Clustering Longitudinal Clinical Marker Trajectories from Electronic Health Data: Applications to Phenotyping and Endotype Discovery

Peter Schulam
Department of Computer Science
Johns Hopkins University
3400 N. Charles St.
Baltimore, MD 21218
pschulam@jhu.edu

Fredrick Wigley
Division of Rheumatology
Johns Hopkins School of Medicine
733. N. Broadway
Baltimore, MD 21205
fwig@jhmi.edu

Suchi Saria
Department of Computer Science
Department of Health Policy & Mgmt.
Johns Hopkins University
3400 N. Charles St.
Baltimore, MD 21218
ssaria@cs.jhu.edu

Abstract

Diseases such as autism, cardiovascular disease, and the autoimmune disorders are difficult to treat because of the remarkable degree of variation among affected individuals. Subtyping research seeks to refine the definition of such complex, multi-organ diseases by identifying homogeneous patient subgroups. In this paper, we propose the Probabilistic Subtyping Model (PSM) to identify subgroups based on clustering individual clinical severity markers. This task is challenging due to the presence of nuisance variability—variations in measurements that are not due to disease subtype—which, if not accounted for, generate biased estimates for the group-level trajectories. Measurement sparsity and irregular sampling patterns pose additional challenges in clustering such data. PSM uses a hierarchical model to account for these different sources of variability. Our experiments demonstrate that by accounting for nuisance variability, PSM is able to more accurately model the marker data. We also discuss novel subtypes discovered using PSM and the resulting clinical hypotheses that are now the subject of follow up clinical experiments.

Introduction and Background

Disease subtyping is the process of developing criteria for stratifying a population of individuals with a shared disease into subgroups that exhibit similar traits; a task that is analogous to clustering in machine learning. Under the assumption that individuals with similar traits share an underlying disease mechanism, disease subtyping can help to propose candidate subgroups of individuals that should be investigated for biological differences. Uncovering such differences can shed light on the mechanisms specific to each group. Observable traits useful for identifying subpopulations of similar patients are called *phenotypes*. When such traits have been linked to a distinct underlying pathobiological mechanism, these are then referred to as *endotypes* (Anderson 2008).

Traditionally, disease subtyping research has been conducted as a by-product of clinical experience. A clinician may notice the presence of subgroups, and may perform a more thorough retrospective or prospective study to confirm their existence (e.g. Barr et al. 1999). Recently, however, literature in the medical community has noted the need for more objective methods for discovering subtypes (De Keulenaer and Brutsaert 2009). Growing repositories of health data stored in electronic health record (EHR) databases and patient registries (Blumenthal 2009; Shea and Hripcsak 2010) present an exciting opportunity to identify disease subtypes in an objective, data-driven manner using tools from machine learning that can help to tackle the problem of combing through these massive databases. In this work, we propose such a tool, the Probabilistic Subtyping Model (PSM), that is designed to discover subtypes of complex, systemic diseases using longitudinal clinical markers collected in EHR databases and patient registries.

Discovering and refining disease subtypes can benefit both the practice and the science of medicine. Clinically, disease subtypes can help to reduce uncertainty in expected outcome of an individual's case, thereby improving treatment. Subtypes can inform therapies and aid in making prognoses and forecasts about expected costs of care (Chang, Clark, and Weiner 2011). Scientifically, disease subtypes can help to improve the effectiveness of clinical trials (Gundlapalli et al. 2008), drive the design of new genome-wide association studies (Kho et al. 2011; Kohane 2011), and allow medical scientists to view related diseases through a more fine-grained lens that can lead to insights that connect their causes and developmental pathways (Hoshida et al. 2007). Disease subtyping is considered to be especially useful for complex, systemic diseases where mechanism is often poorly understood. Examples of disease subtyping research include work in autism (State and Sestan 2012), cardiovascular disease (De Keulenaer and Brutsaert 2009), and Parkinson's disease (Lewis et al. 2005).

Complex, systemic diseases are characterized using the level of disease activity present in an array of organ systems. Clinicians typically measure the influence of a disease on an organ using clinical tests that quantify the extent to which that organ's function has been affected by the disease. The results of these tests, which we refer to as *illness severity markers* (s-markers for short), are being routinely collected over the course of care for large numbers of patients within EHR databases and patient registries. For a single individual, the time series formed by the sequence of these s-markers can be interpreted as a *disease activity trajectory*. Operating under the hypothesis that individuals with similar disease activity trajectories are more likely to

share mechanism, our goal in this work is to cluster individuals according to their longitudinal clinical marker data and learn the associated *prototypical disease activity trajectories* (i.e. a continuous-time curve characterizing the expected s-marker values over time) for each subtype.

The s-marker trajectories recorded in EHR databases are influenced by factors such as age and co-existing conditions that are unrelated to the underlying disease mechanism (Lötvall et al. 2011). We call the effects of these additional factors *nuisance variability*. In order to correctly cluster individuals and uncover disease subtypes that are likely candidates for endotyping (Lötvall et al. 2011), it is important to model and explain away nuisance variability. In this work, we account for nuisance variability in the following ways. First, we use a population-level regression on to observed covariates—such as demographic characteristics or co-existing conditions—to account for variability in s-marker values across individuals. For example, lung function as measured by the forced expiratory volume (FEV) test is well-known to be worse in smokers than in non-smokers (Camilli et al. 1987). Second, we use individual-specific parameters to account for variability across individuals that is not predicted using the observed covariates. This form of variability may last throughout the course of an individual's disease (e.g. the individual may have an unusually weak respiratory system) or may be episodic (e.g. periods during which an individual is recovering from a cold). Finally, the subtypes' prototypical disease activity trajectories must be inferred from measurement sequences that vary widely between individuals in the time-stamp of the first measurement, the duration between consecutive measurements, and the time-stamp of the last measurement. After accounting for nuisance variability, our goal is to cluster the time series formed by the residual activity. We hypothesize that differences across such clusters are more likely candidates for endotype investigations.

A large body of work has focused on identifying subtypes using genetic data (e.g. Chen et al. 2011). Enabled by the increasing availability of EHRs, researchers have also recently started to leverage clinical markers to conduct subtype investigations. Chen et al. 2007 use the clinarray—a vector containing summary statistics of all available clinical markers for an individual—to discover distinct phenotypes among patients with similar diseases. The clinarray summarizes longitudinal clinical markers using a single statistic, which ignores the pattern of progression over time. More recently, Ho et al. have used tensor factorization as an alternative approach to summarizing high-dimensional vectors created from EHRs (Ho, Ghosh, and Sun 2014). Others have used cross-sectional data to piece together disease progressions (e.g. Ross and Dy 2013), but do not model longitudinal data. Another approach to phenotyping using time series data, often applied in the acute care setting, is to segment an individual's time series into windows in order to discover transient traits that are expressed over shorter durations—minutes, hours, or days. For example, Saria et al. propose a probabilistic framework for discovering traits from physiologic time series that have similar shape characteristics (Saria, Duchi, and Koller 2011) or dynamics characteristics (Saria, Koller, and Penn 2010). Lasko et al. use deep learning to induce an over-complete dictionary in order to define traits observed in shorter segments of clinical markers (Lasko, Denny, and Levy 2013).

