MODEL-BASED DATA MINING METHODS FOR IDENTIFYING PATTERNS IN MEDICAL AND HEALTH DATA

A Thesis
Presented to
The Academic Faculty

by

Ross P. Hilton

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in the
School of Industrial and Systems Engineering

Georgia Institute of Technology
December 2015

MODEL-BASED DATA MINING METHODS FOR IDENTIFYING PATTERNS IN MEDICAL AND HEALTH DATA

Approved by:

Professor Nicoleta Serban, Advisor
School of Industrial and Systems Engineering
Georgia Institute of Technology

Professor Julie Swann
School of Industrial and Systems Engineering
Georgia Institute of Technology

Professor Brani Vidakovic
School of Industrial and Systems Engineering
Georgia Institute of Technology

Professor Mark Braunstein
College of Computing
Georgia Institute of Technology

Professor Paul Griffin
School of Industrial and Systems Engineering
Georgia Institute of Technology

Date Approved: 7 October 2015

ACKNOWLEDGEMENTS

I have many people to thank who have helped me to the complete this thesis. I would like to thank Dr. Serban, without whose advisement none of the research in this thesis would be possible. To my committee members, Dr. Swann, Dr. Vidakovic, Dr. Braunstein, and Dr. Griffin, thank you for your comments and criticisms that have helped to improve this thesis.

Thank you to Matt Sanders and Paul Diedrich who assisted with the data safeguards and information technology infrastructure, and who were always willing to help with other technology issues at a moment's notice. A special thank you to Richard Starr and Richard Zheng, who have helped me save enormous amounts of time over the last two years by lending their vast database and SQL knowledge.

To my colleagues both current and past including Matt, Christine, Kevin, Richard, Dr. Goldsman, Dr. Parker, Pam Morrison, just to name a few, thanks for the support over the years at ISyE. To Mr. Shubert, who initially instilled in me a love for mathematics, thanks so much. Thank you to Pat and Brian for their steadfast encouragement and friendship, and for supporting me throughout this whole process. To the rugby community, Emory, Life, and Wake Rugby teams, I owe an enormous thank you for the diversion of a lifetime and for the lessons I learned from such a wonderful game.

I have to thank my parents and siblings for encouraging me and never doubting me. Finally, it is appropriate that if my master's thesis was dedicated to my parents, then this thesis should be dedicated to my amazing wife, Miranda. Her words over the last few years, listening to me, supporting me, and sometimes shoving me in the right direction daily saw me through the ups and downs of my graduate studies. Thank you so much.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

SUMMARY

Every day humans and machines are responsible for the creation of massive amounts of data. Alongside the growth of these data banks, a new field of study, *data science*, has emerged. The central role of data science is to infer knowledge on the data in the form of models and estimates employing methods at the intersection of computer science, data mining, mathematics, and statistics. In this thesis we provide statistical and model-based data mining methods for pattern detection with applications to biomedical and healthcare data sets. In particular, we examine applications in costly acute or chronic disease management. Health data are extremely varied: at the macro-level, one can examine the healthcare utilization of millions of patients in the insurance systems like Medicare and Medicaid, while at the micro-level, a single snapshot from a medical imaging device may be used to diagnose cancerous cells in the body. In all, statisticians can contribute methods that extract structure from large, noisy data.

In Chapter II, we consider NMR experiments in which we seek to locate and de-mix smooth, yet highly localized components in a noisy two-dimensional signal. By using wavelet-based methods we are able to separate components from the noisy background, as well as from other neighboring components. In Chapter III, we pilot methods for identifying profiles of patient utilization of the healthcare system from large, highly-sensitive, patient-level data. We combine model-based data mining methods with clustering analysis in order to extract longitudinal utilization profiles. We transform these profiles into simple visual displays that can inform policy decisions and quantify the potential cost savings of interventions that improve adherence to recommended care guidelines. In Chapter IV, we

propose new methods integrating survival analysis models and clustering analysis to profile patient-level utilization behaviors while controlling for variations in the population's demographic and healthcare characteristics and explaining variations in utilization due to different state-based Medicaid programs, as well as access and urbanicity measures.

CHAPTER I

INTRODUCTION

Every day humans and machines are responsible for the creation of massive amounts of data. With every electronic transaction, social media interaction, business or healthcare event, data are recorded and stored for use in future analysis at the rate of 2.5 quintillion bytes per day (50). Alongside the growth of these data banks, a new field of study, *data science*, has emerged. Data scientists are present in academia, government, and industry, wherever entities recognize the potential value in extracting knowledge from often complex and large data sets that may have previously been unused.

The central role of data science is to infer knowledge on the data in the form of models and estimates employing methods at the intersection of computer science, data mining, mathematics, and statistics (80). The National Research Council states that it is the role of statisticians to justify the inferential leap from data to knowledge. However, traditional statistical analysis will not suffice when studying data sets of extraordinary size and/or complexity. It is the goal of this thesis to provide and inspire new methodologies that are both computationally attractive and statistically sound, with a potential impact beyond the applications we present.

In this thesis we focus on applications to biomedical and healthcare data sets. In particular, we examine applications in costly acute or chronic disease management. Health data are extremely varied: at the macro-level, one can examine the healthcare utilization of millions of patients in the insurance systems like Medicare and Medicaid, while at the micro-level, a single snapshot from a medical imaging device may be used to diagnose cancerous cells in the body. In all, statisticians can contribute methods that extract structure from large, noisy data. Traditionally, statisticians derive inference in the form of

cause-and-effect relationships (e.g. regression or network analysis), the drivers of variation in a phenomenon (e.g. principal component analysis), and grouping objects based on some notion of similarity (cluster analysis). In this thesis we combine traditional statistical inference with model-based data mining techniques including nonparametric regression, stochastic modelling, and clustering analysis to perform meaningful analysis and provide sound, principled methods for translating data into knowledge.

In Chapter II we present methods for identifying peaks (*components*) in nuclear magnetic resonance (NMR) data for biomolecular structure determination. In particular, we use the continuous wavelet transform together with the theory of convolutions, to identify smooth, yet highly-localized components in a noisy signal without the need for denoising or the inverse wavelet transform. The methods in this chapter are applicable to any experiment where the target features to be identified are assumed to have a smooth, symmetric shape. In our application of interest, the discovery of true components can lead to the correct structure determination. Using NMR spectroscopy to correctly identify protein structures from medical images can inform the use of proactive treatments of tumors and cancers in the human body.

In Chapter III we present methods for deriving inferences on longitudinal patient healthcare utilization profiles from large, highly-sensitive medical claims data, and quantify the cost-saving effects of interventions that bend patient-level utilization towards more effective practices. We pilot our methods using the CMS Medicaid Analytic Extract (MAX) data for five years (2005-2009) from two states, Georgia and North Carolina. By considering neighboring states with similar pediatric populations, we are able to determine the effects of factors such as different managed care organizations or geography on healthcare utilization and expenditure. We chose pediatric asthma as the health condition of interest because it is a common chronic childhood condition, allowing for the investigation of potentially costly patient-level utilization behaviors. Furthermore, by considering the Medicaid population, we can determine interventions to prevent costly and ineffective treatments for a

subpopulation most susceptible to disparate healthcare utilization.

In this chapter we assume that a patient's utilization sequence follows a Markov renewal process. We employ a single-linkage tree to search for the cluster division that produces a maximal increase in the Bayesian Information Criterion (BIC) score at each step. We utilize the expectation-maximization (EM) algorithm to simultaneously determine model parameters, perform cluster membership, and infer provider networks. Additionally, we analyze "what-if" scenarios for different case studies on the effects of potential changes to healthcare utilization. We provide a case study where we bend the realized utilization patterns to follow the recommended care guidelines for pediatric asthma and determine the new expected cost per patient.

In Chapter IV, we introduce new models for integrating survival analysis and clustering analysis to profile patient-based utilization behaviors. In this study we add patient histories from four more states to the study sample in Chapter III, and integrate advanced statistical modelling, estimation, and cluster analysis procedures in order to address some of the limitations of the study in Chapter III. In particular, we address the effects of censoring, the multivariate counting process nature of the patient utilization histories, and incorporate basic demographic, geographic, and health characteristics of the patients in the clustering algorithm. Survival analysis provides statistical methods that allow for the inclusion of possibly censored interarrival times between events and the study of the effects of demographic, geographic, and health-related covariates in the model. The outcomes of this study are a population-level mixture model that identifies patient cluster membership, while controlling for the effects of "fixed" variables such as age, race, and health status, for instance, and estimates the effects of access measures, urbanicity, and state-based programs on healthcare utilization.

CHAPTER II

THEORETICAL LIMITS OF COMPONENT IDENTIFICATION IN A SEPARABLE NONLINEAR LEAST SQUARES PROBLEM

In this chapter we provide theoretical insights into component identification in a separable nonlinear least squares problem in which the model is a linear combination of nonlinear functions (called *components* in this chapter). Within this research, we assume that the number of components is unknown. The objective of this chapter is to understand the limits of component discovery under the assumed model. We focus on two aspects. One is sensitivity analysis referring to the ability of separating regression components from noise. The second is resolution analysis referring to the ability of de-mixing components that have similar location parameters.We use a wavelet transformation that allows the researcher to zoom in at different levels of detail in the observed data. We further apply these theoretical insights to provide a road map on how to detect components in more realistic settings such as a two-dimensional Nuclear Magnetic Resonance (NMR) experiment for protein structure determination.

2.1 Introduction

The separable nonlinear least squares problem encompasses many variations from its general form of a linear combination of nonlinear functions. In this chapter, we cast our methodology within this framework but focus on a particular form of this regression model. Specifically, the regression is a linear combination of location-scale, highly localized, nonlinear components, where the shape of the components is a unimodal symmetric function specified up to the unknown location and scale parameters. With two components, the

model becomes

$$Y_i = A_1 s\left(x_i; \omega_1, \tau_1\right) + A_2 s\left(x_i; \omega_2, \tau_2\right) + \sigma\epsilon_i, \qquad i = 1, \ldots, M$$

where A_1 and A_2 are the separable parameters, ω_1 and ω_2 are the location parameters, τ_1 and τ_2 are the scale coefficients, and $s(x)$ is the shape function. The separable parameter of each component plays the role of an amplitude. The location parameter represents the center or the mode of the component and the scale parameter is a measure of the width of the component. The values x_i are assumed to be observed over a grid within a d-dimensional space. The error terms ϵ_i's are assumed independent and identically distributed. This model can be extended to $L \geq 2$ components all sharing the same shape function.

In most of the existing research on separable nonlinear regression, the number of components is assumed known, an assumption that does not hold in many applications. Moreover, regardless of the estimation algorithm, variable projection or alternating two-step approach, the estimation of such models relies on iterative algorithms requiring the input of good initial estimates (38). To obtain initial estimates, the number of components needs to be estimated accurately. The presence of a large number of false positives or falsely discovered components could lead to an ill-conditioned estimation problem whereas a large number of false negatives or undiscovered components could result in a incomplete understanding of the underlying science behind the data. False positive components may arise due to high variance noise or other artifacts in the data. False negative components could arise when the number of components is large, and the distance between the location parameters of some components ($\theta = \|\omega_1 - \omega_2\|$) is small such that some components will mix partially or totally.

In this chapter, we provide theoretical and methodological insights in obtaining an estimate of the number of components along with initial estimates of the location parameters without the need of estimating the width or separable parameters. Particularly, we focus on two aspects, separating components from the noisy background and identifying components that are mixed. Separating components from the noisy background is a widely

researched problem; methodological approaches and heuristic algorithms have been introduced in numerous fields particularly in biomedical sciences, e.g., mass spectroscopy, Nuclear Magnetic Resonance biomolecular structure determination and tumor-spread description using CT scan (40; 54; 101; 122), although often within the framework of detecting features from a noisy image. The problem of identifying mixed components has been recently introduced within the framework of a hypothesis testing problem using a penalized regression test statistic (102).

The objective of this study is not to yet introduce other such methods, but to provide an understanding of the theoretical limits to what can be detectable at given levels of *sensitivity* (measured as the signal-to-noise ratio) and *resolution* (often measured by the separation between components). Two questions that we address are:

• Sensitivity Limit: What is the amplitude level at which we can still detect a signal component given the noise level specified by σ?

Figure 1 shows two examples of components, with large (left plot) and small (right plot) amplitudes. The high amplitude component (often called *peak*) stands out clearly whereas the low amplitude component can be confounded with a spike in the noise. A common approach to identify low amplitude components or peaks is to compare its highest value to some threshold that is a function of the error variance. However, this approach has limited power as it only uses information about one observed value of the component disregarding neighboring values. Moreover, a low amplitude component may simply be too similar to the noise behavior that may not be distinguishable from noise spikes. However, if we could transform the data in such a way we can differentiate between the behavior of noise and the behavior of components, then we may be able to detect low amplitude components within some measurable limits of "separation" of signal from the noise.

• Resolution Limit: What is the smallest distance between the location of two components given the widths of the components specified by τ_1 and τ_2?

Figure 2 shows two examples of mixed components. These two examples are different

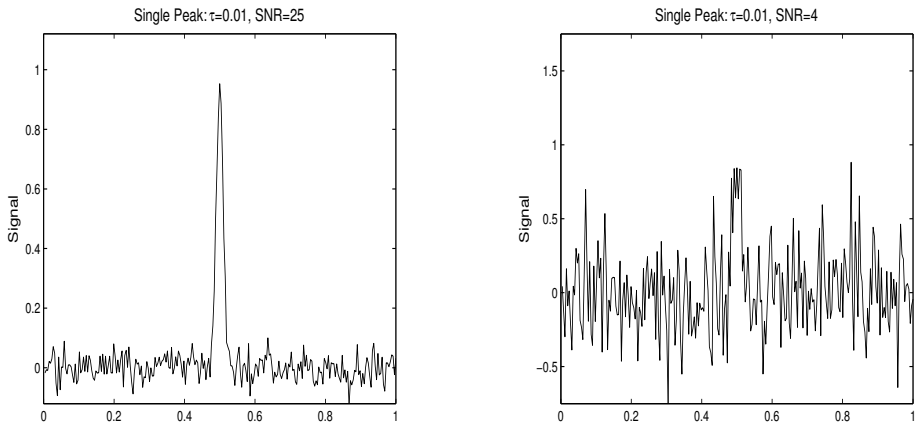

Figure 1: Gaussian-shaped components with different amplitudes.

in that they show partial mixing, with two components mapping into two modes (left plot), and total mixing, with two components forming one mode (right plot). The emphasis of this study is on correctly discovering the location of the components in the wavelet domain, as opposed to the modes in the signal domain. Once the distance between the location of the components is smaller than the sum of their widths, the two components will merge to become unimodal as opposed to bimodal as in the left plot. Although the two components in the right plot are overlapped into one mode or local maximum, can we apply a data transformation that allows separation of the two components into two distinct modes? When do we reach the limits of "de-mixing" components?

A natural choice for data transformation that allows both separation of signal from the noise and identification of mixed components is the wavelet transform. The majority of existing research in wavelet transform analysis focuses on two areas: signal denoising and feature detection (29; 30; 66; 116). In this chapter, our primary objective is in feature detection. Wavelet transform analysis has been applied to detection of various features, including edges, corners, or blobs, depending on the context of the data (66; 117). The components in the assumed regression model are smooth, referred to as *blobs* in (60). However, in our application of interest the widths of the components are assumed to be sufficiently small, such that hundreds of components can coexist within the unit square.

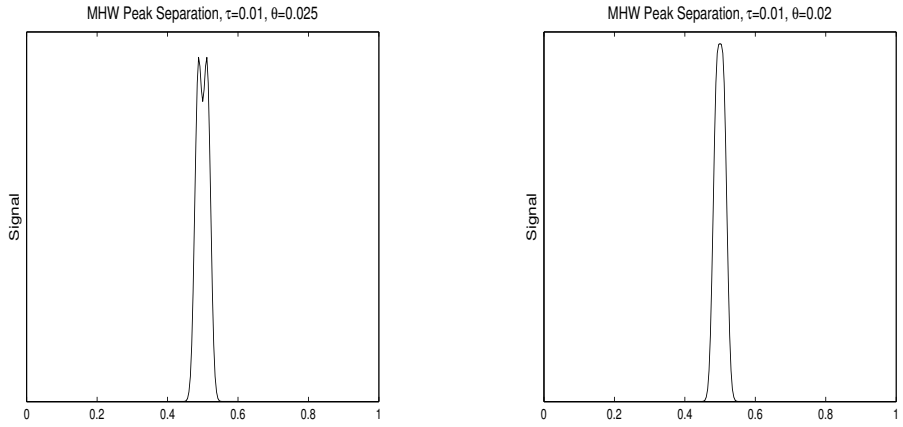

Figure 2: Two Gaussian-shaped components with equal scale parameter $\tau = .01$ and distance $\theta = 0.025$ (left) and $\theta = 0.02$ (right).

We make use of two important properties of wavelet analysis. One is the ability to 'zoom in' at different levels of detail in the data using a multiresolution decomposition. This property is particularly useful in separating components from the noise because the noise behaves differently than the smooth, yet localized components at different 'zooming' resolutions. The second important property is that the wavelet coefficients are a measure of the similarity of the wavelet and the signal to be analyzed. Thus, the choice of the wavelet basis is critical in modulating the signal through the wavelet transform such that detailed information (e.g. modality) will be clearly amplified and detailed features detected. This property is particularly useful in de-mixing components that have similar location parameters.

In Figure 3, we compare the application of one commonly used wavelet basis in signal denoising, the discrete wavelet transform (DWT) with the Daubechies basis (4 vanishing moments in this example), and our choice of wavelet basis, the continuous wavelet transform (CWT) with the Mexican Hat basis. The orthogonal property of the DWT yields the sparsest representation possible in the wavelet domain, which is desirable when thresholding the wavelet coefficients for achieving locally adaptive denoising, but is not particularly useful when trying to determine the location of multiple components. The sparse property

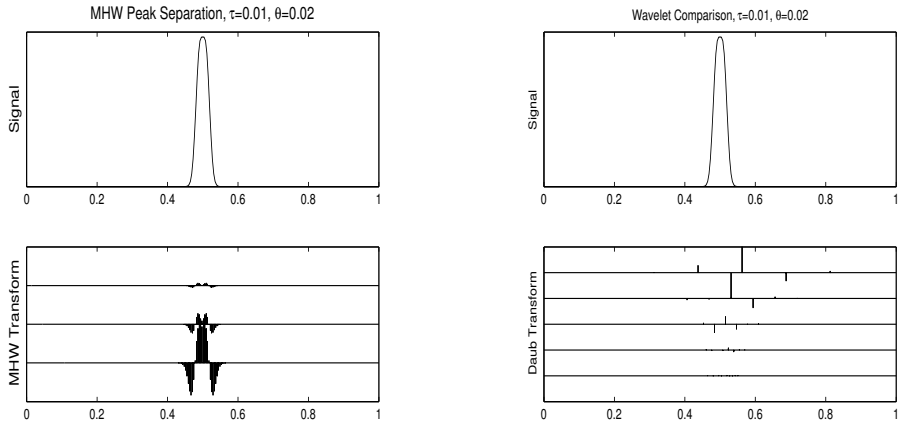

Figure 3: Two Gaussian-shaped components with parameters $\tau = .01$ and $\theta = .02$. The signal is plotted on top. The three finest scales of the CWT with Mexican Hat wavelet basis is on the bottom of the display on the left and the DWT with the Daubechies wavelet basis (4 vanishing moments) is on the bottom of the display on the right.

of the Daubechies DWT coefficients does not effectively locate the center of the components since the sampling density of the wavelet coefficients is cut in half at each scale. On the other hand, although the Mexican Hat CWT coefficients are not sparse, this particular transform allows locating and separating components as we will illustrate further in this chapter.

We derive both sensitivity and resolution limits for detecting components in the assumed model using the Mexican Hat CWT. We further apply these theoretical insights to provide a road map on how to detect features in more realistic settings, particularly, when the number of components is large and when the signal-to-noise ratio is low enough such that some components are not clearly distinguishable from the noisy background. We will therefore develop an algorithm that takes information from all wavelet transform scales to perform both feature detection and separation.

We will investigate our theoretical and methodological insights in a real case application after building a case for our methods within a simulation study. More specifically, we are interested in detecting "peaks" in a two-dimensional nuclear magnetic resonance (NMR) data that was generated to aid in the process of biomolecule structure identification. The

data generated from NMR experiments are generally modeled using a separable nonlinear model as described in this introduction, where the shape function is an amplitude-Fscaled Cauchy probability density function called *Lorentzian* (48; 101).

This chapter is organized as follows. Section 2.2 introduces the framework for feature detection using the Mexican Hat Wavelet (MHW) transform. Section 2.3 provides theoretical insights for the sensitivity and resolution limits of feature detection under the assumption that the shape function of the components is Gaussian while discussing the extension of these results for other symmetric unimodal shape functions. Section 2.4 provides an algorithm for feature detection while using these theoretical results along with a simulation study for evaluating the performance of the algorithm. Our application is presented in Section 2.5. Some technical details and additional simulation studies are deferred to the Appendices.

2.2 Mexican Hat Wavelet Analysis for Feature Detection

In this section we discuss the choice of the MHW for the CWT drawing inspiration from existing research. Upon arriving at our choice for the MHW, we begin with a basic introduction of the wavelet transform and describe the behavior of the coefficients from an intuitive standpoint.

2.2.1 Background

In this chapter we draw inspiration from many research sources on feature detection including image processing, discrete wavelet theory and continuous wavelet theory for both sharp and smooth signal features. In image processing, features can take many forms including edges, corners, or blobs, depending on the context of the problem. The type of changes that we are interested in here are blob-type features. In (60) the author uses a scale-space representation to extract such features and introduces the convolution of the image with the 2nd derivative of the Gaussian function to determine amplitude and size of smooth image structures. This convolution idea motivates our selection for the wavelet basis as we expand

in this section.

Mallat and Hwang formalized many concepts such as Lipschitz exponents and the 'cone of influence' as well as their relation to the wavelet transform (66). We use many of these ideas, however, because the primary application of their work was for edge detection we must approach the problem differently. Specifically, in (66), the authors relied on a finite Lipschitz exponent, or singular signal, whereas the Lipschitz exponent of a Gaussian-shaped component is unbounded (see Appendix A.1).

Finally, Nenadic and Burdick proposed the use of the CWT with biorthogonal wavelets for 'spike' detection using a wavelet that approximately matches the signal shape while restricting signal analysis to a set of scales and translations to identify the location of spikes based on prior knowledge of the data (81). They state that the wavelet coefficients are, in fact, a measure of similarity between the wavelet and the signal. Lindeberg states that the values of the convolution will be maximized when the scale of the wavelet approximately matches the width of the signal (62), see Figure 4 for a demonstration of this property. Using these ideas on scale selection, we consider appropriate scales in order to capture smooth components of varying widths.

With these facts we choose the MHW because of its analytic and statistical properties and its similarity in shape to the Gaussian function as shown in Figure 5.

2.2.2 Preliminary Definitions

A wavelet, ψ, is a function that satisfies

$$\int_0^\infty \frac{|\hat{\psi}(\omega)|^2}{\omega} d\omega = \int_{-\infty}^0 \frac{|\hat{\psi}(\omega)|^2}{\omega} d\omega < \infty,$$

where $\hat{\psi}$ denotes the Fourier transform of the wavelet function (42). If this condition is satisfied we have that

$$\int_{-\infty}^{\infty} \psi(u) du = 0.$$

ψ is typically referred to as the *mother wavelet*. The wavelet function is well localized in both the time and frequency domains, i.e. $\psi \in L^2(R)$ and integrates to 0. It is not

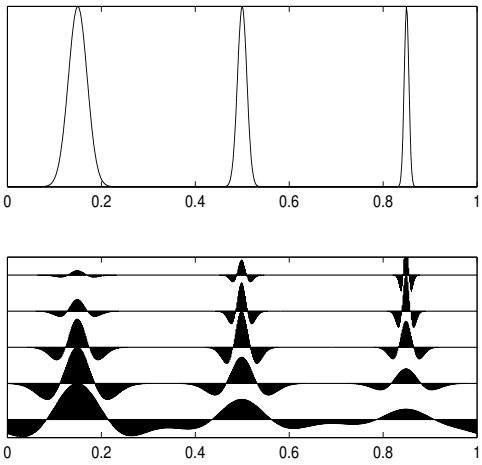

Figure 4: The MHW transformation of three components of different widths. The wavelet coefficients are normalized at each scale so that the maximum value of the coefficients is 1. Notice the change in relative maximum of coefficients from the widest component to the narrowest as the scales decrease.

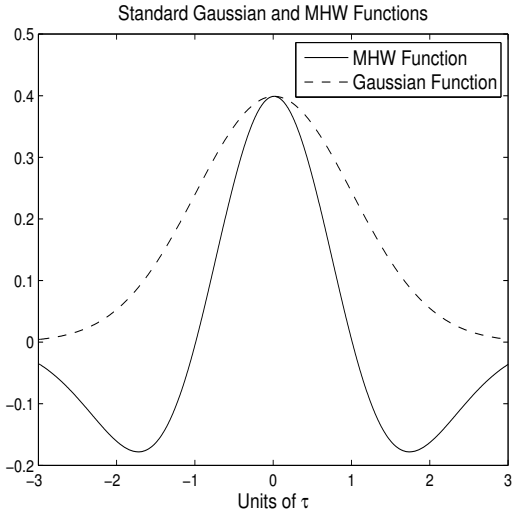

Figure 5: Comparison of the Gaussian and MHW functions.

sufficient for the wavelet simply to integrate to zero, it must also be bounded in L^2 space, implying that it has finite 'energy'. It is also a desirable property for the wavelet to have small support, thus allowing for analysis of a small portion of the signal at a time.

In order to determine information at various scales of the signal by 'zooming in' on the signal, we dilate the wavelet at scale s by:

$$\psi_s(x) = \frac{1}{s}\psi\left(\frac{x}{s}\right).$$

A *translation* of the wavelet is $\psi_{s,t}(x) = \psi_s(x-t)$. From the scaling and translation formulas we are able to determine a family of wavelet functions, $\psi_{s,t}(x)$, for $s, t \in R^+$.

The wavelet transform of a one dimensional signal, f, at scale s is the convolution of the signal with the dilated wavelet function:

$$Wf(s,x) = f \star \psi_s(x) = \int_{-\infty}^{\infty} f(u)\psi_s(x-u)du.$$

Since we can only observe the data in a discrete sample we are limited to a discrete set of translations. Let $t \in T$, where $T = \{0, 1, \cdots, N-1\}$, where $N = 2^K$ is the number of sampling points. Then the set of translations is limited to $\psi_s(x-t)$. The wavelet transform $Wf(s,x)$ is the convolution of the signal and the wavelet in discrete time, therefore,

$$Wf(s,x_i) = f \star \psi_s(x_i) = \sum_{t\in T} < f[t], \psi_s(x_i - t) > .$$

When deriving the theoretical properties of the MHW, we assume that we have a continuous signal. However, the results apply for a discrete sampling also.

The MHW is a member of the Hermitian family of continuous wavelets:

$$\psi_n(x) = c_n H_n(x)\phi(x),$$

where c_n is a normalization constant, $H_n(x)$ is the nth order Hermite polynomial, and $\phi(x)$ is the standard normal density. The wavelet is normalized so that $||\psi||_2 = 1$. The MHW is defined as follows:

$$\psi_{MHW}(x) \stackrel{def}{=} \psi(x) = -c_2\frac{d^2}{dx^2}\phi(x) = c_2(1-x^2)\phi(x).$$

Despite the fact that the wavelet has infinite support, the values of the MHW function decay to 0 exponentially. As stated before, this is an important characteristic of wavelet transform because it allows analysis of small sections of the signal at a time.

2.2.3 Preliminary Insights on the Wavelet Coefficients

For symmetrically-shaped components as assumed in our model, it is intuitive to only analyze the finest scales. This is due to the behavior of the components or blobs across multiple resolution levels. Because the Lipschitz exponent is unbounded for Gaussian-shaped components, we can expect that the wavelet coefficients should have maxima present across all scales, while the coefficients corresponding to noise will be smoothed out at coarser scales and will have multiple sporadic spikes appearing at fine scales. Furthermore, the convolution of the MHW with a Gaussian component will have large amplitude positive coefficients at the component center and will be followed with high amplitude negative valleys.

2.3 Theoretical Results

2.3.1 Resolution Analysis

One-Dimensional Signal. By assuming that the data follow a model in which the regression function is a linear combination of Gaussian components, we are able to derive, up to scale, the wavelet coefficients in a noise-free environment. In doing so, we provide a theoretical basis for the resolution limits later in this section. We begin with the simplest setting, that of a continuous one-dimensional, noise free model,

$$f(x) = \sum_{l=1}^{L} A_l s\left(x; \omega_l, \tau_l\right) = \sum_{l=1}^{L} A_l \exp\left\{-\frac{1}{2}\left(\frac{x-\omega_l}{\tau_l}\right)^2\right\}.$$

By assuming the components to be Gaussian-shaped, we derive the closed-form expression of the wavelet coefficients using the proof of (13) on the convolution of Gaussian functions

and properties of convolutions. The wavelet coefficients are proportional to (up to scale)

$$Wf(s,x) \propto -\sum_{l=1}^{L} \frac{A_l \tau_l}{\sqrt{\tau_l^2 + s^{-2}}} H_2\left(\frac{x-\omega_l}{\sqrt{\tau_l^2+s^{-2}}}\right) \exp\left\{-\frac{1}{2}\left(\frac{x-\omega_l}{\sqrt{\tau_l^2+s^{-2}}}\right)^2\right\}.$$

Furthermore, using the explicit form of the wavelet coefficients, we are able to analytically derive the theoretical limits for component separation. As shown in Figure 3, the MHW transform can distinguish mixed components in the wavelet domain even when they are unimodal in the signal domain. However, there is a limit to this de-mixing; the theoretical limit for the distance between the mixed components depends on the widths of the components and it is different across scales as provided in the next theorem. The proof relies heavily on the roots of the Hermitian polynomials and it is provided in the Appendix A.1.

Theorem 2.3.1 *Suppose we have two Gaussian components with width parameters τ_1 and τ_2, satisfying $\tau_1 \leq \tau_2$ without loss of generality, and amplitude parameters, A_1 and A_2. Then for the components to be de-mixed at scale s the distance between them must satisfy*

$$\theta > min\left\{\sqrt{3(\tau_1^2+s^{-2})},\ .742\sqrt{\tau_1^2+s^{-2}} + .742\sqrt{\tau_2^2+s^{-2}}\right\}.$$

While this theorem provides a necessary condition for de-mixing, it is not a sufficient condition except in the following corollary.

Corollary 2.3.2 *Suppose we have two Gaussian components with width parameters $\tau_1 \leq \tau_2$ and equal amplitude parameters $A_1 = A_2$. Then the components can be de-mixed at scale s if the distance between them satisfies*

$$\theta > min\left\{\sqrt{3(\tau_1^2+s^{-2})},\ .742\sqrt{\tau_1^2+s^{-2}} + .742\sqrt{\tau_2^2+s^{-2}}\right\}.$$

Multivariate Extension. We can extend the previous results to a d-dimensional signal. To simplify the derivation of the coefficients, we illustrate their close-form expression for $d = 2$ while the limit bound for de-mixing are for any $d \geq 2$. We first use the following definitions of Mallat (67) for 2-dimensional wavelets:

$$\psi^1(x,y) = \frac{d^2}{dx^2}\phi(x,y),\ \psi^2(x,y) = \frac{d^2}{dy^2}\phi(x,y),\ \text{and}\ \psi^3(x,y) = \frac{d^4}{dx^2dy^2}\phi(x,y).$$

That is, we can take the derivative of the Gaussian function in either the vertical, horizontal, or diagonal orientation resulting in three wavelet functions. In the general case of $d \geq 2$ dimensions, there are $2^d - 1$ possible wavelet orientations. Then the dilated wavelet in any orientation becomes

$$\psi_s^m(x,y) = \left(\frac{1}{s}\right)^2 \psi^m\left(\frac{x}{s}, \frac{y}{s}\right), \qquad m = 1, \ldots, 2^d - 1.$$

The wavelet function is now a function of the scale variable, s, and the two coordinate values x and y, that is,

$$W^m f(s,x,y) = f \star \psi_s^m(x,y).$$

Then following similar arguments from the 1-dimensional wavelet transform, we have that the wavelet coefficients for the horizontal orientation are, up to scale,

$$-\sum_{l=1}^{L} \frac{A_l \tau_l}{\sqrt{\tau_{l,1}^2 + s^{-2}}} H_2\left(\frac{x - \omega_{l,1}}{\sqrt{\tau_{l,1}^2 + s^{-2}}}\right) \times \exp\left\{-\frac{1}{2}\left[\left(\frac{x - \omega_{l,1}}{\sqrt{\tau_{l,1}^2 + s^{-2}}}\right)^2 + \left(\frac{y - \omega_{l,2}}{\sqrt{\tau_{l,2}^2 + s^{-2}}}\right)^2\right]\right\}.$$

Similar formulations for the vertical and diagonal orientations of the wavelet transform can be easily derived. As illustrated in Figure 6, by combining knowledge from the vertical and horizontal orientations our algorithm can discern the existence of four true Gaussian components. In the diagonal orientation, however, our algorithm, which tracks maxima across wavelet scales, would select eight components, resulting in the false discovery of four components. Therefore, to decrease the false discovery rate, we exclude information from the diagonal orientation.

Using the closed-form expressions for the wavelet coefficients we can derive the resolution limits for a d-dimensional signal.

