logLik	df	hypvol	nobs
VVV	3	-73796.91	-36373.56	107		18236

Table A.16: GMM Model Summary

component	size	proportion	visit_count	no_show	med_count	messages	er_visit	specialty_visit	pt_call
1	6871	0.38	1.29	0.99	1.01	0.67	1.16	1.00	0.76
2	3554	0.19	-1.01	-1.22	-1.37	-1.45	-1.28	-1.43	-1.43
3	7811	0.43	0.82	0.21	0.51	0.69	0.09	0.44	0.18

Table A.17: GMM PSE Provider Aggregation

Provider	Practice	pat	pse	z
M	Mayland	2410	2085.70	0.52
L	Mayland	2227	2063.94	0.28
H	Nelson	1475	1567.79	-0.22
B	Tanglewood	1337	1425.51	-0.24

Provider	Practice	pat	pse	z
K	Nelson	1006	1052.89	-0.17
F	Nelson	896	954.47	-0.24
A	HEWHC	831	929.69	-0.43
J	HEWHC	829	905.59	-0.34
G	HEWHC	675	715.43	-0.23
C	Nelson	620	652.16	-0.20
N	Mayland	731	648.81	0.47
D	Nelson	584	605.96	-0.15
I	HEWHC	502	584.87	-0.61
O	Nelson	503	531.64	-0.22
E	Nelson	362	395.38	-0.36
P	HEWHC	372	370.27	0.02
Q	Nelson	259	291.26	-0.49

Table A.18: GMM Internal Correlation Model

term	estimate	std.error	statistic	p.value
visit_count	0.16	0.030	5.31	0.0004904
no_show	0.72	0.158	4.57	0.0013565
med_count	-0.01	0.004	-3.52	0.0065411
messages	0.02	0.007	2.84	0.0195303
er_visit	0.04	0.103	0.44	0.6725645
specialty_visit	0.10	0.014	6.84	0.0000753
pt_call	-0.07	0.013	-5.54	0.0003611

Table A.19: GMM Internal Correlation ANOVA

variable	Sum Sq	prsq	Df	F value	Pr(>F)
visit_count	25269.13	0.180	1	28.15	0.0004904
no_show	18708.93	0.133	1	20.84	0.0013565
med_count	11107.86	0.079	1	12.37	0.0065411
messages	7220.60	0.052	1	8.04	0.0195303
er_visit	171.27	0.001	1	0.19	0.6725645
specialty_visit	42038.23	0.300	1	46.83	0.0000753
pt_call	27553.29	0.197	1	30.69	0.0003611
Residuals	8079.87	0.058	9		

Table A.20: GMM Internal Correlation Model

term	estimate	std.error	statistic	p.value
visit_count	0.26	0.033	8.05	0.0000111
no_show	0.64	0.181	3.52	0.0055257
med_count	-0.02	0.004	-4.64	0.0009272
messages	0.01	0.005	2.73	0.0210231
er_visit	-0.46	0.115	-3.99	0.0025702
specialty_visit	0.10	0.012	8.03	0.0000114
pt_call	-0.02	0.010	-1.60	0.1417815

Table A.21: GMM Internal Correlation ANOVA

variable	Sum Sq	prsq	Df	F value	Pr(>F)
visit_count	61622.26	0.325	1	64.81	0.0000111
no_show	11790.05	0.062	1	12.40	0.0055257
med_count	20437.84	0.108	1	21.50	0.0009272
messages	7110.65	0.038	1	7.48	0.0210231
er_visit	15116.91	0.080	1	15.90	0.0025702
specialty_visit	61321.96	0.324	1	64.49	0.0000114
pt_call	2419.07	0.013	1	2.54	0.1417815
Residuals	9508.02	0.050	10		

A.2.4 Patient Burden Scoring

Table A.22: PBS Score Distribution

cut	n	median	mean	sd	q1	q3
1	4853	0.281	0.264	0.152	0.160	0.396
2	7081	0.757	0.756	0.140	0.641	0.875
3	5065	1.204	1.216	0.138	1.096	1.324
4	1237	1.622	1.644	0.110	1.548	1.715
overall	18236	0.804	0.813	0.440	0.473	1.134

Table A.23: PBS Cluster Summary Table

cluster	utilization	min	q1	mean	median	q3	max
1	visit_count	0	1	1.55	1	2	11
2	visit_count	1	2	2.99	2	4	11
3	visit_count	1	3	5.24	5	7	11
4	visit_count	1	7	8.50	9	11	11
1	specialty_visit	0	0	0.52	0	1	11
2	specialty_visit	0	0	3.55	2	5	26
3	specialty_visit	0	3	9.09	7	13	26
4	specialty_visit	0	11	17.80	19	26	26
1	pt_call	0	0	0.17	0	0	6
2	pt_call	0	0	0.78	0	1	10
3	pt_call	0	0	2.53	1	4	10
4	pt_call	0	3	6.50	8	10	10
1	no_show	0	0	0.07	0	0	2

cluster	utilization	min	q1	mean	median	q3	max
2	no_show	0	0	0.17	0	0	2
3	no_show	0	0	0.32	0	0	2
4	no_show	0	0	0.69	0	1	2
1	messages	0	0	0.30	0	0	8
2	messages	0	0	2.91	0	4	28
3	messages	0	0	7.73	2	14	28
4	messages	0	0	13.03	10	28	28
1	med_count	0	0	2.38	1	3	29
2	med_count	0	3	10.80	7	13	97
3	med_count	0	13	34.24	25	48	97
4	med_count	1	54	73.40	83	97	97
1	er_visit	0	0	0.08	0	0	3
2	er_visit	0	0	0.35	0	0	3
3	er_visit	0	0	0.81	0	1	3
4	er_visit	0	0	1.62	2	3	3

Table A.24: PBS PSE Provider Aggregation

Provider	Practice	pat	pse	z
L	Mayland	2227	1670.24	1.06
M	Mayland	2410	1660.72	1.37
H	Nelson	1475	1253.07	0.61
B	Tanglewood	1337	1135.08	0.62
K	Nelson	1006	958.10	0.19
F	Nelson	896	852.47	0.19

Provider	Practice	pat	pse	z
A	HEWHC	831	809.82	0.10
J	HEWHC	829	718.96	0.56
G	HEWHC	675	609.98	0.40
C	Nelson	620	581.98	0.25
I	HEWHC	502	523.93	-0.17
D	Nelson	584	489.95	0.71
O	Nelson	503	477.17	0.21
N	Mayland	731	458.90	1.88
E	Nelson	362	386.88	-0.27
Q	Nelson	259	308.28	-0.73
P	HEWHC	372	297.93	0.93

Table A.25: PBS Internal Correlation Model

term	estimate	std.error	statistic	p.value
visit_count	0.10	0.017	5.69	0.0002998
no_show	0.25	0.089	2.80	0.0206552
med_count	0.00	0.002	-1.12	0.2932609
messages	0.03	0.004	6.67	0.0000913
er_visit	0.01	0.058	0.17	0.8665502
specialty_visit	0.06	0.008	7.94	0.0000235
pt_call	0.00	0.007	0.24	0.8161995

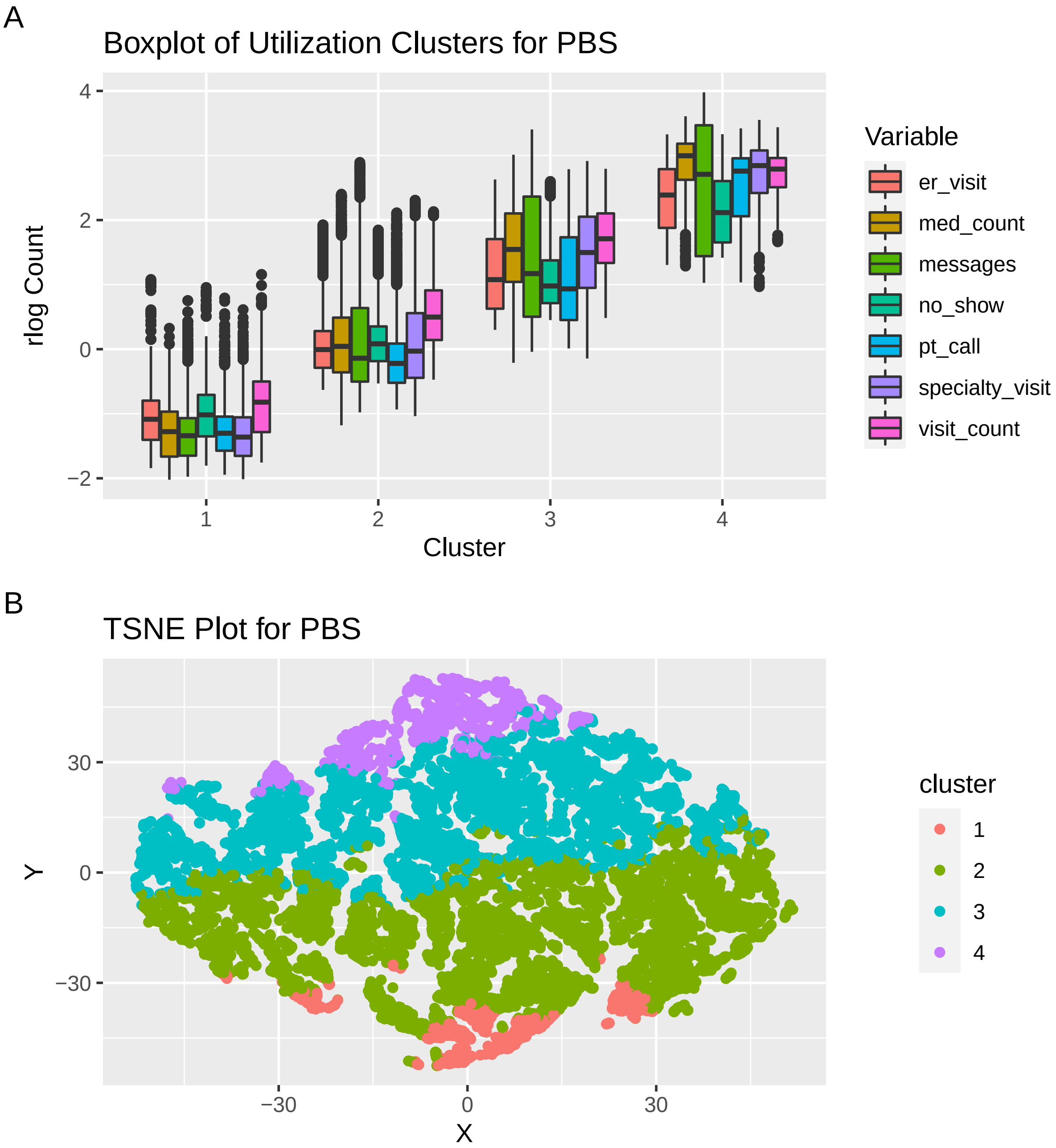


Figure A.5: Patient Burden Scoring Clustering Plots

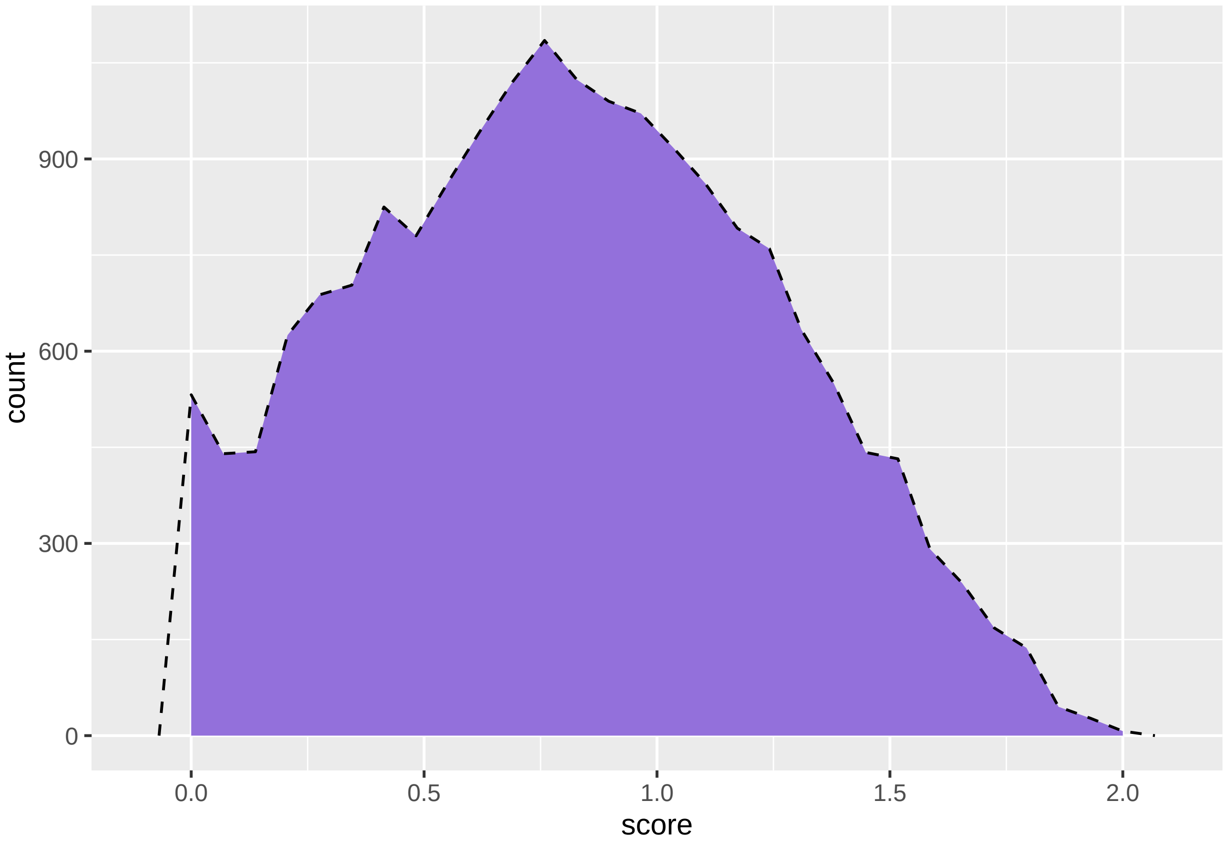


Figure A.6: Patient Burden Scoring Distribution

Table A.26: PBS Internal Correlation ANOVA

variable	Sum Sq	prsq	Df	F value	Pr(>F)
visit_count	9360.98	0.204	1	32.32	0.0002998
no_show	2273.30	0.050	1	7.85	0.0206552
med_count	360.81	0.008	1	1.25	0.2932609
messages	12896.93	0.282	1	44.53	0.0000913
er_visit	8.66	0.000	1	0.03	0.8665502
specialty_visit	18258.37	0.399	1	63.04	0.0000235
pt_call	16.59	0.000	1	0.06	0.8161995
Residuals	2606.50	0.057	9		

Table A.27: PBS Internal Correlation Model

term	estimate	std.error	statistic	p.value
visit_count	0.12	0.026	4.58	0.0010149
no_show	0.36	0.144	2.50	0.0311813
med_count	0.00	0.004	-0.50	0.6271139
messages	0.03	0.004	6.38	0.0000801
er_visit	-0.02	0.091	-0.17	0.8671396
specialty_visit	0.05	0.010	5.28	0.0003577
pt_call	0.02	0.008	2.13	0.0588424

Table A.28: PBS Internal Correlation ANOVA

variable	Sum Sq	prsq	Df	F value	Pr(>F)
visit_count	12605.72	0.189	1	20.95	0.0010149
no_show	3775.31	0.057	1	6.27	0.0311813
med_count	151.11	0.002	1	0.25	0.6271139
messages	24507.13	0.368	1	40.73	0.0000801
er_visit	17.73	0.000	1	0.03	0.8671396
specialty_visit	16772.33	0.252	1	27.88	0.0003577
pt_call	2734.32	0.041	1	4.54	0.0588424
Residuals	6016.65	0.090	10		

A.3 Simulation

A.3.1 Simulation Setup and Details

Table A.29: Simulation Variations

Simulation Number	Number of Doctors	Patient Count	Means	Variance
1	2	same	same	same
2	2	same	diff	same
3	2	same	same	diff
4	2	same	diff	diff
5	2	diff	same	same
6	2	diff	diff	same
7	2	diff	same	diff
8	2	diff	diff	diff

Simulation Number	Number of Doctors	Patient Count	Means	Variance
9	3	same	same	same
10	3	same	diff	same
11	3	same	same	diff
12	3	same	diff	diff
13	3	diff	same	same
14	3	diff	diff	same
15	3	diff	same	diff
16	3	diff	diff	diff

Table A.30: Simulation Parameters for Scenario 1

Doc	group	n	theta1	theta2	theta3	theta4	lam1	lam2	lam3	lam4
A	1	100	1.1	2.25	4	10	0.1	0.15	0.40	0.25
A	2	600	0.1	2.25	9	27	0.0	0.20	0.40	0.25
A	3	1500	0.1	0.25	1	4	0.0	-0.01	0.05	0.40
A	4	800	0.1	0.25	2	8	0.0	0.00	0.10	0.10
B	1	100	1.1	2.25	4	10	0.1	0.15	0.40	0.25
B	2	600	0.1	2.25	9	27	0.0	0.20	0.40	0.25
B	3	1500	0.1	0.25	1	4	0.0	-0.01	0.05	0.40
B	4	800	0.1	0.25	2	8	0.0	0.00	0.10	0.10

Table A.31: Simulation Parameters for Scenario 2

Doc	group	n	theta1	theta2	theta3	theta4	lam1	lam2	lam3	lam4
A	1	100	1.10	2.25	4.0	10	0.1	0.15	0.40	0.25

Doc	group	n	theta1	theta2	theta3	theta4	lam1	lam2	lam3	lam4
A	2	600	0.10	2.25	9.0	27	0.0	0.20	0.40	0.25
A	3	1500	0.10	0.25	1.0	4	0.0	-0.01	0.05	0.40
A	4	800	0.10	0.25	2.0	8	0.0	0.00	0.10	0.10
B	1	100	1.08	1.20	2.5	7	0.1	0.15	0.40	0.25
B	2	600	0.08	1.20	6.0	18	0.0	0.20	0.40	0.25
B	3	1500	0.08	0.20	0.8	3	0.0	-0.01	0.05	0.40
B	4	800	0.08	0.20	1.0	4	0.0	0.00	0.10	0.10

Table A.32: Simulation Parameters for Scenario 3

Doc	group	n	theta1	theta2	theta3	theta4	lam1	lam2	lam3	lam4
A	1	100	1.1	2.25	4	10	0.1	0.15	0.40	0.25
A	2	600	0.1	2.25	9	27	0.0	0.20	0.40	0.25
A	3	1500	0.1	0.25	1	4	0.0	-0.01	0.05	0.40
A	4	800	0.1	0.25	2	8	0.0	0.00	0.10	0.10
B	1	100	1.1	2.25	4	10	0.1	0.05	0.20	0.10
B	2	600	0.1	2.25	9	27	0.0	0.10	0.20	0.10
B	3	1500	0.1	0.25	1	4	0.0	-0.01	0.01	0.10
B	4	800	0.1	0.25	2	8	0.0	-0.01	0.05	0.10

Table A.33: Simulation Parameters for Scenario 4

Doc	group	n	theta1	theta2	theta3	theta4	lam1	lam2	lam3	lam4
A	1	100	1.10	2.25	4.0	10	0.1	0.15	0.40	0.25
A	2	600	0.10	2.25	9.0	27	0.0	0.20	0.40	0.25

Doc	group	n	theta1	theta2	theta3	theta4	lam1	lam2	lam3	lam4
A	3	1500	0.10	0.25	1.0	4	0.0	-0.01	0.05	0.40
A	4	800	0.10	0.25	2.0	8	0.0	0.00	0.10	0.10
B	1	100	1.08	1.20	2.5	7	0.1	0.05	0.20	0.10
B	2	600	0.08	1.20	6.0	18	0.0	0.10	0.20	0.10
B	3	1500	0.08	0.20	0.8	3	0.0	-0.01	0.01	0.10
B	4	800	0.08	0.20	1.0	4	0.0	-0.01	0.05	0.10

Table A.34: Simulation Parameters for Scenario 5

Doc	group	n	theta1	theta2	theta3	theta4	lam1	lam2	lam3	lam4
A	1	100	1.1	2.25	4	10	0.1	0.15	0.40	0.25
A	2	600	0.1	2.25	9	27	0.0	0.20	0.40	0.25
A	3	1500	0.1	0.25	1	4	0.0	-0.01	0.05	0.40
A	4	800	0.1	0.25	2	8	0.0	0.00	0.10	0.10
B	1	75	1.1	2.25	4	10	0.1	0.15	0.40	0.25
B	2	450	0.1	2.25	9	27	0.0	0.20	0.40	0.25
B	3	1125	0.1	0.25	1	4	0.0	-0.01	0.05	0.40
B	4	600	0.1	0.25	2	8	0.0	0.00	0.10	0.10

Table A.35: Simulation Parameters for Scenario 6

Doc	group	n	theta1	theta2	theta3	theta4	lam1	lam2	lam3	lam4
A	1	100	1.10	2.25	4.0	10	0.1	0.15	0.40	0.25
A	2	600	0.10	2.25	9.0	27	0.0	0.20	0.40	0.25
A	3	1500	0.10	0.25	1.0	4	0.0	-0.01	0.05	0.40

Doc	group	n	theta1	theta2	theta3	theta4	lam1	lam2	lam3	lam4
A	4	800	0.10	0.25	2.0	8	0.0	0.00	0.10	0.10
B	1	75	1.08	1.20	2.5	7	0.1	0.15	0.40	0.25
B	2	450	0.08	1.20	6.0	18	0.0	0.20	0.40	0.25
B	3	1125	0.08	0.20	0.8	3	0.0	-0.01	0.05	0.40
B	4	600	0.08	0.20	1.0	4	0.0	0.00	0.10	0.10

Table A.36: Simulation Parameters for Scenario 7

Doc	group	n	theta1	theta2	theta3	theta4	lam1	lam2	lam3	lam4
A	1	100	1.1	2.25	4	10	0.1	0.15	0.40	0.25
A	2	600	0.1	2.25	9	27	0.0	0.20	0.40	0.25
A	3	1500	0.1	0.25	1	4	0.0	-0.01	0.05	0.40
A	4	800	0.1	0.25	2	8	0.0	0.00	0.10	0.10
B	1	75	1.1	2.25	4	10	0.1	0.05	0.20	0.10
B	2	450	0.1	2.25	9	27	0.0	0.10	0.20	0.10
B	3	1125	0.1	0.25	1	4	0.0	-0.01	0.01	0.10
B	4	600	0.1	0.25	2	8	0.0	-0.01	0.05	0.10

Table A.37: Simulation Parameters for Scenario 8

Doc	group	n	theta1	theta2	theta3	theta4	lam1	lam2	lam3	lam4
A	1	100	1.10	2.25	4.0	10	0.1	0.15	0.40	0.25
A	2	600	0.10	2.25	9.0	27	0.0	0.20	0.40	0.25
A	3	1500	0.10	0.25	1.0	4	0.0	-0.01	0.05	0.40
A	4	800	0.10	0.25	2.0	8	0.0	0.00	0.10	0.10

Doc	group	n	theta1	theta2	theta3	theta4	lam1	lam2	lam3	lam4
B	1	75	1.08	1.20	2.5	7	0.1	0.05	0.20	0.10
B	2	450	0.08	1.20	6.0	18	0.0	0.10	0.20	0.10
B	3	1125	0.08	0.20	0.8	3	0.0	-0.01	0.01	0.10
B	4	600	0.08	0.20	1.0	4	0.0	-0.01	0.05	0.10

Table A.38: Simulation Parameters for Scenario 9

Doc	group	n	theta1	theta2	theta3	theta4	lam1	lam2	lam3	lam4
A	1	100	1.1	2.25	4	10	0.1	0.15	0.40	0.25
A	2	600	0.1	2.25	9	27	0.0	0.20	0.40	0.25
A	3	1500	0.1	0.25	1	4	0.0	-0.01	0.05	0.40
A	4	800	0.1	0.25	2	8	0.0	0.00	0.10	0.10
B	1	100	1.1	2.25	4	10	0.1	0.15	0.40	0.25
B	2	600	0.1	2.25	9	27	0.0	0.20	0.40	0.25
B	3	1500	0.1	0.25	1	4	0.0	-0.01	0.05	0.40
B	4	800	0.1	0.25	2	8	0.0	0.00	0.10	0.10
C	1	100	1.1	2.25	4	10	0.1	0.15	0.40	0.25
C	2	600	0.1	2.25	9	27	0.0	0.20	0.40	0.25
C	3	1500	0.1	0.25	1	4	0.0	-0.01	0.05	0.40
C	4	800	0.1	0.25	2	8	0.0	0.00	0.10	0.10

Table A.39: Simulation Parameters for Scenario 10

Doc	group	n	theta1	theta2	theta3	theta4	lam1	lam2	lam3	lam4
A	1	100	1.10	2.25	4.0	10	0.1	0.15	0.40	0.25