Beyond s-markers, others have used ICD-9 codes—codes indicating the presence or absence of a condition—to study comorbidity patterns over time among patients with a shared disease (e.g. Doshi-Velez, Ge, and Kohane 2014). ICD-9 codes are further removed from the biological processes measured by quantitative tests. Moreover, the notion of disease severity is more difficult to infer from codes.

Latent class mixed models (LCMMs) are a family of methods designed to discover subgroup structure in longitudinal datasets using fixed and random effects (e.g., Muthén and Shedden 1999, McCulloch et al. 2002, and Nagin and Odgers 2010). Random effects are typically used in linear models where an individual's coefficients may be probabilistically perturbed from the group's, which alters the model's fit to the individual over the entire observation period. Modeling s-marker data for chronic diseases where data are collected over tens of years requires accounting for additional influences such as those due to transient disease activity. The task of modeling variability between related time series has been explored in other contexts (e.g. Listgarten et al. 2006 and Fox et al. 2011). These typically assume regularly sampled time series and model properties that are different from those in our application.

Work in the machine learning literature has also looked at relaxing the assumption of regularly sampled data. Marlin et al. cluster irregular clinical time series from in-hospital patients to improve mortality prediction using Gaussian process priors that allow unobserved measurements to be marginalized (Marlin et al. 2012). Lasko et al. also address irregular measurements by using MAP estimates of Gaussian processes to impute sparse time series (Lasko, Denny, and Levy 2013).

In this paper, we propose the Probabilistic Subtyping Model (PSM), a novel model for discovering disease subtypes and associated prototypical disease activity trajectories using observational data that is routinely collected in electronic health records (EHRs). In particular, our model is geared towards identifying homogeneous patient subgroups from s-marker data for chronic diseases, and is particularly useful for characterizing complex, systemic diseases that are often poorly understood. Towards this end, PSM uncovers subtypes and their prototypical disease activity trajectories, while explaining away variability unrelated to disease subtyping; namely (1) covariate-dependent nuisance variability, (2) individual-specific long-term nuisance variability, and (3) individual-specific short-term nuisance variability. In addition, PSM is able to infer these prototypical trajectories using clinical time series that vary with respect to their endpoints and sampling patterns, which is common in observational EHR data. To evaluate PSM, we use real and simulated data to demonstrate that, by accounting for these levels of nuisance variability, PSM is able to both accurately predict s-markers and accurately recover prototypical disease activity trajectories. Finally, we discuss novel subtypes discovered using PSM.

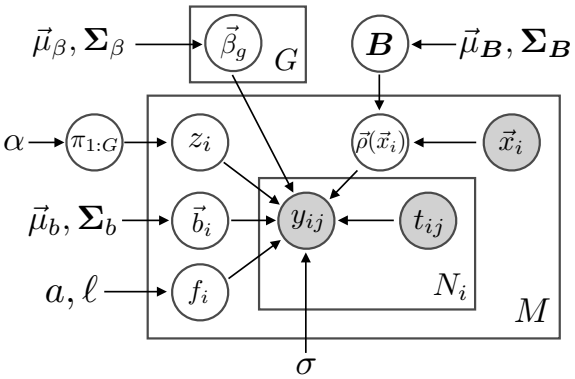

Figure 1: Graphical model for PSM.

Probabilistic Subtyping Model

We define a generative model for a collection of M individuals with associated s-marker sequences. For each individual i, the s-marker sequence has N_i measurement times and values, which are denoted as $\vec{t}_i \in \mathbb{R}^{N_i}$ and $\vec{y}_i \in \mathbb{R}^{N_i}$ respectively. In addition to the measurement times and values, each individual is assumed to have a vector of d covariates $\vec{x}_i \in \mathbb{R}^d$. The s-marker values $\vec{y}_{1:M}$ are random variables, and we assume that $\vec{t}_{1:M}$ and $\vec{x}_{1:M}$ are fixed and known. The major conceptual pieces of the model are the subtype mixture model, covariate-dependent nuisance variability, individual-specific long-term nuisance variability, and individual-specific short-term nuisance variability. We describe each of these pieces in turn. The graphical model in Figure 1 shows the relevant hyperparameters, random variables, and dependencies.

Subtype mixture model. Each individual is assumed to belong to one of G latent groups (representing a disease subtype). The random variable $z_i \in \{1, \ldots, G\}$ encodes subtype membership, and is drawn from a multinomial distribution with probability vector $\pi_{1:G}$. The probabilities $\pi_{1:G}$ are modeled as a Dirichlet random variable with symmetric concentration parameter α. Formally, we have:

$$p\left(\pi_{1:G}\right) = \text{Dir}\left(\pi_{1:G}; \alpha\right) \tag{1}$$

$$p\left(z_i | \pi_{1:G}\right) = \text{Mult}\left(z_i; \pi_{1:G}\right). \tag{2}$$

We model each of the subtype prototypical disease activity trajectories using B-splines. The P B-spline basis function values at times $\vec{t}$ are arranged into columns of a feature matrix $\boldsymbol{\phi}\left(\vec{t}\right) = \left[\phi_1\left(\vec{t}\right), \ldots, \phi_P\left(\vec{t}\right)\right]$, and the prototypical disease activity trajectory for each subtype g is parameterized by a coefficient vector $\vec{\beta}_g$. The coefficients $\vec{\beta}_g$ are modeled as vectors drawn from a multivariate normal distribution:

$$p\left(\vec{\beta}_g\right) = \mathcal{N}\left(\vec{\beta}_g; \vec{\mu}_\beta, \mathbf{\Sigma}_\beta\right). \tag{3}$$

Covariate-dependent nuisance variability. When one or more covariates are available that are known to influence clinical test results, but are posited to be unrelated to the task of discovering underlying disease mechanisms, the influences can be accounted for using covariate-dependent effects. For example, smoking status can partially explain why an individual's forced expiratory volume declines more rapidly, or African American race may be associated with especially severe scleroderma-related skin fibrosis. By fitting a standard mixture of B-splines to s-marker data directly, the subtype random variable z_i may incorrectly capture correlations among groups of individuals with similar covariates. By including covariate-dependent effects, these correlations are removed, thereby freeing the subtype mixture model to capture *residual correlations* that are more likely to be due to shared underlying disease mechanism. Covariate-dependent effects are modeled using a polynomial function $\gamma\left(\vec{t}\right)\vec{\rho}(\vec{x})$ with feature matrix $\gamma\left(\vec{t}\right)$ and coefficients $\vec{\rho}(\vec{x})$. We use first order polynomials (lines) with feature matrix for times $\vec{t}$ defined to be $\gamma\left(\vec{t}\right) = \left[\vec{1}, \vec{t} - \bar{t}\right]$, where $\bar{t}$ is the midpoint of the period over which the subtype trajectories are being modeled. The coefficients are modeled as a linear function of the individual's covariates:

$$\vec{\rho}(\vec{x}) = \boldsymbol{B}\vec{x}, \tag{4}$$

where $\boldsymbol{B} \in \mathbb{R}^{2,d}$ is a loading matrix linearly linking the covariates to the intercept and slope values. Therefore, conditioned on the loading matrix $\boldsymbol{B}$, individuals with similar covariates will have similar coefficients $\vec{\rho}(\vec{x})$. We model the rows of the loading matrix $\boldsymbol{B}_{c,\cdot} : c \in \{1, 2\}$ as multivariate normal random variables:

$$p\left(\boldsymbol{B}_{c,\cdot}^\top\right) = \mathcal{N}\left(\boldsymbol{B}_{c,\cdot}^\top; \vec{\mu}_B, \mathbf{\Sigma}_B\right). \tag{5}$$