Theorem 2.3.3 *A necessary condition for de-mixing two d-dimensional Gaussian components at scale s is that there must exist some $d' \in \{1, \ldots, d\}$ such that*

$$\theta_{d'} = |\omega_{1,d'} - \omega_{2,d'}| > min\left\{\sqrt{3(\tau_{d',1}^2 + s^{-2})},\ .742\sqrt{\tau_{d',1}^2 + s^{-2}} + .742\sqrt{\tau_{d',2}^2 + s^{-2}}\right\}.$$

Similar to the 1-dimensional case we have the following corollary:

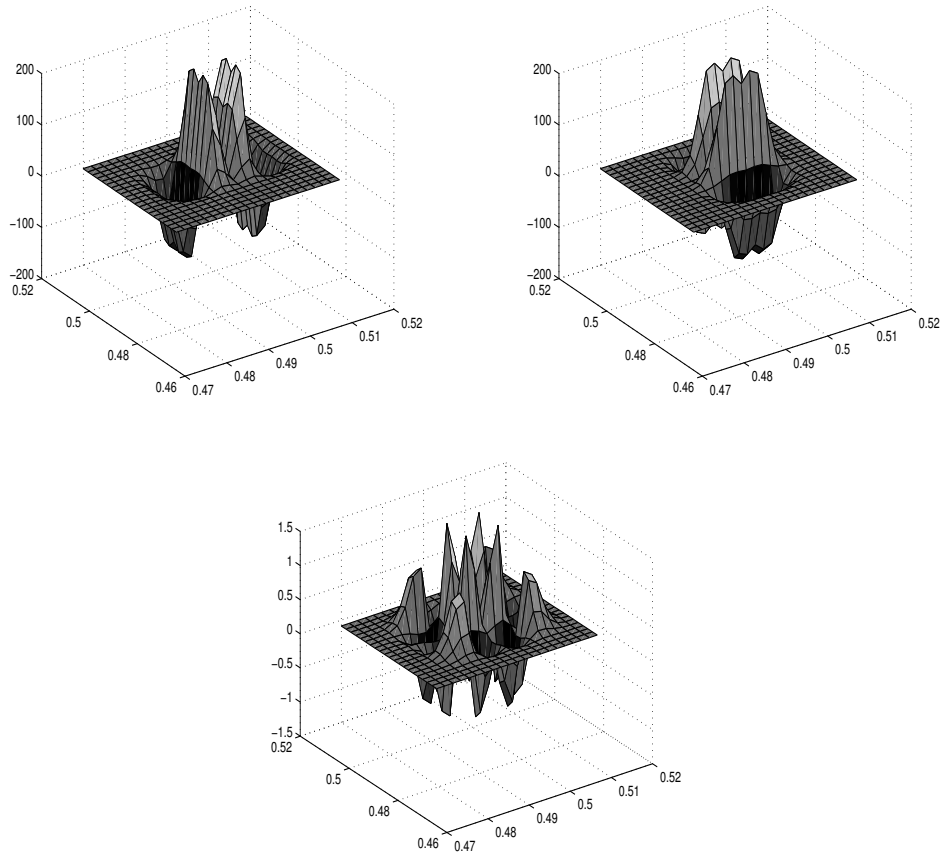

Figure 6: Wavelet transform of four Gaussian components in close proximity, comparison across orientations.

Corollary 2.3.4 *If $A_1 = A_2$, then a sufficient condition for de-mixing two d-dimensional Gaussian components is that there must exist some $d' \in \{1, \ldots, d\}$, such that*

$$\theta_{d'} > min\left\{\sqrt{3(\tau_{d',1}^2 + s^{-2})},\ .742\sqrt{\tau_{d',1}^2 + s^{-2}} + .742\sqrt{\tau_{d',2}^2 + s^{-2}}\right\}.$$

2.3.2 Pitfalls of the Resolution Limits

As is often the case when moving from theoretical derivations to working with real data, situations arise where the results may not fully line up with theoretical derivations. That is, not all components within the theoretical limits of de-mixing are discoverable in the wavelet domain. When are the resolution limits guaranteed to hold? If they do not, why? The answers to these questions are twofold: interference between clusters of components and the discrete time sampling density.

Interference Between Component Clusters. The resolution limits previously derived apply strictly to two components in isolation from other components. To illustrate, consider Figure 7. Here we plot four components that are pair-wise separable according to the resolution limits previously described. However, we would only discover two instead of four components in the wavelet domain.

Since the resolution limits rely on the derivative of the MHW function, we can determine when another cluster of components is in close enough proximity to interfere with the resolution results. Following similar arguments from the previous section using the roots of the Hermitian polynomials we have the following lemma providing only a necessary condition.

Lemma 2.3.5 *Suppose we have a set of 3 components $s(x; \omega_1, \tau_1)$, $s(x; \omega_2, \tau_2)$, and $s(x; \omega_3, \tau_3)$, with inter-component distances $\theta_1 = |\omega_1 - \omega_2|$ and $\theta_2 = |\omega_2 - \omega_3|$, where θ_1 satisfies the resolution limits of Theorem 2.3.1 and $\theta_2 > \theta_1$, without loss of generality. Then the resolution limits from Theorem 2.3.1 hold for the pair of components 1 and 2 if θ_2 satisfies $\theta_2 > 2.334\sqrt{\tau_2^2 + s^{-2}} + 2.334\sqrt{\tau_3^2 + s^{-2}}$.*

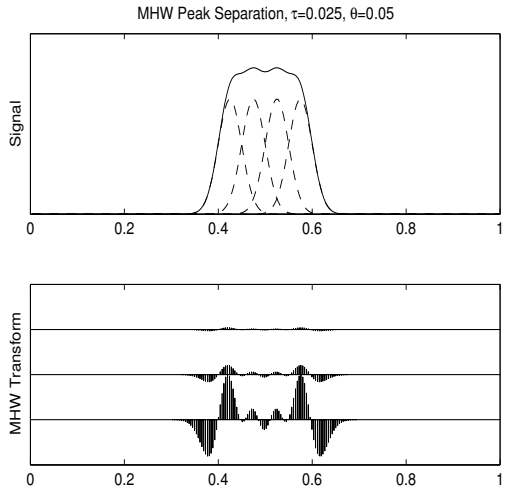

Figure 7: Four Gaussian components in close proximity. The individual components are plotted with a dashed line, as well as their sum and wavelet transform.

Lemma 2.3.6 *Suppose we have a set of 3 d-dimensional components $s(x;\omega_1,\tau_1)$, $s(x;\omega_2,\tau_2)$, and $s(x;\omega_3,\tau_3)$, with inter-component distances $\theta_1 = |\omega_1 - \omega_2|$ and $\theta_2 = |\omega_2 - \omega_3|$, where θ_1 satisfies the resolution limits of Theorem 2.3.1 and $\theta_2 > \theta_1$, without loss of generality. Then the resolution limits from Theorem 2.3.3 hold for components 1 and 2 if $\forall\ d' \in \{1,\ldots,d\}$, $\theta_{2,d'} > 2.334\sqrt{\tau_{2,d'}^2 + s^{-2}} + 2.334\sqrt{\tau_{3,d'}^2 + s^{-2}}$.*

Discrete Sampling Limitations. As stated previously, all the results have been derived under the assumption of a continuous signal. The results for a discrete signal still apply in the sense that the wavelet coefficients will follow the closed-form expressions derived in the previous section but at discrete sampled coordinates. However, the resolution limits may not hold because the sampling density may not be fine enough. Therefore, we offer the following lemma to address the issue of discrete sampling.

Lemma 2.3.7 *The results from Theorem 1 and Corollary 1 hold for a discrete sample if the distance and sampling density satisfy*

$$\theta > min\left\{\sqrt{3(\tau_1^2 + 2^{-2K})},\ .742\sqrt{\tau_1^2 + 2^{-2K}} + .742\sqrt{\tau_2^2 + 2^{-2K}}\right\},$$

where 2^K is the sampling density.

A minimum scale, $s = 2^K$, is recommended otherwise the wavelet will be too fine and only detect noise spikes due to the wavelet 'fitting' between sample points. Although not particularly difficult to derive in most cases, we do not add Lemmas for the multivariate sampling density results because the necessary and sufficient conditions become overwhelmingly intricate due to the large number of combinations of widths, orientations and sampling densities to consider.

Proofs of the results in this section are available in Appendix A.1.

Effect of Noise on the Resolution Limits. When $\sigma > 0$, the wavelet transform of a white noise process is itself white noise while the additivity of the wavelet transform is preserved,

$$Wf(s,x) = \sum_{l=1}^{L} Wf_l(s,x) + WN(0,\sigma_W).$$

Furthermore, because the Gaussian function belongs to a family of smoothing functions and the MHW is similarly shaped, it is reasonable to assume that the wavelet function will smooth out the noise. Finally, by beginning our search for components at coarser scales and working to finer scales, the number of noise peaks in the signal domain will be reduced as described in the section on sensitivity results. Therefore, in all but the extreme cases the resolution limits will hold.

2.3.3 Sensitivity Analysis

Mallat and Hwang show that when using a wavelet with the appropriate number of vanishing moments, the wavelet coefficients of a singular function will decay exponentially at a rate α:

$$|Wf(s,x)| \leq Cs^{\alpha},$$

where C is a constant and α is the Lipschitz exponent (66). Since the Gaussian function does not have a bounded Lipschitz exponent (see Appendix A.1), we expect a different behavior when compared with the surrounding noise coefficients. According to the results of Lindeberg (60), the wavelet coefficients will increase until $s^{-1} \approx \tau$ in proportion to the

coefficients related to their noise counterparts. We demonstrate this property in a simple example in Figure 4. Furthermore, the Lipschitz exponent of a white noise process is $\alpha = -1/2 - \epsilon$, for $\epsilon > 0$. This implies that the number of maxima at scale $s = 2^k + 1$ will be approximately double the number of maxima at $s = 2^k$ (66). Therefore, to detect components we begin at coarser scales to avoid false positives and progress to finer scales to ensure greater accuracy in locating components. However, how coarse of a scale should we consider? How do we ensure that we are tracking the maxima of signal components and not noise?

When performing the CWT we are not limited by the choice of scale as in the case of the DWT. In theory, any choice of scale, s, is possible, but in practice we are limited by the sampling density and the type of feature we are trying to detect. Assume that we have $N = 2^K$ sampling points. Then an upper limit for the scale is $s \leq 2^K$, otherwise the wavelet function may fit between sample points. (In keeping with the DWT, we choose to limit the selection of scale to dyadic levels with intermediate half-levels on the $\log_2$ scale.) In selecting the appropriate scale, we refer to two results. Lindeberg has shown that the wavelet coefficients, $Wf(s, x)$, will be maximized when $s^{-1} = \tau$, (60), and Nenadic and Burdick propose choosing the set of scales to be uniformly distributed within the known range of the signal width (81). Therefore, we consider the following set of scales: $\left\{s_k = 2^{K+k}, k \in \{-2, -1.5, \ldots, 0\}\right\}$, where 2^K is the sampling density, while limiting the scales in the NMR application using prior knowledge of a lower bound of the number of components that need to be discovered. We also employ ideas from (61) by tracking the maxima of the wavelet transform across scales. This mitigates the doubling effect of the noise spikes and takes advantage of the multi-scale properties of smooth 'blob'-like components.

2.4 Component Identification Algorithm

In this section we describe an algorithm for component identification informed by the previous theoretical insights. To begin, consider the example in the left plots of Figure 3. Here the components are completely mixed and are not separable at the coarser scales of the wavelet transform. Transitioning from coarse to fine scales, the separation of the components becomes evident. In order to determine an approximate location for the components we trace the behavior of predominant features across scales in the wavelet transform. We use dyadic scaling with an intermediate half-scale at each level. The coarsest scale considered is 2^{K-2} and the finest is 2^K, where 2^K is the sampling density.

For simplicity, we present the algorithm for 2-dimensional signals but the same algorithm can be extended for general dimensionality d. The algorithm consists of three steps:

1. Initial thresholding at each individual scale and orientation.

2. Tracking maxima across scales beginning with the coarsest scale.

3. Combining candidate components across orientations.

We begin with a simple thresholding of the wavelet coefficients to reduce the computational effort in tracking the behavior of predominant features or components across scales. The wavelet coefficients take positive values in the neighborhood of the mode of a component and they are of larger scale than the coefficients corresponding to noise. Thus we use a low enough positive threshold such that we do not miss components with low amplitude but large enough such that we carve out smaller regions to reduce the computation effort.

For each scale s_k and orientation $m \in \{1, 2\}$ consider the wavelet coefficients $W^m f(s_k, x_i, y_j)$, referred to simply as $\beta_{i,j,k,m}$. If $\beta_{i,j,k,m} > 2.5\sigma_{s_k,m}$, then we consider the area neighboring the coefficient to further search for features corresponding to components in the assumed model. The parameter $\sigma^2_{s_k,m}$ is the variance of the errors at the s_k scale and orientation m. We estimate $\sigma_{s_k,m}$ at each scale and orientation by the median absolute difference (MAD)

estimator,

$$\hat{\sigma}_{s_k,m} = \frac{med\{|\beta_{i,j,k,m} - \overline{\beta}_{k,m}|\}}{.6745},$$

where $\overline{\beta}_{k,m}$ is the average of the wavelet coefficients at scale s_k and orientation m. This estimator is robust to outliers (29; 47). Outliers arise due to large amplitude coefficients at the mode and surrounding the mode of a component. The choice for the starting threshold $2.5\sigma_{s_k,m}$ is conservative in selecting candidate components at the scales we consider. At the sampling densities we consider, with over 250,000 sample points, we are not concerned with eliminating a large portion of the wavelet coefficients (99.38%, in the case of normally distributed errors). In our simulation studies in examples with 500 components, there are routinely about 1500 candidate components above the threshold at the finest scale. However, by combining knowledge across scales, we greatly reduce the number of false positives.

After performing this initial thresholding step, we start in a 6×6 neighborhood around each candidate component at the coarsest scale. It is common to begin with the coarsest scale in order to reduce the number of false positives (81). This mitigates the doubling effect of the noise spikes and takes advantage of the multi-scale properties of smooth 'blob'-like components (60). For each orientation, we keep only the candidate regions that present a candidate component across the 5 finest scales and half-scales.

In order to combine information across orientations we first point out that the transform in the horizontal orientation may capture features that the vertical orientation may not, and vice versa. Therefore we take the union of identified components across the vertical and horizontal orientations only.

We highlight that for estimating the number of components, our algorithm does not require information on the width or amplitude of the components. By considering the three finest dyadic scales we capture components with a large variety of widths as demonstrated in the simulation trials. If the conditions of the resolution limits are met then the wavelet transform will separate mixed components without the additional computational expense

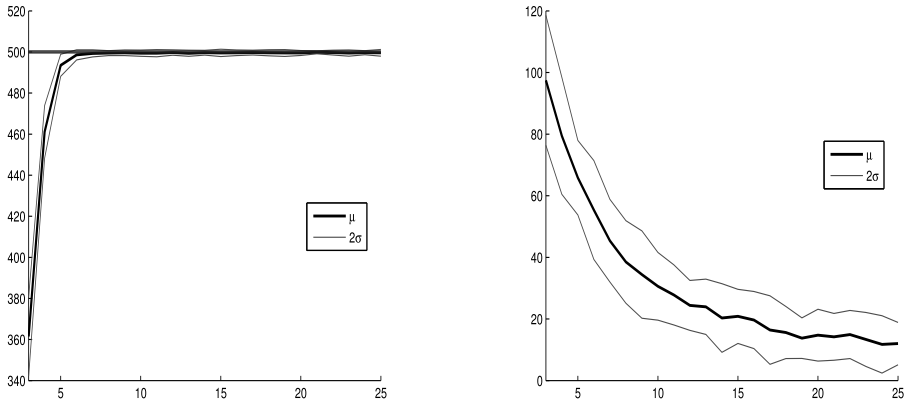

Figure 8: Plots of the mean and 2σ confidence intervals of the number of true and false positives at SNR levels 3-25.

of estimating the component widths or transforming back into the signal domain.

2.4.1 Simulation Studies

In this section we perform three simulation studies in which we test the sensitivity of the algorithm under varying signal-to-noise ratio (SNR) levels, we test the resolution limits at low noise levels while varying the width parameters of the components, and finally we combine these two settings. Throughout the simulation studies we set the sampling density along each axis at 2^9 in order to mimic the NMR data in our application.

2.4.1.1 Sensitivity Simulation

We simulate from the assumed model with $L = 500$ components within the unit square and 2^9 sampling density. We set $\tau = .0025$ and allow the SNR levels to vary from 3 to 25. We assume that τ is unknown in the simulation trial but *do not allow* mixing in this case, so we can solely examine the sensitivity limits. That is, all components are pairwise separable. We plot in Figure 8 the mean and 95% confidence intervals of the power and coverage of the algorithm, measured by the number of true and false positives detected. There is a very apparent sharp increase in sensitivity once a SNR level of 4 is reached with the mean being 460 (about 92%) and tending towards 500 once the SNR reaches 5.

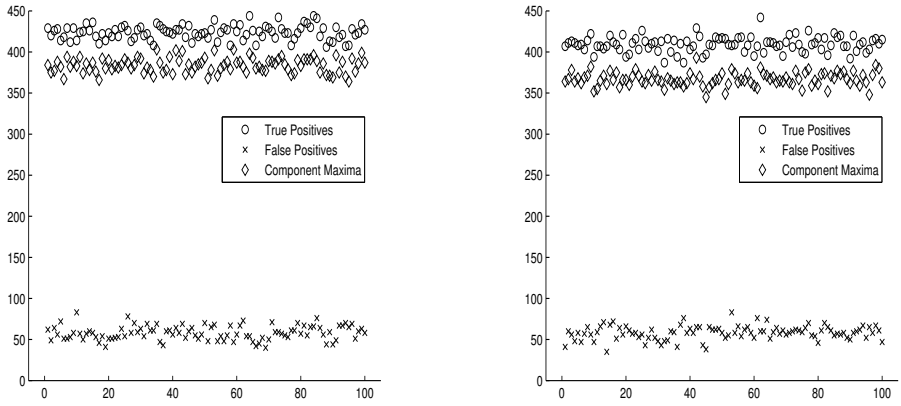

Figure 9: Plots of the results of 100 simulations with $L = 500$ Gaussian components. In the left plot, we fixed the SNR to 25, while in the right we allow the SNR vary by component uniformly between 4 and 25.

2.4.1.2 Resolution Simulation

We simulate from the assumed model with $L = 500$ components within the unit square and 2^9 sampling density. We set the SNR level to 25, and we allow τ to vary in the set $\{.0015, .0025, .005, .006, .0075\}$ with 100 components for each width value. We also allow for mixing to occur and plot the number of true and false positives as well as the number of 'undiscoverable' components. We assume that the widths τ are unknown and reasonably small, so that 500 components will fit within the unit square.

We plot the simulation results in Figure 9. Despite the variation in τ and the presence of complete mixing, the simulation results show a 10% increase in the number of components identifiable from the signal domain to the wavelet domain, with means of 384 and 424 components respectively.

2.4.1.3 Sensitivity and Resolution Simulation

In this setting we combine the simulation parameters of the previous two settings, where the SNR levels vary uniformly by component between 4 and 25 and the widths vary in the set $\{.0015, .0025, .005, .006, .0075\}$ with 100 components for each width. We also allow for mixing to occur. This is the 'worst-case scenario', where both τ and the amplitude are

allowed to vary and remain unknown. Even in this more extreme case, we have a 12% increase in the mean number of discovered components with means of 409 and 367 in the wavelet and signal domains, respectively.

2.5 Case Study: Two-dimensional NMR Data

We consider the analysis of a 2-dimensional Nuclear Magnetic Resonance (NMR) data generated for a doubly-labelled sample of a 130-residue RNA binding protein, rho130, using Heteronuclear Single Quantum Coherence (HSQC). The NMR signals were processed with FELIX (Felix NMR, Inc) using apodization and linear prediction methods that are typical for these types of experiments. After Fourier Transformation of the processed NMR signals, the 2D NMR data generated from this experiment follow a separable nonlinear model where the model components are approximately Lorentzian-shaped (100). (Appendix A.2 discusses the extension of the properties of the MHW transform apply to the Lorentzian shape also.) The data are observed over a two-dimensional 512×256 grid of points. The primary sampling density of $2^9 = 512$ was chosen by the experimenters and is adequate to understand the protein of interest, while the sampling density along the secondary axis is limited by the number of frequency combinations considered. These data have been previously analyzed by Serban (101).

In the NMR application, it is important to identify most components, even when they are totally mixed with other components, as they provide specific information about the structure of the protein. In certain cases, the lack of a small number of essential components can lead to a significant deviation in the predicted structure (44). Importantly, because the identification of the regression components is one of the first steps in the overall approach for structure determination using NMR, inaccuracy at this step will be perpetuated at further steps (43).

Commonly, the number of components for such data is large. For the NMR data analyzed in this chapter and displayed in Figure 10, the biomolecule (rho130) is rather small

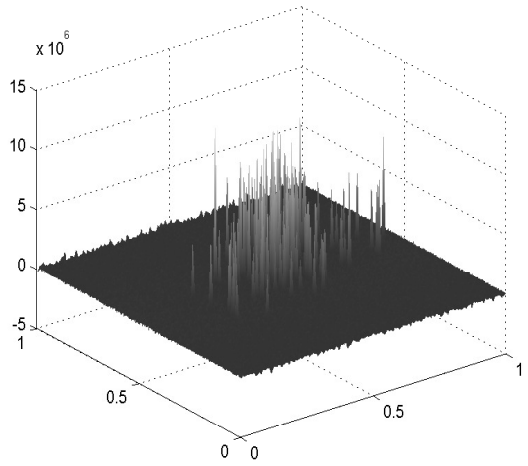

Figure 10: Perspective plot of the NMR Fourier Transformed data.

with about 130 detectable components. The majority of components are located closely in the center of the grid space and vary greatly in amplitude. The SNR level appears to be quite high although it is possible that some components to be buried among the larger amplitude ones.

Because we have prior knowledge of the components, specifically, that they are extremely thin, we only obtain the wavelet coefficients for the Mexican Hat CWT at the 3 finest scales and half-scales using the algorithm in Section 2.4. Because the NMR components will be clustered around the center of the grid space, we can select the threshold level at individual scales and orientations in order to limit false positives around the border of the signal.

In order to assess the impact of the threshold level on the number of identified components, we consider multiple threshold levels. We perform the thresholding at 15 different levels uniformly distributed between 3 and 10 standard deviations and plot the number of components against the threshold level in Figure 11. Due to prior knowledge that there will exist approximately 130 detectable components, we consider threshold levels of 5σ and 10σ initially. An experimenter could manipulate the threshold levels depending on the

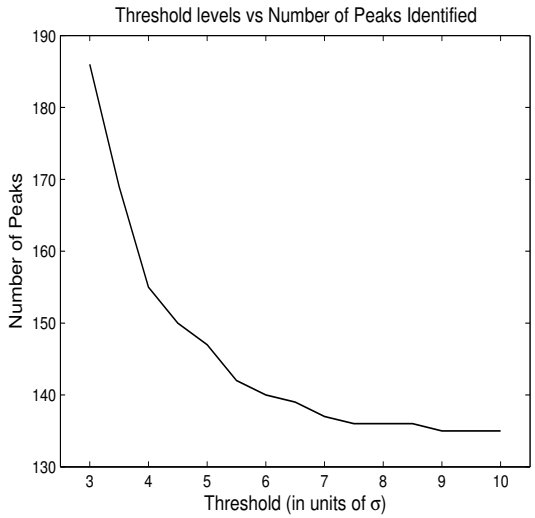

Figure 11: The number of components identified at various threshold levels in units of σ. Because we know that the number of components is approximately 130 from biological knowledge we consider higher threshold values in step (1) of our algorithm.

Table 1: Breakdown of Components Identified by Method

Method	Component Numbers
Discoverable	130
Both Methods	128 (98.5%)
MICE Only	4
MHW Transform Only	2
Possible Mixtures	7
Noise Peaks (5σ Thr. Only)	7

purpose of the study: if a false negative is more costly than a false positive, then a lower threshold may be considered, for example.

For the purposes of our study, we compare the set of identified components with those identified by Serban (101), which discovered 132 components plus additional mixtures. Based on our algorithm, we identify between 135 to 147 as we move from threshold levels with 10 to 5 standard deviations. Table 1 provides the results based on this comparison at a threshold level of 5σ.

We begin by estimating the number of theoretically separable components satisfying

the requirements of θ from Theorem 2.3.3. Setting τ in the limit to 0, thus greatly underestimating the lower limits of discovery, we have two pairs of components fall under the lower limit being unseparable in the wavelet transform. Setting τ equal to the width of the sampling density, 2^{-K}, we have six pairs under the lower limit.

The two sets of components provided by the algorithm in this chapter and by the comparison approach have 128 components in common. Following our theoretical results, all of the identified components are within the resolution and sensitivity limits and thus "detectable". There are four cases not detected by the MHW transform algorithm, two of which are below the resolution limits and thus are undetectable following our theoretical insights. The two additional cases detected by the MHW algorithm are false positives. Most importantly, the MHW transform has seven possible cases where the algorithm demixes a pair of components following from our resolution limits. The question remains then, whether the vertical and horizontal orientations are identifying the same component in a slightly different location, or if it is a true mixture. In four of these cases it seems that the algorithm is detecting a mixture because the distance between the components is more than two sampling points.

2.6 Discussion

The primary emphasis of this study is on understanding when components in the assumed separable nonlinear regression model are 'detectable' while providing a road map on how to use these results to identify components. We distinguish between sensitivity and resolution analysis, the former separating the components from noise (noise-component interference) and the latter separating components that have similar location parameters (inter-component interference). We propose using a wavelet transformation using the Mexican Hat wavelet function to perform the two analyses jointly.

Our theoretical results are rooted in the existing research from the fields of image processing, discrete and continuous wavelet theory. Drawing from the ideas of (59; 60; 62), we

are able to derive closed-form expressions for wavelet coefficients and determine behavior of component features across wavelet scales. Additionally, we maintain a fundamental principal of matching the shape of the wavelet function with that of the features that are to be detected; a high similarity between the two results in large amplitude wavelet coefficients (81). Finally, we consider discrete wavelet transform methods to sketch an algorithm for component identification that is grounded in our theoretical results.

The validity and applicability of our theoretical resolution and sensitivity limits have been demonstrated in a series of simulation studies presented in Section 2.4.1. In these simulations, we start with simple single-component settings in which we assess the coverage and power of our algorithm then move to more realistic multiple-component settings in which a model with 500 components is generated. We find that in simpler settings the rate of discovery is higher than 99.4% in the sensitivity analysis and in average of 95.8% in the resolution analysis across varying signal-to-noise (SNR) ratios. This result confirms the validity of our theoretical limits. In addition, the multiple-component simulations show that the presence of other components reduce our ability to discover components when the SNR is low. When the SNR is in the range of 15, the rate of false positives is around 1% whereas when it decreases at 10, the rate of false positives is around 10%. This significant increase in the false positive rate comes from two sources, the discrete sampling and the interference between clusters of components.

One important limitation of the derivations in this chapter is that we assume the width parameters to be known. The theoretical limits are functions of the widths. Thus, in order to be able to evaluate the resolution and sensitivity limits, the widths need to be estimated with some degree of accuracy. In the NMR applications, it is commonly assumed that the components have similar widths and thus one could borrow information across well defined and separated features to first estimate the common width parameter. In other applications, prior information about the experiment could be used to obtain estimates for these parameters. Generally, in a setting without restrictions on these parameters it is theoretically

infeasible to determine when the algorithm is separating a cluster of components or when different orientations are simply estimating the location of the same component in a slightly different location.

CHAPTER III

UNCOVERING LONGITUDINAL HEALTHCARE UTILIZATION FROM PATIENT-LEVEL MEDICAL CLAIMS DATA

In this chapter the objective is to study longitudinal claims data observed at the patient level, with inference on the heterogeneity of healthcare utilization behaviors, and to quantify cost-saving interventions in improving outcomes within large healthcare systems such as Medicaid. The proposed approach is model-based, allowing for visualization of longitudinal utilization behaviors and manipulation of utilization profiles in order to evaluate "what-if" scenarios using simple stochastic graphical networks. The approach is general, providing a framework for the study of other chronic conditions wherever longitudinal healthcare utilization data are available. Our methods are inspired by and applied to patient-level Medicaid claims for asthma-diagnosed children diagnosed observed over a period of five years, with a comparison of two neighboring states, Georgia and North Carolina.

3.1 Introduction

Healthcare can be thought of as a continual series of information-processing experiments: from the initial collection of data (the patient's history, physical exam, and diagnostic tests), a hypothesis (diagnosis) is formed and then validated by further data collection (94). Data in healthcare are generated at every patient's encounter with the healthcare system, at every implementation of medical processes, with every decision made by healthcare organizations, and with every policy implementation in the healthcare ecosystem, resulting in billions of data points every day. Every patient in any medical setting generates an invaluable data point that can contribute to understanding what works, for who and where.

One health-related information technology that has provided substantive opportunities

to study healthcare data across large populations and across many years is the medical claims system. Information coded in claims data is standardized to a great extent (12), hence making such data amenable to large scale studies. Developing methods to translate medical claims data into meaningful data is the first crucial step in deriving knowledge useful to make inferences about the healthcare system. Further development of adaptive and scalable data mining and statistical methods provide the means for analyzing these data. However, there are a series of challenges associated with mining data derived from medical claims, including the derivation of knowledge for decision support while maintaining computational efficiency and complying with privacy regulations.

Two common methodologies for mining healthcare data are network analysis and cluster modeling. Network analysis investigates the structure of relationships between different entities, i.e. healthcare providers or patients, defined as nodes in the network, in order to determine the extent of relationships between different nodes and groups of nodes (52; 82; 103). It is often applied in healthcare analytics to produce visual summaries of large healthcare datasets and to detect the strength of the connection between different event types (19; 56; 104). However, most network studies only model the strength of the connection between two event types without considering a rigorous treatment of the time domain. Furthermore, most network analysis models seek to determine clusters of nodes within a single network, not allowing for the heterogeneity in the population. Statistical clustering analysis is commonly used to characterize heterogeneity or similarity among patients with respect to a set of predefined features (99; 109; 120), but it has not been applied to model sequences of discrete healthcare events as proposed in this study. We propose a method that combines the benefits of network analysis and model-based clustering for discrete event sequences, assuming the discrete-event sequences follow a stochastic process. Thus one contribution is a model-based data mining algorithm that has the ability to scale to massive data while producing meaningful stochastic networks that can then be used in decision support through visualization and evaluation of different "what-if" scenarios. The

second contribution is the application of the modeling approach to derive inferences on utilization behaviors from highly-sensitive, large patient-level claims data.

We pilot our methodology using Medicaid Analytic Extract (MAX) data acquired from the Centers of Medicare and Medicaid Services (CMS) for five years (2005-2009). We consider one specific chronic condition, pediatric asthma, and we compare utilization for two states, Georgia (GA) and North Carolina (NC). While GA and NC have similar pediatric populations, the two states deliver care under different coordinated-care Medicaid systems (5; 51). This pilot study provides insight into the effects of such different state-based Medicaid systems. We chose pediatric asthma as the health condition of interest because it is a common chronic childhood condition, with more than 9% of American children affected by the disease (8). The MAX data consist of 1.8 and 2.4 millions claims for GA and NC, respectively. We evaluated the computational complexity of our methodology and tested its implementation for much larger number of claims, validating the applicability of our methodological framework to larger states, such as California and New York, and to larger healthcare benefits systems.

This chapter is organized as follows. Section 3.2 introduces the data science framework with a focus on the information translational process, as applied to the MAX claims data. Section 3.3 presents the model-based clustering procedures. We apply the methodology and provide results and findings in Section 3.4. We conclude with overall policy implications and discussions in Section 3.5. Difficult derivations and further data summaries beyond the scope of this article are deferred to the Appendices.

3.2 From Information to Meaningful Data

The increasing availability of large amounts of data over the last two decades has resulted in a new field of study, *data science*, dedicated to knowledge discovery from large data sets. Data science goes beyond statistical data analysis (80; 119), particularly for massive, complex data sets, where the priorities now shift from simply getting and analyzing data to

making them manageable and understandable. Because the advancements in data science have not kept pace with the size and complexity of the data available, there is a clear emergence of methodologies to overcome what Tien & Goldschmidt-Clermont (108) call the 'data rich, information poor conundrum.' Particularly in healthcare, the derivation of knowledge is especially limited by the availability of information. When considering large amounts of information, it is critical not only to decide the appropriate data to use but also to determine *how* to use them. Knowledge discovery relies and builds entirely on this initial translation step (80).

In this section, we expand on the derivation of the patient-level utilization sequences from the CMS Medicaid Analytic Extract claims data, as an illustration of the translational process of information into data. The data are made available as a set of large flat files, with an extensive data dictionary including highly-specialized coded information. The flat files of medical claims must be reshaped in order to analyze longitudinal utilization sequences, requiring extensive database structuring and use of data dictionaries together with information from various other sources. Parsing through large, flat text files is extremely computationally intensive, therefore we reconstruct the flat files into a relational database, with keys and indices to accelerate the data extraction process. We use a combination of SQL queries and scripting language to manipulate and analyze the extracted data.

Our emphasis is on a subset of patients, particularly the Medicaid-enrolled children ages 4-18 with an asthma-related primary diagnoses. We filtered the data based on the ICD-9 diagonosis codes provided with each claim (given in Appendix B.1) and their date of birth. (The age group 0-3 is excluded from this study because of the difficulty and inaccuracy of diagnosing asthma at this age.) Moreover, in order to capture longitudinal utilization behaviors, we only consider those patients that are of the appropriate age to qualify for Medicaid for at least four of the five years. Thus starting with a dataset including a total of 316 and 457 millions of claims for Georgia and North Carolina, respectively, we derive utilization and cost data from 1.8 and 2.4 millions of claims for this subset of patients.