Doc	group	n	theta1	theta2	theta3	theta4	lam1	lam2	lam3	lam4
A	2	600	0.10	2.25	9.0	27	0.0	0.20	0.40	0.25
A	3	1500	0.10	0.25	1.0	4	0.0	-0.01	0.05	0.40
A	4	800	0.10	0.25	2.0	8	0.0	0.00	0.10	0.10
B	1	100	1.08	1.20	2.5	7	0.1	0.15	0.40	0.25
B	2	600	0.08	1.20	6.0	18	0.0	0.20	0.40	0.25
B	3	1500	0.08	0.20	0.8	3	0.0	-0.01	0.05	0.40
B	4	800	0.08	0.20	1.0	4	0.0	0.00	0.10	0.10
C	1	100	1.20	2.50	5.0	13	0.1	0.15	0.40	0.25
C	2	600	0.20	3.50	10.0	32	0.0	0.20	0.40	0.25
C	3	1500	0.20	0.50	2.0	7	0.0	-0.01	0.05	0.40
C	4	800	0.20	0.50	2.0	10	0.0	0.00	0.10	0.10

Table A.40: Simulation Parameters for Scenario 11

Doc	group	n	theta1	theta2	theta3	theta4	lam1	lam2	lam3	lam4
A	1	100	1.1	2.25	4	10	0.10	0.15	0.40	0.25
A	2	600	0.1	2.25	9	27	0.00	0.20	0.40	0.25
A	3	1500	0.1	0.25	1	4	0.00	-0.01	0.05	0.40
A	4	800	0.1	0.25	2	8	0.00	0.00	0.10	0.10
B	1	100	1.1	2.25	4	10	0.10	0.05	0.20	0.10
B	2	600	0.1	2.25	9	27	0.00	0.10	0.20	0.10
B	3	1500	0.1	0.25	1	4	0.00	-0.01	0.01	0.10
B	4	800	0.1	0.25	2	8	0.00	-0.01	0.05	0.10
C	1	100	1.1	2.25	4	10	0.15	0.25	0.50	0.30
C	2	600	0.1	2.25	9	27	0.05	0.25	0.50	0.30
C	3	1500	0.1	0.25	1	4	0.05	0.06	0.15	0.30

Doc	group	n	theta1	theta2	theta3	theta4	lam1	lam2	lam3	lam4
C	4	800	0.1	0.25	2	8	0.05	0.06	0.13	0.15

Table A.41: Simulation Parameters for Scenario 12

Doc	group	n	theta1	theta2	theta3	theta4	lam1	lam2	lam3	lam4
A	1	100	1.10	2.25	4.0	10	0.10	0.15	0.40	0.25
A	2	600	0.10	2.25	9.0	27	0.00	0.20	0.40	0.25
A	3	1500	0.10	0.25	1.0	4	0.00	-0.01	0.05	0.40
A	4	800	0.10	0.25	2.0	8	0.00	0.00	0.10	0.10
B	1	100	1.08	1.20	2.5	7	0.10	0.05	0.20	0.10
B	2	600	0.08	1.20	6.0	18	0.00	0.10	0.20	0.10
B	3	1500	0.08	0.20	0.8	3	0.00	-0.01	0.01	0.10
B	4	800	0.08	0.20	1.0	4	0.00	-0.01	0.05	0.10
C	1	100	1.20	2.50	5.0	13	0.15	0.25	0.50	0.30
C	2	600	0.20	3.50	10.0	32	0.05	0.25	0.50	0.30
C	3	1500	0.20	0.50	2.0	7	0.05	0.06	0.15	0.30
C	4	800	0.20	0.50	2.0	10	0.05	0.06	0.13	0.15

Table A.42: Simulation Parameters for Scenario 13

Doc	group	n	theta1	theta2	theta3	theta4	lam1	lam2	lam3	lam4
A	1	100	1.1	2.25	4	10	0.1	0.15	0.40	0.25
A	2	600	0.1	2.25	9	27	0.0	0.20	0.40	0.25
A	3	1500	0.1	0.25	1	4	0.0	-0.01	0.05	0.40
A	4	800	0.1	0.25	2	8	0.0	0.00	0.10	0.10

Doc	group	n	theta1	theta2	theta3	theta4	lam1	lam2	lam3	lam4
B	1	120	1.1	2.25	4	10	0.1	0.15	0.40	0.25
B	2	720	0.1	2.25	9	27	0.0	0.20	0.40	0.25
B	3	1800	0.1	0.25	1	4	0.0	-0.01	0.05	0.40
B	4	960	0.1	0.25	2	8	0.0	0.00	0.10	0.10
C	1	85	1.1	2.25	4	10	0.1	0.15	0.40	0.25
C	2	510	0.1	2.25	9	27	0.0	0.20	0.40	0.25
C	3	1275	0.1	0.25	1	4	0.0	-0.01	0.05	0.40
C	4	680	0.1	0.25	2	8	0.0	0.00	0.10	0.10

Table A.43: Simulation Parameters for Scenario 14

Doc	group	n	theta1	theta2	theta3	theta4	lam1	lam2	lam3	lam4
A	1	100	1.10	2.25	4.0	10	0.1	0.15	0.40	0.25
A	2	600	0.10	2.25	9.0	27	0.0	0.20	0.40	0.25
A	3	1500	0.10	0.25	1.0	4	0.0	-0.01	0.05	0.40
A	4	800	0.10	0.25	2.0	8	0.0	0.00	0.10	0.10
B	1	120	1.08	1.20	2.5	7	0.1	0.15	0.40	0.25
B	2	720	0.08	1.20	6.0	18	0.0	0.20	0.40	0.25
B	3	1800	0.08	0.20	0.8	3	0.0	-0.01	0.05	0.40
B	4	960	0.08	0.20	1.0	4	0.0	0.00	0.10	0.10
C	1	85	1.20	2.50	5.0	13	0.1	0.15	0.40	0.25
C	2	510	0.20	3.50	10.0	32	0.0	0.20	0.40	0.25
C	3	1275	0.20	0.50	2.0	7	0.0	-0.01	0.05	0.40
C	4	680	0.20	0.50	2.0	10	0.0	0.00	0.10	0.10

Table A.44: Simulation Parameters for Scenario 15

Doc	group	n	theta1	theta2	theta3	theta4	lam1	lam2	lam3	lam4
A	1	100	1.1	2.25	4	10	0.10	0.15	0.40	0.25
A	2	600	0.1	2.25	9	27	0.00	0.20	0.40	0.25
A	3	1500	0.1	0.25	1	4	0.00	-0.01	0.05	0.40
A	4	800	0.1	0.25	2	8	0.00	0.00	0.10	0.10
B	1	120	1.1	2.25	4	10	0.10	0.05	0.20	0.10
B	2	720	0.1	2.25	9	27	0.00	0.10	0.20	0.10
B	3	1800	0.1	0.25	1	4	0.00	-0.01	0.01	0.10
B	4	960	0.1	0.25	2	8	0.00	-0.01	0.05	0.10
C	1	85	1.1	2.25	4	10	0.15	0.25	0.50	0.30
C	2	510	0.1	2.25	9	27	0.05	0.25	0.50	0.30
C	3	1275	0.1	0.25	1	4	0.05	0.06	0.15	0.30
C	4	680	0.1	0.25	2	8	0.05	0.06	0.13	0.15

Table A.45: Simulation Parameters for Scenario 16

Doc	group	n	theta1	theta2	theta3	theta4	lam1	lam2	lam3	lam4
A	1	100	1.10	2.25	4.0	10	0.10	0.15	0.40	0.25
A	2	600	0.10	2.25	9.0	27	0.00	0.20	0.40	0.25
A	3	1500	0.10	0.25	1.0	4	0.00	-0.01	0.05	0.40
A	4	800	0.10	0.25	2.0	8	0.00	0.00	0.10	0.10
B	1	120	1.08	1.20	2.5	7	0.10	0.05	0.20	0.10
B	2	720	0.08	1.20	6.0	18	0.00	0.10	0.20	0.10
B	3	1800	0.08	0.20	0.8	3	0.00	-0.01	0.01	0.10
B	4	960	0.08	0.20	1.0	4	0.00	-0.01	0.05	0.10

Doc	group	n	theta1	theta2	theta3	theta4	lam1	lam2	lam3	lam4
C	1	85	1.20	2.50	5.0	13	0.15	0.25	0.50	0.30
C	2	510	0.20	3.50	10.0	32	0.05	0.25	0.50	0.30
C	3	1275	0.20	0.50	2.0	7	0.05	0.06	0.15	0.30
C	4	680	0.20	0.50	2.0	10	0.05	0.06	0.13	0.15

A.3.2 Simulation Results

Table A.46: Kmeans Simulation Results

sim	Bias	MSE	Precision	Coverage	ARI	ClusterPurity	ClassErrorRate
1	-0.111	85.204	0.012	0.502	0.543	0.871	0.255
2	404.175	163586.672	0.004	1.000	0.396	0.814	0.341
3	110.814	12417.717	0.007	1.000	0.536	0.869	0.259
4	586.338	344323.640	0.002	1.000	0.344	0.782	0.389
5	-1.333	94.436	0.011	1.000	0.543	0.871	0.254
6	420.939	177462.608	0.004	1.000	0.410	0.820	0.331
7	112.961	12924.851	0.006	1.000	0.535	0.869	0.261
8	613.254	376411.583	0.003	1.000	0.358	0.791	0.374
9	0.089	113.728	0.009	0.335	0.542	0.871	0.255
10	516.480	267161.705	0.002	1.000	0.396	0.806	0.343
11	163.728	27020.832	0.005	1.000	0.514	0.858	0.269
12	804.997	649057.933	0.001	1.000	0.319	0.760	0.398
13	-1.660	129.771	0.008	1.000	0.543	0.871	0.255
14	242.645	59164.619	0.003	1.000	0.380	0.800	0.355
15	-205.047	42389.532	0.003	1.000	0.516	0.858	0.268

sim	Bias	MSE	Precision	Coverage	ARI	ClusterPurity	ClassErrorRate
16	369.133	136601.914	0.003	0.426	0.303	0.751	0.416

Table A.47: GMM Simulation Results

sim	Bias	MSE	Precision	Coverage	ARI	ClusterPurity	ClassErrorRate
1	0.063	120.440	0.008	0.496	0.636	0.880	0.195
2	257.487	68099.039	0.001	1.000	0.340	0.775	0.352
3	42.502	1934.029	0.008	0.999	0.622	0.874	0.207
4	341.478	118349.699	0.001	1.000	0.274	0.726	0.381
5	-1.066	129.138	0.008	1.000	0.637	0.878	0.195
6	256.477	67524.782	0.001	1.000	0.363	0.790	0.338
7	43.439	2009.996	0.008	1.000	0.625	0.876	0.204
8	342.377	119564.086	0.000	1.000	0.294	0.741	0.370
9	0.021	103.962	0.010	0.330	0.634	0.874	0.198
10	386.344	153165.535	0.000	0.998	0.382	0.789	0.329
11	57.687	3534.318	0.005	0.992	0.582	0.862	0.226
12	459.041	230059.842	0.000	0.999	0.324	0.757	0.351
13	-1.170	129.633	0.008	1.000	0.634	0.874	0.199
14	84.572	12041.699	0.000	0.196	0.351	0.780	0.349
15	-73.772	5752.673	0.003	1.000	0.584	0.862	0.226
16	176.438	34734.500	0.000	0.947	0.300	0.743	0.360

Table A.48: PBS Simulation Results

sim	Bias	MSE	Precision	Coverage	ARI	ClusterPurity	ClassErrorRate
1	647.536	419828.7	0.002	0.498	0.609	0.841	0.206
2	754.778	570150.4	0.002	1.000	0.441	0.767	0.302
3	738.940	546572.6	0.002	1.000	0.603	0.837	0.208
4	795.452	633062.7	0.003	1.000	0.375	0.751	0.348
5	809.161	655477.1	0.001	1.000	0.609	0.841	0.206
6	904.787	819550.9	0.001	1.000	0.458	0.770	0.291
7	885.876	785504.9	0.001	1.000	0.603	0.837	0.210
8	960.083	922533.1	0.001	1.000	0.394	0.752	0.330
9	649.346	422072.6	0.002	0.332	0.608	0.841	0.206
10	952.589	907666.5	0.004	1.000	0.455	0.766	0.280
11	815.325	665441.6	0.001	1.000	0.570	0.825	0.218
12	1048.532	1099680.8	0.004	1.000	0.344	0.719	0.347
13	777.970	605803.2	0.002	1.000	0.608	0.841	0.206
14	612.392	375266.6	0.004	0.822	0.435	0.756	0.297
15	601.766	362777.2	0.002	1.000	0.573	0.827	0.216
16	623.949	389631.6	0.003	1.000	0.325	0.717	0.365

Appendix B

APPENDIX FOR CHAPTER 2

Table B.1: Four Cut Methodology

Cut	Assignment	Description
1	Sole provider	Patient has only seen one provider in the past year
2	Majority provider	Patient has seen multiple providers but has seen one a majority of the time
3	Provider who performed last physical	No majority provider seen by a patient in past year
4	Last seen provider	First 3 cuts do not settle pcp assignment so it goes to the last seen provider

B.1 Algorithms

B.2 VCU Data Analysis

B.2.1 Overall

Table B.2: VCU Health Systems Demographic Summary

covariate	factor	num	pct
age	0-17	1651	9.1%
age	18-34	3848	21%

covariate	factor	num	pct
age	35-49	4081	22%
age	50-64	5131	28%
age	65+	3525	19%
insurance	None/Other	1098	6%
insurance	Medicaid	3295	18%
insurance	Medicare	3559	20%
insurance	Private	10284	56%
practice	HEWHC FamMed	3791	21%
practice	Mayland FamMed	5360	29%
practice	Nelson FamMed	6035	33%
practice	Tanglewood FamMed	3050	17%
race	Other/Unknown	2692	15%
race	Black	6756	37%
race	White	8788	48%
sex	F	10159	56%
sex	M	8077	44%

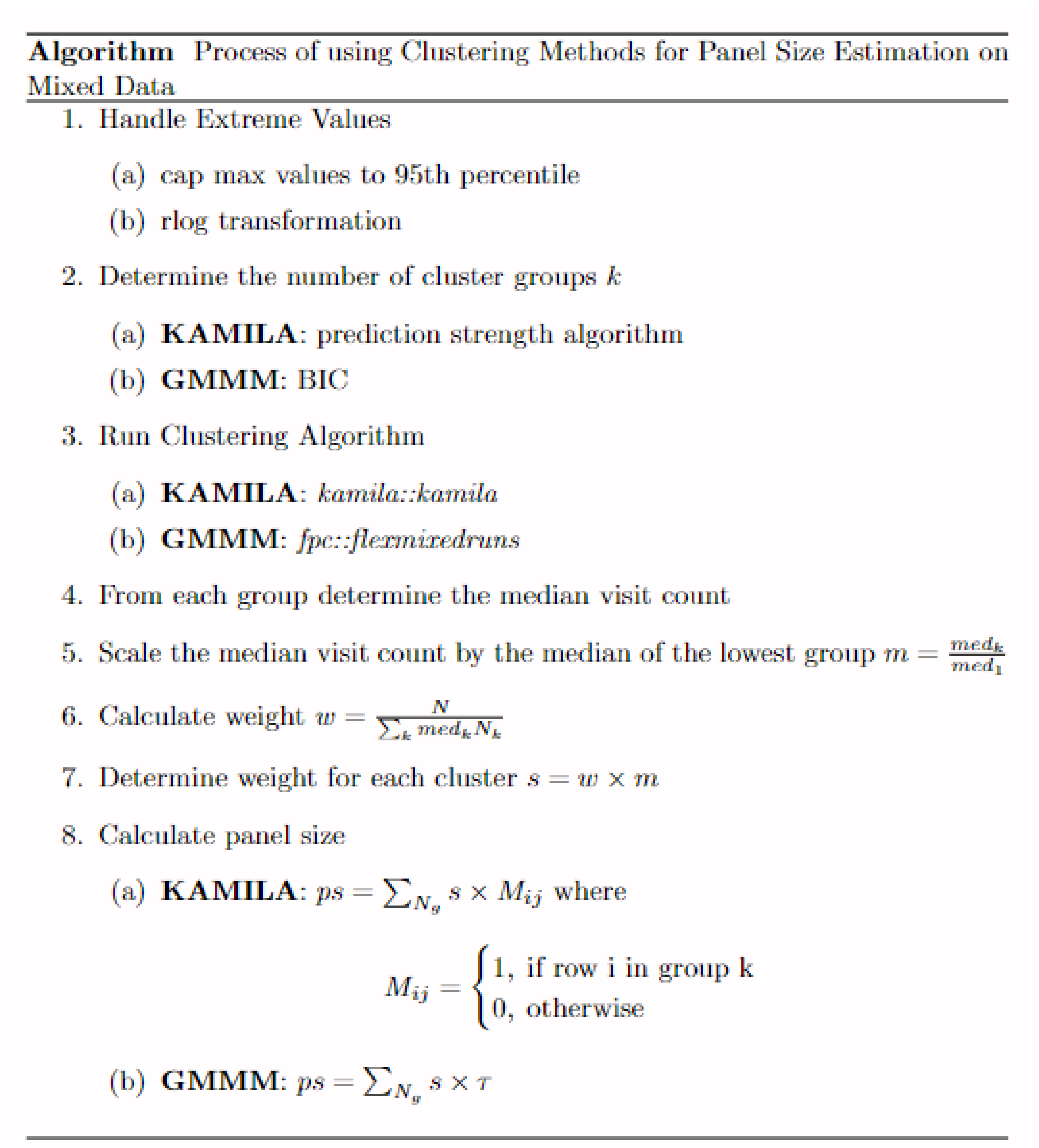
Table B.3: VCU Health Systems Utilization Summary

utilization	min	p25	median	mean	p75	max
visit_count	0	1	3	3.80	5	66
specialty_visit	0	0	2	5.94	7	176
pt_call	0	0	0	1.95	2	114
no_show	0	0	0	0.26	0	12
messages	0	0	0	5.55	5	532

utilization	min	p25	median	mean	p75	max
med_count	0	2	8	22.95	24	917
er_visit	0	0	0	0.68	1	97

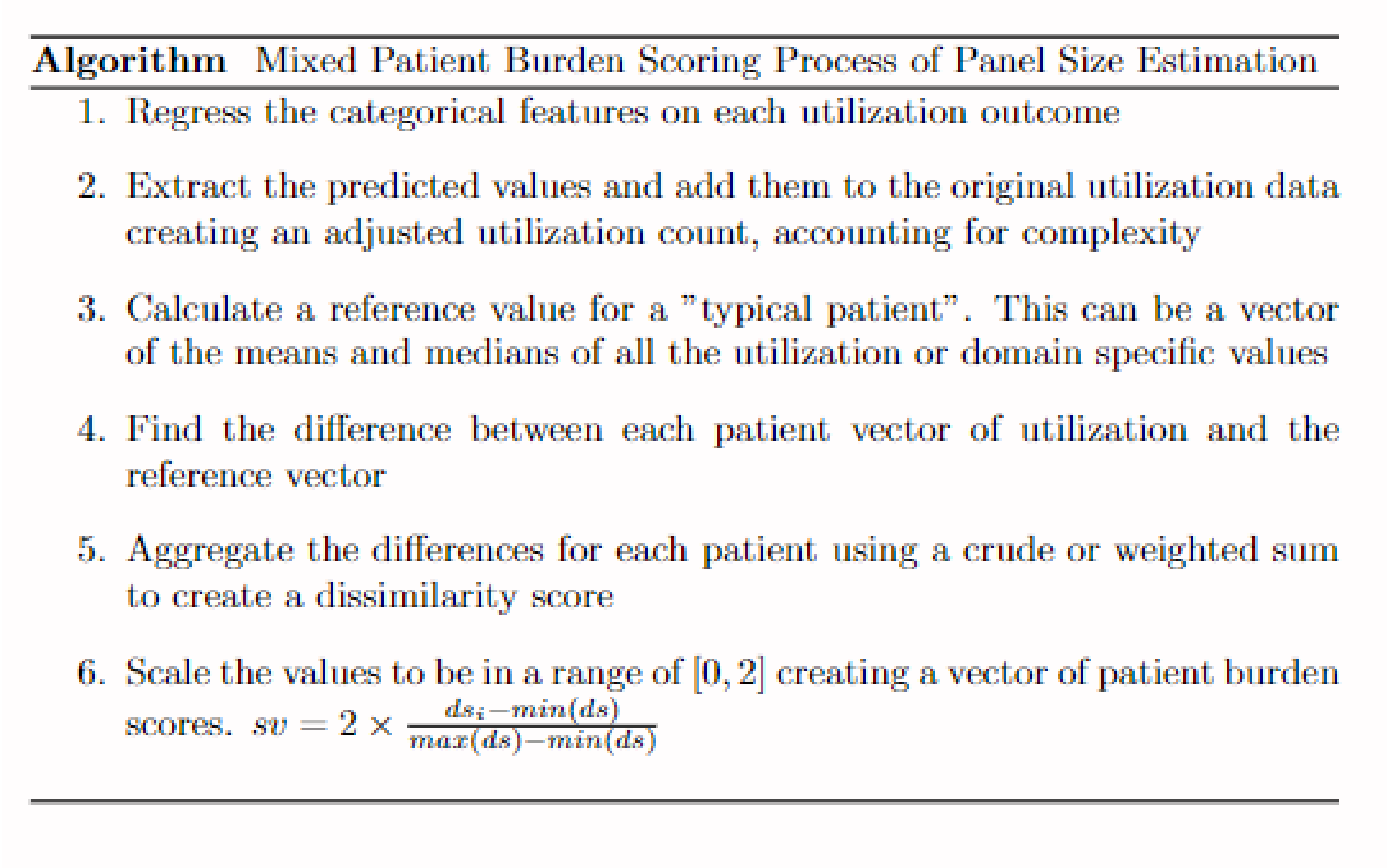
Table B.4: VCU Health Systems Provider Aggregation Summary

Provider	Practice	pat	visit_count	no_show	med_count	messages	er_visit	specialty_visit	pt_call
M	Mayland	2410	7065	320	35848	14681	382	9833	90
L	Mayland	2227	6523	322	34614	17197	377	9177	104
H	Nelson	1475	3783	280	20205	5919	1187	11098	2845
B	Tanglewood	1337	8058	299	53248	306	1017	6096	25
K	Nelson	1006	3392	203	19131	7494	514	6761	2530
F	Nelson	896	3218	156	16254	6164	503	5726	2028
A	HEWHC	831	4012	335	24368	1021	618	5544	3274
J	HEWHC	829	3687	386	16269	902	592	4233	2612
N	Mayland	731	1599	128	7285	2585	168	3253	31
G	HEWHC	675	2931	250	15608	1793	423	3820	2459
C	Nelson	620	2060	103	12010	3504	270	4113	1568
D	Nelson	584	1745	112	7708	2931	375	3824	1235
O	Nelson	503	1818	76	9362	3110	250	3455	1141
I	HEWHC	502	2424	312	16083	1024	406	3527	2282
P	HEWHC	372	1673	126	6893	1191	167	1401	1175
E	Nelson	362	1518	72	9910	3519	214	2561	1307
Q	Nelson	259	1493	61	10404	1750	144	2212	1235



```
Algorithm Process of using Clustering Methods for Panel Size Estimation on Mixed Data
1. Handle Extreme Values
    (a) cap max values to 95th percentile
    (b) rlog transformation
2. Determine the number of cluster groups $k$
    (a) KAMILA: prediction strength algorithm
    (b) GMMM: BIC
3. Run Clustering Algorithm
    (a) KAMILA: kamila::kamila
    (b) GMMM: fpc::flexmixedruns
4. From each group determine the median visit count
5. Scale the median visit count by the median of the lowest group $m = \frac{med_k}{med_1}$
6. Calculate weight $w = \frac{N}{\sum_k med_k N_k}$
7. Determine weight for each cluster $s = w \times m$
8. Calculate panel size
    (a) KAMILA: $ps = \sum_{N_g} s \times M_{ij}$ where
            $M_{ij} = \begin{cases} 1, & \text{if row i in group k} \\ 0, & \text{otherwise} \end{cases}$
    (b) GMMM: $ps = \sum_{N_g} s \times \tau$
```

Figure B.1: PSE Algorithm for mixed data in EMR



Algorithm Mixed Patient Burden Scoring Process of Panel Size Estimation

1. Regress the categorical features on each utilization outcome
2. Extract the predicted values and add them to the original utilization data creating an adjusted utilization count, accounting for complexity
3. Calculate a reference value for a "typical patient". This can be a vector of the means and medians of all the utilization or domain specific values
4. Find the difference between each patient vector of utilization and the reference vector
5. Aggregate the differences for each patient using a crude or weighted sum to create a dissimilarity score
6. Scale the values to be in a range of $[0, 2]$ creating a vector of patient burden scores. $sv = 2 \times \frac{ds_i - min(ds)}{max(ds) - min(ds)}$