Individual-specific long-term nuisance variability. An individual may express additional variability over the entire observation period beyond what is explained away using covariates. For example, an individual may have an unusually weak respiratory system and so may have a lower baseline value (intercept) than other individuals with similar covariates. This form of variability is also modeled using first order polynomials with feature matrix $\gamma(\vec{t})$. The coefficient vector is a multivariate normal random variable:

$$p\left(\vec{b}_i | \boldsymbol{B}, \vec{x}_i\right) = \mathcal{N}\left(\vec{b}_i; 0, \mathbf{\Sigma}_b\right). \tag{6}$$

Individual-specific short-term nuisance variability. Finally, an individual may experience episodic disease activity that only affects a handful of measurements within a small time window. For example, there may be periods during which an individual is recovering from a cold, which temporarily weakens his or her respiratory system and causes FEV to become depressed over a short period of time. We model these transient deviations nonparametrically using a Gaussian process with mean 0 and a squared exponential covariance function with hyperparameters a and ℓ:

$$p\left(f_i\right) = \text{GP}\left(f_i; 0, k(\cdot, \cdot)\right) \tag{7}$$

$$k(t_1, t_2) = a^2 \exp\left\{-\frac{(t_1 - t_2)^2}{2\ell^2}\right\}. \tag{8}$$

When treatments exist that can alter long-term course, and the points at which they are administered vary widely across

individuals, treatments become additional sources of nuisance variability. In scleroderma, our disease of interest, and many other systemic diseases, no drugs known to modify the long-term course of the disease exist, and, therefore, we do not tackle this issue. Moreover, treatments are challenging to fully account for, and subtyping studies that do so often focus on data from practices that follow a similar treatment protocol so that variability due to treatment is controlled.

Collapsed likelihood. Conditioned on these four components, an individual's s-marker sequence is modeled as a spherical multivariate normal with standard deviation σ:

$$p\left(\vec{y}_i|z_i,\vec{b}_i,f_i,\vec{\beta}_{1:G},\vec{t}_i\right) \tag{9}$$
$$= \mathcal{N}\left(\vec{y}_i;\boldsymbol{\phi}\left(\vec{t}_i\right)\vec{\beta}_{z_i}+\boldsymbol{\gamma}\left(\vec{t}_i\right)\left(\boldsymbol{B}\vec{x}_i+\vec{b}_i\right)+f_i\left(\vec{t}_i\right),\sigma^2\boldsymbol{I}\right).$$

Assuming independence between all individuals conditioned on the model parameters $\vec{\beta}_{1:G}$, $\pi_{1:G}$, and $\boldsymbol{B}$, the complete likelihood is:

$$\prod_{i=1}^{M} p\left(z_i|\pi_{1:G}\right)p\left(\vec{b}_i\right)p\left(f_i\right)p\left(\vec{y}_i|z_i,\vec{b}_i,f_i,\vec{\beta}_{1:G},\vec{t}_i\right).$$

By marginalizing over the latent variables z_i, $\vec{b}_i$, and f_i, the likelihood for each individual can be written as a mixture of multivariate normals:

$$p\left(\vec{y}_i|\vec{\beta}_{1:G},\pi_{1:G},\boldsymbol{B},\vec{t}_i,\vec{x}_i\right)=\mathbb{E}_{z_i}\left[\mathcal{N}\left(\vec{\mu}_i^{(z_i)},\boldsymbol{\Sigma}_i\right)\right],$$

where

$$\vec{\mu}_i^{(z_i)}=\boldsymbol{\phi}\left(\vec{t}_i\right)\vec{\beta}_{z_i}+\boldsymbol{\gamma}\left(\vec{t}_i\right)\boldsymbol{B}\vec{x}_i$$
$$\boldsymbol{\Sigma}_i=\sigma^2\boldsymbol{I}+k(\vec{t}_i,\vec{t}_i)+\boldsymbol{\gamma}\left(\vec{t}_i\right)\boldsymbol{\Sigma}_b\boldsymbol{\gamma}\left(\vec{t}_i\right)^\top.$$

The final joint distribution becomes:

$$\underbrace{\prod_{i=1}^{M} p\left(\vec{y}_i|\vec{\beta}_{1:G},\pi_{1:G},\boldsymbol{B},\vec{t}_i,\vec{x}_i\right)}_{\text{likelihood}}\underbrace{p\left(\pi_{1:G}\right)\prod_{g=1}^{G}p\left(\vec{\beta}_g\right)p\left(\boldsymbol{B}\right)}_{\text{prior}}. \tag{10}$$

Learning

We want to learn the parameters $\boldsymbol{\Theta}=\left\{\vec{\beta}_{1:G},\pi_{1:G},\boldsymbol{B}\right\}$ and the hyperparameters $\{a,\ell,\sigma\}$. Conditioned on choices for the hyperparameters, we will compute MAP estimates of $\boldsymbol{\Theta}$ by optimizing the log of the joint distribution in Equation 10 with respect to $\boldsymbol{\Theta}$. The likelihood is a product of sums, which makes direct optimization difficult. We will instead use the EM algorithm to maximize the sum of the log prior and the expected complete-data log likelihood. To choose $\{a,\ell,\sigma\}$, we use a grid search, often shown to work well in low dimensions (e.g. Bardenet, Kégl, and others 2010).

Expectation Step Conditioned on estimates of $\vec{\beta}_{1:G}$, $\pi_{1:G}$, and $\boldsymbol{B}$ at time τ, the posterior distribution over z_i for each individual is:

$$q_i\left(z_i\right)=p\left(z_i|\vec{y}_i,\vec{\beta}_{1:G},\pi_{1:G},\boldsymbol{B},\vec{t}_i,\vec{x}_i\right)$$
$$=\frac{1}{Z_i}p\left(z_i\right)\mathcal{N}\left(\vec{\mu}_i^{(z_i)},\boldsymbol{\Sigma}_i\right), \tag{11}$$

where Z_i is the normalizing constant of the multinomial.