The MAX claims are structured into inpatient care (IP), long-term care (LT), other care including outpatient services (OT), patient summary (PS) and prescription claim summary (RX) files. Included for each claim are data entries specifying the date of service, the Medicaid Statistical Information System identification (MSIS ID) of each patient, the International Classification of Diseases, Ninth Revision (ICD-9) codes for diagnosis or services provided, and the type and place of services rendered. We use the IP and OT files to determine the visits to a specific provider type, and the RX file to determine the medication type and date of the prescription being filled. We abbreviate our derived provider types as follows: clinic visits (CL), emergency room visits and outpatient hospitalizations (ER), inpatient hospitalizations (HO), physician's office visits (PO), nurse practitioner services (NP), and filling of medication prescriptions (RX). These provider types are derived from the place of service code and type of service code in the IP and OT files. We consider long-term asthma controller medications, derived from the National Drug Code in the RX files, as an event type in the sequential analysis due to its significance in treating asthma symptoms.

In short, we are able to extract the utilization-specific data and transform claims records into patient-level utilization sequences. We include a table in Appendix B.1 detailing the roadmap between the entries in these files and our categorizations.

3.3 From Data to Knowledge: Uncovering Utilization Profiles

In this section we describe our methods for translating patient-level utilization sequences into knowledge about underlying utilization behaviors via model-based data mining techniques. We compare our method with other approaches and provide our contributions, then present our modeling approach along with details on our choice of model estimation and selection techniques as well as how we quantify cost-saving interventions.

3.3.1 Model-based Data Mining

The goal of this study is to cluster patients using model-based methods according to their healthcare utilization behaviors and to produce meaningful visualization of utilization profiles through stochastic network modelling. Patient-level utilization is observed in the form of *sequential data*, referring to the observation of a discrete set of events over a period of time. In sequential data, the events may be ordinal or categorical, and the time domain may be discrete or continuous. Examples of such data can be found in pattern recognition of text and speech (118), in process mining where business workflows must be inferred (10), and in the area of genetics, where sequential clustering is a primary research interest (58).

The proposed methods for modeling sequential data are inspired by the large body of existing research in network analysis, process mining and claims mining literature. While network analysis is useful in determining the strength of connections between event types and in producing meaningful visual outputs, it has not been applied to model longitudinal sequences of events and it has not been considered jointly with clustering analysis to derive distinct networks for heterogeneous groups of members or patients (19; 56; 104). Process mining techniques are applied in business and healthcare settings to extract meaningful patterns from data logs that document events making up the workflow (7; 10; 24; 35; 53; 57; 93). Typically, these methods only model the order of the sequence of events without consideration of the interarrival time between events (35; 93). Finally, in the existing research for modeling longitudinal claims data, stochastic models are primarily used to identify outlying utilization behaviors, particularly in fraud detection (63; 68; 84; 86; 106; 110; 123). In contrast our objective is to inform policy decision making on major underlying utilization profiles, not outlying individuals or providers, by simultaneously grouping probabilistically similar patients and estimating the distribution parameters in order to produce useful model summaries for visualization.

Our algorithm has the following novel features: adaptability due to the hierarchical tree-based step, scalability due to our model assumptions, without the need for costly Markov

chain Monte Carlo (MCMC) experiments to initialize the algorithm, and a rigorous treatment of likelihood theory and model complexity. Model-based clustering approaches do use an expectation-maximization (EM) algorithm for maximizing the posterior likelihood of the cluster membership, but without the guarantee of producing consistent results with each run (35; 93) that is possible with hierarchical methods. Others use hierarchical methods employing statistical measures of complexity, but may not necessarily maximize the posterior likelihood (90; 91). Our approach combines important properties of hierarchical methods and the EM algorithm to find a clustering outcome that maximizes the tradeoff between posterior likelihood and model complexity as measured by the Bayesian information criterion (BIC) score. Additionally, by performing hierarchical clustering in a top-down approach we are able to quickly identify the large underlying profiles of care. This is in contrast to the computationally extensive bottoms-up approach of grouping together similar patients or employing costly MCMC experiments to initialize the algorithm (95).

3.3.2 Clustering Analysis: The Model

In this section we describe how we use a Markov renewal process (MRP) framework to model longitudinal utilization sequences. This model-based algorithm simultaneously estimates model parameters, groups patients into distinct profiless, and improves the BIC score at each iteration. By using the MRP model we take advantage of properties of stochastic processes to provide simple model estimation procedures with minimal computational complexity. Particularly, Markov processes provide a manner for aggregating large amounts of sensitive data so that it may be shared in the form of attractive visual displays.

3.3.2.1 The MRP Model

We begin introducing our approach by presenting the model for one sequential realization of the patient's utilization of the healthcare system (85; 98). We extend this model to multiple sequences corresponding to multiple patients in the next section.

Let $\vec{X} = (X_1, \ldots, X_L)$ refer to the sequence of events and $\vec{T} = (T_1, \ldots, T_L)$ to the set of "arrival" times, times that an event occurs, where L is the length of the patient healthcare utilization sequence. An example of a longitudinal utilization sequence could be: patient A visits the emergency room for an asthma attack on January 1st, 2005, is given a prescription for an inhaler which she fills one month later, and is referred to a primary care physician. Subsequent visits to the same physician and refills of her asthma prescriptions occur at 3-month intervals. The sequence $(X_1, T_1), \ldots, (X_6, T_6)$ is given by $(\text{ER}, 0.00)$, $(\text{RX}, 0.08)$, $(\text{PO}, 0.25)$, $(\text{PO}, 0.50)$, $(\text{RX}, 0.75)$, $(\text{PO}, 1.00)$.

The MRP is the continuous-time analog of a discrete-time Markov chain (DTMC). The primary assumption of any Markov process is that it is 'memoryless', i.e. future states are only dependent on the current state of the system. Define $\tau_n = T_n - T_{n-1}$. Then we have that

$$Pr(\tau_{L+1} \leq t, X_{L+1} = s_j | X_1, T_1, \ldots, X_L, T_L) = Pr(\tau_{L+1} \leq t, X_{L+1} = s_j | X_L = s_i).$$

In an MRP, the concept of memoryless-ness arises twice. Not only are the events memoryless, as in the DTMC, but so are the interarrival time distributions. While the memoryless property may not be a reasonable assumption in the case of longitudinal healthcare utilization our clustering algorithm profiles patients based on the complete patient history, so that the clustering outputs are representative of underlying utilization behaviors from start to finish.

3.3.2.2 Parameter Estimation

Consider again the sequence $\vec{X}, \vec{T}$. Let $s_i, i \in \{1, \ldots, S\}$ be all possible events in the sequence (in our case CL, ER, HO, PO, NP, and RX), where S is the number of states. In an MRP, the sequence $\vec{X}$ is itself a DTMC, with corresponding transition matrix P, where P_{ij} denotes the transition probability between s_i and s_j, and $\sum_{j=1}^{S} P_{ij} = 1$. The likelihood

function for a single realization of a DTMC is given by

$$L(P|\vec{X}_L = \vec{s}_L) = Pr(X_1 = s_{i_1}, \ldots, X_L = s_{i_L}) = \prod_{l=2}^{L} P_{i_{l-1}, i_l}, \tag{1}$$

with the derivation given in Appendix B.2. We estimate each P_{ij} via maximum likelihood estimation: for each state s_i and s_j, $\hat{P}_{ij}$ is the number of transitions from s_i to s_j divided by the total number of transitions out of s_i.

Now we define the distributions for the sequence of interarrival times, $\tau_l = T_{l+1} - T_l$. We assume that for each pair $i, j \in \{1, \ldots, S\}$, the distribution of the interarrival time between states s_i and s_j is given by F_{ij}. We assume that F_{ij} follows an exponential distribution with rate parameter λ_{ij}. To estimate λ_{ij} we use maximum likelihood estimation. The likelihood function of the interarrival times is given by

$$L(\Lambda|\vec{T}) = \prod_{l=2}^{L} \lambda_{ij} \exp\{-\lambda_{ij}\tau_l\} I(X_l = s_i, X_{l+1} = s_j),$$

and the MLE is the reciprocal of the average interarrival times between any pair of states s_i and s_j. We will use the matrix $\{\Lambda\}_{ij}$ to denote the inverse of the average interarrival times, λ_{ij}, between states s_i and s_j.

The assumption of exponentially distributed interarrival times is restrictive, however it is a reasonable approximation in that it has an appropriate time domain starting at 0 and with a long tail towards infinity. Additionally, the MLEs are easy to compute in our model, an important aspect within a large-data analysis context. Furthermore, if it were the case that the distribution of interarrival times is multi-modal, then it is within the realm of our algorithm to separate such subsets of patients by forcing the interarrival times to be unimodal.

Now we can define the likelihood function for a set of patient utilization sequences. For patients $r \in \{1, \ldots, R\}$, the likelihood function of P is:

$$L(P|\vec{X}_1, \ldots, \vec{X}_R) = L(P|\vec{X}_{\vec{R}}) = \prod_{r=1}^{R} \prod_{l=2}^{L_r} P_{i_{l-1_r}, i_{l_r}}.$$

Likewise, the likelihood function of Λ is:

$$L(\Lambda|\vec{T}_1,\ldots,\vec{T}_R) = L(\Lambda|\vec{T}_{\vec{R}}) = \prod_{r=1}^{R}\prod_{l=2}^{L_r} \lambda_{ij}\exp\{-\lambda_{ij}\tau_{l_r}\}.$$

Therefore, the joint likelihood function of P and Λ is:

$$L(P,\Lambda|\vec{X}_{\vec{R}},\vec{T}_{\vec{R}}) = L(P|\vec{X}_{\vec{R}}) \times L(\Lambda|\vec{T}_{\vec{R}}), \tag{2}$$

with the derivation given in Appendix B.2. Together, the set of all possible transitions and interarrival times out of state s_i form a probability distribution which we refer to as the *transition distribution* out of s_i. Each transition distribution is a mixture of exponential distributions.

Remark: There is no significance to the observational timeframe in our study, 2005 through 2009, other than these are the endpoints of our study. It is entirely possible that we miss visits and referrals to providers before and after the time period of our study. Likewise, the estimates for the first arrival time and the last arrival time are going to be extremely biased. Therefore, we leave the first and last interarrival times out of the estimation and calculation of the posterior distribution. We revise the likelihood function to be:

$$L(P|\vec{X}_{\vec{R}}) \times L(\Lambda|\vec{T}_{\vec{R}}) \times \prod_{r=1}^{R} P_{LC,i_{1_r}} \times \prod_{r=1}^{R} P_{i_{L_r},RC},$$

where LC is the left censor (Jan. 1st, 2005) and RC is the right censor (Dec. 31st, 2009).

3.3.2.3 Determining Cluster Membership

In our algorithm we assign each patient to a profile based on the maximum posterior likelihood of the patient for each profile. Let $\vec{Z}_{\vec{R}}$ be a latent variable vector $(\vec{Z}_1,\vec{Z}_2,\ldots,\vec{Z}_R)$, following a multinomial distribution and containing the latent profile membership of patient r, for $r \in \{1,\ldots,R\}$. Together the vectors $(\vec{X}_{\vec{R}},\vec{T}_{\vec{R}},\vec{Z}_{\vec{R}})$ form the complete data on the patient population under our model assumptions. However, because $\vec{Z}_{\vec{R}}$ is unknown, we must infer the $\vec{Z}_r$ from $\vec{X}_r$ and $\vec{T}_r$, specifically the posterior (conditional) likelihood $P(\vec{X}_r,\vec{T}_r|Z_{rk}=1)$, the probability that patient r belongs to profile k given $\vec{X}_r,\vec{T}_r$:

$$P(Z_{rk}=1|\vec{X}_r,\vec{T}_r) = \frac{P(\vec{X}_r,\vec{T}_r|Z_{rk=1})P(Z_{rk}=1)}{P(\vec{X}_r,\vec{T}_r)} \propto P(\vec{X}_r,\vec{T}_r|Z_{rk}=1).$$

Here, $P(\vec{X}_r, \vec{T}_r)$ will be constant for all k and thus can be ignored. Likewise, without any *a priori* knowledge of the system, we set $P(Z_{r1} = 1) = P(Z_{r2} = 1) = \cdots = P(Z_{rK} = 1)$. Therefore, profile membership will be solely determined by the posterior likelihood $P(\vec{X}, \vec{T}|Z_{rk} = 1)$. That is, each observation is assumed to belong to the profile which produces it with the greatest posterior likelihood.

3.3.2.4 Model Selection

We seek to find the optimal clustering of sequences, given by $\vec{Z}_{\vec{R}}$, such that the BIC score is maximized. The BIC is an objective function that balances the tradeoff between maximizing the likelihood function while minimizing model size. For a model M,

$$BIC(M) = \ell(M) - |M| \cdot \log(R)/2,$$

where $\ell(M)$ is the log-likelihood of the model M, $|M|$ is the model size and R is the number of patients. Given the transition and interarrival parameters for the set of patients in profile k, P_k and Λ_k, for $k \in 1, \ldots, K$, $\ell(M)$ is given by taking the log of the likelihood function, (5),

$$\ell(M) = \sum_{k=1}^{K} \sum_{r=1}^{R} \ell(P_k|\vec{X}_r) + \ell(\Lambda_k|\vec{T}_r) + \sum_{r=1}^{R} P_{LC,i_{r1}} + \sum_{r=1}^{R} P_{i_{rL_r},RC}.$$

For model M with K profiles, we will estimate $KS(S+1)-1$ parameters for the transition matrices, $P_k, k \in \{1, \ldots, K\}$, and KS^2 in the interarrival matrices, $\Lambda_k, k \in \{1, \ldots, K\}$.

A common approach for model estimation is to use the EM algorithm. However, such an algorithm requires the user to pre-specify the number of profiles, K, regardless of the number of true profiles. Additionally, each initialization may produce a different outcome, implying that a global optimum is not necessarily reached with each clustering result. However, with a satisfactory initialization the output will produce a high likelihood without complex calculation.

Other researchers favor a tree-based algorithm, where a distance metric is used to determine splits in the set of observation (91). Ramoni, et al., use the BIC in conjunction

with the KL distance to perform agglomerative hierarchical clustering. However, a top-down approach is warranted in our case since we can choose a reasonable stopping point in the algorithm where the smallest number of splits explain the predominant patterns in the system. In contrast with the EM algorithm, the benefit of such a tree-based algorithm is that the the number of clusters can be determined after the clustering analysis is performed. However, it may not be guaranteed to maximize posterior likelihood of cluster membership. Therefore, we propose a joint tree-based, EM optimization algorithm that maximizes the BIC criterion.

3.3.2.5 The Algorithm

As K and R increase, it becomes computationally intractable to consider all possible partitions to find the maximum BIC score. Therefore, we present an algorithm that searches for a nearly maximal BIC at each iteration. Our algorithm, as in (90; 91), is guided by the Kullback-Leibler (KL) distance:

$$KL(Q_1||Q_1) = \int Q_1(x)\log\left(Q_1(x)/Q_2(x)\right)\partial x,$$

where Q_1 and Q_2 are the probability distributions under comparison. Specifically, we find the KL distance between the transition distribution out of each of the s_i for each individual sequence and the entire set of sequences in a given profile and then average across the $s_i, i \in \{1, \ldots, S\}$. (We provide the derivation of the KL distance in Appendix B.2.) We then order the average KL distances and find a nearly optimal partition in the observations to use as the initialization of the EM algorithm to maximize the posterior likelihood function. An overview of the algorithm is given below:

1. We begin with the null assumption, H_0, that all patients in a set belong to one profile. Find the population MLEs, $\bar{\Lambda}_{ij}$, and the transition matrix $\bar{P}_{ij}$ under the null hypothesis. Calculate the BIC_0 value.

2. Calculate the average KL distances between individual sequences and the one profile

(null hypothesis), $D_{ave}(P, \Lambda || \bar{P}, \bar{\Lambda})$.

3. For a sufficiently large, equally-spaced set of the ordered average KL distances, (say, 50), $D_{(i)}$, let $W^{-}_{D_{(i)}}$ be the set of patients with average KL distances from the null distribution less than $D_{(i)}$, and $W^{+}_{D_{(i)}}$ be the set of patient with average KL distances from the null distribution greater than $D_{(i)}$. For each partition, $\{W^{-}_{D_{(i)}}, W^{+}_{D_{(i)}}\}$, calculate the BIC_A corresponding to the BIC value of the alternative hypothesis, H_A, that the set of sequences should be partitioned into two profiles. This step is a heuristic search for the best initialization for the EM algorithm in the next.

4. Consider the partition $\{W^{*-}_{D_{(i)}}, W^{*+}_{D_{(i)}}\}$, such that the BIC score is maximized. Let this partition be the initialization for the EM algorithm. Recalculate the BIC score, call it BIC^{*}_{A} after the iterations of the EM algorithm.

5. If $BIC^{*}_{A} > BIC_0$, then divide the sequences into distinct profiles. Repeat steps (1)-(4) until no more divisions are made.

3.3.3 Deriving Simple Utilization Profile Visualizations

By employing stochastic models for clustering utilization sequences we can further derive stochastic provider networks via the transition matrices, allowing visualization of the utilization behaviors as networks across providers of different types. The primary inputs for the stochastic provider networks are the transition matrices. Specifically, the six provider types, CL, ER, HO, NP, PO, and RX, are the nodes in a directed graph. The directed edges represent transition probabilities between two provider types, for example, the transition of patients from the emergency room to a physician's office visit. For a simplified representation, the networks only include nodes such that a total of 90% volume is represented. We use different types of arcs for different levels of transition probabilities to better identify nodes that are most visited within each profile.

3.3.4 Assessing our Clustering Algorithm

We highlight five important properties of our clustering algorithm (4):

- *Robustness*: Defined as the ability to detect outliers. Our algorithm will place every observation within one profile, but as more divisions are made, the outlying observations become evident in low-membership profiles.

- *Minimum user-specified input*: By combining the EM algorithm with a hierarchical framework we do not need predefine parameters such as the number of profiles in the algorithm.

- *Scalability*: We simulated 5 different settings of R patients ($R = 100K$, $300K$, $500K$, $1M$, $1.5M$) and determined the run time of a single iteration of the algorithm. See Figure 12 for results on the runtime of the algorithm. In our study with over 100K patients in each state, the algorithm ran to completion through 8 iterations in approximately 3 hours.

- *Computational complexity*: The primary computational steps involved in fitting a patient sequence to an MRP rely on simple counting and averaging, while the computation of posterior likelihood relies on multiplication. All of these steps have computational complexity of order $O(n)$. The sorting step of the posterior likelihoods is the most computationally expensive with order $O(n\log n)$. Therefore, the computational complexity of our algorithm is $O(n\log n)$.

- *Visualization feasibility*: We translate the transition matrices into stochastic provider networks to produce simple visualizations of the utilization behaviors with each profile. The ability to quickly digest information on the similarities and differences between the different stochastic provider networks is a major advantage over simply providing the resulting estimated transition matrices as it can play an integral role

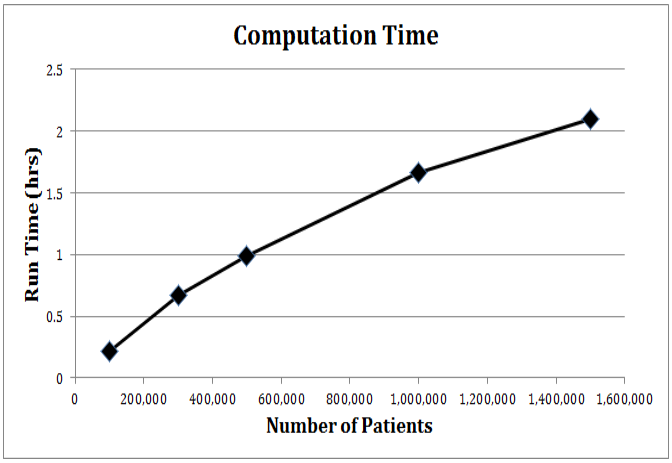

Figure 12: The runtime in hours of a single iteration of our algorithm plotted against the number of simulated patients.

in decision support systems. Moreover, we allow for different levels of visualization granularity of potentially complex healthcare systems. That is, the clusters of utilization behaviors can be split further into distinct profiles to reach a desirable balance between the number of profiles and intra-profile complexity. This is especially important when there are potentially a large number of event types.

3.3.5 Interventions for Recommended Care Adherence

By manipulating the weights of the transition matrices, we can analyze the cost-savings of an intervention using analytical derivations by linear algebra techniques on the transition matrix. Details are provided in Appendix B.3.

One illustration of such an intervention is targeting reduced readmission rates to the ER or hospitalizations. We quantify the cost-savings of an intervention that leads greater adherence to recommended care guidelines, namely an increase in utilization of asthma controller medications and follow-up from the ER or hospitalizations to a physician's office. We assume that we cannot prevent a patient's first visit to the ER or a hospitalization, and that ER visits or hospitalizations that occur after a PO visit or prescription fill are due to an emergency, and thus is not preventable. Therefore, we simply reduce the probability

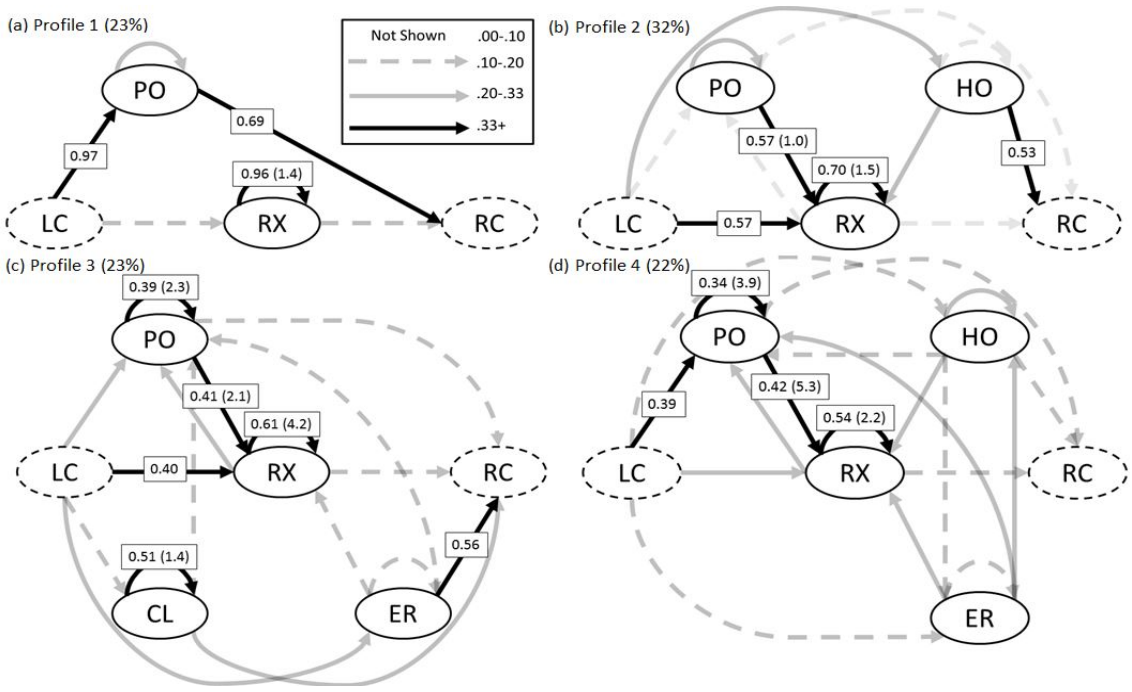

Figure 13: Network graphs of estimated utilization profiles of GA. Transition probabilities are given on each edge along with the average interarrival times measured in months in parentheses.

of readmission into ER and HO and re-allocate that probability into PO and RX. We also leave transitions from ER or HO to CL and NP unchanged if this is the patient's preferred source of treatment.

3.4 Results

In this section we summarize the results of our pilot study on pediatric asthma patients on Medicaid in GA and NC for the years 2005 through 2009. We begin with 1.8 and 2.4 million total claims in Georgia (GA) and North Carolina (NC) for patients with a primary diagnosis of asthma which are translated into 754,597 and 1,224,579 visits for GA and NC, respectively.

3.4.1 Graphical Representations

Figures 13 and 14 are visual representations of the estimated utilization profiles as probabilistic network graphs. We only include high-traffic nodes in these graphs, such that 90%

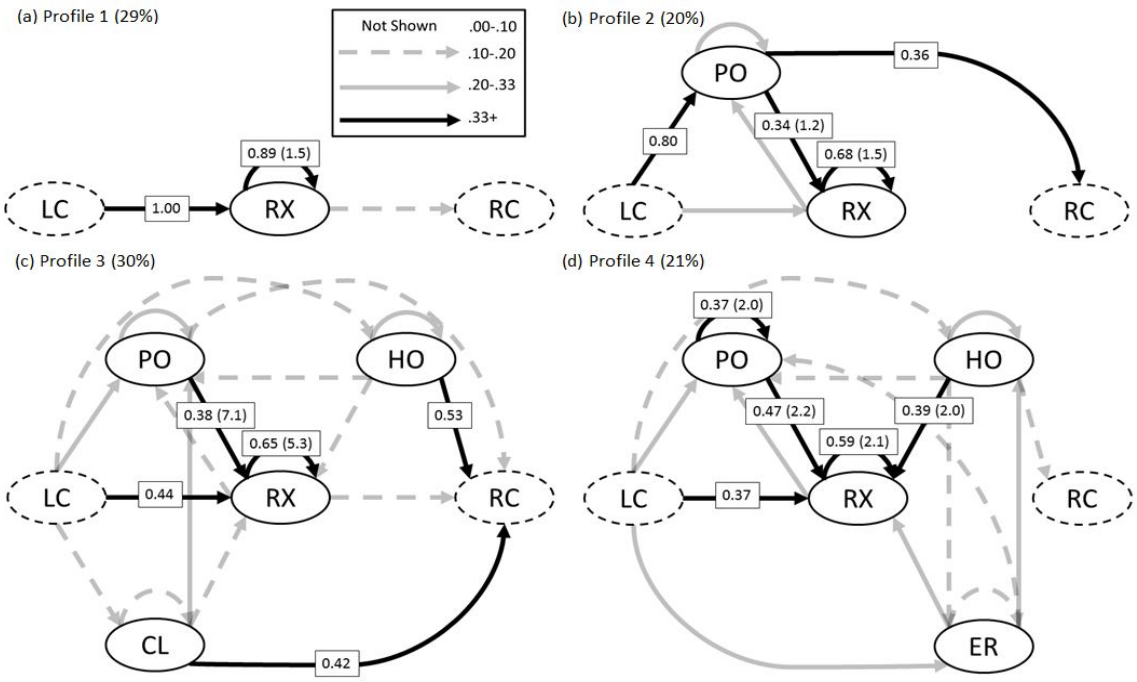

Figure 14: Network graphs of estimated utilization profiles of NC. Transition probabilities are given on each edge along with the average interarrival times measured in months in parentheses.

of the overall volume of encounters is summarized. The nodes are labeled by the provider types corresponding to the contributing states of the utilization sequences in each underlying profile and edges are a visual representation of the the estimated transition probabilities between nodes or states. We provide the complete set of interarrival times between the different states in the tables in Appendix B.4. The legend describes our choice for visualization of the transitions between providers based on the transition probabilities between states. Across all networks, we include the LC (left censoring) and RC (right censoring) nodes specifying the beginning (2005) and the end (2009) year of the study.

3.4.2 Utilization Networks for GA

The network graphs of the four utilization profiles we highlight from GA are displayed in Figure 14. Our decision to highlight these four utilization profiles is described in Appendix B.5.

Profile 1: For patients in this profile, the initial probability of visiting PO is extremely high (0.97) while the probability of having repeat visits to PO is low, with average interarrival time of 8.2 months. Likewise, the initial probability of a RX prescription is low but the probability of repeat encounters is extremely high (0.96), with an average interarrival time of 1.4 months. There are no directed edges between PO and RX indicating that this profile consists of those patients who either visit PO or RX but not both, and would likely be divided into separate profiles in later iterations of the algorithm.

Profile 2: Patients in this profile have a high expected number of RX encounters, equal to 4.31, and PO visits, equal to 1.09, with a low expected number of HO, equal to 0.52. There are many directed edges into RX with high probability (0.57 - PO $\rightarrow$ RX, 0.70 - repeated RX encounters), with no directed edges between PO and HO. The interarrival times into RX are also low (1.0 month for HO $\rightarrow$ RX, 1.1 months for PO $\rightarrow$ RX, 1.5 months for RX refills), and the interarrival time from HO to PO is 0.6 months. Although HO is present in this profile, repeat admissions into HO are infrequent, with an average interarrival time of 7.3 months.

Profile 3: The expected number of visits is 0.46 for CL visits, 0.5 for ER, 1.84 for PO and 3.30 for RX prescriptions, with many directed edges into PO and RX. The high number of RX prescriptions is due to many directed edges into RX from ER and PO as well as the high probability of repeat encounters with relatively high interarrival times compared to the other profiles of 4.2 months. PO likewise receives a high number of visits because of a large number of directed edges, although with low probability, from the other three provider types. Although ER is present in this profile, the readmission into the ER are infrequent, with 8.5 months on average between visits.

Profile 4: The expected number of visits to ER, HO, and PO are higher than in the previous profiles with 0.72, 1.13, 2.53, respectively, while RX still has a high number of encounters, equal to 4.03. The interarrival times between consecutive RX encounters are low on average at 2.2 months and interarrival times into ER and HO are overall high, with

the lowest being HO $\rightarrow$ HO at 4.0 months.

RX is present in all four profiles, with high expected number of encounters in Profiles 2-4. The PO/RX relationship is highly prevalent, judging by the high transition probabilities between the two.

3.4.3 Utilization Networks for NC

The network graphs of the four utilization profiles we highlight from NC are displayed in Figure 14. Our decision to highlight these four utilization profiles is described in Appendix B.5.

Profile 1: This profile consists of patients primarily on asthma-controlled medication, where the expected number of RX (re)fills is equal to 9.34 over the study period. The probability of RX refills is high at 0.89, while the interarrival time between consecutive RX encounters is low (1.5 months). These patients rarely visit physician offices (less that 10% of the utilization in this profile and hence not present) and they almost never visit ER or have hospitalization. This group of patients could be used as a baseline to compare patients with other utilization profiles.

Profile 2: The expected number of RX encounters are lower in this profile (3.14) than Profile 1, with more expected visits to PO, equal to 2.35. A strong connection between PO and RX is clear, with a stronger directed edge going from PO to RX, implying RX prescription fills after a physician office visit. The probability of RX refills is high, equal to 0.68, with a low average interarrival time of 1.5 months. The average interarrival time between PO visits is higher (6.3 months). Hence, patients in this profile tend to visit physician office more often than those in Profile 1, with insignificant ER utilization or hospitalizations.

Profile 3: Patients in this profile have an overall lower number of visits to RX and PO (equal to 2.51 and 0.97, respectively), while CL and HO add more visits, with an expected number equal to 0.30 and 0.36, respectively. This profile has many similarities to Profile 3 of GA, with much higher average interarrival times between RX encounters equal to 5.3

months, but also with high average interarrival times between HO readmissions equal to 5.8 months. Transitions from HO to PO and to RX have high interarrival times at 9.9 months and 9.7 months, indicating non-adherence to follow-up treatment for controlling asthma.

Profile 4: This profile primarily consists of RX and PO visits, with expected numbers equal to 7.52 and 4.16, respectively, while ER and HO add fewer, with expected numbers equal to 0.99 and 1.16, respectively. There are strong connections between PO and RX, and many directed edges with high probability into RX. Here the average interarrival times between consecutive readmissions to the ER and HO are 4.0 and 2.8, respectively, while the interarrival times between PO and RX encounters are lower, equal to 2.0 and 2.1, respectively. All the interarrival times from ER and HO to PO and to RX are low ranging between 2.2 months and 2.5 months. Hence, patients in this profile display higher variation in their healthcare utilization for asthma than in the other three profiles.

3.4.4 Comparing Utilization in GA & NC

The network graphs for the two states show remarkable similarities between the longitudinal utilization profiles across both states; particularly, Profile 1 of GA and Profile 2 of NC are similar as well as Profiles 3 and 4 of both GA and NC. Other commonalities include the apparent prominent relationship between PO visits and subsequent RX encounters, with high probabilities, indicating well-managed asthma patients. In all but Profile 1 of GA and NC there are directed edges between the two provider types, routinely with high probability and low average interarrival times. Likewise, as shown in Figures 13 and 14, there are no connections between PO or RX and HO or ER with transition probability greater than 0.33. By examining the visits by provider type bar chart in Figure 15, we find that GA has more uniformity and variation between the provider types across the four profiles. The major differences between the two states lie in the high concentration of RX visits in NC (67% versus 54% in GA), and the relatively high proportion of ER and HO visits in GA (13% versus 8% in NC).

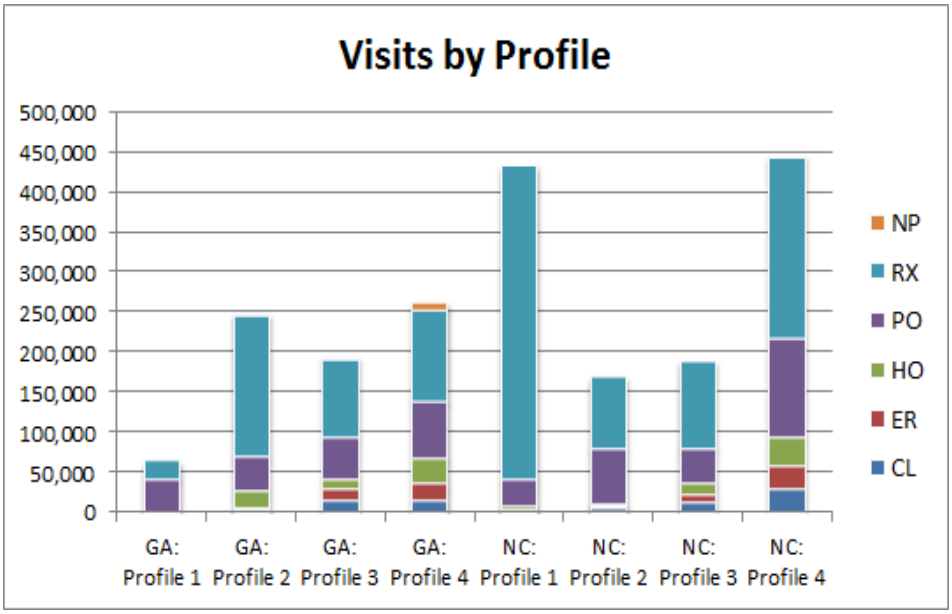

Figure 15: A chart plotting the total number of visits to each provider type from all patients per profile during the years 2005 - 2009.