Figure B.2: Mixed Patient Burden Scoring Algorithm

B.2.2 KAMILA

Table B.5: KAMILA Utilization Summary by Cluster

cluster	utilization	min	q1	mean	median	q3	max
1	visit_count	1	5	7.31	7	10	11
2	visit_count	1	2	3.25	3	4	11
3	visit_count	0	1	1.44	1	2	11
4	visit_count	1	3	4.94	4	7	11
5	visit_count	1	1	2.55	2	3	11
1	specialty_visit	0	7	14.47	14	24	26
2	specialty_visit	0	1	5.87	4	8	26
3	specialty_visit	0	0	0.38	0	0	7
4	specialty_visit	0	1	6.96	4	10	26
5	specialty_visit	0	0	2.53	1	3	26
1	pt_call	0	1	5.12	5	10	10
2	pt_call	0	0	0.71	0	1	10
3	pt_call	0	0	0.12	0	0	5
4	pt_call	0	0	2.07	1	3	10
5	pt_call	0	0	0.58	0	1	10
1	no_show	0	0	0.56	0	1	2
2	no_show	0	0	0.13	0	0	2
3	no_show	0	0	0.06	0	0	2
4	no_show	0	0	0.33	0	0	2
5	no_show	0	0	0.14	0	0	2
1	messages	0	0	9.74	4	21	28
2	messages	2	7	14.37	13	21	28
3	messages	0	0	0.20	0	0	5

cluster	utilization	min	q1	mean	median	q3	max
4	messages	0	0	0.41	0	0	9
5	messages	0	0	1.76	0	2	28
1	med_count	0	32	59.17	57	97	97
2	med_count	0	5	14.15	10	19	97
3	med_count	0	0	1.96	1	3	17
4	med_count	0	10	29.22	21	40	97
5	med_count	0	2	7.10	5	10	57
1	er_visit	0	0	1.34	1	3	3
2	er_visit	0	0	0.22	0	0	3
3	er_visit	0	0	0.07	0	0	3
4	er_visit	0	0	0.84	0	1	3
5	er_visit	0	0	0.26	0	0	3

Table B.6: KAMILA Demographic Summary by Cluster

	Cluster1	Cluster2	Cluster3	Cluster4	Cluster5	Overall
	(N=2853)	(N=2641)	(N=3977)	(N=3438)	(N=5327)	(N=18236)
race.f						
Other/Unknown	152 (5.3%)	257 (9.7%)	964 (24.2%)	338 (9.8%)	981 (18.4%)	2692 (14.8%)
Black	1690 (59.2%)	670 (25.4%)	1011 (25.4%)	1687 (49.1%)	1698 (31.9%)	6756 (37.0%)
White	1011 (35.4%)	1714 (64.9%)	2002 (50.3%)	1413 (41.1%)	2648 (49.7%)	8788 (48.2%)
insurance.f						

	Cluster1	Cluster2	Cluster3	Cluster4	Cluster5	Overall
None/Other	131 (4.6%)	41 (1.6%)	316 (7.9%)	248 (7.2%)	362 (6.8%)	1098 (6.0%)
Medicaid	568 (19.9%)	80 (3.0%)	874 (22.0%)	796 (23.2%)	977 (18.3%)	3295 (18.1%)
Medicare	1119 (39.2%)	213 (8.1%)	189 (4.8%)	1361 (39.6%)	677 (12.7%)	3559 (19.5%)
Private	1035 (36.3%)	2307 (87.4%)	2598 (65.3%)	1033 (30.0%)	3311 (62.2%)	10284 (56.4%)
sex.f						
F	2008 (70.4%)	1594 (60.4%)	1752 (44.1%)	1990 (57.9%)	2815 (52.8%)	10159 (55.7%)
M	845 (29.6%)	1047 (39.6%)	2225 (55.9%)	1448 (42.1%)	2512 (47.2%)	8077 (44.3%)
age.f						
0-17	23 (0.8%)	8 (0.3%)	769 (19.3%)	258 (7.5%)	593 (11.1%)	1651 (9.1%)
18-34	259 (9.1%)	661 (25.0%)	1419 (35.7%)	315 (9.2%)	1194 (22.4%)	3848 (21.1%)
35-49	545 (19.1%)	889 (33.7%)	825 (20.7%)	475 (13.8%)	1347 (25.3%)	4081 (22.4%)
50-64	1174 (41.1%)	769 (29.1%)	709 (17.8%)	1097 (31.9%)	1382 (25.9%)	5131 (28.1%)
65+	852 (29.9%)	314 (11.9%)	255 (6.4%)	1293 (37.6%)	811 (15.2%)	3525 (19.3%)

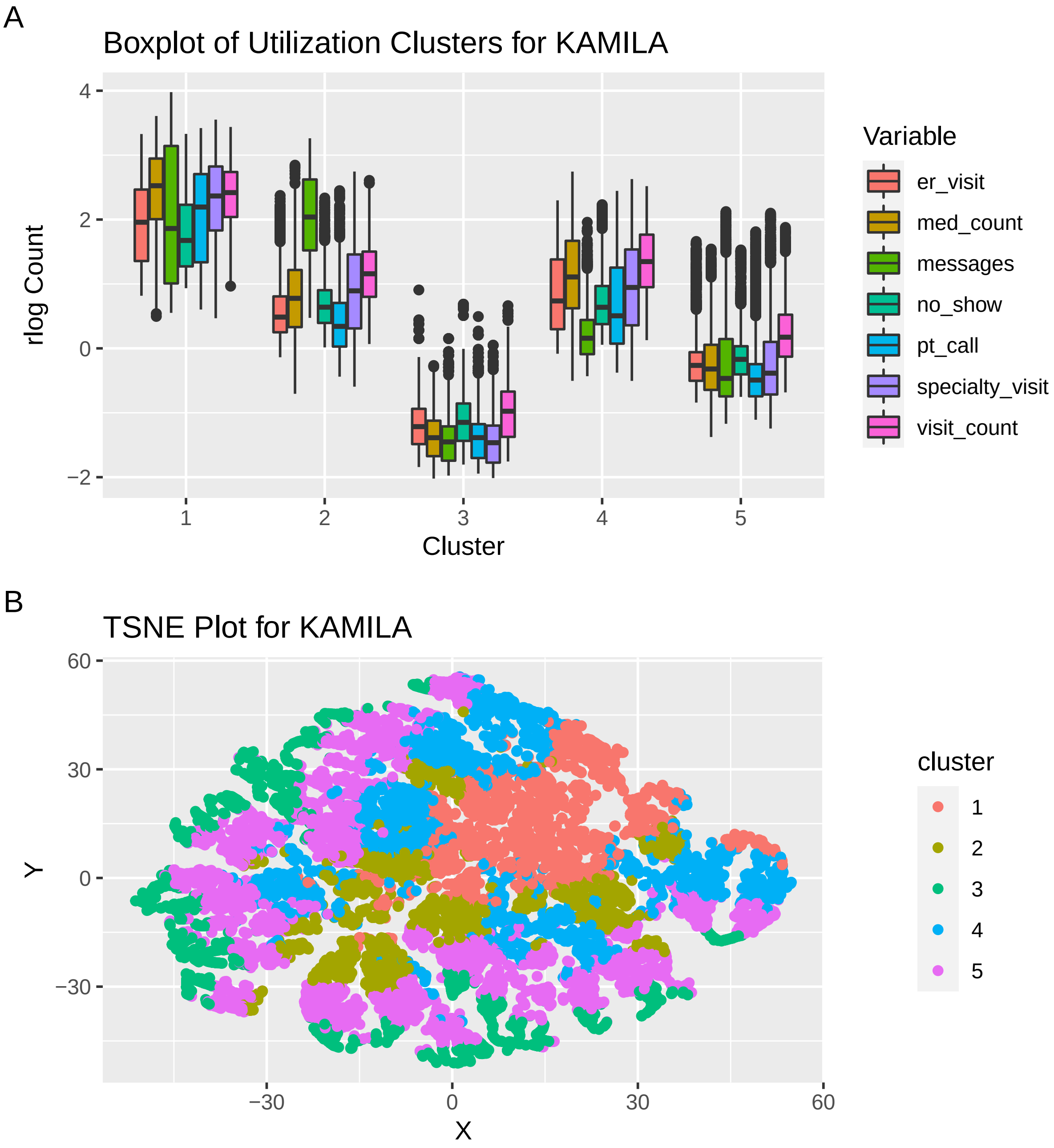


Figure B.3: KAMILA Clustering Plots

Table B.7: KAMILA Panel Weight Mapping

clusterId	medianVisit	clusterSize	scale	weight
1	7	2853	7	2.27
2	3	2641	3	0.97
3	1	3977	1	0.32
4	4	3438	4	1.30
5	2	5327	2	0.65

Table B.8: KAMILA PSE Provider Aggregation

Provider	Practice	pat	pse	z
M	Mayland	2410	1922.21	0.82
L	Mayland	2227	1860.31	0.66
H	Nelson	1475	1540.49	-0.16
B	Tanglewood	1337	1467.25	-0.34
K	Nelson	1006	1179.83	-0.60
A	HEWHC	831	1074.19	-0.97
F	Nelson	896	1033.69	-0.54
J	HEWHC	829	950.41	-0.52
G	HEWHC	675	815.61	-0.73
C	Nelson	620	710.62	-0.53
I	HEWHC	502	694.09	-1.27
D	Nelson	584	593.64	-0.06
O	Nelson	503	578.09	-0.55
N	Mayland	731	532.07	1.27
E	Nelson	362	479.58	-1.13

Provider	Practice	pat	pse	z
Q	Nelson	259	411.85	-1.89
P	HEWHC	372	382.69	-0.12

Table B.9: KAMILA Internal Correlation Model

term	estimate	std.error	statistic	p.value
visit_count	0.13	0.016	8.46	0.0000141
no_show	0.20	0.083	2.46	0.0362760
med_count	0.00	0.002	-1.25	0.2438465
messages	0.02	0.004	4.71	0.0011120
er_visit	-0.03	0.054	-0.51	0.6196714
specialty_visit	0.08	0.007	10.76	0.0000019
pt_call	0.03	0.007	4.71	0.0011041

Table B.10: KAMILA Internal Correlation ANOVA

variable	Sum Sq	prsq	Df	F value	Pr(>F)
visit_count	17831.37	0.288	1	71.65	0.0000141
no_show	1503.64	0.024	1	6.04	0.0362760
med_count	387.04	0.006	1	1.56	0.2438465
messages	5509.65	0.089	1	22.14	0.0011120
er_visit	65.74	0.001	1	0.26	0.6196714
specialty_visit	28838.42	0.466	1	115.88	0.0000019
pt_call	5521.56	0.089	1	22.19	0.0011041
Residuals	2239.82	0.036	9		

Table B.11: KAMILA Internal Correlation Model

term	estimate	std.error	statistic	p.value
visit_count	0.15	0.023	6.62	0.0000592
no_show	0.29	0.128	2.26	0.0469899
med_count	0.00	0.003	-0.76	0.4624437
messages	0.02	0.004	5.05	0.0004977
er_visit	-0.04	0.081	-0.46	0.6560436
specialty_visit	0.06	0.009	7.43	0.0000223
pt_call	0.03	0.007	4.65	0.0009054

Table B.12: KAMILA Internal Correlation ANOVA

variable	Sum Sq	prsq	Df	F value	Pr(>F)
visit_count	20945.71	0.270	1	43.84	0.0000592
no_show	2450.25	0.032	1	5.13	0.0469899
med_count	278.94	0.004	1	0.58	0.4624437
messages	12193.71	0.157	1	25.52	0.0004977
er_visit	100.65	0.001	1	0.21	0.6560436
specialty_visit	26401.27	0.341	1	55.26	0.0000223
pt_call	10338.02	0.133	1	21.64	0.0009054
Residuals	4777.27	0.062	10		

B.2.3 Gaussian Multinomial Mixture Model

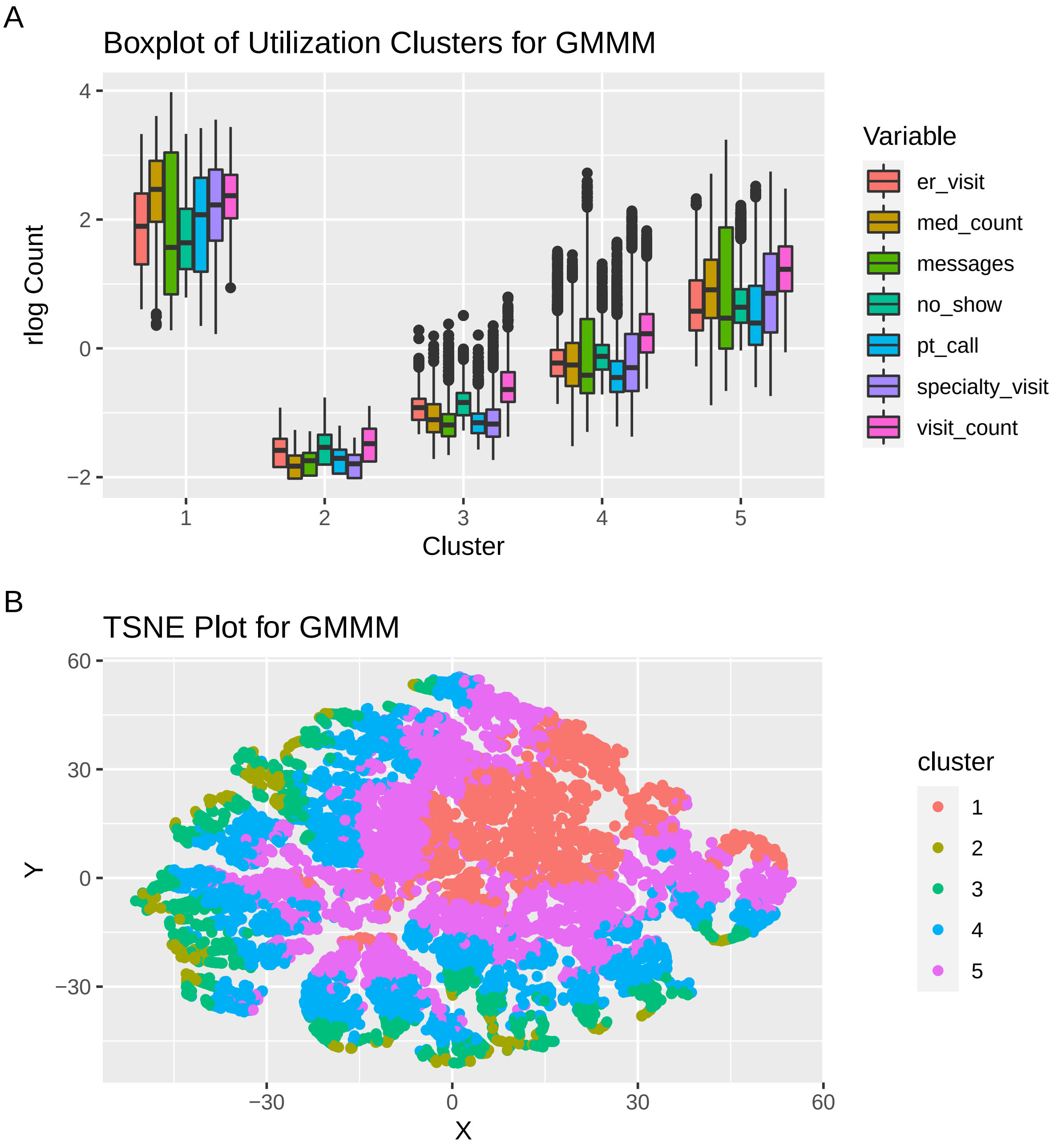


Figure B.4: Gaussian Multinomial Mixture Model Clustering Plots

Table B.13: GMMM Utilization Summary by Cluster

cluster	utilization	min	q1	mean	median	q3	max
1	visit_count	1	5	7.38	7	10	11
2	visit_count	1	1	1.12	1	1	2
3	visit_count	0	1	1.75	1	2	11
4	visit_count	1	1	2.47	2	3	11
5	visit_count	1	2	4.08	4	5	11
1	specialty_visit	0	6	13.43	12	22	26
2	specialty_visit	0	0	0.08	0	0	1
3	specialty_visit	0	0	0.74	0	1	9
4	specialty_visit	0	0	2.87	1	4	26
5	specialty_visit	0	1	6.38	4	9	26
1	pt_call	0	0	4.88	5	10	10
2	pt_call	0	0	0.03	0	0	1
3	pt_call	0	0	0.18	0	0	4
4	pt_call	0	0	0.53	0	1	10
5	pt_call	0	0	1.47	0	2	10
1	no_show	0	0	0.57	0	1	2
2	no_show	0	0	0.02	0	0	1
3	no_show	0	0	0.07	0	0	2
4	no_show	0	0	0.10	0	0	2
5	no_show	0	0	0.26	0	0	2
1	messages	0	0	8.91	2	19	28
2	messages	0	0	0.03	0	0	1
3	messages	0	0	0.40	0	0	6
4	messages	0	0	2.60	0	4	28

cluster	utilization	min	q1	mean	median	q3	max
5	messages	0	0	6.11	1	10	28
1	med_count	0	32	58.71	56	97	97
2	med_count	0	0	0.87	0	1	5
3	med_count	0	1	3.12	2	4	26
4	med_count	0	2	6.93	5	9	55
5	med_count	0	7	21.01	14	27	97
1	er_visit	0	0	1.32	1	3	3
2	er_visit	0	0	0.02	0	0	1
3	er_visit	0	0	0.09	0	0	2
4	er_visit	0	0	0.24	0	0	3
5	er_visit	0	0	0.57	0	1	3

Table B.14: GMMM Demographic Summary by Cluster

	Cluster1	Cluster2	Cluster3	Cluster4	Cluster5	Overall
	(N=3207)	(N=1670)	(N=2807)	(N=4825)	(N=5727)	(N=18236)
race.f						
Other/Unknown	160 (5.0%)	479 (28.7%)	582 (20.7%)	875 (18.1%)	596 (10.4%)	2692 (14.8%)
Black	1894 (59.1%)	385 (23.1%)	760 (27.1%)	1479 (30.7%)	2238 (39.1%)	6756 (37.0%)
White	1153 (36.0%)	806 (48.3%)	1465 (52.2%)	2471 (51.2%)	2893 (50.5%)	8788 (48.2%)
insurance.f						

	Cluster1	Cluster2	Cluster3	Cluster4	Cluster5	Overall
None/Other	146 (4.6%)	139 (8.3%)	201 (7.2%)	315 (6.5%)	297 (5.2%)	1098 (6.0%)
Medicaid	630 (19.6%)	379 (22.7%)	569 (20.3%)	807 (16.7%)	910 (15.9%)	3295 (18.1%)
Medicare	1292 (40.3%)	53 (3.2%)	217 (7.7%)	590 (12.2%)	1407 (24.6%)	3559 (19.5%)
Private	1139 (35.5%)	1099 (65.8%)	1820 (64.8%)	3113 (64.5%)	3113 (54.4%)	10284 (56.4%)
sex.f						
F	2216 (69.1%)	687 (41.1%)	1345 (47.9%)	2517 (52.2%)	3394 (59.3%)	10159 (55.7%)
M	991 (30.9%)	983 (58.9%)	1462 (52.1%)	2308 (47.8%)	2333 (40.7%)	8077 (44.3%)
age.f						
0-17	31 (1.0%)	344 (20.6%)	475 (16.9%)	486 (10.1%)	315 (5.5%)	1651 (9.1%)
18-34	287 (8.9%)	697 (41.7%)	834 (29.7%)	1089 (22.6%)	941 (16.4%)	3848 (21.1%)
35-49	594 (18.5%)	332 (19.9%)	622 (22.2%)	1269 (26.3%)	1264 (22.1%)	4081 (22.4%)
50-64	1289 (40.2%)	223 (13.4%)	613 (21.8%)	1259 (26.1%)	1747 (30.5%)	5131 (28.1%)
65+	1006 (31.4%)	74 (4.4%)	263 (9.4%)	722 (15.0%)	1460 (25.5%)	3525 (19.3%)

Table B.15: GMMM Panel Weight Mapping

clusterId	medianVisit	clusterSize	scale	weight
1	7.0	3207	7.0	2.10
2	1.0	1670	1.0	0.30
3	1.5	2807	1.5	0.45
4	2.0	4825	2.0	0.60
5	4.0	5727	4.0	1.20

Table B.16: GMMM PSE Provider Aggregation

Provider	Practice	pat	pse	z
M	Mayland	2410	1975.22	0.72
L	Mayland	2227	1922.66	0.53
H	Nelson	1475	1508.75	-0.08
B	Tanglewood	1337	1483.29	-0.38
K	Nelson	1006	1165.37	-0.55
A	HEWHC	831	1048.56	-0.88
F	Nelson	896	1040.17	-0.57
J	HEWHC	829	934.30	-0.46
G	HEWHC	675	795.78	-0.64
C	Nelson	620	693.50	-0.44
I	HEWHC	502	672.09	-1.14
D	Nelson	584	590.47	-0.04
O	Nelson	503	573.40	-0.52
N	Mayland	731	537.01	1.23
E	Nelson	362	474.56	-1.09

Provider	Practice	pat	pse	z
Q	Nelson	259	398.79	-1.76
P	HEWHC	372	377.82	-0.06

Table B.17: GMMM Internal Correlation Model

term	estimate	std.error	statistic	p.value
visit_count	0.15	0.015	9.80	0.0000042
no_show	0.21	0.078	2.65	0.0263751
med_count	0.00	0.002	-1.91	0.0879980
messages	0.02	0.003	6.20	0.0001591
er_visit	-0.02	0.051	-0.30	0.7679980
specialty_visit	0.07	0.007	10.61	0.0000022
pt_call	0.02	0.006	3.45	0.0073182

Table B.18: GMMM Internal Correlation ANOVA

variable	Sum Sq	prsq	Df	F value	Pr(>F)
visit_count	21257.80	0.345	1	96.07	0.0000042
no_show	1556.62	0.025	1	7.03	0.0263751
med_count	810.00	0.013	1	3.66	0.0879980
messages	8501.88	0.138	1	38.42	0.0001591
er_visit	20.46	0.000	1	0.09	0.7679980
specialty_visit	24890.57	0.404	1	112.49	0.0000022
pt_call	2627.93	0.043	1	11.88	0.0073182
Residuals	1991.45	0.032	9		

Table B.19: GMMM Internal Correlation Model

term	estimate	std.error	statistic	p.value
visit_count	0.14	0.020	7.08	0.0000336
no_show	0.37	0.110	3.34	0.0075475
med_count	0.00	0.003	0.14	0.8921321
messages	0.02	0.003	6.53	0.0000666
er_visit	-0.08	0.069	-1.15	0.2756990
specialty_visit	0.06	0.007	8.12	0.0000103
pt_call	0.04	0.006	7.24	0.0000280

Table B.20: GMMM Internal Correlation ANOVA

variable	Sum Sq	prsq	Df	F value	Pr(>F)
visit_count	17552.22	0.215	1	50.17	0.0000336
no_show	3892.55	0.048	1	11.13	0.0075475
med_count	6.77	0.000	1	0.02	0.8921321
messages	14905.65	0.182	1	42.60	0.0000666
er_visit	465.18	0.006	1	1.33	0.2756990
specialty_visit	23084.36	0.282	1	65.98	0.0000103
pt_call	18322.94	0.224	1	52.37	0.0000280
Residuals	3498.74	0.043	10		

B.2.4 Mixed Patient Burden Scoring

Table B.21: Mixed Pbs Score Distribution

cut	n	median	mean	sd	q1	q3
1	2829	0.365	0.346	0.108	0.271	0.435
2	7498	0.771	0.765	0.141	0.648	0.889
3	6391	1.208	1.222	0.139	1.104	1.331
4	1518	1.624	1.643	0.105	1.555	1.712
overall	18236	0.926	0.933	0.392	0.635	1.219

Table B.22: Mixed PBS Cluster Summary Table

cluster	utilization	min	q1	mean	median	q3	max
1	visit_count	1	1	1.52	1	2	11
2	visit_count	0	1	2.55	2	3	11
3	visit_count	1	3	4.69	4	6	11
4	visit_count	1	6	8.16	9	11	11
1	specialty_visit	0	0	0.32	0	0	10
2	specialty_visit	0	0	2.56	1	3	26
3	specialty_visit	0	2	7.89	5	12	26
4	specialty_visit	0	10	16.59	17	26	26
1	pt_call	0	0	0.13	0	0	6
2	pt_call	0	0	0.54	0	1	10
3	pt_call	0	0	2.04	0	3	10
4	pt_call	0	3	6.40	8	10	10
1	no_show	0	0	0.06	0	0	2

cluster	utilization	min	q1	mean	median	q3	max
2	no_show	0	0	0.15	0	0	2
3	no_show	0	0	0.29	0	0	2
4	no_show	0	0	0.62	0	1	2
1	messages	0	0	0.19	0	0	15
2	messages	0	0	2.84	0	3	28
3	messages	0	0	6.59	1	11	28
4	messages	0	0	8.80	2	18	28
1	med_count	0	0	1.30	1	2	18
2	med_count	0	2	6.91	4	9	97
3	med_count	0	11	29.46	21	40	97
4	med_count	1	50	71.46	79	97	97
1	er_visit	0	0	0.07	0	0	3
2	er_visit	0	0	0.27	0	0	3
3	er_visit	0	0	0.68	0	1	3
4	er_visit	0	0	1.57	2	3	3

Table B.23: Mixed PBS PSE Provider Aggregation

Provider	Practice	pat	pse	z
L	Mayland	2227	1880.48	0.62
M	Mayland	2410	1880.24	0.90
H	Nelson	1475	1429.89	0.12
B	Tanglewood	1337	1370.15	-0.09
K	Nelson	1006	1080.15	-0.27
F	Nelson	896	937.26	-0.17

Provider	Practice	pat	pse	z
A	HEWHC	831	899.27	-0.30
J	HEWHC	829	811.54	0.08
G	HEWHC	675	698.75	-0.14
C	Nelson	620	661.93	-0.26
I	HEWHC	502	575.88	-0.54
N	Mayland	731	561.53	1.05
D	Nelson	584	553.29	0.21
O	Nelson	503	536.01	-0.25
E	Nelson	362	420.67	-0.61
Q	Nelson	259	339.51	-1.12
P	HEWHC	372	308.98	0.78