Maximization Step Using the posterior distribution computed with parameters $\boldsymbol{\Theta}^\tau$, the expected complete-data log likelihood for a single individual as a function of the parameters at step $\tau+1$ is:

$$L_i(\boldsymbol{\Theta}^{\tau+1}|\boldsymbol{\Theta}^\tau)= \tag{12}$$
$$\mathbb{E}_{q_i}\left[\log p\left(z_i|\pi_{1:G}^{\tau+1}\right)\right]+$$
$$\mathbb{E}_{q_i}\left[\log p\left(\vec{y}_i|z_i,\vec{\beta}_{1:G}^{\tau+1},\boldsymbol{B}^{\tau+1},\vec{t}_i,\vec{x}_i\right)\right].$$

The expected complete-data log likelihood and log priors on the parameters then form the objective of the maximization step:

$$Q(\boldsymbol{\Theta}^{\tau+1}|\boldsymbol{\Theta}^\tau)=\sum_{i=1}^{M}L_i(\boldsymbol{\Theta}^{\tau+1}|\boldsymbol{\Theta}^\tau)+\ldots \tag{13}$$
$$\log p\left(\pi_{1:G}^{\tau+1}\right)+\sum_{g=1}^{G}\log p\left(\vec{\beta}_g^{\tau+1}\right)+\log p\left(\boldsymbol{B}^{\tau+1}\right),$$

which we optimize with respect to $\boldsymbol{\Theta}^{\tau+1}$ to obtain the next set of parameter estimates. We begin by maximizing with respect to $\pi_{1:G}$. Dropping terms that do not include $\pi_{1:G}$ from Equation 13, we find that the objective is maximized using the standard Dirichlet-multinomial posterior estimates. We use a block coordinate ascent algorithm to maximize the objective with respect to $\beta_{1:G}$ and $\boldsymbol{B}$. We break the parameters into $G+2$ blocks; one block for each of the subtype-specific coefficients β_g and one block for each row of $\boldsymbol{B}$. Each block appears in a quadratic term in the likelihood of each patient, and so it is easy to see that holding all other blocks fixed, the coordinate ascent step can be maximized exactly. We cycle through the block updates until the value of Equation 13 converges, and iterate over the EM updates until the joint shown in Equation 10 converges. The updates for each of the parameters $\pi_{1:G}$, $\vec{\beta}_{1:G}$, and $\boldsymbol{B}$ are listed in Section A1 of the supplementary material.[1]

Scalability PSM is designed to aid in the subtype discovery process when large electronic health databases are available for analysis. Thus, the scalability of the learning algorithm is a natural concern. The primary computational bottleneck of the PSM learning procedure is the E-step, which may be expensive due to (1) the number of individuals in the analysis, or (2) the inversion of the individual covariance

[1] www.cs.jhu.edu/~ssaria/psm_supp.pdf.

matrices Σ_i (e.g., in Equation 11). The computational complexity due to large M can be offset by parallelizing the E-step because the individual-specific latent variables are conditionally independent given Θ. Inversion of Σ_i has computational complexity $\mathcal{O}(N_i^3)$. Because we study disease activity over the course of 10-20 years and because visit rates typically do not exceed 12 per year, the number of measurements N_i is typically on the order of 100-200 measurements. Inversion of Σ_i is therefore inexpensive.

Inference

To obtain a posterior prediction of an individual's disease activity trajectory, a subtype is first chosen using the most likely value of z_i under the distribution in Equation 11. Conditioned on subtype membership, we estimate $\vec{b}_i$ and f_i by first conditioning on $f_i = 0$, and computing the MAP estimate of $\vec{b}_i$. From Figure 1, it is clear that after conditioning on z_i, f_i and $\mathbf{\Theta}$, the joint distribution over $\vec{b}_i$ and $\vec{y}_i$ is a linear Gaussian system. We can therefore use standard equations to obtain the MAP estimate of $\vec{b}_i$:

$$\vec{b}_i = \mathbf{\Sigma}_{\vec{b}_i} \left(\sigma^{-2} \boldsymbol{\gamma} \left(\vec{t}_i \right)^\top \vec{y}_i^{(\vec{b}_i | z_i)} + \mathbf{\Sigma}_b^{-1} \vec{\mu}_b \right), \text{ where}$$

$$\mathbf{\Sigma}_{\vec{b}_i} = \left(\sigma^{-1} \boldsymbol{\gamma} \left(\vec{t}_i \right)^\top \boldsymbol{\gamma} \left(\vec{t}_i \right) + \mathbf{\Sigma}_b^{-1} \right)^{-1}$$

$$\vec{y}_i^{(\vec{b}_i | z_i)} = \vec{y}_i - \boldsymbol{\phi} \left(\vec{t}_i \right) \vec{\beta}_{z_i} - \boldsymbol{\gamma} \left(\vec{t}_i \right) \boldsymbol{B} \vec{x}_i.$$

We choose to temporarily condition on $f_i = 0$ in order to use the long-term variability to explain as much of the individual variation as possible. This encodes our preference for a parsimonious posterior prediction that uses the simpler long-term variability before relying on the more complex Gaussian process.

Finally, conditioned on z_i and $\vec{b}_i$, the individual's short term variation is defined at time t^* to be

$$f_i(t^*) = k(t^*, \vec{t}_i) \left(k(\vec{t}_i, \vec{t}_i) + \sigma^2 \boldsymbol{I} \right)^{-1} \vec{y}_i^{(f_i | \vec{b}_i, z_i)}, \text{ where}$$

$$\vec{y}_i^{(f_i | \vec{b}_i, z_i)} = \vec{y}_i - \boldsymbol{\phi} \left(\vec{t}_i \right) \vec{\beta}_{z_i} - \boldsymbol{\gamma} \left(\vec{t}_i \right) \boldsymbol{B} \vec{x}_i - \boldsymbol{\gamma} \left(\vec{t}_i \right) \vec{b}_i.$$

This is simply the MAP prediction of a Gaussian process. A new measurement y^* taken at time t^* is then predicted to be:

$$y^* = \underbrace{\boldsymbol{\phi} \left(t^* \right) \vec{\beta}_{z_i}}_{\text{subtype}} + \underbrace{\boldsymbol{\gamma} \left(t^* \right) \boldsymbol{B} \vec{x}_i}_{\text{covariate}} + \underbrace{\boldsymbol{\gamma} \left(t^* \right) \vec{b}_i}_{\text{long-term}} + \underbrace{f_i(t^*)}_{\text{short-term}}. \quad (14)$$

It is also useful to characterize posterior uncertainty in prototypical trajectory estimates when drawing qualitative inferences from the model. Estimates of posterior uncertainty are described in Section A2 of the supplement.