3.4.5 Evaluating Interventions for Adherence to Recommended Care

The results for quantifying cost-saving interventions are given in Figure 16 and the table in Appendix B.3. Noteworthy findings are that Profiles 4 of GA and NC, both of which are extremely varied in terms of utilization of different provider types, have the most potential for cost savings. Overall the potential for cost-saving interventions seems higher in GA for Profiles 2 and 3. Notably, because the patients in Profile 1 of GA only have PO and RX visits there are no cost-savings for these patients, highlighting the fact that cost-saving interventions should be applied to those most at risk for variational healthcare utilization. The overall potential cost-savings for a 25% increase in adherence to recommended care guidelines results in a cost-savings of $2.24M in GA and $2.18M in NC.

3.5 Conclusion

In this chapter we introduce a data science framework for extracting, analysing and integrating large, highly-sensitive claims information for deriving simple graphical interpretations of healthcare utilization. The objective is to characterize and visualize underlying profiles

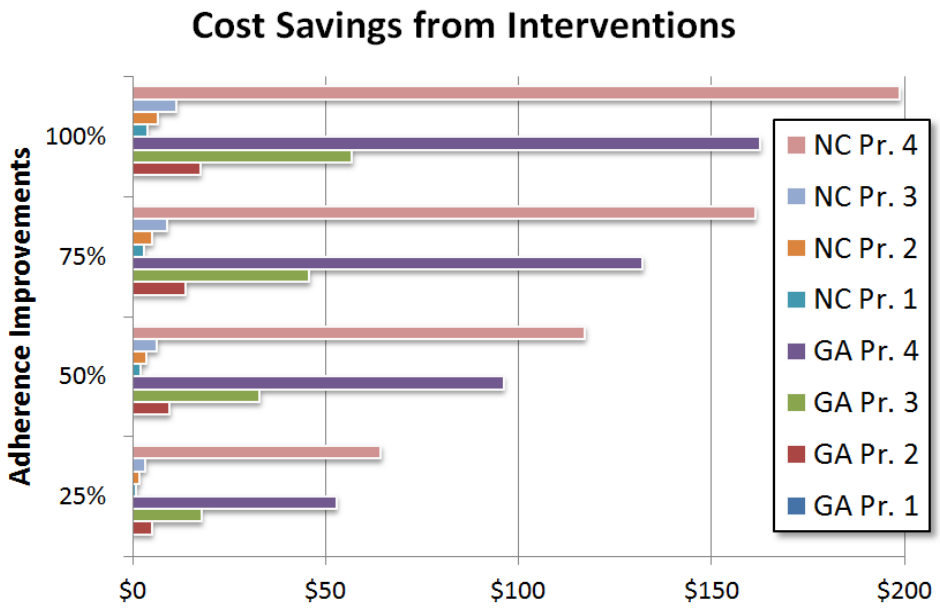

Figure 16: A chart plotting potential cost-savings at different levels of adherence improvements to recommended care guidelines.

of patient-level utilization behaviors. Because the approach is model-based, it allows discovery of underlying probabilistic relationships of patients' transitions between different provider types and can be used to analyze "what-if" interventional scenarios to examine the effects of changes in the network of care and the associated expected cost-savings. Our framework begins with manipulation and processing of large flat files of administratively coded claims into meaningful data in the form of streamlined utilization sequences. The patient-level utilization sequences are then the input for a scalable model-based clustering analysis for discovering the underlying utilization profiles. Our methods are both rigorous and general, with applicability beyond the case study in this chapter.

We pilot our study with Medicaid claims data across five years, 2005-2009. We extract data for only a subset of patients, particularly, asthma-diagnosed children older than 3, and we focus on two states, Georgia and North Carolina.

Our study emphasis is on healthcare utilization as it is at the core of critical aspects of healthcare delivery, including healthcare access, expenditure and cost, prevention and chronic disease management (98). We also focus on the Medicaid system as the test bed

for our analysis because caring for the disadvantaged populations, particularly Medicaid children, is one priority of the current health policies in the United States, with potential impact on reducing health and healthcare disparities, and on containing the associated costs (112). Medicaid constitutes the primary source of coverage for low-income children in the United States.

An important aspect of the Medicaid benefits system is that its implementation and reimbursement structure vary by state. Due to these state-based differences in the implementation, the effectiveness of the program also varies greatly by state. Thus, by comparing utilization of care across states one can reveal the impact of these variations on the care ecosystem.

Particularly, we chose Georgia and North Carolina for this comparison because the demographics of the pediatric populations are very similar (30-50% minority population (111) and approximately $37,000 Per Capita Personal Income 2012 (17)), although they have different care-coordinated systems. While North Carolina has a state-coordinated Medicaid system, Georgia's Medicaid patients are primarily managed by three Medicaid Managed Care Organizations with a reasonably small percentage of children under the fee-for-service care practice (51). According to the 2007 ranking of states based on the Medicaid eligibility, scope of services, quality of care, and reimbursement obtained by the Public Citizen Health Research Group (5), North Carolina is ranked in the second quartile and Georgia is ranked in the third quartile.

With similar Medicaid populations but different care coordination systems and effectiveness rankings, we find some striking similarities in the longitudinal (multi-year) utilization behaviors for pediatric asthma care.

Both states have an underlying profile including patients primarily visiting a physician's office (Profile 1 in GA and Profile 2 in NC). Likewise Profile 2 in GA and Profile 1 in NC have a high probability of filling a prescription for asthma-controlled medication (higher for North Carolina than for Georgia) but a lower probability of PO visits (lower for North

Carolina than for Georgia). The transition probabilities are low connecting ER/HO to PO and vice versa, with stronger links between medication and physician offices, suggesting that the more variational clusters also include a proportion of patients that primarily visit physician office with sporadic (low probability) visits to the emergency department or with hospitalizations. This indicates that the majority of patients utilizing the physician's office in the variational profiles adhere to a great extent to evidence-based practices for asthma care.

A third noteworthy finding is the prevalence of clinic visits in Profile 3 for both states, where clinics refer to federally-qualified and rural health clinics. This is not surprising since Medicaid children rely heavily on care from clinics located in underserved areas. Importantly, patients with clinic visits have a higher probability to follow up with a phyician's office visit rather than visit ER or have an hospitalization for both states.

Finally, in both GA and NC, patients belonging to Profile 4 are not only the most costly, but are also those with the most potential for cost-savings. Cost-savings from interventions targeting a higher follow-up from the ER or HO to PO or RX visits can lead to a 5% reduction in GA and a 3% reduction in NC in cost for patients belonging to this profile with just a 25% adherence improvement.

An important dissimilarity across the two states is the proportion of patients with regular PO visits (physician's office visits make up 59% of non-RX (re)fills in Georgia, while they make up 64% of non-RX (re)fills in North Carolina). Profiles 1 & 2 in North Carolina contain almost 50% of patients where those patients primarily utilize the physician's office along with RX encounters; in contrast, Profile 1 (21%) and approximately half of Profile 2 (roughly 16%) patients in Georgia utilize the physician's office almost exclusively. Hence, in aggregate, North Carolina has around 50% more of the patient population than Georgia that visit physician's offices to the exclusion of other provider types on a regular basis.

In both states, the average interarrival time of RX fills is very similar averaging 1.5 months in Profiles 1 and 2 of both states, 2.2 months in Profile 4, with Profiles 3 having

the longest average interarrival times of 4.2 months in GA and 5.3 months in NC. When comparing the graphical networks, we also find that physician's office visits and medication fills nodes are represented strongly for both North Carolina and Georgia across all four profiles. On the other hand, emergency department or hospitalization nodes appear to serve only as intermediary connections for both Georgia and North Carolina, with a stronger presence in Georgia.

This study has several limitations. One shortcoming in using claims data to infer utilization is that while we seek to make inference on an entire subpopulation, we capture realized and not potential utilization of the system (28; 98). First, the MAX files only include claims that have been reimbursed. Second, not all Medicaid-eligible children are enrolled or they have intermittent enrollment. Moreover, there will be a percentage of Medicaid-enrolled children who are undiagnosed due primarily to lack of healthcare access. Therefore, estimates on the healthcare utilization are likely to be to be biased, particularly for the Medicaid population, where certain subgroups have difficulty in maintaining Medicaid coverage or are susceptible to particularly disparate utilization (28; 87). We provide further analysis of the enrollment patterns of the Medicaid children in our study in Appendix B.6.

While our model and its estimation and selection methods are computationally attractive, they can be extended further for relaxing some of the underlying assumptions. First, we do not include the mean times until the first event and the mean times between the last event because they are biased estimates of complete lifetimes due to the censored nature of our data. In doing so, we are unable to completely determine the consistency with which patients visit providers. For instance, with unbiased estimates of the arrival to the first event it would be clear if a patient waits a long time between groups of consecutive visits or utilizes the system at a fairly homogeneous rate across the complete study time span. Furthermore, in order to produce simple visualizations and minimize computational costs we assume the interarrival times to be exponentially distributed, conditional on the visit

type. More importantly, it is likely that covariates including age, condition severity, comorbidities, enrollment status and access play a role in the frequency of the visits. However, this method does not capture the potential effects of these covariates on utilization.

Despite these shortcomings, our model allows for reduction of high-dimensional utilization data into a one-dimensional vector containing cluster memberships, thus providing the means for policy-makers to easily simulate or visualize healthcare utilization and further study explanatory variables that could explain the variations across patient-level utilization profiles.

Even though this study has several limitations, it has some important implications for health care providers and policy makers. Importantly, following the care practice recommendations, if a child visits the emergency department for asthma care then he/she needs to be referred back to primary care (79). In both Georgia and North Carolina, the transition from emergency department or from hospitalization to physician's office varies across utilization profiles, with very low probability of physician's office follow-up visits for the patients using emergency department and hospitalization regularly. Those follow-up visits vary with the patient's profile, indicating that different interventions should be considered for each of the profile of patients. More importantly, in both states, patients who are visiting emergency department regularly for asthma care are few, with long periods of time between readmissions.

Asthma-controlled medication uptake is strongly connected with physician's office visits across three profiles, and in one profile where it is not, patients are regularly taking medication with no significant severe outcomes recorded. From the strength of the links between physician's office and medication (re-)fills, and lack of connection of those two event types to the emergency department, those patients who visit a physician's office on a regular basis while staying on asthma-controlled medication are unlikely to have emergency department visits in both states. This finding provides evidence that asthma can be controlled with regular physician's office visits and medication, with the potential of

eliminating costly emergency department visits.

CHAPTER IV

MODELING HETEROGENEITY IN HEALTHCARE UTILIZATION USING MASSIVE MEDICAL CLAIMS DATA

In this chapter we introduce a modeling approach for characterizing heterogeneity in healthcare utilization using massive medical claims data. We first translate the medical claims observed for a large number of patients and across five years into patient-level discrete events of care called *utilization sequences*. We model the utilization sequences using an exponential proportional hazards mixture model to capture heterogeneous behaviors in patients' healthcare utilization. The objective is to cluster patients according to their longitudinal utilization behaviors and to determine the main drivers of variation in healthcare utilization while controlling for the demographic, geographic, and health characteristics of the patients. Due to the computational infeasibility of fitting a parametric proportional hazards model for high-dimensional, large sample size data we use an iterative one-step procedure to estimate the model parameters and impute the cluster membership. The approach is used to draw inferences on utilization behaviors of children in the Medicaid system with persistent asthma across six states. We conclude with policy implications for targeted interventions to improve adherence to recommended care practices for pediatric asthma.

4.1 Introduction

Appropriate utilization of the healthcare system is a positive tenet in preempting severe health outcomes and is the basis for more effective healthcare practices (20; 69; 87). A well-managed health condition and adherence to recommended care practices typically result in reduced use of the emergency room and hospitalizations, thus leading to better health outcomes and less costly care for patients with chronic diseases (69). Characterizing

utilization behaviors and studying the drivers of variations in healthcare utilization can suggest targeted interventions for improving chronic disease management.

Understanding and managing healthcare utilization is now possible with the advent of patient-detailed health records and claims data, available from healthcare providers, and public or private insurers (18; 27; 64; 72). The largest insurer in the United States, the Centers for Medicare and Medicaid Services (CMS), has provided a platform for acquiring such data in a standardized format across all states. Typically, CMS claims data include not only healthcare services information such as the type and place of care, services provided, diagnosis and procedure codes but also patient-specific information such as demographics for more than 100 million patients. The claims data are presented in formats that are amenable for administrative purposes rather than research, hence they require substantive efforts of data manipulation and processing.

Particularly for the Medicaid program, the CMS claims data are only available as identifiable patient health information divided into multiple files depending on the healthcare services provided, by year and by state. The patient identification is unique across all files allowing researchers to trace patients longitudinally across their Medicaid-reimbursed healthcare encounters. Thus, in order to characterize longitudinal healthcare utilization at the patient level, the Medicaid claims data need to be mapped into longitudinal sequences of care events and joined with patient characteristics requiring multiple sources of information and database manipulations. After this initial translational process, statistical modeling can be applied to make inference on the heterogeneity in healthcare utilization.

In this study, we seek to make inferences on healthcare utilization for Medicaid-enrolled children diagnosed with persistent asthma across six states, including five southeast states, Georgia, Louisiana, Mississippi, North Carolina and Tennessee with comparison to Minnesota. Medicaid-eligible patients typically belong to disadvantaged socioeconomic groups and are, therefore, more likely to utilize the healthcare system disparately (89). We focus on asthma as it is the most prevalent respiratory chronic condition for children (26). The

study population includes more than 400,000 children with approximately 6 million asthma events. The utilization sequences are complemented by patient characteristics including demographics, enrollment characteristics, urbanization environment of their residence, spatial access to primary care (37) and clinical risk group (CRG) derived using the 3M Core Grouping Software (1) among others. Substantive computational challenges arise in deriving inferences from such high-dimensional, massive datasets within a restrictive data environment in place for identifiable protected health information (PHI).

By studying the CMS claims data we are able to infer utilization patterns from a large, standardized data source for a large number of children. Healthcare utilization has been the primary topic of interest for many healthcare studies with most explaining the frequency of utilization with respect to patient characteristics and other determinants of utilization for various conditions, typically relying on statistical methods such as regression or general linear models, see (6; 41; 49; 96; 97) among many others.

In Chapter III, we used model-based clustering approaches for characterizing heterogeneity in patient utilization where the underlying model is a Markov model with a finite state space. While this model lends itself to computationally feasible inference and visualization, it is limited in many ways. The first limitation is that it does not account for missing data in utilization sequences. Utilization sequences derived from claims data are often subject to data *censoring*, referring to missed events when a patient may not be eligible for Medicaid benefits or events occurring outside the study time period. The second limitation is an insufficient treatment of the effects of different event types on the prevalence of visits to a specific provider type. Each patient potentially visits multiple provider types repeatedly over the time period of interest. Thus, we have a competing-risks, repeated-events framework. A third limitation involves incorporating demographic and health-related covariates into the model.

To address these limitations, we will combine techniques from survival analysis and

statistical clustering analysis to measure the rate at which patients in the study population receive treatment for asthma from various provider types. A central theme in survival analysis is that of handling censored data. Cox's proportional hazards model allows for the inclusion of possibly censored survival times in the likelihood function while also incorporating knowledge on characteristics of the patient, modelled as a linear function of covariates. In this study, we will fit a parametric proportional hazards model to find the rate at which pediatric asthma patients visit different provider types given variables such as access to care, their current overall health condition, demographic variables, differences in state-based Medicaid programs, and history of healthcare utilization. We assume a mixture of proportional hazard models to capture heterogeneity in utilization behaviors. Using this model, we will derive three primary outputs from which we aim to determine the main contributors to variations in healthcare utilization: the posterior probabilities that a patient belongs to a specific *cluster* of patients given a set of control variables and utilization history, parameter estimates of the control variables that measure the effects of control covariates on the event hazard rates, and parameter estimates for the explanatory variables that can be used to evaluate the impact of potential interventions on the rate of healthcare visits.

This method was inspired by the complexity of the healthcare data set that we study and has roots in the survival analysis literature, particularly an adaptation of the Cox model to parametric counting process data (11) and models for heterogeneity in discrete choice models and survival analysis (9; 14; 31; 39; 45; 46; 92; 114). Two areas that are closely related to the proposed methodology are those of determining 'long-term' survivors in a cohort (34; 55; 70; 107) as well as the use of the multivariate Weibull mixture model to capture heterogeneity in duration data (15; 33; 65; 75; 76; 77; 78). We look to extend the contributions of these authors by generalizing the proportional hazards cure model to allow for different rates for multiple (more than two) subpopulations. Furthermore, while mixture

modelling is prevalent in the literature, few authors incorporate explanatory and/or controlling factors, see (15; 65; 78). By bringing the computational feasibility of the estimation algorithm to bear, we can analyze massive, high-dimensional datasets. This is a promising contribution in light of the exponential growth of healthcare data (32) and demonstrates the ability to apply these methods wherever high-dimensional counting process data are available.

The remaining structure of this chapter is as follows: in Section 4.2 we further summarize the target population and the covariates we include in this study, in Section 4.3 we present the model and model estimation techniques, in Section 4.4 we present results from our application to pediatric asthma patients in the Medicaid system, and we conclude with a discussion in Section 4.5. We provide additional derivations and details on the results for the motivating case study in the Appendices.

4.2 Data

We begin by translating the Medicaid Analytic Extract (MAX) claims data into patient utilization data. Our study population consists of all Medicaid-enrolled children ages 4-18 with persistent asthma (115) from Georgia (GA), Louisiana (LA), Mississippi (MS), Minnesota (MN), North Carolina (NC), and Tennessee (TN) between 2005 and 2009. Children age 0-3 are not included in the study due to inconsistency in asthma diagnosis at this age. We only include children with persistent asthma, that is, children that have at least one emergency room visit or hospitalization with a diagnosis of asthma, at least three outpatient visits with a diagnosis of asthma, or a prescription fill for asthma controller medications. In total we have 426,400 patients, approximately 4 million healthcare events, and approximately 1.5 million patient-years in this study. For a table with summary statistics on the data see Tables 8-10 in Appendix C.1.

In order to specify the provider type, we use a combination of *Place of Service* Code and *Type of Service* Code from the MAX data files. We abbreviate the provider types in

the following manner: clinic visits (CL), emergency room and outpatient hospitalizations (ER), inpatient hospitalizations (HO), physician's office visits (PO), and nurse practitioner care in a physician's office (NP). In addition, we model a claim where a patient visits the pharmacy to fill a prescription for asthma controller medication (RX) as a unique event type.

In addition to the provider types and timestamps of the visits, we also extract patient demographic, zip code, and health-related information such as age, Medicaid eligibility status and health condition or clinical risk group (CRG) derived using the 3M Core Grouping Software (version 2014.3.2 with the Clinical Risk Groups version 1.12) from the MAX data. Using the zip code of the patient, we include additional variables such as the state of residence, urbanization level of a patient's residence zip code derived using the RUCA categorization (74) and travel distance to pediatric primary care derived using optimization models (37). We only consider access to primary care since it is the most prevalent non-emergency care type for Medicaid-insured children diagnosed with asthma as we show in Chapter III.

We divide the covariates into two groups: *control* and *explanatory*. The control variables are: age group (4-5, 6-14, 15-17), race (white, black, and other), overall health condition of the patient (healthy: CRG 1, minor chronic: CRG 2-4, chronic: CRG 5-7, and severe: CRG 8-9, determined by the 3M software), reason for Medicaid eligibility (disabled, foster care and income-based) and the last event type to account for the patient's healthcare history. The explanatory variables include the state of residence of the patient, urbanicity categorized as urban (RUCA 1-3), suburban (RUCA 4-6) and rural (RUCA 7-10) and travel distance to pediatric primary care.

A summary of the observed vectors of data is given below. Throughout this chapter bold typeface will be used for vectors and matrices.

- $\boldsymbol{H_r}(t) = \{H_{r1}(t), \ldots, H_{r|S|}(t)\}$ the count of visits for each patient r to providers of type $s \in \{1, \ldots, |S|\}$ over the time t with a maximum of five years. In our study,

we consider $|S| = 6$ event types (CL, ER, HO, PO, NP, and RX). $\vec{\boldsymbol{H}}$ contains all counting processes for all patients.

- $\boldsymbol{D_r}$ and $\boldsymbol{E_r}$ are column vectors of observed covariates corresponding to the control and explanatory variables, respectively. $\vec{\boldsymbol{D}}$ and $\vec{\boldsymbol{E}}$ contain all covariates for all patients.

A Word on the Time Domain: In this chapter we are interested in the time-to-event data and the effect of patient historical and demographic information on the times between events of the same type. We will denote the standard time domain with t and the time since the last event or re-enrollment time as τ. Changes from enrolled to unenrolled are considered to be censored lifetimes. See Appendix C.2 for an example.

4.3 The Latent Variable Proportional Hazards Model

In this section we begin by motivating the use of survival analysis for this particular problem. We then introduce the mixture model formulation and demonstrate the use of the expectation-maximization (EM) algorithm to estimate the mixture model parameters. Finally, we present a computationally efficient, iterative algorithm to estimate the proportional hazard and utilization-choice model parameters, which applies to high-dimensional, large sample size data.

4.3.1 The Proportional Hazards Model

Consider a counting process $N(t)$ counting the number of events up to time t. Then Aalen (2) and Andersen (3) show that $N(t)$ has a random hazard process $\lambda(t)$ defined as

$$\lambda(t) = \lim_{h \to 0} \Pr(T < t + h | T > t),$$

where T is a random variable for the time of the event. Let $f(t)$ be the probability density function for an event at time t and $S(t)$ be the survival function up to time t. Then we can

relate the three functions with the following formula:

$$f(t) = \lambda(t)S(t).$$

The Cox regression model (21) specifies the hazard rate given a set of time-varying covariates $x(t)$ via the equation

$$\lambda(t|x(t)) = \lambda_0 \exp\{\boldsymbol{\beta}^\intercal x(t)\},$$

where λ_0 is a fixed underlying baseline hazard function. This model is typically referred to as the 'proportional-hazards' model due to the fact that the hazard rate of an event at time t for different subpopulations are proportional to each other.

Our model of the hazard rate for an event of type s can be written as

$$\lambda_s(\tau|\boldsymbol{D_r}(\tau)) = \lambda_{rs}(\tau) = \exp\{\boldsymbol{\beta_s}^\intercal \boldsymbol{D_r}(\tau)\},$$

where $\boldsymbol{\beta_s} = [\beta_{0s}, \beta_{1s}, \ldots, \beta_{Ps}]$. Thus, the baseline hazard function is $\lambda_0 = \exp(\beta_0)$. Furthermore, the vector $\boldsymbol{D_r}$ may vary with time because it includes dummy variables for the last event type as well as the health status of the patient which may change annually. Therefore,

$$S_s(\tau|\boldsymbol{D_r}(\tau)) = S_{rs}(\tau) = \exp\{-\tau \exp[\boldsymbol{\beta_s}^\intercal \boldsymbol{D_r}(\tau)]\},$$

and

$$f_s(\tau|\boldsymbol{D_r}(\tau)) = f_{rs}(\tau) = \lambda_{rs}(\tau)S_{rs}(\tau).$$

4.3.1.1 Choice of Baseline Hazard Function

Our choice of the exponential for the interarrival times distribution is due to the distributional characteristics and favorable analytic properties of the exponential survival model. In the field of survival analysis there are two primary choices for the baseline hazard model: a nonparametric baseline hazard function and a parametric baseline hazard function such as the exponential, Weibull or log-logistic, for instance. We choose a parametric baseline

because we must force the baseline hazard function to be unimodal, otherwise heterogeneous subpopulations may be incorrectly grouped together. The favorable properties of the exponential proportional hazards model are discussed in Appendix C.2 and C.3.

4.3.2 The Latent Cluster Model

The problem we are trying to solve is that of clustering similar patients based on their utilization patterns, estimating the coefficients corresponding to the control variables, and determining the factors that explain the variations in longitudinal utilization behaviors. Let $\boldsymbol{Z_r}$ be a multinomial random variable denoting the latent cluster membership of patient r taking values 0 and 1, where $Z_{rk} = 1$ if patient r belongs to cluster k. Given that $Z_{rk} = 1$ the probability that each patient contributes is

$$\Pr(\boldsymbol{H_r}|\boldsymbol{D_r}, Z_{rk} = 1) = \prod_{s=1}^{|S|} \prod_{l_r=1}^{L_r} f_{rks}(\tau_{l_r})^{\delta_s(\tau_{l_r})} \times S_{rks}(\tau_{l_r})^{1-\delta_s(\tau_{l_r})}, \tag{3}$$

where τ_{l_r} is the l^{th} interarrival time between consecutive events, censoring, or re-enrollment times for patient r, and $\delta_s(\tau)$ is an indicator function taking value 1 if patient r visits provider type s at time τ and 0 otherwise.

Following the cure model of (34; 55; 70; 107), we want to model the probability that patient r belongs to cluster k given the explanatory variables $\boldsymbol{E_r}$. Let $Z_{rk|\boldsymbol{E_r}}$ be a multinomial random variable denoting the latent cluster membership of patient r with explanatory variables $\boldsymbol{E_r}$. We assume that the probability that $Z_{rk|\boldsymbol{E_r}} = 1$ follows a multinomial logistic regression model:

$$\Pr(Z_{rk|\boldsymbol{E_r}} = 1) = \Pr(Z_{rk} = 1|\boldsymbol{E_r}) = \pi_{rk} = \frac{\exp\{\boldsymbol{E_r}^\intercal \boldsymbol{b_k}\}}{1 + \sum_{\kappa=1}^{K-1} \exp\{\boldsymbol{E_r}^\intercal \boldsymbol{b_\kappa}\}}, \text{ for } k < K,$$
$$\text{and } \pi_{rK} = \frac{1}{1 + \sum_{k=1}^{K-1} \exp\{\boldsymbol{E_r}^\intercal \boldsymbol{b_k}\}}. \tag{4}$$

Combining Equations 3 and 4, we can derive the likelihood function for $\vec{\boldsymbol{b}}$ and $\vec{\boldsymbol{\beta}}$ as

$$L(\vec{\boldsymbol{b}}, \vec{\boldsymbol{\beta}}) = \prod_{r=1}^{R} \prod_{k=1}^{K} \pi_{rk} \Pr(\boldsymbol{H_r}|\boldsymbol{D_r}, Z_{rk} = 1). \tag{5}$$

This model *controls* for the effects of the control covariates, $\boldsymbol{D_r}$, and allows the cluster-specific baseline hazard of an event to vary while *explaining* the causes of variations due to the explanatory variables $\boldsymbol{E_r}$.

4.3.3 The EM Algorithm

Together $[\vec{\boldsymbol{Z}}, \vec{\boldsymbol{H}}, \vec{\boldsymbol{D}}, \vec{\boldsymbol{E}}]$ with $\vec{\boldsymbol{Z}} = [\boldsymbol{Z_1}, \ldots, \boldsymbol{Z_R}]$ form the complete information on a patient's utilization history. However, $\vec{\boldsymbol{Z}}$ is unknown and must be inferred from $[\vec{\boldsymbol{H}}, \vec{\boldsymbol{D}}, \vec{\boldsymbol{E}}]$. We will use the EM algorithm (25) to estimate the probability that patient r belongs to cluster k, $\Pr(Z_{rk} = 1)$, for all r, k.

Under the framework of complete information we can revise Equation 5 to get the complete likelihood function:

$$L_C(\vec{\boldsymbol{b}}, \vec{\boldsymbol{\beta}} | \vec{\boldsymbol{Z}}) = \prod_{r=1}^{R} \prod_{k=1}^{K} \pi_{rk}^{Z_{rk}} \prod_{s=1}^{|S|} \prod_{l_r=1}^{L_r} \left[f_{rks}(\tau_{l_r})^{\delta_s(\tau_{l_r})} \times S_{rks}(\tau_{l_r})^{1-\delta_s(\tau_{l_r})} \right]^{Z_{rk}}$$
$$= L_C(\vec{\boldsymbol{b}} | \vec{\boldsymbol{Z}}) \times L_C(\vec{\boldsymbol{\beta}} | \vec{\boldsymbol{Z}}).$$

Due to the fact that the complete likelihood function can be split between a likelihood for $\vec{\boldsymbol{b}}$ and $\vec{\boldsymbol{\beta}}$ we can divide the model estimation procedures into three parts: estimating the probability that patient r belongs to cluster k (E-step), and estimating separately the proportional hazards coefficients and the multinomial logistic coefficients (M-step).

4.3.3.1 The E-Step

In the E-step we find the expected values of the missing values $\vec{\boldsymbol{Z}}$ with respect to the distribution given the current estimates for the model parameters, $\vec{\boldsymbol{b}}^{(\boldsymbol{m})}$ and $\vec{\boldsymbol{\beta}}^{(m)}$:

$$Z_{rk}^{(m+1)} = E(Z_{rk} | \vec{\boldsymbol{b}}^{(\boldsymbol{m})}, \vec{\boldsymbol{\beta}}^{(\boldsymbol{m})}) = P(Z_{rk} = 1 | \vec{\boldsymbol{b}}^{(\boldsymbol{m})}, \vec{\boldsymbol{\beta}}^{(\boldsymbol{m})})$$
$$= \frac{\prod_r \pi_{rk}^{(m)} \prod_s \prod_{l_r} f_{rks}^{(m)}(\tau_{l_r})^{\delta_s(\tau_{l_r})} \times S_{rks}^{(m)}(\tau_{l_r})^{1-\delta_s(\tau_{l_r})}}{\sum_{\kappa=1}^{K} \prod_r \pi_{r\kappa}^{(m)} \prod_s \prod_{l_r} f_{r\kappa s}^{(m)}(\tau_{l_r})^{\delta_s(\tau_{l_r})} \times S_{r\kappa s}^{(m)}(\tau_{l_r})^{1-\delta_s(\tau_{l_r})}}.$$

After performing the E-step we take the current estimates, $\vec{\boldsymbol{Z}}^{(\boldsymbol{m+1})}$, and use them to calculate the next step estimates for the parameters in the proportional hazards and multinomial logistic regression model.