Table B.24: Mixed PBS Internal Correlation Model

term	estimate	std.error	statistic	p.value
visit_count	0.11	0.030	3.59	0.0058358
no_show	0.30	0.154	1.95	0.0825679
med_count	0.00	0.004	-0.49	0.6325781
messages	0.02	0.007	3.36	0.0084509
er_visit	0.01	0.101	0.07	0.9420699
specialty_visit	0.08	0.014	5.95	0.0002166
pt_call	-0.02	0.013	-1.57	0.1508311

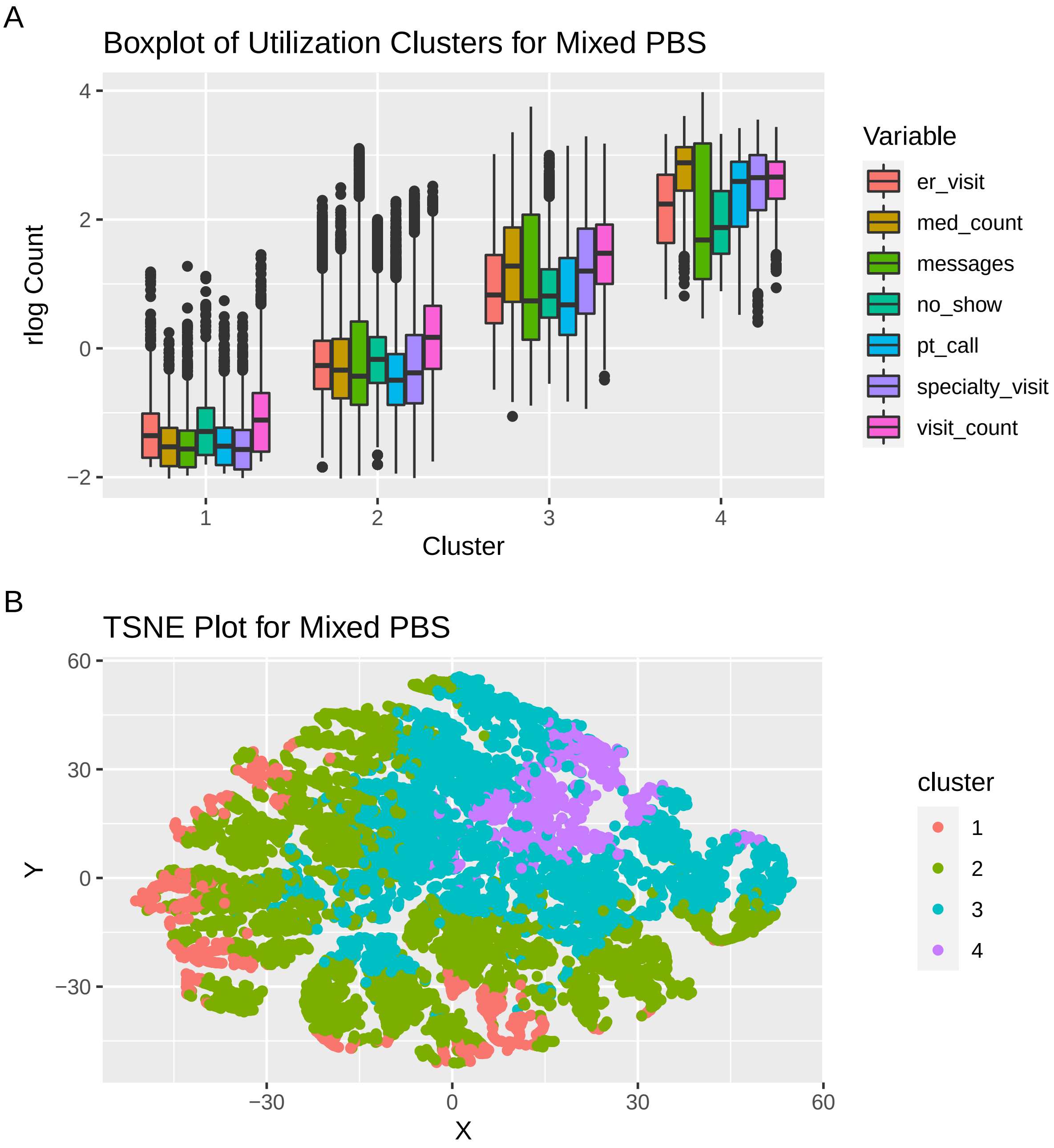


Figure B.5: Mixed Patient Burden Scoring Clustering Plots

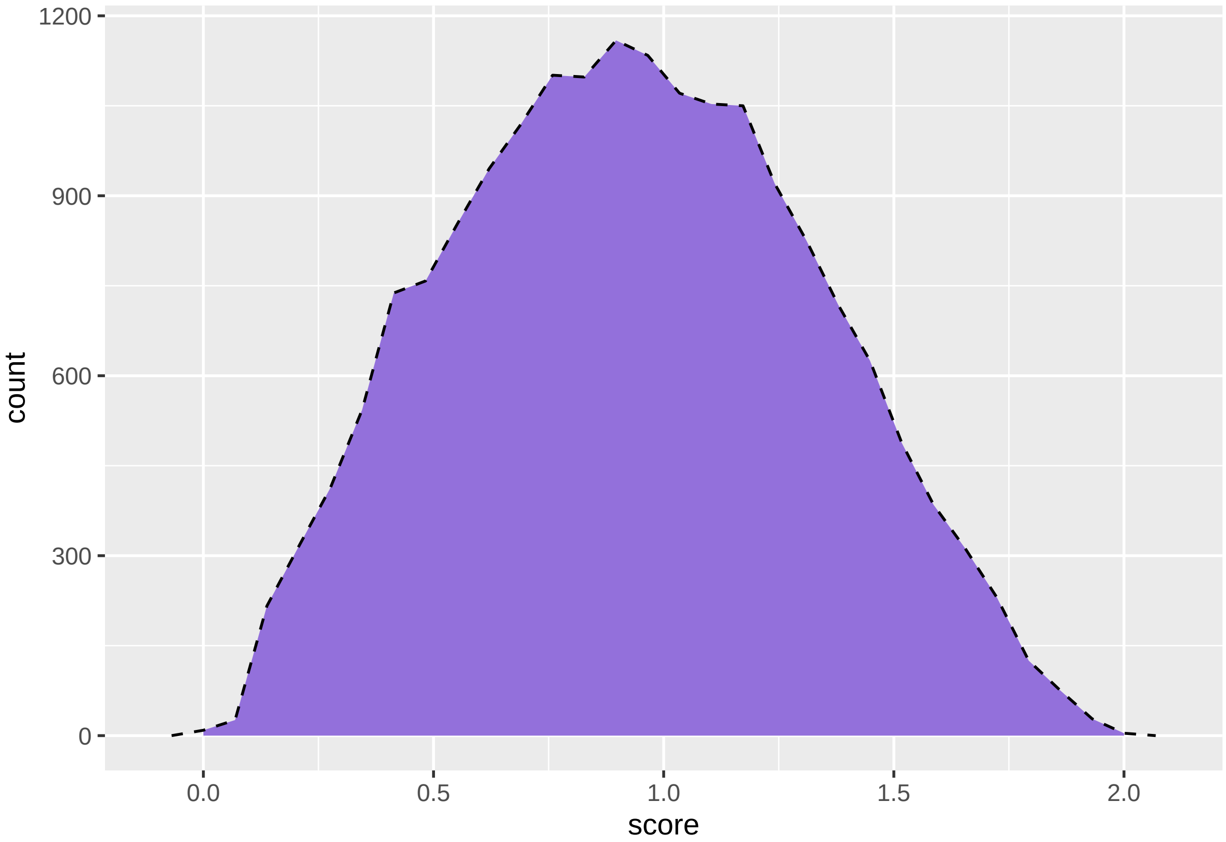


Figure B.6: Patient Burden Scoring Distribution

Table B.25: Mixed PBS Internal Correlation ANOVA

variable	Sum Sq	prsq	Df	F value	Pr(>F)
visit_count	11053.56	0.172	1	12.89	0.0058358
no_show	3270.68	0.051	1	3.81	0.0825679
med_count	209.96	0.003	1	0.24	0.6325781
messages	9654.26	0.150	1	11.26	0.0084509
er_visit	4.79	0.000	1	0.01	0.9420699
specialty_visit	30308.48	0.471	1	35.35	0.0002166
pt_call	2114.01	0.033	1	2.47	0.1508311
Residuals	7717.49	0.120	9		

Table B.26: Mixed PBS Internal Correlation Model

term	estimate	std.error	statistic	p.value
visit_count	0.11	0.039	2.95	0.0144327
no_show	0.39	0.215	1.81	0.1000437
med_count	0.00	0.005	0.11	0.9115019
messages	0.03	0.006	4.10	0.0021502
er_visit	-0.03	0.137	-0.25	0.8094566
specialty_visit	0.06	0.015	4.21	0.0018011
pt_call	0.01	0.012	0.87	0.4021765

Table B.27: Mixed PBS Internal Correlation ANOVA

variable	Sum Sq	prsq	Df	F value	Pr(>F)
visit_count	11781.65	0.152	1	8.73	0.0144327
no_show	4433.64	0.057	1	3.28	0.1000437
med_count	17.54	0.000	1	0.01	0.9115019
messages	22676.45	0.293	1	16.80	0.0021502
er_visit	82.76	0.001	1	0.06	0.8094566
specialty_visit	23923.94	0.309	1	17.72	0.0018011
pt_call	1033.29	0.013	1	0.77	0.4021765
Residuals	13500.56	0.174	10		

B.3 Simulation

B.3.1 Setup and Settings

Table B.28: Simulation Settings

Simulation Number	Number of Doctors	Patient Count	Utilization	Demographics
1	2	same	same	same
2	2	same	diff	same
3	2	same	same	diff
4	2	same	diff	diff
5	2	diff	same	same
6	2	diff	diff	same
7	2	diff	same	diff
8	2	diff	diff	diff

Simulation Number	Number of Doctors	Patient Count	Utilization	Demographics
9	3	same	same	same
10	3	same	diff	same
11	3	same	same	diff
12	3	same	diff	diff
13	3	diff	same	same
14	3	diff	diff	same
15	3	diff	same	diff
16	3	diff	diff	diff

B.3.2 Results

Table B.29: KAMILA Simulation Results

sim	Bias	MSE	Precision	Coverage	ARI	ClusterPurity	ClassErrorRate
1	0.147	72.793	0.014	0.507	0.557	0.883	0.204
2	194.303	39381.102	0.001	1.000	0.540	0.868	0.230
3	4.823	500.702	0.002	0.591	0.513	0.885	0.298
4	190.815	36589.566	0.006	1.000	0.502	0.879	0.306
5	-1.782	71.966	0.015	1.000	0.557	0.883	0.204
6	188.577	37530.765	0.001	1.000	0.554	0.871	0.226
7	1.936	452.524	0.002	1.000	0.514	0.885	0.294
8	184.641	34263.267	0.006	1.000	0.512	0.884	0.297
9	-0.003	94.726	0.011	0.338	0.557	0.883	0.204
10	635.610	404501.541	0.002	1.000	0.384	0.798	0.347
11	13.358	700.507	0.002	0.813	0.526	0.884	0.248
12	546.212	300762.488	0.000	1.000	0.373	0.802	0.319

sim	Bias	MSE	Precision	Coverage	ARI	ClusterPurity	ClassErrorRate
13	-1.764	106.522	0.010	1.000	0.557	0.883	0.204
14	730.641	534178.491	0.003	1.000	0.381	0.793	0.348
15	12.458	946.318	0.001	1.000	0.521	0.885	0.260
16	657.485	436359.288	0.000	1.000	0.366	0.798	0.340

Table: GMMM Simulation Results

|| || || ||

Table B.30: Mixed PBS Simulation Results

sim	Bias	MSE	Precision	Coverage	ARI	ClusterPurity	ClassErrorRate
1	678.486	469490.8	0	0.506	0.646	0.853	0.152
2	932.555	878768.1	0	1.000	0.631	0.847	0.156
3	821.048	687188.8	0	1.000	0.639	0.850	0.153
4	1181.570	1406383.2	0	1.000	0.504	0.791	0.210
5	813.628	674785.7	0	1.000	0.646	0.853	0.152
6	1098.774	1220288.3	0	1.000	0.641	0.851	0.151
7	1014.709	1049135.8	0	1.000	0.630	0.846	0.156
8	1417.844	2025015.7	0	1.000	0.484	0.780	0.221
9	680.974	469455.6	0	0.338	0.654	0.856	0.148
10	953.351	914596.4	0	1.000	0.442	0.762	0.245
11	739.358	553890.1	0	0.982	0.644	0.853	0.157
12	1181.395	1402377.2	0	1.000	0.373	0.729	0.279
13	814.143	671235.1	0	1.000	0.653	0.856	0.148
14	1124.880	1273614.2	0	1.000	0.439	0.761	0.245
15	906.008	831926.6	0	1.000	0.642	0.852	0.157

sim	Bias	MSE	Precision	Coverage	ARI	ClusterPurity	ClassErrorRate
16	1431.631	2057255.1	0	1.000	0.350	0.716	0.291

B.4 Synthea

B.4.1 Overview

Table B.31: Synthea Continuous Summary Table

covariateName	min	p25	median	mean	p75	max
Charlson index - Romano adaptation	1	1	1	1.73	2	9
Diabetes Comorbidity Severity Index (DCSI)	1	2	2	2.28	2	6
CHADS2VASc	0	0	1	1.16	2	6
condition_era distinct concept count during day -730 through 0 concept_count relative to index	1	1	1	1.73	2	9
drug_era distinct concept count during day -730 through 0 concept_count relative to index	1	2	3	3.53	5	17
procedure_occurrence distinct concept count during day -730 through 0 concept_count relative to index	1	1	2	3.20	3	19
measurement distinct concept count during day -730 through 0 concept_count relative to index	1	12	21	22.02	29	72
observation distinct concept count during day -730 through 0 concept_count relative to index	1	1	1	1.14	1	10
visit_occurrence concept count during day -730 through 0 concept_count relative to index: Inpatient Visit	1	1	1	2.50	2	31
visit_occurrence concept count during day -730 through 0 concept_count relative to index: Outpatient Visit	1	3	4	6.67	8	252

covariateName	min	p25	median	mean	p75	max
visit_occurrence concept count during day -730 through 0 concept_count relative to index: Emergency Room Visit	1	1	1	1.45	1	37

Table B.32: Synthea Categorical Summary Table

covariateName	num	pct
age group: 15 - 19	353	4.4%
age group: 20 - 24	929	12%
age group: 25 - 29	1105	14%
age group: 30 - 34	789	9.8%
age group: 35 - 39	563	7%
age group: 40 - 44	603	7.5%
age group: 45 - 49	615	7.6%
age group: 50 - 54	636	7.9%
age group: 55 - 59	651	8.1%
age group: 60 - 64	593	7.3%
age group: 65 - 69	384	4.8%
age group: 70 - 74	269	3.3%
age group: 75 - 79	223	2.8%
age group: 80 - 84	136	1.7%
age group: 85 - 89	61	0.76%
age group: 90 - 94	44	0.54%
age group: 95 - 99	37	0.46%
age group: 100 - 104	56	0.69%
age group: 105 - 109	30	0.37%

covariateName	num	pct
gender = MALE	3807	47%
race = Asian	214	2.6%
race = Black or African American	4071	50%
race = White	3742	46%
gender = FEMALE	4270	53%
condition_era group during day -730 through 0 days relative to index: Osteoporosis	38	0.47%
condition_era group during day -730 through 0 days relative to index: Neoplasm of breast	18	0.22%
condition_era group during day -730 through 0 days relative to index: Sepsis	12	0.15%
condition_era group during day -730 through 0 days relative to index: Neoplasm of digestive tract	38	0.47%
condition_era group during day -730 through 0 days relative to index: Diabetes mellitus	37	0.46%
condition_era group during day -730 through 0 days relative to index: Neoplasm of respiratory tract	5	0.062%
condition_era group during day -730 through 0 days relative to index: Chronic obstructive lung disease	10	0.12%
condition_era group during day -730 through 0 days relative to index: Chronic sinusitis	84	1%
condition_era group during day -730 through 0 days relative to index: Heart failure	28	0.35%
condition_era group during day -730 through 0 days relative to index: Hypertensive disorder	189	2.3%
condition_era group during day -730 through 0 days relative to index: Asthma	43	0.53%
condition_era group during day -730 through 0 days relative to index: Migraine	12	0.15%

covariateName	num	pct
condition_era group during day -730 through 0 days relative to index: Heart disease	127	1.6%
condition_era group during day -730 through 0 days relative to index: Cerebrovascular disease	33	0.41%
condition_era group during day -730 through 0 days relative to index: Complication of pregnancy, childbirth and/or the puerperium	199	2.5%
condition_era group during day -730 through 0 days relative to index: Chronic pain	77	0.95%
condition_era group during day -730 through 0 days relative to index: Anemia	297	3.7%
condition_era group during day -730 through 0 days relative to index: Anxiety	3	0.037%
condition_era group during day -730 through 0 days relative to index: Concussion injury of brain	173	2.1%
condition_era group during day -730 through 0 days relative to index: Neoplasm of digestive system	38	0.47%
condition_era group during day -730 through 0 days relative to index: Neoplasm of trunk	75	0.93%
condition_era group during day -730 through 0 days relative to index: Neoplasm of intra-abdominal organs	53	0.66%
condition_era group during day -730 through 0 days relative to index: Neoplasm of intrathoracic organs	5	0.062%
condition_era group during day -730 through 0 days relative to index: Neoplasm of pelvis	15	0.19%
condition_era group during day -730 through 0 days relative to index: Chronic urinary tract infection	6	0.074%
observation during day -730 through 0 days relative to index: Body mass index 30+ - obesity	164	2%

covariateName	num	pct
condition_era group during day -730 through 0 days relative to index: Neoplasm of thorax	23	0.28%
condition_era group during day -730 through 0 days relative to index: Neoplasm of digestive organ	38	0.47%
condition_era group during day -730 through 0 days relative to index: Neoplasm of lower respiratory tract	5	0.062%
condition_era group during day -730 through 0 days relative to index: Neoplasm of lung	5	0.062%
condition_era group during day -730 through 0 days relative to index: Neoplasm of male genital organ	15	0.19%
condition_era group during day -730 through 0 days relative to index: Neoplasm of prostate	15	0.19%
condition_era group during day -730 through 0 days relative to index: Neoplasm of gastrointestinal tract	38	0.47%
condition_era group during day -730 through 0 days relative to index: Neoplasm of large intestine	38	0.47%
condition_era group during day -730 through 0 days relative to index: Neoplasm of colon	38	0.47%
condition_era group during day -730 through 0 days relative to index: Neoplasm of intestinal tract	38	0.47%
condition_era group during day -730 through 0 days relative to index: Neoplasm of abdomen	53	0.66%
condition_era group during day -730 through 0 days relative to index: Dementia	31	0.38%
condition_era group during day -730 through 0 days relative to index: Drug overdose	25	0.31%

covariateName	num	pct
condition_era group during day -730 through 0 days relative to index: Normal pregnancy	1195	15%
procedure_occurrence during day -730 through 0 days relative to index: Colonoscopy	999	12%
condition_era group during day -730 through 0 days relative to index: Arthritis	8	0.099%
drug_era group during day -730 through 0 days relative to index: INSULINS AND ANALOGUES	221	2.7%
drug_era group during day -730 through 0 days relative to index: BLOOD GLUCOSE LOWERING DRUGS, EXCL. INSULINS	326	4%
drug_era group during day -730 through 0 days relative to index: ANTITHROMBOTIC AGENTS	549	6.8%
drug_era group during day -730 through 0 days relative to index: BETA BLOCKING AGENTS	28	0.35%
drug_era group during day -730 through 0 days relative to index: CALCIUM CHANNEL BLOCKERS	1923	24%
drug_era group during day -730 through 0 days relative to index: ACE INHIBITORS, PLAIN	1631	20%
drug_era group during day -730 through 0 days relative to index: LIPID MODIFYING AGENTS, PLAIN	1101	14%
drug_era group during day -730 through 0 days relative to index: SEX HORMONES AND MODULATORS OF THE GENITAL SYSTEM	1996	25%
drug_era group during day -730 through 0 days relative to index: DRUGS FOR OBSTRUCTIVE AIRWAY DISEASES	554	6.9%
drug_era group during day -730 through 0 days relative to index: Antibiotics	862	11%
drug_era group during day -730 through 0 days relative to index: OPIOIDS	1078	13%

covariateName	num	pct
drug_era group during day -730 through 0 days relative to index: ANTIDEPRESSANTS	14	0.17%
condition_era group during day -730 through 0 days relative to index: Disorder caused by alcohol	2	0.025%
condition_era group during day -730 through 0 days relative to index: Chronic kidney disease	15	0.19%

B.4.2 KAMILA

Table B.33: KAMILA Continuous Summary for Synthea

rowname	cluster1	cluster2	cluster3	cluster4	cluster5
outpatientVisitCount	3	13	2	5	7
medicationCount	2	4	1	3	5
conditionCount	1	2	1	1	2
procedureCount	1	11	1	2	2
measurementCount	19	30	12	21	31
inpatientVisitCount	1	1	1	1	2
panelWeights	0.574	2.489	0.383	0.957	1.34
clusterSize	2465	1077	1226	2410	899

Table B.34: KAMILA Categorical Summary for Synthea

covariateName	clus	num	tot	pct
age group: 15 - 19	1	113	353	32%

covariateName	clus	num	tot	pct
age group: 15 - 19	2	56	353	16%
age group: 15 - 19	3	76	353	22%
age group: 15 - 19	4	108	353	31%
age group: 20 - 24	1	292	929	31.43%
age group: 20 - 24	2	193	929	20.78%
age group: 20 - 24	3	191	929	20.56%
age group: 20 - 24	4	246	929	26.48%
age group: 20 - 24	5	7	929	0.75%
age group: 25 - 29	1	300	1105	27.1%
age group: 25 - 29	2	315	1105	28.5%
age group: 25 - 29	3	256	1105	23.2%
age group: 25 - 29	4	222	1105	20.1%
age group: 25 - 29	5	12	1105	1.1%
age group: 30 - 34	1	220	789	27.88%
age group: 30 - 34	2	240	789	30.42%
age group: 30 - 34	3	150	789	19.01%
age group: 30 - 34	4	172	789	21.80%
age group: 30 - 34	5	7	789	0.89%
age group: 35 - 39	1	177	563	31.4%
age group: 35 - 39	2	108	563	19.2%
age group: 35 - 39	3	87	563	15.5%
age group: 35 - 39	4	172	563	30.6%
age group: 35 - 39	5	19	563	3.4%
age group: 40 - 44	1	217	603	36.0%
age group: 40 - 44	2	74	603	12.3%
age group: 40 - 44	3	122	603	20.2%

covariateName	clus	num	tot	pct
age group: 40 - 44	4	153	603	25.4%
age group: 40 - 44	5	37	603	6.1%
age group: 45 - 49	1	199	615	32.4%
age group: 45 - 49	2	64	615	10.4%
age group: 45 - 49	3	104	615	16.9%
age group: 45 - 49	4	187	615	30.4%
age group: 45 - 49	5	61	615	9.9%
age group: 50 - 54	1	217	636	34.1%
age group: 50 - 54	2	24	636	3.8%
age group: 50 - 54	3	73	636	11.5%
age group: 50 - 54	4	234	636	36.8%
age group: 50 - 54	5	88	636	13.8%
age group: 55 - 59	1	250	651	38.40%
age group: 55 - 59	2	2	651	0.31%
age group: 55 - 59	3	90	651	13.82%
age group: 55 - 59	4	210	651	32.26%
age group: 55 - 59	5	99	651	15.21%
age group: 60 - 64	1	211	593	35.58%
age group: 60 - 64	2	1	593	0.17%
age group: 60 - 64	3	57	593	9.61%
age group: 60 - 64	4	217	593	36.59%
age group: 60 - 64	5	107	593	18.04%
age group: 65 - 69	1	114	384	29.7%
age group: 65 - 69	3	9	384	2.3%
age group: 65 - 69	4	166	384	43.2%
age group: 65 - 69	5	95	384	24.7%

covariateName	clus	num	tot	pct
age group: 70 - 74	1	60	269	22%
age group: 70 - 74	3	8	269	3%
age group: 70 - 74	4	102	269	38%
age group: 70 - 74	5	99	269	37%
age group: 75 - 79	1	50	223	22%
age group: 75 - 79	4	95	223	43%
age group: 75 - 79	5	78	223	35%
age group: 80 - 84	1	22	136	16%
age group: 80 - 84	4	48	136	35%
age group: 80 - 84	5	66	136	49%
age group: 85 - 89	1	7	61	11.5%
age group: 85 - 89	3	1	61	1.6%
age group: 85 - 89	4	22	61	36.1%
age group: 85 - 89	5	31	61	50.8%
age group: 90 - 94	1	5	44	11%
age group: 90 - 94	4	16	44	36%
age group: 90 - 94	5	23	44	52%
age group: 95 - 99	1	4	37	10.8%
age group: 95 - 99	3	1	37	2.7%
age group: 95 - 99	4	9	37	24.3%
age group: 95 - 99	5	23	37	62.2%
age group: 100 - 104	1	6	56	11%
age group: 100 - 104	4	17	56	30%
age group: 100 - 104	5	33	56	59%
age group: 105 - 109	1	1	30	3.3%
age group: 105 - 109	3	1	30	3.3%