Experiments

The purpose of PSM is to discover subtypes, useful for tasks such as developing tailored treatment plans and advancing understanding of underlying disease mechanisms. Exploratory clustering method evaluations are two-fold. From a quantitative perspective, we want to measure the model's fit to data. From a qualitative standpoint, we want to judge the insights that the model conveys. Consequently, we assess the merit of PSM using two experiments. First, we investigate PSM's ability to predict unobserved s-marker measurements. In the absence of ground-truth subtypes that we can use to compute a cluster-based metric, we instead measure the generalizability of PSM by evaluating posterior predictions of held-out s-marker observations.[2] Our second experiment involves qualitative analyses of the subtypes discovered by PSM. We evaluate the clinical merit of the prototypical disease activity trajectories discovered, and discuss follow-up clinical investigations that have stemmed from our results.

Scleroderma S-markers

Scleroderma is a multi-system autoimmune disease resulting in insults that include damage to the skin, pulmonary system, and circulatory system. For our experiments we choose four datasets, each containing the s-marker corresponding to one of the major organ systems implicated in scleroderma. *Total skin score* (TSS) scores the thickness of the skin, which is used as a surrogate measure for the degree of fibrosis. An increased TSS indicates more extensive fibrosis across the individual's body. *Percent of predicted forced vital capacity* (pFVC) measures the volume of air expelled from the lung after maximum inhalation. pFVC is a measure of restricted ventilatory defect, which may reflect the severity of interstitial lung disease (ILD). *Percent of predicted diffusing capacity* (pDLCO) measures the efficiency of oxygen diffusion from the lungs to the bloodstream. Decreases in pDLCO indicate a defect in gas exchange, which is associated with development of pulmonary arterial hypertension (PAH). Finally, *right ventricular systolic pressure* (RVSP) measures systolic pressure in the chamber of the heart that directly pumps blood through the pulmonary vasculature. Significantly increased systolic pressure suggests an increased risk of heart failure due to pulmonary arterial hypertension. We focus here on the analysis of subtypes for each of the complications individually. If subtypes exist, then a natural follow-up is to identify whether a common mechanism might jointly influence trajectories for two or more organ systems, but this relies on first developing an understanding of the individual s-markers; the focus of our analyses below.

Unobserved S-marker Prediction

Our first experiment evaluates the accuracy of s-marker value predictions at unobserved times for models that incorporate different types of nuisance variability. The first model includes covariate effects and group/subtype effects (C+G). The covariates provided to us were gender, African American race, and age at disease onset, which are well-known risk factors for severity in scleroderma (Varga, Denton, and Wigley 2012). The second model uses individual-

[2] Alternatively, one may use held-out data log-likelihood, but we choose predictive accuracy because the task of prediction is natural in the clinical setting and it is therefore easier to interpret the significance of the results.

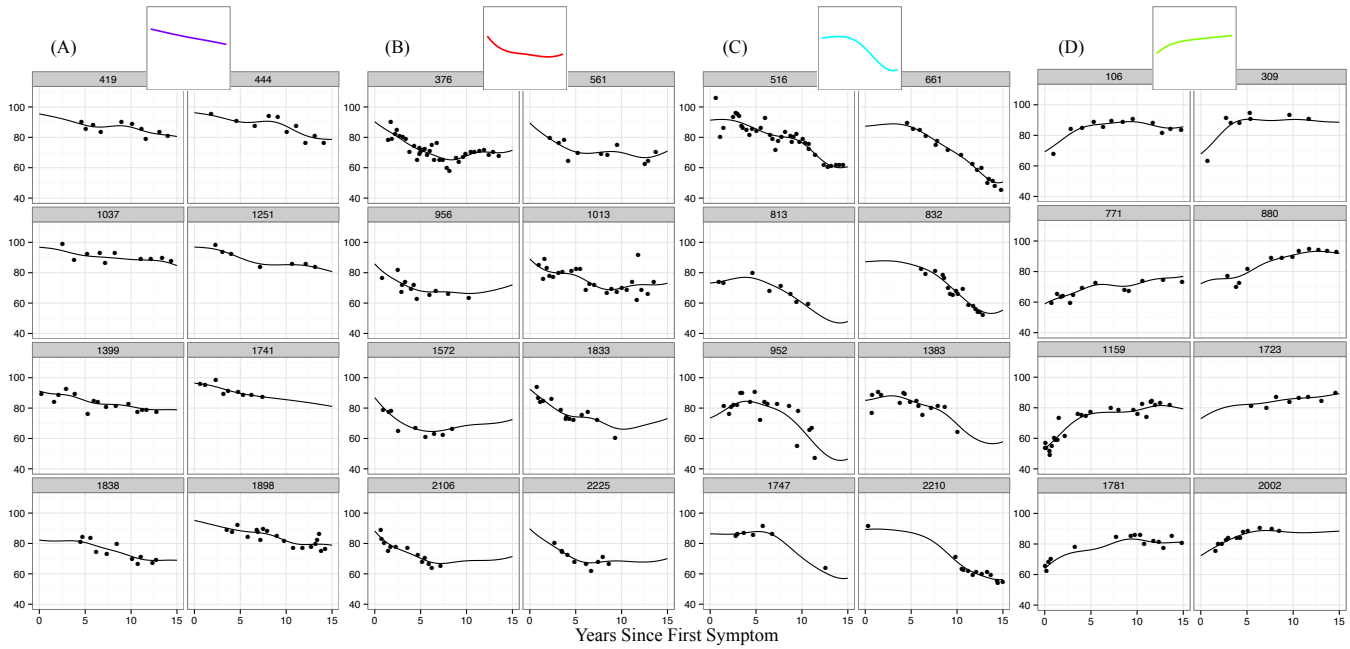

Figure 2: Example model fits to individual pFVC trajectories. Samples from four subtype candidates are displayed; one subtype per 4 by 2 block of individuals. Solid lines show full model fit (computed using Equation 14). Dots show observed pFVC values. Solid lines at the top of each block show the prototypical s-marker trajectory for that subtype.

specific long-term effects in addition to covariate and group effects (C+G+L). Finally, PSM uses covariate, group, long-term, and short-term effects.

To choose the number of groups G for each model, we use BIC as follows: we randomly generate five folds of the data by subsampling 75% of the individuals without replacement. For each model and for each choice of G, we compute the average BIC across the five folds and choose the number of clusters that results in the largest sequential drop in BIC (i.e. we search for the "elbow").

For each model, we use $P = 5$ bases and non-informative priors for the parameters; $\alpha = 2$, $\vec{\mu}_\beta = \vec{\mu}_B = 0$, and $\mathbf{\Sigma}_\beta = \mathbf{\Sigma}_B = \text{diag}(1 \times 10^5)$. The prior variance on $\vec{b}_i$ was kept relatively small to limit the amount of variability that could be explained using long-term individual effects: $\mathbf{\Sigma}_b = \text{diag}(v_b)$ where v_b was swept along the log-10 scale from $\{-1, \ldots, -5\}$. Similarly, to limit the amount of variability explained by the Gaussian process, the GP hyperparameters $\{a, \ell, \sigma\}$ were each swept from $\{1, \ldots, 5\}$. Years since first scleroderma related symptom was used to align each individual on the time axis. We truncate time at 15 years following the first scleroderma related symptom. An individual was included in the experiment if there were at least 4 measurements available. We used 1,011, 1,177, 1,114, and 504 individuals for TSS, pFVC, pDLCO, and RVSP respectively. Within this subset of individuals, the average number of measurements for each s-marker over the 15 year period is: 9.1 for TSS, 8.6 for pFVC, 8.3 for pDLCO, and 6.0 for RVSP. The average time in years between observed measurements for each s-marker is: 0.9 for TSS, 0.9 for pFVC, 1.0 for pDLCO, and 1.3 for RVSP.