4.3.3.2 The M-Step

Assuming that the probability distribution of events follows an exponential distribution we have that $f_{rks}(\tau) = \lambda_{rks}(\tau)\exp\{-\tau\lambda_{rks}(\tau)\}$ and $S_{rks}(\tau) = \exp\{-\tau\lambda_{rks}(\tau)\}$, where $\lambda_{rks}(\tau) = \exp\{\boldsymbol{\beta_{ks}}^\intercal \boldsymbol{D_r}(\tau)\}$. Then the total likelihood function for all patients, $L_C(\vec{\boldsymbol{\beta}})$, can be written as

$$\begin{aligned} L_C(\vec{\boldsymbol{\beta}}|\boldsymbol{Z_r}) &= \prod_r \sum_k Z_{rk} \prod_s \prod_{l_r} \exp\{\delta_s(\tau_{l_r})\boldsymbol{\beta_{ks}}^\intercal \boldsymbol{D_r}(\tau_{l_r}) - \tau_{l_r}\exp\{\boldsymbol{\beta_{ks}}^\intercal \boldsymbol{D_r}(\tau_{l_r})\} \\ &= \exp\left\{\sum_r \sum_k \sum_s \sum_{l_r} Z_{rk}\delta_s(\tau_{l_r})\boldsymbol{\beta_{ks}}^\intercal \boldsymbol{D_r}(\tau_{l_r}) - \tau_{l_r}\exp\{\boldsymbol{\beta_{ks}}^\intercal \boldsymbol{D_r}(\tau_{l_r})\right\}, \end{aligned}$$

where the equality holds in the second line because for $\boldsymbol{Z_r}$ only one entry is equal to one and all others are zero. Now, set $\boldsymbol{\beta_{ks}}^\intercal = [\beta_{0ks}, \boldsymbol{\beta_s}^\intercal]$, where $\boldsymbol{\beta_s}^\intercal = [\beta_{1s}, \ldots, \beta_{Ps}]$. Recall that $\boldsymbol{\beta_s}$ are common across all clusters $k \in \{1, \ldots, K\}$. Then the complete log likelihood function can be written as:

$$\ell_C(\vec{\boldsymbol{\beta}}|\vec{\boldsymbol{Z}}) = \sum_{r,k,s,l_r} \left[\delta_s(\tau_{l_r})Z_{rk}\boldsymbol{\beta_{ks}}^\intercal \boldsymbol{D_r}(\tau_{l_r}) - \tau_{l_r}Z_{rk}\exp\left\{\boldsymbol{\beta_{ks}}^\intercal \boldsymbol{D_r}(\tau_{l_r})\right\}\right].$$

Before moving onto the iterative procedure for estimating $\vec{\boldsymbol{\beta}}$ and $\vec{\boldsymbol{b}}$ we must perform some derivations first on the complete likelihood function for $\vec{\boldsymbol{b}}$, following the arguments of (23):

$$\begin{aligned} L_C(\vec{\boldsymbol{b}}|\vec{\boldsymbol{Z}}) &= \prod_r \prod_k \pi_{rk}^{Z_{rk}} = \prod_r \left[\left(\prod_{k=1}^{K-1} \pi_{rk}^{Z_{rk}}\right) \times \pi_{rK}^{1-\sum_{k=1}^{K-1} Z_{rk}}\right] \\ &= \prod_r \left[\left(\prod_{k=1}^{K-1} \pi_{rk}^{Z_{rk}}\right) \times \frac{\pi_{rK}}{\sum_{k=1}^{K-1} \pi_{rK}^{Z_{rk}}}\right] = \prod_r \left[\left(\prod_{k=1}^{K-1} \frac{\pi_{rk}}{\pi_{rK}}\right)^{Z_{rk}} \times \pi_{rK}\right] \\ &= \prod_r \left[\left(\prod_{k=1}^{K-1} \exp\{\boldsymbol{b_k}^\intercal \boldsymbol{E_r}\}^{Z_{rk}}\right) \times \left(1 + \sum_{k=1}^{K-1} \exp\{\boldsymbol{b_k}^\intercal \boldsymbol{E_r}\}\right)^{-1}\right]. \end{aligned}$$

Therefore the log likelihood function is

$$\ell_C(\vec{\boldsymbol{b}}|\vec{\boldsymbol{Z}}) = \sum_r \left[\sum_{k=1}^{K-1} (Z_{rk}\boldsymbol{b_k}^\intercal \boldsymbol{E_r}) - \log\left(1 + \sum_{k=1}^{K-1} \exp\{\boldsymbol{b_k}^\intercal \boldsymbol{E_r}\}\right)\right].$$

4.3.3.3 An Iterative Solution to the Likelihood Equations

Now we employ the iterative procedure of (36; 71; 73; 121) to estimate the parameters $\vec{\boldsymbol{b}}$ and $\vec{\boldsymbol{\beta}}$. The main idea of the algorithm is to split the large, computationally extensive task of estimating $\vec{\boldsymbol{\beta}}$ and $\vec{\boldsymbol{b}}$ into many single estimation steps. Therefore, in order to find the next-step estimate for $\beta_{ps}, p \in \{1, \ldots, P\}$ given the current estimates $\vec{\boldsymbol{\beta}}^{(\boldsymbol{m})}$ and $\vec{\boldsymbol{Z}}^{(\boldsymbol{m+1})}$, we take the derivative of $\ell_C(\vec{\boldsymbol{\beta}})$ with respect to a single β_{ps}:

$$\ell_C^{(1)}(\beta_{ps}) = \left.\frac{\partial \ell_C(\vec{\boldsymbol{\beta}}|\vec{\boldsymbol{Z}}^{(\boldsymbol{m+1})})}{\partial \beta_{ps}}\right|_{\vec{\boldsymbol{\beta}}=\vec{\boldsymbol{\beta}}^{(\boldsymbol{m})}}$$
$$= \sum_{r,k,l_r} \left[\delta_s(\tau_{l_r}) Z_{rk}^{(m+1)} D_{rp}(\tau_{l_r}) - \tau_{l_r} Z_{rk}^{(m+1)} D_{rp}(\tau_{l_r}) \exp\left\{\boldsymbol{\beta}_{\boldsymbol{ks}}^{(\boldsymbol{m})\intercal} \boldsymbol{D}_{\boldsymbol{r}}(\tau_{l_r})\right\}\right]$$
$$= \sum_{r,l_r} \left[\delta_s(\tau_{l_r}) D_{rp}(\tau_{l_r})\right] - \sum_{r,k,l_r} \left[\tau_{l_r} Z_{rk}^{(m+1)} D_{rp}(\tau_{l_r}) \exp\left\{\boldsymbol{\beta}_{\boldsymbol{ks}}^{(\boldsymbol{m})\intercal} \boldsymbol{D}_{\boldsymbol{r}}(\tau_{l_r})\right\}\right].$$

Likewise, the second derivative is:

$$\ell_C^{(2)}(\beta_{ps}) = \left.\frac{\partial^2 \ell_C(\vec{\boldsymbol{\beta}}|\vec{\boldsymbol{Z}}^{(\boldsymbol{m+1})})}{\partial \beta_{ps}^2}\right|_{\vec{\boldsymbol{\beta}}=\vec{\boldsymbol{\beta}}^{(\boldsymbol{m})}}$$
$$= -\sum_{r,k,l_r} \left[\tau_{l_r} Z_{rk}^{(m+1)} D_{rp}^2(\tau_{l_r}) \exp\left\{\boldsymbol{\beta}_{\boldsymbol{ks}}^{(\boldsymbol{m})\intercal} \boldsymbol{D}_{\boldsymbol{r}}(\tau_{l_r})\right\}\right].$$

Using Taylor's expansion, we have that the one-step update for β_{ps} is

$$\beta_{ps}^{(m+1)} = \beta_{ps}^{(m)} + \triangle'_{ps} = \beta_{ps}^{(m)} - \frac{\ell_C^{(1)}(\beta_{ps})}{\ell_C^{(2)}(\beta_{ps})}.$$

Following similar arguments for β_{0ks} we have that

$$\ell_C^{(1)}(\beta_{0ks}) = \left.\frac{\partial \ell_C(\vec{\boldsymbol{\beta}}|\vec{\boldsymbol{Z}}^{(\boldsymbol{m+1})})}{\partial \beta_{0ks}}\right|_{\vec{\boldsymbol{\beta}}=\vec{\boldsymbol{\beta}}^{(\boldsymbol{m})}}$$
$$= \sum_{r,l_r} \left[\delta_s(\tau_{l_r}) Z_{rk}^{(m+1)} - Z_{rk}^{(m+1)} \tau_{l_r} \exp\left\{\boldsymbol{\beta}_{\boldsymbol{ks}}^{(\boldsymbol{m})\intercal} \boldsymbol{D}_{\boldsymbol{r}}(\tau_{l_r})\right\}\right],$$

$$\ell_C^{(2)}(\beta_{0ks}) = \left.\frac{\partial^2 \ell_C(\vec{\boldsymbol{\beta}}|\vec{\boldsymbol{Z}}^{(\boldsymbol{m+1})})}{\partial \beta_{0ks}^2}\right|_{\vec{\boldsymbol{\beta}}=\vec{\boldsymbol{\beta}}^{(\boldsymbol{m})}} = -\sum_{r,l_r} \left[Z_{rk}^{(m+1)} \tau_{l_r} \exp\left\{\boldsymbol{\beta}_{\boldsymbol{ks}}^{(\boldsymbol{m})\intercal} \boldsymbol{D}_{\boldsymbol{r}}(\tau_{l_r})\right\}\right],$$

and

$$\beta_{0ks}^{(m+1)} = \beta_{0ks}^{(m)} + \triangle'_{0ks} = \beta_{0ks}^{(m)} - \frac{\ell_C^{(1)}(\beta_{0ks})}{\ell_C^{(2)}(\beta_{0ks})}.$$

As in (36; 73; 121), we perform a complete sweep over all parameters in $\vec{\boldsymbol{\beta}}$ multiple times instead of performing multiple iterations of a single parameter and moving onto the next.

Following the arguments of (23) one can show that the first and second derivatives of Equation 6 with respect to a single b_{jk} is

$$\ell_C^{(1)}(b_{jk}) = \left.\frac{\partial \ell_C(\vec{\boldsymbol{b}}|\vec{\boldsymbol{Z}}^{(\boldsymbol{m+1})})}{\partial b_{jk}}\right|_{\vec{\boldsymbol{b}}=\vec{\boldsymbol{b}}^{(\boldsymbol{m})}} = \sum_{r=1}^{R}(Z_{rk}^{(m+1)} - \pi_{rk}^{(m)})E_{rj},$$

and

$$\ell_C^{(2)}(b_{jk}) = \left.\frac{\partial^2 \ell_C(\vec{\boldsymbol{b}}|\vec{\boldsymbol{Z}}^{(\boldsymbol{m+1})})}{\partial b_{jk}^2}\right|_{\vec{\boldsymbol{b}}=\vec{\boldsymbol{b}}^{(\boldsymbol{m})}} = -\sum_{r=1}^{R}\pi_{rk}^{(m)}(1-\pi_{rk}^{(m)})E_{rj}^2.$$

The one-step update for b_{jk} is

$$b_{jk}^{(m+1)} = b_{jk}^{(m)} + \triangle'_{jk} = b_{jk}^{(m)} - \frac{\ell_C^{(1)}(b_{jk})}{\ell_C^{(2)}(b_{jk})}.$$

As with the proportional hazards coefficients we perform multiple sweeps over all model parameters instead of multiple iterations for a single parameter.

When performing these one-step estimation algorithms it is important that a single step does not go too far. This can occur when the log-likelihood function is not locally quadratic and can lead to ill-fitting results. Therefore, we employ the trust region algorithm of (36; 121). Furthermore, we only perform a maximum of five sweeps for the proportional hazards model coefficients in the M-Step, as the likelihood function will still sufficiently increase. The pseudocode is provided in Algorithm 1.

4.4 Application

In this section we present the results of our study on uncovering utilization patterns among the asthma diagnosed patients in the Medicaid system. We provide the estimated model along with a practical interpretation and various visualizations of the results in Appendix **??**. The model selected using the approach in this appendix presents five clusters of patients according to their utilization behavior. Statistical significance of the covariate effects and multinomial logistic parameters are investigated in Appendix C.3.

Algorithm 1 M Step for PH and MN Coefficients

function M STEP($\vec{\boldsymbol{H}}, \vec{\boldsymbol{D}}, \vec{\boldsymbol{E}}, \vec{\boldsymbol{Z}}^{(m+1)}, \vec{\boldsymbol{\beta}}^{(m)}, \vec{\boldsymbol{b}}^{(m)}$)

$\triangle_{0ks} = \triangle_{ps} = \triangle_{jk} = 1, \forall\ k \in \{1, \ldots, K\}, s \in \{1, \ldots, |S|\}, \forall\ p \in \{1, \ldots, P\}, \forall\ j \in \{1, \ldots, J\}$

for $n = 1, 2, \ldots\ 5$ or until convergence **do**

for $\forall\ k, \forall\ p$ **do**

compute $\triangle'_{0ks}$, $\triangle'_{ps}$

$\triangle''_{0ks} = \text{sign}(\triangle'_{0ks}) \times \min(\triangle_{0ks}, \triangle'_{0ks})$, $\triangle''_{ps} = \text{sign}(\triangle'_{ps}) \times \min(\triangle_{ps}, \triangle'_{ps})$

$\beta^{(m+1)}_{0ks} = \beta^{(m)}_{0ks} + \triangle''_{0ks}$, $\beta^{(m+1)}_{ps} = \beta^{(m)}_{ps} + \triangle''_{ps}$

$\triangle_{0ks} = \max(2\triangle''_{0ks}, \triangle_{0ks}/2)$, $\triangle_{ps} = \max(2\triangle''_{ps}, \triangle_{ps}/2)$

end for

end for

for $n = 1, 2, \ldots$ until convergence **do**

for $j = 1, \ldots, J$ **do**

for $k = 1, \ldots, K$ **do**

compute $\triangle'_{jk}$

$\triangle''_{jk} = \text{sign}(\triangle'_{jk}) \times \min(\triangle_{jk}, \triangle'_{jk})$

$b^{(m+1)}_{jk} = b^{(m)}_{jk} + \triangle''_{jk}$

$\triangle_{jk} = \max(2\triangle''_{jk}, \triangle_{jk}/2)$

end for

end for

end for

end function

4.4.1 Proportional Hazards Model

4.4.1.1 Baseline Rates

We begin by presenting the baseline rate of events per year for each provider type. The baseline group of patients represents the population of children who are *white, chronically ill, aged 4-5, have not visited a healthcare provider yet in our study and are not eligible for Medicaid for reasons including blindness, disability or foster care*. The baseline rates are in Figure 17. The proportion of patients belonging to each cluster are 55.74% (Cluster 1), 16.10% (Cluster 2), 15.09% (Cluster 3), 10.32% (Cluster 4), and 2.75% (Cluster 5).

The baseline rate changes for each subpopulation within a cluster, and thus, should not be interpreted solely on their absolute value but on their relative values across clusters also. For instance, patients in Cluster 4 are more than twice as likely to fill a prescription than patients in any other cluster. Likewise, patients in Cluster 5 are more than six times as likely to visit a healthcare clinic than other patients. Because the effects of the control variables are the same regardless of cluster membership, these statements will hold regardless of age, demographics, or health status.

Cluster 1, with the greatest proportion of the population, has the least number of RX visits per year, less than one third of the cluster with the next lowest RX rate. Patients in Cluster 2 rely almost solely on RX visits, with low rates of visits to all other provider types. Cluster 3 patients have the highest rate of PO visits but the second lowest number of RX visits. Patients belonging to Cluster 4 have the greatest number of RX visits per year, but also have the second highest rate of HO visits. Finally, Cluster 5, with the fewest patients, has the greatest number of CL, ER and HO visits, with the third highest rate of PO visits and second highest rate of RX visits.

4.4.1.2 Covariate Effects

Now we describe the effects of the control covariates on the baseline visitation rates. In Figure 18 we provide the rate multipliers for the different covariate values. Thus, the rates

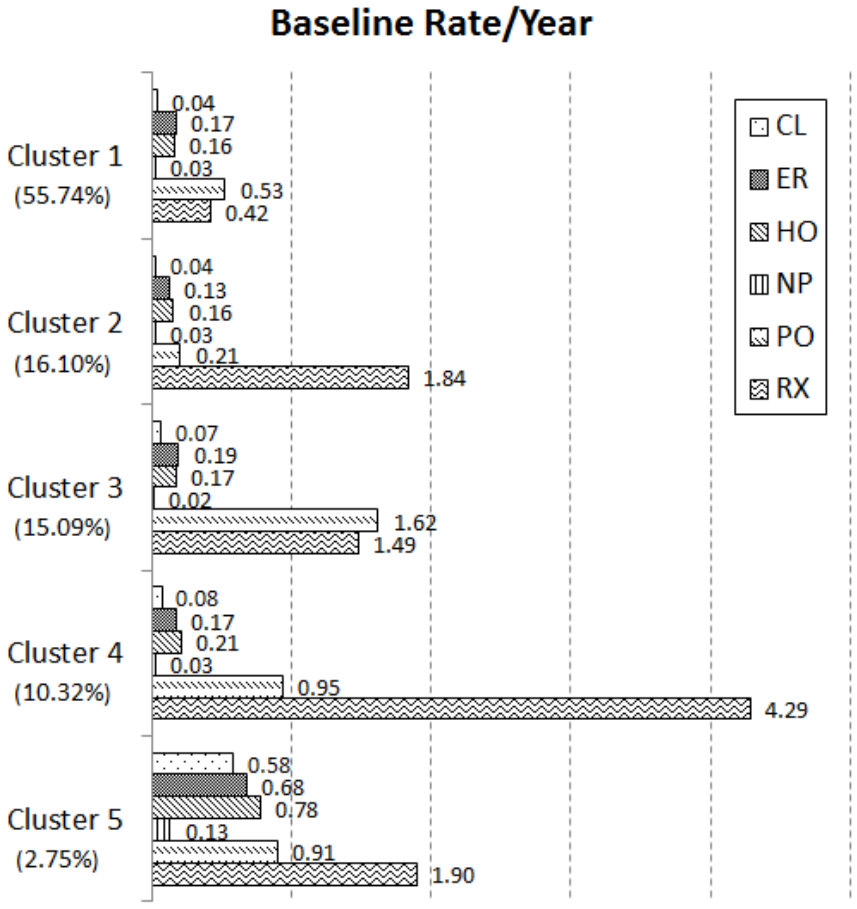

Figure 17: Baseline rate of events per year for white, chronically ill patients, aged 4-5, who are not eligible for Medicaid for blindness/disability or foster care, and without a prior observed event.

of visits for different subpopulations can be found by multiplying the baseline rate by the rate multipliers from this chart. For instance, to find the rates for a black patient, age 16, one would multiply the baseline rate from Figure 17 by the rate multipliers from the *black* and *age 15-17* covariates in Figure 18. It is important to remember that the effects of these covariates are the same across all clusters. That is, a severely ill patient will have 6.79 times more hospitalizations regardless of whether they belong to Cluster 1 or Cluster 5.

We find that the effects of health status or clinical risk group (healthy, minor or severely ill) have the greatest practically significant effect on the baseline rate. While the clinical risk group is an overall evaluation of the health condition, it can also reflect the severity of asthma. For example, a patient categorized as healthy will have mild asthma. A severely ill patient has a higher rate for all provider types but the rate of hospitalizations is 6.79 times higher than a chronically ill patient. Patients with a minor chronic illness have little relative change, while healthy patients have drastically less events of all types. Other findings include higher utilization of the CL, ER, and HO and lower utilization of RX for patients that are non-white, while patients who are eligible for Medicaid due to being blind or disabled or in foster care have overall lower rates of visits. Finally, the effects of age seem to have little practical difference for patients in age group 6-14, while patients age 15-17 have higher rates of visits to all provider types except CL and RX.

4.4.1.3 Provider Networks

Now we demonstrate how our model outputs can be used to visualize the provider transition networks for patients in different subpopulations and/or clusters. In this example, we compare the effects of the patient's clinical risk group on healthcare utilization for the baseline group of patients. We chose this example for illustration purposes because of the drastic multiplicative effects of health status on the baseline visit rates as shown in Figure 18. In Figure 19, we compare the network plots of healthy, chronically ill, and severely ill patients, leaving out patients with a minor illness due to the small change from those that

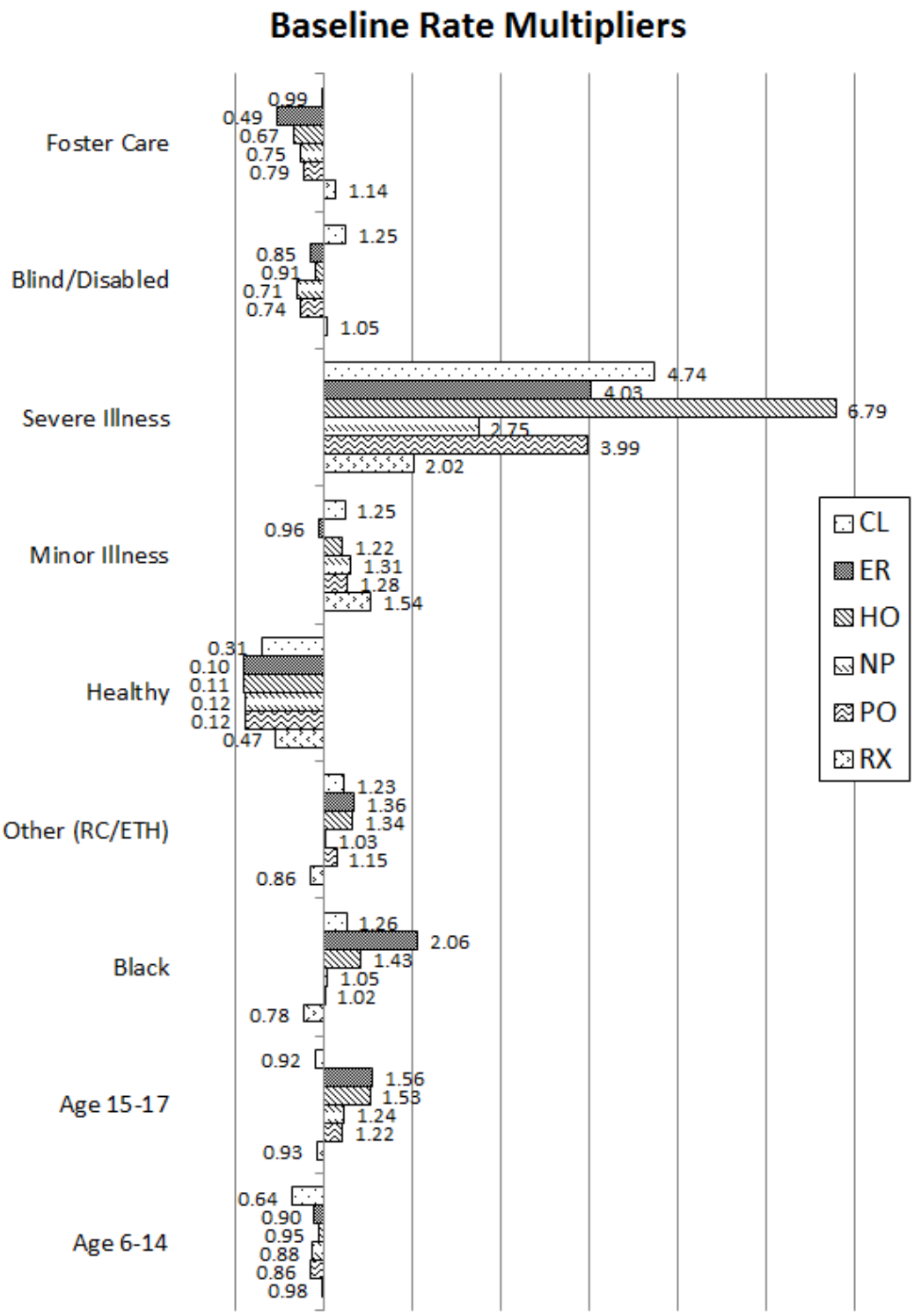

Figure 18: Baseline rate multipliers for each subpopulation

are chronically ill. The derivation of the transition probabilities and the average interarrival times for the baseline group across clusters can be found in Appendix C.3.

Clusters 2, 3, and 4 networks have strong connections from all nodes leading to RX visits. However, as a patient's health condition becomes more severe, utilization becomes more variational, with a greater number of connections between different provider types for the chronic and severe illness columns. Patients in Cluster 1 have high probability transitions into PO and RX provider types, with a higher probability of readmission into HO for chronically and severely ill patients. Clusters 2 and 4 are similar for healthy and chronically ill patients, except more transitions into HO in Cluster 2 and PO in Cluster 4. Cluster 3 healthy patients have similar networks as Cluster 2 and 4 healthy patients but with much greater variation for patients with a chronic or severe illness. Patients in cluster 2 route into RX regardless of overall health condition. Chronic and Severe patients in Clusters 1 and 3 have high probability transitions into PO from all nodes, while severe patients in Cluster 4 have some transitions from CL to PO. Cluster 5, with the smallest percentage of patients, consists of those who more frequently utilize ER and HO with significant transitions into HO for both chronically ill and severely ill patients, while severely ill patients having more than 50% chance of readmission into HO. Across all clusters, NP is insignificant and primarily routes patients back to NP or into PO or RX visits.

4.4.2 Latent Variable Model

Now we provide visualizations for the effects of the explanatory variables on cluster membership. We provide the parameter outputs from the model chosen in Appendix C.4.

In Figure 20 we plot the proportion of patients from each state by urbanicity category and by cluster. That is, for a given state and urbanicity level, the sum of the values in the chart across clusters will be one. The black dashed lines indicate the overall proportion of patients belonging to a given cluster regardless of state and urbanicity.

We can see that while the urbanicity level of the patient's residence does affect cluster

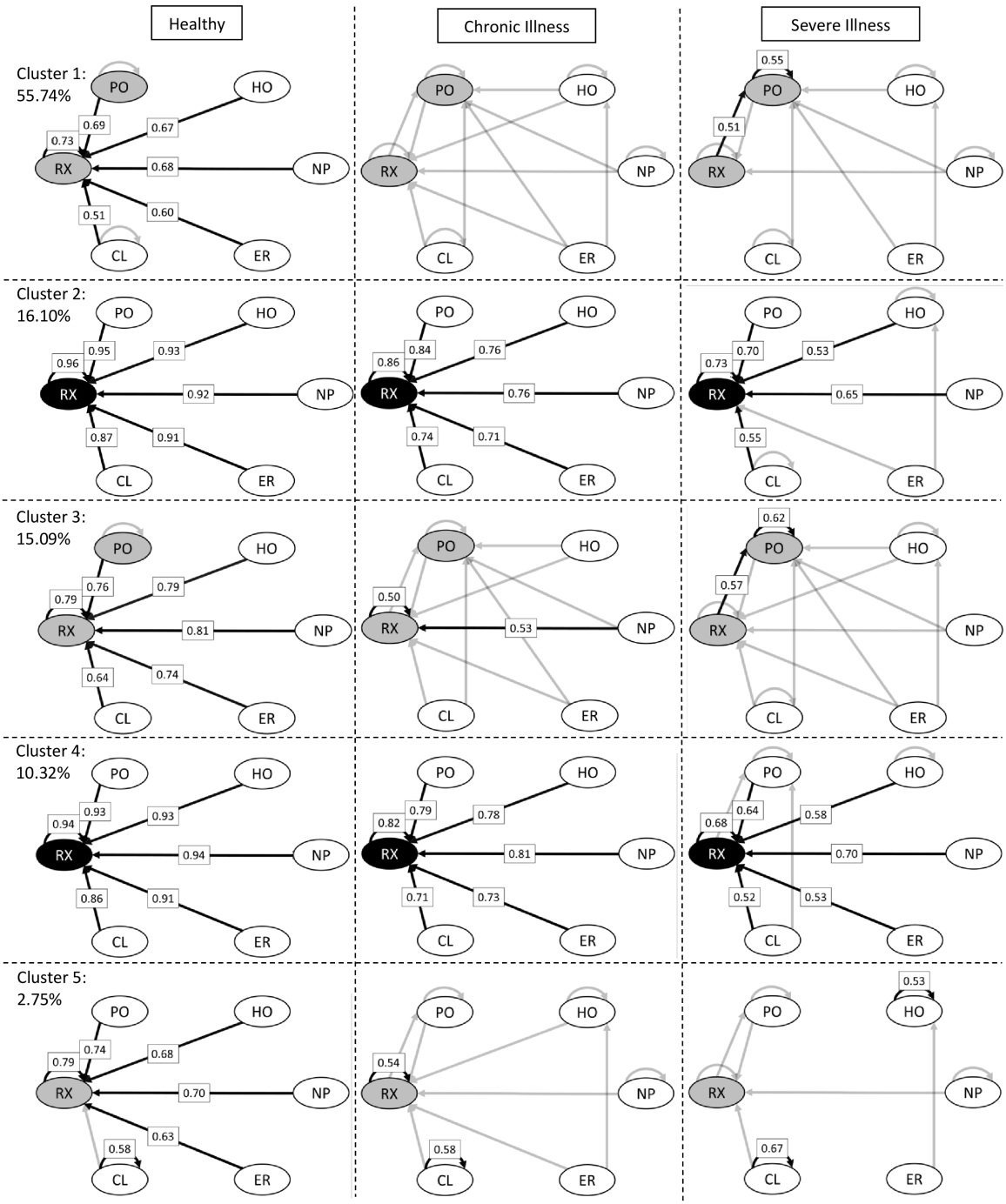

Figure 19: Provider networks inferred from the proportional hazards coefficients. The following rules were used in setting the grayscale of the coefficients and nodes: $< 0.2 \rightarrow$ not shown/white, $[0.2, 0.5) \rightarrow$ gray, and $\geq 0.5 \rightarrow$ black.

membership, it is the patient's residence state that is the main driver of variation in utilization behaviors. Furthermore, it appears that within each state, urban and suburban patients act similarly while rural patients behave differently. Clusters 1 and 3 have a higher proportion of urban and suburban patients relative to rural patients while Clusters 2 and 4 have the opposite. Cluster 5 appears to be evenly divided among the three urbanicity measures.

GA and MS seem to behave differently than the other states while LA and MN, and NC and TN behave similarly. Recall that Cluster 1 patients rely on PO and RX visits, Clusters 2 and 4 rely almost solely on RX for healthy and chronically ill patients, Cluster 3 has a high rate of PO visits and some RX visits, and Cluster 5 utilizes more ER and HO visits than the others. From Figure 20, it becomes clear that GA patients are overall more variational, relying less on RX visits than the overall average and more on other provider types, having the greatest proportion of patients in Cluster 5. MS has the highest proportion of patients belonging to clusters dominated by RX visits, namely Clusters 2 and 4, with MN, NC, and TN patients also having relatively high proportions in those clusters. LA has the highest proportion of patients belonging to Cluster 1 and the lowest belonging to Cluster 5.

The third explanatory variable in our study is a measure of travel distance to primary care, which is the main source of care for asthma for the Medicaid-insured children. Interpreting the effects of travel distance on cluster membership is more difficult because the variable is numerical instead of categorical. However, we provide an example of the effects of increased travel time on cluster membership, assuming that the baseline probability of belonging to Clusters 1-5 *are equal* (this is not always the case as state and urbanicity also factor in greatly). In Figure 21, we demonstrate the change in probability for this hypothetical example for patients in Clusters 1-5 with travel distances ranging from 0-25 miles.

This graph should be interpreted by the relative change in probability across clusters. We find that higher travel distances increase the probability of membership in Clusters 1, 3, and 5, with 5 being the greatest, while probabilities decrease for Clusters 2 and 4 as travel distance increases. Incidentally, Clusters 1, 3, and 5 tend to be more variational when

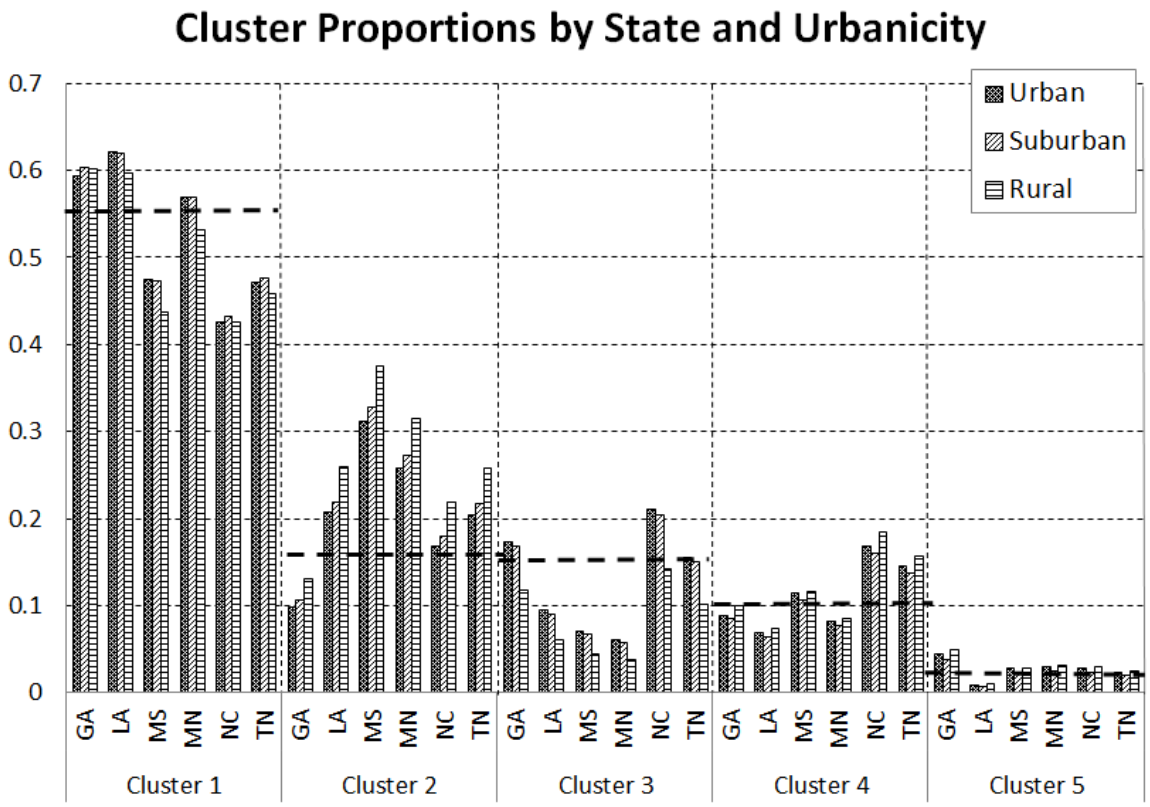

Figure 20: Proportions of patients belonging to each cluster stratified by state and urbanicity

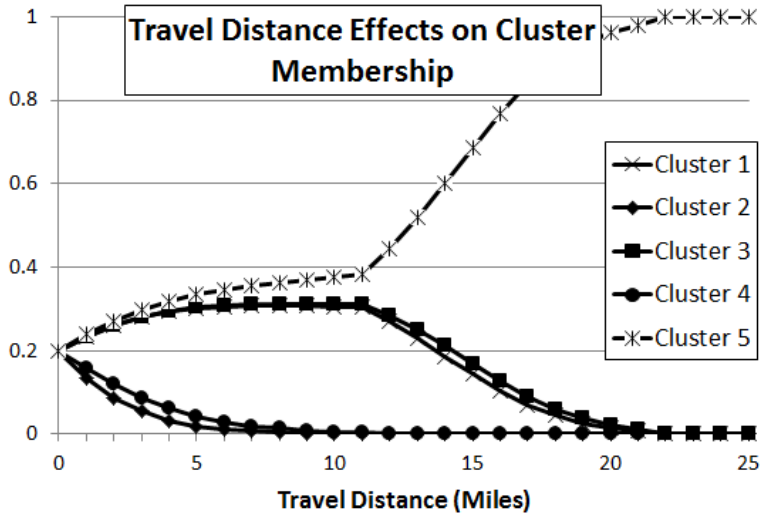

Figure 21: Plot of the change in probability for Clusters 1-5 with travel time ranging from 0-10.

compared to Clusters 2 and 4, which primarily rely on RX utilization.

4.5 Discussion

In this chapter we introduce a model-based clustering analysis via a parametric proportional hazards model that allows for derivation of model parameters, cluster membership probabilities and visualizations. We also demonstrate the applicability of the methodology to policy-making on healthcare utilization. Our algorithm for the model estimation is computationally attractive, allowing for complete model estimation in 2-3 hours for a set of more than 400,000 patients and 6 million interarrival times. By studying pediatric asthma patients from six states, we are able to determine the drivers of inter-cluster variation while controlling for the effects of controlling covariates such as age, race and ethnicity, and overall health status.