covariateName	clus	num	tot	pct
age group: 105 - 109	4	14	30	46.7%
age group: 105 - 109	5	14	30	46.7%
gender = MALE	1	1469	3807	38.59%
gender = MALE	2	12	3807	0.32%
gender = MALE	3	1062	3807	27.90%
gender = MALE	4	843	3807	22.14%
gender = MALE	5	421	3807	11.06%
race = Asian	1	62	214	29.0%
race = Asian	2	31	214	14.5%
race = Asian	3	40	214	18.7%
race = Asian	4	60	214	28.0%
race = Asian	5	21	214	9.8%
race = Black or African American	1	1238	4071	30%
race = Black or African American	2	529	4071	13%
race = Black or African American	3	629	4071	15%
race = Black or African American	4	1178	4071	29%
race = Black or African American	5	497	4071	12%
race = White	1	1150	3742	31%
race = White	2	506	3742	14%
race = White	3	549	3742	15%
race = White	4	1159	3742	31%
race = White	5	378	3742	10%
gender = FEMALE	1	996	4270	23.3%
gender = FEMALE	2	1065	4270	24.9%
gender = FEMALE	3	164	4270	3.8%
gender = FEMALE	4	1567	4270	36.7%

covariateName	clus	num	tot	pct
gender = FEMALE	5	478	4270	11.2%
condition_era group during day -730 through 0 days relative to index: Osteoporosis	1	2	38	5.3%
condition_era group during day -730 through 0 days relative to index: Osteoporosis	4	17	38	44.7%
condition_era group during day -730 through 0 days relative to index: Osteoporosis	5	19	38	50.0%
condition_era group during day -730 through 0 days relative to index: Neoplasm of breast	2	3	18	16.7%
condition_era group during day -730 through 0 days relative to index: Neoplasm of breast	4	1	18	5.6%
condition_era group during day -730 through 0 days relative to index: Neoplasm of breast	5	14	18	77.8%
condition_era group during day -730 through 0 days relative to index: Sepsis	2	6	12	50%
condition_era group during day -730 through 0 days relative to index: Sepsis	4	4	12	33%
condition_era group during day -730 through 0 days relative to index: Sepsis	5	2	12	17%
condition_era group during day -730 through 0 days relative to index: Neoplasm of digestive tract	1	2	38	5.3%
condition_era group during day -730 through 0 days relative to index: Neoplasm of digestive tract	4	5	38	13.2%
condition_era group during day -730 through 0 days relative to index: Neoplasm of digestive tract	5	31	38	81.6%

covariateName	clus	num	tot	pct
condition_era group during day -730 through 0 days relative to index: Diabetes mellitus	1	4	37	11%
condition_era group during day -730 through 0 days relative to index: Diabetes mellitus	2	10	37	27%
condition_era group during day -730 through 0 days relative to index: Diabetes mellitus	4	10	37	27%
condition_era group during day -730 through 0 days relative to index: Diabetes mellitus	5	13	37	35%
condition_era group during day -730 through 0 days relative to index: Neoplasm of respiratory tract	5	5	5	100%
condition_era group during day -730 through 0 days relative to index: Chronic obstructive lung disease	1	3	10	30%
condition_era group during day -730 through 0 days relative to index: Chronic obstructive lung disease	2	1	10	10%
condition_era group during day -730 through 0 days relative to index: Chronic obstructive lung disease	4	4	10	40%
condition_era group during day -730 through 0 days relative to index: Chronic obstructive lung disease	5	2	10	20%
condition_era group during day -730 through 0 days relative to index: Chronic sinusitis	1	15	84	17.9%
condition_era group during day -730 through 0 days relative to index: Chronic sinusitis	2	17	84	20.2%
condition_era group during day -730 through 0 days relative to index: Chronic sinusitis	3	2	84	2.4%
condition_era group during day -730 through 0 days relative to index: Chronic sinusitis	4	42	84	50.0%

covariateName	clus	num	tot	pct
condition_era group during day -730 through 0 days relative to index: Chronic sinusitis	5	8	84	9.5%
condition_era group during day -730 through 0 days relative to index: Heart failure	2	3	28	10.7%
condition_era group during day -730 through 0 days relative to index: Heart failure	4	1	28	3.6%
condition_era group during day -730 through 0 days relative to index: Heart failure	5	24	28	85.7%
condition_era group during day -730 through 0 days relative to index: Hypertensive disorder	1	29	189	15.3%
condition_era group during day -730 through 0 days relative to index: Hypertensive disorder	2	100	189	52.9%
condition_era group during day -730 through 0 days relative to index: Hypertensive disorder	4	56	189	29.6%
condition_era group during day -730 through 0 days relative to index: Hypertensive disorder	5	4	189	2.1%
condition_era group during day -730 through 0 days relative to index: Asthma	1	7	43	16%
condition_era group during day -730 through 0 days relative to index: Asthma	2	18	43	42%
condition_era group during day -730 through 0 days relative to index: Asthma	4	18	43	42%
condition_era group during day -730 through 0 days relative to index: Migraine	1	5	12	41.7%
condition_era group during day -730 through 0 days relative to index: Migraine	4	6	12	50.0%

covariateName	clus	num	tot	pct
condition_era group during day -730 through 0 days relative to index: Migraine	5	1	12	8.3%
condition_era group during day -730 through 0 days relative to index: Heart disease	1	4	127	3.1%
condition_era group during day -730 through 0 days relative to index: Heart disease	2	9	127	7.1%
condition_era group during day -730 through 0 days relative to index: Heart disease	4	27	127	21.3%
condition_era group during day -730 through 0 days relative to index: Heart disease	5	87	127	68.5%
condition_era group during day -730 through 0 days relative to index: Cerebrovascular disease	1	1	33	3.0%
condition_era group during day -730 through 0 days relative to index: Cerebrovascular disease	2	3	33	9.1%
condition_era group during day -730 through 0 days relative to index: Cerebrovascular disease	4	6	33	18.2%
condition_era group during day -730 through 0 days relative to index: Cerebrovascular disease	5	23	33	69.7%
condition_era group during day -730 through 0 days relative to index: Complication of pregnancy, childbirth and/or the puerperium	2	184	199	92.5%
condition_era group during day -730 through 0 days relative to index: Complication of pregnancy, childbirth and/or the puerperium	4	14	199	7.0%
condition_era group during day -730 through 0 days relative to index: Complication of pregnancy, childbirth and/or the puerperium	5	1	199	0.5%
condition_era group during day -730 through 0 days relative to index: Chronic pain	1	18	77	23.4%

covariateName	clus	num	tot	pct
condition_era group during day -730 through 0 days relative to index: Chronic pain	2	18	77	23.4%
condition_era group during day -730 through 0 days relative to index: Chronic pain	4	38	77	49.4%
condition_era group during day -730 through 0 days relative to index: Chronic pain	5	3	77	3.9%
condition_era group during day -730 through 0 days relative to index: Anemia	1	31	297	10.4%
condition_era group during day -730 through 0 days relative to index: Anemia	2	153	297	51.5%
condition_era group during day -730 through 0 days relative to index: Anemia	4	88	297	29.6%
condition_era group during day -730 through 0 days relative to index: Anemia	5	25	297	8.4%
condition_era group during day -730 through 0 days relative to index: Anxiety	1	1	3	33%
condition_era group during day -730 through 0 days relative to index: Anxiety	2	1	3	33%
condition_era group during day -730 through 0 days relative to index: Anxiety	5	1	3	33%
condition_era group during day -730 through 0 days relative to index: Concussion injury of brain	1	39	173	22.5%
condition_era group during day -730 through 0 days relative to index: Concussion injury of brain	2	27	173	15.6%
condition_era group during day -730 through 0 days relative to index: Concussion injury of brain	3	14	173	8.1%

covariateName	clus	num	tot	pct
condition_era group during day -730 through 0 days relative to index: Concussion injury of brain	4	71	173	41.0%
condition_era group during day -730 through 0 days relative to index: Concussion injury of brain	5	22	173	12.7%
condition_era group during day -730 through 0 days relative to index: Neoplasm of digestive system	1	2	38	5.3%
condition_era group during day -730 through 0 days relative to index: Neoplasm of digestive system	4	5	38	13.2%
condition_era group during day -730 through 0 days relative to index: Neoplasm of digestive system	5	31	38	81.6%
condition_era group during day -730 through 0 days relative to index: Neoplasm of trunk	1	2	75	2.7%
condition_era group during day -730 through 0 days relative to index: Neoplasm of trunk	2	3	75	4.0%
condition_era group during day -730 through 0 days relative to index: Neoplasm of trunk	4	11	75	14.7%
condition_era group during day -730 through 0 days relative to index: Neoplasm of trunk	5	59	75	78.7%
condition_era group during day -730 through 0 days relative to index: Neoplasm of intra-abdominal organs	1	2	53	3.8%
condition_era group during day -730 through 0 days relative to index: Neoplasm of intra-abdominal organs	4	10	53	18.9%
condition_era group during day -730 through 0 days relative to index: Neoplasm of intra-abdominal organs	5	41	53	77.4%
condition_era group during day -730 through 0 days relative to index: Neoplasm of intrathoracic organs	5	5	5	100%

covariateName	clus	num	tot	pct
condition_era group during day -730 through 0 days relative to index: Neoplasm of pelvis	4	5	15	33%
condition_era group during day -730 through 0 days relative to index: Neoplasm of pelvis	5	10	15	67%
condition_era group during day -730 through 0 days relative to index: Chronic urinary tract infection	2	2	6	33%
condition_era group during day -730 through 0 days relative to index: Chronic urinary tract infection	4	4	6	67%
observation during day -730 through 0 days relative to index: Body mass index 30+ - obesity	1	51	164	31.10%
observation during day -730 through 0 days relative to index: Body mass index 30+ - obesity	2	38	164	23.17%
observation during day -730 through 0 days relative to index: Body mass index 30+ - obesity	3	32	164	19.51%
observation during day -730 through 0 days relative to index: Body mass index 30+ - obesity	4	42	164	25.61%
observation during day -730 through 0 days relative to index: Body mass index 30+ - obesity	5	1	164	0.61%
condition_era group during day -730 through 0 days relative to index: Neoplasm of thorax	2	3	23	13.0%
condition_era group during day -730 through 0 days relative to index: Neoplasm of thorax	4	1	23	4.3%
condition_era group during day -730 through 0 days relative to index: Neoplasm of thorax	5	19	23	82.6%
condition_era group during day -730 through 0 days relative to index: Neoplasm of digestive organ	1	2	38	5.3%

covariateName	clus	num	tot	pct
condition_era group during day -730 through 0 days relative to index: Neoplasm of digestive organ	4	5	38	13.2%
condition_era group during day -730 through 0 days relative to index: Neoplasm of digestive organ	5	31	38	81.6%
condition_era group during day -730 through 0 days relative to index: Neoplasm of lower respiratory tract	5	5	5	100%
condition_era group during day -730 through 0 days relative to index: Neoplasm of lung	5	5	5	100%
condition_era group during day -730 through 0 days relative to index: Neoplasm of male genital organ	4	5	15	33%
condition_era group during day -730 through 0 days relative to index: Neoplasm of male genital organ	5	10	15	67%
condition_era group during day -730 through 0 days relative to index: Neoplasm of prostate	4	5	15	33%
condition_era group during day -730 through 0 days relative to index: Neoplasm of prostate	5	10	15	67%
condition_era group during day -730 through 0 days relative to index: Neoplasm of gastrointestinal tract	1	2	38	5.3%
condition_era group during day -730 through 0 days relative to index: Neoplasm of gastrointestinal tract	4	5	38	13.2%
condition_era group during day -730 through 0 days relative to index: Neoplasm of gastrointestinal tract	5	31	38	81.6%
condition_era group during day -730 through 0 days relative to index: Neoplasm of large intestine	1	2	38	5.3%
condition_era group during day -730 through 0 days relative to index: Neoplasm of large intestine	4	5	38	13.2%

covariateName	clus	num	tot	pct
condition_era group during day -730 through 0 days relative to index: Neoplasm of large intestine	5	31	38	81.6%
condition_era group during day -730 through 0 days relative to index: Neoplasm of colon	1	2	38	5.3%
condition_era group during day -730 through 0 days relative to index: Neoplasm of colon	4	5	38	13.2%
condition_era group during day -730 through 0 days relative to index: Neoplasm of colon	5	31	38	81.6%
condition_era group during day -730 through 0 days relative to index: Neoplasm of intestinal tract	1	2	38	5.3%
condition_era group during day -730 through 0 days relative to index: Neoplasm of intestinal tract	4	5	38	13.2%
condition_era group during day -730 through 0 days relative to index: Neoplasm of intestinal tract	5	31	38	81.6%
condition_era group during day -730 through 0 days relative to index: Neoplasm of abdomen	1	2	53	3.8%
condition_era group during day -730 through 0 days relative to index: Neoplasm of abdomen	4	10	53	18.9%
condition_era group during day -730 through 0 days relative to index: Neoplasm of abdomen	5	41	53	77.4%
condition_era group during day -730 through 0 days relative to index: Dementia	1	1	31	3.2%
condition_era group during day -730 through 0 days relative to index: Dementia	4	7	31	22.6%
condition_era group during day -730 through 0 days relative to index: Dementia	5	23	31	74.2%

covariateName	clus	num	tot	pct
condition_era group during day -730 through 0 days relative to index: Drug overdose	1	1	25	4%
condition_era group during day -730 through 0 days relative to index: Drug overdose	2	4	25	16%
condition_era group during day -730 through 0 days relative to index: Drug overdose	4	17	25	68%
condition_era group during day -730 through 0 days relative to index: Drug overdose	5	3	25	12%
condition_era group during day -730 through 0 days relative to index: Normal pregnancy	1	15	1195	1.26%
condition_era group during day -730 through 0 days relative to index: Normal pregnancy	2	992	1195	83.01%
condition_era group during day -730 through 0 days relative to index: Normal pregnancy	3	2	1195	0.17%
condition_era group during day -730 through 0 days relative to index: Normal pregnancy	4	183	1195	15.31%
condition_era group during day -730 through 0 days relative to index: Normal pregnancy	5	3	1195	0.25%
procedure_occurrence during day -730 through 0 days relative to index: Colonoscopy	1	286	999	28.6%
procedure_occurrence during day -730 through 0 days relative to index: Colonoscopy	2	18	999	1.8%
procedure_occurrence during day -730 through 0 days relative to index: Colonoscopy	3	44	999	4.4%
procedure_occurrence during day -730 through 0 days relative to index: Colonoscopy	4	391	999	39.1%

covariateName	clus	num	tot	pct
procedure_occurrence during day -730 through 0 days relative to index: Colonoscopy	5	260	999	26.0%
condition_era group during day -730 through 0 days relative to index: Arthritis	1	1	8	12%
condition_era group during day -730 through 0 days relative to index: Arthritis	2	2	8	25%
condition_era group during day -730 through 0 days relative to index: Arthritis	4	5	8	62%
drug_era group during day -730 through 0 days relative to index: INSULINS AND ANALOGUES	4	7	221	3.2%
drug_era group during day -730 through 0 days relative to index: INSULINS AND ANALOGUES	5	214	221	96.8%
drug_era group during day -730 through 0 days relative to index: BLOOD GLUCOSE LOWERING DRUGS, EXCL. INSULINS	1	1	326	0.31%
drug_era group during day -730 through 0 days relative to index: BLOOD GLUCOSE LOWERING DRUGS, EXCL. INSULINS	2	12	326	3.68%
drug_era group during day -730 through 0 days relative to index: BLOOD GLUCOSE LOWERING DRUGS, EXCL. INSULINS	4	55	326	16.87%
drug_era group during day -730 through 0 days relative to index: BLOOD GLUCOSE LOWERING DRUGS, EXCL. INSULINS	5	258	326	79.14%
drug_era group during day -730 through 0 days relative to index: ANTITHROMBOTIC AGENTS	1	56	549	10.20%
drug_era group during day -730 through 0 days relative to index: ANTITHROMBOTIC AGENTS	2	22	549	4.01%
drug_era group during day -730 through 0 days relative to index: ANTITHROMBOTIC AGENTS	3	2	549	0.36%

covariateName	clus	num	tot	pct
drug_era group during day -730 through 0 days relative to index: ANTITHROMBOTIC AGENTS	4	177	549	32.24%
drug_era group during day -730 through 0 days relative to index: ANTITHROMBOTIC AGENTS	5	292	549	53.19%
drug_era group during day -730 through 0 days relative to index: BETA BLOCKING AGENTS	2	3	28	10.7%
drug_era group during day -730 through 0 days relative to index: BETA BLOCKING AGENTS	4	1	28	3.6%
drug_era group during day -730 through 0 days relative to index: BETA BLOCKING AGENTS	5	24	28	85.7%
drug_era group during day -730 through 0 days relative to index: CALCIUM CHANNEL BLOCKERS	1	453	1923	23.6%
drug_era group during day -730 through 0 days relative to index: CALCIUM CHANNEL BLOCKERS	2	221	1923	11.5%
drug_era group during day -730 through 0 days relative to index: CALCIUM CHANNEL BLOCKERS	3	34	1923	1.8%
drug_era group during day -730 through 0 days relative to index: CALCIUM CHANNEL BLOCKERS	4	730	1923	38.0%
drug_era group during day -730 through 0 days relative to index: CALCIUM CHANNEL BLOCKERS	5	485	1923	25.2%
drug_era group during day -730 through 0 days relative to index: ACE INHIBITORS, PLAIN	1	445	1631	27.3%
drug_era group during day -730 through 0 days relative to index: ACE INHIBITORS, PLAIN	2	212	1631	13.0%
drug_era group during day -730 through 0 days relative to index: ACE INHIBITORS, PLAIN	3	45	1631	2.8%

covariateName	clus	num	tot	pct
drug_era group during day -730 through 0 days relative to index: ACE INHIBITORS, PLAIN	4	611	1631	37.5%
drug_era group during day -730 through 0 days relative to index: ACE INHIBITORS, PLAIN	5	318	1631	19.5%
drug_era group during day -730 through 0 days relative to index: LIPID MODIFYING AGENTS, PLAIN	1	231	1101	20.98%
drug_era group during day -730 through 0 days relative to index: LIPID MODIFYING AGENTS, PLAIN	2	24	1101	2.18%
drug_era group during day -730 through 0 days relative to index: LIPID MODIFYING AGENTS, PLAIN	3	7	1101	0.64%
drug_era group during day -730 through 0 days relative to index: LIPID MODIFYING AGENTS, PLAIN	4	464	1101	42.14%
drug_era group during day -730 through 0 days relative to index: LIPID MODIFYING AGENTS, PLAIN	5	375	1101	34.06%
drug_era group during day -730 through 0 days relative to index: SEX HORMONES AND MODULATORS OF THE GENITAL SYSTEM	1	350	1996	17.5%
drug_era group during day -730 through 0 days relative to index: SEX HORMONES AND MODULATORS OF THE GENITAL SYSTEM	2	793	1996	39.7%
drug_era group during day -730 through 0 days relative to index: SEX HORMONES AND MODULATORS OF THE GENITAL SYSTEM	3	39	1996	2.0%
drug_era group during day -730 through 0 days relative to index: SEX HORMONES AND MODULATORS OF THE GENITAL SYSTEM	4	757	1996	37.9%
drug_era group during day -730 through 0 days relative to index: SEX HORMONES AND MODULATORS OF THE GENITAL SYSTEM	5	57	1996	2.9%
drug_era group during day -730 through 0 days relative to index: DRUGS FOR OBSTRUCTIVE AIRWAY DISEASES	1	148	554	26.7%

covariateName	clus	num	tot	pct
drug_era group during day -730 through 0 days relative to index: DRUGS FOR OBSTRUCTIVE AIRWAY DISEASES	2	94	554	17.0%
drug_era group during day -730 through 0 days relative to index: DRUGS FOR OBSTRUCTIVE AIRWAY DISEASES	3	24	554	4.3%
drug_era group during day -730 through 0 days relative to index: DRUGS FOR OBSTRUCTIVE AIRWAY DISEASES	4	213	554	38.4%
drug_era group during day -730 through 0 days relative to index: DRUGS FOR OBSTRUCTIVE AIRWAY DISEASES	5	75	554	13.5%
drug_era group during day -730 through 0 days relative to index: Antibiotics	1	222	862	25.8%
drug_era group during day -730 through 0 days relative to index: Antibiotics	2	147	862	17.1%
drug_era group during day -730 through 0 days relative to index: Antibiotics	3	29	862	3.4%
drug_era group during day -730 through 0 days relative to index: Antibiotics	4	341	862	39.6%
drug_era group during day -730 through 0 days relative to index: Antibiotics	5	123	862	14.3%
drug_era group during day -730 through 0 days relative to index: OPIOIDS	1	300	1078	27.8%
drug_era group during day -730 through 0 days relative to index: OPIOIDS	2	155	1078	14.4%
drug_era group during day -730 through 0 days relative to index: OPIOIDS	3	52	1078	4.8%
drug_era group during day -730 through 0 days relative to index: OPIOIDS	4	401	1078	37.2%

covariateName	clus	num	tot	pct
drug_era group during day -730 through 0 days relative to index: OPIOIDS	5	170	1078	15.8%
drug_era group during day -730 through 0 days relative to index: ANTIDEPRESSANTS	1	4	14	28.6%
drug_era group during day -730 through 0 days relative to index: ANTIDEPRESSANTS	2	1	14	7.1%
drug_era group during day -730 through 0 days relative to index: ANTIDEPRESSANTS	4	7	14	50.0%
drug_era group during day -730 through 0 days relative to index: ANTIDEPRESSANTS	5	2	14	14.3%
condition_era group during day -730 through 0 days relative to index: Disorder caused by alcohol	1	1	2	50%
condition_era group during day -730 through 0 days relative to index: Disorder caused by alcohol	4	1	2	50%
condition_era group during day -730 through 0 days relative to index: Chronic kidney disease	2	1	15	6.7%
condition_era group during day -730 through 0 days relative to index: Chronic kidney disease	5	14	15	93.3%

Table B.35: KAMILA PSE Table for Synthea

assigned_pcp	patientCount	pse	z
Arnulfo Carter	531	560.72	-0.22
Denny Yundt	488	418.68	0.62
Florinda Gutmann	624	722.49	-0.57

assigned_pcp	patientCount	pse	z
Halley Hamill	434	284.67	1.77
Harlan Beer	629	751.78	-0.69
Houston Beier	401	266.48	1.73
Jeri Hansen	700	832.19	-0.67
Kathie Zulauf	503	381.54	1.13
Marilou Conn	622	760.59	-0.78
Modesto Trantow	426	271.65	1.90
Paulene Schultz	642	746.42	-0.59
Sharyl Hilpert	667	815.34	-0.78
Shonta Sanford	631	733.98	-0.59
Taylor Wilderman	394	290.03	1.29
Wilber Harris	385	240.45	2.01

B.4.3 Gaussian Multinomial Mixture Model

Table B.36: GMMM Continuous Summary for Synthea

rowname	cluster1	cluster2	cluster3	cluster4	cluster5
outpatientVisitCount	3	13	2	6	5
medicationCount	2	4	1	5	3
conditionCount	1	2	1	2	1
procedureCount	1	11	1	2	2
measurementCount	19	30	12	30	21
inpatientVisitCount	1	1	1	1	1
panelWeights	0.58	2.514	0.387	1.16	0.967

rowname	cluster1	cluster2	cluster3	cluster4	cluster5
clusterSize	2168	1006	1152	1127	2624