To estimate prediction error, we split the individual trajectories into 10 groups and use 10-fold cross validation. PSM is trained on nine of the ten folds, and predictions are made for the tenth fold as follows. We select four s-marker observations from each held-out trajectory and assign each to a point-level fold. For each of the point-level folds, we condition on the remaining s-marker observations and use the MAP estimates of the held-out observations as our predictions. We compute root mean squared error (RMSE) for each point-level fold across the trajectory-level folds, and finally compute the mean RMSE and standard errors across point-level folds.

The prediction results and standard errors are displayed in Table 1. We see that PSM significantly outperforms the alternative models for TSS, pFVC, and RVSP. Although an improvement is demonstrated for pDLCO, the difference is not statistically significant. Figure 2 displays model fits to individual trajectories sampled from four of the discovered candidate subtypes (one subtype per 4 by 2 block of individuals) for the pFVC s-marker. Note that blocks display overall similar behavior, but that long-term and short-term variability tailor predictions for each individual using the observed markers.

S-marker	C+G	C+G+L	PSM
TSS	5.32 ± 0.18	5.41 ± 0.07	$^*\mathbf{4.43} \pm 0.14$
pFVC	9.27 ± 0.49	9.34 ± 0.46	$^*\mathbf{7.69} \pm 0.39$
pDLCO	15.03 ± 1.82	15.13 ± 1.93	$\mathbf{14.08} \pm 1.77$
RVSP	12.21 ± 0.50	12.11 ± 0.44	$^*\mathbf{10.89} \pm 0.27$

Table 1: RMSE with standard errors for s-marker prediction. Bold shows best performance on s-marker; * shows statistical significance ($p \leq 0.05$).

Simulated Data Trajectory Estimate Accuracy Another natural question is whether modeling these sources of vari-

ability reduces bias in our estimates of the prototypical trajectories. In other words, do the individual deviations cancel out so that PSM offers no benefit over less expressive models like C+G? For this, we turn to simulated data and investigate whether PSM recovers prototypical trajectories more accurately than C+G and C+G+L.

The simulation model samples observation time-stamps by sampling the N_i from a Poisson distribution, and, conditioned on N_i, samples $\vec{t}_i$ from a Gaussian mixture model. The $\vec{y}_i$ are then sampled from a subtype mixture model with a hierarchy of individual-specific long-term, short-term, and iid noise as employed within PSM. The parameters used for the simulation are specified in Section A3 of the supplementary material. We compare the same three models used above, but do not simulate covariates and so they are not included. To measure bias, we find the alignment between estimated and true trajectories that minimizes the RMSE averaged across each estimated-true pair; to compute RMSE between two curves, we use a discrete approximation.

The iid noise $\sigma = 0.1$ for all simulations, and the left column of Table 2 shows the amplitude a of short-term individual variability and standard deviation of individual-specific intercept terms σ_b. Individual specific slope variance was set to 1×10^{-4} for all simulations.

We see that as more nuisance variability is added, PSM is able to recover less biased trajectory estimates. In Section A4 of the supplementary material, we provide plots that show example individual trajectories from these experiments and how they contribute to the bias of prototypical trajectory estimates.

(a, σ_b)	G	G+L	PSM
$(0.00, 0.10)$	0.27 ± 0.06	$^{*}\mathbf{0.04} \pm 0.01$	0.07 ± 0.06
$(0.00, 0.15)$	0.34 ± 0.05	$^{*}\mathbf{0.05} \pm 0.01$	0.09 ± 0.06
$(0.10, 0.15)$	0.34 ± 0.04	$\mathbf{0.11} \pm 0.05$	0.14 ± 0.08
$(0.15, 0.15)$	0.34 ± 0.05	0.18 ± 0.04	$^{*}\mathbf{0.13} \pm 0.06$
$(0.20, 0.15)$	0.36 ± 0.05	0.25 ± 0.07	$^{*}\mathbf{0.14} \pm 0.07$
$(0.25, 0.15)$	0.36 ± 0.06	0.32 ± 0.07	$^{*}\mathbf{0.18} \pm 0.04$

Table 2: Estimated trajectory RMSE and standard errors (computed over 20 replications) for simulations. Bold indicates best performance, statistical significance is indicated using * ($p \leq 0.05$).

Discovered Subtypes

We now present a qualitative discussion of the discovered subtypes. For all results below, we use non-informative priors for the parameters; $\alpha = 2$, $\vec{\mu}_\beta = \vec{\mu}_B = 0$, and $\mathbf{\Sigma}_\beta = \mathbf{\Sigma}_B = \text{diag}(1 \times 10^5)$. The prior variance on $\vec{b}_i$ was kept relatively small to limit the amount of variability that could be explained using long-term individual effects: $\mathbf{\Sigma}_b = \text{diag}(v_b)$ where v_b was swept along the log-10 scale from $\{-1, \ldots, -5\}$. Similarly, to limit the amount of variability explained by the Gaussian process, the GP hyperparameters $\{a, \ell, \sigma\}$ were each swept from $\{1, \ldots, 5\}$.

We begin with pFVC. Figure 3A displays the clusters learned using PSM on pFVC trajectories of 1,177 individuals with $G = 9$ (chosen using BIC), and Figure 2 displays individual trajectory fits from 4 of the 9 clusters. We first focus on the subtypes displayed in panels (A), (B), and (C) of Figure 2. Each of these show distinct patterns of decline: individuals in (A) have a steady, linear progression, those in (B) decline quickly within the first five years and then stabilize, and those in (C) are stable for the first five to ten years and then decline rapidly. Many of these individuals have at least one measurement that drops by more than 7% from the previous observation, which is clinically considered to suggest interstitial lung disease (see, for example, Beretta et al. 2007). It is clear, however, that they display unique patterns of decline, which has raised the question of whether individuals with these different subtypes differ with respect to their antibody profiles (since scleroderma is an autoimmune disease).

We now turn to panel (D). Fibrosis in the lungs due to end-stage interstitial lung disease is thought of as non-reversible damage. The individuals shown in panel (D), however, begin with inhibited pulmonary function, but slowly recover. This pattern of recovery warrants additional investigation; it is possible that in these patients, the initial insult is not due to end-stage lung disease, but rather due to other causes of restricted ventilatory defect such as inflammation. Recognizing this pattern may alter clinical management of these patients. Moreover, if it is the case that these patients all share a common comorbidity at the onset of disease, then it may suggest the presence of another subtype whose underlying mechanism triggers the comorbid condition.