The primary outputs from our model consists of the rate of visits by event type for patients belonging to the baseline group; the effects of the control covariates on the baseline rates in the form of rate multipliers indicating the variations of utilization that cannot be impacted by interventions; and the effects of the last visit type on future utilization choice. We show how these effects can be used to determine a one-step provider network. We finish with visualizations of the effects of the explanatory variables on cluster membership.

The baseline rate per year shows that the majority of patients, those belonging to Cluster 1 (55%), utilize asthma controller medications the least but also have few emergency room visits or hospitalizations. The provider networks across health conditions show that as the patient's condition worsens, patients tend to utilize the physician's office more, indicating that the majority of asthma patients are well-managed and require minimal, routine care to control asthmatic conditions. Cluster 3 (15%) is similar to Cluster 1 just with more visits to the physician's office and prescription fills for asthma controller medication and relatively few emergency room visits or hospitalizations, also indicating patients who require minimal care. Higher travel distances increase the probability of membership within these

two clusters. From Figure 20 we see that GA and LA have above average representation in Cluster 1, with MS, NC, and TN having well below average. Cluster 3 has above average representation of NC patients while LA, MS, and MN are well below average.

Cluster 2, with 16% of the population consists of those patients who rely heavily on medication and little else, thus representing those patients with the least utilization of the system of care, hence with a well controlled asthma. The effects of health status on the provider networks are minimal with slightly more admission into hospitalizations for patients with a severe health condition. Despite the fact that lower travel distances increase the probability of membership in this cluster, these patients rarely utilize the physician's office. GA has well below average representation in this cluster while MS has the greatest.

Cluster 4 (10%) patients are the highest utilizers of medication with relatively high rates of visits physician office visits but also relatively low baseline rates of ER and HO visits, hence another cluster of patients with well controlled asthma. This cluster has the least variation when comparing across health status with the severely ill patients having high transition rates to the physician's office. NC and TN have above average representation in this cluster. Lower travel distances increase membership probability in this cluster which could explain the high baseline rate of visits to physician office. While NC has average representation in Cluster 2, it has the highest proportion of patients in Cluster 4.

Finally, Cluster 5 (3%) consists of those patients who have the highest utilization of the emergency department and hospitalizations, and are likely to be those patients with the most severe asthmatic conditions requiring high-end care. Both chronically and severely ill patients have higher rates of emergency department visits and hospitalizations as indicated by the provider networks. GA has the most patients in this cluster and LA the least. These patients also tend to have the highest travel times to a physician's office.

Some important findings drawn from this study are:

• The most influential factor on the differences between children at the baseline or entry point in the system is the overall health condition.

• Older children are higher utilizers of the system, particularly of both emergency departments and hospitalizations. One explanation is that asthma in older children can interfere with sleep, school, sports and social activities.

• Children in foster care are lower utilizers of the system with a lower rate of both emergency departments and hospitalizations. This is expected because such visits require the presence of a social worker and possibly a member of the foster care agency if one is involved. This additional requirements may discourage utilization of emergency services.

• The black population has twice the rate of emergency department visits. Prior research has not found a statistically significant association of the percentage of non-white population to geographic access while controlling for income in Georgia (83).

• Patients who are categorized as severely ill using the clinical risk group classification have the highest utilization across all provider types and of being prescribed medication. This is not unexpected because other comorbidities could lead to more severe outcomes for asthma. Moreover, these patients are most challenging to control because of the preexistence of other conditions that could more severely affect a patient than asthma.

• The clustering of the patients reflects different utilization behaviors. While the majority of the patients utilize the system disparately (Cluster 1), others have a high rate of medication uptake with little interaction with the system (Cluster 2), with some utilization of the physician office (Cluster 3) or with high utilization of the physician's office and high rate of medication uptake (Cluster 4). There is also a small percentage of patients (3%) who are higher utilizers of the system, not necessarily with a high medication uptake, that visit the emergency department or hospital at a higher rate with a 0.2-0.5 probability of being followed by a hospitalization for most subpopulations.

• The probability of follow-up visits once a patient visits the emergency department or has a hospitalization is lower than 0.2 for most subpopulations that are not severely ill across all clusters except for some subpopulations in Clusters 1 and 5. Additionally, except for healthy patients, the probability of filling a prescription for an asthma controller medication

after an emergency department visit or hospitalization is lower than 0.5 except for Clusters 2 and 4.

• Most of all visits to a healthcare provider, including a hospital, a clinic or physician, result in a medication prescription being filled, with a high probability of a refill.

• There are some variations across different urbanicity levels although the variations are higher between states. GA, LA and MN have a larger percentage of patients who utilize the system disparately (Cluster 1) while NC and TN have a higher percentage of patients who are high utilizers of medication (Cluster 4).

CHAPTER V

CONCLUSION

In this thesis we have provided adaptive and scalable model-based data mining and statistics methods allowing for the extraction of useful knowledge from large, complex datasets. We have included methods from nonparametric statistics, statistical clustering analysis and stochastic process modelling, adapting the methods to the data set of interest. With an exponential increasing rate of healthcare data acquisition, data scientists must continue to adapt previous useful methodologies to make inference on data of increasing size and complexity. By studying noisy nuclear magnetic resonance and extremely large, complex Medicaid claims data in this thesis, we provide methods that could inspire advanced statistical and model-based data mining methods across many types of applications, including business and finance sectors, for instance. Additionally, our methods make use of model assumptions that provide for computational ease and elegant simplicity in order that our inferences are useful and timely for decision makers and practicioners.

In Chapter II our component identification methods provide researchers with potentially more accurate methods for protein structure identification. We have contributed methods that provide estimates for the location of components that are potentially obscured by noise and interference from other components.

In Chapters III and IV we pilot structure identification methods for large-scale claims data sets using model-based data mining. In particular, in Chapter III we studied the pediatric asthma population with Medicaid coverage in Georgia and North Carolina. By using a Markov renewal process model we are able to not only estimate underlying utilization profiles but also produce useful visualizations to inform potential cost-saving interventions that lead to increased adherence to recommended care guidelines. Additionally, we are able

to easily infer the potential cost-savings of such an intervention by manipulating model parameters. By studying neighboring states with different state-based Medicaid systems we show that despite the fact that demographic and geographic characteristics are similar between the two states, Georgia asthma patients tend to have more disparate utilization and higher admission rates into the emergency room or hospitalizations due to asthma with more potential for cost-saving interventions.

In Chapter IV we provide methods that overcome some of the shortcoming of the study in Chapter III, primarily, the inclusion of demographic, geographic, and health characteristics and a statistically sound handling of missing data. We provide visual summaries of the effects of controlling and explanatory covariates on healthcare utilization. In particular, we demonstrate the effects of age, race, health status, Medicaid eligibility reason and patient history on future healthcare utilization, and provide explanation on future utilization choices from state, urbanicity, and access measures. We conclude with policy implications for targeted interventions to improve adherence to recommended care for pediatric asthma.

APPENDIX A

CHAPTER II: SUPPLEMENTARY MATERIALS

A.1 Proofs

A.1.1 Derivation of the Wavelet Coefficients

One-Dimensional Signal Assume the simplest setting, that of a continuous one-dimensional, noise free model,

$$f(x) = \sum_{l=1}^{L} A_l s\left(x; \omega_l, \tau_l\right) = \sum_{l=1}^{L} A_l \exp\left\{-\frac{1}{2}\left(\frac{x-\omega_l}{\tau_l}\right)^2\right\}.$$

By the fact that the wavelet function is defined to be $\psi_s(x) = s^2 \frac{d^2}{dx^2}\phi_s(x)$ and from the results of Mallat and Hwang (66), we have for a single Gaussian component,

$$Wf(s,x) \propto A_l s_{\omega_l,\tau_l} \star \frac{d^2}{dx^2}\phi_s(x) = A_l \frac{d^2}{dx^2}(s_{\omega_l,\tau_l} \star \phi_s)(x).$$

We are able to exchange the order of the derivative and the convolution in the previous equality because the convolution is simply an integral. Bromiley shows that the convolution of two Gaussian functions is itself a Gaussian function in (13). Therefore we have

$$Wf(s,x) \propto -A_l \tau_l \frac{d^2}{dx^2} \exp\left\{-\frac{1}{2}\left(\frac{x-\omega_l}{\sqrt{\tau_l^2 + s^{-2}}}\right)^2\right\}.$$

The derivative property of Gaussian functions, $\frac{d^n}{dx^n}\frac{1}{\tau}\phi\left(\frac{x}{\tau}\right) = (-1)^n H_n\left(\frac{x}{\tau}\right)\frac{1}{\tau}\phi\left(\frac{x}{\tau}\right)$, gives us the result:

$$Wf(s,x) \propto -\frac{A_l \tau_l}{\sqrt{\tau_l^2 + s^{-2}}} H_2\left(\frac{x-\omega_l}{\sqrt{\tau_l^2 + s^{-2}}}\right) \exp\left\{-\frac{1}{2}\left(\frac{x-\omega_l}{\sqrt{\tau_l^2 + s^{-2}}}\right)^2\right\} \qquad (6)$$

where $H_2(x)$ is the 2nd Hermite polynomial (the negative in front is the traditional form of the Mexican hat function).

By the additivity of the wavelet transform we then have that the wavelet coefficients of a sum of Gaussian components are

$$Wf(s,x) = \sum_{l=1}^{L} Wf_l(s,x) \propto -\sum_{l=1}^{L} \frac{A_l \tau_l}{\sqrt{\tau_l^2 + s^2}} H_2\left(\frac{x-\omega_l}{\sqrt{\tau_l^2+s^2}}\right) \exp\left\{-\frac{1}{2}\left(\frac{x-\omega_l}{\tau_l}\right)^2\right\}. \tag{7}$$

Extension to d-Dimensions Now we extend the derivation results to d-dimensional Gaussian components. We assume that the noise-free signal takes the form,

$$f(x) = \sum_{l=1}^{L} A_l s\left(x; \omega_l, \tau_l\right) = \sum_{l=1}^{L} A_l \exp\left\{-\frac{1}{2}\left(\sum_{d'=1}^{d} \frac{x_{d'} - \omega_{l,d'}}{\tau_{l,d'}}\right)^2\right\}.$$

We assume that all off-diagonal values for the width matrix τ are 0. With this assumption, we can extend the results of Bromiley (13) to d-dimensions, so that the width component for dimension d' is $\sqrt{\tau_{d',l}^2 + s^{-2}}$. Using this result and the interchangeability of the convolution and derivative we can determine that the wavelet coefficients of the sum of d-dimensional components in the horizontal orientation are, up to scale,

$$Wf(s,x) \propto -\sum_{l=1}^{L} A_l \frac{d^2}{dx_1^2} \exp\left\{-\frac{1}{2}\left(\sum_{d'=1}^{d} \frac{x_{d'} - \omega_{l,d'}}{\sqrt{\tau_{d',l}^2 + s^{-2}}}\right)^2\right\}.$$

Again, using the Gaussian derivative result we have,

$$Wf_l(s,x) \propto -\sum_{l=1}^{L} A_l H_2\left(\frac{x_1 - \omega_{l,1}}{\sqrt{\tau_{1,l}^2 + s^{-2}}}\right) \exp\left\{-\frac{1}{2}\left(\sum_{d'=1}^{d} \frac{x_{d'} - \omega_{l,d'}}{\sqrt{\tau_{d',l}^2 + s^{-2}}}\right)^2\right\}. \tag{8}$$

Derivations for dimensions $d' \in \{2, \ldots, d\}$ can be easily found.

A.1.2 Resolution Result

One-Dimensional Signal This proof relies heavily on the roots of the Hermite polynomials and the Gaussian derivative result. Recall that $\frac{d^n}{dx^n}\frac{1}{\tau}\phi\left(\frac{x}{\tau}\right) = (-1)^n H_n\left(\frac{x}{\tau}\right)\frac{1}{\tau}\phi\left(\frac{x}{\tau}\right)$, and from Equation (6) that the wavelet coefficients of a single Gaussian component are, up to a shift in location and change in scale,

$$-CH_2\left(\frac{x}{\tau}\right)\exp\left\{-\frac{1}{2}\left(\frac{x}{\tau}\right)^2\right\},$$

where C is a positive constant. Then for a pair of of components we have,

$$-C_1 H_2\left(\frac{x}{\tau_1}\right)\exp\left\{-\frac{1}{2}\left(\frac{x}{\tau_1}\right)^2\right\} - C_2 H_2\left(\frac{x-\theta}{\tau_2}\right)\exp\left\{-\frac{1}{2}\left(\frac{x-\theta}{\tau_2}\right)^2\right\}. \quad (9)$$

Proofs of Theorem 1 and Corollary 1:

Suppose we have two Gaussian peaks, with location parameters ω_l such that $\omega_1 < \omega_2$, width parameters τ_l such that $\tau_1 \leq \tau_2$ and amplitude $A_1 = A_2$. Let $\theta = |\omega_1 - \omega_2|$. We have already derived a closed-form expression of the wavelet coefficients. Let's begin by finding the local maxima and minima of the wavelet transform of a single Gaussian component. Taking the derivative of the wavelet coefficients, by the Gaussian derivative property, we have

$$\frac{d}{dx}Wf(s,x) = CH_3\left(\frac{x}{\tau}\right)\exp\left\{-\frac{1}{2}\left(\frac{x}{\tau}\right)^2\right\}. \quad (10)$$

Since C and $\exp(x)$ are both positive, we know that the roots of Equation (10) must be the roots of the third Hermite polynomial, $x^3 - 3x$, $0, \pm\sqrt{3}$. Call this set of points ξ_0.

Consider Figure 22. The minimums and maximums of the 3rd derivative can be found at the roots of the 4th Hermite polynomial, $x^4 - 6x^2 + 3$, $\pm\sqrt{3 \pm 6} \approx \pm.742, \pm 2.334$. Call this set of points ξ_1. Furthermore, we have the following property:

$$\forall\, x \in R,\ \left|\frac{d}{dx}Wf(s,\xi_1)\right| \geq \left|\frac{d}{dx}Wf(s,x)\right|, \quad (11)$$

and we can determine the sign of the derivative function.

The derivative of the wavelet coefficients of a pair of Gaussian components is then, up to a shift in location,

$$\frac{d}{dx}Wf(s,x) \propto H_3\left(\frac{x}{\tau_1}\right)\exp\left\{-\frac{1}{2}\left(\frac{x}{\tau_1}\right)^2\right\} + H_3\left(\frac{x-\theta}{\tau_2}\right)\exp\left\{-\frac{1}{2}\left(\frac{x-\theta}{\tau_2}\right)^2\right\}. \quad (12)$$

Since $A_1 = A_2$ and $\theta > 0$, we are then guaranteed that Equation (12) will be non-negative at $\omega_l - .742\tau_l$, and non-positive at $\omega_l + .742\tau_l$ because of the property from Equation (11). This is because $\left|\frac{d}{dx}Wf_1(s, \omega_1 \pm .742\tau_1)\right| \geq \left|\frac{d}{dx}Wf_2(s, \omega_1 \pm .742\tau_1)\right|$, and vice

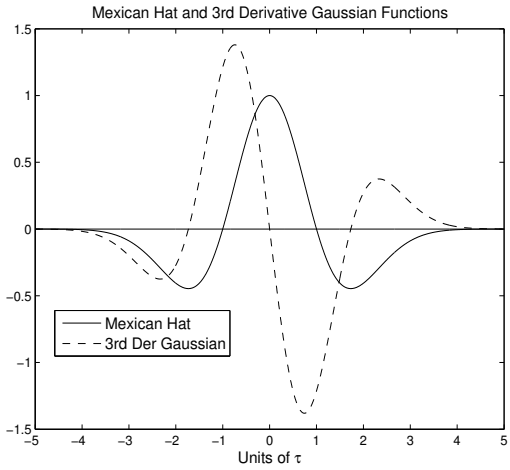

Figure 22: Comparison of the Mexican hat function and the 3rd derivative of the Gaussian function.

versa, and therefore the sign at $\omega_l \pm .742\tau_l$ will determine the sign of the sum of the derivatives. Furthermore, the limit towards $-\infty$ is negative while the limit towards ∞ is positive. Therefore, we should be able to find limits on θ such that the derivative goes from negative, to positive, to negative, back to positive, indicating that two local maxima exist in the wavelet transform. Both functions in Equation (12) are continuous, ensuring that the sum is continuous so no discontinuities in the wavelet transform or derivative exist.

Recalling that $\omega_1 < \omega_2$, $\tau_1 \leq \tau_2$ and $\theta = |\omega_1 - \omega_2|$, if $\theta > .742\tau_1 + .742\tau_2$ then the 3rd derivative will be negative towards $-\infty$, to positive at $-.742\tau_1$, to negative at $.742\tau_1$ to positive at $\theta - .742\tau_2$, back to negative at $\theta + .742\tau_2$, and finally to positive towards ∞. Therefore we will have two maxima around component locations and a minima in between, thus de-mixing the two components.

Now suppose that $\tau_2 > \frac{\sqrt{3}-.742}{.742}\tau_1 \approx 1.75\tau_1$. (This limit will make sense shortly). Recalling the roots, ξ_0, we know that the values of Equation (9) will be negative from $-\infty$, positive at $\omega_l - .742\tau_l$, negative at $\omega_l + .742\tau_l$ and back to positive towards ∞. Then if $\theta > \sqrt{3}\tau_1$, we are assured that between that the Equation (9) will go from positive for $x < \omega_1$, to negative at $x = \omega + .742\tau_1$, to positive for $x \in (\omega_1 + \sqrt{3}\tau_1, \omega_2)$, and back

to negative at $\omega_2 + .742\tau_2$. Therefore, we are insured to have two maxima in the wavelet transform with a minima in between, thus de-mixing the pair of components.

We now have two conditions for de-mixing a pair of components, and we take the minimum of each to derive a necessary and sufficient condition for component de-mixing. If $\tau_2 > \frac{\sqrt{3}-.742}{.742}\tau_1 \approx 1.75\tau_1$ then we use the latter condition and vice versa. This concludes the proof for Corollary 1. These results are not guaranteed to hold when the amplitudes, A_1 and A_2 differ because we can no longer assume that $|\frac{d}{dx}Wf_l(s,\xi_1)| \geq |\frac{d}{dx}Wf_l(s,x)|$ holds for $l = 1, 2$. Therefore, as stated in Theorem 1 we only have a necessary condition for de-mixing a pair of components.

Recalling that the wavelet coefficients will have width $\sqrt{\tau^2 + s^{-2}}$ instead of τ due to the convolution result we have now derived the limits of Theorem 1 and Corollary 1.

Multivariate Extension Extending the previous resolution results is straightforward when comparing Equations (7) and (8). Recall the following results:

Theorem A.1.1 *A necessary condition for de-mixing two d-dimensional Gaussian components at scale s, is that there must exist some $d' \in \{1, \ldots, d\}$ such that*

$$\theta_{d'} = |\omega_{1,d'} - \omega_{2,d'}| > min\left\{\sqrt{3(\tau_{d',1}^2 + s^{-2})},\ .742\sqrt{\tau_{d',1}^2 + s^{-2}} + .742\sqrt{\tau_{d',2}^2 + s^{-2}}\right\}.$$

Corollary A.1.2 *If $A_1 = A_2$, then a sufficient condition for de-mixing two d-dimensional Gaussian components is that there must exist some $d' \in \{1, \ldots, d\}$, such that*

$$\theta_{d'} > min\left\{\sqrt{3(\tau_{d',1}^2 + s^{-2})},\ .742\sqrt{\tau_{d',1}^2 + s^{-2}} + .742\sqrt{\tau_{d',2}^2 + s^{-2}}\right\}.$$

Because the structure of the wavelet coefficients are the same except for the $\exp(x)$ portion of the equation we can rely on all of the same arguments. The roots and derivatives of the Hermite polynomials will be the same as in the one-dimensional proof. Furthermore, we only need to meet the minimum distance requirements for θ in a single dimension $d' \in \{1, \ldots, d\}$ and to apply the wavelet transform in that orientation for the results to hold. Theorem 2 and Corollary 2 easily follow.

A.1.3 Interference Results

Now we move onto the following interference results,

Lemma A.1.3 *Suppose we have a set of 3 components* $s(x;\omega_1,\tau_1)$, $s(x;\omega_2,\tau_2)$, *and* $s(x;\omega_3,\tau_3)$, *with inter-component distances* $\theta_1 = |\omega_1 - \omega_2|$ *and* $\theta_2 = |\omega_2 - \omega_3|$, *where* $\theta_1 << \theta_2$ *without loss of generality. Then the resolution limits from Theorem 2.3.1 are assured to hold for the pair of components 1 and 2 if* θ_2 *satisfies* $\theta_2 > 2.334\sqrt{\tau_2^2 + s^{-2}} + 2.334\sqrt{\tau_3^2 + s^{-2}}$.

Lemma A.1.4 *Suppose we have a set of 3* d*-dimensional components* $s(x;\omega_1,\tau_1)$, $s(x;\omega_2,\tau_2)$, *and* $s(x;\omega_3,\tau_3)$, *with inter-component distances* $\theta_1 = |\omega_1 - \omega_2|$ *and* $\theta_2 = |\omega_2 - \omega_3|$, *where* $\theta_1 << \theta_2$ *without loss of generality. Then the resolution limits from Theorem 2.3.3 are assured to hold for components 1 and 2 if* $\forall\ d' \in \{1,\ldots,d\}$, $\theta_{2,d'} > 2.334\sqrt{\tau_{2,d'}^2 + s^{-2}} + 2.334\sqrt{\tau_{3,d'}^2 + s^{-2}}$.

The proof of these follow from the fact that the derivative of the 3rd component will take a negative value at $\omega_3 - 2.334\tau_3$. If $A_3 > A_2$, then the sum of the derivatives of components 2 and 3 are guaranteed to take a negative value at $x = \omega_3 - 2.334\tau_3$, by the property in Equation (15), and there will be a minimum between components 2 and 3. If $A_3 < A_2$, then the resolution results will not be affected because the amplitude of the interfering coefficient is too small to interfere with the wavelet coefficients of components 1 and 2.

A.1.4 Lipschitz Exponent of Gaussian

It is well known from Mallat and Hwang that a differentiable function, $f(x)$ is Lipschitz 1, and that its primitive $g(x)$ is Lipschitz 2 (66). We know that the Gaussian function is infinitely continuously differentiable, specifically, $\frac{d^n}{dx^n}\phi(x) = (-1)^n H_n(x)\phi(x)$ for all $n > 0$. Together with the result on the Lipschitz exponent of primitives it is easy to show by induction on n that the Lipschitz exponent of a Gaussian function is unbounded.

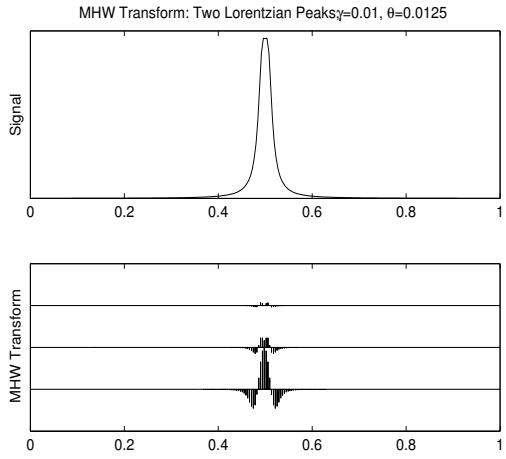

Figure 23: Two Lorentzian components with parameters $\tau = .01$ and $\theta = .0125$. The signal is plotted on top and the three finest scales of the CWT are on the bottom.

A.2 Robustness of the Algorithm: Lorentzian-Shaped Components

In this section we illustrate that the properties of the MHW transform not only apply to Gaussian-shaped components but to other smooth, symmetric components such as Lorentzian-shaped components

$$a(x; x_0, \gamma) = \frac{\gamma}{(x - x_0)^2 + \gamma^2},$$

where x_0 is a location parameter and γ is the scale parameter. Define θ as before to be the distance between two components. We show in Figure 23 the ability of the MHW to both locate and separate two Lorentzian components.

At the coarsest scale it is clear that a component exists in the center of the domain and in the finer scales, the MHW transform contains two maxima at the locations of the two Lorentzian peaks. Also, notably, the MHW transform can discern Lorentzian-shaped components at a much smaller θ than is possible with Gaussian-shaped components.

APPENDIX B

CHAPTER III: SUPPLEMENTARY MATERIALS

B.1 Derivation of the Utilization Sequences

A utilization sequence for an asthma-diagnosed child is derived based on the following set of information available in the Centers for Medicare and Medicaid (CMS) Medicaid Analytical Extract (MAX) medical claims data:

• *Primary and secondary asthma diagnosis ICD-9 codes*: We extract only those claims with the following asthma-related ICD-9 codes: 493.00, 493.01, 493.02, 493.10, 493.11, 493.12, 493.20, 493.21, 493.22, 493.81, 493.82, 493.90, 493.91, 493.92. These are the only diagnosis codes corresponding to an asthma diagnosis available in the MAX files.

• *Type of service and Place of Service codes*: Both codes are available for each claim from the IP and OT files of the MAX extract. They are used to derive a provider type for each medical visit as shown in Table B.1.

• *National Drug Code of Long-term asthma control medications*: These medications are taken regularly to control chronic symptoms and prevent asthma attacks and can be

Table 2: Provider Type Crosswalk

Type of Service Code	Logic	Place of Service Code	Provider Type
12: Clinic	OR	50: Federally qualified health center; 71: State or local public health clinic; 72: Rural health clinic	CL
11: Outpatient hospital	OR	23: Emergency room	ER
01: Inpatient hospital	AND	Any	HO
37: Nurse practicioner services	AND	11: Office	NP
08: Physicians	AND	11: Office	PO

found in the RX file of the MAX extract. The medication types include: Inhaled Corticosteroids, Long-Acting Beta-Agonists (LABAs), Cromolyn and Tehophylline, Leukotriene Modifiers and Immunomodulators. We consider the claim for asthma control medication as an event type (RX) in the stochastic network analysis.

• *Service begin date as the start time of the event*: Multiple claims in a single day are considered as one visit and the corresponding type of provider is defined by the care provider based on the first claim for each date.

B.2 Markov Renewal Process: Proofs and Derivations

B.2.1 Likelihood Function Derivations

In this first subsection we provide derivations for the likelihood functions in Equations (1) and (2) in Section 3.3.

Discrete Time Markov Chain Likelihood Derivation Consider a discrete time Markov chain (DTMC) with a sequence of events denoted by $\vec{X}_L = (X_1, \ldots, X_L)$. The derivation of the likelihood function in equation (2) from Section 3.3 is given below:

$$\begin{aligned} L(P|\vec{X}_L) = \Pr(\vec{X}_L = \vec{s}_L) &= \Pr(X_L = s_{i_L}|\vec{X}_{L-1} = \vec{s}_{L-1}) \\ \times \Pr(X_{L-1} = s_{i_{L-1}}|\vec{X}_{L-2} = \vec{s}_{L-2}) &\times \cdots \times P(X_2 = s_{i_2}|X_1 = s_{i_1}) \times P(X_1 = s_{i_1}) \\ &= \Pr(X_L = s_{i_L}|X_{L-1} = s_{i_{L-1}}) \times \Pr(X_{L-1} = s_{i_{L-1}}|X_{L-2} = s_{i_{L-2}}) \\ &\quad \times \cdots \times \Pr(X_2 = s_{i_2}|X_1 = s_{i_1}) \times \Pr(X_1 = s_{i_1}) \\ &= P_{s_{i_{L-1}},s_{i_L}} \times \cdots \times P_{s_{i_1},s_{i_2}} \times P_{LC,s_{i_1}} = \prod_{l=1}^{L} P_{s_{i_{l-1}},s_{i_l}} \times P_{LC,s_{i_1}} \end{aligned}$$

Markov Renewal Process Likelihood Derivation: Assuming that a patient sequence of events with timestamps follow a Markov renewal process, denoted by $(\vec{X}_L, \vec{T}_L)$, we provide

the derivation of the likelihood function from equation (2) in Section 3.3:

$$\begin{aligned} L(P,\Lambda|\vec{X}_L,\vec{T}_L) &= \Pr(\vec{X}_L = \vec{s}_L, \vec{T}_L = \vec{\tau}_L) \\ &= P(X_L = s_{i_L}, T_L = \tau_L | \vec{X}_{L-1} = \vec{s}_{L-1}, \vec{T}_{L-1} = \vec{\tau}_{L-1}) \\ &\times \cdots \times P(X_2 = s_{i_2}, T_2 = \tau_2 | X_1 = s_{i_1}, T_1 = \tau_1) \times P(X_1 = s_{i_1}, T_1 = \tau_1) \\ &= \Pr(X_L = s_{i_L}, T_L = \tau_L | X_{L-1} = s_{i_{L-1}}) \\ &\times \Pr(X_{L-1} = s_{i_{L-1}}, T_{L-1} = \tau_{L-1} | X_{L-2} = s_{i_{L-2}}) \\ &\times \cdots \times \Pr(X_2 = s_{i_2}, T_2 = \tau_2 | X_1 = s_{i_1}) \times \Pr(X_1 = s_{i_1}) \end{aligned}$$

Next we make use of the following conditional probability rule:

$$\begin{aligned} &\Pr(X_l = s_{i_l}, T_l = \tau_l | X_{l-1} = s_{i_{l-1}}) \\ &= \Pr(T_l = \tau_l | X_{l-1} = s_{i_{l-1}}, X_l = s_{i_l}) \times \Pr(X_l = s_{i_l} | X_{l-1} = s_{i_{l-1}}). \end{aligned}$$

Combining the two previous equations completes the derivation.

B.2.2 Derivation of the KL Distance

Step (2) of the algorithm in Section 3.3 requires the calculation of the KL distance between the estimated one-step transition distributions of each patient sequence and the overall population. Let $\bar{P}$ be the transition matrix corresponding to profile k, and P be the transition matrix of observation r belonging to profile k. Likewise, let $\bar{\Lambda}$ contain the MLEs for the exponentialy distributed interarrival times for profile k, and Λ contain the MLEs for observation r belonging to profile k. Let $\bar{P}_{i,j}$, $P_{i,j}$, $\bar{\Lambda}_{i,j}$, and $\Lambda_{i,j}$ denote the transition probabilities and expected interarrival times between states s_i and s_j. We can now derive a closed-form solution of the average KL distance between the transition distributions out of state s_i for an individual and a population.

Consider a utilization sequence such that $X_l = s_i$ at time t. We want to compare the probability of transition at time $T + \tau$ to state s_j of the patient to that of the all patients within the profile, where T was the last arrival time. This distribution is a finite mixture of

exponential distributions: given that the next event is state s_j occuring with probability P_{ij}, the interarrival time is given by $Exp(\lambda_{ij})$. Using the P and Λ matrices we derive the KL distance between the individual and cluster distributions:

$$d(P,\Lambda||\bar{P},\bar{\Lambda}) = \sum_j \int_0^\infty P_{i,j}\lambda_{i,j}\exp\{-\lambda_{i,j}\tau\} \times \log\left(\frac{P_{i,j}\lambda_{i,j}\exp\{-\lambda_{i,j}\tau\}}{\bar{P}_{i,j}\bar{\lambda}_{i,j}\exp\{-\bar{\lambda}_{i,j}\tau\}}\right) d\tau$$

$$= \sum_j P_{i,j}\int_0^\infty \lambda_{i,j}\exp\{-\lambda_{i,j}\tau\} \times \left[\log\left(\frac{P_{i,j}\lambda_{i,j}}{\bar{P}_{i,j}\bar{\lambda}_{i,j}}\right) + \log\left(\exp\{\bar{\lambda}_{i,j}\tau - \lambda_{i,j}\tau\}\right)\right] d\tau$$

$$= \sum_j P_{i,j}\log\left(P_{i,j}/\bar{P}_{i,j}\right) + \log\left(\lambda_{i,j}/\bar{\lambda}_{i,j}\right) + \bar{\lambda}_{i,j}\int_0^\infty \tau\lambda_{i,j}\exp\{-\lambda_{i,j}\tau\}d\tau$$

$$-\lambda_{i,j}\int_0^\infty \tau\lambda_{i,j}\exp\{-\lambda_{i,j}\tau\}d\tau$$

$$= \sum_j P_{i,j}\left[\log\left(P_{i,j}/\bar{P}_{i,j}\right) + \log\left(\lambda_{i,j}/\bar{\lambda}_{i,j}\right) + \bar{\lambda}_{i,j}/\lambda_{i,j} - 1\right]$$

Finally, we want to average across all states $s_i, i \in \{1,\ldots,S\}$, so we use the measure:

$$D_{ave}(P,\Lambda||\bar{P},\bar{\Lambda}) = \frac{\sum_{i=1}^S d(P,\Lambda||\bar{P},\bar{\Lambda})}{S}.$$

B.3 Quantifying Cost-Saving Interventions

In this section we describe the methods for quantifying the cost-savings of interventions that bend the adherence levels of patients to recommended care guidelines.