Table B.37: GMMM Categorical Summary for Synthea

covariateName	clus	num	tot	pct
age group: 15 - 19	1	94	353	26.63%
age group: 15 - 19	2	59	353	16.71%
age group: 15 - 19	3	73	353	20.68%
age group: 15 - 19	4	2	353	0.57%
age group: 15 - 19	5	125	353	35.41%
age group: 20 - 24	1	269	929	29.0%
age group: 20 - 24	2	186	929	20.0%
age group: 20 - 24	3	176	929	18.9%
age group: 20 - 24	4	11	929	1.2%
age group: 20 - 24	5	287	929	30.9%
age group: 25 - 29	1	266	1105	24.1%
age group: 25 - 29	2	297	1105	26.9%
age group: 25 - 29	3	245	1105	22.2%
age group: 25 - 29	4	30	1105	2.7%
age group: 25 - 29	5	267	1105	24.2%
age group: 30 - 34	1	206	789	26.1%
age group: 30 - 34	2	222	789	28.1%
age group: 30 - 34	3	138	789	17.5%
age group: 30 - 34	4	29	789	3.7%
age group: 30 - 34	5	194	789	24.6%

covariateName	clus	num	tot	pct
age group: 35 - 39	1	159	563	28.2%
age group: 35 - 39	2	100	563	17.8%
age group: 35 - 39	3	84	563	14.9%
age group: 35 - 39	4	33	563	5.9%
age group: 35 - 39	5	187	563	33.2%
age group: 40 - 44	1	191	603	31.7%
age group: 40 - 44	2	66	603	10.9%
age group: 40 - 44	3	117	603	19.4%
age group: 40 - 44	4	59	603	9.8%
age group: 40 - 44	5	170	603	28.2%
age group: 45 - 49	1	181	615	29.4%
age group: 45 - 49	2	56	615	9.1%
age group: 45 - 49	3	98	615	15.9%
age group: 45 - 49	4	76	615	12.4%
age group: 45 - 49	5	204	615	33.2%
age group: 50 - 54	1	192	636	30.2%
age group: 50 - 54	2	20	636	3.1%
age group: 50 - 54	3	70	636	11.0%
age group: 50 - 54	4	101	636	15.9%
age group: 50 - 54	5	253	636	39.8%
age group: 55 - 59	1	209	651	32%
age group: 55 - 59	3	82	651	13%
age group: 55 - 59	4	119	651	18%
age group: 55 - 59	5	241	651	37%
age group: 60 - 64	1	185	593	31.2%
age group: 60 - 64	3	51	593	8.6%

covariateName	clus	num	tot	pct
age group: 60 - 64	4	141	593	23.8%
age group: 60 - 64	5	216	593	36.4%
age group: 65 - 69	1	91	384	23.7%
age group: 65 - 69	3	8	384	2.1%
age group: 65 - 69	4	111	384	28.9%
age group: 65 - 69	5	174	384	45.3%
age group: 70 - 74	1	50	269	19%
age group: 70 - 74	3	8	269	3%
age group: 70 - 74	4	110	269	41%
age group: 70 - 74	5	101	269	38%
age group: 75 - 79	1	38	223	17%
age group: 75 - 79	4	95	223	43%
age group: 75 - 79	5	90	223	40%
age group: 80 - 84	1	17	136	12%
age group: 80 - 84	4	74	136	54%
age group: 80 - 84	5	45	136	33%
age group: 85 - 89	1	6	61	9.8%
age group: 85 - 89	3	1	61	1.6%
age group: 85 - 89	4	34	61	55.7%
age group: 85 - 89	5	20	61	32.8%
age group: 90 - 94	1	4	44	9.1%
age group: 90 - 94	4	24	44	54.5%
age group: 90 - 94	5	16	44	36.4%
age group: 95 - 99	1	3	37	8.1%
age group: 95 - 99	4	25	37	67.6%
age group: 95 - 99	5	9	37	24.3%

covariateName	clus	num	tot	pct
age group: 100 - 104	1	6	56	11%
age group: 100 - 104	4	36	56	64%
age group: 100 - 104	5	14	56	25%
age group: 105 - 109	1	1	30	3.3%
age group: 105 - 109	3	1	30	3.3%
age group: 105 - 109	4	17	30	56.7%
age group: 105 - 109	5	11	30	36.7%
gender = MALE	1	1324	3807	34.78%
gender = MALE	2	9	3807	0.24%
gender = MALE	3	996	3807	26.16%
gender = MALE	4	499	3807	13.11%
gender = MALE	5	979	3807	25.72%
race = Asian	1	57	214	27%
race = Asian	2	26	214	12%
race = Asian	3	39	214	18%
race = Asian	4	30	214	14%
race = Asian	5	62	214	29%
race = Black or African American	1	1095	4071	27%
race = Black or African American	2	491	4071	12%
race = Black or African American	3	584	4071	14%
race = Black or African American	4	605	4071	15%
race = Black or African American	5	1296	4071	32%
race = White	1	1001	3742	27%
race = White	2	478	3742	13%
race = White	3	522	3742	14%
race = White	4	487	3742	13%

covariateName	clus	num	tot	pct
race = White	5	1254	3742	34%
gender = FEMALE	1	844	4270	19.8%
gender = FEMALE	2	997	4270	23.3%
gender = FEMALE	3	156	4270	3.7%
gender = FEMALE	4	628	4270	14.7%
gender = FEMALE	5	1645	4270	38.5%
condition_era group during day -730 through 0 days relative to index: Osteoporosis	1	2	38	5.3%
condition_era group during day -730 through 0 days relative to index: Osteoporosis	4	21	38	55.3%
condition_era group during day -730 through 0 days relative to index: Osteoporosis	5	15	38	39.5%
condition_era group during day -730 through 0 days relative to index: Neoplasm of breast	4	18	18	100%
condition_era group during day -730 through 0 days relative to index: Sepsis	2	5	12	42%
condition_era group during day -730 through 0 days relative to index: Sepsis	4	4	12	33%
condition_era group during day -730 through 0 days relative to index: Sepsis	5	3	12	25%
condition_era group during day -730 through 0 days relative to index: Neoplasm of digestive tract	1	1	38	2.6%
condition_era group during day -730 through 0 days relative to index: Neoplasm of digestive tract	4	37	38	97.4%
condition_era group during day -730 through 0 days relative to index: Diabetes mellitus	4	26	37	70%

covariateName	clus	num	tot	pct
condition_era group during day -730 through 0 days relative to index: Diabetes mellitus	5	11	37	30%
condition_era group during day -730 through 0 days relative to index: Neoplasm of respiratory tract	4	5	5	100%
condition_era group during day -730 through 0 days relative to index: Chronic obstructive lung disease	1	3	10	30%
condition_era group during day -730 through 0 days relative to index: Chronic obstructive lung disease	2	1	10	10%
condition_era group during day -730 through 0 days relative to index: Chronic obstructive lung disease	4	2	10	20%
condition_era group during day -730 through 0 days relative to index: Chronic obstructive lung disease	5	4	10	40%
condition_era group during day -730 through 0 days relative to index: Chronic sinusitis	1	11	84	13.1%
condition_era group during day -730 through 0 days relative to index: Chronic sinusitis	2	19	84	22.6%
condition_era group during day -730 through 0 days relative to index: Chronic sinusitis	3	2	84	2.4%
condition_era group during day -730 through 0 days relative to index: Chronic sinusitis	4	13	84	15.5%
condition_era group during day -730 through 0 days relative to index: Chronic sinusitis	5	39	84	46.4%
condition_era group during day -730 through 0 days relative to index: Heart failure	4	28	28	100%
condition_era group during day -730 through 0 days relative to index: Hypertensive disorder	1	24	189	12.7%

covariateName	clus	num	tot	pct
condition_era group during day -730 through 0 days relative to index: Hypertensive disorder	2	99	189	52.4%
condition_era group during day -730 through 0 days relative to index: Hypertensive disorder	4	9	189	4.8%
condition_era group during day -730 through 0 days relative to index: Hypertensive disorder	5	57	189	30.2%
condition_era group during day -730 through 0 days relative to index: Asthma	1	7	43	16%
condition_era group during day -730 through 0 days relative to index: Asthma	2	18	43	42%
condition_era group during day -730 through 0 days relative to index: Asthma	5	18	43	42%
condition_era group during day -730 through 0 days relative to index: Migraine	1	1	12	8.3%
condition_era group during day -730 through 0 days relative to index: Migraine	4	1	12	8.3%
condition_era group during day -730 through 0 days relative to index: Migraine	5	10	12	83.3%
condition_era group during day -730 through 0 days relative to index: Heart disease	1	2	127	1.6%
condition_era group during day -730 through 0 days relative to index: Heart disease	4	103	127	81.1%
condition_era group during day -730 through 0 days relative to index: Heart disease	5	22	127	17.3%
condition_era group during day -730 through 0 days relative to index: Cerebrovascular disease	2	3	33	9.1%

covariateName	clus	num	tot	pct
condition_era group during day -730 through 0 days relative to index: Cerebrovascular disease	4	23	33	69.7%
condition_era group during day -730 through 0 days relative to index: Cerebrovascular disease	5	7	33	21.2%
condition_era group during day -730 through 0 days relative to index: Complication of pregnancy, childbirth and/or the puerperium	2	174	199	87.4%
condition_era group during day -730 through 0 days relative to index: Complication of pregnancy, childbirth and/or the puerperium	4	12	199	6.0%
condition_era group during day -730 through 0 days relative to index: Complication of pregnancy, childbirth and/or the puerperium	5	13	199	6.5%
condition_era group during day -730 through 0 days relative to index: Chronic pain	1	8	77	10.4%
condition_era group during day -730 through 0 days relative to index: Chronic pain	2	21	77	27.3%
condition_era group during day -730 through 0 days relative to index: Chronic pain	4	4	77	5.2%
condition_era group during day -730 through 0 days relative to index: Chronic pain	5	44	77	57.1%
condition_era group during day -730 through 0 days relative to index: Anemia	1	23	297	7.7%
condition_era group during day -730 through 0 days relative to index: Anemia	2	142	297	47.8%
condition_era group during day -730 through 0 days relative to index: Anemia	4	39	297	13.1%
condition_era group during day -730 through 0 days relative to index: Anemia	5	93	297	31.3%

covariateName	clus	num	tot	pct
condition_era group during day -730 through 0 days relative to index: Anxiety	2	1	3	33%
condition_era group during day -730 through 0 days relative to index: Anxiety	4	1	3	33%
condition_era group during day -730 through 0 days relative to index: Anxiety	5	1	3	33%
condition_era group during day -730 through 0 days relative to index: Concussion injury of brain	1	37	173	21.4%
condition_era group during day -730 through 0 days relative to index: Concussion injury of brain	2	29	173	16.8%
condition_era group during day -730 through 0 days relative to index: Concussion injury of brain	3	13	173	7.5%
condition_era group during day -730 through 0 days relative to index: Concussion injury of brain	4	25	173	14.5%
condition_era group during day -730 through 0 days relative to index: Concussion injury of brain	5	69	173	39.9%
condition_era group during day -730 through 0 days relative to index: Neoplasm of digestive system	1	1	38	2.6%
condition_era group during day -730 through 0 days relative to index: Neoplasm of digestive system	4	37	38	97.4%
condition_era group during day -730 through 0 days relative to index: Neoplasm of trunk	1	1	75	1.3%
condition_era group during day -730 through 0 days relative to index: Neoplasm of trunk	4	74	75	98.7%
condition_era group during day -730 through 0 days relative to index: Neoplasm of intra-abdominal organs	1	1	53	1.9%

covariateName	clus	num	tot	pct
condition_era group during day -730 through 0 days relative to index: Neoplasm of intra-abdominal organs	4	52	53	98.1%
condition_era group during day -730 through 0 days relative to index: Neoplasm of intrathoracic organs	4	5	5	100%
condition_era group during day -730 through 0 days relative to index: Neoplasm of pelvis	4	15	15	100%
condition_era group during day -730 through 0 days relative to index: Chronic urinary tract infection	2	2	6	33%
condition_era group during day -730 through 0 days relative to index: Chronic urinary tract infection	5	4	6	67%
observation during day -730 through 0 days relative to index: Body mass index 30+ - obesity	1	47	164	28.7%
observation during day -730 through 0 days relative to index: Body mass index 30+ - obesity	2	40	164	24.4%
observation during day -730 through 0 days relative to index: Body mass index 30+ - obesity	3	29	164	17.7%
observation during day -730 through 0 days relative to index: Body mass index 30+ - obesity	4	2	164	1.2%
observation during day -730 through 0 days relative to index: Body mass index 30+ - obesity	5	46	164	28.0%
condition_era group during day -730 through 0 days relative to index: Neoplasm of thorax	4	23	23	100%
condition_era group during day -730 through 0 days relative to index: Neoplasm of digestive organ	1	1	38	2.6%
condition_era group during day -730 through 0 days relative to index: Neoplasm of digestive organ	4	37	38	97.4%

covariateName	clus	num	tot	pct
condition_era group during day -730 through 0 days relative to index: Neoplasm of lower respiratory tract	4	5	5	100%
condition_era group during day -730 through 0 days relative to index: Neoplasm of lung	4	5	5	100%
condition_era group during day -730 through 0 days relative to index: Neoplasm of male genital organ	4	15	15	100%
condition_era group during day -730 through 0 days relative to index: Neoplasm of prostate	4	15	15	100%
condition_era group during day -730 through 0 days relative to index: Neoplasm of gastrointestinal tract	1	1	38	2.6%
condition_era group during day -730 through 0 days relative to index: Neoplasm of gastrointestinal tract	4	37	38	97.4%
condition_era group during day -730 through 0 days relative to index: Neoplasm of large intestine	1	1	38	2.6%
condition_era group during day -730 through 0 days relative to index: Neoplasm of large intestine	4	37	38	97.4%
condition_era group during day -730 through 0 days relative to index: Neoplasm of colon	1	1	38	2.6%
condition_era group during day -730 through 0 days relative to index: Neoplasm of colon	4	37	38	97.4%
condition_era group during day -730 through 0 days relative to index: Neoplasm of intestinal tract	1	1	38	2.6%
condition_era group during day -730 through 0 days relative to index: Neoplasm of intestinal tract	4	37	38	97.4%
condition_era group during day -730 through 0 days relative to index: Neoplasm of abdomen	1	1	53	1.9%

covariateName	clus	num	tot	pct
condition_era group during day -730 through 0 days relative to index: Neoplasm of abdomen	4	52	53	98.1%
condition_era group during day -730 through 0 days relative to index: Dementia	1	1	31	3.2%
condition_era group during day -730 through 0 days relative to index: Dementia	4	25	31	80.6%
condition_era group during day -730 through 0 days relative to index: Dementia	5	5	31	16.1%
condition_era group during day -730 through 0 days relative to index: Drug overdose	2	5	25	20%
condition_era group during day -730 through 0 days relative to index: Drug overdose	4	4	25	16%
condition_era group during day -730 through 0 days relative to index: Drug overdose	5	16	25	64%
condition_era group during day -730 through 0 days relative to index: Normal pregnancy	1	11	1195	0.92%
condition_era group during day -730 through 0 days relative to index: Normal pregnancy	2	917	1195	76.74%
condition_era group during day -730 through 0 days relative to index: Normal pregnancy	3	2	1195	0.17%
condition_era group during day -730 through 0 days relative to index: Normal pregnancy	4	51	1195	4.27%
condition_era group during day -730 through 0 days relative to index: Normal pregnancy	5	214	1195	17.91%
procedure_occurrence during day -730 through 0 days relative to index: Colonoscopy	1	233	999	23.3%

covariateName	clus	num	tot	pct
procedure_occurrence during day -730 through 0 days relative to index: Colonoscopy	2	13	999	1.3%
procedure_occurrence during day -730 through 0 days relative to index: Colonoscopy	3	40	999	4.0%
procedure_occurrence during day -730 through 0 days relative to index: Colonoscopy	4	304	999	30.4%
procedure_occurrence during day -730 through 0 days relative to index: Colonoscopy	5	409	999	40.9%
condition_era group during day -730 through 0 days relative to index: Arthritis	2	2	8	25%
condition_era group during day -730 through 0 days relative to index: Arthritis	5	6	8	75%
drug_era group during day -730 through 0 days relative to index: INSULINS AND ANALOGUES	4	220	221	99.55%
drug_era group during day -730 through 0 days relative to index: INSULINS AND ANALOGUES	5	1	221	0.45%
drug_era group during day -730 through 0 days relative to index: BLOOD GLUCOSE LOWERING DRUGS, EXCL. INSULINS	4	289	326	89%
drug_era group during day -730 through 0 days relative to index: BLOOD GLUCOSE LOWERING DRUGS, EXCL. INSULINS	5	37	326	11%
drug_era group during day -730 through 0 days relative to index: ANTITHROMBOTIC AGENTS	1	30	549	5.5%
drug_era group during day -730 through 0 days relative to index: ANTITHROMBOTIC AGENTS	2	16	549	2.9%
drug_era group during day -730 through 0 days relative to index: ANTITHROMBOTIC AGENTS	4	326	549	59.4%

covariateName	clus	num	tot	pct
drug_era group during day -730 through 0 days relative to index: ANTITHROMBOTIC AGENTS	5	177	549	32.2%
drug_era group during day -730 through 0 days relative to index: BETA BLOCKING AGENTS	4	28	28	100%
drug_era group during day -730 through 0 days relative to index: CALCIUM CHANNEL BLOCKERS	1	379	1923	19.7%
drug_era group during day -730 through 0 days relative to index: CALCIUM CHANNEL BLOCKERS	2	205	1923	10.7%
drug_era group during day -730 through 0 days relative to index: CALCIUM CHANNEL BLOCKERS	3	27	1923	1.4%
drug_era group during day -730 through 0 days relative to index: CALCIUM CHANNEL BLOCKERS	4	571	1923	29.7%
drug_era group during day -730 through 0 days relative to index: CALCIUM CHANNEL BLOCKERS	5	741	1923	38.5%
drug_era group during day -730 through 0 days relative to index: ACE INHIBITORS, PLAIN	1	391	1631	24.0%
drug_era group during day -730 through 0 days relative to index: ACE INHIBITORS, PLAIN	2	197	1631	12.1%
drug_era group during day -730 through 0 days relative to index: ACE INHIBITORS, PLAIN	3	39	1631	2.4%
drug_era group during day -730 through 0 days relative to index: ACE INHIBITORS, PLAIN	4	381	1631	23.4%
drug_era group during day -730 through 0 days relative to index: ACE INHIBITORS, PLAIN	5	623	1631	38.2%
drug_era group during day -730 through 0 days relative to index: LIPID MODIFYING AGENTS, PLAIN	1	174	1101	15.80%

covariateName	clus	num	tot	pct
drug_era group during day -730 through 0 days relative to index: LIPID MODIFYING AGENTS, PLAIN	2	20	1101	1.82%
drug_era group during day -730 through 0 days relative to index: LIPID MODIFYING AGENTS, PLAIN	3	4	1101	0.36%
drug_era group during day -730 through 0 days relative to index: LIPID MODIFYING AGENTS, PLAIN	4	432	1101	39.24%
drug_era group during day -730 through 0 days relative to index: LIPID MODIFYING AGENTS, PLAIN	5	471	1101	42.78%
drug_era group during day -730 through 0 days relative to index: SEX HORMONES AND MODULATORS OF THE GENITAL SYSTEM	1	302	1996	15.1%
drug_era group during day -730 through 0 days relative to index: SEX HORMONES AND MODULATORS OF THE GENITAL SYSTEM	2	746	1996	37.4%
drug_era group during day -730 through 0 days relative to index: SEX HORMONES AND MODULATORS OF THE GENITAL SYSTEM	3	38	1996	1.9%
drug_era group during day -730 through 0 days relative to index: SEX HORMONES AND MODULATORS OF THE GENITAL SYSTEM	4	116	1996	5.8%
drug_era group during day -730 through 0 days relative to index: SEX HORMONES AND MODULATORS OF THE GENITAL SYSTEM	5	794	1996	39.8%
drug_era group during day -730 through 0 days relative to index: DRUGS FOR OBSTRUCTIVE AIRWAY DISEASES	1	124	554	22%
drug_era group during day -730 through 0 days relative to index: DRUGS FOR OBSTRUCTIVE AIRWAY DISEASES	2	82	554	15%
drug_era group during day -730 through 0 days relative to index: DRUGS FOR OBSTRUCTIVE AIRWAY DISEASES	3	22	554	4%
drug_era group during day -730 through 0 days relative to index: DRUGS FOR OBSTRUCTIVE AIRWAY DISEASES	4	100	554	18%

covariateName	clus	num	tot	pct
drug_era group during day -730 through 0 days relative to index: DRUGS FOR OBSTRUCTIVE AIRWAY DISEASES	5	226	554	41%
drug_era group during day -730 through 0 days relative to index: Antibiotics	1	194	862	22.5%
drug_era group during day -730 through 0 days relative to index: Antibiotics	2	138	862	16.0%
drug_era group during day -730 through 0 days relative to index: Antibiotics	3	28	862	3.2%
drug_era group during day -730 through 0 days relative to index: Antibiotics	4	153	862	17.7%
drug_era group during day -730 through 0 days relative to index: Antibiotics	5	349	862	40.5%
drug_era group during day -730 through 0 days relative to index: OPIOIDS	1	253	1078	23.5%
drug_era group during day -730 through 0 days relative to index: OPIOIDS	2	154	1078	14.3%
drug_era group during day -730 through 0 days relative to index: OPIOIDS	3	48	1078	4.5%
drug_era group during day -730 through 0 days relative to index: OPIOIDS	4	199	1078	18.5%
drug_era group during day -730 through 0 days relative to index: OPIOIDS	5	424	1078	39.3%
drug_era group during day -730 through 0 days relative to index: ANTIDEPRESSANTS	1	4	14	28.6%
drug_era group during day -730 through 0 days relative to index: ANTIDEPRESSANTS	2	1	14	7.1%

covariateName	clus	num	tot	pct
drug_era group during day -730 through 0 days relative to index: ANTIDEPRESSANTS	4	2	14	14.3%
drug_era group during day -730 through 0 days relative to index: ANTIDEPRESSANTS	5	7	14	50.0%
condition_era group during day -730 through 0 days relative to index: Disorder caused by alcohol	1	1	2	50%
condition_era group during day -730 through 0 days relative to index: Disorder caused by alcohol	5	1	2	50%
condition_era group during day -730 through 0 days relative to index: Chronic kidney disease	4	15	15	100%

Table B.38: GMMM PSE Table for Synthea

assigned_pcp	patientCount	pse	z
Arnulfo Carter	531	560.99	-0.22
Denny Yundt	488	426.01	0.55
Florinda Gutmann	624	710.47	-0.51
Halley Hamill	434	292.39	1.66
Harlan Beer	629	740.44	-0.63
Houston Beier	401	274.79	1.59
Jeri Hansen	700	827.46	-0.65
Kathie Zulauf	503	386.95	1.07
Marilou Conn	622	752.05	-0.74
Modesto Trantow	426	286.20	1.67
Paulene Schultz	642	744.12	-0.57

assigned_pcp	patientCount	pse	z
Sharyl Hilpert	667	809.87	-0.75
Shonta Sanford	631	724.59	-0.54
Taylor Wilderman	394	292.19	1.25
Wilber Harris	385	248.49	1.86

B.4.4 Mixed PBS

Table B.39: Mixed Pbs Continuous Summary for Synthea

rowname	cluster1	cluster2	cluster3	cluster4
outpatientVisitCount	2	4	9	16
medicationCount	1	3	4	7
conditionCount	1	1	2	3
procedureCount	1	1	5	10
measurementCount	14	20	29	38
inpatientVisitCount	1	1	1	2
clusterSize	1498	4283	2074	222