Figure 3B displays the clusters learned by PSM for Total Skin Score (TSS) using $M = 1,011$ individuals with $G = 5$ (selected using BIC). TSS is a well-studied s-marker in scleroderma, and is used for one of the primary clinical classification criteria. Individuals with more extensive skin disease have higher TSS scores. Traditionally, skin disease in scleroderma is defined as limited (minimal involvement at the system level) or diffuse (systemic level involvement) (Varga, Denton, and Wigley 2012). Here we see five clusters: clusters 4 and 5 exhibit limited involvement, and clusters 1, 2, and 3 indicate different patterns of diffuse skin disease.

Figure 3C displays the clusters learned using PSM for pDLCO using $M = 1,114$ individuals with $G = 11$ (chosen using BIC). Clinicians monitor pDLCO to detect the onset of pulmonary arterial hypertension (PAH), one of the most prominent sources of mortality among patients with scleroderma. Once pDLCO is low enough, an individual will typically be screened using additional diagnostic tests (Steen and Medsger 2003). It is clear from Figure 3D, however, that there are several patterns of decline (seen in clusters 2, 8, 9, 10, and 11). Understanding whether particular patterns of pDLCO decline are more predictive of PAH may help to develop more effective clinical heuristics.

Finally, Figure 3D displays the clusters learned using PSM for RVSP using $M = 504$ individuals with $G = 5$ (chosen using BIC). The RVSP results are noisier than the others, which may be due to the inherent noise in the measurement process. RVSP is measured using an echocardiogram, and is inaccurate when the true underlying systolic pressure is between 30 and 45 mmHg (millimeters mercury). We note that there are two groups (1 and 4) with stable, healthy pressures, which are presumably individuals with no serious circulatory complications. The remaining three clus-

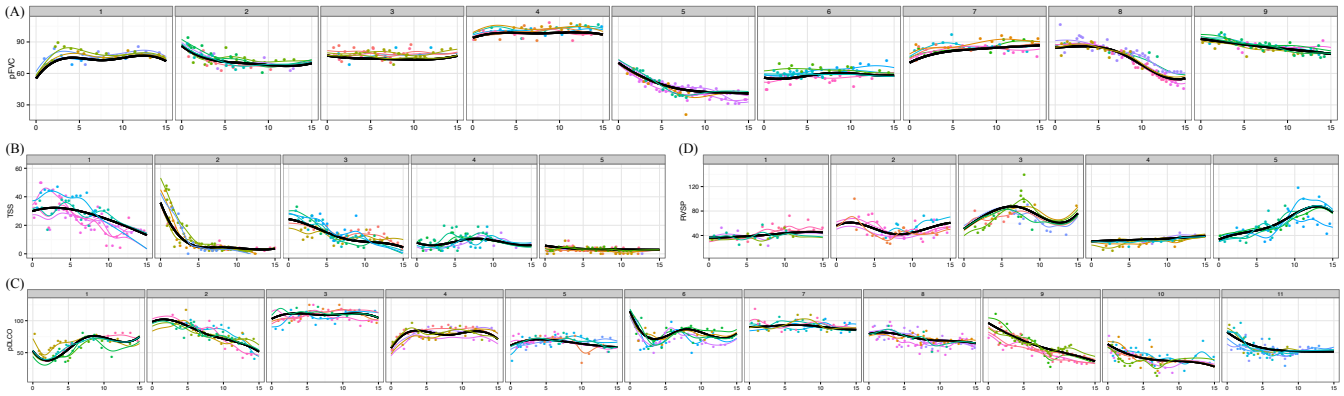

Figure 3: Discovered subtypes for all four s-markers. Panel (A) shows pFVC, panel (B) shows TSS, panel (C) shows pDLCO, and panel (D) shows RVSP. Prototypical s-marker trajectories are shown in black, and individuals sampled from the subtype are shown in color. Colored lines show the individualized s-marker trajectory, and colored points show the observed s-markers. Best viewed in color.

ters (2, 3, and 5) are more difficult to interpret because they are not associated with any known patterns of heart involvement; this may be attributable to the noisiness of the observations, or other phenomena that are as yet unknown. These remain the subject of future clinical follow up.

Joint Analysis of S-Markers We have focused our analyses using PSM on single s-marker subtypes. A natural follow-up question is whether we can infer clusters across multiple s-markers. If we are analyzing K s-marker types, then, as presented, PSM assumes the following joint distribution over s-marker sequences $\vec{y}_i^{\,k}$ and memberships $z_i^{\,k}$:

$$\prod_{k=1}^{K} p\left(\vec{y}_i^{\,k} \mid z_i^{\,k}, \Theta\right) p\left(z_i^{\,k}\right). \tag{15}$$

We can replace the fully factored distribution over $z_i^{\,1}, \ldots, z_i^{\,K}$ with a complete joint $p\left(z_i^{\,1}, \ldots, z_i^{\,K}\right)$ to induce correlations across s-marker types. The posterior distribution over $z_i^{\,1}, \ldots, z_i^{\,K}$ can be inspected to discover clusters defined over multiple s-markers.

A simpler alternative when full posterior inference over group memberships is not desired, is to represent each individual using a vector of categorical variables indicating cluster membership for each s-marker as inferred by PSM (e.g. [tss-type-1, pfvc-type-3, dlco-type-2, ...]). Then, using a distance-based clustering method, such as hierarchical agglomerative clustering (HAC), clusters across multiple s-marker types can be obtained. Due to space constraints, we omit analyses of joint clusters.

Discussion

This paper presents the Probabilistic Subtyping Model, a model for clustering time series of clinical markers obtained from routine visits to identify homogeneous patient subgroups. We introduce the concept of nuisance variability—effects due to factors such as age and co-existing conditions—that affect the observed clinical test results, but are unrelated to the underlying disease mechanism. PSM introduces a framework for modeling these sources of nuisance variability to uncover disease activity subtypes that are likely candidates for endotyping.

PSM identified novel subtypes for each of the four key organ systems affected in patients with scleroderma. To identify whether these subtypes are indeed distinct disease mechanisms, our ongoing clinical experiments are investigating whether there exist molecular differences consistent with these subtypes.

PSM provides a new tool to clinicians for studying candidate subtypes, especially useful in complex, chronic diseases such as neuropsychiatric and autoimmune diseases (e.g., scleroderma is one of 80 autoimmune diseases) known to be heterogeneous and poorly understood. As electronic medical records continue to amass data from routine visits, there will be growing need for such tools to refine our characterization of what constitutes a disease.

A natural direction for follow up is to identify ways to quantify effects of variability in treatment protocols across individuals on the estimated subtypes.

Acknowledgements. This work was supported by a Google Faculty Research Award, NSF award #1418590, and a Whiting School of Engineering Centennial Fellowship.

References

Anderson, G. P. 2008. Endotyping asthma: new insights into key pathogenic mechanisms in a complex, heterogeneous disease. *The Lancet* 372(9643):1107–1119.