In Section 3.4 we provide a figure of the cost savings at various levels of improvement of adherence to recommended care guidelines. Now we clarify our approach. We desire to quantify the potential cost savings of a patient who *after* an ER or HO visit follows up with a PO or RX visit. We assume that we cannot change the initial transition probability to an ER or HO visit, and can only intervene afterwards to reduce readmission rates. Furthermore, we assume that any transition into the ER or HO after previous visits to PO and RX are likely to be due to emergencies and cannot be prevented. Under these assumptions, we can manipulate the transition matrix P_k in the following manner:

1. Given a transition matrix P_k, find the total transition probability from ER or HO back into ER or HO.

Table 3: Expected Cost-Savings Per Patient

Adh. Improvement	GA Prof. 1	GA Prof. 2	GA Prof. 3	GA Prof. 4
Current Cost	\$107.35	\$592.74	\$626.10	\$1,054.77
25%	\$0	\$5.03	\$17.89	\$52.91
50%	\$0	\$9.57	\$32.96	\$96.17
75%	\$0	\$13.72	\$45.79	\$132.16
100%	\$0	\$17.53	\$56.90	\$162.63
	NC Prof. 1	NC Prof. 2	NC Prof. 3	NC Prof. 4
Current Cost	\$1268.87	\$699.91	\$654.17	\$2,479.66
25%	\$1.02	\$1.83	\$3.38	\$64.39
50%	\$2.02	\$3.54	\$6.36	\$117.23
75%	\$2.95	\$5.12	\$8.99	\$161.43
100%	\$3.82	\$6.58	\$11.38	\$198.84

2. Re-weight part (or all) of this transition probability into the PO or RX nodes, proportional to the initial transition weights into PO or RX from ER or HO.

3. Re-calculate the expected number of visits to each state.

4. Multiply the expected number of visits to each state by the average cost per visit to each state in order to get the new expected cost per patient over the five year timespan.

The potential cost savings for each profile at 25%, 50%, 75%, and 100% adherence improvements to recommended care guidelines are given in Table B.3.

B.4 Interarrival Times

We provide the average interarrival times of the predominant transitions between the different provider types in Tables 4 and 5.

B.5 Model Selection

In this section we provide clustering trees for GA and NC, respectively in Figure 24. In the clustering trees, we plot the divisions of patients into underlying utilization profiles, represented as nodes, where each node is represented by a pie chart displaying the proportion of the contributions of the non-RX visits. The size of the nodes are proportional to

Table 4: Average interarrival times (in months): GA, Profiles 1-4 from top to bottom, left to right

	CL	ER	HO	PO	RX
CL	–	–	–	–	–
ER	–	–	–	–	–
HO	–	–	–	–	–
PO	–	–	–	8.2	5.0
RX	–	–	–	1.7	1.4

	CL	ER	HO	PO	RX
CL	–	–	–	–	–
ER	–	–	–	–	–
HO	–	–	7.3	0.6	1.0
PO	–	–	1.8	5.2	1.1
RX	–	–	0.7	0.4	1.5

	CL	ER	HO	PO	RX
CL	1.4	–	–	8.1	10.4
ER	–	8.5	–	0.8	8.7
HO	–	–	–	–	–
PO	1.4	9.2	–	2.3	2.1
RX	7.6	7.0	–	4.5	4.2

	CL	ER	HO	PO	RX
CL	–	–	–	–	–
ER	–	4.8	5.1	6.3	2.3
HO	–	5.6	4.0	7.6	5.0
PO	–	5.1	9.1	3.9	5.3
RX	–	2.8	5.0	0.8	2.2

Table 5: Average interarrival times (in months): NC, Profiles 1-4 from top to bottom, left to right

	CL	ER	HO	PO	RX
CL	–	–	–	–	–
ER	–	–	–	–	–
HO	–	–	–	–	–
PO	–	–	–	–	–
RX	–	–	–	–	1.5

	CL	ER	HO	PO	RX
CL	–	–	–	–	–
ER	–	–	–	–	–
HO	–	–	–	–	–
PO	–	–	–	6.3	1.2
RX	–	–	–	0.4	1.5

	CL	ER	HO	PO	RX
CL	0.1	–	–	8.1	8.2
ER	–	–	–	–	–
HO	–	–	5.8	9.9	9.7
PO	7.8	–	10.6	5.7	7.1
RX	5.2	–	9.8	5.2	5.3

	CL	ER	HO	PO	RX
CL	–	–	–	–	–
ER	–	4.0	1.9	2.2	2.4
HO	–	5.6	2.8	2.5	2.4
PO	–	4.0	3.9	2.0	2.2
RX	–	2.7	2	1.0	2.1

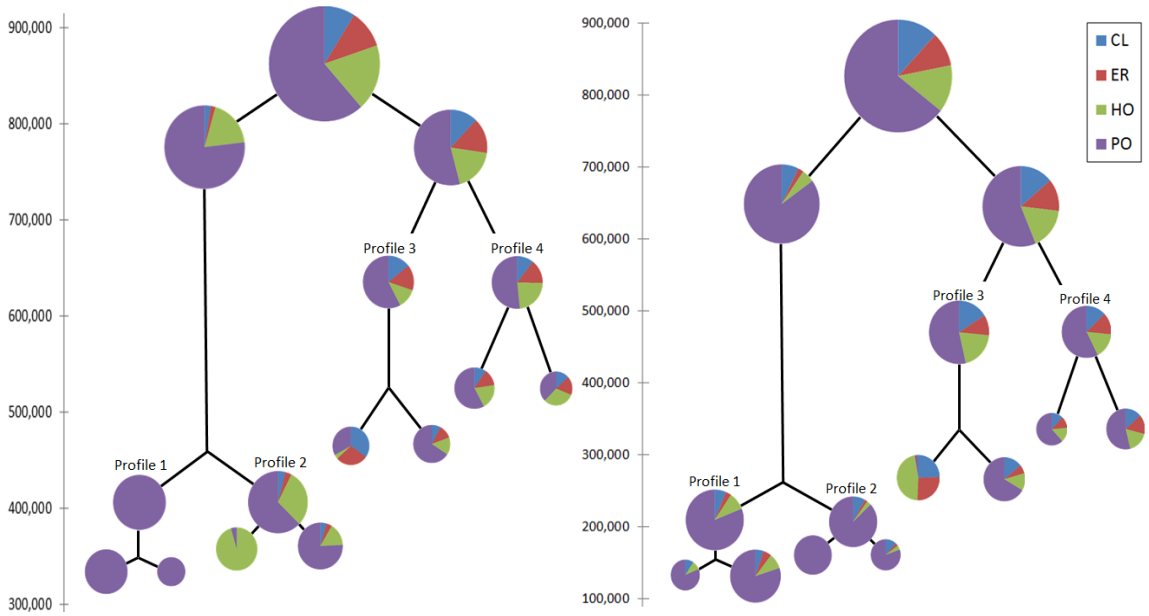

Figure 24: The clustering trees for GA (left) and NC (right). Here the nodes are located along y-axes according to the BIC score prior to the next splitting iteration. (Note: the y-axes in the two graphs are not on the same scale.) The size of each node is determined by the proportion of the population contained within the node. We do not include RX encounters in these charts in order to compare the visits to different provider types. The profiles we examine in our network graphs are labeled.

the percentage of the patient population contained in each node. The tree is laid out such that the splits occur at the value of the BIC score prior to the split. We do not include all splits in the tree graphs because divisions that occur further down in the trees tend to capture outlier profiles with low membership and are redundant in terms of the provider types they contain. Such divisions will be useful if the emphasis is on the identification of outlier behaviors; our objective in this study is to characterize underlying profiles.

The tree graphs for GA and NC have multiple notable characteristics in common. The nodes on the right side of the tree are more variational and result in larger improvements in the BIC score. The nodes on the left side are primarily composed of HO and PO visits and are typically less variational in terms of utilization. In the second partition in Figure 24, the patients belonging to Profile 1 in GA solely utilize PO (23%), while Profile 2 is almost evenly split between HO and PO visits (32%). On the right side of the tree, the delineation between provider types per profile is unclear; however, the networks in Figure

2 of the main text show that Profile 3 (23%) contains more CL visits than Profile 4 (22%), which has more ER and HO visits.

The NC clustering tree splits patients into seemingly similar profiles. Profile 1 (29%) almost solely consists of RX encounters with very little contribution from the other provider types and Profile 2 (20%) is dominated by PO visits. Similarly to GA, the second division on the right side does not separate patients into homogeneous clusters based on provider types; however, from the networks in Figure 3 of the main text, Profile 3 (30%) contains more CL visits, while Profile 4 (21%) is more evenly split between ER and HO visits.

While the division patterns are similar across both states, the pie charts corresponding to the cluster nodes indicate a greater dependence on HO in GA than in NC, where a higher rate of PO utilization occurs. Furthermore, referring to Figure 24, the clustering algorithm has better results in NC as indicated by the faster decrease in BIC scores in earlier divisions, indicating that NC may have more heterogeneous utilization patterns.

Figure 25 plots the (negative) BIC score and improvements with the addition of each profile for GA and NC, respectively. The BIC improves up until there are 126 and 79 profiles for GA and NC, respectively. If one considered all possible profiles, the results of the clustering analysis would be unintelligible simply because of the large number of profiles and parameters. Therefore, we look for the point in the BIC curve where the improvements with each profile division are small enough such that they no longer warrant higher model complexity in terms of additional profiles. In this case, we would select the first 30 profiles for GA and 25 for NC. As stated previously, further splits tend to isolate outlying behaviors while not necessarily capture the underlying profile structures.

B.6 Enrollment Lapse Summary

In this section we summarize the enrollment patterns of Medicaid eligible children within our study. To begin, we provide a summary of the enrollment behaviors and reasons for enrollment in Tables 6 and 7.

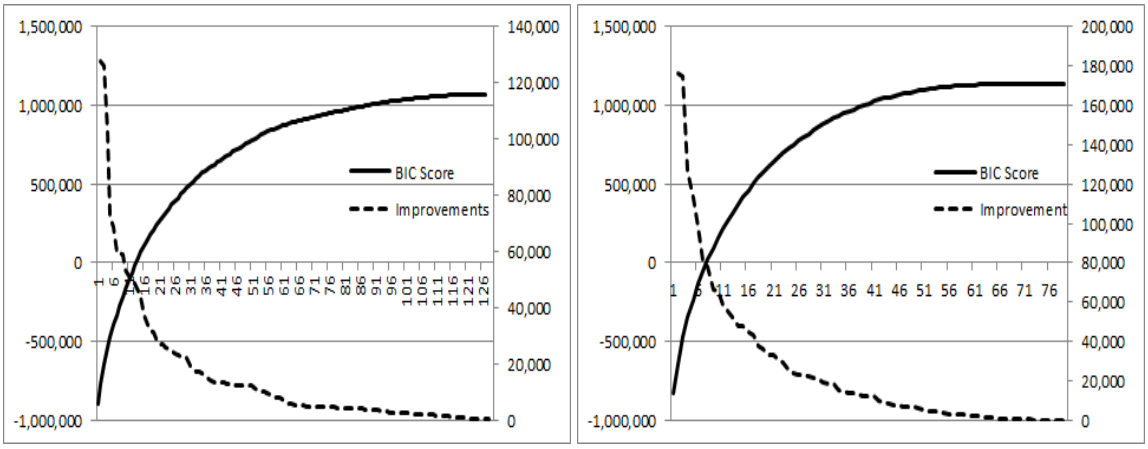

Figure 25: Plot of the negative BIC score and improvements with the number of profiles on the x-axis for GA (left) and NC (right). Note that we use two different y-axes for the BIC score and improvements.

Table 6: Months of Enrollment by Profile and Enrollment Reason (Blind/Disabled, Foster Care, Other): GA

Pr.	BD	FC	Oth	Possible Enr. Mths	Event Rate Mulitplier
1	95,820	47,274	937,282	1,653,120	1.53
2	144,645	57,473	1,227,858	2,298,288	1.61
3	119,792	40,198	1,023,210	1,670,796	1.41
4	139,816	42,735	1,048,028	1,624,152	1.32

As we state in Section 3.5, one of the shortcomings of modelling claims data as longitudinal utilization sequences is the fact that claims data are incomplete. Using the maximum likelihood estimator of the rate of events from a censored survival model where the inter-arrival times between events are assumed to be exponentially distributed we provide an event rate mulitplier (22). This multiplier suggests that if patients have the same utilization behavior when unenrolled as they do when they are enrolled, then the frequency of events will increase by the amount in the event rate multiplier. However due to the fact that the majority of patients are enrolled for reasons other than being blind or disabled it is highly unlikely that they have other types of insurance. We can, therefore, suggest that the majority of patients will be visiting only the emergency room during times of unenrollment.

Table 7: Months of Enrollment by Profile and Enrollment Reason (Blind/Disabled, Foster Care, Other): NC

Pr.	BD	FC	Oth	Possible Enr. Months	Event Rate Multiplier
1	154,400	57,318	1,460,626	2,430,396	1.45
2	101,160	31,022	973,837	1,609,896	1.46
3	171,866	44,560	1,627,954	2,429,628	1.32
4	147,076	26,608	1,126,189	1,692,372	1.30

APPENDIX C

CHAPTER IV: SUPPLEMENTARY MATERIALS

C.1 Patient Data Summaries

This appendix contains basic patient summaries for each of the clusters in Tables 8, 9, and 10. Table 8 gives the number of events for each type and cluster. Table 9 gives the total exposure (in years) across all patients in each cluster for each of the control covariates. We present exposure here instead of patient counts because the primary survival model input is exposure, not patient count. Table 10 gives the total number of patients in each cluster stratified by state and urbanicity.

C.2 Sample Patient Data

In this appendix, we present an example of a virtual patient to demonstrate of how the patient level data is transformed into the model inputs. Consider patient A who enters the system 30 days into the measurement period, visits the emergency room on days 90, 125, and 270 and fills a prescription for an inhaler on days 155 and 300. Suppose that the patient's Medicaid enrollment lapses between days 100 and 115. Let the length of the measurement period be 365 days. Table 11 provides a data summary for patient A. Table 12 demonstrates how the input table would be shaped for our estimation algorithm with

Table 8: Event Counts by Cluster

Event	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Total
CL	32,759	7,003	14,970	10,158	39,359	104,249
ER	86,410	16,975	24,709	13,927	19,410	161,431
HO	95,398	23,884	23,695	19,344	29,154	191,475
NP	18,700	6,704	3,797	3,770	6,141	39,112
PO	290,212	38,676	297,999	126,677	30,940	784,504
RX	450,456	655,085	444,585	1,020,559	104,473	2,675,158

Table 9: Exposure (in years) by Cluster and Control Variable

Var. Family	Var. Value	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
Exposure	Total	930,365	253,665	220,521	143,076	37,538
Age Group	4-5	395,286	107,361	94,376	63,915	14,701
	6-14	513,240	139,965	121,934	76,421	21,991
	15-17	21,839	6,339	4,211	2,740	846
Race	White	393,511	118,066	92,889	71,209	16,700
	Black	433,568	107,513	101,917	53,516	15,710
	Other	103,286	28,086	25,715	18,351	5,128
Health	Healthy	492,493	119,387	103,052	55,877	14,287
Status	Minor	162,819	45,748	30,631	23,521	3,913
	Chronic	268,704	86,037	85,828	61,389	18,787
	Severe	6,349	2,494	1,010	2,290	551
Medicaid	Blind/Disabled	85,004	24,888	18,832	17,970	5,458
Eligibility	Foster Care	21,263	6,390	5,054	4,855	1,067
	Other	824,098	222,387	196,635	120,251	31,013

Table 10: Patient Counts by Cluster

Var. Family	Var. Value	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
Count	Total	237,693	68,640	64,357	44,003	11,707
State	GA	54,196	7,220	15,337	6,755	4,018
	LA	60,892	15,911	9,013	5,500	874
	MS	16,283	9,528	2,155	3,375	954
	MN	1,404	5,396	1,606	1,826	808
	NC	44,757	14,720	21,558	14,616	2,842
	TN	46,162	16,314	14,689	11,930	2,210
Urbanicity	Urban	173,644	45,331	48,435	31,142	8,730
	Suburban	49,403	16,376	13,412	9,511	2,058
	Rural	14,647	6,931	2,510	3,350	919

Table 11: Sample Data Summary

Start Time	Stop Time	Event Type	Health Status
30	90	ER	Chronically Ill
90	100	0	Chronically Ill
115	125	ER	Chronically Ill
125	155	RX	Chronically Ill
155	270	ER	Chronically Ill
270	300	RX	Chronically Ill
300	365	0	Chronically Ill

Table 12: Sample Input for Estimation Algorithm

τ	$\delta_{\text{ER}}(\tau)$	$\delta_{\text{PO}}(\tau)$	$\delta_{\text{RX}}(\tau)$	$D_{r,1}(\tau)$: Last Event	$D_{r,2}(\tau)$: Health Status
60	1	0	0	0	0
10	0	0	0	ER	0
10	1	0	0	ER	0
30	0	0	1	ER	0
115	1	0	0	RX	0
30	0	0	1	ER	0
65	0	0	0	RX	0

last event type as a covariate in the study. This example demonstrates how our algorithm handles censoring, multivariate survival data, and time varying covariates.

One property of the exponential baseline assumption is that we can simply subtract the start time from the stop time due to the memoryless property of the exponential and use the same interarrival time across all event types but with different event indicator values. This allows for extreme computational efficiency that reduces to simple matrix algebra when estimating the coefficients $\vec{\beta}$. For other distributions, such as the Weibull or log-logistic, we would not be able to simply subtract the start times from the stop times in order to get the length of the time period until event or censoring, thus greatly increasing the complexity of the algorithm. Furthermore, as shown in Appendix C.3, with the exponential distribution we can derive simply the provider transition networks.

C.3 Proportional Hazards and Multinomial Coefficients and Interpretations

In this appendix, we provide the parameters from the model and explain how we derive the numerical outputs depicted in the figures in Section 4.4. Table 13 contains the raw proportional hazards parameters from the algorithm. In order to get the baseline rates and their multipliers from Sections 4.1.1 and 4.1.2, respectively, we simply take $\exp(\beta)$, where β is the coefficient value.

In order to determine the provider networks from Section 4.4 we first must determine the event rates for the different subpopulations. The subpopulation in Figure 3 of the main text are white patients, aged 4-5, who are neither blind nor disabled or in foster care. The middle column of networks pertains to chronically ill patients and the rate parameters are $\lambda_{0ks} = \exp(\beta_{0ks})$, where β_{0ks} are the baseline proportional hazard coefficients for cluster k and event s. Let $\beta_{\text{Healthy},s}$ and $\beta_{\text{Severe},s}$ be the coefficients for the healthy and severely ill patients, respectively. Then the event rates for these two groups are $\exp(\beta_{0ks} \times \beta_{\text{Healthy},s})$ and $\exp(\beta_{0ks} \times \beta_{\text{Severe},s})$. Furthermore, we can determine the rates for, say, a healthy patient with a last visit of CL by calculating $\exp(\beta_{0ks} \times \beta_{\text{Healthy},s} \times \beta_{\text{CL},s})$. Now we employ the following result on exponential random variables.

Let $T_1, \ldots, T_{|S|}$ be exponentially distributed random variables for the interarrival times for events $1, \ldots, |S|$ with parameters $\lambda_1, \ldots, \lambda_{|S|}$. Then it can be easily shown that the probability that T_s is the smallest of $T_1, \ldots, T_{|S|}$ is

$$\frac{\lambda_s}{\lambda_1 + \cdots + \lambda_{|S|}}.$$

These probabilites are the transition probabilities depicted in the provider networks.

The multinomial logistic regression model parameters are given in Table 14. Tables 15 through 19 contain the average interarrival times in years between events for the baseline group for Clusters 1-5. This is equal to $\exp(-\beta_{0ks}\beta_{\text{s,Event}})$.

Table 13: Proportional Hazards Coefficients

Var. Family	Var. Value	CL	ER	HO	NP	PO	RX
Baseline	$k = 1$	-2.80	-2.69	-2.15	-4.61	-1.37	-0.92
	$k = 2$	-4.31	-4.26	-2.76	-5.68	-3.58	0.44
	$k = 3$	-1.18	-4.34	-4.09	-5.90	-1.20	0.13
	$k = 4$	-2.55	-1.74	-1.54	-3.67	-0.06	1.46
	$k = 5$	-1.39	-1.73	-1.20	-3.74	-1.89	0.41
Medicaid	Blind/Disabled	0.22	-0.16	-0.10	-0.34	-0.31	0.05
Eligibility	Foster Care	-0.01	-0.72	-0.40	-0.29	-0.24	0.13
Health	Healthy	-1.18	-2.34	-2.24	-2.10	-2.11	-0.75
Condition	Minor Ilness	0.22	-0.04	0.20	0.27	0.24	0.43
	Severe Illness	1.56	1.39	1.92	1.01	1.38	0.70
Race/	Black	0.23	0.72	0.36	0.05	0.02	-0.07
Ethnicity	Other	0.21	0.31	0.29	0.03	0.14	-0.15
Age Group	6-14	-0.44	-0.10	-0.05	-0.12	-0.15	-0.02
	15-18	-0.09	0.44	0.42	0.22	0.20	-0.07
Previous	CL	2.64	-0.37	1.29	0.07	0.42	0.49
Event	ER	0.33	0.99	1.33	0.37	0.31	0.52
	HO	0.23	0.12	0.01	-0.06	0.11	0.58
	NP	0.00	-0.32	-0.36	3.23	0.15	0.76
	PO	0.51	-0.27	-0.16	0.32	0.79	0.80
	RX	0.36	-0.22	0.01	0.92	0.80	1.03

Table 14: Multinomial Logistic Regression Coefficients

Var. Family	Var. Value	$k = 1$	$k = 2$	$k = 3$	$k = 4$
	Baseline	-0.11	1.79	0.56	-0.79
State	LA	-0.99	-0.70	-1.35	-2.34
	MS	-0.90	-1.37	-2.05	-1.60
	MN	-1.04	-1.00	-2.01	-1.36
	NC	0.10	-0.86	-0.34	-1.01
	TN	-0.22	-0.95	-0.83	-1.39
Urbanicity	Suburban	-0.12	-0.05	-0.10	-0.25
	Rural	-0.17	-0.27	-0.66	-0.18
Access	Travel	0.16	0.55	0.55	0.57

Table 15: Average Interarrival Times (in years): Cluster 1

Previous Event	CL	ER	HO	NP	PO	RX
CL	1.71	8.28	6.83	31.57	1.24	1.46
ER	17.12	2.12	1.70	23.23	1.38	1.41
HO	18.96	5.04	1.64	35.83	1.71	1.33
NP	23.73	7.86	8.90	1.33	1.63	1.11
PO	14.30	7.51	7.30	24.56	0.86	1.07
RX	16.67	7.11	6.11	13.42	1.63	1.11

Table 16: Average Interarrival Times (in years): Cluster 2

Previous Event	CL	ER	HO	NP	PO	RX
CL	2.05	11.28	6.98	31.60	3.15	0.33
ER	20.52	2.89	1.74	23.25	3.51	0.32
HO	22.72	6.86	1.68	35.86	4.33	0.30
NP	28.45	10.71	9.09	4.13	1.33	0.25
PO	17.14	10.23	7.46	24.47	2.19	0.24
RX	19.98	9.69	6.25	13.43	2.17	0.19

Table 17: Average Interarrival Times (in years): Cluster 3

Previous Event	CL	ER	HO	NP	PO	RX
CL	1.00	7.43	6.37	44.27	0.40	0.41
ER	10.00	1.91	1.59	32.58	0.45	0.40
HO	11.08	4.52	1.53	50.24	0.55	0.38
NP	13.88	7.06	8.30	1.86	0.53	0.32
PO	8.36	6.74	6.81	34.29	0.28	0.30
RX	9.75	6.38	5.70	18.81	0.28	0.24

Table 18: Average Interarrival Times (in years): Cluster 4

Previous Event	CL	ER	HO	NP	PO	RX
CL	0.91	8.26	5.12	36.75	0.69	0.14
ER	9.17	2.12	1.28	27.04	0.77	0.14
HO	10.16	5.03	1.23	0.95	41.70	0.13
NP	12.72	7.85	6.68	1.54	0.91	0.11
PO	7.66	7.50	5.48	28.47	0.48	0.10
RX	8.92	7.10	4.59	15.62	0.48	0.08

Table 19: Average Interarrival Times (in years): Cluster 5

Previous Event	CL	ER	HO	NP	PO	RX
CL	0.12	2.12	1.42	0.32	6.99	0.72
ER	1.23	0.54	0.35	5.14	0.80	0.31
HO	1.36	1.29	0.34	7.93	0.99	0.29
NP	1.70	2.01	1.85	0.29	0.95	0.25
PO	1.03	1.92	1.52	0.24	5.41	0.50
RX	1.20	1.82	1.27	0.19	2.97	0.50

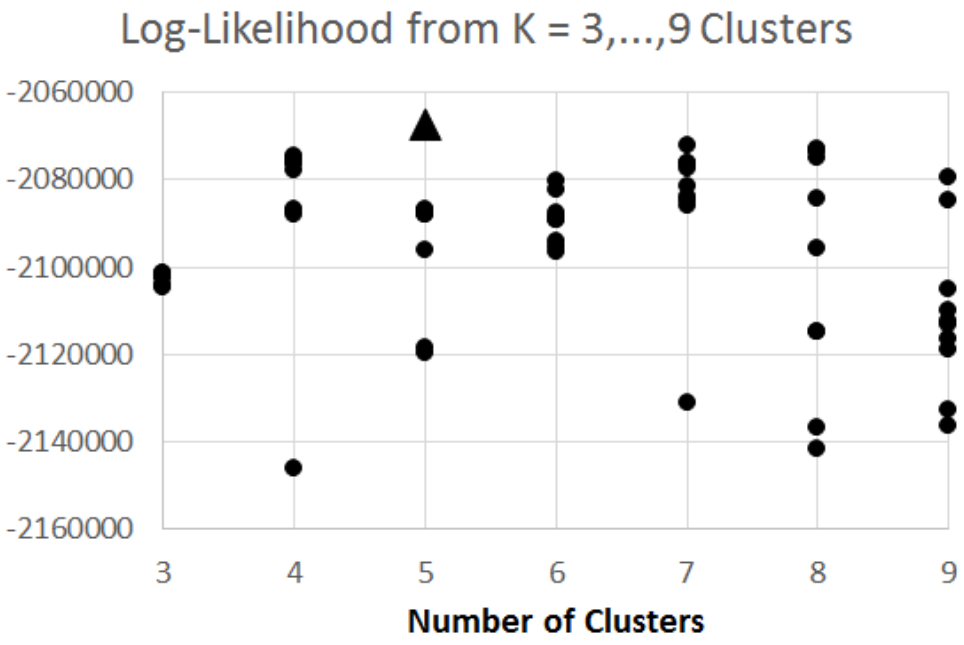

Figure 26: The resulting log-likelihood plotted from $N = 10$ different initializations for $K = 3, \ldots, 9$ clusters. We chose the model that resulted in the highest likelihood after convergence, with $K = 5$ clusters denoted by ▲.

C.4 Proportional Hazards Mixture Model Selection

In this appendix, we describe the initialization of the algorithm in order to produce the results in Section 4.4.1.3. We use a random initialization with $K = 3, 4, \ldots, 9$ clusters. The random initialization begins with a random (hard) clustering assignment for each patient to a cluster k, $k \in \{1, \ldots, K\}$. That is, $\boldsymbol{Z}_r^{(1)}$ has only one entry with value one and all others are zero. The vector $\vec{\boldsymbol{b}}$ begins with all values set to 0, such that $\pi_{rk} = 1/K$ for all k and for all r. The EM algorithm then proceeds through the iterations to find the maximum likelihood given this initialization. We repeat this with 10 different random initializations for each value of K.

Figure 26 displays the resulting likelihood of each of the initializations. In this article we present the model outputs from the initialization that produces the highest likelihood.

C.5 Statistical Significance of Proportional Hazards and Multinomial Coefficients

While the study in Chapter IV focuses on the practical significance of the effects of the covariates on patient utilization we present a summary of the statistical significance of the parameter estimates in this appendix. We calculate the estimates of the variance of the parameters by first finding the estimate for the Fisher information. The Fisher information is

$$I(\theta) = -E\left[\frac{\partial^2}{\partial\theta^2}\log f(X|\theta)\middle|\theta\right],$$

where θ is a model parameter and X is a random variable. In our case, we do not know the parameters $\vec{\boldsymbol{\beta}}$ and $\vec{\boldsymbol{b}}$, and thus we must estimate the Fisher information. We have already calculated the 2nd derivative of the log-likelihood function with respect to β_{ps}, β_{0ks}, and b_{jk} in Section 3.3.3 from the main text. Then the estimates for the Fisher information are

$$\widehat{I(\beta_{ps})} = I(\widehat{\beta}_{ps}) = \sum_{r,k,l_r}\left[\tau_{l_r}\widehat{Z}_{rk}D^2_{rp}(\tau_{l_r})\exp\left\{\widehat{\boldsymbol{\beta}}_{\boldsymbol{ks}}{}^{\intercal}\boldsymbol{D_r}(\tau_{l_r})\right\}\right],$$

$$\widehat{I(\beta_{0ks})} = I(\widehat{\beta}_{0ks}) = \sum_{r,l_r}\left[\widehat{Z}_{rk}\tau_{l_r}\exp\left\{\widehat{\boldsymbol{\beta}}_{\boldsymbol{ks}}{}^{\intercal}\boldsymbol{D_r}(\tau_{l_r})\right\}\right], \text{ and}$$

$$\widehat{I(b_{jk})} = I(\widehat{b}_{jk}) = \sum_{r=1}^{R}\widehat{\pi}_{rk}(1-\widehat{\pi}_{rk})E^2_{rj}.$$

We use Wald's test statistic in order to calculate the p-value of the parameters,

$$\frac{\hat{\theta}-0}{\widehat{V(\theta)}},$$

where $\widehat{V(\theta)} = 1/I(\widehat{\theta})$, and $\theta = \{\beta_{ps}, b_{jk}\}$. That is, we assume that the control and explanatory covariates have no effect on the baseline rate of events or cluster membership, respectively. There are only 6 combinations of event types and control covariates for which the p-value is greater than 0.001: *Race: Other*, NP (p-value = 0.036); *Minor Illness*, ER (p-value = 0.001); *Prior Event: CL*, NP (p-value = 0.099); *Prior Event: HO*, NP (p-value = 0.027), *Prior Event: NP*, CL (p-value = 0.921), and *Foster Care*, CL (p-value = 0.777). All of the multinomial logistic regression parameter estimates are significant at a $\alpha = 0.001$

Table 20: 99% Confidence Intervals for Baseline Rates by Event and Cluster

Event	$k = 1$	$k = 2$	$k = 3$	$k = 4$	$k = 5$
CL	(0.041, 0.043)	(0.034, 0.036)	(0.070, 0.074)	(0.076, 0.081)	(0.572, 0.599)
ER	(0.172, 0.177)	(0.125, 0.132)	(0.190, 0.199)	(0.170, 0.180)	(0.666, 0.701)
HO	(0.159, 0.164)	(0.154, 0.162)	(0.169, 0.177)	(0.210, 0.220)	(0.759, 0.793)
NP	(0.029, 0.031)	(0.028, 0.031)	(0.020, 0.022)	(0.024, 0.027)	(0.128, 0.141)
PO	(0.523, 0.532)	(0.204, 0.211)	(1.611, 1.637)	(0.935, 0.955)	(0.890, 0.922)
RX	(0.419, 0.424)	(1.834, 1.853)	(1.477, 1.494)	(4.266, 4.307)	(1.884, 1.923)

critical value. For the baseline rates, we provide 99% confidence intervals in Table 20. From this table we can see that the following cluster pairs and event types have statistically insignificant differences at the $\alpha = 0.01$ confidence level: Clusters 1 and 2: HO and NP; Clusters 1 and 4: ER. These findings suggest that the practical interpretations provided in the main body of the paper are also statistically significant with few exceptions.

C.6 Additional Transition Networks

In this appendix, we provide the provider transition networks for the other covariate familes: age group, race/ethnicity, and Medicaid eligibility categorization in Figures 27, 28, and 29, respectively. In each case, we consider the baseline group for the other covariates.

The *Age* networks do not show that *Age 6-14* patients show higher probability connections into HO, but less variation otherwise while and more reliance on RX while *Age 15-17* shows greater variation in provider types with more transitions leading to CL, ER, HO, and PO. The *Race* networks have the same pattern with greater variations for *Black* and *Other* groups in Cluster 1, 3, and 5. Clusters 2 and 4 non-white patients utilize more HO and PO, respectively. Finally, it appears that for the baseline group *Blind/Disabled* and *Foster Care* patients have less variation than those in *Other* and have stronger transitions leading to RX. These plots show that except for the *Medicaid Eligibility* variable the baseline group is less variational than others indicating a possibility that white children, age 4-5 are better managed in asthma care.