Table B.40: Mixed Pbs Categorical Summary for Synthea

covariateName	clus	num	tot	pct
age group: 15 - 19	1	93	353	26%
age group: 15 - 19	2	191	353	54%
age group: 15 - 19	3	62	353	18%
age group: 15 - 19	4	7	353	2%

covariateName	clus	num	tot	pct
age group: 20 - 24	1	239	929	25.7%
age group: 20 - 24	2	440	929	47.4%
age group: 20 - 24	3	228	929	24.5%
age group: 20 - 24	4	22	929	2.4%
age group: 25 - 29	1	323	1105	29.2%
age group: 25 - 29	2	400	1105	36.2%
age group: 25 - 29	3	355	1105	32.1%
age group: 25 - 29	4	27	1105	2.4%
age group: 30 - 34	1	203	789	25.7%
age group: 30 - 34	2	307	789	38.9%
age group: 30 - 34	3	249	789	31.6%
age group: 30 - 34	4	30	789	3.8%
age group: 35 - 39	1	104	563	18%
age group: 35 - 39	2	305	563	54%
age group: 35 - 39	3	143	563	25%
age group: 35 - 39	4	11	563	2%
age group: 40 - 44	1	146	603	24.2%
age group: 40 - 44	2	333	603	55.2%
age group: 40 - 44	3	110	603	18.2%
age group: 40 - 44	4	14	603	2.3%
age group: 45 - 49	1	126	615	20.5%
age group: 45 - 49	2	346	615	56.3%
age group: 45 - 49	3	132	615	21.5%
age group: 45 - 49	4	11	615	1.8%
age group: 50 - 54	1	85	636	13.4%
age group: 50 - 54	2	421	636	66.2%

covariateName	clus	num	tot	pct
age group: 50 - 54	3	121	636	19.0%
age group: 50 - 54	4	9	636	1.4%
age group: 55 - 59	1	101	651	15.51%
age group: 55 - 59	2	443	651	68.05%
age group: 55 - 59	3	102	651	15.67%
age group: 55 - 59	4	5	651	0.77%
age group: 60 - 64	1	68	593	11.5%
age group: 60 - 64	2	409	593	69.0%
age group: 60 - 64	3	105	593	17.7%
age group: 60 - 64	4	11	593	1.9%
age group: 65 - 69	1	2	384	0.52%
age group: 65 - 69	2	258	384	67.19%
age group: 65 - 69	3	110	384	28.65%
age group: 65 - 69	4	14	384	3.65%
age group: 70 - 74	1	6	269	2.2%
age group: 70 - 74	2	147	269	54.6%
age group: 70 - 74	3	104	269	38.7%
age group: 70 - 74	4	12	269	4.5%
age group: 75 - 79	2	129	223	57.8%
age group: 75 - 79	3	77	223	34.5%
age group: 75 - 79	4	17	223	7.6%
age group: 80 - 84	2	65	136	47.8%
age group: 80 - 84	3	60	136	44.1%
age group: 80 - 84	4	11	136	8.1%
age group: 85 - 89	1	1	61	1.6%
age group: 85 - 89	2	29	61	47.5%

covariateName	clus	num	tot	pct
age group: 85 - 89	3	25	61	41.0%
age group: 85 - 89	4	6	61	9.8%
age group: 90 - 94	2	18	44	40.9%
age group: 90 - 94	3	23	44	52.3%
age group: 90 - 94	4	3	44	6.8%
age group: 95 - 99	2	13	37	35.1%
age group: 95 - 99	3	22	37	59.5%
age group: 95 - 99	4	2	37	5.4%
age group: 100 - 104	2	17	56	30.4%
age group: 100 - 104	3	34	56	60.7%
age group: 100 - 104	4	5	56	8.9%
age group: 105 - 109	1	1	30	3.3%
age group: 105 - 109	2	12	30	40.0%
age group: 105 - 109	3	12	30	40.0%
age group: 105 - 109	4	5	30	16.7%
gender = MALE	1	1365	3807	35.9%
gender = MALE	2	2003	3807	52.6%
gender = MALE	3	388	3807	10.2%
gender = MALE	4	51	3807	1.3%
race = Asian	1	47	214	22.0%
race = Asian	2	109	214	50.9%
race = Asian	3	51	214	23.8%
race = Asian	4	7	214	3.3%
race = Black or African American	1	766	4071	19%
race = Black or African American	2	2127	4071	52%
race = Black or African American	3	1056	4071	26%

covariateName	clus	num	tot	pct
race = Black or African American	4	122	4071	3%
race = White	1	676	3742	18.1%
race = White	2	2022	3742	54.0%
race = White	3	952	3742	25.4%
race = White	4	92	3742	2.5%
gender = FEMALE	1	133	4270	3.1%
gender = FEMALE	2	2280	4270	53.4%
gender = FEMALE	3	1686	4270	39.5%
gender = FEMALE	4	171	4270	4.0%
condition_era group during day -730 through 0 days relative to index: Osteoporosis	2	13	38	34%
condition_era group during day -730 through 0 days relative to index: Osteoporosis	3	20	38	53%
condition_era group during day -730 through 0 days relative to index: Osteoporosis	4	5	38	13%
condition_era group during day -730 through 0 days relative to index: Neoplasm of breast	3	10	18	56%
condition_era group during day -730 through 0 days relative to index: Neoplasm of breast	4	8	18	44%
condition_era group during day -730 through 0 days relative to index: Sepsis	2	2	12	17%
condition_era group during day -730 through 0 days relative to index: Sepsis	3	5	12	42%
condition_era group during day -730 through 0 days relative to index: Sepsis	4	5	12	42%

covariateName	clus	num	tot	pct
condition_era group during day -730 through 0 days relative to index: Neoplasm of digestive tract	2	5	38	13%
condition_era group during day -730 through 0 days relative to index: Neoplasm of digestive tract	3	26	38	68%
condition_era group during day -730 through 0 days relative to index: Neoplasm of digestive tract	4	7	38	18%
condition_era group during day -730 through 0 days relative to index: Diabetes mellitus	2	11	37	30%
condition_era group during day -730 through 0 days relative to index: Diabetes mellitus	3	18	37	49%
condition_era group during day -730 through 0 days relative to index: Diabetes mellitus	4	8	37	22%
condition_era group during day -730 through 0 days relative to index: Neoplasm of respiratory tract	3	3	5	60%
condition_era group during day -730 through 0 days relative to index: Neoplasm of respiratory tract	4	2	5	40%
condition_era group during day -730 through 0 days relative to index: Chronic obstructive lung disease	2	7	10	70%
condition_era group during day -730 through 0 days relative to index: Chronic obstructive lung disease	3	3	10	30%
condition_era group during day -730 through 0 days relative to index: Chronic sinusitis	1	1	84	1.2%
condition_era group during day -730 through 0 days relative to index: Chronic sinusitis	2	43	84	51.2%
condition_era group during day -730 through 0 days relative to index: Chronic sinusitis	3	35	84	41.7%

covariateName	clus	num	tot	pct
condition_era group during day -730 through 0 days relative to index: Chronic sinusitis	4	5	84	6.0%
condition_era group during day -730 through 0 days relative to index: Heart failure	2	1	28	3.6%
condition_era group during day -730 through 0 days relative to index: Heart failure	3	10	28	35.7%
condition_era group during day -730 through 0 days relative to index: Heart failure	4	17	28	60.7%
condition_era group during day -730 through 0 days relative to index: Hypertensive disorder	2	71	189	38%
condition_era group during day -730 through 0 days relative to index: Hypertensive disorder	3	88	189	47%
condition_era group during day -730 through 0 days relative to index: Hypertensive disorder	4	30	189	16%
condition_era group during day -730 through 0 days relative to index: Asthma	2	22	43	51.2%
condition_era group during day -730 through 0 days relative to index: Asthma	3	17	43	39.5%
condition_era group during day -730 through 0 days relative to index: Asthma	4	4	43	9.3%
condition_era group during day -730 through 0 days relative to index: Migraine	2	8	12	67%
condition_era group during day -730 through 0 days relative to index: Migraine	3	4	12	33%
condition_era group during day -730 through 0 days relative to index: Heart disease	2	25	127	20%

covariateName	clus	num	tot	pct
condition_era group during day -730 through 0 days relative to index: Heart disease	3	69	127	54%
condition_era group during day -730 through 0 days relative to index: Heart disease	4	33	127	26%
condition_era group during day -730 through 0 days relative to index: Cerebrovascular disease	2	4	33	12%
condition_era group during day -730 through 0 days relative to index: Cerebrovascular disease	3	23	33	70%
condition_era group during day -730 through 0 days relative to index: Cerebrovascular disease	4	6	33	18%
condition_era group during day -730 through 0 days relative to index: Complication of pregnancy, childbirth and/or the puerperium	2	1	199	0.5%
condition_era group during day -730 through 0 days relative to index: Complication of pregnancy, childbirth and/or the puerperium	3	155	199	77.9%
condition_era group during day -730 through 0 days relative to index: Complication of pregnancy, childbirth and/or the puerperium	4	43	199	21.6%
condition_era group during day -730 through 0 days relative to index: Chronic pain	2	46	77	59.7%
condition_era group during day -730 through 0 days relative to index: Chronic pain	3	27	77	35.1%
condition_era group during day -730 through 0 days relative to index: Chronic pain	4	4	77	5.2%
condition_era group during day -730 through 0 days relative to index: Anemia	2	100	297	34%
condition_era group during day -730 through 0 days relative to index: Anemia	3	143	297	48%

covariateName	clus	num	tot	pct
condition_era group during day -730 through 0 days relative to index: Anemia	4	54	297	18%
condition_era group during day -730 through 0 days relative to index: Anxiety	2	1	3	33%
condition_era group during day -730 through 0 days relative to index: Anxiety	3	1	3	33%
condition_era group during day -730 through 0 days relative to index: Anxiety	4	1	3	33%
condition_era group during day -730 through 0 days relative to index: Concussion injury of brain	1	15	173	8.7%
condition_era group during day -730 through 0 days relative to index: Concussion injury of brain	2	101	173	58.4%
condition_era group during day -730 through 0 days relative to index: Concussion injury of brain	3	50	173	28.9%
condition_era group during day -730 through 0 days relative to index: Concussion injury of brain	4	7	173	4.0%
condition_era group during day -730 through 0 days relative to index: Neoplasm of digestive system	2	5	38	13%
condition_era group during day -730 through 0 days relative to index: Neoplasm of digestive system	3	26	38	68%
condition_era group during day -730 through 0 days relative to index: Neoplasm of digestive system	4	7	38	18%
condition_era group during day -730 through 0 days relative to index: Neoplasm of trunk	2	6	75	8%
condition_era group during day -730 through 0 days relative to index: Neoplasm of trunk	3	50	75	67%

covariateName	clus	num	tot	pct
condition_era group during day -730 through 0 days relative to index: Neoplasm of trunk	4	19	75	25%
condition_era group during day -730 through 0 days relative to index: Neoplasm of intra-abdominal organs	2	6	53	11%
condition_era group during day -730 through 0 days relative to index: Neoplasm of intra-abdominal organs	3	37	53	70%
condition_era group during day -730 through 0 days relative to index: Neoplasm of intra-abdominal organs	4	10	53	19%
condition_era group during day -730 through 0 days relative to index: Neoplasm of intrathoracic organs	3	3	5	60%
condition_era group during day -730 through 0 days relative to index: Neoplasm of intrathoracic organs	4	2	5	40%
condition_era group during day -730 through 0 days relative to index: Neoplasm of pelvis	2	1	15	6.7%
condition_era group during day -730 through 0 days relative to index: Neoplasm of pelvis	3	11	15	73.3%
condition_era group during day -730 through 0 days relative to index: Neoplasm of pelvis	4	3	15	20.0%
condition_era group during day -730 through 0 days relative to index: Chronic urinary tract infection	2	3	6	50%
condition_era group during day -730 through 0 days relative to index: Chronic urinary tract infection	3	2	6	33%
condition_era group during day -730 through 0 days relative to index: Chronic urinary tract infection	4	1	6	17%
observation during day -730 through 0 days relative to index: Body mass index 30+ - obesity	1	41	164	25%

covariateName	clus	num	tot	pct
observation during day -730 through 0 days relative to index: Body mass index 30+ - obesity	2	80	164	49%
observation during day -730 through 0 days relative to index: Body mass index 30+ - obesity	3	38	164	23%
observation during day -730 through 0 days relative to index: Body mass index 30+ - obesity	4	5	164	3%
condition_era group during day -730 through 0 days relative to index: Neoplasm of thorax	3	13	23	57%
condition_era group during day -730 through 0 days relative to index: Neoplasm of thorax	4	10	23	43%
condition_era group during day -730 through 0 days relative to index: Neoplasm of digestive organ	2	5	38	13%
condition_era group during day -730 through 0 days relative to index: Neoplasm of digestive organ	3	26	38	68%
condition_era group during day -730 through 0 days relative to index: Neoplasm of digestive organ	4	7	38	18%
condition_era group during day -730 through 0 days relative to index: Neoplasm of lower respiratory tract	3	3	5	60%
condition_era group during day -730 through 0 days relative to index: Neoplasm of lower respiratory tract	4	2	5	40%
condition_era group during day -730 through 0 days relative to index: Neoplasm of lung	3	3	5	60%
condition_era group during day -730 through 0 days relative to index: Neoplasm of lung	4	2	5	40%
condition_era group during day -730 through 0 days relative to index: Neoplasm of male genital organ	2	1	15	6.7%

covariateName	clus	num	tot	pct
condition_era group during day -730 through 0 days relative to index: Neoplasm of male genital organ	3	11	15	73.3%
condition_era group during day -730 through 0 days relative to index: Neoplasm of male genital organ	4	3	15	20.0%
condition_era group during day -730 through 0 days relative to index: Neoplasm of prostate	2	1	15	6.7%
condition_era group during day -730 through 0 days relative to index: Neoplasm of prostate	3	11	15	73.3%
condition_era group during day -730 through 0 days relative to index: Neoplasm of prostate	4	3	15	20.0%
condition_era group during day -730 through 0 days relative to index: Neoplasm of gastrointestinal tract	2	5	38	13%
condition_era group during day -730 through 0 days relative to index: Neoplasm of gastrointestinal tract	3	26	38	68%
condition_era group during day -730 through 0 days relative to index: Neoplasm of gastrointestinal tract	4	7	38	18%
condition_era group during day -730 through 0 days relative to index: Neoplasm of large intestine	2	5	38	13%
condition_era group during day -730 through 0 days relative to index: Neoplasm of large intestine	3	26	38	68%
condition_era group during day -730 through 0 days relative to index: Neoplasm of large intestine	4	7	38	18%
condition_era group during day -730 through 0 days relative to index: Neoplasm of colon	2	5	38	13%
condition_era group during day -730 through 0 days relative to index: Neoplasm of colon	3	26	38	68%

covariateName	clus	num	tot	pct
condition_era group during day -730 through 0 days relative to index: Neoplasm of colon	4	7	38	18%
condition_era group during day -730 through 0 days relative to index: Neoplasm of intestinal tract	2	5	38	13%
condition_era group during day -730 through 0 days relative to index: Neoplasm of intestinal tract	3	26	38	68%
condition_era group during day -730 through 0 days relative to index: Neoplasm of intestinal tract	4	7	38	18%
condition_era group during day -730 through 0 days relative to index: Neoplasm of abdomen	2	6	53	11%
condition_era group during day -730 through 0 days relative to index: Neoplasm of abdomen	3	37	53	70%
condition_era group during day -730 through 0 days relative to index: Neoplasm of abdomen	4	10	53	19%
condition_era group during day -730 through 0 days relative to index: Dementia	2	7	31	23%
condition_era group during day -730 through 0 days relative to index: Dementia	3	17	31	55%
condition_era group during day -730 through 0 days relative to index: Dementia	4	7	31	23%
condition_era group during day -730 through 0 days relative to index: Drug overdose	2	13	25	52%
condition_era group during day -730 through 0 days relative to index: Drug overdose	3	10	25	40%
condition_era group during day -730 through 0 days relative to index: Drug overdose	4	2	25	8%

covariateName	clus	num	tot	pct
condition_era group during day -730 through 0 days relative to index: Normal pregnancy	2	44	1195	3.7%
condition_era group during day -730 through 0 days relative to index: Normal pregnancy	3	1029	1195	86.1%
condition_era group during day -730 through 0 days relative to index: Normal pregnancy	4	122	1195	10.2%
procedure_occurrence during day -730 through 0 days relative to index: Colonoscopy	1	39	999	3.9%
procedure_occurrence during day -730 through 0 days relative to index: Colonoscopy	2	616	999	61.7%
procedure_occurrence during day -730 through 0 days relative to index: Colonoscopy	3	305	999	30.5%
procedure_occurrence during day -730 through 0 days relative to index: Colonoscopy	4	39	999	3.9%
condition_era group during day -730 through 0 days relative to index: Arthritis	2	6	8	75%
condition_era group during day -730 through 0 days relative to index: Arthritis	3	2	8	25%
drug_era group during day -730 through 0 days relative to index: INSULINS AND ANALOGUES	2	7	221	3.2%
drug_era group during day -730 through 0 days relative to index: INSULINS AND ANALOGUES	3	155	221	70.1%
drug_era group during day -730 through 0 days relative to index: INSULINS AND ANALOGUES	4	59	221	26.7%
drug_era group during day -730 through 0 days relative to index: BLOOD GLUCOSE LOWERING DRUGS, EXCL. INSULINS	2	52	326	16%

covariateName	clus	num	tot	pct
drug_era group during day -730 through 0 days relative to index: BLOOD GLUCOSE LOWERING DRUGS, EXCL. INSULINS	3	216	326	66%
drug_era group during day -730 through 0 days relative to index: BLOOD GLUCOSE LOWERING DRUGS, EXCL. INSULINS	4	58	326	18%
drug_era group during day -730 through 0 days relative to index: ANTITHROMBOTIC AGENTS	2	188	549	34%
drug_era group during day -730 through 0 days relative to index: ANTITHROMBOTIC AGENTS	3	304	549	55%
drug_era group during day -730 through 0 days relative to index: ANTITHROMBOTIC AGENTS	4	57	549	10%
drug_era group during day -730 through 0 days relative to index: BETA BLOCKING AGENTS	3	9	28	32%
drug_era group during day -730 through 0 days relative to index: BETA BLOCKING AGENTS	4	19	28	68%
drug_era group during day -730 through 0 days relative to index: CALCIUM CHANNEL BLOCKERS	1	49	1923	2.5%
drug_era group during day -730 through 0 days relative to index: CALCIUM CHANNEL BLOCKERS	2	1026	1923	53.4%
drug_era group during day -730 through 0 days relative to index: CALCIUM CHANNEL BLOCKERS	3	735	1923	38.2%
drug_era group during day -730 through 0 days relative to index: CALCIUM CHANNEL BLOCKERS	4	113	1923	5.9%
drug_era group during day -730 through 0 days relative to index: ACE INHIBITORS, PLAIN	1	63	1631	3.9%
drug_era group during day -730 through 0 days relative to index: ACE INHIBITORS, PLAIN	2	929	1631	57.0%

covariateName	clus	num	tot	pct
drug_era group during day -730 through 0 days relative to index: ACE INHIBITORS, PLAIN	3	545	1631	33.4%
drug_era group during day -730 through 0 days relative to index: ACE INHIBITORS, PLAIN	4	94	1631	5.8%
drug_era group during day -730 through 0 days relative to index: LIPID MODIFYING AGENTS, PLAIN	1	7	1101	0.64%
drug_era group during day -730 through 0 days relative to index: LIPID MODIFYING AGENTS, PLAIN	2	605	1101	54.95%
drug_era group during day -730 through 0 days relative to index: LIPID MODIFYING AGENTS, PLAIN	3	423	1101	38.42%
drug_era group during day -730 through 0 days relative to index: LIPID MODIFYING AGENTS, PLAIN	4	66	1101	5.99%
drug_era group during day -730 through 0 days relative to index: SEX HORMONES AND MODULATORS OF THE GENITAL SYSTEM	1	20	1996	1.0%
drug_era group during day -730 through 0 days relative to index: SEX HORMONES AND MODULATORS OF THE GENITAL SYSTEM	2	917	1996	45.9%
drug_era group during day -730 through 0 days relative to index: SEX HORMONES AND MODULATORS OF THE GENITAL SYSTEM	3	946	1996	47.4%
drug_era group during day -730 through 0 days relative to index: SEX HORMONES AND MODULATORS OF THE GENITAL SYSTEM	4	113	1996	5.7%
drug_era group during day -730 through 0 days relative to index: DRUGS FOR OBSTRUCTIVE AIRWAY DISEASES	1	25	554	4.5%
drug_era group during day -730 through 0 days relative to index: DRUGS FOR OBSTRUCTIVE AIRWAY DISEASES	2	323	554	58.3%
drug_era group during day -730 through 0 days relative to index: DRUGS FOR OBSTRUCTIVE AIRWAY DISEASES	3	171	554	30.9%

covariateName	clus	num	tot	pct
drug_era group during day -730 through 0 days relative to index: DRUGS FOR OBSTRUCTIVE AIRWAY DISEASES	4	35	554	6.3%
drug_era group during day -730 through 0 days relative to index: Antibiotics	1	38	862	4.4%
drug_era group during day -730 through 0 days relative to index: Antibiotics	2	471	862	54.6%
drug_era group during day -730 through 0 days relative to index: Antibiotics	3	303	862	35.2%
drug_era group during day -730 through 0 days relative to index: Antibiotics	4	50	862	5.8%
drug_era group during day -730 through 0 days relative to index: OPIOIDS	1	63	1078	5.8%
drug_era group during day -730 through 0 days relative to index: OPIOIDS	2	612	1078	56.8%
drug_era group during day -730 through 0 days relative to index: OPIOIDS	3	344	1078	31.9%
drug_era group during day -730 through 0 days relative to index: OPIOIDS	4	59	1078	5.5%
drug_era group during day -730 through 0 days relative to index: ANTIDEPRESSANTS	2	11	14	79%
drug_era group during day -730 through 0 days relative to index: ANTIDEPRESSANTS	3	3	14	21%
condition_era group during day -730 through 0 days relative to index: Disorder caused by alcohol	2	2	2	100%
condition_era group during day -730 through 0 days relative to index: Chronic kidney disease	2	1	15	6.7%

covariateName	clus	num	tot	pct
condition_era group during day -730 through 0 days relative to index: Chronic kidney disease	3	8	15	53.3%
condition_era group during day -730 through 0 days relative to index: Chronic kidney disease	4	6	15	40.0%

Table B.41: Mixed Pbs PSE Table for Synthea

assigned_pcp	patientCount	pse	z
Arnulfo Carter	531	443.14	0.73
Denny Yundt	488	356.38	1.30
Florinda Gutmann	624	560.16	0.43
Halley Hamill	434	265.33	2.08
Harlan Beer	629	580.48	0.32
Houston Beier	401	249.33	2.02
Jeri Hansen	700	646.17	0.31
Kathie Zulauf	503	360.71	1.37
Marilou Conn	622	589.24	0.21
Modesto Trantow	426	258.18	2.13
Paulene Schultz	642	579.44	0.41
Sharyl Hilpert	667	620.67	0.28
Shonta Sanford	631	576.77	0.36
Taylor Wilderman	394	279.54	1.45
Wilber Harris	385	229.39	2.22