Bardenet, R.; Kégl, B.; et al. 2010. Surrogating the surrogate: accelerating gaussian-process-based global optimization with a mixture cross-entropy algorithm. In *Proceedings of the 27th International Conference on Machine Learning*, 55–62.

Barr, S. G.; ZonanaNacach, A.; Magder, L. S.; and Petri, M. 1999. Patterns of disease activity in systemic lupus erythematosus. *Arthritis and Rheumatism* 42(12):2682–2688.

Beretta, L.; Caronni, M.; Raimondi, M.; Ponti, A.; Viscuso, T.; Origgi, L.; and Scorza, R. 2007. Oral cyclophosphamide improves pulmonary function in scleroderma patients with fibrosing alveolitis: experience in one centre. *Clinical rheumatology* 26(2):168–172.

Blumenthal, D. 2009. Stimulating the adoption of health

information technology. *New England Journal of Medicine* 360(15):1477–1479.

Camilli, A. E.; Burrows, B.; Knudson, R. J.; Lyle, S. K.; and Lebowitz, M. D. 1987. Longitudinal changes in forced expiratory volume in one second in adults. effects of smoking and smoking cessation. *The American review of respiratory disease* 135(4):794–799.

Chang, H. Y.; Clark, J. M.; and Weiner, J. P. 2011. Morbidity trajectories as predictors of utilization: multi-year disease patterns in taiwan's national health insurance program. *Medical care* 49(10):918–923.

Chen, D. P.; Weber, S. C.; Constantinou, P. S.; Ferris, T. A.; Lowe, H. J.; and Butte, A. J. 2007. Clinical arrays of laboratory measures, or "clinarrays", built from an electronic health record enable disease subtyping by severity. *AMIA Annual Symposium Proceedings Archive* 115–119.

Chen, M.; Zaas, A.; Woods, C.; Ginsburg, G. S.; Lucas, J.; Dunson, D.; and Carin, L. 2011. Predicting viral infection from high-dimensional biomarker trajectories. *Journal of the American Statistical Association* 106(496).

De Keulenaer, G. W., and Brutsaert, D. L. 2009. The heart failure spectrum: time for a phenotype-oriented approach. *Circulation* 119(24):3044–3046.

Doshi-Velez, F.; Ge, Y.; and Kohane, I. 2014. Comorbidity clusters in autism spectrum disorders: an electronic health record time-series analysis. *Pediatrics* 133(1):e54–63.

Fox, E. B.; Sudderth, E. B.; Jordan, M. I.; and Willsky, A. S. 2011. Joint modeling of multiple related time series via the beta process. *arXiv preprint arXiv:1111.4226*.

Gundlapalli, A. V.; South, B. R.; Phansalkar, S.; Kinney, A. Y.; Shen, S.; Delisle, S.; Perl, T.; and Samore, M. H. 2008. Application of natural language processing to va electronic health records to identify phenotypic characteristics for clinical and research purposes. *Summit on translational bioinformatics* 2008:36.

Ho, J. C.; Ghosh, J.; and Sun, J. 2014. Marble: high-throughput phenotyping from electronic health records via sparse nonnegative tensor factorization. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, 115–124. ACM.

Hoshida, Y.; Brunet, J.-P.; Tamayo, P.; Golub, T. R.; and Mesirov, J. P. 2007. Subclass mapping: identifying common subtypes in independent disease data sets. *PloS one* 2(11):e1195.

Kho, A. N.; Pacheco, J. A.; Peissig, P. L.; Rasmussen, L.; Newton, K. M.; Weston, N.; Crane, P. K.; Pathak, J.; Chute, C. G.; Bielinski, S. J.; Kullo, I. J.; Li, R.; Manolio, T. A.; Chisholm, R. L.; and Denny, J. C. 2011. Electronic medical records for genetic research: results of the emerge consortium. *Science translational medicine* 3(79):79re1.

Kohane, I. S. 2011. Using electronic health records to drive discovery in disease genomics. *Nature Reviews Genetics* 12(6):417–428.

Lasko, T. A.; Denny, J. C.; and Levy, M. A. 2013. Computational phenotype discovery using unsupervised feature learning over noisy, sparse, and irregular clinical data. *PloS one* 8(6):e66341.

Lewis, S. J.; Foltynie, T.; Blackwell, A. D.; Robbins, T. W.; Owen, A. M.; and Barker, R. A. 2005. Heterogeneity of parkinson's disease in the early clinical stages using a data driven approach. *Journal of neurology, neurosurgery, and psychiatry* 76(3):343–348.

Listgarten, J.; Neal, R. M.; Roweis, S. T.; Puckrin, R.; and Cutler, S. 2006. Bayesian detection of infrequent differences in sets of time series with shared structure. In *Advances in neural information processing systems*, 905–912.

Lötvall, J.; Akdis, C. A.; Bacharier, L. B.; Bjermer, L.; Casale, T. B.; Custovic, A.; Lemanske Jr, R. F.; Wardlaw, A. J.; Wenzel, S. E.; and Greenberger, P. A. 2011. Asthma endotypes: a new approach to classification of disease entities within the asthma syndrome. *Journal of Allergy and Clinical Immunology* 127(2):355–360.

Marlin, B. M.; Kale, D. C.; Khemani, R. G.; and Wetzel, R. C. 2012. Unsupervised pattern discovery in electronic health care data using probabilistic clustering models. In *Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium*, 389–398. ACM.

McCulloch, C.; Lin, H.; Slate, E.; and Turnbull, B. 2002. Discovering subpopulation structure with latent class mixed models. *Statistics in medicine* 21(3):417–429.

Muthén, B., and Shedden, K. 1999. Finite mixture modeling with mixture outcomes using the em algorithm. *Biometrics* 55(2):463–469.

Nagin, D. S., and Odgers, C. L. 2010. Group-based trajectory modeling in clinical research. *Annual Review of Clinical Psychology* 6:109–138.

Ross, J., and Dy, J. 2013. Nonparametric mixture of gaussian processes with constraints. In *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*, 1346–1354.

Saria, S.; Duchi, A.; and Koller, D. 2011. Discovering deformable motifs in continuous time series data. In *International Joint Conference on Artificial Intelligence*, volume 22, 1465.

Saria, S.; Koller, D.; and Penn, A. 2010. Learning individual and population level traits from clinical temporal data. In *Proc. Neural Information Processing Systems (NIPS), Predictive Models in Personalized Medicine workshop*.

Shea, S., and Hripcsak, G. 2010. Accelerating the use of electronic health records in physician practices. *New England Journal of Medicine* 362(3):192–195.

State, M. W., and Sestan, N. 2012. Neuroscience. the emerging biology of autism spectrum disorders. *Science (New York, N.Y.)* 337(6100):1301–1303.

Steen, V., and Medsger, T. A. 2003. Predictors of isolated pulmonary hypertension in patients with systemic sclerosis and limited cutaneous involvement. *Arthritis & Rheumatism* 48(2):516–522.

Varga, J.; Denton, C. P.; and Wigley, F. M. 2012. *Scleroderma: From pathogenesis to comprehensive management*. Springer.