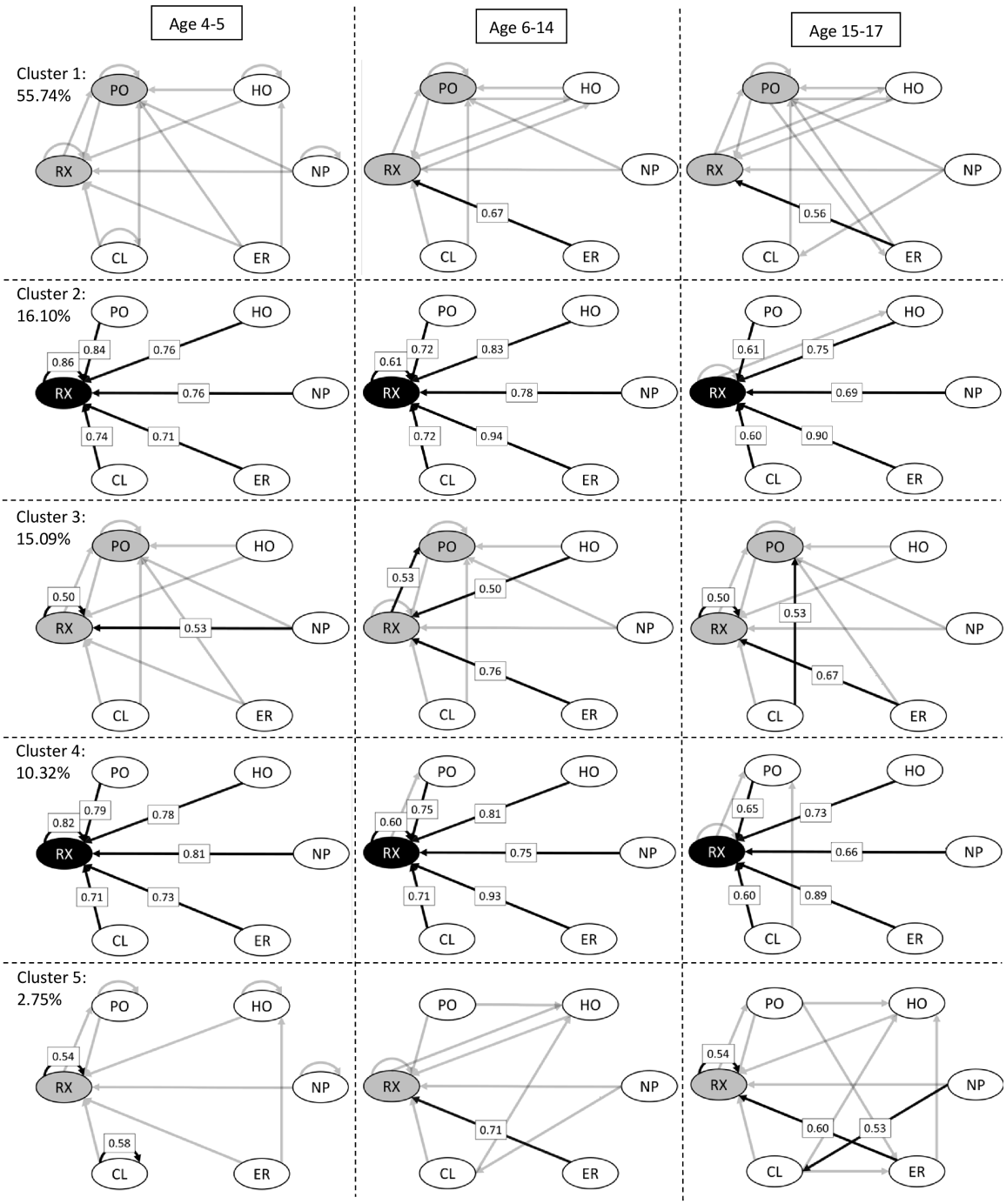

Figure 27: Provider networks for *Age* subgroups induced from the proportional hazards coefficients. The following rules were used in setting the grayscale of the coefficients and nodes: $< 0.2 \rightarrow$ not shown/white, $[0.2, 0.5) \rightarrow$ gray, and $\geq 0.5 \rightarrow$ black.

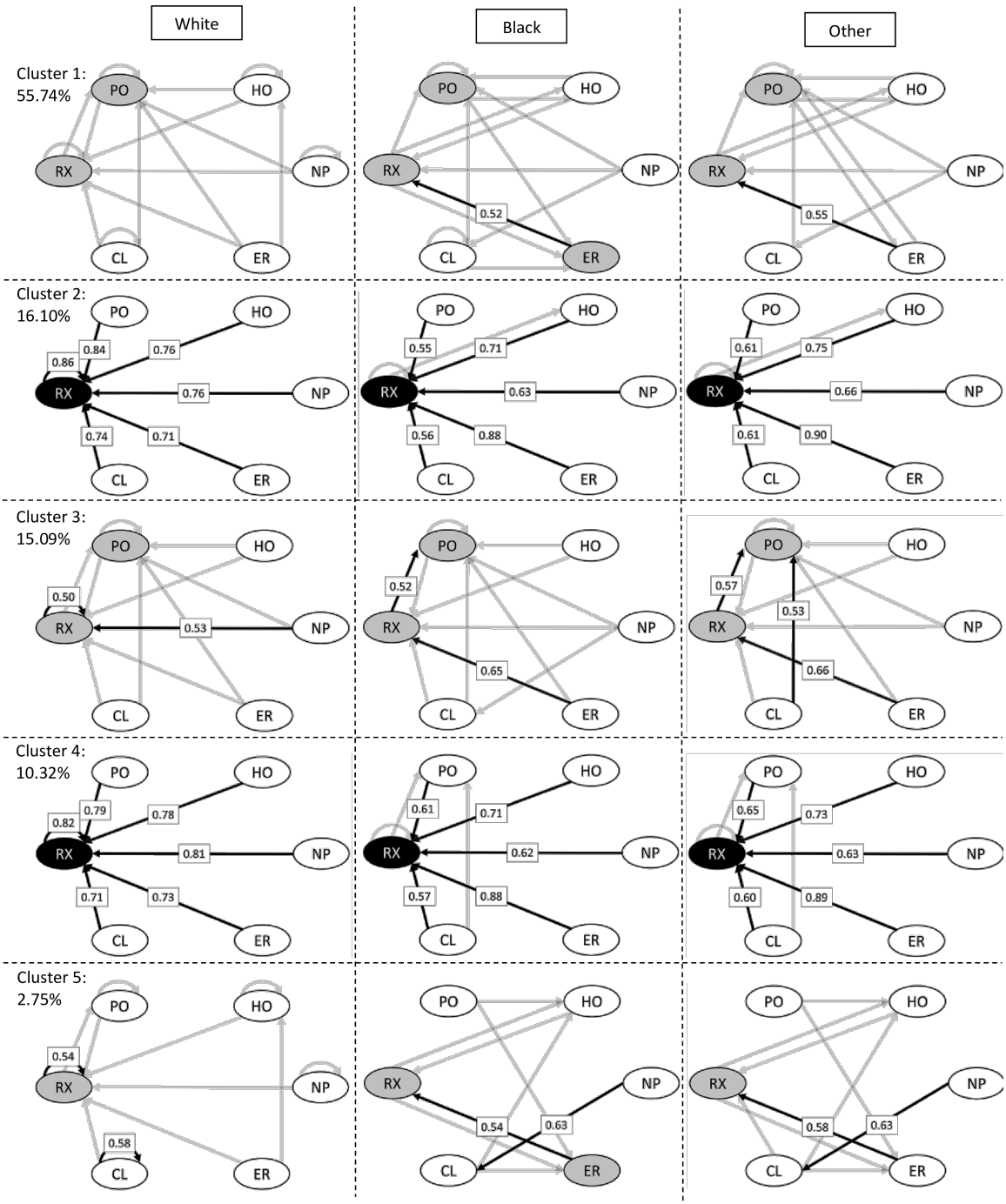

Figure 28: Provider networks for *Race* induced from the proportional hazards coefficients. The following rules were used in setting the grayscale of the coefficients and nodes: $<$ $0.2 \rightarrow$ not shown/white, $[0.2, 0.5) \rightarrow$ gray, and $\geq 0.5 \rightarrow$ black.

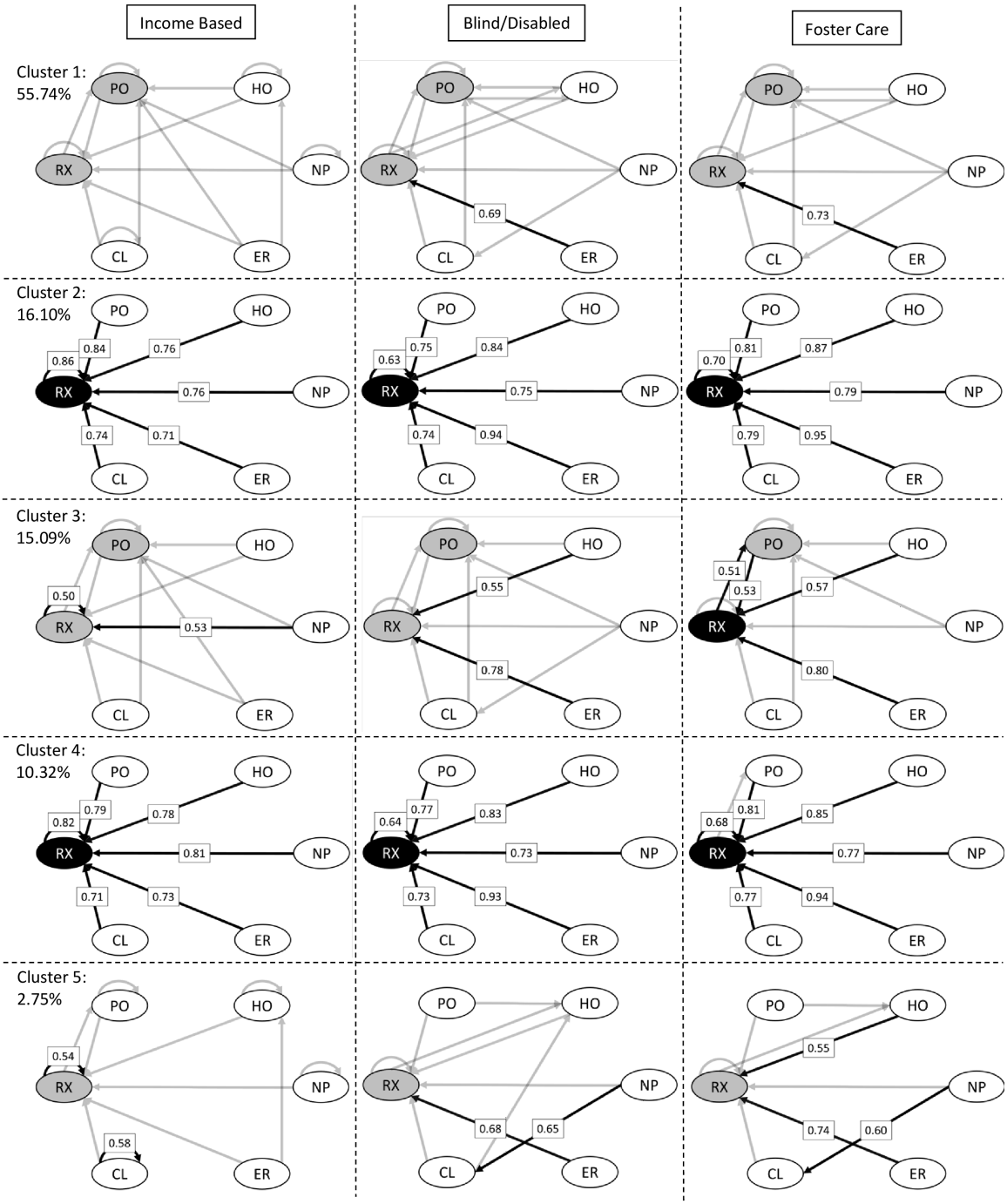

Figure 29: Provider networks for *Medicaid Eligbility* subgroups induced from the proportional hazards coefficients. The following rules were used in setting the grayscale of the coefficients and nodes: $< 0.2 \rightarrow$ not shown/white, $[0.2, 0.5) \rightarrow$ gray, and $\geq 0.5 \rightarrow$ black.

Bibliography

[1] 3M, "Core grouping software," 2014.

[2] AALEN, O., "Nonparametric inference for a family of counting processes," The Annals of Statistics, pp. 701–726, 1978.

[3] ANDERSEN, P. K. and GILL, R. D., "Cox's regression model for counting processes: a large sample study," The Annals of Statistics, pp. 1100–1120, 1982.

[4] ANDREOPOULOS, B., AN, A., WANG, X., and SCHROEDER, M., "A roadmap of clustering algorithms: finding a match for a biomedical application," Briefings in Bioinformatics, vol. 10, no. 3, pp. 297–314, 2009.

[5] ARELLANO, A. R. D. and WOLFE, S., "Unsettling scores: A ranking of state Medicaid programs. 2007," 2011.

[6] BÄHLER, C., HUBER, C. A., BRÜNGGER, B., and REICH, O., "Multimorbidity, health care utilization and costs in an elderly community-dwelling population: a claims data based observational study," BMC Health Services Research, vol. 15, no. 1, p. 23, 2015.

[7] BECKER, M. and LAUE, R., "A comparative survey of business process similarity measures," Computers in Industry, vol. 63, no. 2, pp. 148–167, 2012.

[8] BLOOM, B., JONES, L., and G., F., "Summary health statistics for U.S. children: National health interview survey," Vital Health Statistics, vol. 10, 2013.

[9] BLOSSFELD, H.-P. and HAMERLE, A., "Unobserved heterogeneity in event history models," Quality and Quantity, vol. 26, no. 2, pp. 157–168, 1992.

[10] BLUM, T., PADOY, N., FEUNER, H., and NAVAB, N., "Workflow mining for visualization and analysis of surgeries," International Journal of Computer Assisted Radiology and Surgery, vol. 3, no. 5, pp. 379–386, 2008.

[11] BORGAN, Ø., "Maximum likelihood estimation in parametric counting process models, with applications to censored failure time data," Scandinavian Journal of Statistics, vol. 11, no. 1, pp. 1–16, 1984.

[12] BRAUNSTEIN, M. L., Health informatics in the Cloud. Springer, 2012.

[13] BROMILEY, P., "Products and convolutions of gaussian distributions," tech. rep., Tina Memo, 2003.

[14] BROWNING, M. and CARRO, J. M., "Heterogeneity in dynamic discrete choice models," Econometrics Journal, vol. 13, no. 1, pp. 1–39, 2010.

[15] BUČAR, T., NAGODE, M., and FAJDIGA, M., "Reliability approximation using finite Weibull mixture distributions," Reliability Engineering & System Safety, vol. 84, no. 3, pp. 241–251, 2004.

[16] BUCZAK, A., BABIN, S., and MONIZ, L., "Data-driven approach for creating synthetic electronic medical records," BMC Medical Informatics and Decision Making, vol. 10, no. 1, p. 59, 2010.

[17] BUREAU OF BUSINESS & ECONOMIC RESEARCH, UNM, "Per capita personal income by state." https://bber.unm.edu/econ/us-pci.htm/, 2013.

[18] CENTERS FOR MEDICARE & MEDICAID SERVICES, "Medicaid Analytic eXtract (MAX)," 2012.

[19] CHANDOLA, V., SUKUMAR, S. R., and SCHRYVER, J. C., "Knowledge discovery from massive healthcare claims data," in Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1312–1320, ACM.

[20] CHANG, J., FREED, G. L., PROSSER, L. A., PATEL, I., ERICKSON, S. R., BAGOZZI, R. P., and BALKRISHNAN, R., "Comparisons of health care utilization outcomes in children with asthma enrolled in private insurance plans versus Medicaid," Journal of Pediatric Health Care, vol. 28, no. 1, pp. 71–79, 2014.

[21] COX, D. R., "Regression models and life-tables," Journal of the Royal Statistical Society: Series B (Methodological), vol. 34, no. 2, pp. 187–220, 1972.

[22] COX, D. R. and OAKES, D., Analysis of survival data, vol. 21. CRC Press, 1984.

[23] CZEPIEL, S. A., "Maximum likelihood estimation of logistic regression models: theory and implementation," tech. rep., 2002.

[24] DELIAS, P., DOUMPOS, M., MANOLITZAS, P., GRIGOROUDIS, E., and MATSATSINIS, N., "Clustering healthcare processes with a robust approach," in Proceedings of the 26th European Conference on Operational Research.

[25] DEMPSTER, A. P., LAIRD, N. M., and RUBIN, D. B., "Maximum likelihood from incomplete data via the EM algorithm," Journal of the Royal Statistical Society. Series B (Methodological), pp. 1–38, 1977.

[26] DEPARTMENT OF HEALTH AND HUMAN SERVICES, CDC, "National Health Interview Survey (NHIS) data: 2011 lifetime and current asthma," tech. rep., CDC, 2012.

[27] DEPARTMENT OF VETERANS AFFAIRS, "Patient treatment file (PTF)," 2013.

[28] DEVOE, J. E., GOLD, R., MCINTIRE, P., PURO, J., CHAUVIE, S., and GALLIA, C. A., "Electronic health records vs Medicaid claims: completeness of diabetes preventive care data in community health centers," The Annals of Family Medicine, vol. 9, no. 4, pp. 351–358, 2011.

[29] DONOHO, D. L. and JOHNSTONE, I. M., "Adapting to unknown smoothness via wavelet shrinkage," Journal of the American Statistical Association, vol. 90, no. 432, pp. 1200–1224, 1995.

[30] DONOHO, D. L. and JOHNSTONE, J. M., "Ideal spatial adaptation by wavelet shrinkage," Biometrika, vol. 81, no. 3, pp. 425–455, 1994.

[31] DUNN, R., READER, S., and WRIGLEY, N., "A nonparametric approach to the incorporation of heterogeneity into repeated polytomous choice models of urban shopping behaviour," Transportation Research Part A: General, vol. 21, no. 45, pp. 327–343, 1987.

[32] EMC CORPORATION, "Vertical industry brief: digital universe driving data growth in health care," tech. rep., 2014.

[33] FARCOMENI, A. and NARDI, A., "A two-component Weibull mixture to model early and late mortality in a Bayesian framework," Computational Statistics & Data Analysis, vol. 54, no. 2, pp. 416–428, 2010.

[34] FAREWELL, V. T., "The use of mixture models for the analysis of survival data with long-term survivors," Biometrics, pp. 1041–1046, 1982.

[35] FERREIRA, D. R., "Applied sequence clustering techniques for process mining," Handbook of Research on Business Process Modeling, Information Science Reference, IGI Global, pp. 492–513, 2009.

[36] GENKIN, A., LEWIS, D. D., and MADIGAN, D., "Large-scale Bayesian logistic regression for text categorization," Technometrics, vol. 49, no. 3, pp. 291–304, 2007.

[37] GENTILI, M., SERBAN, N., O'CONOR, J., and SWANN, J., "Quantifying disparities in accessibility and availability of pediatric primary care with implication for policy," under review, 2015.

[38] GOLUB, G. and PEREYRA, V., "Separable nonlinear least squares: the variable projection method and its applications," Inverse Problems, vol. 19, no. 2, p. R1, 2003.

[39] GREENE, W. H. and HENSHER, D. A., "A latent class model for discrete choice analysis: contrasts with mixed logit," Transportation Research Part B: Methodological, vol. 37, no. 8, pp. 681–698, 2003.

[40] GRONWALD, W. and KALBITZER, H. R., "Automated structure determination of proteins by NMR spectroscopy," Progress in Nuclear Magnetic Resonance Spectroscopy, vol. 44, no. 1, pp. 33–96, 2004.

[41] GROSSE, S. D., BOULET, S. L., GRANT, A. M., HULIHAN, M. M., and FAUGHNAN, M. E., "The use of US health insurance data for surveillance of rare disorders: hereditary hemorrhagic telangiectasia," Genetics in Medicine, vol. 16, no. 1, pp. 33–39, 2013.

[42] GROSSMANN, A. and MORLET, J., "Decomposition of Hardy functions into square integrable wavelets of constant shape," SIAM Journal on Mathematical Analysis, vol. 15, no. 4, pp. 723–736, 1984.

[43] GUERRY, P. and HERRMANN, T., "Advances in automated NMR protein structure determination," Quarterly Reviews of Biophysics, vol. 44, no. 03, pp. 257–309, 2011.

[44] GUNTERT, P., "Automated NMR protein structure calculation," Progress in Nuclear Magnetic Resonance Spectroscopy, vol. 43, no. 3, pp. 105–125, 2003.

[45] HECKMAN, J. J., Heterogeneity and State Dependence, pp. 91–140. University of Chicago Press, 1981.

[46] HECKMAN, J. J. and BORJAS, G. J., "Does unemployment cause future unemployment? Definitions, questions and answers from a continuous time model of heterogeneity and state dependence," Economica, vol. 47, no. 187, pp. 247–283, 1980.

[47] HOAGLIN, D. C., MOSTELLER, F., and TUKEY, J. W., Understanding Robust and Exploratory Data Analysis, vol. 3. Wiley New York, 1983.

[48] HOCH, J. C. and STERN, A. S., NMR Data Processing. Wiley-Liss New York:, 1996.

[49] HUBER, C. A., SCHNEEWEISS, S., SIGNORELL, A., and REICH, O., "Improved prediction of medical expenditures and health care utilization using an updated chronic disease score and claims data," Journal of Clinical Epidemiology, vol. 66, no. 10, pp. 1118–1127, 2013.

[50] IBM, "Bringing big data to the enterprise."

[51] KAISER FAMILY FOUNDATION, "State health facts: Medicaid & CHIP." `http://kff.org/state-category/medicaid-chip/`, 2014.

[52] KARRER, B. and NEWMAN, M. E., "Random graph models for directed acyclic networks," Physical Review E, vol. 80, no. 4, p. 046110, 2009.

[53] KHANNA, S., BOYLE, J., GOOD, N., LIND, J., and ZEITZ, K., "Time based clustering for analyzing acute hospital patient flow," in Engineering in Medicine and Biology Society (EMBC), 2012 Annual International Conference of the IEEE, pp. 5903–5906, IEEE.

[54] KORADI, R., BILLETER, M., ENGELI, M., GNTERT, P., and WTHRICH, K., "Automated peak picking and peak integration in macromolecular NMR spectra using AUTOPSY," Journal of Magnetic Resonance, vol. 135, no. 2, pp. 288–297, 1998.

[55] KUK, A. Y. and CHEN, C.-H., "A mixture model combining logistic regression with proportional hazards regression," Biometrika, vol. 79, no. 3, pp. 531–541, 1992.

[56] LEE, B. Y., MCGLONE, S. M., SONG, Y., AVERY, T. R., EUBANK, S., CHANG, C.-C., BAILEY, R. R., WAGENER, D. K., BURKE, D. S., and PLATT, R., "Social network analysis of patient sharing among hospitals in Orange County, California," American Journal of Public Health, vol. 101, no. 4, p. 707, 2011.

[57] LEE, S.-J., Visualization of Clinical Practice Guidelines and Patient Care Process. PhD thesis, 2006.

[58] LI, W., JAROSZEWSKI, L., and GODZIK, A., "Clustering of highly homologous sequences to reduce the size of large protein databases," Bioinformatics, vol. 17, no. 3, pp. 282–283, 2001.

[59] LINDEBERG, T., "Scale-space behaviour of local extrema and blobs," Journal of Mathematical Imaging and Vision, vol. 1, no. 1, pp. 65–99, 1992.

[60] LINDEBERG, T., "Detecting salient blob-like image structures and their scales with a scale-space primal sketch: a method for focus-of-attention," International Journal of Computer Vision, vol. 11, no. 3, pp. 283–318, 1993.

[61] LINDEBERG, T., Scale-space Theory in Computer Vision. Springer, 1994.

[62] LINDEBERG, T., "Feature detection with automatic scale selection," International Journal of Computer Vision, vol. 30, no. 2, pp. 79–116, 1998.

[63] LIOU, F.-M., TANG, Y.-C., and CHEN, J.-Y., "Detecting hospital fraud and claim abuse through diabetic outpatient services," Health Care Management Science, vol. 11, no. 4, pp. 353–358, 2008.

[64] LOVE, D., CUSTER, W., and MILLER, P., "All-payer claims databases: state initiatives to improve health care transparency," Issue Brief, The Commonwealth Fund, September, 2010.

[65] MAIR, P. and HUDEC, M., "Multivariate Weibull mixtures with proportional hazard restrictions for dwell-time-based session clustering with incomplete data," Journal of the Royal Statistical Society: Series C (Applied Statistics), vol. 58, no. 5, pp. 619–639, 2009.

[66] MALLAT, S. and HWANG, W. L., "Singularity detection and processing with wavelets," Information Theory, IEEE Transactions on, vol. 38, no. 2, pp. 617–643, 1992.

[67] MALLAT, S., A Wavelet Tour of Signal Processing. Elsevier, 1999.

[68] MCGEE, M. K., "North Carolina fights Medicaid fraud with analytics," 2012.

[69] MCGRADY, M. E. and HOMMEL, K. A., "Medication adherence and health care utilization in pediatric chronic illness: A systematic review," Pediatrics, vol. 132, no. 4, pp. 730–740, 2013.

[70] MCLACHLAN, G. and MCGIFFIN, D., "On the role of finite mixture models in survival analysis," Statistical Methods in Medical Research, vol. 3, no. 3, pp. 211–226, 1993.

[71] MENG, X.-L. and RUBIN, D. B., "Maximum likelihood estimation via the ECM algorithm: A general framework," Biometrika, vol. 80, no. 2, pp. 267–278, 1993.

[72] MILLER, P. B., LOVE, D., SULLIVAN, E., PORTER, J., and COSTELLO, A., "All-payer claims databases," Robert Woods Johnson Foundation, 2010.

[73] MITTAL, S., MADIGAN, D., CHENG, J. Q., and BURD, R. S., "Large-scale parametric survival analysis," Statistics in Medicine, vol. 32, no. 23, pp. 3955–3971, 2013.

[74] MORRILL, R., CROMARTIE, J., and HART, G., "RUCA data," 2005.

[75] MOSLER, K., "Mixture models in econometric duration analysis," Applied Stochastic Models in Business and Industry, vol. 19, no. 2, pp. 91–104, 2003.

[76] MOSLER, K. and SCHEICHER, C., "Homogeneity testing in a Weibull mixture model," Statistical Papers, vol. 49, no. 2, pp. 315–332, 2008.

[77] MOSLER, K. and SEIDEL, W., "Theory & methods: Testing for homogeneity in an exponential mixture model," Australian & New Zealand Journal of Statistics, vol. 43, no. 2, pp. 231–247, 2001.

[78] NAGODE, M. and FAJDIGA, M., "An improved algorithm for parameter estimation suitable for mixed Weibull distributions," International Journal of Fatigue, vol. 22, no. 1, pp. 75–80, 2000.

[79] NATIONAL INSTITUTE OF HEALTH, "National asthma education and prevention program, third expert panel on the diagnosis and management of asthma," Clinical Practice Guidelines, 2007.

[80] NATIONAL RESEARCH COUNCIL, Frontiers in Massive Data Analysis. Washington, D.C.: The National Academies Press, 2013.

[81] NENADIC, Z. and BURDICK, J. W., "Spike detection using the continuous wavelet transform," Biomedical Engineering, IEEE Transactions on, vol. 52, no. 1, pp. 74–87, 2005.

[82] NEWMAN, M. E. and LEICHT, E. A., "Mixture models and exploratory analysis in networks," Proceedings of the National Academy of Sciences, vol. 104, no. 23, pp. 9564–9569, 2007.

[83] NOBLES, M., SERBAN, N., and SWANN, J., "Spatial accessibility of pediatric primary healthcare: Measurement and inference," The Annals of Applied Statistics, vol. 8, no. 4, pp. 1922–1946, 2014.

[84] ORTEGA, P. A., FIGUEROA, C. J., and RUZ, G. A., "A medical claim fraud/abuse detection system based on data mining: A case study in Chile," Proceedings of the 2006 International Conference on Data Mining, DMIN, vol. 6, pp. 26–29, 2006.

[85] PHILLIPS, K. A., MORRISON, K. R., ANDERSEN, R., and ADAY, L. A., "Understanding the context of healthcare utilization: assessing environmental and provider-related variables in the behavioral model of utilization," Health Services Research, vol. 33, no. 3 Pt 1, p. 571, 1998.

[86] PHUA, C., LEE, V., SMITH, K., and GAYLER, R., "A comprehensive survey of data mining-based fraud detection research," 2010.

[87] PIECORO, L. T., POTOSKI, M., TALBERT, J. C., and DOHERTY, D. E., "Asthma prevalence, cost, and adherence with expert guidelines on the utilization of health care services and costs in a state Medicaid population," Health Services Research, vol. 36, no. 2, pp. 357–371, 2001.

[88] PRESS, G., "A very short history of big data," Forbes, 2013.

[89] PYLYPCHUK, Y. and SARPONG, E. M., "Comparison of health care utilization: United States versus Canada," Health Services Research, vol. 48, no. 2, pp. 560–581, 2013.

[90] RAMONI, M., SEBASTIANI, P., and COHEN, P., "Multivariate clustering by dynamics," in AAAI/IAAI, pp. 633–638.

[91] RAMONI, M., SEBASTIANI, P., and COHEN, P., "Bayesian clustering by dynamics," Machine learning, vol. 47, no. 1, pp. 91–121, 2002.

[92] READER, S., "Unobserved heterogeneity in dynamic discrete choice models," Environment and Planning A, vol. 25, no. 4, pp. 495–519, 1993.

[93] REBUGE, A. J. D. S., "Business process analysis in healthcare environments," 2012.

[94] REID, P. P., COMPTON, W. D., GROSSMAN, J. H., and FANJIANG, G., Building a Better Delivery System: a New Engineering/Health Care Partnership. National Academies Press, 2005.

[95] RIDGEWAY, G., "Finite discrete Markov process clustering," tech. rep., Technical Report TR 97-24, Microsoft Research, Redmond, WA, 1997.

[96] ROEBUCK, M. C., LIBERMAN, J. N., GEMMILL-TOYAMA, M., and BRENNAN, T. A., "Medication adherence leads to lower health care use and costs despite increased drug spending," Health Affairs, vol. 30, no. 1, pp. 91–99, 2011.

[97] ROSS, J. S., MAYNARD, C., KRUMHOLZ, H. M., SUN, H., RUMSFELD, J. S., NORMAND, S.-L. T., WANG, Y., and FIHN, S. D., "Use of administrative claims models to assess 30-day mortality among Veterans Health Administration hospitals," Medical care, vol. 48, no. 7, pp. 652–658, 2010.

[98] ROUSE, W. B. and SERBAN, N., Understanding and Managing the Complexity of Healthcare. MIT Press, 2014.

[99] SABAU, A. S., "Survey of clustering based financial fraud detection research," Informatica Economica, vol. 16, no. 1, pp. 110–122, 2012.

[100] SERBAN, N., Analysis of multiple curves and multiple peaks with application to molecular biology. PhD thesis, 2005.

[101] SERBAN, N., "MICE: Multiple peak identification, characterization, and estimation," Biometrics, vol. 63, no. 2, pp. 531–539, 2007.

[102] SERBAN, N. and LI, P., "A statistical test for mixture detection with application to component identification in multi-dimensional biomolecular NMR studies.," Canadian Journal of Statistics, vol. In Press, 2013.

[103] SHEN, H.-W., WANG, D., SONG, C., and BARABSI, A.-L., "Modeling and predicting popularity dynamics via reinforced Poisson processes," 2014.

[104] SIDEN, H. and URBANOSKI, K., "Using network analysis to map the formal clinical reporting process in pediatric palliative care: a pilot study," BMC Health Services Research, vol. 11, no. 1, p. 343, 2011.

[105] SIMMONS, R. and DAVIS, R., The Roles of Knowledge and Representation in Problem Solving, pp. 27–45. Springer, 1993.

[106] SOKOL, L., GARCIA, B., WEST, M., RODRIGUEZ, J., and JOHNSON, K., "Precursory steps to mining HCFA health care claims," in System Sciences, 2001. Proceedings of the 34th Annual Hawaii International Conference on, p. 10 pp.

[107] SY, J. P. and TAYLOR, J. M. G., "Estimation in a Cox proportional hazards cure model," Biometrics, vol. 56, no. 1, pp. 227–236, 2000.

[108] TIEN, J. M. and GOLDSCHMIDT-CLERMONT, P., "Engineering healthcare as a service system," Information-Knowledge-Systems Management, vol. 8, pp. 277–297, 2009.

[109] TOMAR, D. and AGARWAL, S., "A survey on data mining approaches for healthcare," International Journal of Bio-Science and Bio-Technology, vol. 5, no. 5, pp. 241–266, 2013.

[110] TRAVAILLE, P., MLLER, R. M., THORNTON, D., and HILLEGERSBERG, J., "Electronic fraud detection in the US Medicaid healthcare program: lessons learned from other industries," 2011.

[111] U.S. BUREAU OF THE CENSUS, "Map – U.S. states estimated 2009 minority percentage." http://censuschannel.net/cc/news/u-s-state-2009-estimated-minority-percentage-1146/, 2010.

[112] U.S. CONGRESS, "Patient protection and Affordable Care Act," 2010.

[113] U.S. DEPARTMENT OF HEALTH & HUMAN SERVICES, "Standards for privacy of individually identifiable health information; final rule," 2002.

[114] VAUPEL, J. W. and YASHIN, A. I., "Heterogeneity's ruses: some surprising effects of selection on population dynamics," The American Statistician, vol. 39, no. 3, pp. 176–185, 1985.

[115] WAKEFIELD, D. B. and CLOUTIER, M. M., "Modifications to HEDIS and CSTE algorithms improve case recognition of pediatric asthma," Pediatric Pulmonology, vol. 41, no. 10, pp. 962–971, 2006.

[116] WANG, Y., "Jump and sharp cusp detection by wavelets," Biometrika, vol. 82, no. 2, pp. 385–397, 1995.

[117] WANG, Y., "Change curve estimation via wavelets," Journal of the American Statistical Association, vol. 93, no. 441, pp. 163–172, 1998.

[118] WONG, P. C., COWLEY, W., FOOTE, H., JURRUS, E., and THOMAS, J., "Visualizing sequential patterns for text mining," in Information Visualization, 2000. InfoVis 2000. IEEE Symposium on, pp. 105–111, IEEE.

[119] WU, C. J., "Statistics = data science," 1998.

[120] YOO, I., ALAFAIREET, P., MARINOV, M., PENA-HERNANDEZ, K., GOPIDI, R., CHANG, J.-F., and HUA, L., "Data mining in healthcare and biomedicine: a survey of the literature," Journal of Medical Systems, vol. 36, no. 4, pp. 2431–2448, 2012.

[121] ZHANG, T. and OLES, F. J., "Text categorization based on regularized linear classification methods," Information Retrieval, vol. 4, no. 1, pp. 5–31, 2001.

[122] ZHONG, J., NING, R., and CONOVER, D., "Image denoising based on multiscale singularity detection for cone beam CT breast imaging," Medical Imaging, IEEE Transactions on, vol. 23, no. 6, pp. 696–703, 2004.

[123] ZHU, S., SHI, Q., and CANES, A., "Detecting Medicaid data anomalies using data mining techniques," Report SDA-04, AdvanceMed Corporation, 2010.