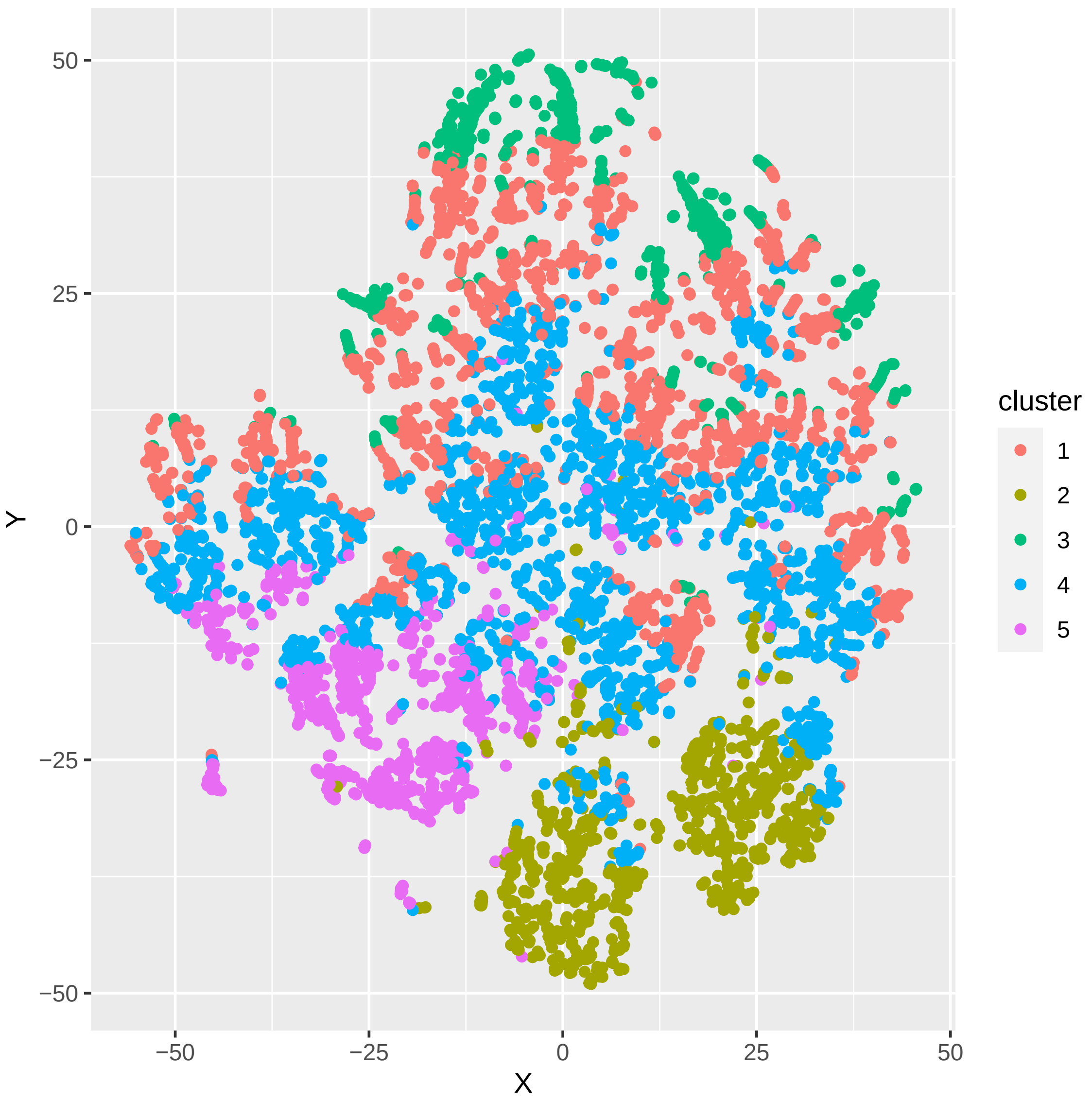


Figure B.7: KAMILA TSNE Plot for Synthea Data

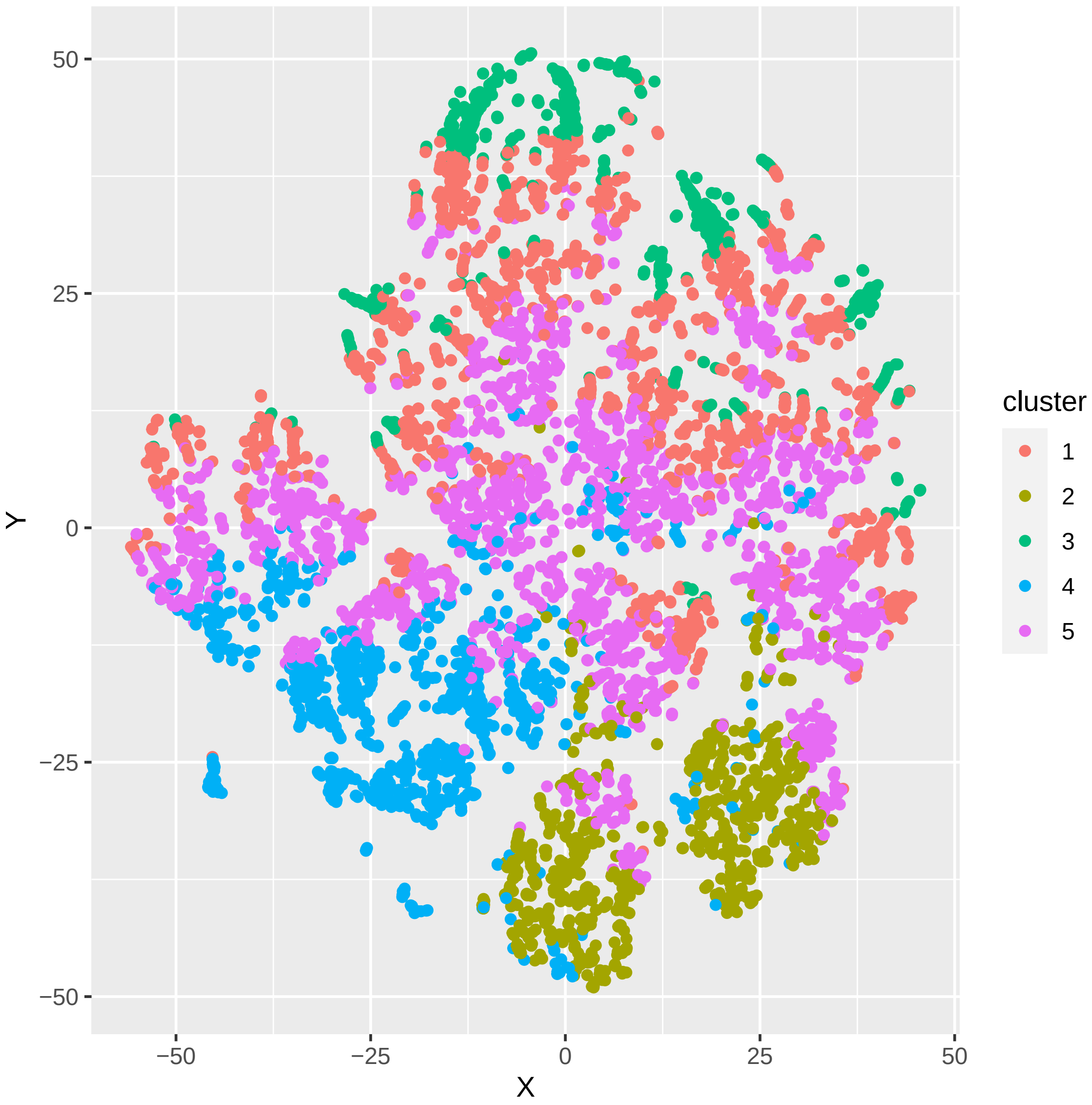


Figure B.8: GMMM TSNE Plot for Synthea Data

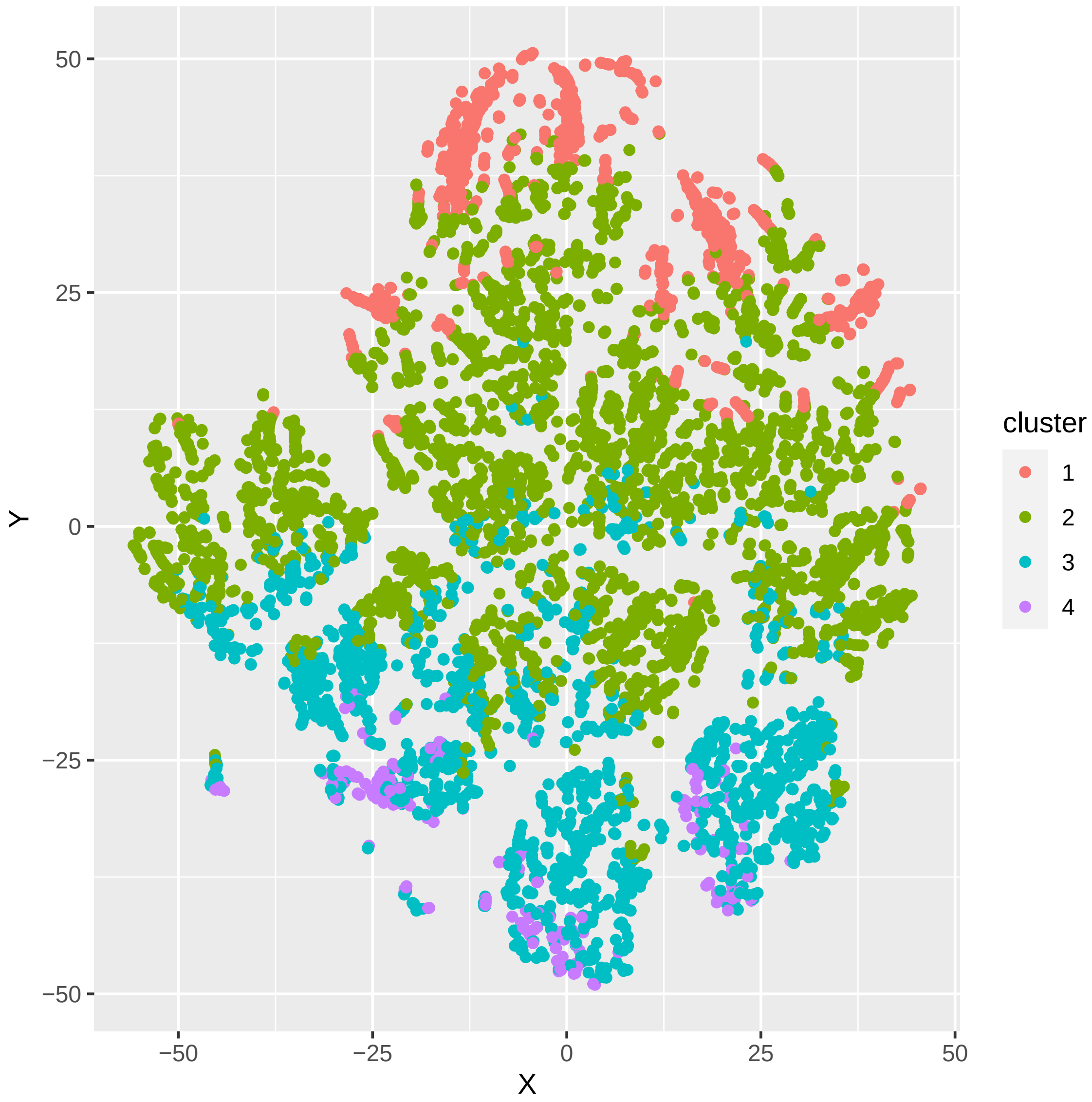


Figure B.9: Mixed PBS TSNE Plot for Synthea Data

COLOPHON

This document is set in EB Garamond, Source Code Pro and Lato. The body text is set at 11pt with *lmr*.

It was written in R Markdown and LaTeX, and rendered into PDF using huskydown and bookdown.

This document was typeset using the XeTeX typesetting system, and the University of Washington Thesis class class created by Jim Fox. Under the hood, the University of Washington Thesis LaTeX template is used to ensure that documents conform precisely to submission standards. Other elements of the document formatting source code have been taken from the Latex, Knitr, and RMarkdown templates for UC Berkeley's graduate thesis, and Dissertate: a LaTeX dissertation template to support the production and typesetting of a PhD dissertation at Harvard, Princeton, and NYU

This version of the thesis was generated on 2022-05-10 12:48:23.

The computational environment that was used to generate this version is as follows:

```
- Session info ---------------------------------------------------------------
 setting  value
 version  R version 4.1.0 (2021-05-18)
 os       Windows 10 x64 (build 19043)
 system   x86_64, mingw32
 ui       RTerm
 language (EN)
 collate  English_United States.1252
 ctype    English_United States.1252
 tz       America/New_York
 date     2022-05-10
 pandoc   2.11.4 @ C:/Program Files/RStudio/bin/pandoc/ (via rmarkdown)

- Packages -------------------------------------------------------------------
 package     * version date (UTC) lib source
 assertthat    0.2.1   2019-03-21 [1] CRAN (R 4.1.0)
 backports     1.4.1   2021-12-13 [1] CRAN (R 4.1.2)
 bookdown      0.25    2022-03-16 [1] CRAN (R 4.1.3)
 brio          1.1.3   2021-11-30 [1] CRAN (R 4.1.3)
 broom         0.7.9   2021-07-27 [1] CRAN (R 4.1.1)
 cachem        1.0.6   2021-08-19 [1] CRAN (R 4.1.1)
 callr         3.7.0   2021-04-20 [1] CRAN (R 4.1.0)
 cellranger    1.1.0   2016-07-27 [1] CRAN (R 4.1.0)
 class         7.3-19  2021-05-03 [2] CRAN (R 4.1.0)
 cli           3.2.0   2022-02-14 [1] CRAN (R 4.1.3)
 cluster       2.1.2   2021-04-17 [2] CRAN (R 4.1.0)
 colorspace    2.0-2   2021-06-24 [1] CRAN (R 4.1.0)
 crayon        1.5.1   2022-03-26 [1] CRAN (R 4.1.3)
 data.table    1.14.2  2021-09-27 [1] CRAN (R 4.1.1)
 DBI           1.1.2   2021-12-20 [1] CRAN (R 4.1.3)
 dbplyr        2.1.1   2021-04-06 [1] CRAN (R 4.1.0)
 DEoptimR      1.0-9   2021-05-24 [1] CRAN (R 4.1.1)
 desc          1.4.1   2022-03-06 [1] CRAN (R 4.1.3)
 devtools    * 2.4.3   2021-11-30 [1] CRAN (R 4.1.3)
 digest        0.6.29  2021-12-01 [1] CRAN (R 4.1.2)
 diptest       0.76-0  2021-05-04 [1] CRAN (R 4.1.1)
```

```
dplyr       * 1.0.8   2022-02-08 [1] CRAN (R 4.1.3)
ellipsis      0.3.2   2021-04-29 [1] CRAN (R 4.1.0)
evaluate      0.15    2022-02-18 [1] CRAN (R 4.1.3)
fansi         1.0.3   2022-03-24 [1] CRAN (R 4.1.3)
farver        2.1.0   2021-02-28 [1] CRAN (R 4.1.0)
fastmap       1.1.0   2021-01-25 [1] CRAN (R 4.1.0)
flexmix       2.3-17  2020-10-12 [1] CRAN (R 4.1.2)
forcats     * 0.5.1   2021-01-27 [1] CRAN (R 4.1.0)
Formula       1.2-4   2020-10-16 [1] CRAN (R 4.1.1)
fpc           2.2-9   2020-12-06 [1] CRAN (R 4.1.2)
fs            1.5.2   2021-12-08 [1] CRAN (R 4.1.3)
generics      0.1.2   2022-01-31 [1] CRAN (R 4.1.3)
ggplot2     * 3.3.5   2021-06-25 [1] CRAN (R 4.1.1)
git2r         0.30.1  2022-03-16 [1] CRAN (R 4.1.3)
glue          1.6.2   2022-02-24 [1] CRAN (R 4.1.3)
gtable        0.3.0   2019-03-25 [1] CRAN (R 4.1.0)
haven         2.4.1   2021-04-23 [1] CRAN (R 4.1.0)
here          1.0.1   2020-12-13 [1] CRAN (R 4.1.1)
highr         0.9     2021-04-16 [1] CRAN (R 4.1.0)
hms           1.1.1   2021-09-26 [1] CRAN (R 4.1.1)
htmltools     0.5.2   2021-08-25 [1] CRAN (R 4.1.1)
httr          1.4.2   2020-07-20 [1] CRAN (R 4.1.0)
huskydown   * 0.0.5   2022-04-06 [1] Github (benmarwick/huskydown@addb48e)
jsonlite      1.8.0   2022-02-22 [1] CRAN (R 4.1.3)
kernlab       0.9-29  2019-11-12 [1] CRAN (R 4.1.1)
knitr         1.38    2022-03-25 [1] CRAN (R 4.1.3)
labeling      0.4.2   2020-10-20 [1] CRAN (R 4.1.0)
lattice       0.20-44 2021-05-02 [2] CRAN (R 4.1.0)
```

```
lifecycle     1.0.1   2021-09-24 [1] CRAN (R 4.1.1)
lubridate     1.8.0   2021-10-07 [1] CRAN (R 4.1.0)
magrittr      2.0.3   2022-03-30 [1] CRAN (R 4.1.3)
MASS          7.3-54  2021-05-03 [2] CRAN (R 4.1.0)
mclust        5.4.7   2020-11-20 [1] CRAN (R 4.1.1)
memoise       2.0.1   2021-11-26 [1] CRAN (R 4.1.3)
modelr        0.1.8   2020-05-19 [1] CRAN (R 4.1.0)
modeltools    0.2-23  2020-03-05 [1] CRAN (R 4.1.1)
munsell       0.5.0   2018-06-12 [1] CRAN (R 4.1.1)
nnet          7.3-16  2021-05-03 [2] CRAN (R 4.1.0)
patchwork   * 1.1.1   2020-12-17 [1] CRAN (R 4.1.1)
pillar        1.7.0   2022-02-01 [1] CRAN (R 4.1.3)
pkgbuild      1.3.1   2021-12-20 [1] CRAN (R 4.1.3)
pkgconfig     2.0.3   2019-09-22 [1] CRAN (R 4.1.0)
pkgload       1.2.4   2021-11-30 [1] CRAN (R 4.1.3)
png           0.1-7   2013-12-03 [1] CRAN (R 4.1.0)
prabclus      2.3-2   2020-01-08 [1] CRAN (R 4.1.2)
prettyunits   1.1.1   2020-01-24 [1] CRAN (R 4.1.0)
processx      3.5.3   2022-03-25 [1] CRAN (R 4.1.3)
ps            1.6.0   2021-02-28 [1] CRAN (R 4.1.0)
purrr       * 0.3.4   2020-04-17 [1] CRAN (R 4.1.0)
R6            2.5.1   2021-08-19 [1] CRAN (R 4.1.1)
Rcpp          1.0.8.3 2022-03-17 [1] CRAN (R 4.1.3)
readr       * 2.1.2   2022-01-30 [1] CRAN (R 4.1.3)
readxl        1.3.1   2019-03-13 [1] CRAN (R 4.1.0)
remotes       2.4.2   2021-11-30 [1] CRAN (R 4.1.2)
reprex        2.0.1   2021-08-05 [1] CRAN (R 4.1.1)
rlang         1.0.2   2022-03-04 [1] CRAN (R 4.1.3)
```

```
rmarkdown      2.13    2022-03-10 [1] CRAN (R 4.1.3)
robustbase     0.93-9  2021-09-27 [1] CRAN (R 4.1.1)
rprojroot      2.0.3   2022-04-02 [1] CRAN (R 4.1.3)
rstudioapi     0.13    2020-11-12 [1] CRAN (R 4.1.0)
rvest          1.0.2   2021-10-16 [1] CRAN (R 4.1.1)
scales         1.1.1   2020-05-11 [1] CRAN (R 4.1.1)
sessioninfo    1.2.2   2021-12-06 [1] CRAN (R 4.1.3)
stringi        1.7.6   2021-11-29 [1] CRAN (R 4.1.2)
stringr      * 1.4.0   2019-02-10 [1] CRAN (R 4.1.0)
table1       * 1.4.2   2021-06-06 [1] CRAN (R 4.1.2)
testthat       3.1.3   2022-03-29 [1] CRAN (R 4.1.3)
tibble       * 3.1.6   2021-11-07 [1] CRAN (R 4.1.2)
tidyr        * 1.2.0   2022-02-01 [1] CRAN (R 4.1.3)
tidyselect     1.1.2   2022-02-21 [1] CRAN (R 4.1.3)
tidyverse    * 1.3.1   2021-04-15 [1] CRAN (R 4.1.1)
tzdb           0.3.0   2022-03-28 [1] CRAN (R 4.1.3)
usethis      * 2.1.5   2021-12-09 [1] CRAN (R 4.1.3)
utf8           1.2.2   2021-07-24 [1] CRAN (R 4.1.1)
vctrs          0.4.0   2022-03-30 [1] CRAN (R 4.1.3)
withr          2.5.0   2022-03-03 [1] CRAN (R 4.1.3)
xfun           0.30    2022-03-02 [1] CRAN (R 4.1.3)
xml2           1.3.3   2021-11-30 [1] CRAN (R 4.1.3)
yaml           2.2.1   2020-02-01 [1] CRAN (R 4.1.0)

[1] C:/Users/mdlav/Documents/R/win-library/4.1
[2] C:/Program Files/R/R-4.1.0/library

------------------------------------------------------------------------------
```

REFERENCES

10 Ajorlou, S., Shams, I., & Yang, K. (2015). An analytics approach to designing patient centered medical homes. *Health Care Management Science*, *18*(1), 3–18. http://doi.org/10.1007/s10729-014-9287-x

Arndt, B., Tuan, W.-J., White, J., & Schumacher, J. (2014). Panel workload assessment in US primary care: Accounting for nonface-to-face panel management activities. *The Journal of the American Board of Family Medicine*, *27*(4), 530–537. http://doi.org/10.3122/jabfm.2014.04.130236

Bodenheimer, T., Ghorob, A., Willard-Grace, R., & Grumbach, K. (2014). The 10 building blocks of high-performing primary care. *The Annals of Family Medicine*, *12*(2), 166–171. http://doi.org/10.1370/afm.1616

Chung, S., Eaton, L. J., & Luft, H. S. (2012). Standardizing primary care physician panels: Is age and sex good enough? *The American Journal of Managed Care*, *18*(7), e262–e268.

Demirtas, H., & Gao, R. (2020). Mixed data generation packages and related computational tools in r. *Communications in Statistics - Simulation and Computation*, *0*(0), 1–44. http://doi.org/10.1080/03610918.2020.1745841

Demirtas, Hakan, Hedeker, D., & Mermelstein, R. J. (2012). Simulation of massive public health data by power polynomials. *Statistics in Medicine*, *31*(27), 3337–3346. http://doi.org/10.1002/sim.5362

Foss, A. H., & Markatou, M. (2018). Kamila: Clustering mixed-type data in r and hadoop. *Journal of Statistical Software, Articles*, *83*(13), 1–44. http://doi.org/10.18637/jss.v083.i13

Foss, A. H., Markatou, M., & Ray, B. (2019). Distance metrics and clustering methods for mixed-type data. *International Statistical Review*, *87*(1), 80–109. http://doi.org/https://doi.org/10.1111/insr.12274

Foss, A., Markatou, M., Ray, B., & Heching, A. (2016). A semiparametric method for clustering mixed data. *Machine Learning*, *105*(3), 419–458. http://doi.org/10.1007/s10994-016-5575-7

Gower, J. C. (1971). A general coefficient of similarity and some of its properties. *Biometrics*, *27*(4), 857–871. Retrieved from http://www.jstor.org/stable/2528823

Grumbach, K., & Olayiwola, J. N. (2015). Patient empanelment: The importance of understanding who is at home in the medical home. *The Journal of the American Board of Family Medicine*, *28*(2), 170–172. http://doi.org/10.3122/jabfm.2015.02.150011

Hartigan, J. A., & Wong, M. A. (1979). A k-means clustering algorithm. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *28*(1), 100–108. http://doi.org/https://doi.org/10.2307/2346830

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning*. Springer. http://doi.org/https://doi.org/10.1007/978-0-387-84858-7

Hennig, C. (2020). *Fpc: Flexible procedures for clustering*. Retrieved from https://CRAN.R-project.org/package=fpc

Hennig, C., & Liao, T. F. (2013). How to find an appropriate clustering for mixed-type variables with application to socio-economic stratification. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *62*(3), 309–369. http://doi.org/https://doi.org/10.1111/j.1467-9876.2012.01066.x

Hunt, L., & Jorgensen, M. (1999). Theory & methods: Mixture model clustering using the MULTIMIX program. *Australian & New Zealand Journal of Statistics*, *41*(2), 154–171. http://doi.org/https://doi.org/10.1111/1467-842X.00071

Hunt, L., & Jorgensen, M. (2011). Clustering mixed data. *WIREs Data Mining and Knowledge Discovery*, *1*(4), 352–361. http://doi.org/https://doi.org/10.1002/widm.33

Inan, G., Demirtas, H., & Gao, R. (2021). *PoisBinOrd: Data generation with poisson, binary and ordinal components*. Retrieved from https://CRAN.R-project.org/package=PoisBinOrd

Kamnetz, S., Trowbridge, E., Lochner, J., Koslov, S., & Pandhi, N. (2018). A simple framework for weighting panels across primary care disciplines: Findings from a large US multidisciplinary group practice. *Quality Management in Healthcare*, *27*(4). Retrieved from https://journals.lww.com/qmhcjournal/Fulltext/2018/10000/A_Simple_Framework_for_Weighting_Panels_Across.1.aspx

Kaufman, L., & Rousseuw, P. J. (1990). *Finding groups in data: An introduction to cluster analysis*. John Wiley & Sons. http://doi.org/https://doi.org/10.1002/9780470316801

Kivlahan, C., Pellegrino, K., Grumbach, K., Skootsky, S. A., Raja, N., Gupta, R., … Todoki, E. (2017). *Calculating primary care panel size*. University of California Center for Health Quality; Innovation. Retrieved from https://www.ucop.edu/uc-health/_files/uch-chqi-white-paper-panel-size.pdf

Krijthe, J. H. (2015). *Rtsne: T-distributed stochastic neighbor embedding using barnes-hut implementation*. Retrieved from https://github.com/jkrijthe/Rtsne

Krzanowski, W. J. (1993). The location model for mixtures of categorical and continuous variables. *Journal of Classification*, *10*(1), 25–49. http://doi.org/10.1007/BF02638452

Lavallee, M. (2021). Estimating weighted panel sizes for primary care providers: An assessment of clustering and novel methods of panel size estimation on electronic medical records. Dissertation Proposal.

Lavallee, M., Evans, L., & Black, A. (2021). *Capr: Cohort definition application programming in r*. Retrieved from https://github.com/OHDSI/Capr

Li, H., Chen, R., Nguyen, H., Chung, Y.-C., Gao, R., & Demirtas, H. (2020). *RNGforGPD: Random number generation for generalized poisson distribution*. Retrieved from https://CRAN.R-project.org/package=RNGforGPD

Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, *15*(12), 550. http://doi.org/10.1186/s13059-014-0550-8

Manning, C. D., Raghavan, P., & Schutze, H. (2009). An introduction to information retrieval. Cambridge University Press.

McLachlan, G., & Peel, D. (2000). *Finite mixture models*. John Wiley & Sons. http://doi.org/https://doi.org/10.1002/0471721182

McNicholas, P. D. (2016). *Mixture model-based classification*. Chapman; Hall/CRC. http://doi.org/https://doi.org/10.1201/9781315373577

Murray, M., Davies, M., & Boushon, B. (2007). Panel size: How many patients can one doctor manage? *Family Practice Management*, *14*(4), 44–51.

Peterson, L. E., Cochrane, A., Bazemore, A., Baxley, E., & Phillips, R. L. (2015). Only one third of family physicians can estimate their patient panel size. *The Journal of the American Board of Family Medicine*, *28*(2), 173–174. http://doi.org/10.3122/jabfm.2015.02.140276

Potts, B., Adams, R., & Spadin, M. (2011). Sustaining primary care practice: A model to calculate disease burden and adjust panel size. *The Permanente Journal*, *15*(1), 53–56. http://doi.org/10.7812/tpp/10-077

R Core Team. (2021). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from `https://www.R-project.org/`

Rajkomar, A., Yim, J. W. L., Grumbach, K., & Parekh, A. (2016). Weighting primary care patient panel size: A novel electronic health record-derived measure using machine learning. *JMIR Med Inform*, *4*(4), e29. http://doi.org/10.2196/medinform.6530

Robinson, D., Hayes, A., & Couch, S. (2021). *Broom: Convert statistical objects into tidy tibbles*. Retrieved from `https://CRAN.R-project.org/package=broom`

Schuemie, M., Suchard, M., Ryan, P., Reps, J., & Sena, A. (2021). *FeatureExtraction: Generating features for a cohort*. Retrieved from `https://github.com/OHDSI/FeatureExtraction`

Schwarz, G. (1978). Estimating the Dimension of a Model. *The Annals of Statistics*, *6*(2), 461–464. http://doi.org/10.1214/aos/1176344136

Sciennces, O. H. D., & Informatics. (2021). *The book of OHDSI*. Retrieved from `https://ohdsi.github.io/TheBookOfOhdsi/`

Scrucca, L., Fop, M., Murphy, T. B., & Raftery, A. E. (2016). Mclust 5: Clustering, classification and density estimation using gaussian finite mixture models. *The R Journal*, *8*(1), 289–317. Retrieved from `https://pubmed.ncbi.nlm.nih.gov/27818791`

Tibshirani, R., & Walther, G. (2005). Cluster validation by prediction strength. *Journal of Computational and Graphical Statistics*, *14*(3), 511–528. http://doi.org/10.1198/106186005X59243

Tibshirani, R., Walther, G., & Hastie, T. (2001). Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *63*(2), 411–423. http://doi.org/https://doi.org/10.1111/1467-9868.00293

van der Maaten, L. J. P., & Hinton, G. E. (2008). Visualizing high-dimensional data using t-SNE. *Journal of Machine Learning Research*, *9*, 2579–2605.

Wagner, S., & Wagner, D. (2007, January). Comparing clusterings - an overview. https://i11www.iti.kit.edu/extra/publications/ww-cco-06.pdf.

Walonoski, J., Kramer, M., Nichols, J., Quina, A., Moesel, C., Hall, D., … McLachlan, S. (2017). Synthea: An approach, method, and software mechanism for generating synthetic patients and the synthetic electronic health care record. *Journal of the American Medical Informatics Association*, *25*(3), 230–238. http://doi.org/10.1093/jamia/ocx079

Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. Springer-Verlag New York. Retrieved from https://ggplot2.tidyverse.org